



*engineering
proceedings*

Proceedings Reprint

The 9th International Conference on Time Series and Forecasting

Volume II

Edited by
Ignacio Rojas, Hector Pomares, Luis Javier Herrera,
Fernando Rojas and Olga Valenzuela

mdpi.com/journal/engproc



**The 9th International Conference on
Time Series and Forecasting-Volume II**

The 9th International Conference on Time Series and Forecasting-Volume II

Editors

Ignacio Rojas

Hector Pomares

Luis Javier Herrera

Fernando Rojas

Olga Valenzuela



Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Editors

Ignacio Rojas
University of Granada
Granada, Spain

Hector Pomares
University of Granada
Granada, Spain

Luis Javier Herrera
University of Granada
Granada, Spain

Fernando Rojas
University of Granada
Granada, Spain

Olga Valenzuela
University of Granada
Granada, Spain

Editorial Office

MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Proceedings published online in the open access journal *Engineering Proceedings* (ISSN 2673-4591) (available at: <https://www.mdpi.com/2673-4591/39/1>).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

Volume II

ISBN 978-3-0365-9730-0 (Hbk)

ISBN 978-3-0365-9731-7 (PDF)

doi.org/10.3390/books978-3-0365-9731-7

Set

ISBN 978-3-0365-9726-3 (Hbk)

ISBN 978-3-0365-9727-0 (PDF)

© 2024 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license.

Contents

Grzegorz Dudek

Combining Forecasts of Time Series with Complex Seasonality Using LSTM-Based Meta-Learning
Reprinted from: *Eng. Proc.* **2023**, 39, 53, doi:10.3390/engproc2023039053 1

Jaime B. Fernandez, Suzanne Little and Noel E. O'Connor

Moving Object Path Prediction in Traffic Scenes Using Contextual Information
Reprinted from: *Eng. Proc.* **2023**, 39, 54, doi:10.3390/engproc2023039054 11

Amir Aieb, Antonio Liotta and Ismahen Kadri

Downscaling Fusion Model for CMIP5 Rainfall Projection under RCP Scenarios: The Case of Trentino-Alto Adige
Reprinted from: *Eng. Proc.* **2023**, 39, 55, doi:10.3390/engproc2023039055 25

Gerardo Covarrubias and Xuedong Liu

Resolution of Systems of Difference Equations and Its Implications for the VAR Model
Reprinted from: *Eng. Proc.* **2023**, 39, 56, doi:10.3390/engproc2023039056 39

Álvaro Quintanar, Rubén Izquierdo, Ignacio Parra and David Fernández-Llorca

Goal-Oriented Transformer to Predict Context-Aware Trajectories in Urban Scenarios
Reprinted from: *Eng. Proc.* **2023**, 39, 57, doi:10.3390/engproc2023039057 47

Martín Solís and Luis-Alexander Calvo-Valverde

A Proposal of Transfer Learning for Monthly Macroeconomic Time Series Forecast
Reprinted from: *Eng. Proc.* **2023**, 39, 58, doi:10.3390/engproc2023039058 59

Abdulnasser Hatemi-J and Alan Mustafa

A Simulation Package in VBA to Support Finance Students for Constructing Optimal Portfolios
Reprinted from: *Eng. Proc.* **2023**, 39, 59, doi:10.3390/engproc2023039059 67

Jackson B. Renteria-Mena, Douglas Plaza and Eduardo Giraldo

Multivariable NARX Based Neural Networks Models for Short-Term Water Level Forecasting
Reprinted from: *Eng. Proc.* **2023**, 39, 60, doi:10.3390/engproc2023039060 77

Maria Koshkareva and Anton Kovantsev

Enhancement of Consumption Forecasting by Customers' Behavioral Predictability Segregation
Reprinted from: *Eng. Proc.* **2023**, 39, 61, doi:10.3390/engproc2023039061 87

Tamara Kuvek, Ivica Pervan and Maja Pervan

Improving the Accuracy of Firm Failure Forecasting Using Non-Financial Variables: The Case of Croatian SME
Reprinted from: *Eng. Proc.* **2023**, 39, 62, doi:10.3390/engproc2023039062 97

Shuvro Ahmed, Joy Karmoker, Rajesh Mojumder, Md. Mahmudur Rahman, Md. Golam Rabiul Alam and Md Tanzim Reza

Hyperautomation in Super Shop Using Machine Learning
Reprinted from: *Eng. Proc.* **2023**, 39, 63, doi:10.3390/engproc2023039063 105

Marta Ferreira

Measuring Extremal Clustering in Time Series
Reprinted from: *Eng. Proc.* **2023**, 39, 64, doi:10.3390/engproc2023039064 115

Arnold Hien, Nicolas Beldiceanu, Claude-Guy Quimper and María-I. Restrepo Automata Based Multivariate Time Series Analysis for Anomaly Detection over Sliding Time Windows Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 65, doi:10.3390/engproc2023039065	123
Ana García-Burgos, Beatriz González-Alzaga, María José Giménez-Asensio, Marina Lacasaña, Nuria Rico-Castro and Desirée Romero-Molina Growth Curves Modelling and Its Application Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 66, doi:10.3390/engproc2023039066	133
Mahdy Kouka and David Cuesta-Frau Slope Entropy Characterisation: Adding Another Interval Parameter to the Original Method Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 67, doi:10.3390/engproc2023039067	137
Maryam Movahedifar, Hossein Hassani and Mahdi Kalantari Recurrent Forecasting in Singular Spectrum Decomposition Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 68, doi:10.3390/engproc2023039068	143
Andreas Heller, Peter Glösekötter, Lukas Buntkiel, Sebastian Reinecke and Sven Annas Sim-to-Real Transfer in Deep Learning for Agitation Evaluation of Biogas Power Plants Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 69, doi:10.3390/engproc2023039069	155
Gueï Cyrille Okou and Amine Amar Modeling Contagion of Financial Markets: A GARCH-EVT Copula Approach Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 70, doi:10.3390/engproc2023039070	165
Rodrigo Hernandez-Mazariegos, Jose Ortiz-Bejar and Jesus Ortiz-Bejar Evaluation of Heuristics for Taken's Theorem Hyper-Parameters Optimization in Time Series Forecasting Tasks Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 71, doi:10.3390/engproc2023039071	173
Kalle Saastamoinen, Antti Rissanen, Juho Suni, Juho Hyttinen, Petteri Paakkunainen and Aaro Liakka Simulation of the Queuing Situation of Patients at a Health Center Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 72, doi:10.3390/engproc2023039072	183
Georgina González-González, Jesús Cerezo-Román and Guillermo Satamaría-Bonfil Development of Methodology for the Evaluation of Solar Energy through Hybrid Models for the Energy Sector Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 73, doi:10.3390/engproc2023039073	189
Antônio Augusto Rodrigues de Camargo and Mauri Aparecido de Oliveira Analysis of the Application of Different Forecasting Methods for Time Series in the Context of the Aeronautical Industry Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 74, doi:10.3390/engproc2023039074	199
John Anderson Torres Mosquera, Carlos Julio Vidal Holguín, Alexander Kressner and Edwin Loaiza Acuña Forecasting System for Inbound Logistics Material Flows at an International Automotive Company Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 75, doi:10.3390/engproc2023039075	215
Cinzia Graziani, Annalisa Lucarelli, Maurizio Lucarelli, Emilia Matera and Andrea Spizzichino Integrating Seasonal Adjustment Approaches of Official Surveys on Labor Supply and Demand Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 76, doi:10.3390/engproc2023039076	229

Diana Manjarrés, Erik Maqueda and Itziar Landa-Torres Online Pentane Concentration Prediction System Based on Machine Learning Techniques Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 77, doi:10.3390/engproc2023039077	239
Federico Delrio, Vincenzo Randazzo, Giansalvo Cirrincione and Eros Pasero Non-Invasive Arterial Blood Pressure Estimation from Electrocardiogram and Photoplethysmography Signals Using a Conv1D-BiLSTM Neural Network Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 78, doi:10.3390/engproc2023039078	247
Mateusz Buczyński, Marcin Chlebus, Katarzyna Kopczewska and Marcin Zajenkowski Financial Time Series Models—Comprehensive Review of Deep Learning Approaches and Practical Recommendations Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 79, doi:10.3390/engproc2023039079	257
Aner Martinez-Soto, Johannes Fürle and Alexander Zipf Urban Heat Island Intensity Prediction in the Context of Heat Waves: An Evaluation of Model Performance Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 80, doi:10.3390/engproc2023039080	271
Fernando García, Francisco Guijarro, Javier Oliver and Rima Tamošiūnienė Foreign Exchange Forecasting Models: ARIMA and LSTM Comparison Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 81, doi:10.3390/engproc2023039081	277
Kento Kotera, Akihiro Yamaguchi and Ken Ueno Learning Local Patterns of Time Series for Anomaly Detection Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 82, doi:10.3390/engproc2023039082	285
Julián David Pastrana-Cortés, David Augusto Cardenas-Peña, Mauricio Holguín-Londoño, Germán Castellanos-Dominguez and Álvaro Angel Orozco-Gutiérrez Multi-Output Variational Gaussian Process for Daily Forecasting of Hydrological Resources Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 83, doi:10.3390/engproc2023039083	295
Paola Barba, Nely Pérez-Méndez, Javier Ramírez-Zelaya, Belén Rosado, Vanessa Jiménez and Manuel Berrocoso Geodynamic Modeling in Central America Based on GNSS Time Series Analysis—Special Case: The Nicoya Earthquake (Costa Rica, 2012) Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 84, doi:10.3390/engproc2023039084	307
Christos Katris Investigation of FIBA World Cup 2019: Evidence Using Advanced Statistical Analysis and Quantitative Tools Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 85, doi:10.3390/engproc2023039085	315
Olena Rayevnyeva, Kostyantyn Stryzhychenko and Silvia Matúšová Impact of Migration Processes on GDP Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 86, doi:10.3390/engproc2023039086	325
Kalle Saastamoinen, Tuomas E. Alanen, Pasi Leskinen, Kai Pihlainen and Joona Jehkonen Defining Sports Performance by Using Automated Machine Learning System Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 87, doi:10.3390/engproc2023039087	335
Daniel Ullrich and Sarah Diefenbach Forecasting Transitions in Digital Society: From Social Norms to AI Applications Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 88, doi:10.3390/engproc2023039088	343

Rodrigo B. Ventura, Filipe M. Santos, Ricardo M. Magalhães, Cátia M. Salgado, Vera Dantas, Matilde V. Rosa, et al. Forecasting Neonatal Mortality in Portugal Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 89, doi:10.3390/engproc2023039089	353
Johannes Korte, Jan Martin Brockmann and Wolf-Dieter Schuh A Comparison between Successive Estimate of TVAR(1) and TVAR(2) and the Estimate of a TVAR(3) Process Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 90, doi:10.3390/engproc2023039090	361
Cristian Alejandro Blanco-Martínez, David Augusto Cardenas-Peña, Mauricio Holguín-Londoño, Andrés Marino Álvarez-Meza and Álvaro Angel Orozco-Gutiérrez Approximation of Weymouth Equation Using Mathematical Programs with Complementarity Constraints for Natural Gas Transportation Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 91, doi:10.3390/engproc2023039091	371
Alejandro Polo-Molina, Eugenio F. Sánchez-Úbeda, José Portela, Rafael Palacios, Carlos Rodríguez-Morcillo, Antonio Muñoz, et al. Analyzing Mobility Patterns of Complex Chronic Patients Using Wearable Activity Trackers: A Machine Learning Approach Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 92, doi:10.3390/engproc2023039092	381
Manuchehr Aminian and Michael Kirby Reduced Order Modeling with Skew-Radial Basis Functions for Time Series Prediction Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 93, doi:10.3390/engproc2023039093	393
Syed Kabir, David Wood and Simon Waller A Deep Learning Model for Generalized Surface Water Flooding across Multiple Return Periods Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 94, doi:10.3390/engproc2023039094	401
Joseph L. Breeden and Yevgeniya Leonova Macroeconomic Adverse Selection in Machine Learning Models of Credit Risk Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 95, doi:10.3390/engproc2023039095	411
Sachin Shetty, Valentina Gori, Gianni Bagni and Giacomo Veneri Sensor Virtualization for Anomaly Detection of Turbo-Machinery Sensors—An Industrial Application Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 96, doi:10.3390/engproc2023039096	423
Bibiana Lanzilotta, Gabriela Mordecki, Pablo Tapie and Joaquín Torres Uncertainty and Business Cycle: An Empirical Analysis for Uruguay Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 97, doi:10.3390/engproc2023039097	431
Mandar Tabib, Kristoffer Skare, Endre Bruaset and Adil Rasheed Data-Driven Spatio-Temporal Modelling and Optimal Sensor Placement for a Digital Twin Set-Up Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 98, doi:10.3390/engproc2023039098	447
Asko Mononen, Ari Alamäki, Janne Kauttonen, Aarne Klemetti, Anu Passi-Rauste and Harri Ketamo Forecasted Self: AI-Based Careerbot-Service Helping Students with Job Market Dynamics Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 99, doi:10.3390/engproc2023039099	457
Jonathan D. Oladeji, Benita G. Zulch and Joseph A. Yacim Predictive Accuracy of Logit Regression for Data-Scarce Developing Markets: A Nigeria and South Africa Study Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 100, doi:10.3390/engproc2023039100	465

Samuel J. Edwards, Matthew D. Collette and Armin W. Troesch Extreme Characteristics of a Stochastic Non-Stationary Duffing Oscillator Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 102, doi:10.3390/engproc2023039102	473
Javier Ramírez-Zelaya, Vanessa Jiménez, Paola Barba, Belén Rosado, Jorge Gárate and Manuel Berrocoso Treatment and Analysis of Multiparametric Time Series from a Seismogeodetic System for Tectonic Monitoring of the Gulf of Cadiz, Spain Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 46, doi:10.3390/engproc2023039046	491
Paola Barba, Javier Ramírez-Zelaya, Vanessa Jiménez, Belén Rosado, Elena Jaramillo, Mario Moreno and Manuel Berrocoso Tropospheric and Ionospheric Modeling Using GNSS Time Series in Volcanic Eruptions (La Palma, 2021) Reprinted from: <i>Eng. Proc.</i> 2023 , 39, 47, doi:10.3390/engproc2023039047	501

Combining Forecasts of Time Series with Complex Seasonality Using LSTM-Based Meta-Learning [†]

Grzegorz Dudek

Electrical Engineering Faculty, Czestochowa University of Technology, Al. AK 17, 42-200 Czestochowa, Poland; grzegorz.dudek@pcz.pl

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: In this paper, we propose a method for combining forecasts generated by different models based on long short-term memory (LSTM) ensemble learning. While typical approaches for combining forecasts involve simple averaging or linear combinations of individual forecasts, machine learning techniques enable more sophisticated methods of combining forecasts through meta-learning, leading to improved forecasting accuracy. LSTM's recurrent architecture and internal states offer enhanced possibilities for combining forecasts by incorporating additional information from the recent past. We define various meta-learning variants for seasonal time series and evaluate the LSTM meta-learner on multiple forecasting problems, demonstrating its superior performance compared to simple averaging and linear regression.

Keywords: ensemble forecasting; LSTM; machine learning; multiple seasonal patterns; short-term load forecasting

1. Introduction

Real-world time series can exhibit various complex properties such as time-varying trends, multiple seasonal patterns, random fluctuations, and structural breaks. Given this complexity, it can be challenging to identify a single best model to accurately approximate the underlying data-generating process [1]. To address this issue, a common approach is to combine multiple forecasting models to capture the multiple drivers of the data-generating process and mitigate uncertainties regarding model form and parameter specification [2]. This approach, known as ensemble forecasting or combining forecasts, has been shown to be effective in improving the accuracy and reliability of time series forecasts. By combining forecasts, the aim is to take advantage of the strengths of multiple models and reduce the impact of their individual weaknesses.

There are several potential explanations for the strong performance of forecast combinations. Firstly, by combining forecasts, the resulting ensemble can capture a broader range of information and better handle the forecasting problem complexity. It can leverage the strengths of individual models, as each model may capture different aspects of the underlying data-generating process. Therefore, the resulting ensemble can incorporate partial and incompletely overlapping information, leading to improved accuracy and robustness. Secondly, in the presence of structural breaks and other instabilities, combining forecasts from models with different degrees of misspecification and adaptability can mitigate the problem. This is because individual models may perform well under certain conditions but poorly under others, and by combining them, the ensemble can better handle a range of potential scenarios [3]. Finally, forecast combinations can improve stability compared to using a single model, as the ensemble is less sensitive to the idiosyncrasies of individual models. This means that the resulting forecasts are less likely to be influenced by outliers or errors in individual models, leading to more reliable predictions.

Citation: Dudek, G. Combining Forecasts of Time Series with Complex Seasonality Using LSTM-Based Meta-Learning. *Eng. Proc.* **2023**, *39*, 53. <https://doi.org/10.3390/engproc2023039053>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 5 July 2023



Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

In a classical way, by combining the predictions from multiple models, the resulting ensemble prediction can be thought of as an average of the individual predictions. The variance of the average of multiple independent random variables is typically lower than the variance of a single random variable, assuming that the individual predictions are diverse. Therefore, a key issue in ensemble learning is ensuring diversity among the individual models being combined. If the models are too similar, the ensemble may not be able to capture the full range of possible outcomes and may not improve predictive performance. In this work, we ensure high diversity among models by using non-interfering models with different operating principles and architectures, including statistical, machine learning (ML), and hybrid models (see Section 3.2).

A simple arithmetic average of forecasts based on equal weights is a popular and surprisingly robust combination rule, outperforming more complicated weighting schemes in many cases [4,5]. Other strategies, such as using the median, mode, trimmed means, and winsorized means, are also applied [6]. To differentiate weights assigned to individual models, linear regression can be used, where the vector of past observations is the response variable and the matrix of past individual forecasts is the predictor variable. Combination weights can be estimated using ordinary least squares. The weights can reflect individual models' performance on historical data [7]. Time-varying weights can be used to improve forecasting ability in the presence of instabilities, and principal components regression can be used as a solution for multicollinearity [8]. Weights can also be derived from information criteria such as AIC [9].

Linear combination approaches assume a linear dependence between constituent forecasts and the variable of interest, and may not result in the best forecast, especially if the individual forecasts come from nonlinear models or if the true relationship between base forecasts and the target has a nonlinear form [10]. In contrast, ML models can combine the base forecasts nonlinearly using a stacking procedure.

Stacking is an ensemble ML algorithm that learns how to best combine predictions from multiple models, using the concept of meta-learning to boost forecasting accuracy beyond that achieved by the individual models. Neural networks (NNs) are often used in stacking to estimate the nonlinear mapping between the target value and its forecasts produced by multiple models [11]. The power of ensemble learning for forecasting was demonstrated in [12], where several meta-learning approaches were evaluated on a large and diverse set of time series data. Ensemble methods were found to provide a benefit in overall forecasting accuracy, with simple ensemble methods leading to good results on average. However, there was no single meta-learning method that was suitable for all time series.

The main contributions of this study can be summarized in the following three aspects:

1. A meta-learning approach based on LSTM is proposed for combining forecasts. This approach incorporates past information accumulated in the internal states, improving accuracy, especially in cases where there is a temporal relationship between base forecasts for successive time points.
2. Various meta-learning variants for time series with multiple seasonal patterns are proposed, such as the use of the full training set, including base forecasts for successive time points, and the use of selected training points that reflect the seasonal structure of the data.
3. Extensive experiments are conducted on 35 time series with triple seasonality using 16 base models to validate the efficacy of the proposed approach. The experimental results demonstrate the high performance of the LSTM meta-learner and its potential to combine forecasts more accurately than simple averaging and linear regression methods.

The remainder of this work is structured as follows. Section 2 presents the proposed LSTM meta-model and introduces both the global and local meta-learning variants. Section 3 provides application examples for time series with complex seasonality and

discusses the results obtained from the conducted experiments. Finally, in Section 4, we conclude our work by summarizing the key findings and contributions.

2. LSTM for Combining Forecasts

The problem of forecast combinations refers to the task of finding regression function f that aggregates the forecasts for time t produced by n forecasting models. The function can use all the available information up to time $t - h$, where h is a forecast horizon, but in this study, we limit this information to the base forecasts expressed by vector $\hat{\mathbf{y}}_t = [\hat{y}_{1,t}, \dots, \hat{y}_{n,t}]$. The combined forecast is $\tilde{y}_t = f(\hat{\mathbf{y}}_t; \theta_t)$, where θ_t is a vector of meta-model parameters.

The model learns using training set $\Phi = \{\hat{\mathbf{y}}_\tau, y_\tau\}_{\tau \in \Xi}$, where y_τ is a target value and Ξ is a set of selected time indexes from interval $T = 1, \dots, t - h$ (selection of this set is considered in Section 2.2).

The class of regression functions f encompasses both linear and nonlinear mappings, as well as series-specific and cross-learning mappings. In the latter approach, the parameters of the function are selected through a learning process over multiple time series, which enhances the generalization capability of the model. Furthermore, the parameters can either be static or time-varying throughout the forecasting horizon. To maximize the performance of the ensemble, we adopt an approach where we learn the meta-model parameters for each forecasting task individually, using a specific training set for each task (see Section 2.2).

2.1. LSTM Model

LSTM is a modern recurrent NN that incorporates gating mechanisms [13]. This NN architecture was specifically designed to handle sequential data and is capable of learning short and long-term relationships in time series [14]. LSTM is composed of recurrent cells that can maintain their internal states over time, i.e., cell state \mathbf{c} and hidden state \mathbf{h} . These cells are regulated by nonlinear gating mechanisms that control the flow of information within the cell, allowing it to adapt to the dynamics of the current process.

In our implementation, the LSTM network consists of two layers: the LSTM layer and the linear layer, see Figure 1. The LSTM layer is responsible for approximating temporal nonlinear dependencies in sequential data and generating state vectors. On the other hand, the linear layer converts hidden state vector \mathbf{h} into the output value. The aggregation function implemented in the LSTM network can be written as:

$$f(\hat{\mathbf{y}}_t) = \mathbf{v}^T \mathbf{h}_t(\hat{\mathbf{y}}_t) + v_0 \tag{1}$$

$$\mathbf{h}_t(\hat{\mathbf{y}}_t) = \text{LSTM}(\hat{\mathbf{y}}_t, \mathbf{c}_{t-1}, \mathbf{h}_{t-1}; \mathbf{w}) \in \mathbb{R}^m \tag{2}$$

where \mathbf{w} and \mathbf{v} are the weights of the LSTM and linear layers, respectively.

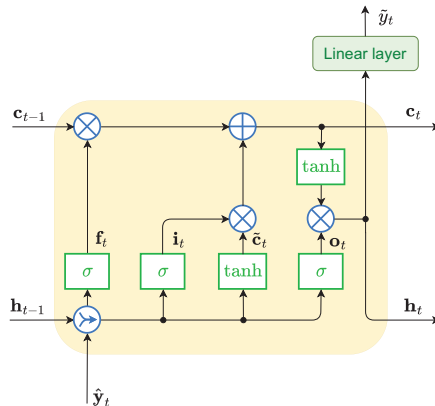


Figure 1. LSTM model.

The number of nodes in each gate, m , is the most critical hyperparameter. It determines the amount of information stored in the states. For more intricate temporal relationships, a higher number of nodes is necessary.

In contrast to non-recurrent ML models such as feed-forward NNs, tree-based models, and support vector regression, to calculate output \hat{y}_t , LSTM uses not only the information included in the base forecasts for time t , \hat{y}_t , but also in the base forecasts for previous time steps, $t - 1, t - 2, \dots$. This is achieved through states \mathbf{c}_{t-1} and \mathbf{h}_{t-1} , which accumulate information from the past steps.

2.2. Meta-Learning Variants

The forecasting models generate forecasts for the successive time points $T = 1, \dots, t - h$. To obtain an ensemble forecast for time t , we can train the meta-model using all available data from the historical period, i.e., $\Xi = T$, which is referred to as the global approach. Using this method, the model can utilize all available information to generate a forecast for the current time point.

In local learning, we restrict the training sequence to the last k points, i.e., $\Xi = t - h - k, \dots, t - h$, allowing the LSTM to model the relationship for the query pattern \hat{y}_t based on the most recent sequence of length k . We refer to this approach as v1.

When ensembling seasonal time series, training the LSTM model on points from the same phase of the cycle as the forecasted point can improve forecast accuracy. In this approach, the training set consists of points $\Xi = \{t - ks_1, t - (k - 1)s_1, \dots, t - s_1\}$, where s_1 denotes the period of the seasonal cycle and k is a predefined size of the training set. It is worth noting that this training set retains the time structure of the data, but simplifies it by only including points that are in the same phase of the seasonal cycle as the forecasted point. We refer to this approach as v2.

In the case of double seasonality with periods s_1 and s_2 (assuming that s_2 is a multiple of s_1), we can create the training set by selecting points from the same phase of both seasonal patterns as the forecasted point. Specifically, the training set is composed of points $\Xi = \{t - ks_2, t - (k - 1)s_2, \dots, t - s_2\}$. We refer to this approach as v3. Figure 2 visualizes the training target points for each variant of LSTM learning.

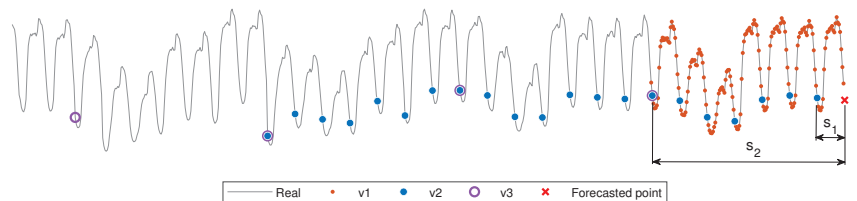


Figure 2. Selection of training points for LSTM.

Note that approaches v2 and v3 remove the training points that are not in the same phase as the forecasted point. This simplifies the relationship between the new training points and the forecasted point, making it easier to model. However, this simplification comes at the cost of potentially losing some of the information related to the seasonal patterns that occur outside of the selected phase. Therefore, it is important to carefully consider which approach to use depending on the specific characteristics of the data.

3. Experimental Study

We evaluate the performance of our proposed approach, combining forecasts generated by 16 forecasting models described in Section 3.2. The forecasting problem is short-term load forecasting for 35 European countries.

3.1. Data, Forecasting Problem and Research Design

We use the real-world data collected from the ENTSO-E repository (www.entsoe.eu/data/power-stats accessed on 6 April 2016). The dataset includes hourly electricity loads spanning from 2006 to 2018, representing 35 European countries. It offers a diverse set of time series, each exhibiting unique properties such as distinct levels and trends, variance stability over time, intensity and regularity of seasonal fluctuations spanning different periods (annual, weekly, and daily), and varying degrees of random fluctuations.

The forecasting models were optimized using data from 2006 to 2017 and applied to generate hourly forecasts for the year 2018, day by day. To evaluate the performance of the combining model, 100 hours for each country were chosen from the second half of 2018 (evenly spaced across the period) and the forecasts for each of these hours were combined using LSTM. The LSTM model was trained separately for each selected hour, with preceding data spanning from 1 January 2018 up to the hour preceding the forecasted hour ($h = 1$) used for optimization and training across three variants (v1, v2, and v3). This resulted in a total of 10,500 training sessions ($35 \cdot 100 \cdot 3$). In variant v2, we assumed daily seasonality period $s_1 = 24$ h, while in variant v3 we assumed weekly period $s_2 = 7 \cdot 24 = 168$ h.

This study utilized Matlab implementation of the LSTM model. Some LSTM hyperparameters were set to default values, while others were determined through experimentation. The latter include the number of nodes $m = 128$, and the number of epochs—200.

As performance metrics, the following measures were used: MAPE—mean absolute percentage error, MdAPE—median of absolute percentage error, MSE—mean square error, MPE—mean absolute percentage error, and StdPE—standard deviation of percentage error.

3.2. Forecasting Models

As the base forecasting models, we use a set of statistical models and classical ML models, as well as recurrent, deep, and hybrid NN architectures from [15]:

- ARIMA—auto-regressive integrated moving average model,
- ETS—exponential smoothing model,
- Prophet—modular additive regression model with nonlinear trend and seasonal components,
- N-WE—Nadaraya–Watson estimator,
- GRNN—general regression NN,
- MLP—perceptron with a single hidden layer and sigmoid nonlinearities,
- SVM—linear epsilon insensitive support vector machine (ϵ -SVM),
- LSTM—long short-term memory,
- ANFIS—adaptive neuro-fuzzy inference system,
- MTGNN—graph NN for multivariate time series forecasting,
- DeepAR—autoregressive recurrent NN model for probabilistic forecasting,
- WaveNet—autoregressive deep NN model combining causal filters with dilated convolutions,
- N-BEATS—deep NN with hierarchical doubly residual topology,
- LGBM—Light Gradient-Boosting Machine,
- XGB—eXtreme Gradient-Boosting algorithm,
- cES-adRNN—contextually enhanced hybrid and hierarchical model combining ETS and dilated RNN with attention mechanism.

3.3. Results and Discussion

Table 1 shows the forecasting quality metrics for the base forecasting models. Note the significant difference in results between the various models, with MAPE ranging from 1.70 for cES-adRNN to 3.83 for Prophet. The overall mean MAPE across all models was 2.53.

Table 2 shows forecasting quality metrics for different ensemble approaches. Mean and Median are just the mean and median of 16 forecasts produced by the base models. LinReg is a linear combination of these forecasts with weights estimated on the training samples $\Xi = T$. As can be seen from Table 2, the most accurate approach is variant v1

of LSTM for $k = 168$. This variant, which involves meta-learning on the full sequence restricted to the last 168 points, provided the most accurate results as measured by MAPE, MdAPE, and MSE errors. Note the significant difference in errors between this variant and the second most accurate ensembling method, LinReg, which achieved about 5% in MAPE and 35% in MSE.

Table 1. Forecasting quality metrics for the base models.

	MAPE	MdAPE	MSE	MPE	StdPE
ARIMA	2.86	1.82	777,012	0.0556	4.60
ETS	2.83	1.79	710,773	0.1639	4.64
Prophet	3.83	2.53	1,641,288	−0.5195	6.24
N-WE	2.12	1.34	357,253	0.0048	3.47
GRNN	2.10	1.36	372,446	0.0098	3.42
MLP	2.55	1.66	488,826	0.2390	3.93
SVM	2.16	1.33	356,393	0.0293	3.55
LSTM	2.37	1.54	477,008	0.0385	3.68
ANFIS	3.08	1.65	801,710	−0.0575	5.59
MTGNN	2.54	1.71	434,405	0.0952	3.87
DeepAR	2.93	2.00	891,663	−0.3321	4.62
WaveNet	2.47	1.69	523,273	−0.8804	3.77
N-BEATS	2.14	1.34	430,732	−0.0060	3.57
LGBM	2.43	1.70	409,062	0.0528	3.55
XGB	2.32	1.61	376,376	0.0529	3.37
cES-adRNN	1.70	1.10	224,265	−0.1860	2.57

Note that using the simplest method of combining forecasts, Mean or Median, resulted in significantly larger errors compared to LSTM v1. Unfortunately, variants v2 and v3, which excluded seasonality from the training sequence, were found to be inaccurate and did not perform well. This suggests that excluding seasonality from the training sequence could lead to the loss of important information related to the seasonal patterns in the data, resulting in deteriorated forecasting performance.

Figure 3 displays the MAPE boxplots for LSTM in three variants with varying lengths of the training sequence k . Additionally, the boxplots for the baseline methods, namely Mean, Median, and LinReg, are shown for comparison. As shown in the figure, LSTM in variants v2 and v3 are highly sensitive to the length of the training sequence. It achieved the lowest errors when trained on all available data points. Extending the training sequence may potentially further reduce errors. In contrast, for LSTM v1, the training sequences of length 168 hours (one week) provided the lowest errors.

MPE in Table 2 provides information about the forecast bias, which is the lowest for LinReg, but LSTM v1, with MPE = 0.0247, is in second place. It is worth noting that Mean and Median produce more biased forecasts. The lowest value of StdPE for LSTM v1 indicates the least dispersed predictions compared to other approaches for combining forecasts.

Table 2. Forecasting quality metrics for different ensemble approaches (best results in bold).

	Variant	MAPE	MdAPE	MSE	MPE	StdPE
Mean	-	1.91	1.23	316,943	−0.0775	3.11
Median	-	1.82	1.13	287,284	−0.0682	3.05
LinReg	-	1.63	1.11	213,428	0.0131	2.38
LSTM	v1, $k = 168$	1.55	1.09	139,667	0.0247	2.26
LSTM	v2, global	1.95	1.34	270,266	−0.1046	2.89
LSTM	v3, global	2.97	1.84	726,108	−0.3628	4.84

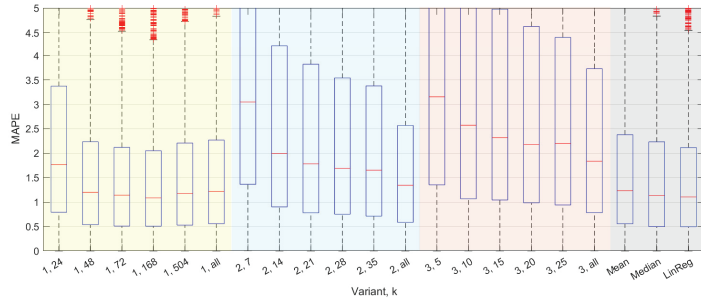


Figure 3. MAPE boxplots for the various ensemble variants.

Figure 4 depicts examples of forecasts for selected countries and test points. It is worth noting that LSTM v1 was able to achieve forecasts close to the target values, which were outside the interval of the base models’ forecasts (let us denote this interval for the i -th test point by Z_i) and despite the fact that no base model even came close to these targets (see test point no. 94 for FR and 99 for GB in Figure 4). One possible explanation for this ability of LSTM is the incorporation of additional information from the immediate past through internal states c and h (see (2)). LinReg, having no internal states, cannot use such information. Mean and Median approaches cannot even go beyond the interval Z_i .

To test the ability of LSTM v1 and LinReg to produce forecasts outside the interval Z_i , we counted the number of such cases out of the 3500 forecasts produced by each model. The results are shown in column N_1 of Table 3. Column N_2 counts how many of these N_1 cases concern the situation where the target value also lay outside the Z -interval, on the same side as the meta-model forecast. Column N_3 counts the number of cases out of N_1 for which the meta-model produces more accurate predictions than the Median approach. It is evident from Table 3 that LSTM generates many more forecasts outside of Z_i than LinReg. This may indicate better extrapolation properties of LSTM, but on the other hand, it may also suggest an increased susceptibility to overfitting.

Table 3. Extrapolation properties of LSTM v1 and LinReg.

	N_1	N_2	N_3
LinReg	48	13	27
LSTM v1	447	192	244

In summary, our research findings suggest that LSTM, as a meta-learner, exhibits sensitivity to the length of the training sequence and achieves optimal performance when trained in global mode. However, it is important to note that the overall performance also depends on the accuracy and correlation of the base forecasts. In this study, we did not delve into the analysis of interdependence between the base forecasts or select the optimal set of base models. These aspects present opportunities for further optimization and improvement in future research.

LSTM poses greater challenges compared to classical ML methods such as MLP or random forests. It involves a larger number of hyperparameters and parameters that need to be tuned, making the optimization and training process more complex. Additionally, LSTM typically requires a larger amount of data to achieve optimal performance due to its ability to capture intricate temporal dependencies. In local versions of training, where shorter training sequences are used, accurate predictions with LSTM can be challenging to obtain. This highlights the importance of having sufficient training data to effectively capture the underlying patterns and dynamics of the sequential data.

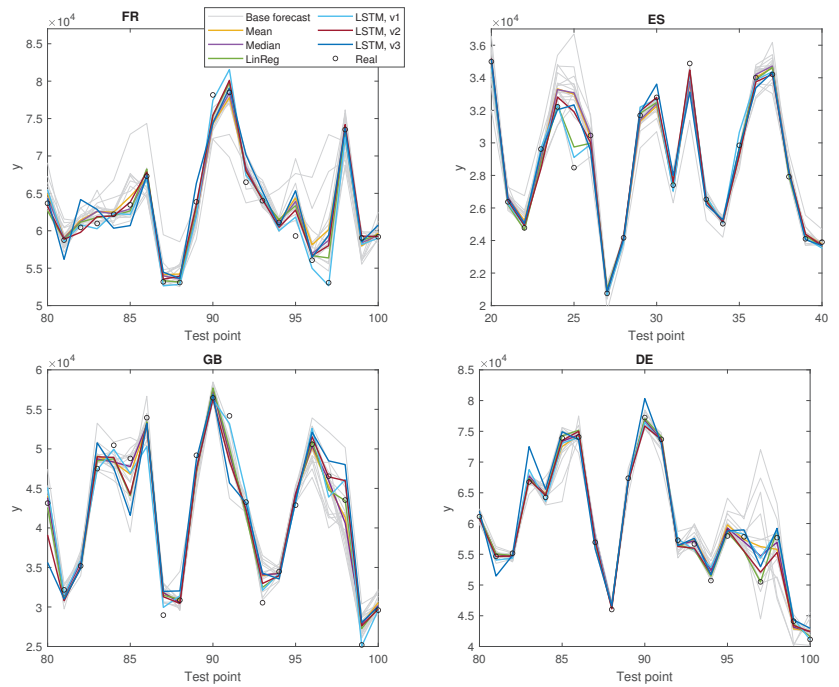


Figure 4. Base and ensemble forecasts.

4. Conclusions

This study proposes a meta-learning approach for combining forecasts based on LSTM, which has the potential to improve accuracy, particularly in cases where there is a temporal relationship between base forecasts. The study also proposes different variants of the approach for time series with multiple seasonal patterns.

The experimental results clearly demonstrate that the LSTM meta-learner outperforms simple averaging, median, and linear regression methods in terms of forecasting accuracy. In addition, LSTM has distinct advantages over non-recurrent ML models as it is capable of leveraging its internal states to model dependencies between base forecasts for consecutive time points and capture patterns in the sequential data.

Further studies could compare LSTM with other meta-learning approaches, such as feed-forward and randomized NNs, random forests, and boosted trees to determine which approach is best suited for a given forecasting problem. Moreover, selecting a pool of base models and controlling their diversity is an interesting topic that requires further investigation.

Funding: This researcher was supported by grant 020/RID/2018/19 “Regional Initiative of Excellence” from the Polish Minister of Science and Higher Education, 2019–23.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We use real-world data collected from www.entsoe.eu (accessed on 6 April 2016).

Acknowledgments: The author thanks Slawek Smyl and Paweł Pełka for providing forecasts from the base models.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANFIS	Adaptive Neuro-Fuzzy Inference System
ARIMA	Auto-Regressive Integrated Moving Average
cES-adRNN	contextually enhanced hybrid and hierarchical model combining ETS and dilated RNN with attention mechanism
DE	Germany
DeepAR	Auto-Regressive Deep recurrent NN model for probabilistic forecasting
ES	Spain
ETS	Exponential Smoothing
FR	France
GB	Great Britain
GRNN	General Regression Neural Network
LinReg	Linear Regression
LGBM	Light Gradient-Boosting Machine
LSTM	Long Short-Term Memory Neural Network
MAPE	Mean Absolute Percentage Error
MdAPE	Median of Absolute Percentage Error
ML	Machine Learning
MLP	Multilayer Perceptron
MPE	Mean Percentage Error
MSE	Mean Square Error
MTGNN	Graph Neural Network for Multivariate Time series forecasting
N-BEATS	deep NN with hierarchical doubly residual topology
N-WE	Nadaraya—Watson Estimator
NN	Neural Network
PE	Percentage Error
PL	Poland
RNN	Recurrent Neural Network
StdPE	Standard Deviation of Percentage Error
SVM	Support Vector Machine
STLF	Short-Term Load Forecasting
WaveNet	Auto-Regressive deep NN model combining causal filters with dilated convolutions
XGB	eXtreme Gradient Boosting

References

- Clements, M.; Hendry, D. *Forecasting Economic Time Series*; Cambridge University Press: Cambridge, UK, 1998.
- Wang, X.; Hyndman, R.; Li, F.; Kang, Y. Forecast combinations: An over 50-year review. *Int. J. Forecast.* **2022**, *in press*. [CrossRef]
- Rossi, B. Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them. *J. Econ. Lit.* **2021**, *59*, 1135–1190. [CrossRef]
- Blanc, S.; Setzer, T. When to choose the simple average in forecast combination. *J. Bus. Res.* **2016**, *69*, 3951–3962. [CrossRef]
- Genre, V.; Kenny, G.; Meyler, A.; Timmermann, A. Combining expert forecasts: Can anything beat the simple average? *Int. J. Forecast.* **2013**, *29*, 108–121. [CrossRef]
- Jose, V.; Winkler, R. Simple robust averages of forecasts: Some empirical results. *Int. J. Forecast.* **2008**, *24*, 163–169. [CrossRef]
- Pawlikowski, M.; Chorowska, A. Weighted ensemble of statistical models. *Int. J. Forecast.* **2020**, *36*, 93–97. [CrossRef]
- Poncela, P.; Rodriguez, J.; Sanchez-Mangas, R.; Senra, E. Forecast combination through dimension reduction techniques. *Int. J. Forecast.* **2011**, *27*, 224–237. [CrossRef]
- Kolassa, S. Combining exponential smoothing forecasts using Akaike weights. *Int. J. Forecast.* **2011**, *27*, 238–251. [CrossRef]
- Babikir, A.; Mwambi, H. Evaluating the combined forecasts of the dynamic factor model and the artificial neural network model using linear and nonlinear combining methods. *Empir. Econ.* **2016**, *51*, 1541–1556. [CrossRef]
- Zhao, S.; Feng, Y. For2For: Learning to forecast from forecasts. *arXiv* **2020**, arXiv:2001.04601.
- Gastinger, J.; Nicolas, S.; Stepić, D.; Schmidt, M.; Schülke, A. A study on ensemble learning for time series forecasting and the need for meta-learning. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
- Hochreiter S.; Schmidhuber J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]

14. Hewamalage H.; Bergmeir C.; Bandara K. Recurrent neural networks for time series forecasting: Current status and future directions. *Int. J. Forecast.* **2021**, *37*, 388–427. [CrossRef]
15. Smyl, S.; Dudek, G.; Pełka, P. Contextually enhanced ES-dRNN with dynamic attention for short-term load forecasting. *arXiv* **2022**, arXiv:2212.09030.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Moving Object Path Prediction in Traffic Scenes Using Contextual Information †

Jaime B. Fernandez *, Suzanne Little and Noel E. O'Connor

Insight SFI Research Centre for Data Analytics, Dublin City University, Glasnevin, Dublin 9, Ireland; suzanne.little@dcu.ie (S.L.); noel.loconnor@dcu.ie (N.E.O.)

* Correspondence: jaimeboanerjes.fernandezroblero@dcu.ie

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Moving object path prediction in traffic scenes from the perspective of a moving vehicle can improve safety on the road, which is the aim of Advanced Driver Assistance Systems (ADAS). However, this task still remains a challenge. Work has been carried out on the use of x, y positional information of the moving objects only. However, besides positional information there is more information that surrounds a vehicle that can be leveraged in the prediction along with the x, y features. This is known as contextual information. In this work, a deep exploration of these features is carried out by evaluating different types of data, using different fusion strategies. The core architectures of this model are CNN and LSTM architectures. It is concluded that in the prediction task, not only are the features important, but the way they are fused in the developed architecture is also of importance.

Keywords: time series; path prediction; traffic scenes; LSTMs

1. Introduction

In previous work [1,2], only the past positions of the observed object in a scene have been used to predict its future path. However, in traffic scenarios there is a rich set of additional information available about the environment of the ego vehicle and each object in the scene. For example, this information could be an image of a moving object, the velocity of the ego vehicle, the position of other objects or an image of the scene itself. Nowadays, instrumented vehicles are capable of sensing and providing this information that could be leveraged in the path prediction task. This contextual information is used along with the $tr(x, y)$ positional information of the object whose future path is to be predicted.

However, using contextual information still poses a challenge. The information comes in different types of data, e.g., numerical, images. Data also derive from different sources. As a result, the following problems have to be faced: synchronization and availability of data, feature extraction, multimodal data management, and data fusion strategies. This work presents an approach developed to use this contextual information in the path prediction task.

In the remainder of this work, Section 2 presents the related works in path prediction. Section 3 presents our approach. Section 4 describes the experimental setup. In Section 5, a deeper exploration of multimodal features is detailed. Finally, in Section 6, conclusions are provided.

2. Related Works

A variety of techniques for path prediction have been developed, from the well known Kalman Filter (KF) [3], some probabilistic approaches [4], approaches based on prototype trajectories [5] to Recurrent Neural Networks (RNNs) and its variants that have shown good performance on sequential or time series data [6].

Citation: Fernandez, J.B.; Little, S.; O'Connor, N.E. Moving Object Path Prediction in Traffic Scenes Using Contextual Information. *Eng. Proc.* **2023**, *39*, 54. <https://doi.org/10.3390/engproc2023039054>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 5 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

LSTMs have been used as well. LSTMs are capable of obtaining information from sequences and then performing prediction using that previous information. The use of LSTMs in different ways has been explored, from stacked LSTMs [7] to encoder–decoders [8]. One interesting work is presented in [7], where they use LSTMs to predict the trajectory of vehicles in highways from a fixed top-view. In [9], multiple cameras were used to predict the trajectory of people in crowded scenes and [10] predicts the trajectory of vehicles in an occupancy grid from the point of view of an ego vehicle. A more closely related work to this paper is presented in [11], where they predict the future path of pedestrians using RNNs as encoder–decoders and also include the prediction of the odometry of the ego-vehicle.

Nowadays, due to the availability of more data, approaches using contextual information have also been explored. Ref. [12] predicts the trajectory of a cyclist based on the image of the cyclist, the distance of the cyclist to the vehicle and information about the scene. They refer to this information as the object, dynamic and static context, respectively. Ref. [13] predicts the trajectory of pedestrians using the observed trajectory, an occupancy map and visual information of a scene which they call the person, social and scene scale, respectively. Ref. [14] also makes use of visual information from a map along with information from the objects such as velocity, acceleration, and heading change rate as inputs. A complete work is shown in [15], where an end-to-end architecture is used to process different contextual features such as the dynamics of the agents, scene context and interaction between agents. An extended survey is presented in [16], with traditional and deep learning approaches used in the path prediction task. Something interesting about [12–15] is that CNNs along with LSTMs are used to extract features from images, which is required when multi-modal features are processed. The mentioned works present different architectures using different features. Looking at them, it can be seen that there are three important tasks that need to be faced. (1) Multimodal data, (2) feature extraction, and (3) fusion strategies. In this work, different to the ones presented in this section, we aim to explore and evidence the importance of those three processes that should be taken into account when predicting the future path of objects in traffic scenes. For this, we carried out some further research on CNNs and fusion Strategies.

One of the limitations when dealing with a sequence of images is that the CNN model can be highly resource consuming, particularly with large images. Therefore, in this work, we focus on working with subsampled images of sizes such as 40×40 , 64×64 and 124×37 pixels to accommodate for the original image size. For that reason, some research was performed on tiny images classification. An interesting work is detailed in [17], where they present a 4Block-4CNN model that performs well on tiny images of 32×32 pixels from the CINIC-10 dataset. Another related work is provided in [18], where they used the Tiny ImageNet dataset with images of 64×64 . In this work, they conclude that the use of deep models (8CNNs) leads to better performance. Similar conclusions were made in [19]. In these works, the use of several layers of CNNs in the models is common. Since here the task is different to classification, another point to consider is the use of a Global Average Pooling Layer (GAP) that allows the model to generate more generic features and also helps to expose the regions in which the CNNs are focused [20]. This is a desirable characteristic since we are not aiming to classify an image but to predict the future path of objects based on them. Finally, because of the work in [21,22], we decided to use batch normalization.

Finally, knowing how the available features from different sources will be combined is another challenge that is still open to further investigation. Fusion strategies define the way that different streams of features will be joined in the model architecture. Three main fusion strategies are described in [23,24]. Early fusion, mid-level fusion and late fusion. In this work, we also explored some of these fusion strategies.

3. Approach

A path, P , is a set of tracks, tr , that contains information such as $tr(x, y)$ spatial coordinates of an object traveling in a given space or scene, $P = \{tr_{t1}, tr_{t2}, \dots, tr_{tlength}\}$. However, besides the spatial location of the object, there are other features that can be extracted from

the object itself, the ego vehicle, the other objects and the scene. This information can also be extracted sequentially so a track can be represented as a set of features such as $tr(x, y, rgbobject, otherobjects, egovehicle, scene, time)$. Each track feature is a measure given for a sensor in intervals of time and in an ordered manner. This means that a path is a sequence of measurements of the same variables collected over time, where the order matters, resulting in a time series. Due to this fact, a path can be seen as a multivariate time series that has several time-dependent variables. Each variable depends on its past values and this dependency is used for forecasting future values. So the task of path prediction can be seen as multivariate multi-step time series forecasting. LSTMs have shown good performance when dealing when time series; thus, in this approach LSTMs are used for path prediction. We used the sliding window approach to create observed tracklets: $tr^O = [tr_{t_1}^O, tr_{t_2}^O, \dots, tr_{t_{obs}}^O]$ and its respective ground truth tracklet: $tr^G = [tr_{t_1}^G, tr_{t_2}^G, \dots, tr_{t_{pred}}^G]$. The predicted vector of each tr^O is called $tr^P = [tr_{t_1}^P, tr_{t_2}^P, \dots, tr_{t_{pred}}^P]$. The aim of this work is to predict a tr^P based on the observed tracks tr^O .

3.1. Multimodal Data

The information we are processing derives from different sources and is of different types, as shown in Figure 1. We want to build an end-to-end model capable of processing all this information. To achieve this requirement, we are using the Keras functional API (https://keras.io/guides/functional_api/ (accessed on 20 January 2023)). The functional API can handle models with non-linear topology, shared layers, and multiple inputs or outputs.

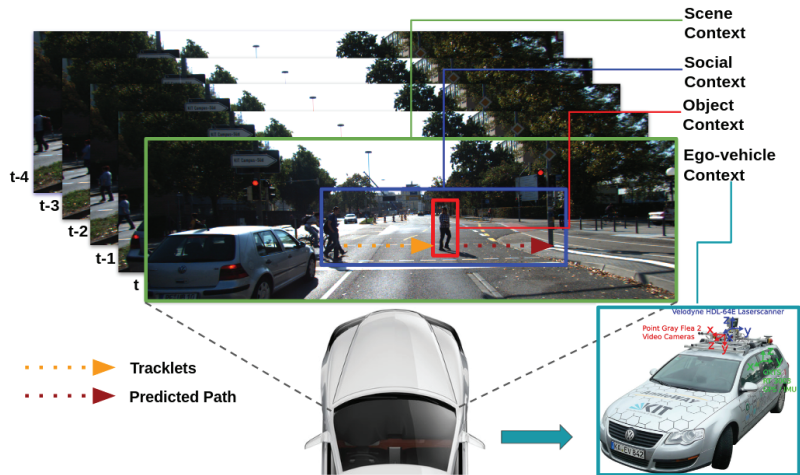


Figure 1. Contextual information.

For traffic scenes, and specifically from cameras mounted on a moving vehicle, the following features were used: object, ego-vehicle telemetry data, scene, and interaction-aware or social context features (other objects), as visualised in Figure 1. These four types of features were selected since they cover most of the information that can be obtained from the perspective of an ego vehicle.

Object. Besides the x, y position of the object, regions of the objects from the RGB images are used. These cropped images can potentially provide a better understanding of the object, such as the pose or orientation that can be leveraged as additional features.

Ego Vehicle. The following features were selected because they are more related to the movement of the vehicle on the ground. Yaw: heading (rad), VF: forward velocity, VL: leftward velocity, VU: upward velocity, AF: forward acceleration, AL: leftward acceleration, AU: upward acceleration.

Scene. Including information about the scene can potentially help to understand the type of traffic scene where the prediction is happening. In this work, RGB images are used.

Interaction Aware. Taking into account the position of other objects is an important feature for path prediction that is called interaction awareness. In this work, three representations of the position of the objects in a scene are used—a grid map, a polar map, and a local BEV map of objects. The grid and the polar map are implemented as in [13]. A Local Bird’s-Eye View Map represents in pixels the position of the objects in the real world. The map encodes the type of each object in a specific color; red for pedestrian, green for vehicles, and blue for cyclists. Here, we used a scale of 1 px:0.3 m because it helps to better separate objects that are too close to one another. However, the BEV contains *all* the objects in a scene, but for predicting the path of an object, nearby objects are the most influential. Therefore, a local BEV is extracted from the global BEV, as in Figure 2.

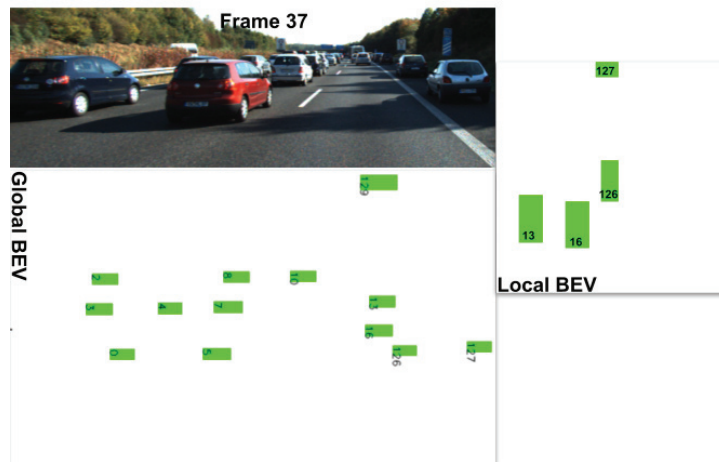


Figure 2. Local BEV for object ID 126.

3.2. Feature Extraction

Taking into account the related works, three CNN models were created to process each type of image in this work. Figure 3 details the model used to process the object image. The architecture is the same for the other type of images—the only difference is the shape, since the sizes of the other types of images are different. For comparison, based on information in [20]—“The responses from the higher-level layers of CNN (e.g., fc6, fc7 from AlexNet) have been shown to be very effective generic features with state-of-the-art performance on a variety of image datasets” (p. 5)—and for its use in [13], we decided to use AlexNet to compare performance with our proposed CNN models. Figure 4 presents the AlexNet architecture for the object image. The architecture is the same for the other type of images, the difference is the stride in the conv1 and maxPool1 layers which is as follows: the interaction-aware image strides for conv1 and maxPool1 are 4 and 1, respectively. For the scene image, the strides for conv1 and maxPool1 are 3 and 2, respectively.

Layer	K	K.S	S	Shape
conv1	32	3×3	1	64×64×32
conv2	32	3×3	2	32×32×32
conv3	32	3×3	1	32×32×32
maxPool1	-	3×3	2	16×16×32
batchNorm1	-	-	-	16×16×32
conv4	64	3×3	1	16×16×64
conv5	64	3×3	1	16×16×64
conv6	64	3×3	2	8×8×64
maxPool2	-	3×3	2	4×4×64
batchNorm2	-	-	-	4×4×64
conv7	128	3×3	1	4×4×128
conv8	128	3×3	1	4×4×128
conv9	128	3×3	1	4×4×128
batchNorm3	-	-	-	4×4×128
conv10	256	3×3	1	4×4×256
conv11	256	3×3	1	4×4×256
conv12	256	3×3	1	4×4×256
GAP	-	-	-	256
dense1	-	-	-	256

Figure 3. 4block3convGAP.

Layer	K	K.S	S	Shape
conv1	96	11×11	4	16×16×96
maxPool1	-	3×3	2	7×7×96
batchNorm1	-	-	-	7×7×96
conv2	256	5×5	1	7×7×256
maxPool2	-	3×3	2	3×3×256
batchNorm2	-	-	-	3×3×256
conv3	384	3×3	1	3×3×384
conv4	384	3×3	1	3×3×384
conv5	256	3×3	1	3×3×256
GAP	-	-	-	256
dense1	-	-	-	256

Figure 4. AlexNet.

3.3. Fusion Strategies

The following three fusion strategies are evaluated: (1) early fusion of raw features (RF), (2) early fusion of latent space (LS) features, (3) middle fusion of latent space (LS) features. The architecture for each type of fusion is presented in the experiments.

3.4. Evaluation Metrics

As in [25,26], the Average Displacement Error (ADE) and Final Displacement Error (FDE) are used:

$$ADE = \frac{\sum_{i=1}^n \sum_{t=1}^{t^{pred}} \sqrt{(\hat{x}_i^t - x_i^t)^2 + (\hat{y}_i^t - y_i^t)^2}}{n(t^{pred})} \quad (1)$$

$$FDE = \frac{\sum_{i=1}^n \sqrt{(\hat{x}_i^{t^{pred}} - x_i^{t^{pred}})^2 + (\hat{y}_i^{t^{pred}} - y_i^{t^{pred}})^2}}{n} \quad (2)$$

where $(\hat{x}_i^t, \hat{y}_i^t)$ are the predicted positions of the tracklet i at time t , (x_i^t, y_i^t) are the actual position (ground truth) of the tracklet i at time t , t^{pred} is the last prediction step, and n is the number of tracklets in the testing set. We used the weighted sum of ADE (WSADE) and weighted sum of FDE (WSFDE) as metrics, as shown in (<http://apolloscape.auto/trajectory.html> (accessed on 10 February 2023)).

3.5. Model Architecture

The focus of this research is to demonstrate the improvement of the path prediction task when using more features, which is why two foundational architectures were used.

Each LSTM has 128 units using the default Keras' configuration. The Adam optimizer was also used with the default values. For regression loss, the Mean Squared Error (MSE) was used. All experiments were run for 1000 epochs, except for those where an initial evaluation of CNN models was performed (300 epochs). The batch size used was 32 due to hardware limitations and because using a small batch size leads to a better trained model. The architecture of the models changes according to the features and are illustrated in their respective sections below:

- Vanilla LSTM: an LSTM using the default Keras' configuration.
- Encoder–Decoder: the encoder comprises a vanilla LSTM for numerical data and a CNN+LSTM for image data. The decoder is a vanilla LSTM that is fed with the features or concatenation of features from the encoder.

4. Experimental Setup

The dataset used to evaluate the models was KITTI due to its realistic scenes, such as highways, inner cities, vehicles standing/moving, and its different objects. All this information is obtained from cameras mounted on a vehicle, which is the main focus of this research work. There are other datasets, however, that do not provide RGB images of the scenarios and ego vehicle features as KITTI does. KITTI provides information from different type of sensors that can be fused and used together. The size of the dataset is 20,141 samples with a class distribution of cyclists (802), pedestrians (6312) and vehicles (13,027). The evaluation was carried out using three objects—pedestrian, vehicles, and cyclists—for a prediction horizon (P.H.) of ± 20 steps. The available dataset is not large and no standard test/train split is available. For this reason, 5-fold cross validation was carried out and the mean results are reported.

4.1. Data Pre-Processing

KITTI provides the data in different files; to use these features together, all the information was synchronized, with the time stamp/ID reference taken as the frame number of each measurement. Each type of data was processed as follows:

- Object: (x, y, z) position in metres were extracted. RGB features: patches of the objects were extracted and resized to 64×64 pixels.
- Ego vehicle: orientation, velocity in $[x, y, z]$, acceleration in $[x, y, z]$ features were extracted from the oxt files which are the dynamics of the ego vehicle.
- Scene: RGB images from the scenes were resized to 124×37 pixels.
- Interaction aware: grid and polar map were flattened first to feed an LSTM. The local image BEV map was resized to 40×40 pixels.

4.2. Training and Prediction

During training, cross validation with $K = 5$ was carried out and for each fold a training and test split was performed with 70% of the data used for training and the rest for testing. Additionally, to feed the model correctly, the following steps were performed:

1. Select the features to use: object image, ego-vehicle information, scene image and interaction-aware information can be used.
2. Scale data: tracklets were scaled to $[0, 1]$. Images were divided by 255 before being fed to the CNN models.
3. Split data into observed tracklet, tr^O , and ground truth tracklet (path to be predicted), tr^G . tr^O is shaped as $[N_{samples}, tobs, features-in-tr^O]$ and tr^G as $[N_{samples}, OutputLength]$. $OutputLength$ is the size of the output of the model. In this case, $OutputLength = t_{pred} * features-in-tr^G$. Here, t_{pred} represents the steps to predict in the future and $features-in-tr^G$ is the number of features to predict in each step.

In [2], it was shown that LSTMs can be extended to predict multiple paths by combining them with Mixture Density Models (MDMs) as a final layer. However, in this work, a single path was predicted to better analyse the impact of the contextual features.

The model outputs a single array, tr^P , per tr^O of size *OutputLength*. tr^P is re-shaped to $[tpred, features-in-tr^G]$.

5. Exploration of Multimodal Features

Several combinations of pairs of features were tested; then, those features that led to better performance were selected to create a combination of three features. Due to hardware limitations and the lack of improvement when fusing three type of features, no further combinations were tested.

The results are visualised using a colour scaling in a range of white to red. A white background means that the model achieved the best performance and a red background means that model showed the worst performance. A reddish shade means the model’s performance is between the best and the worst. The second column shows WSADE, followed by the individual ADE for each type of object. The fifth column presents WSFDE, followed by the individual FDE for each type of object.

Experiments and Results

X, Y Features. This experiment consists of feeding an LSTM with raw features versus first encoding the raw features in a latent space with an LSTM (the encoder) and then feeding this into another LSTM (the decoder). The results depicted in Figure 5 demonstrate that encoding the raw features in latent space improves performance over using the raw features to feed the Vanilla LSTM.

Model	WADE	Ped.	Veh.	Cyc.	WFDE	Ped.	Veh.	Cyc.	Fusion
VLSTM: X,Y	0.674	0.562	0.950	0.719	1.476	1.199	2.150	1.595	RF.
Enc-Dec: X,Y	0.692	0.599	0.936	0.718	1.380	1.144	1.994	1.447	LS.

Figure 5. Raw features vs. latent space features: x, y features.

X, Y, Ego-vehicle Features. Three combinations of ego vehicle information were evaluated—1. $[x,y,VF,VL]$, 2. $[x,y,VF,VL,AF,AL]$ and 3. $[x,y, HEADING, VF,VL,AF,AL]$. To explore these features further, three fusion strategies were tested. From the results shown in Figure 6, two observations can be drawn: (1) the combination of features: the results indicate that using x, y and VF, VL, AF and AL reduces the error in the prediction. (2) Fusion strategy: in most of the cases, using the middle fusion of features’ latent space improves performance. On the contrary, using the raw features directly to feed the LSTM leads to a greater error.

Model	WADE	Ped.	Veh.	Cyc.	WFDE	Ped.	Veh.	Cyc.	Fusion
Enc-Dec: X, Y, VF, VL	0.766	0.545	0.987	1.148	1.694	1.152	2.193	2.671	Middle. LS
Enc-Dec: X, Y, VF, VL, AF, AL	0.673	0.519	0.909	0.866	1.442	1.077	2.019	1.878	Middle. LS
Enc-Dec: X, Y, HEADING, VF, VL, AF, AL	0.729	0.575	1.025	0.864	1.532	1.169	2.261	1.825	Middle. LS
VLSTM: X, Y, VF, VL	0.810	0.616	1.060	1.094	1.603	1.214	2.274	2.018	Early. RF.
VLSTM: X, Y, VF, VL, AF, AL	0.850	0.547	1.057	1.460	1.615	1.076	2.216	2.490	Early. RF.
VLSTM: X, Y, HEADING, VF, VL, AF, AL	0.770	0.581	1.016	1.044	1.585	1.103	2.122	2.367	Early. RF.
Enc-Dec: X, Y, HEADING, VF, VL, AF, AL	0.812	0.635	0.984	1.120	1.563	1.191	2.084	2.070	Early. LS
Enc-Dec: X, Y, VF, VL, AF, AL	0.748	0.556	0.941	1.078	1.526	1.089	2.015	2.234	Early. LS
Enc-Dec: X, Y, VF, VL	0.744	0.571	0.977	0.987	1.508	1.131	2.083	1.981	Early. LS

Figure 6. Early fusion raw features, early fusion latent space, middle fusion latent space: x, y , ego-vehicle features.

The models used to fuse the information in this experiment in three different strategies—early fusion no latent space, early fusion latent space and, middle fusion latent space—are presented in Figures 7–9, respectively. The figures show the fusion of four features—VF, VL, AF, AL of the ego vehicle—the fusion with the other two combinations—VF, VL and Heading, VF, VL, AF, AL—follow the same architecture.

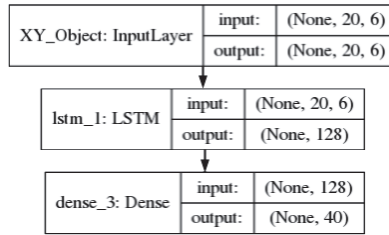


Figure 7. Early fusion no latent space: x, y and vf, vl, af, al features.

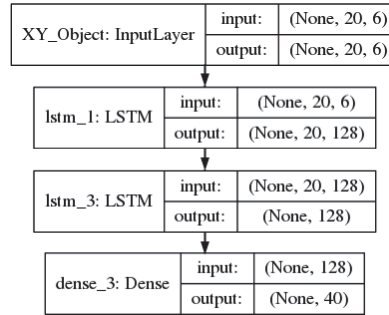


Figure 8. Early fusion latent space: x, y and vf, vl, af, al features.

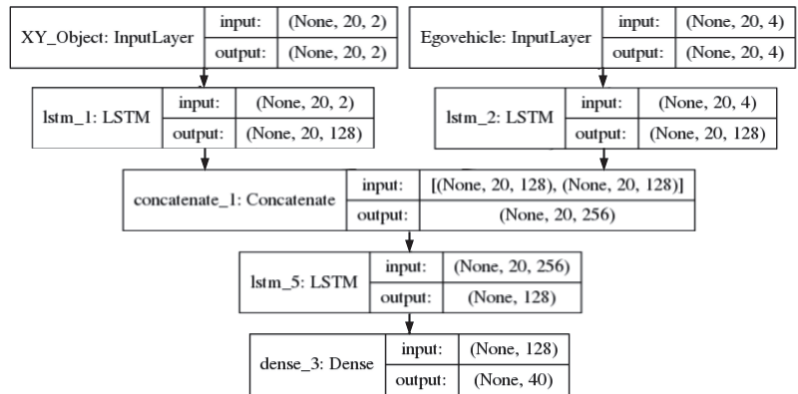


Figure 9. Middle fusion latent space: x, y and vf, vl, af, al features.

$X, Y, Object$ Image. The initial approach to include deep features was to use the model from SSLSTM [13], where the code (<https://github.com/xuehaouwa/SS-LSTM> (accessed on 15 February 2023)) is provided. They use AlexNet without the two layers of Conv2D (384), which creates a shallow and narrow CNN model (AlexNet light). The results of this evaluation show no improvement in comparison to the baseline (using only x, y features). The next step was to find out if using a different CNN would improve the performance; therefore, based on the available literature [13,17], we selected two CNNs. (1) 4block3convGAP for object image and (2) AlexNetGAP for object image. Due to the computational requirements, to initially test the performance of the models they were run for 300 epochs, using the k-fold numbers 1, 3 and 5. The results shown in Figure 10 indicate that using the CNN model 4block3convGAP leads to better performance.

Model / 300 Epochs	WADE	Ped.	Veh.	Cyc.	WFDE	Ped.	Veh.	Cyc.	Fusion
Enc-Dec: X, Y, Object image (4block3convGAP)	0.764	0.611	1.158	0.807	1.464	1.077	2.374	1.658	Middle. LS.
Enc-Dec: X, Y, Object Image (AlexNetGAP)	0.813	0.672	1.204	0.832	1.618	1.311	2.530	1.599	Middle. LS.

Figure 10. 4block3convGAP vs. AlexNetGAP for object image.

For the second stage of this experiment, the model 4block3convGAP was selected to be run for 1000 epochs and for all the k-folds. The results are shown in Figure 11. It can be observed that the 4block3convGAP CNN significantly improves the performance of the model compared to using the CNN from SSLSTM. Additionally, the 4block3convGAP CNN also improves the performance against the baseline model (using only x, y). The improvement is mostly seen for FDE, which leads to an error reduction in the prediction of the final position of an object. The improvement reached by using the 4block3convGAP in contrast to the model from SSLSTM provides evidence that refining the CNN model used leads to an improvement in the prediction. The encoder–decoder model used to fuse x, y positional information and the object image is presented in Figure 12.

Model	WADE	Ped.	Veh.	Cyc.	WFDE	Ped.	Veh.	Cyc.	Fusion
Enc-Dec: X, Y	0.692	0.599	0.936	0.718	1.380	1.144	1.994	1.447	LS.
Enc-Dec: X, Y, Object image (AlexNet light)	0.726	0.637	1.037	0.680	1.408	1.209	2.123	1.281	Middle. LS.
Enc-Dec: X, Y, Object image (4block3convGAP)	0.688	0.576	0.971	0.723	1.308	1.068	1.993	1.317	Middle. LS.

Figure 11. x, y only vs. AlexNet light vs. 4block3convGAP for object image.

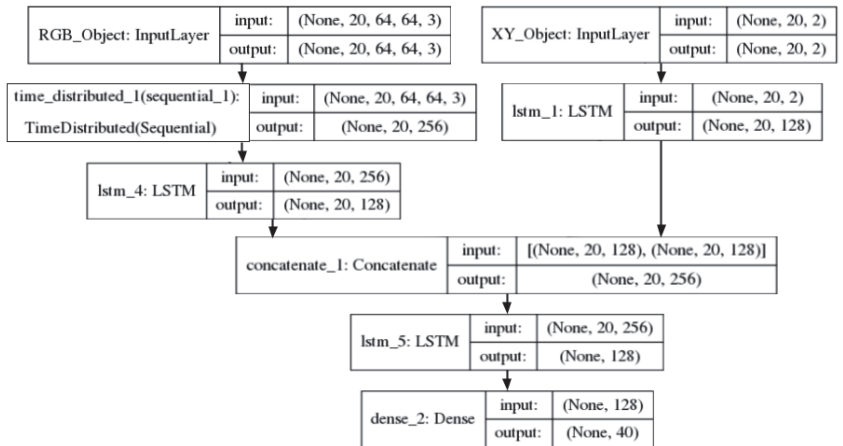


Figure 12. Fusion model for x, y and object image.

X, Y, Interaction-aware Map. Three types of maps were explored to include other objects’ positional information, a grid map, a polar map, and one local BEV. The results are presented in Figure 13. The following observations can be made: (1) Grid Map vs. Polar Map: using a grid map shows better performance. (2) Fusion strategy: there is a significant error reduction when using middle fusion of features in the latent space. Early fusion, however, leads to a higher error. This is observed for both the grid and polar map.

Model	WADE	Ped.	Veh.	Cyc.	WFDE	Ped.	Veh.	Cyc.	Fusion
VLSTM: X, Y, Polar map	1.666	2.062	1.161	1.080	2.854	3.255	2.519	2.104	Early. RF.
Enc-Dec: X, Y, Polar map	2.126	2.715	1.149	1.463	3.361	4.036	2.450	2.410	Early. LS.
Enc-Dec: X, Y, Polar map	0.924	0.872	1.096	0.905	1.904	1.820	2.302	1.765	Middle. LS.
VLSTM: X, Y, Grid map	2.111	2.552	1.490	1.522	3.363	3.831	3.009	2.453	Early. RF.
Enc-Dec: X, Y, Grid map	2.572	3.154	1.610	1.909	3.979	4.627	3.135	3.039	Early. LS.
Enc-Dec: X, Y, Grid map	0.850	0.822	1.031	0.761	1.888	1.817	2.336	1.666	Middle. LS.

Figure 13. Early fusion raw features, Early fusion latent space, Middle fusion latent space: x, y , int-aware features.

For the third map BEV, as in $X, Y, Object Image$, two CNN models were selected—4block3convGAP and AlexNetGAP—for the BEV image. The two CNNs were run for 300 epochs using the k-folds numbers 1, 3 and 5 and the mean are reported in Figure 14. It can be seen that AlexNetGAP is slightly better than 4block3convGAP.

Model / 300 Epochs	WADE	Ped.	Veh.	Cyc.	WFDE	Ped.	Veh.	Cyc.	Fusion
Enc-Dec: X, Y, Local BEV map (4block3convGAP)	1.021	0.894	1.395	1.015	1.719	1.430	2.547	1.729	Middle. LS.
Enc-Dec: X, Y, Local BEV map (AlexNetGAP)	0.909	0.762	1.422	0.830	1.750	1.381	2.847	1.725	Middle. LS.

Figure 14. 4block3convGAP vs. AlexNetGAP for interaction-aware local BEV map.

The next step was to train the AlexNetGAP for 1000 epochs on the 5 folds. The results are depicted in Figure 15. It can be observed that using the local BEV with AlexNetGAP improves the performance slightly over the use of a grid with middle fusion on latent space. However, the use of local BEV with AlexNetGAP does not lead to error reduction over the base line model (using only x, y features). The reason could be because the KITTI dataset does not have many crowded scenes where the objects interact with each other.

Model	WADE	Ped.	Veh.	Cyc.	WFDE	Ped.	Veh.	Cyc.	Fusion
Enc-Dec: X, Y	0.692	0.599	0.936	0.718	1.380	1.144	1.994	1.447	LS.
VLSTM: X, Y, Polar map	1.666	2.062	1.161	1.080	2.854	3.255	2.519	2.104	Early. RF.
Enc-Dec: X, Y, Polar map	2.126	2.715	1.149	1.463	3.361	4.036	2.450	2.410	Early. LS.
Enc-Dec: X, Y, Polar map	0.924	0.872	1.096	0.905	1.904	1.820	2.302	1.765	Middle. LS.
VLSTM: X, Y, Grid map	2.111	2.552	1.490	1.522	3.363	3.831	3.009	2.453	Early. RF.
Enc-Dec: X, Y, Grid map	2.572	3.154	1.610	1.909	3.979	4.627	3.135	3.039	Early. LS.
Enc-Dec: X, Y, Grid map	0.850	0.822	1.031	0.761	1.888	1.817	2.336	1.666	Middle. LS.
Enc-Dec: X, Y, Local BEV Map (AlexNetGAP)	0.839	0.751	1.101	0.832	1.800	1.573	2.383	1.869	Middle. LS.

Figure 15. x, y only vs. polar vs. grid vs. local BEV map.

Three different models were used to fuse positional information and the created handcrafted maps—early fusion with no latent space, early fusion with latent space and middle fusion with latent space. These three models are similar to those used in X, Y , ego-vehicle features (Figures 7–9, respectively.) The only difference is that in this case the features used are from the created handcrafted maps (grid or polar map). The encoder–decoder model used to fuse positional information and the local BEV map image is similar to the one presented in Figure 12. The only difference is the CNN model, which in this case is used for the interaction-aware image.

X, Y , Scene Image. This experiment combines RGB images of the scenes with the x, y features of the objects. Again, in this set of experiments, we first tested two CNN models (4block3convGAP and AlexNetGAP) for 300 epochs on the k-fold numbers 1, 3 and 5. The mean results are presented in Figure 16. Since AlexNetGAP demonstrated better performance, it was trained for 1000 epochs on the 5 folds. The results are depicted in Figure 17. It can be observed that the use of AlexNetGAP to extract features from scene images leads to a slight error reduction in comparison to the baseline model. The improvement is observed for FDE and mostly for the object cyclist. The encoder–decoder model used to fuse x, y positional information and the scene image is similar to the one presented in Figure 12. The only difference is the CNN model, which in this case is the one used for the scene image.

Model / 300 Epochs	WADE	Ped.	Veh.	Cyc.	WFDE	Ped.	Veh.	Cyc.	Fusion
Enc-Dec: X, Y, Scene image (4block3convGAP)	0.736	0.574	1.165	0.772	1.388	0.997	2.370	1.528	Middle. LS.
Enc-Dec: X, Y, Scene image (AlexNetGAP)	0.667	0.503	1.083	0.721	1.396	1.027	2.309	1.540	Middle. LS.

Figure 16. 4block3convGAP vs. AlexNetGAP for scene image.

Model	WADE	Ped.	Veh.	Cyc.	WFDE	Ped.	Veh.	Cyc.	Fusion
Enc-Dec: X, Y	0.692	0.599	0.936	0.718	1.380	1.144	1.994	1.447	LS.
Enc-Dec: X, Y, scene image (AlexNetGAP)	0.693	0.598	0.964	0.696	1.372	1.148	1.992	1.399	Middle. LS.

Figure 17. x, y and scene image: x, y only vs. AlexNetGAP for scene image.

Combinations of Features that Lead to a Better Performance. The Figure 18 shows the performance of the baseline model in the first row compared to the two best combinations of features. It can be seen that combining the x, y features with the ego vehicle information (VF, VL, AF, AL) (second row) leads to better performance. The combination of x, y features with the object image also improves the performance against the baseline.

Model	WADE	Ped.	Veh.	Cyc.	WFDE	Ped.	Veh.	Cyc.	Fusion
Enc-Dec: X, Y	0.692	0.599	0.936	0.718	1.380	1.144	1.994	1.447	LS.
Enc-Dec: X, Y, VF, VL, AF, AL	0.673	0.519	0.909	0.866	1.442	1.077	2.019	1.878	Middle. LS.
Enc-Dec: X, Y, Object image (4block3convGAP)	0.688	0.576	0.971	0.723	1.308	1.068	1.993	1.317	Middle. LS.

Figure 18. Combinations of features that lead to a better performance.

X, Y, Object Image, Ego-vehicle Features. Previous experiments shown that ego vehicle information and visual information of the object lead to an error reduction. In this experiment, those two features are combined with the x, y positional information. The results are presented in Figure 19, it can be seen that no improvement was reached when combining these three features. The model architecture is similar to the one shown in Figure 12 but in this case we added another branch of information coming from the ego-vehicle feature, as in Figure 9.

Model	WADE	Ped.	Veh.	Cyc.	WFDE	Ped.	Veh.	Cyc.	Fusion
Enc-Dec: X, Y	0.692	0.599	0.936	0.718	1.380	1.144	1.994	1.447	LS.
Enc-Dec: X, Y, Object image, VF, VL, AF, AL (4block3convGAP)	0.782	0.652	1.057	0.872	1.531	1.219	2.127	1.812	Middle. LS.

Figure 19. Combination of three features: x, y , object image, and ego-vehicle.

Ensembles. Three ensembles were built using the models that showed improvement in performance compared to our baseline methodology. Ensemble 1: [Enc-Dec: X, Y], [Enc-Dec: X, Y, VF, VL, AF, AL]. Ensemble 2: [Enc-Dec: X, Y], [Enc-Dec: X, Y, Object image (4block3convGAP)]. Ensemble 3: [Enc-Dec: X, Y], [Enc-Dec: X, Y, VF, VL, AF, AL], [Enc-Dec: X, Y, Object image (4block3convGAP)]. In each ensemble, the output of each model is combined by averaging in the final output layer. The results are presented in Figure 20. The three ensembles achieve better performance than the baseline model (row 1) and the two individual combinations of features (rows 2 and 3).

Model	WADE	Ped.	Veh.	Cyc.	WFDE	Ped.	Veh.	Cyc.	Fusion
Enc-Dec: X, Y	0.692	0.599	0.936	0.718	1.380	1.144	1.994	1.447	LS.
Enc-Dec: X, Y, VF, VL, AF, AL	0.673	0.519	0.909	0.866	1.442	1.077	2.019	1.878	Middle. LS
Enc-Dec: X, Y, Object image (4block3convGAP)	0.688	0.576	0.971	0.723	1.308	1.068	1.993	1.317	Middle. LS.
Ensemble 1: [Enc-Dec: X, Y], [Enc-Dec: X,Y, VF, VL, AF, AL]	0.567	0.460	0.805	0.634	1.207	0.942	1.800	1.369	Middle. LS.
Ensemble 2: [Enc-Dec: X, Y], [Enc-Dec: X,Y, Object image (4block3convGAP)]	0.601	0.501	0.864	0.626	1.255	1.026	1.887	1.284	Middle. LS.
Ensemble 3: [Enc-Dec: X, Y], [Enc-Dec: X,Y, VF, VL, AF, AL], [Enc-Dec: X,Y, Object image (4block3convGAP)]	0.525	0.423	0.783	0.559	1.145	0.901	1.764	1.227	Middle. LS.

Figure 20. Results of ensembles.

6. Discussion and Conclusions

The following observations can be noted from exploring multimodal features on the path prediction task:

Best Combination of Features. Combining x, y with the ego-vehicle features leads to a better performance mostly for ADE. The combination of x, y with Object image also improves the performance against the baseline model for both ADE and FDE.

Latent Space Representation and Fusion Strategies. It can be noted that the way in which the features are fused in the model architecture has a significant impact on the prediction. The initial evidence was presented in X, Y features in Figure 5; the results show that representing the features in a latent space improves the performance compared to using the features directly to the LSTM. Then, evidence was presented in X, Y, Ego-vehicle features; the results in Figure 6 indicate that the middle fusion of features represented in latent space had better performance. On the contrary, early fusion led to a higher error. Next, evidence was presented in X, Y, interaction-aware map; the results exhibited in Figure 13 revealed that there is an error reduction when using middle fusion in the latent space. Again, early fusion led to a higher error. Looking at the impact of the fusion strategies, it would be interesting to explore more in this area, since not only are the features important, but the way these features are fused in a model architecture is also of importance.

CNN Models. From the three experiments in which deep features were used, something important to point out is the impact that a CNN model has in the overall path prediction task. When dealing with images, an evaluation of different CNNs is desirable. It would be interesting to evaluate more CNNs and to see if some of them with common characteristics such as the number of CNN layers, filter size, or specific type of layers (GAP, Attention, etc.) can improve the performance.

Ensembles. It is known that ensembles improve the performance of individual classifiers or methods; in this work, there is evidenced in the context of path prediction in traffic scenes. The three ensembles improved the performance of the baseline methodology and the two other best individual models.

Author Contributions: Conceptualization, J.B.F. and S.L.; methodology, J.B.F. and S.L.; data acquisition, J.B.F.; software tools, J.B.F.; validation, J.B.F.; investigation, J.B.F.; writing—original draft preparation, J.B.F.; writing—review and editing, J.B.F., S.L. and N.E.O.; visualization, J.B.F.; supervision, S.L. and N.E.O. All authors have read and agreed to the published version of the manuscript.

Funding: This publication has emanated from research conducted with the financial support of Science Foundation Ireland [12/RC/2289_P2] at Insight the SFI Research Centre for Data Analytics at Dublin City University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The GPU GeForce GTX 980 used for this research was donated by the NVIDIA Corporation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fernandez, J.B.; Little, S.; O'Connor, N.E. A Single-Shot Approach Using an LSTM for Moving Object Path Prediction. In Proceedings of the 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turkey, 6–9 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
2. Fernandez, J.B.; Little, S.; O'Connor, N.E. Multiple Path Prediction for Traffic Scenes using LSTMs and Mixture Density Models. In Proceedings of the Proceedings of the 6th International Conference on Vehicle Technology and Intelligent Transport Systems—VEHITS, INSTICC, Prague, Czech Republic, 2–4 May 2020; SciTePress: Setubal, Portugal, 2020; pp. 481–488. [CrossRef]
3. Jin, X.B.; Su, T.L.; Kong, J.L.; Bai, Y.T.; Miao, B.B.; Dou, C. State-of-the-Art Mobile Intelligence: Enabling Robots to Move Like Humans by Estimating Mobility with Artificial Intelligence. *Appl. Sci.* **2018**, *8*, 379. [CrossRef]
4. Keller, C.G.; Gavrilu, D.M. Will the pedestrian cross? A study on pedestrian path prediction. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 494–506. [CrossRef]
5. Bian, J.; Tian, D.; Tang, Y.; Tao, D. A survey on trajectory clustering analysis. *arXiv* **2018**, arXiv:1802.06971.
6. Zyner, A.; Worrall, S.; Nebot, E. Naturalistic driver intention and path prediction using recurrent neural networks. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 1584–1594. [CrossRef]
7. Althché, F.; de La Fortelle, A. An LSTM network for highway trajectory prediction. In Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 353–359.
8. Park, S.H.; Kim, B.; Kang, C.M.; Chung, C.C.; Choi, J.W. Sequence-to-sequence prediction of vehicle trajectory via LSTM encoder-decoder architecture. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Suzhou, China, 26–30 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1672–1678.
9. Bartoli, F.; Lisanti, G.; Ballan, L.; Del Bimbo, A. Context-aware trajectory prediction. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1941–1946.
10. Kim, B.; Kang, C.M.; Kim, J.; Lee, S.H.; Chung, C.C.; Choi, J.W. Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network. In Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 399–404.
11. Bhattacharyya, A.; Fritz, M.; Schiele, B. Long-term on-board prediction of pedestrians in traffic scenes. In Proceedings of the 1st Conference on Robot Learning, Mountain View, CA, USA, 13–15 November 2017.
12. Pool, E.A.; Kooij, J.F.; Gavrilu, D.M. Context-based cyclist path prediction using recurrent neural networks. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 824–830.
13. Xue, H.; Huynh, D.Q.; Reynolds, M. SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2017; IEEE: Piscataway, NJ, USA, 2018; pp. 1186–1194.
14. Cui, H.; Radosavljevic, V.; Chou, F.C.; Lin, T.H.; Nguyen, T.; Huang, T.K.; Schneider, J.; Djuric, N. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 2090–2096.
15. Lee, N.; Choi, W.; Vernaza, P.; Choy, C.B.; Torr, P.H.; Chandraker, M. Desire: Distant future prediction in dynamic scenes with interacting agents. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 336–345.
16. Bighashdel, A.; Dubbelman, G. A survey on path prediction techniques for vulnerable road users: From traditional to deep-learning approaches. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1039–1046.
17. Sharif, M.; Kausar, A.; Park, J.; Shin, D.R. Tiny image classification using Four-Block convolutional neural network. In Proceedings of the 2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 16–18 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
18. Yao, L.; Miller, J. Tiny imagenet classification with convolutional neural networks. *CS 231N* **2015**, *2*, 8.
19. Abai, Z.; Rajmalwar, N. DenseNet Models for Tiny ImageNet Classification. *arXiv* **2019**, arXiv:1904.10429.
20. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
21. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; PMLR: Cambridge, MA, USA, 2015; pp. 448–456.
22. Garbin, C.; Zhu, X.; Marques, O. Dropout vs. batch normalization: An empirical study of their impact to deep learning. *Multimed. Tools Appl.* **2020**, *79*, 12777–12815. [CrossRef]
23. Hu, Y.; Lu, M.; Lu, X. Spatial-Temporal Fusion Convolutional Neural Network for Simulated Driving Behavior Recognition. In Proceedings of the 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV), Singapore, 18–21 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1271–1277.

24. Roitberg, A.; Pollert, T.; Haurilet, M.; Martin, M.; Stiefelhagen, R. Analysis of deep fusion strategies for multi-modal gesture recognition. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 15–20 June 2019.
25. Hou, L.; Xin, L.; Li, S.E.; Cheng, B.; Wang, W. Interactive trajectory prediction of surrounding road users for autonomous driving using structural-LSTM network. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 4615–4625. [CrossRef]
26. Ma, Y.; Zhu, X.; Zhang, S.; Yang, R.; Wang, W.; Manocha, D. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6120–6127.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Proceeding Paper

Downscaling Fusion Model for CMIP5 Rainfall Projection under RCP Scenarios: The Case of Trentino-Alto Adige [†]

Amir Aieb ^{1,*}, Antonio Liotta ¹ and Ismahen Kadri ²

¹ Faculty of Engineering, Free University of Bozen-Bolzano, 39100 Bolzano, Italy; antonio.liotta@unibz.it

² Laboratory LGCH, Department of Civil Engineering and Hydraulics, 8 Mai 1945 Guelma University, Guelma 24000, Algeria; ismahenk15@gmail.com

* Correspondence: amir.aieb@stud-inf.unibz.it

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Climate parameter projections obtained by global and regional models (GCM and RCM, respectively) offer a challenge to many researchers in terms of controlling the quality of the outcome data using several scales. In the literature, the proposed models, namely statistical downscaled and regression-based models, are mostly used to adjust the RCM data series. Contrariwise, in practice, these conceptual models perform poorly in certain cases and at certain scales. In this regard, a new downscaling model is proposed herein for annual rainfall projection, based on fusion models, namely polynomial regression (Poly_R), classification and regression tree (CRT), and principal component regression (PCR). The proposed model downscales the rainfall data projected by the coupled model intercomparison phase five (CMIP5) under different representative concentration pathway (RCP) scenarios (2.6, 4.5, 6.0, and 8.5) using overlapping data between the observation and the CMIP5 historical data. This process aims to define the framework for how to use the output equations and algorithm to correct data forecasting by RCM. Generally, the model can be summarized into three levels of analysis, starting with an iterative downscaling using a trendline model that is obtained by Poly_R fitting. Then, the CRT is used to classify and predict the data in subsets. Finally, multiple regression is given by a PCR model using principal components and standardized variables. The final model is also used to downscale the predicted data obtained by both previous models. The results provide the best performance of the fusion model in all RCP cases, compared to the delta change correction and linear scale models. This performance is proved by R^2 scores which range between 0.87 and 0.95.

Keywords: CMIP5 rainfall projection; RCP scenario; downscaling fusion model

Citation: Aieb, A.; Liotta, A.; Kadri, I. Downscaling Fusion Model for CMIP5 Rainfall Projection under RCP Scenarios: The Case of Trentino-Alto Adige. *Eng. Proc.* **2023**, *39*, 55. <https://doi.org/10.3390/engproc2023039055>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 5 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Evaluating hydroclimatic indices under future climate change conditions is crucial for informing decision-makers about future water availability. This information must also be considered in future plans for civil construction and development [1,2]. Future studies on rainfall variability often use data from global climate models (GCMs), which explore the causes of climate change and the relationship between natural climate variability and human activities. GCMs are widely used in climate research to study global and regional climate patterns, assess the potential impacts of climate change, and evaluate the effectiveness of different mitigation and adaptation strategies. Unfortunately, GCMs have a very low spatial resolution, which can lead to limitations in accurately representing local climate characteristics [3]. In order to address these limitations, researchers often use certain techniques such as downscaling, which involves refining the GCM outputs to a finer scale to better represent local conditions.

Downscaling can be performed using statistical methods or by using regional climate models (RCM) that possess a higher spatial resolution. Generally, the RCM model uses

incorporating topography, land–sea contrast, surface heterogeneities, and certain information about physical processes using a spatial resolution from 20 to 50 km [4]. The statistical downscaling (SD) method is based on historical observed data to create an empirical relationship between the GCM and observed data [5,6]. The most commonly used SD methods are regression-based [7–9] and correction-based equations [10].

Artificial neural networks (ANNs) offer another regression-based method, which is able to capture non-linear relationships, and these networks tend to perform better than the multiple linear regression method in certain cases [11]. ANNs also yield physically interpretable linkages to surface climate. However, ANN models require large time series data and are incapable of predicting values outside of the historical dataset [4].

This work aims to propose a new model to downscale the annual rainfall data projected by the coupled model intercomparison phase five (CMIP5) using representative concentration pathway (RCP) scenarios, such as RCP 2.6, RCP 4.5, RCP 6.0, and RCP 8.0 [12]. The idea is to fuse two different results obtained by polynomial and tree-regression (Poly_R and CRT, respectively) methods using principal components regression (PCR). In general, the new model follows three steps of processing using overlapping data series of observed and simulated historical data. The Trentino-Alto Adige region (in northern Italy) has been selected in our study due to the significant temporal variability of annual rainfall observed there during the 17 years under study, as well as for the high diversity of elevation which characterizes this region.

In the experimental part, the new method shows in detail its efficiency in correcting the large errors between CMIP5 and real data, followed by a comparative study to explain its performance compared to other models mostly used in the state of art. This technique provides an improvement when applying consecutive processing on a downscaled output using different classifications by CRT and PCR models.

2. Related Work

In the literature, several statistical methods are based on regression techniques using linear and machine-learning models to correct future RCM multiscale data. Generally, these models use historical observed and RCM data to define a new factor or equation for climate downscaling in each specific region. Examples of these methods are those implemented by the soil and water assessment tool (SWAT) [13], which are the linear scaling (LS) model and delta change correction (DCC) model.

2.1. Linear Scaling Model

The LS model is one of the statistical methods applied to downscale the rainfall and temperature projection data obtained by RCM, based on the change factors α and β , respectively [14]. Both factors are obtained by dividing the overlapping data of real observation on the data projection, as is shown by Equations (1) and (2), respectively. Using these factors allows to correct the future RCM rainfall and temperature data.

$$P_{D,F} = \alpha \times P_{RCM,F}, \text{ where } \alpha = \frac{P_{Obs,Hist}}{P_{RCM,Hist}} \quad (1)$$

$$T_{D,F} = \beta \times T_{RCM,F}, \text{ where } \beta = \frac{T_{Obs,Hist}}{T_{RCM,Hist}} \quad (2)$$

where $P_{D,F}$ and $T_{D,F}$ are the future rainfall and temperature downscaled data, respectively; $P_{RCM,F}$ and $T_{RCM,F}$ are the future rainfall and temperature data projection by RCM, respectively; $P_{Obs,Hist}$ and $T_{Obs,Hist}$ are the historical rainfall and temperature real data, respectively; and $P_{RCM,Hist}$ and $T_{RCM,Hist}$ are the historical rainfall and temperature data projection, respectively.

2.2. Delta Change Correction Model

This method uses the extreme rainfall or temperature values (P^E or T^E , respectively) obtained during T years by the generalized extreme value distribution model (GEV). The extreme values are used to determine the correction factor to downscale RCM future data using the following functions [15]:

$$P_{D,F} = \gamma \times P_{RCM,F}, \text{ where } \gamma = \frac{P_{Obs,Hist}^E}{P_{RCM,Hist}^E}, P^E = GEV (P_{Hist}) \quad (3)$$

$$T_{D,F} = \delta \times T_{RCM,F}, \text{ where } \delta = \frac{T_{Obs,Hist}^E}{T_{RCM,Hist}^E}, T^E = GEV (T_{Hist}) \quad (4)$$

where $P_{D,F}$ and $T_{D,F}$ are the future rainfall and temperature downscaled data, respectively; $P_{RCM,F}$ and $T_{RCM,F}$ are the future rainfall and temperature data projection by RCM, respectively; $P_{Obs,Hist}^E$ and $T_{Obs,Hist}^E$ are the extreme rainfall and temperature real data, respectively; and $P_{RCM,Hist}^E$ and $T_{RCM,Hist}^E$ are the extreme rainfall and temperature data projection, respectively.

3. Material and Methods

In this part, the used methods, the study area, and the newly proposed data-driven downscaling model are detailed. This is used to adjust the CMIP5 annual rainfall projections given by RCM, under different scenarios, as detailed in the following sub-sections.

3.1. Used Method

3.1.1. Polynomial regression

Polynomial regression (Ploy_R) is a case of a multiple non-linear regression model with only one independent variable (X). In this function, we regress the variable X on powers (i) [16], as follows:

$$Y = \beta_0 + \sum_{i=1}^n \beta_i \times X_i, \text{ where } i = 1, 2, \dots, n \quad (5)$$

3.1.2. Classification Regression Tree

The tree-driven regression algorithm is one of the family of machine-learning models. In this regard, four models were developed using the tree technique, including a classification and regression tree (CRT), random forest (RF), gradient-boosting decision tree (GBDT), and extreme gradient boosting (XGB) [17]. CRT and RF belong to supervised learning. In time series analysis, both models provide a stable performance for data downscaling under different scales [18].

3.1.3. Principal Component Regression

Generally, this model performs better than a simple regression. It is based on data classification and regression of each subset using principal components (Y_i). This method regresses the independent variable (X_i) using a standardized variable (X'_i) and principal component (C_i) [19], as follows:

$$X'_i = \frac{(X_i - \bar{X}_i)}{S_i} \quad (6)$$

$$C_i = \sum_{j=1}^p \alpha_{ij} \times X'_j, \text{ where } j = 1, 2, \dots, P \quad (7)$$

$$Y_i = \sum_{j=1}^p \beta_j \times C_j \text{ where } j = 1, 2, \dots, P \quad (8)$$

3.2. Metrics of Performance

A set of statistical parameters has been applied in the experimental part of this paper to control the quality of performance provided by each sub-model used in this proposal. These metrics are the coefficient of determination (R^2), adjusted coefficient of determination (R^2_{Adj}), root mean square error (RMSE), and residual analysis [20–22]. These were applied to compare the predicted values with the observed values.

4. Study Area and Data

Trentino-Alto Adige is located in the northern part of Italy. It has an approximate total surface area of 13,612 km² and a demography of 523,000 people. Alto Adige is located between a latitude of 45.67° N and 47.10° N, and a longitude of 10.37° E and 12.48° E. The region is well known for its diverse geography, which includes the towering Dolomite Mountains and rolling hills dotted with vineyards and apple orchards. The climate in Alto Adige is continental, with warm summers and cold winters. The average maximum temperature in the summer months, especially during July and August, is around 25 °C, while the average minimum temperature in the winter months is around −5 °C. The region experiences an average amount of annual rainfall of around 895 mm. This region is characterized by diverse elevations ranging from 200 m to 4565 m, which are distributed relatively from the central to the northern part of the region, between an urban and a mountainous area, respectively [23]. The maximum altitude in the Trentino-Alto Adige region is more observed in the western part between 2295 m and 4565 m (Figure 1). In this work, the annual rainfall historical data of the Trentino-Alto Adige region observed by the monitoring station between 2005 and 2022 are presented as the response variable for the proposed model. On the other side, the CMIP5 data projection for the same region is obtained by the RCM according to a multi-model ensemble under different RCPs scenarios, which are 2.6, 4.5, 6.0, and 8.5. The CMIP5 data are supported by the IPCC's fifth assessment report, which is available on the climate knowledge portal website of the World Bank Group: <https://climateknowledgeportal.worldbank.org/country/italy/cmip5> (accessed on 20 January 2023).

Figure 2 shows quantitative and qualitative statistical tools to describe the data variability and distribution of each dataset used in the experimental part of this paper. This part gives information about a comparative analysis between the CMIP5 data series under each scenario and the observed data, using the histogram of density, the curve of the values compared to the mean, and the quantile values (first quantile, median, and third quantile). According to real observations that were obtained by the Trentino-Alto Adige meteorological station, the region underwent a humid period between 2005 and 2015, where the annual rainfall exceeded an average of 958.20 mm. In 2009 and 2010, the rainfall accumulation reached the maximum value during this period. However, between 2016 and 2022, a drought phase was observed in the region, while the minimum extreme value is 920 mm. This was observed during 2019 and 2022 (Figure 2(B1)). Generally, the rainfall pattern in the Trentino-Alto Adige region is non-stationary, demonstrated in Figure 2(C1) by a median value above the average with a variability equal to 0.031. All cases where rainfall data were projected by the RCM under RCP scenarios exhibit a high diversity in rainfall variability and distribution (Table S1). According to Scenario 2.6, the rainfall data follow the Weibull distribution during the whole period (Figure 2(B2)), followed by a periodic variability between wet and dry; during the periods of 2005–2008 and 2013–2019, the region exhibited two phases of humidity, demonstrated by a maximum rainfall value equal to 970 mm which was observed in 2016.

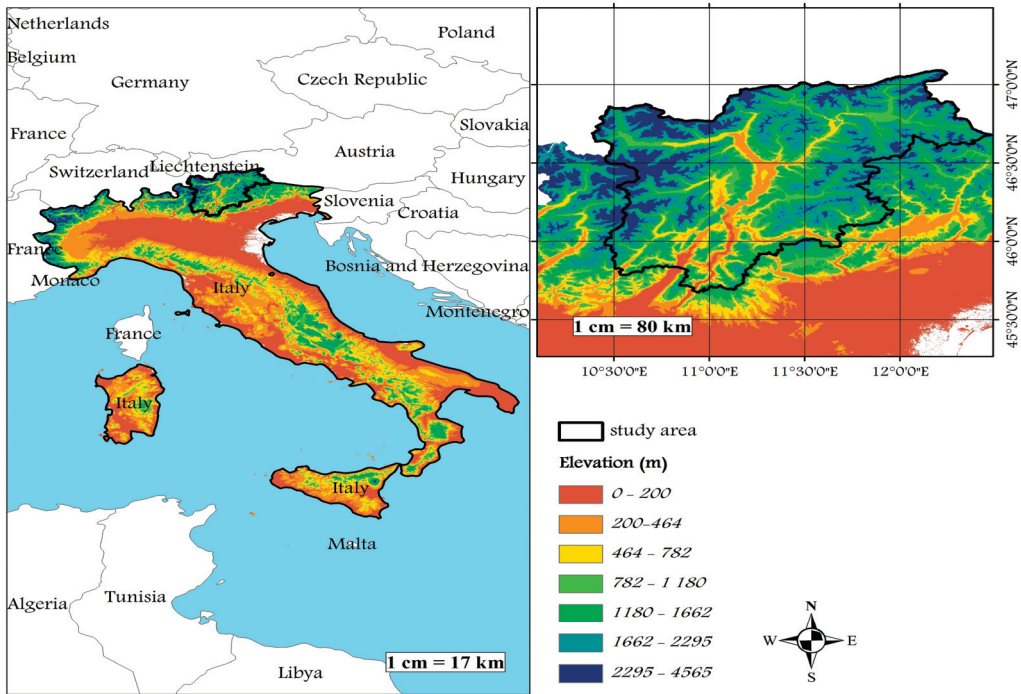


Figure 1. Map of the Trentino-Alto Adige region showing the spatial elevations and its regional location in Italy.

In addition, during the time range of 2009–2012 and 2020–2022, a drought was observed in this region (Figure 2(B2)), which reached a minimum value of 920 mm (Figure 2(C2)). The data projected by the RCP 2.6 scenario provide a large gap when compared with the actual observation. Contrariwise, the RCP 4.5 rainfall data have a symmetric variability compared to the real observation, in which the series started with a dryness phase between 2005 and 2017. Then, a period of humidity was observed between 2018 and 2022, in which the rainfall showed a maximum value of 980 mm in 2021. During this period, the projected data follow a GEV distribution (Figure 2(C1)). Moreover, the data obtained under RCP 6.0 have the same distribution as the previous projection. The average value of this series is close to the mean actual data. Generally, this series is the best which provides a near variability to actual data (Table S1). The only difference is the temporal data distribution where the data have a symmetric distribution compared to actual observation (Figure 2(B1–B5)). The data obtained by the RCP 8.5 scenario also exhibit similarity with downscaled data under the 4.5 scenario, where the variability of both series is very close, demonstrated by a CV equal to 0.16 and 0.15, respectively.

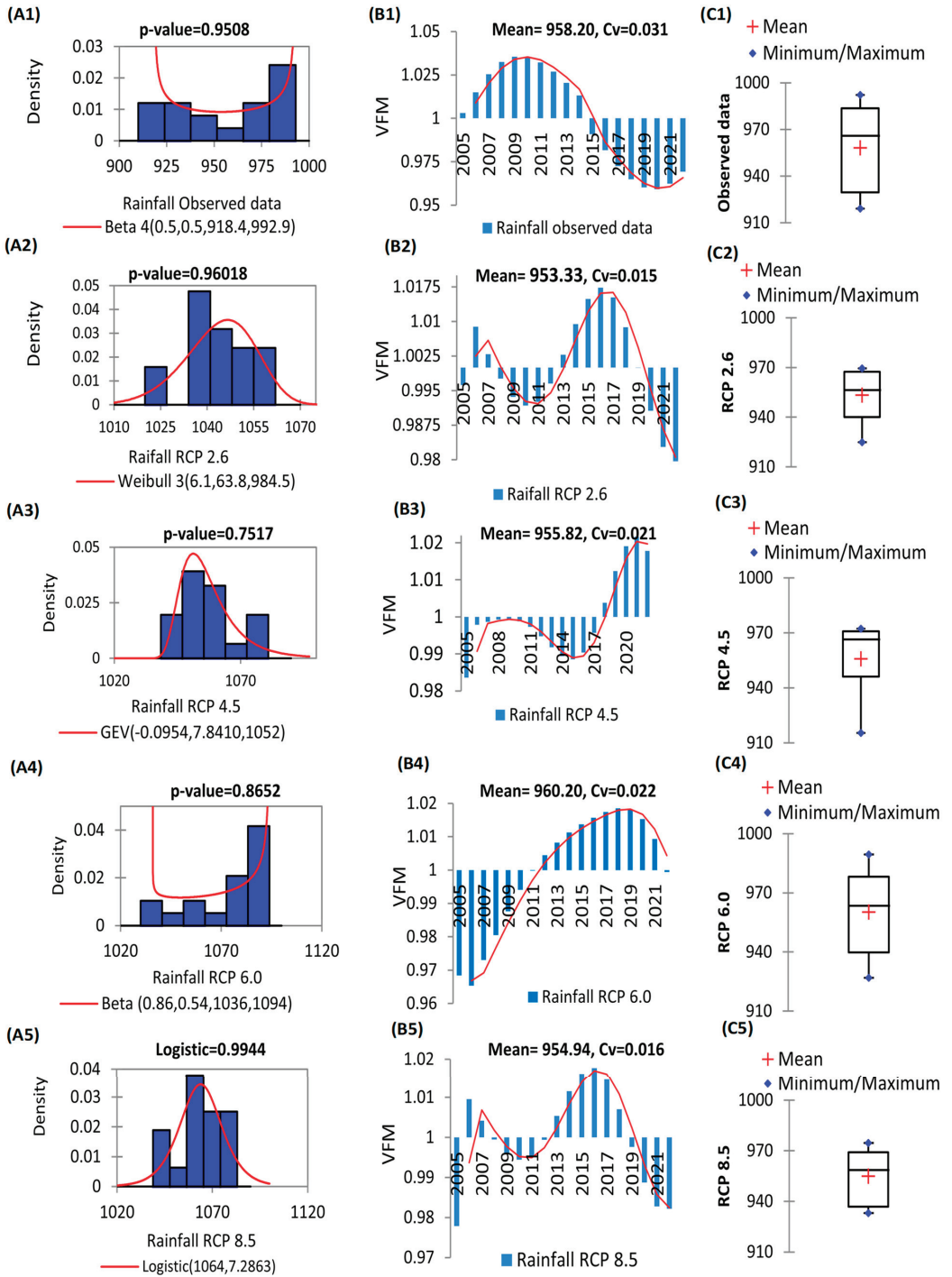


Figure 2. Statistical description of observed and projected annual rainfall data in the Trentino-Alto Adige region using RCP scenarios, given by density curve (A), histograms of values from the mean (B), and boxplots (C). CV: coefficient of variation.

5. Experimental Part

In this section, the model proposed to downscale CMIP5 rainfall data projection, obtained under different RCP scenarios, is represented in Figure 3. A flowchart summarizes in detail the three fundamental analyses step by step. The model process is a form of framework used to increase the quality of the input data. The observed data histories measured by the meteorological station are used firstly as the response variable of the downscaling model and secondly to control the performance of the outcome, while the overlapping data simulated by the RCM are selected in the first and second steps as the independent variable (X_i) of each sub-model.

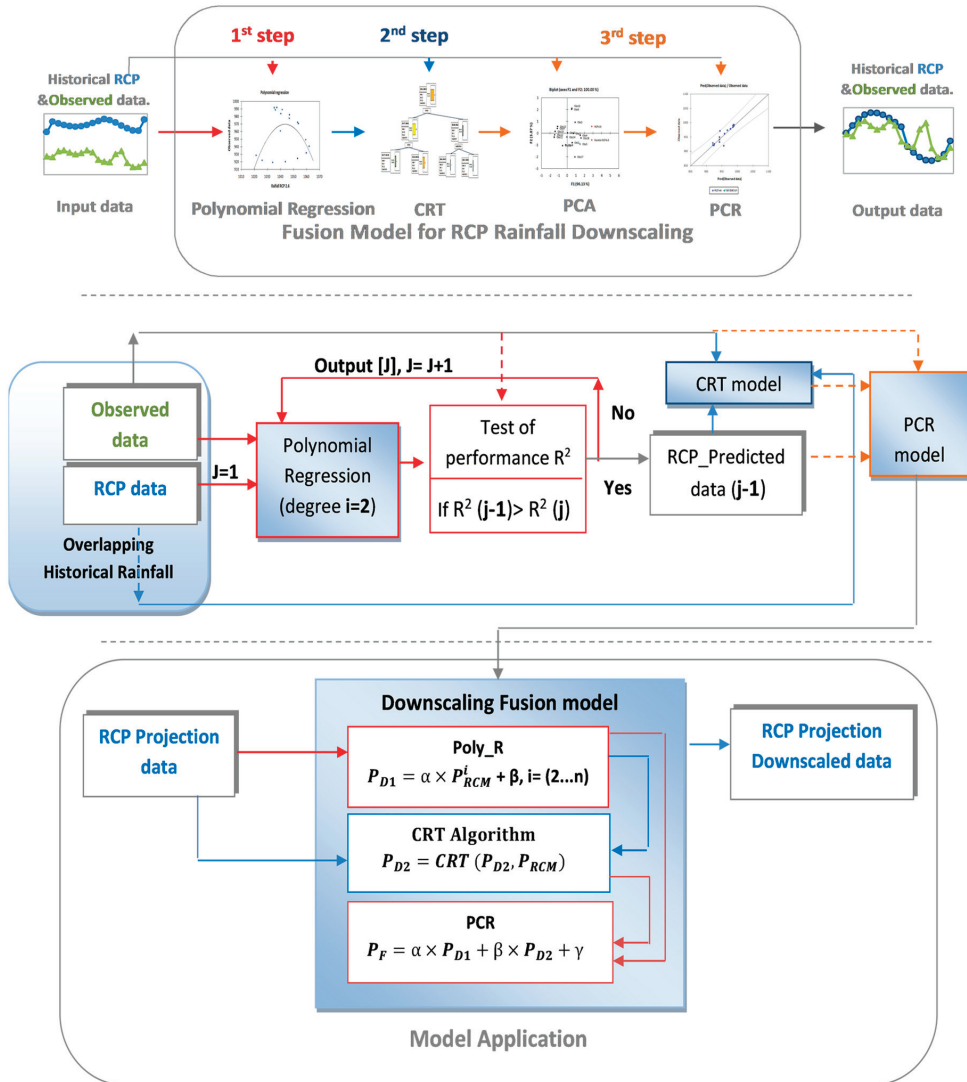


Figure 3. Flowchart summarizes the fusion model steps proposed for CMIP5 annual rainfall downscaling. R^2 : coefficient of determination; j : number of iterations. P_{D1} , P_{D2} , and P_F are the downscaled data by each sub-model. Poly_R: polynomial regression; PCA: principal component analysis; PCR: principal component regression; CRT: classification and regression tree.

The proposed model starts the procedure of data correction by using a non-linear adjustment between the real and simulated data for each RCP scenario to define the trend equation that will be applied for CMIP5 data downscaling (Table S2). The polynomial method produces a good response to rainfall data distribution. However, the model parameters vary from one scenario to another. For this reason, we have defined an iterative process by this method using the second degree of power. In the first iteration, the sub-model uses the projection data of each scenario as univariate. Then, a validation test will be applied to the predicted data using the R^2 to verify how the data fit with the real observation. The procedure iteratively uses the outcome results as input for the next step (Table S2). The downscale analysis stops when the R^2 of iteration (j) shows a value lower than the one obtained in iteration (j-1).

According to Figure 4, the application of the second degree of the polynomial model in two iterations gives a good fit, provided by an R^2 ranging between 0.52 and 0.61. On the other side, during the first iteration of applying Poly_R, the results show a good fit of 0.69 (Figure 4(A3)).

According to the graphs shown in Figure 4(B1–B4), a significant improvement is observed by the predicted data when compared with the CMIP5 data projected by the four scenarios. In the second step, the proposed model uses a multivariate classification via the application of CRT to the data predicted by the Poly_R model and projected by the RCP scenario as an independent variable of the CRT model. This classification helps to provide satisfactory results of data downscaling compared to the previous model. According to Figure 5(A1–A4), a good fit is observed between the actual data and the predicted series by the CRT model, which is demonstrated by an R^2 equal between 0.6194 and 0.8019.

In all scenarios, an improvement of the downscaled model was observed when comparing results to the first step of data correction. The application of the RCP model to downscale CMIP5 data by using the results obtained by both correction models (Poly_R and CRT) shows very good results. The PCR classifies the outcome data obtained by the previous models into clusters to estimate the standardized variable. This step helps to provide a very good estimation, which is proved by an R^2 ranging between 0.894 and 0.9466. The regression plots and the residual analysis show that the PCR model exhibits the best performance and provides a good response in all RCP scenarios, where the adjustment values fall within the confidence interval better than the CRT outcome. As a result, the fusion model produces a set of equations that will be used to downscale the CMIP5 rainfall data forecasting in the application phase.

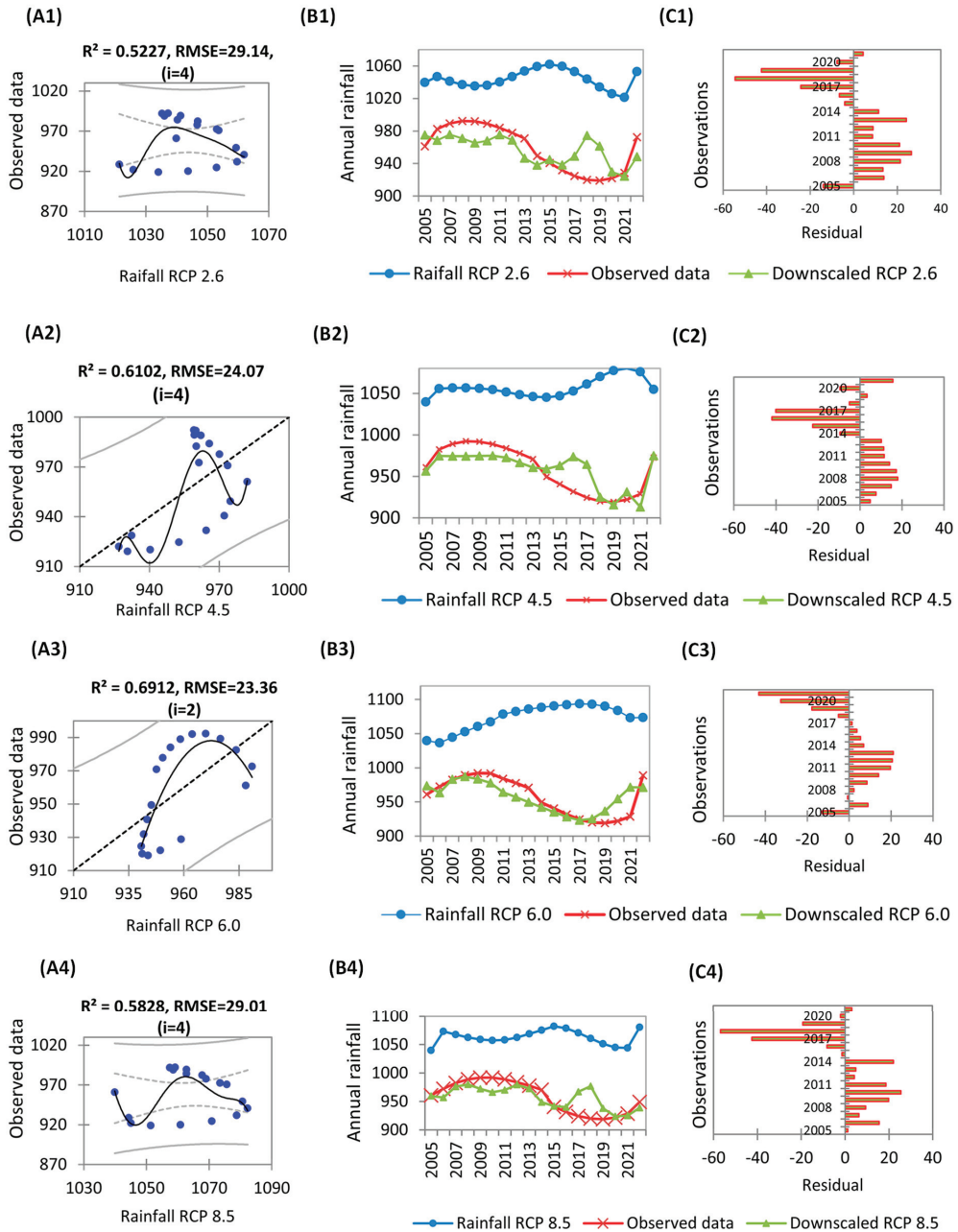


Figure 4. Non-linear regression scatter graphs (A,B) followed by residual histograms (C) to compare actual and RCP rainfall downscaling obtained by the Poly_R function. R^2 : coefficient of determination; RMSE: root mean square error.

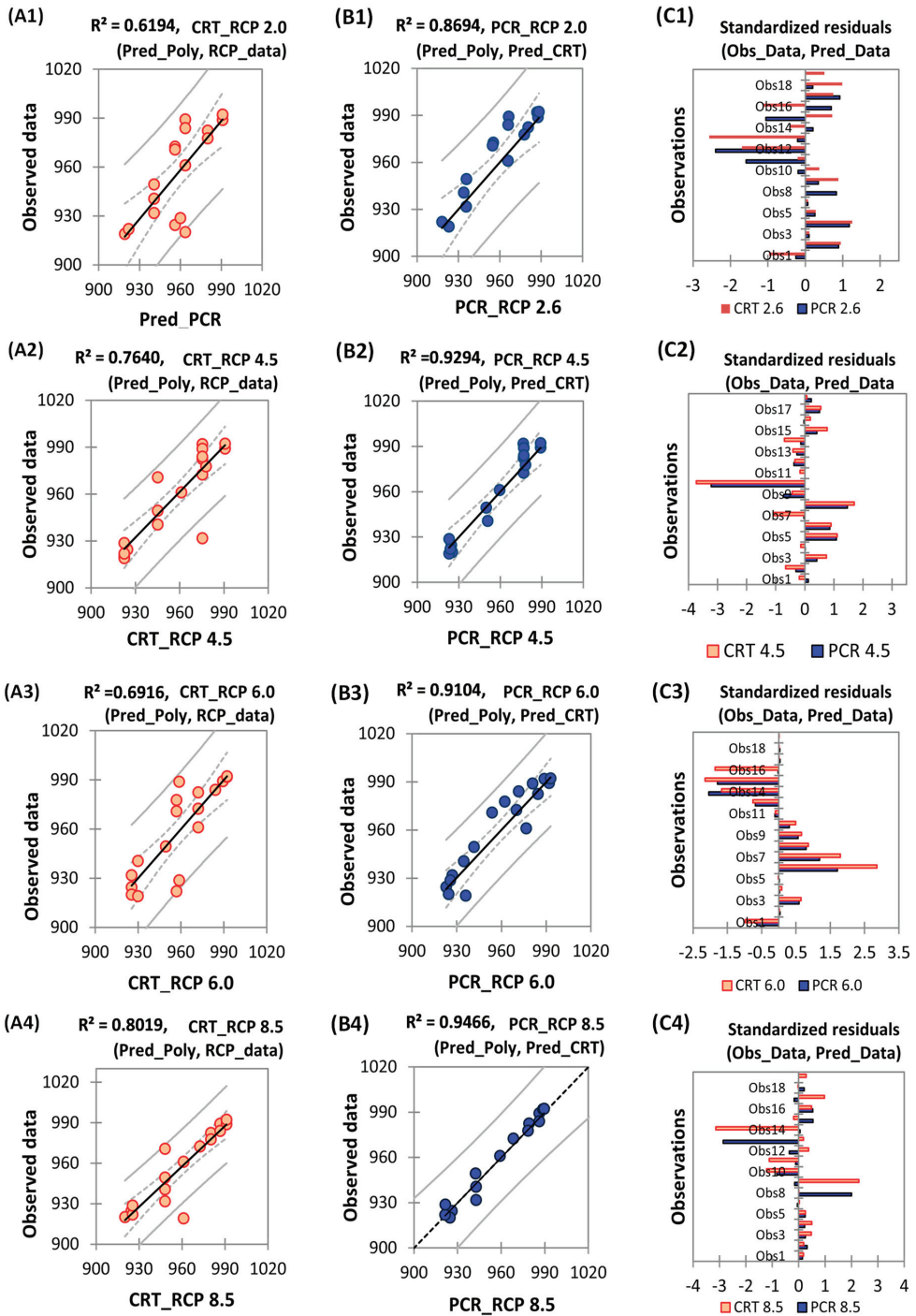


Figure 5. Regression scatter graph (A,B) and histogram of the residuals (C) between the actual and downscaled rainfall obtained by the CRT and PCR models. R^2 : coefficient of determination.

6. Validation and Performance

The performance and the validity of the proposed model are provided in this section by comparing the outcome results for each RCP scenario with predicted data obtained by LS and DCC downscaling models. Both models were applied using the SWAT software.

We used statistics metrics including R^2 , R^2_{Adj} , and RMSE to control the performance and the error tendency given by each model. A graphical representation for the whole predicted rainfall series is also given to monitor and compare in detail the downscaled values with the real one during the time period. In this part, the performance analysis applied to the CMIP5 data assessment under all RCP scenarios (2.6, 4.5, 6.0, and 8.5) is well explained in Figure 6.

The results show a very good performance of the proposed model with all projected rainfall data series, given by an R^2_{Adj} between 0.87 and 0.94. The model provides very low errors of RMSE, which vary between 5 mm to 10 mm. On the other hand, the LS model performs better than the DCC model when using data series obtained under the RCP 2.6 and 8.5 scenarios. However, with the data provided by the RCP 4.5 and 6.0 scenarios, the LS model produces fewer errors than the DCC downscaling model in each case of data processing. The histogram of the downscaled rainfall projection versus actual observation shows that the new model provides the best estimation between 2005 and 2022. When using rainfall data simulated under the 2.6 and 4.5 scenarios, the models gave only one underestimated value each which were observed in 2016 and 2014, respectively. On the other hand, the proposed model showed 2/14 underestimations of data when processing the data series obtained under the RCP 6.0 and RCP 8.5 scenarios. This bias estimation was observed in 2005 and 2018, respectively.

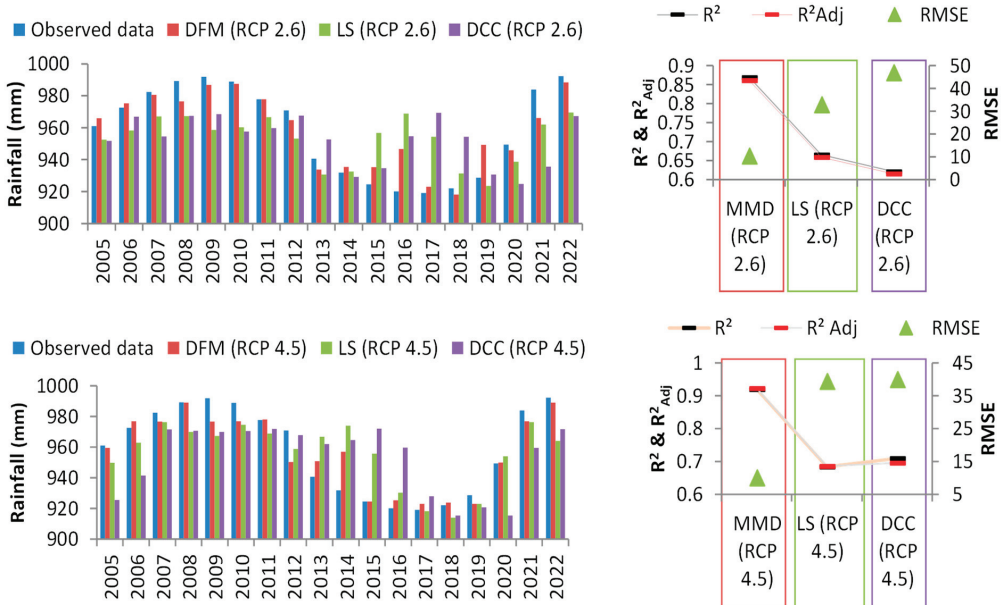


Figure 6. Cont.

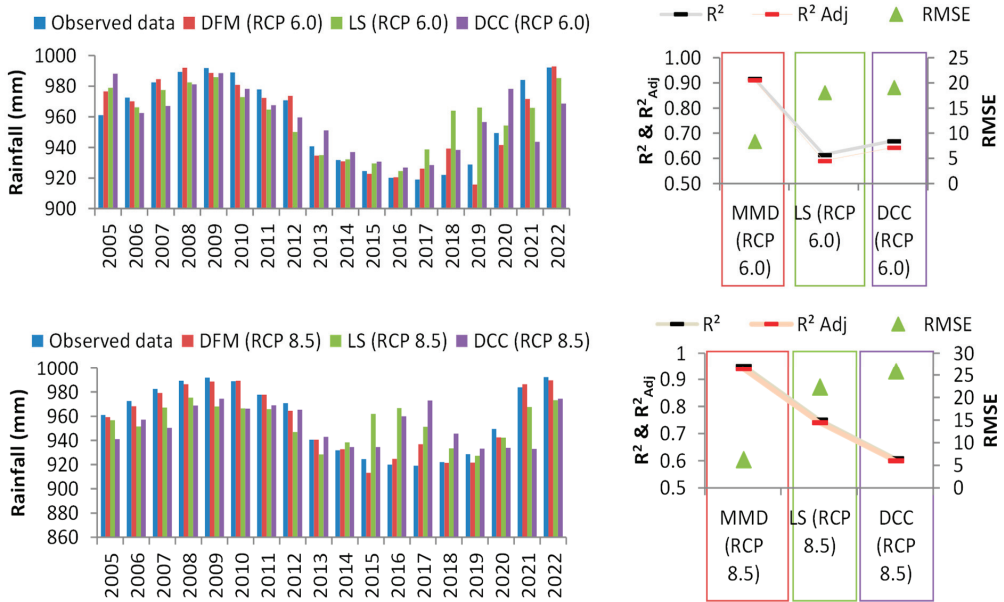


Figure 6. Histograms of actual and downscaled annual rainfall data obtained by the proposed downscaling fusion model (DFM) compared with LS and DCC downscaled models of SWAT, followed by statistical performance. R^2 : coefficient of determination; R^2 Adj: adjusted coefficient of determination; RMSE: root mean square error.

7. Conclusions and Summary

Climate change significantly impacts future biodiversity and the ecosystem. Good knowledge of several natural phenomena is based mainly on the good quality of the climate data projected by GCM and RCM. This work aims to propose a new method to downscale the CMIP5 rainfall data, under different RCP scenarios.

The new proposition is a fusion of three sub-models of the machine-learning family, which were applied to annual rainfall data observed in the Trentino-Alto Adige region between 2005 and 2022.

The first step was to iteratively apply a Poly_R model of a second-degree power on the rainfall simulated data by each scenario. A performance of 0.69 was observed after the first iteration of adjusting projected data by the Poly_R model. Then, improvements of RCP 2.6, 4.5, and 8.5 data downscaling by the Poly_R model were remarked in the second iteration, where the R^2 equaled between 0.52 and 0.62. The CRT model which was applied to the outcome data obtained by the previous model showed a good adjustment between 0.60 and 0.80. This performance was more noticeable when using rainfall data under RCP 4.5 and 8.5. Moreover, the application of the PCR model to downscaling data provided by both previous sub-models gave the best performance, which was proven by an R^2 between 0.86 and 0.94. The quality of the performance was also approved and compared against the LS and DCC models, where the proposed model proved the most efficient assessment in all RCP scenarios.

The good performance of this method using different scenarios shows its capacity for multiscale application. The method does not depend on the region of study because no physical parameters were used as input variables. This technique can also be employed to correct the estimation biases of several models in the hydro climatological field.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/engproc2023039055/s1>, Table S1: Statistic of observed and CMIP5 projected rainfall data in Trentino-Alto Adige under different RCP scenarios; Table S2: Performance analysis of polynomial regression model to downscale CMIP5 annual rainfall data projection using different iterations, followed by model's equation.

Author Contributions: All authors of this manuscript have directly participated in this study. A.A. worked on data collection, statistical analysis, modeling, validation, and comparison. A.L. worked on the research method, supervision, co-editing, and reviewing. I.K. worked on mapping. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study can be requested from the corresponding author. They are also available in the Climate Change Knowledge Portal, <https://climateknowledgeportal.worldbank.org/country/italy/cmip5> (accessed on 20 January 2023).

Acknowledgments: We wish to thank the Free University of Bozen-Bolzano, Italy, for providing advice and feedback.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sachindra, D.; Ahmed, K.; Rashid, M.; Shahid, S.; Perera, B. Statistical downscaling of precipitation using machine learning techniques. *Atmos. Res.* **2018**, *212*, 240–258. [CrossRef]
2. Noor, M.; Ismail, T.; Chung, E.-S.; Shahid, S.; Sung, J.H. Uncertainty in Rainfall Intensity Duration Frequency Curves of Peninsular Malaysia under Changing Climate Scenarios. *Water* **2018**, *10*, 1750. [CrossRef]
3. Onyutha, C.; Tabari, H.; Rutkowska, A.; Nyeko-Ogiramo, P.; Willems, P. Comparison of different statistical downscaling methods for climate change rainfall projections over the Lake Victoria basin considering CMIP3 and CMIP5. *J. Hydro-Environ. Res.* **2016**, *12*, 31–45. [CrossRef]
4. Trzaska, S.; Schnarr, E. *A Review of Downscaling Methods for Climate Change Projections*; United States Agency for International Development by Tetra Tech ARD: Washington, DC, USA, 2014; pp. 1–42.
5. Liu, D.L.; Zuo, H. Statistical downscaling of daily climate variables for climate change impact assessment over New South Wales, Australia. *Clim. Chang.* **2012**, *115*, 629–666. [CrossRef]
6. Busuioc, A.; Chen, D.; Hellström, C. Performance of statistical downscaling models in GCM validation and regional climate change estimates: Application for Swedish precipitation. *Int. J. Clim.* **2001**, *21*, 557–578. [CrossRef]
7. Hessami, M.; Gachon, P.; Ouara, T.B.; St-Hilaire, A. Automated regression-based statistical downscaling tool. *Environ. Model. Softw.* **2008**, *23*, 813–834. [CrossRef]
8. Bürger, G.; Chen, Y. Regression-based downscaling of spatial variability for hydrologic applications. *J. Hydrol.* **2005**, *311*, 299–317. [CrossRef]
9. Chen, J.; Brissette, F.P.; Leconte, R. Assessing regression-based statistical approaches for downscaling precipitation over North America. *Hydrol. Process.* **2013**, *28*, 3482–3504. [CrossRef]
10. Mami, A.; Raimonet, M.; Yebdri, D.; Sauvage, S.; Zettam, A.; Perez, J.M.S. Future climatic and hydrologic changes estimated by bias-adjusted regional climate model outputs of the Cordex-Africa project: Case of the Tafna basin (North-Western Africa). *Int. J. Glob. Warm.* **2021**, *23*, 58–90. [CrossRef]
11. Vu, M.T.; Aribarg, T.; Supratid, S.; Raghavan, S.V.; Liang, S.-Y. Statistical downscaling rainfall using artificial neural network: Significantly wetter Bangkok? *Theor. Appl. Climatol.* **2016**, *126*, 453–467. [CrossRef]
12. Laddimath, R.S.; Patil, N.S. Assessment of Future Meteorological Drought in Bhima basin based on CMIP5 Multi-model Projections. *Int. J. Future Gener. Commun. Netw.* **2020**, *13*, 2903–2911.
13. Krysanova, V.; White, M. Advances in water resources assessment with SWAT—An overview. *Hydrol. Sci. J.* **2015**, *60*, 771–783. [CrossRef]
14. Mahmood, R.; Jia, S. An extended linear scaling method for downscaling temperature and its implication in the Jhelum River basin, Pakistan, and India, using CMIP5 GCMs. *Theor. Appl. Clim.* **2016**, *130*, 725–734. [CrossRef]
15. Sarr, M.; Seidou, O.; Trambly, Y.; El Adlouni, S. Comparison of downscaling methods for mean and extreme precipitation in Senegal. *J. Hydrol. Reg. Stud.* **2015**, *4*, 369–385. [CrossRef]
16. Ostertagová, E. Modelling using Polynomial Regression. *Procedia Eng.* **2012**, *48*, 500–506. [CrossRef]
17. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.

18. Liu, Y.; Yao, L.; Jing, W.; Di, L.; Yang, J.; Li, Y. Comparison of two satellite-based soil moisture reconstruction algorithms: A case study in the state of Oklahoma, USA. *J. Hydrol.* **2020**, *590*, 125406. [CrossRef]
19. Liu, R.; Kuang, J.; Gong, Q.; Hou, X. Principal component regression analysis with spss. *Comput. Methods Programs Biomed.* **2002**, *71*, 141–147. [CrossRef] [PubMed]
20. LeGates, D.R.; McCabe, G.J., Jr. Evaluating the use of “goodness-of-fit” Measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **1999**, *35*, 233–241. [CrossRef]
21. Rosa, D.P.; Cantú-Lozano, D.; Luna-Solano, G.; Polachini, T.C.; Telis-Romero, J. Mathematical modeling of orange seed drying kinetics. *Ciênc. Agrotecnol.* **2015**, *39*, 291–300. [CrossRef]
22. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
23. Nikolopoulos, E.I.; Borga, M.; Marra, F.; Crema, S.; Marchi, L. Debris flows in the eastern Italian Alps: Seasonality and atmospheric circulation patterns. *Nat. Hazards Earth Syst. Sci.* **2015**, *15*, 647–656. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Resolution of Systems of Difference Equations and Its Implications for the VAR Model [†]

Gerardo Covarrubias * and Xuedong Liu *

Faculty of Higher Studies Aragon, National Autonomous University of Mexico, Ciudad Netzahualcóyotl 57000, Mexico

* Correspondence: jgcovarrubias@economia.unam.mx (G.C.); xdong@comunidad.unam.mx (X.L.)

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Systems of difference equations frequently present dynamically unstable solutions in the long term, which could imply the appearance of complications in the application of vector autoregressive (VAR) models in the Johansen sense, regardless of the precision required. In this work the necessary conditions are presented to guarantee the dynamical convergence of the solutions from the approach of the systems in discrete time series with the stochastic processes. The main aim is to show the importance of dynamic stability in structural-type models with respect to estimator bias.

Keywords: VAR model; systems of differential equations; cointegration in the long run

1. Introduction

In order to analyze the structure and/or forecast realizations of a stochastic process, or an observation within the time series, the models currently developed by econometricians have a certain degree of complexity. Thus, econometric estimation and the analysis of stochastic processes to explain economic phenomena through models are increasingly relevant, mainly in the context of time series, taking into consideration that various processes focus on understanding the dynamic structure of the series and on the possibility of forecasting its dynamic pattern of temporal behavior or the extrapolation of a stochastic process [1–4] where the lags of the variables involved play a key role in terms of the autoregressive models, as is the case in the estimation of autoregressive vector models (VARs), the central theme in this work.

These models are expressed through differential equations, since each variable is explained by the lags of both itself and the remaining variables.

It should be noted that each of the variables involved must meet the assumption of stationarity as a particular state of statistical equilibrium, where their probability distributions remain stable over time [5,6]. This implies that once the system is interrupted by some type of shocks, it will adjust back to equilibrium [7] or the shocks gradually disappear.

In estimating these models, it is accepted and often required that, if the estimators meet the tests, then they are the best linearly unbiased estimators (BLUEs). But what happens if the tests applied to the model are not fulfilled? Are the estimators not valid for the analysis? Does the model have to be scrapped?

2. Empirical Obtaining of Estimators in a VAR Model

It is important to point out that when there are large samples, the assumptions of normality, homoscedasticity, and the absence of autocorrelation in the errors are hardly fulfilled. This occurs regularly when using short duration data, for example monthly, quarterly, as well as long periods in the analysis. This could mean a limitation of the model that leads to strong criticism in this regard; however, as Wooldridge mentioned in his modern approach, given the law of large numbers, an asymptotic normality is

Citation: Covarrubias, G.; Liu, X. Resolution of Systems of Difference Equations and Its Implications for the VAR Model. *Eng. Proc.* **2023**, *39*, 56. <https://doi.org/10.3390/engproc2023039056>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 5 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

assumed, due to the size of the sample and in terms of the homoscedasticity and absence of autocorrelation in the errors, the results obtained allow themselves to be the best linearly unbiased estimators, pointed out by Guarati and Porter [8].

In this regard, it has agreed in common that for the forecasting purposes, the VAR models are required to fulfill the assumptions of the estimation: normality, homoscedasticity, and the absence of autocorrelation in the errors. However, if the estimated model is used only to analyze the structural changes of the economic variables, the requirement could be relaxed to the stability of the solutions; that is, the convergence in terms of the dynamic analysis, which can be determined by estimating the inverse roots of the characteristic polynomial of the autoregressive vector.

It should be noted that, since these are unrestricted models, the main advantage is that there will be no specification errors in the empirical estimation, in addition to the fact that the long-term cointegration solution is exempt from the problem of spuriousness or meaningless regressions, as it is defined by Granger and Newbold [9], with the initial idea owed to Yule [10].

Therefore, the VAR models not only provide a better estimate of forecasts compared to static ones, but also could be analyzed in a dynamic and structural manner where the importance of a shock of one variable on the others is revealed, with the relaxation of the related assumptions.

To apply the structural analysis of systems of simultaneous differential equations derived from the VAR approach, the series corresponding to Mexican exports to the US market and imports of Chinese origin are used with an annual periodicity from 2001 to 2020 (Figure 1).

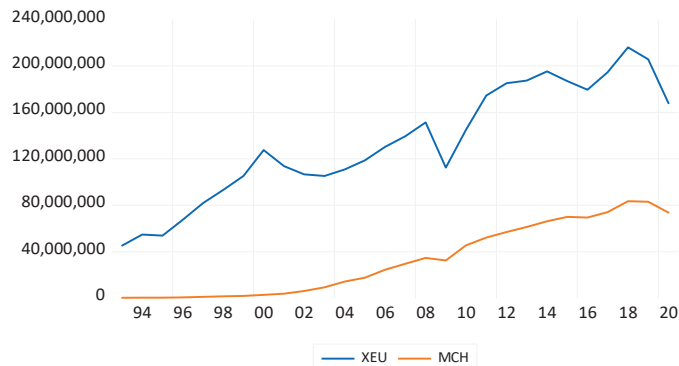


Figure 1. Mexican exports to the United States and Chinese imports to Mexico in millions of dollars.

It should be noted that the study Mexican trade with its two most important trading partners during the last 20 years is of great importance given the two facts. Firstly, China's entry into the World Trade Organization, an unprecedented event that contributed to the expansion of the Chinese products around the world markets, and eventually has become the second largest trading partner for Mexico since 2003. Secondly, at the beginning of 2020 COVID-19 pandemic has generated an important structural change in the trading among the three countries.

It is evident that both series are non-stationary since they have a trend. In order to estimate a VAR model, it is necessary for the series to be integrated in the same order to be transformed into stationarity; in this particular case, they are $I(1)$. Figure 2 shows the stationarity of the two series after the respective first difference, which was confirmed by augmented Dickey–Fuller [11] and Phillips–Perron [12] tests.

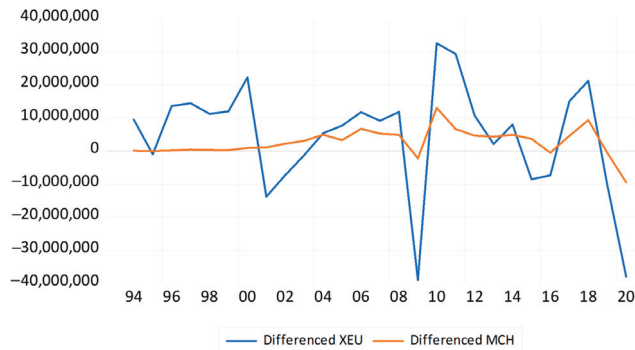


Figure 2. Mexican exports to the United States and Chinese imports to Mexico by first differences.

Subsequently, the VAR models were estimated with one and two lags respectively, estimations that accredited the stability tests within the unit circle and the residuals model’s presented normality, homoscedasticity and an absence of autocorrelation.

2.1. Solution of the VAR System of Equations with One Lag

The equation derived from the normalized equation estimated in the Eviews 12 Student Version Lite of S&P Global, New York, NY, USA with a model that shows stability within the unit circle (inverse roots of the autoregressive characteristic polynomial) is as follows:

$$XEU_t = 0.0613XEU_{t-1} + 1.142MCH_{t-1} + 98,954,092.05 \tag{1}$$

$$MCH_t = -0.1816XEU_{t-1} + 1.1944MCH_{t-1} + 23,399,484.6987 \tag{2}$$

$$XEU_t = 1.76 MCH_t \tag{3}$$

To simplify the notation, suppose that $XEU_t = A_t$ and $MCH_t = B_t$, so that

$$A_t = 0.0613A_{t-1} + 1.142B_{t-1} + 98,954,092.05 \tag{4}$$

$$B_t = -0.1816A_{t-1} + 1.1944B_{t-1} + 23,399,484.6987 \tag{5}$$

This implies that.

$$A_t - 0.0613A_{t-1} - 1.142B_{t-1} = 98,954,092.05 \tag{6}$$

$$B_t + 0.1816A_{t-1} - 1.1944B_{t-1} = 23,399,484.7 \tag{7}$$

Particular solutions that can be supposed for the two variables are $A_t = k_1$ and $B_t = k_2$, which implies that $A_{t-1} = k_1$ and $B_{t-1} = k_2$; then,

$$k_1 - 0.0613k_1 - 1.142k_2 = 98,954,092.05 \tag{8}$$

$$k_2 + 0.1816k_1 - 1.1944k_2 = 23,399,484.7 \tag{9}$$

Reducing yields are the following:

$$0.9387k_1 - 1.142k_2 = 98,954,092.05 \tag{10}$$

$$0.1816k_1 - 0.1944k_2 = 23,399,484.7 \tag{11}$$

Solving the system yields $k_1 = 300,576,585.66$ and $k_2 = 160,417,818.7$, while normalizing yields the following:

$$\frac{k_1}{k_2} = \frac{300,576,585.66}{160,417,818.7} = 1.8 \approx 1.76$$

Note that 1.8 is very close to the solution provided by the software (1.76).

To obtain the complementary solutions, the following can be carried out:

We assume that $A_t = \gamma\beta^t$ and $B_t = \delta\beta^t$ where γ and δ are constants; therefore, $A_{t-1} = \gamma\beta^{t-1}$ and $B_{t-1} = \delta\beta^{t-1}$.

Substituting into the difference equations that are now homogeneous,

$$\gamma\beta^t - 0.0613\gamma\beta^{t-1} - 1.142\delta\beta^{t-1} = 0 \tag{12}$$

$$\delta\beta^t + 0.1816\gamma\beta^{t-1} - 1.1944\delta\beta^{t-1} = 0 \tag{13}$$

Multiply everything by β^{1-t} , arriving at the following:

$$\gamma\beta - 0.0613\gamma - 1.142\delta = 0 \tag{14}$$

$$\delta\beta + 0.1816\gamma - 1.1944\delta = 0 \tag{15}$$

Grouping then yields

$$(\beta - 0.0613)\gamma - 1.142\delta = 0$$

$$0.1816\gamma + (\beta - 1.1944)\delta = 0$$

Obtaining the following:

$$(\beta - 0.0613)(\beta - 1.1944) - (0.1816)(-1.1944) = 0 \tag{16}$$

$$\beta^2 - 1.2557\beta + 0.2805 = 0 \tag{17}$$

Solving the equation, the values of β are $\beta_1 = 0.965$ and $\beta_2 = 0.2907$.

The complementary solutions are as follows:

$$A_t = \gamma_1(0.965)^t + \gamma_2(0.2907)^t \tag{18}$$

$$B_t = \delta_1(0.965)^t + \delta_2(0.2907)^t \tag{19}$$

Consequently, the general solutions are as follows:

$$A_t = \gamma_1(0.965)^t + \gamma_2(0.2907)^t + 300,576,585.66 \tag{20}$$

$$B_t = \delta_1(0.965)^t + \delta_2(0.2907)^t + 160,417,818.7 \tag{21}$$

2.2. Solution of the VAR System of Equations with Two Lags

Regrouping yields the following:

$$A_t = 0.176A_{t-1} - 0.451A_{t-2} + 1.48B_{t-1} + 0.064B_{t-2} + 131,598,419.78 \tag{22}$$

$$B_t = -0.1626A_{t-1} - 0.0724A_{t-2} + 1.236B_{t-1} + 0.023B_{t-2} + 28,642,607.5672 \tag{23}$$

$$1.27k_1 - 1.544k_2 = 131,598,419.78 \tag{24}$$

$$0.235k_1 - 0.259k_2 = 28,642,607.5672 \tag{25}$$

Solving the system results in $k_1 = 308,456,389.5$ and $k_2 = 169,284,339.7$. Normalizing then results in the following:

$$\frac{k_1}{k_2} = \frac{308,456,389.5}{169,284,339.7} = 1.8$$

In this case, the solution given by the software is 1.67; therefore, it is consistent.

Following this reasoning, we could generalize as follows:

Let be a VAR model of j variables with i lags:

$$\begin{aligned} X_{1t} &= \alpha_{11} X_{1t-1} + \alpha_{12} X_{1t-2} + \dots + \alpha_{1i} X_{1t-i} + \\ &\alpha_{21} X_{2t-1} + \alpha_{22} X_{2t-2} + \dots + \alpha_{2i} X_{2t-i} + \dots + \\ &\alpha_{j1} X_{jt-1} + \alpha_{j2} X_{jt-2} + \dots + \alpha_{ji} X_{jt-i} + C_1 \\ X_{2t} &= \beta_{11} X_{1t-1} + \beta_{12} X_{1t-2} + \dots + \beta_{1i} X_{1t-i} + \\ &\beta_{21} X_{2t-1} + \beta_{22} X_{2t-2} + \dots + \beta_{2i} X_{2t-i} + \dots + \\ &\beta_{j1} X_{jt-1} + \beta_{j2} X_{jt-2} + \dots + \beta_{ji} X_{jt-i} + C_2 \\ &\vdots \\ X_{jt} &= \gamma_{11} X_{1t-1} + \gamma_{12} X_{1t-2} + \dots + \gamma_{1i} X_{1t-i} + \\ &\gamma_{21} X_{2t-1} + \gamma_{22} X_{2t-2} + \dots + \gamma_{2i} X_{2t-i} + \dots + \\ &\gamma_{j1} X_{jt-1} + \gamma_{j2} X_{jt-2} + \dots + \gamma_{ji} X_{jt-i} + C_j \end{aligned}$$

We express the VAR model in difference equations:

$$\begin{aligned} X_{1t} - \alpha_{11} X_{1t-1} - \alpha_{12} X_{1t-2} - \dots - \alpha_{1i} X_{1t-i} - \\ \alpha_{21} X_{2t-1} - \alpha_{22} X_{2t-2} - \dots + \alpha_{2i} X_{2t-i} - \dots - \\ \alpha_{j1} X_{jt-1} - \alpha_{j2} X_{jt-2} - \dots - \alpha_{ji} X_{jt-i} &= C_1 \\ X_{2t} - \beta_{11} X_{1t-1} - \beta_{12} X_{1t-2} - \dots - \beta_{1i} X_{1t-i} - \\ \beta_{21} X_{2t-1} - \beta_{22} X_{2t-2} - \dots - \beta_{2i} X_{2t-i} - \dots - \\ \beta_{j1} X_{jt-1} - \beta_{j2} X_{jt-2} - \dots - \beta_{ji} X_{jt-i} &= C_2 \\ &\vdots \\ X_{jt} - \gamma_{11} X_{1t-1} - \gamma_{12} X_{1t-2} - \dots - \gamma_{1i} X_{1t-i} - \\ \gamma_{21} X_{2t-1} - \gamma_{22} X_{2t-2} - \dots - \gamma_{2i} X_{2t-i} - \dots - \\ \gamma_{j1} X_{jt-1} - \gamma_{j2} X_{jt-2} - \dots - \gamma_{ji} X_{jt-i} &= C_j \end{aligned}$$

Supposing that $X_{1t} = k_1$, $X_{2t} = k_2$ and $X_{jt} = k_j$, then $X_{1t-1} = k_1$, $X_{2t-1} = k_2$, and $X_{jt-1} = k_j$.

Analogously, if $X_{1t-i} = k_1$, $X_{2t-2} = k_2$ and $X_{jt-1} = k_j$ then,

$$\begin{aligned} k_1 - \alpha_{11} k_1 - \alpha_{12} k_1 - \dots - \alpha_{1i} k_1 - \\ \alpha_{21} k_2 - \alpha_{22} k_2 - \dots + \alpha_{2i} k_2 - \dots - \\ \alpha_{j1} k_j - \alpha_{j2} k_j - \dots - \alpha_{ji} k_j &= C_1 \end{aligned}$$

$$\begin{aligned}
 &k_1 - \beta_{11} k_1 - \beta_{12} k_2 - \dots - \beta_{1i} k_i - \\
 &\beta_{21} k_2 - \beta_{22} k_2 - \dots - \beta_{2i} k_i - \dots - \\
 &\beta_{j1} k_j - \beta_{j2} k_j - \dots - \beta_{ji} k_j = C_2 \\
 &\quad \vdots \\
 &k_1 - \gamma_{11} k_1 - \gamma_{12} k_2 - \dots - \gamma_{1i} k_i - \\
 &\gamma_{21} k_2 - \gamma_{22} k_2 - \dots - \gamma_{2i} k_i - \dots - \\
 &\gamma_{j1} k_j - \gamma_{j2} k_j - \dots - \gamma_{ji} k_j = C_j
 \end{aligned}$$

Factoring the constants yields the following:

$$\begin{aligned}
 &(1 - \sum_i \alpha_{1i}) k_1 - (\sum_i \alpha_{2i}) k_2 - \dots - (\sum_i \alpha_{ji}) k_j = C_1 \\
 &-(\sum_i \beta_{1i}) k_1 + (1 - \sum_i \beta_{2i}) k_2 - \dots - (\sum_i \beta_{ji}) k_j = C_2 \\
 &\quad \vdots \\
 &-(\sum_i \gamma_{1i}) k_1 - (\sum_i \gamma_{2i}) k_2 - \dots + (1 - \sum_i \gamma_{ji}) k_j = C_j
 \end{aligned}$$

In matrix form, this is represented as

$$\begin{bmatrix}
 (1 - \sum_i \alpha_{1i}) & -\sum_i \alpha_{2i} & \dots & -\sum_i \alpha_{ji} \\
 -\sum_i \beta_{1i} & (1 - \sum_i \beta_{2i}) & \dots & -\sum_i \beta_{ji} \\
 \vdots & \vdots & \ddots & \vdots \\
 -\sum_i \gamma_{1i} & -\sum_i \gamma_{2i} & \dots & (1 - \sum_i \gamma_{ji})
 \end{bmatrix}
 \begin{bmatrix}
 k_1 \\
 k_2 \\
 \vdots \\
 k_j
 \end{bmatrix}
 =
 \begin{bmatrix}
 C_1 \\
 C_2 \\
 \vdots \\
 C_j
 \end{bmatrix}$$

where $Ak = c, k = A^{-1}c$.

This results in a system of simultaneous linear equations with the number of variables equal to the number of equations, i.e., it will always be a square matrix.

The solution of the system constitutes the vector k of the long-term cointegration equation.

To achieve the above, that is, the general solutions converging to their particular solutions, the complementary solutions would have to be dynamically stable, or they would converge to nullity.

$X_{it} = A_{ii} b_i^t, i = 1, 2, \dots, n$ represent the number of variables.

Since all the solutions of the characteristic equation are real numbers, $b_j \in R$.

To fulfill the above, it is required that $|b_j| < 1$, and the particular solutions will be as follows:

$$\lim_{t \rightarrow \infty} X_{it} = \lim_{t \rightarrow \infty} A_{ii} b_i^t = 0 \tag{26}$$

On the other hand, to achieve $|b_i| < 1$, no doubt the necessary and sufficient condition is that the series involved in the system of difference equations has to be stationary, which is consistent with a single time series.

3. Concluding Remarks

Finally, it is possible to observe the way in which the solution of the VAR cointegration models was generalized for i variables and j lags where the estimators obtained are very useful in observing the structural components of the phenomenon to be analyzed, so that the idea of obtaining BLUEs is not a necessary condition. Even so, in the case of heterodasticity and autocorrelation in the errors, it is accepted that the estimator loses efficiency; that is, it is a good estimator, although it is not the best.

Author Contributions: Conceptualization, G.C. and X.L.; methodology, G.C. and X.L.; software, G.C.; validation, X.L.; formal analysis, G.C. and X.L.; investigation G.C. and X.L.; resources, X.L.; data curation, G.C.; writing—original draft preparation, G.C. and X.L.; writing—review and editing, G.C. and X.L.; supervision, X.L.; project administration, G.C.; funding acquisition, G.C. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Consejo Nacional de Ciencia y Tecnología grant number I1200/320/2022.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used for the estimates in this research are available on the website of the Secretary of Economy (Mexico): http://www.economia-snci.gob.mx/sic_php/pages/estadisticas/ accessed on 15 February 2023.

Acknowledgments: Conahcyt is widely recognized, specifically the Program of National Postdoctoral Stays and the Faculty of Higher Studies Aragón of the National Autonomous University of Mexico who provided the means to carry out this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Harvey, A. *Time Series Model*, 2nd ed.; Harvester Wheatsheaf: Hemel Hemstead, UK, 1993.
2. Maddala, G. *Introduction to Econometrics*, 3rd ed.; John Wiley and Sons Chichester: London, UK, 2001.
3. Guerrero, C. *Introducción a la Econometría Aplicada*; Editorial Trillas: Mexico City, Mexico, 2011.
4. Enders, W. *Applied Econometrics Time Series*, 4th ed.; Wiley: Hoboken, NJ, USA, 2015.
5. Box, G.; Jenkins, G. *Time Series Analysis: Forecasting and Control*; Holden Day: San Francisco, CA, USA, 1976.
6. Wooldridge, J. *Introducción a la Econometría: Un Enfoque Moderno*, 4th ed.; Cengage Learning: Mexico City, Mexico, 2010.
7. Juselius, K. *The Cointegrated VAR Model. Methodology and Applications*; Advanced Texts in Econometrics; Oxfröd University Press: Oxfröd, NY, USA, 2006.
8. Gujarati, D.N.Y.; Porter, D. *Econometría*, 5th ed.; McGrawHill: Mexico City, Mexico, 2010.
9. Granger, C.W.; Newbold, P. Spurious regressions in econometrics. *J. Econom.* **1974**, *2*, 111–120. Available online: <https://www.sciencedirect.com/science/article/pii/0304407674900347?via%3Dihub> (accessed on 16 July 2021). [CrossRef]
10. Yule, G. Why do we Sometimes get Nonsense-Correlations between Time-Series? A Study in Sampling and the Nature of Time-Series. *J. R. Stat. Soc.* **1926**, *89*, 1–63. Available online: <https://www.jstor.org/stable/2341482> (accessed on 25 June 2020). [CrossRef]
11. Dickey, D.A.; Fuller, W.A. Distribution of Estimators for Autoregressive Time Series with a Unit Root. *J. Am. Stat. Assoc.* **1979**, *74*, 427–431. Available online: <https://www.jstor.org/stable/2286348> (accessed on 4 July 2020).
12. Phillips, P.; Perron, P. Testing for a Unit Root in Time Series Regression. *Biometrika* **1988**, *75*, 335–346. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Goal-Oriented Transformer to Predict Context-Aware Trajectories in Urban Scenarios [†]

Álvaro Quintanar ^{1,*}, Rubén Izquierdo ¹, Ignacio Parra ¹ and David Fernández-Llorca ^{1,2}

¹ Computer Engineering Department, Universidad de Alcalá, 28801 Alcalá de Henares, Spain; ruben.izquierdo@uah.es (R.I.); ignacio.parra@uah.es (I.P.); david.fernandez-llorca@ec.europa.eu (D.F.-L.)

² Joint Research Centre, European Commission, 41092 Seville, Spain

* Correspondence: alvaro.quintanar@uah.es

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: The accurate prediction of road user behaviour is of paramount importance for the design and implementation of effective trajectory prediction systems. Advances in this domain have recently been centred on incorporating the social interactions between agents in a scene through the use of RNNs. Transformers have become a very useful alternative to solve this problem by making use of positional information in a straightforward fashion. The proposed model leverages positional information together with underlying information of the scenario through goals in the digital map, in addition to the velocity and heading of the agent, to predict vehicle trajectories in a prediction horizon of up to 5 s. This approach allows the model to generate multimodal trajectories, considering different possible actions for each agent, being tested on a variety of urban scenarios, including intersections, and roundabouts, achieving state-of-the-art performance in terms of generalization capability, providing an alternative to more complex models.

Keywords: trajectory prediction; urban scenarios; transformer; intelligent transportation systems

Citation: Quintanar, Á.; Izquierdo, R.; Parra, I.; Fernández-Llorca, D.

Goal-Oriented Transformer to Predict Context-Aware Trajectories in Urban Scenarios. *Eng. Proc.* **2023**, *39*, 57.

<https://doi.org/10.3390/engproc2023039057>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 5 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Motion forecasting is a vital component in the pipeline of an autonomous vehicle. It involves predicting the future motion of other vehicles, pedestrians, bicycles, and other objects in the environment in which the autonomous vehicle is operating. This information is crucial for the vehicle to make safe and efficient decisions, such as determining when to change lanes, slow down, or stop. Without accurate motion forecasting, the autonomous vehicle may make unsafe decisions or fail to respond in a timely manner to the actions of other road users. Moreover, forecasting is necessary and currently used for the creation of realistic simulations to test and validate the performance of autonomous vehicles before hitting the road, as well as essential for the development of cooperative systems, where multiple agents, both autonomous and human-driven, share the road. It allows the autonomous vehicle to anticipate the actions of other road users and plan its own motion accordingly, ensuring safe and efficient interactions.

In autonomous driving, it is essential to understand each driving situation in order to anticipate the trajectories of other agents. In each driving scenario, agents will react differently depending on traffic conditions and road structure. By knowing the behaviour of an agent a certain number of seconds in advance, it is possible to anticipate decisions, increasing safety and comfort for subsequent manoeuvres. Usually, agents will tend to take trajectories that are ideal for their goal, avoiding collisions and being socially accepted, i.e., following traffic rules and interacting with other agents on the road.

The problem of pedestrian trajectory prediction has been broadly explored by the community in the past years, being generally classified into two categories according to the type of analysis: pedestrians in crowded areas, where there may be erratic movements

due to low speed and avoidance of potential collisions, and environments shared with vehicles and other agents, where the traffic density is reduced but inter-class interaction is incorporated. This work, summarized as shown in Figure 1, continues the evolution of the previous one [1], essentially inspired by the initial research developed on pedestrians [2].

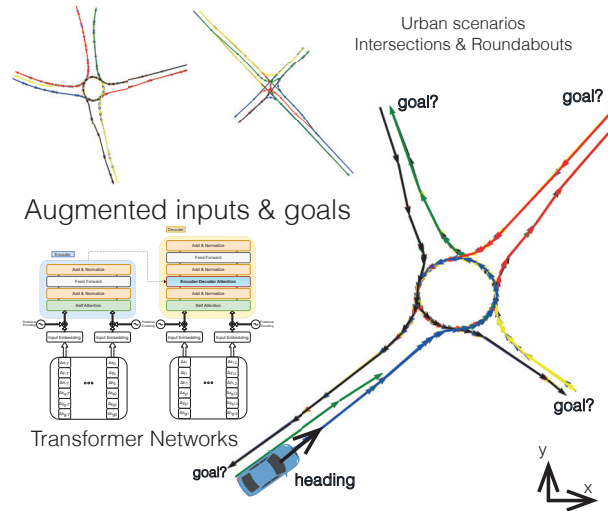


Figure 1. System overview.

2. Related Work

In the early stages of trajectory prediction, classical approaches relied essentially on linear regression, Bayesian filtering or Markov decision process. These methods performed properly, but since they are based on physical variables, their scaling and generalization are quite limited. After the arrival of deep learning, and specifically RNNs and LSTMs, it was found that they could successfully model the relationships between agents, exploiting their time dependency to predict future vehicle manoeuvres [3,4] and trajectories [5]. In this context of social approaches, S-LSTM [6] was proposed, connecting neighbouring LSTMs using a social pooling layer, predicting trajectories for multiple pedestrians. A similar approach was presented in [7] for vehicles. This was refined by SR-LSTM, making use of a message passing framework to enhance social nature [8]. Some models also propose an occupancy grid to define the interaction between agents [9,10]. Other authors have followed the line of generating a set of acceptable trajectories using architectures such as GAN [11,12] and CVAE [13]. In GAN, the generator and discriminator are used in a complementary way to improve the generation and detection of valid trajectories, while CVAE is used to encode in a latent space and generate multi-path trajectories based on the observed paths.

Using the Vanilla-TF as a model, the context-augmented Transformer network [14] uses interaction and semantic information as the input to provide robust prediction on datasets with strong pedestrian–vehicle interactions, similar to the inD dataset.

In parallel to these deep learning-based approaches, OSP [15] proposes a traditional probabilistic approach, developing a pedestrian–vehicle interaction model that outperforms models such as S-GAN and MATF with real-time execution speed that is really convenient.

Although LSTMs seem to be a good model for learning trajectory sequences, they are inefficient at modelling data in long temporal sequences, and thus suffer more from the lack of input data in observations, a very common issue in real systems involving physical sensors. In this way, Transformer models [16] have been successfully adapted to predict pedestrian trajectories in crowded spaces [2], achieving state-of-the-art results in TrajNet benchmark [17], by relying only on self positional information (i.e., without adding any

social or interactive data). Moving beyond pedestrians, this paper will focus on vehicle trajectories, whose interaction is rather intense in the environments analysed (intersections and roundabouts).

Recent work has explored including the road graph, history and interaction between agents using more sophisticated models and a bespoke architecture for each type of input [18]. Whereas, others have employed images and detections of mixed traffic environments to provide an explainable nature to their model, developing an important analysis concerning this issue [19].

In this work, a Transformer model is used in its simplest form, exploiting its nature to adapt the inputs and improve the results without major changes in the architecture that could lead to greater complexity in its training and use, exploring its capabilities with augmented input data such as velocity and orientation, analysing its performance on vehicles in various datasets, and performing cross tests to assess its generalization capability.

3. Methodology

This section addresses the methodology used to deploy the model, starting with the selection of the input and output data, the preprocessing and analysis of the input data for the BEV datasets used in the study, and the creation of the enhanced model, analysing the different transformations made to adopt the new information. In addition, the use of context information through data provided by the digital maps present in each scenario will be covered, using the lanelet2 library to compute positions with respect to lanes, off-road zones and routes to goals, among others. The approach of the “post hoc” multimodality paradigm using the potential goals for each agent in the scenario is fully discussed at the end of this section.

3.1. Introducing the Problem

Let $X_t = \{x_t, v_t, a_t\}$ be the state of the vehicle at time t , where x_t is the position, v_t is the velocity, and a_t is the acceleration. Let Y_t be the set of environmental conditions at time t . The goal of trajectory prediction is to estimate the future trajectory $T_t = \{T_{t1}, T_{t2}, \dots, T_{tN}\}$ given the observations $O_t = \{X_t, Y_t\}$ up to time t and a prediction horizon N .

3.2. Inputs and Outputs

In our work, these inputs are the velocity (position increments) and the heading increment of the agent under study itself, in combination with the same information with respect to the possible goal it may have in the testing scenario. Thus, $X_t = \{\Delta x_e, \Delta y_e, \Delta h_e\}$, and $Y_t = \{\Delta x_g, \Delta y_g, \Delta h_g\}$. Currently, there are several benchmarks that consider different time horizons, both for visualization and prediction. TrajNet was followed in the previous work: a benchmark in which datasets are measured at 2.5 Hz, observing 8 frames (3.2 s) and predicting 12 frames (4.8).

It is important to highlight the importance of working with increments of positions and headings, rather than directly with the absolute data. Previous tests showed that the model failed to learn with this approach, yielding a sub-par performance on the datasets under analysis. This can show the nature of the data being used, allowing minor variations in velocity that make it easier to predict a more constant output, aside from data filtering with Kalman-derived filters, which will tend to follow the preceding frame velocity. Thus, comparatively, we have also considered the input of the heading increment in degrees in absolute form, without any previous adaptation that could pre-normalize it, since it will be performed in the training and testing process. We think that this pre-normalization developed in 2021 could have affected and worsened the results, as explored in the ongoing experiments. The complete model overview is depicted in Figure 2.

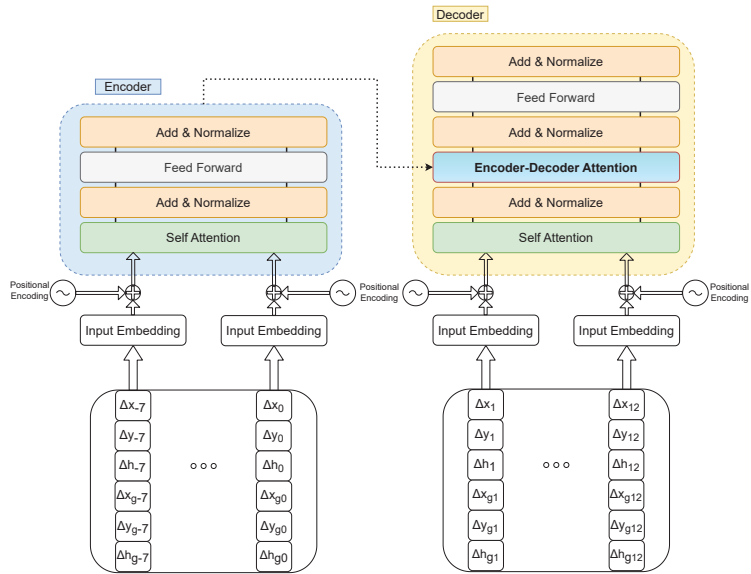


Figure 2. Architectural overview: the addition of new inputs.

3.3. Exploring Context Information

After studying the scenarios' topology for each dataset, and the available data, we considered the option of incorporating contextual data, taking into consideration that the datasets used have digital maps within the lanelet2 library framework, allowing access to context information for each lanelet, such as the distance to the centre of the lane, distance to the nearest curbs, no-go zones for driving in the case of vehicle agents, etc. However, powerful information included in these maps concerns the routing graph, such that knowing the position of an agent permits delimiting the possible routes it can follow in the scenario, according to the traffic rules. The exact knowledge of the map and traffic rules also allows to extend to the social factor, where metrics such as IDM (car-following model [20]) or RSS [21] can be computed to analyse possible dangerous situations involving near agents.

3.4. The Architecture

3.4.1. Data Preprocessing

For the datasets used it was necessary to carry out a prior stage of data analysis and extraction in order to properly format them for model input and planned experiments. During this stage, the parked vehicles present in some recordings were removed, and the frame rate was taken into account to adapt it to the desired frame rate for the study, with sequences at 2.5 Hz. Thus, the initial input consists of the *location, frame, track, x, y, heading* structure, to then go through the feature addition module, where the corresponding increments are calculated and the goal information is introduced based on the map. After this process, the input to the model includes $\Delta x, \Delta y, \Delta h, \Delta x_g, \Delta y_g$ and Δh_g . The heading data are introduced in absolute values between 0 and 360° , adapting the entries of each dataset appropriately, whereas the distance to the centre of the lane in the corresponding tests are entered in modulus and SI units.

3.4.2. Details

The architecture from [1] was maintained, with the addition of an L2 loss that includes position increments for improved independence of each position, as well as normalized heading. The d_{model} was set to 512, with 6 layers and 8 attention heads. A warm-up period

of 10 epochs was implemented, employing a decaying learning rate in the subsequent epochs.

3.5. Post Hoc Multimodality

To assess the model’s ability to know the intrinsic structure of the scenario without receiving explicit information about it, a “post hoc multimodality” approach was adopted. This consisted of generating five trajectories for each of the goals existing in the test scenario. This was calculated through the routes present in the route graph for each scenario of the inD and rounD datasets. Table 1 shows the number of routes and goals present in each scenario.

Table 1. Number of routes and global goals per scenario and dataset.

Dataset Scenario	# of Routes	# of Goals
inD: 1	13	4
inD: 2	12	4
inD: 3	6	3
inD: 4	12	3
rounD: 0	36	4
rounD: 1	17	4
rounD: 2	17	4

4. Experiments and Results

4.1. Datasets

In addition to pedestrian-centric approaches, the NGSIM datasets [22,23] were pioneers in covering highway areas, with information obtained from cameras mounted on a skyscraper. Several multi-agent datasets have been developed over the past few years, with a focus on highway scenarios, such as the highD dataset [24] for highway vehicle trajectory prediction. This dataset provides aerial images obtained using a drone located over various locations of the German autobahn, with vehicle labelling ensuring an error below 10 cm. The dataset provides a total of 147 h of drive time on over 100,000 vehicles. The authors of this dataset went further and expanded the concept to urban scenarios, with the inD [25] and rounD [26] datasets recording different intersections and roundabouts, respectively, as well as the novel exiD [27], that covers some stretches at mergings. The Interaction dataset [28] combines all these scenarios, including ramp merging, signalized intersections, and roundabouts. This dataset also provides a diverse range of driving behaviours, including critical manoeuvres, and even accidents. These situations add value to a trajectory prediction solution and should be evaluated in a qualitative manner. Table 2 overviews the datasets used to develop the experiments.

Finally, while 2D datasets taken from drones or fixed locations from a bird’s eye view are relatively easy to create and label, the ultimate goal is to train models that can be ported to vehicles equipped with onboard sensors and tested in datasets such as NuScenes [29], Argoverse [30] or Prevention [3].

4.2. Goal Analysis

Goal evaluation for each dataset was carried out automatically on the routes contained in the digital map graphs for each scenario. The training was carried out with the real target, and then tests were performed for each of the scenario targets, generating five trajectories for each one and choosing the ones with the lowest error. This approach brings variability to the results and a “post hoc multimodality” method similar to that conducted in other published research, differing in that in this case we are sampling directly in the tests with the possible targets present in each map rather than using a distribution for each mode.

Table 2. Datasets used in this work.

Dataset	inD	rounD	Interaction
Country	Germany	Germany	USA Germany China
Locations	urban intersections (4)	(sub-)urban roundabouts (3)	roundabout (5), intersection (4), highway (2)
# of Tracks	11,500	13,746	40,054
Road User Types	pedestrian, bicycle, car, truck, bus	pedestrian, bicycle, motorcycle, car, van, truck, bus, trailer	pedestrian/bicycle, car, truck
Data Frequency	25 Hz	25 Hz	10 Hz
Maps	yes	yes	yes

4.3. Evaluation Metrics

The metrics employed are the state-of-the-art standards for the datasets considered here, average displacement error (ADE) and final displacement error (FDE). The ADE/MAD calculates the difference in the L2 norm between the 12 points of the predicted trajectory and compares them with the respective ground truth in metres, while the FDE/FAD only accounts for the last observation of this prediction. Thus, the ADE indicates a general fit of the predicted and actual trajectories. This can be questionable, as the predicted trajectories cannot deviate too far from the actual trajectory but enter prohibited zones for the corresponding agent, leading to situations where the predictions for vehicles end up entering pedestrian pavements. Due to this, other metrics are considered in this work, such as the off-road rate or miss-rate, that will be explored in future tests with the datasets that embody them. The experiments performed in this case (i.e., for quantitative analysis) have been deployed with the real goal corresponding to each agent, while the complete analysis of the “post hoc multimodality” is reserved for the qualitative analysis. Implementing typical metrics, such as min-ADE, are more commonly used in other datasets than the ones involved in our work.

4.4. inD: Comparative Results

Using the same data split used by the authors of the DCENet to make an objective comparison, we obtained comparative analysis results for the inD dataset, as shown in the Table 3. These results include all agent types, not just vehicles, meaning the goal approach is less effective than splits that include vehicles only, as discussed later. This dataset includes parts of the test scenarios in the training split, so the model is already familiar regarding the trajectories that agents can perform, which, when combined with the fact that pedestrians are also being evaluated, reduces global errors. The Goal-TF model still outperforms the “typical” architectures, S-LSTM and S-GAN, and improves the results of their homonymous TFs which include less information; however, the model still underperforms against AMENet and DCENet. However, we can appreciate that the inclusion of the target has been positive, reducing the FAD by more than 20 cm with respect to the Oriented-TF. In the following experiments the set of tested agents will be reduced to vehicles (cars, trucks, vans, trails, buses, etc.).

Table 3. General performance.

InD	Average (MAD/FAD)
S-LSTM	1.88/4.47
S-GAN	2.38/4.66
AMENet	0.73/1.59
DCENET	0.69/1.52
Vanilla-TF	1.07/2.65
Oriented-TF	1.02/2.57
Goal-TF	0.94/2.34

4.5. Testing in Different Datasets

4.5.1. Single Dataset Tests

This section reports the results of the leave-one-out (LOO) technique for the inD, round and Interaction datasets for their intersection and roundabout variants, compared with the Vanilla and Oriented models, where the heading is introduced as additional information. As shown in Table 4, the Goal-TF model is better in all cases than the other models, except in the Interaction-GL scenario, where the Oriented model stands out. The improvements are substantial, with a difference greater than 4 m in the FDE in some experiments. The inclusion of the target is considered beneficial in terms of the additional information that the network is able to learn and understand for trajectory prediction.

Table 4. Single dataset tests.

Training // Test	Vanilla-TF ADE / FDE	Oriented-TF ADE / FDE	Goal-TF ADE / FDE
inD: 123 // 4	7.67/17.22	7.71/16.83	6.61/13.90
inD: 134 // 2	2.80/7.46	3.47/9.02	2.62/6.43
inD: 234 // 1	1.91/5.18	1.89/5.14	1.61/3.97
round: 01 // 2	6.59/16.87	6.62/17.09	5.26/11.81
round: 02 // 1	6.64/17.04	6.88/17.53	5.05/11.76
round: 12 // 0	6.68/16.71	7.98/19.82	7.50/15.06
INT-intersection: EP0-EP1-MA // GL	2.54/6.95	2.10/5.66	2.36/6.18
INT-roundabout: SR-FR-EP-OF // LN	4.46/11.65	3.81/9.51	2.56/6.27
INT-intersection: MA-GL-EP0 // EP1	3.27/8.17	2.80/7.16	1.96/4.94
INT-roundabout: LN-SR-FT-EP // OF	4.27/11.63	3.68/10.11	2.75/6.66

4.5.2. Mixing Datasets: Similar Scenarios of Different Datasets

In the case of cross-dataset generalization, it seems that the choice of method when introducing additional information may penalize the Goal model, with the Vanilla model remaining the best option if transfer learning between datasets becomes the preferred method, as shown in the Table 5.

Table 5. Equivalent scenario tests (training on an entire dataset).

Training // Test	Vanilla-TF ADE/FDE	Oriented-TF ADE/FDE	Goal-TF ADE/FDE
inD // INT-int	3.12/8.10	4.89/10.87	3.57/8.58
INT-int // inD	4.04/10.10	4.24/10.32	3.09/7.52
round // INT-round	3.19/8.34	5.18/11.72	5.69/12.59
INT-round // round	5.30/14.13	6.99/16.54	3.48/8.58

4.6. Qualitative Results

Apart from the quantitative results measured by the corresponding metrics, it is necessary to assess thoroughly how an agent actually behaves in practice when a specific situation occurs in a particular scenario; for example, at an intersection with different exits. Figure 3 shows an instance prior to a turn where the vehicle has slowed down when approaching the intersection. Thus, it can be seen how the prediction can yield various results depending on the target in question. In one case, the vehicle will continue straight ahead, while in another the prediction outputs the vehicle turning in one way or the other. However, at other times the model will also be completely wrong, leading to completely erroneous predictions, such as when the vehicle is meant to continue straight ahead and the prediction is a turning prediction, or vice versa. Figure 4 briefly shows multiple trajectories generated at the approach of a roundabout according to the selected goal, including the option of a complete turn to change direction.

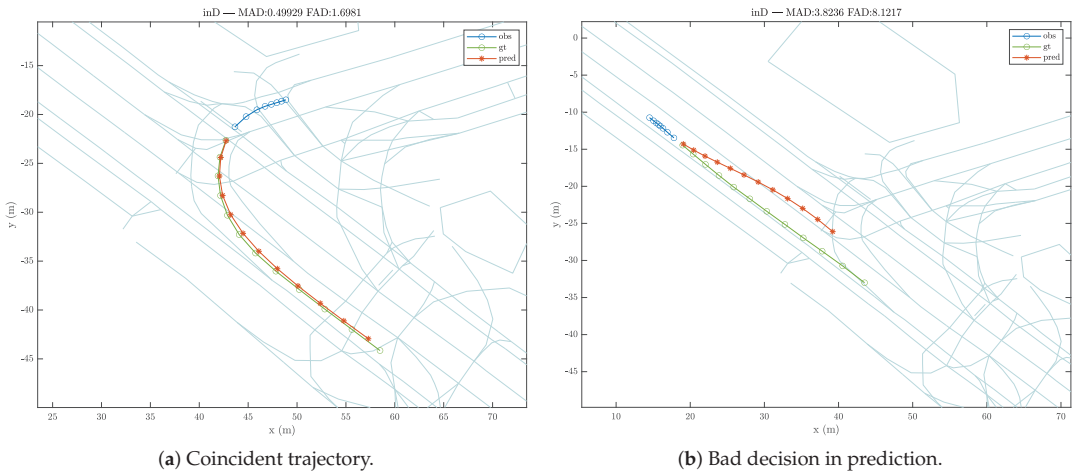
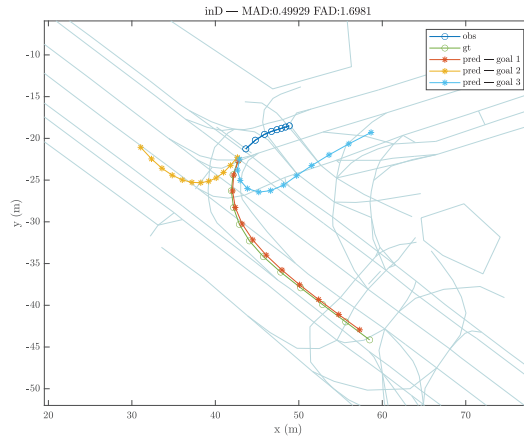


Figure 3. Cont.



(c) Trajectories predicted when selecting another goal.

Figure 3. Sample outputs for leave-one-out experiments using the inD dataset | location 3. Observed trajectory is depicted in blue, ground truth in green and predicted trajectory in orange (view legend).

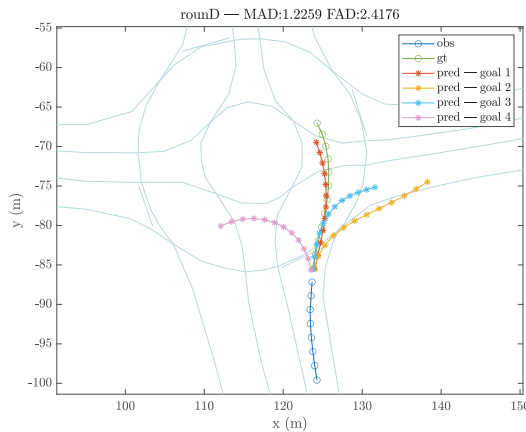


Figure 4. Instances before entering a roundabout depending on the goal in the roundD dataset.

5. Conclusions and Future Work

The experiments performed showed that the inclusion of context variables relative to the goal obtained from the digital map routes linked to each scenario improved the results compared to models that did not use them. This allowed for multimodal trajectory generation, an important point that should be developed in future work. The generalization of this model was also discussed, with tests on different datasets highlighting its high versatility. In future, the challenge to integrate social information needs to be addressed, exploring a way to introduce simultaneous data from several agents to allow for the generation of socially aware trajectories. Furthermore, tasks such as the extension of multimodality tests to all datasets, providing specific metrics, or extending the datasets to other existing datasets in the field, such as NuScenes or Argoverse, is still pending. Finally, a viability study would be beneficial for the implementation of such a system for real-time inference using real-time information collected by an vehicle to demonstrate whether these models are ready to be deployed in the real world.

Author Contributions: Conceptualization, Á.Q., I.P. and D.F.-L.; methodology, Á.Q., I.P. and D.F.-L.; software, Á.Q.; validation, Á.Q. and R.I.; investigation, Á.Q. and R.I.; resources, I.P. and D.F.-L.; writing Á.Q., R.I., I.P. and D.F.-L.; supervision, I.P. and D.F.-L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Research Grants PRE2018-084256, PID2020-114924RB-I00 and PDC2021-121324-I00 (Spanish Ministry of Science and Innovation) and partially by S2018/EMT-4362 SEGVAUTO 4.0-CM and CM/JIN/2021-005 (Community of Madrid). D.F.-L. acknowledges the funding from the HUMAINT project by the Directorate-General Joint Research Centre of the European Commission.

Data Availability Statement: Datasets were obtained from [25,26] and are available at their corresponding websites (with their permission of their authors).

Conflicts of Interest: The authors declare no conflict of interest. The views expressed in this article are purely those of the authors and may not, under any circumstances, be regarded as an official position of the European Commission.

References

1. Quintanar, A.; Fernández-Llorca, D.; Parra, I.; Izquierdo, R.; Sotelo, M.A. Predicting Vehicles Trajectories in Urban Scenarios with Transformer Networks and Augmented Information. In Proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV), Nagoya, Japan, 11–17 July 2021; pp. 1051–1056.
2. Giuliari, F.; Hasan, I.; Cristani, M.; Galasso, F. Transformer Networks for Trajectory Forecasting. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 10335–10342.
3. Izquierdo, R.; Quintanar, A.; Parra, I.; Fernández-Llorca, D.; Sotelo, M.A. The PREVENTION dataset: A novel benchmark for PREDiction of VEHicles iNTentIONS. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; pp. 3114–3121.
4. Biparva, M.; Fernández-Llorca, D.; Gonzalo, R.I.; Tsotsos, J.K. Video Action Recognition for Lane-Change Classification and Prediction of Surrounding Vehicles. *IEEE Trans. Intell. Veh.* **2022**, *7*, 569–578. [CrossRef]
5. Izquierdo, R.; Quintanar, A.; Parra, I.; Fernández-Llorca, D.; Sotelo, M.A. Vehicle Trajectory Prediction in Crowded Highway Scenarios Using Bird Eye View Representations and CNNs. In Proceedings of the IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020.
6. Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; Savarese, S. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 961–971.
7. Deo, N.; Trivedi, M.M. Convolutional Social Pooling for Vehicle Trajectory Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1549–15498.
8. Zhang, P.; Ouyang, W.; Zhang, P.; Xue, J.; Zheng, N. SR-LSTM: State Refinement for LSTM Towards Pedestrian Trajectory Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12077–12086.
9. Pfeiffer, M.; Paolo, G.; Sommer, H.; Nieto, J.; Siegart, R.; Cadena, C. A Data-driven Model for Interaction-Aware Pedestrian Motion Prediction in Object Cluttered Environments. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 5921–5928.
10. Manh, H.; Alagband, G. Scene-LSTM: A Model for Human Trajectory Prediction. *arXiv* **2019**, arXiv:1808.04018.
11. Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; Alahi, A. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2255–2264.
12. Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; Savarese, S. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1349–1358.
13. Zhao, T.; Xu, Y.; Monfort, M.; Choi, W.; Baker, C.; Zhao, Y.; Wang, Y.; Wu, Y.N. Multi-Agent Tensor Fusion for Contextual Trajectory Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12118–12126.
14. Saleh, K. Pedestrian Trajectory Prediction using Context-Augmented Transformer Networks. *arXiv* **2020**, arXiv:2012.01757.
15. Anderson, C.; Vasudevan, R.; Johnson-Roberson, M. Off The Beaten Sidewalk: Pedestrian Prediction In Shared Spaces For Autonomous Vehicles. *IEEE Robot. Autom. Lett.* **2020**, *5*, 6892–6899. [CrossRef]
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Łukasz, K.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5999–6009.

17. Sadeghian, A.; Kosaraju, V.; Gupta, A.; Savarese, S.; Alahi, A. Trajnet: Towards a benchmark for human trajectory prediction. *arXiv* **2018**, submitted for publication.
18. Nayakanti, N.; Al-Rfou, R.; Zhou, A.; Goel, K.; Refaat, K.S.; Sapp, B. Wayformer: Motion Forecasting via Simple & Efficient Attention Networks. *arXiv* **2022**, arXiv:2207.05844.
19. Zhang, Z.; Tian, R.; Sherony, R.; Domeyer, J.; Ding, Z. Attention-Based Interrelation Modeling for Explainable Automated Driving. *IEEE Trans. Intell. Veh.* **2023**, *8*, 1564–1573. [CrossRef]
20. Treiber, M.; Hennecke, A.; Helbing, D. Microscopic Simulation of Congested Traffic. In *Traffic and Granular Flow '99*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 365–376.
21. Shalev-Shwartz, S.; Shammah, S.; Shashua, A. On a Formal Model of Safe and Scalable Self-driving Cars. *arXiv* **2017**, arXiv:1708.06374v6.
22. Halkias, J.; Colyar, J. NGSIM—Interstate 80 Freeway Dataset. 2006. Available online: <https://www.fhwa.dot.gov/publications/research/operations/06137/06137.pdf> (accessed on 31 May 2023).
23. Colyar, J.; Halkias, J. NGSIM—US Highway 101 Dataset. 2007. Available online: <https://www.fhwa.dot.gov/publications/research/operations/07030/07030.pdf> (accessed on 31 May 2023).
24. Krajewski, R.; Bock, J.; Kloeker, L.; Eckstein, L. The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems. In Proceedings of the IEEE Conference on Intelligent Transportation Systems, ITSC, Maui, HI, USA, 4–7 November 2018; pp. 2118–2125.
25. Bock, J.; Krajewski, R.; Moers, T.; Runde, S.; Vater, L.; Eckstein, L. The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections. In Proceedings of the IEEE Intelligent Vehicles Symposium, Las Vegas, NV, USA, 19 October–13 November 2020; pp. 1929–1934.
26. Krajewski, R.; Moers, T.; Bock, J.; Vater, L.; Eckstein, L. The round Dataset: A Drone Dataset of Road User Trajectories at Roundabouts in Germany. In Proceedings of the IEEE 23rd International Conference on Intelligent Transportation Systems, ITSC, Rhodes, Greece, 20–23 September 2020.
27. Moers, T.; Vater, L.; Krajewski, R.; Bock, J.; Zlocki, A.; Eckstein, L. The exiD Dataset: A Real-World Trajectory Dataset of Highly Interactive Highway Scenarios in Germany. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 4–9 June 2022; pp. 958–964.
28. Zhan, W.; Sun, L.; Wang, D.; Shi, H.; Clause, A.; Naumann, M.; Kummerle, J.; Konigshof, H.; Stiller, C.; de La Fortelle, A.; et al. INTERACTION Dataset: An INTERNATIONAL, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps. *arXiv* **2019**, arXiv:1910.03088.
29. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A multimodal dataset for autonomous driving. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
30. Chang, M.F.; Lambert, J.; Sangkloy, P.; Singh, J.; Bak, S.; Hartnett, A.; Wang, D.; Carr, P.; Lucey, S.; Ramanan, D.; et al. Argoverse: 3D Tracking and Forecasting with Rich Maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8748–8749.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

A Proposal of Transfer Learning for Monthly Macroeconomic Time Series Forecast [†]

Martín Solís ^{1,*} and Luis-Alexander Calvo-Valverde ²¹ Business School, Tecnológico de Costa Rica, Cartago 159-7050, Costa Rica² Computer Engineering School, Tecnológico de Costa Rica, Cartago 159-7050, Costa Rica; lcalvo@tec.ac.cr

* Correspondence: marsolis@itcr.ac.cr

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Transfer learning has not been widely explored with time series. However, it could boost the application and performance of deep learning models for predicting macroeconomic time series with few observations, like monthly variables. In this study, we propose to generate a forecast of five macroeconomic variables using deep learning and transfer learning. The models were evaluated with cross-validation on a rolling basis and the metric MAPE. According to the results, deep learning models with transfer learning tend to perform better than deep learning models without transfer learning and other machine learning models. The difference between statistical models and transfer learning models tends to be small. Although, in some series, the statistical models had a slight advantage in terms of the performance metric, the results are promising for the application of transfer learning to macroeconomic time series.

Keywords: macroeconomic forecast; deep learning; transfer learning; benchmark

1. Introduction

It is necessary to generate forecasts of macroeconomic variables for national economic policy and financial decision-making [1,2]. For this reason, Central Banks, international institutions, and economic research bodies allocate time and resources to generate accurate forecasts [3,4]. The prediction of macroeconomic and financial variables is regarded as one of the most challenging applications of modern time series forecasting [5].

In recent years, deep learning models have gained relevance in the time series field because they have shown high prediction capacity in several tasks, like the forecasting of electricity consumption [6], weather conditions [7] and other tasks. Nevertheless, its implementation in macroeconomic forecasts has been scarce [8–11]. A possible reason is that deep learning performs better with large datasets and some macroeconomic variables have a low number of observations [4], especially monthly or quarterly time series. For example, 30 years of observations would only amount to 360 observations for a monthly time series. This issue is still more relevant in some emerging countries where it is difficult to find a long history of information [11].

An alternative to building deep learning models with short time series is to train a model using a handful of diverse time series. Then, the pre-trained model can be used for transfer learning with the target time series. Transfer learning has not been widely explored with time series; however, there is evidence of good results [12–14].

In this study, we applied transfer learning with monthly macroeconomic variables to analyze the performance of deep learning models for forecasting macroeconomic time series. Our main contributions are the following:

- ✓ As far as we know, this is the first study that explores and proposes the generation of pre-trained models that can be used with transfer learning to make predictions in any country using monthly macroeconomic variables.

Citation: Solís, M.; Calvo-Valverde, L.-A. A Proposal of Transfer Learning for Monthly Macroeconomic Time Series Forecast. *Eng. Proc.* **2023**, *39*, 58. <https://doi.org/10.3390/engproc2023039058>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 5 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

- ✓ Some studies compare the performance of several models with macroeconomic variables; however, as far as we know, this is the first that makes a benchmark between deep learning, statistical, and machine learning models.
- ✓ This study compares the application of deep learning models without transfer learning, as economic researchers tend to do, and the application of deep learning models with transfer learning.

Our results suggest that the latter procedure is the better practice. Therefore, the study provides findings that can be relevant to economic and financial forecast research.

2. Deep Learning Applied to Macroeconomic Variables Forecast

Some studies have used deep learning to generate forecast models for macroeconomic variables such as exchange rate [5], inflation [2], unemployment rate [15], GDP [16], interest rate [17], and exports [11]. The periodicity of the time series used is diverse, and ranges from daily to annual. When the periodicity is less frequent, such as monthly and quarterly, the time series are short; however, deep learning models without transfer learning have shown good results e.g., [4,5]. For this reason, we also used deep learning without transfer learning.

The deep learning architectures for macroeconomic predictions tend to be based on Long Short-Term Memory, which is expected, because, unlike other networks, the LSTM takes into account the sequential pattern of the time series. Some researchers have used simple LSTM [18] or LSTM encoder–decoder [9], and others have created their own architectures. For example, in [19] they created a model named DC-LSTM, which is based on a first layer using two LSTM models to learn the features. In the second layer, a coupled LSTM is built on the features learned from the first layer. Then, the learned features are fed into a fully connected layer to make the forecast. For their part, in [20] they generated an ensemble of LSTM using bagging to predict the daily exchange rate. For the prediction, they computed the median value of the k replicas. Besides LSTM, other recurrent neuronal networks have been used as a gated recurrent unit [15]. Some researchers have used non-recurrent networks such as Fully Connected Architecture, Convolutional Neuronal Networks with residual connections [21], Neural Network Autoregression (NNAR) models [10], multilayer perceptron (MLP) [15], and stacked autoencoders [8].

The models are trained using different inputs. Some models only use the lagged values of the time series as an input to the neural network e.g., [10,18], while others also use the lagged values of other time series e.g., [2,4]. A particular case is the model of [17], who incorporated Twitter sentiment related to multiple events happening around the globe into interest rate prediction.

Concerning macroeconomic prediction performance, deep learning has shown good results. For example, [5] found that LSTM outperformed VAR and SVM for predicting the monthly USD–INR foreign exchange rate. In [8] they concluded that stacked autoencoders achieve more accurate results than support vector machines for predicting the 0 daily EUR–USD exchange rate. In [2] they found that LSTM had the best performance for the inflation prediction of more than one month compared to random forests, extreme gradient boosting, and k -nearest neighbors. According to [11], the deep learning approach showed better prediction powers than conventional econometric approaches such as VECM.

3. Method

3.1. Macroeconomic Time Series

Five economic time series variables from five countries were chosen for the analysis. The variables were: Consumer Price Index (CPI), Industrial Production Index (IPI), Value of Exports in US Dollars (TVE), Average Monthly Exchange Rates of domestic Currency per US Dollar (ER), and Producer Price Index (PPI). The countries analyzed were Costa Rica (CR), the United States (UE), the United Kingdom (UK), South Korea (SR), and Bulgaria

(BU). In the case of the United States, we did not analyze domestic currency. Thus, 24 time series were used.

3.2. Datasets

Three datasets were built and used: (a) Dataset with the 24 target time series that were taken from the International Monetary Fund in May 2022; (b) Dataset with the economic variables mentioned before, from the countries shown by the FMI web page in May 2022 (<https://data.imf.org/?sk=388DFA60-1D26-4ADE-B505-A05A558D9A42&sid=1479329132316>, accessed on 22 May 2022), except for the countries used as the target. We deleted time series that did not show variability or had missing values. This dataset had 515-time series, and was named the macroeconomic dataset; (c) Dataset with 1000 time series taken randomly from the M4 competition [22]. This was named the M4 subsample. The last two datasets were built to generate the pre-trained models used for transfer learning.

3.3. Models

We analyzed five types of models:

- ✓ Three statistical models (St): Auto Arima (arima), ETS (ets), and Theta (theta);
- ✓ Machine learning models (MI): Support vector regression (svr), random forest (rf) and XGBoost (xgb);
- ✓ Deep learning models without transfer learning (DI/wh): Long short-term memory (lstm), temporal convolutional network (tcn), convolutional neuronal network (cnn);
- ✓ First proposal of deep learning with transfer Learning (DI/t_M4). We applied the same deep learning models mentioned previously, but in this case, the models were trained with the 1000 time series concatenated from the M4 subsample dataset. The concatenation was in the input and output. For example, for the prediction of the next three months based on the previous twelve months, the all-time series was transformed into a matrix of k rows and fifteen columns; then, the matrices were concatenated to obtain the final dataset. Finally, when the models were trained, we used them to apply transfer learning for each of the target time series (the 24 time series)
- ✓ The second proposal of deep learning with transfer learning (DI/t). The methodology was like that of the previous models, but the models were trained using the macroeconomic dataset.

We generated models to predict three periods and twelve periods ahead. Two types of input sizes were proved in models B, C, D, and E. For the forecast horizon of three periods, we used the previous three periods and the previous twelve periods as input, and for the forecast horizon of twelve periods, we used the previous twelve periods and the previous fifteen periods. Finally, we only kept the model with the input size that achieved the best performance. The deep learning models were trained using the Adam optimizer and a stop criterion which consisted of stopping after two epochs without improvement in the validation sample's loss function, which was the mean squared error. The hyperparameters of models B, C, D, and E and the architectures of the neuronal networks were defined in the training phase through Bayesian optimization. The search grid for the Bayesian optimization is in Table 1.

The transfer learning for models D and E was realized in the last two layers of the model. The weights were updated using the target time series and a learning rate of 0.000005, which is less than that used in the training phase. A maximum of 75 epochs were set; however, the model could stop earlier if, after two epochs, there were no improvements in the mean square error of the validation sample.

Table 1. Search grid for the Bayesian optimization of Machine Learning Models.

XGBoost: ✓ Max depth between 2 and 12 ✓ Learning rate between 0.01 and 1 ✓ Estimators between 10 and 150	Support Vector Machines: ✓ C between 0.01 and 10 ✓ Gamma between 0.001 and 0.33	Random Forest: ✓ Estimators between 10 and 250 ✓ Max_features between 1 and 15 ✓ Min sample leaf between 2 and 8 ✓ Max samples between 0.70 and 0.99
CNN: ✓ Number of convolutional layers between 1 and 2 (Batch normalization is applied after the first layer) ✓ Filters between 12 and 132 with a step of 24 ✓ Kernel size between 2 and 12 with a step of 2 ✓ Max pooling with a size of 2 or without max pooling ✓ Activation function among linear, relu, and tanh ✓ Learning rate among 0.001, 0.0001, 0.00001	TCN: ✓ Filters between 12 and 132 with a step of 24 ✓ Kernel size between 2 and 12 with a step of 2 ✓ Activation function among linear, relu, and tanh ✓ Return sequences True or False ✓ Learning rate among 0.001, 0.0001, 0.00001 ✓ Learning rate among 0.001, 0.0001, 0.00001 ✓ Dilations between [1,2,4,8] or [1,2,4,8,16]	LSTM: ✓ Recurrent units between 12 and 132 with a step of 24 ✓ Activation function among linear, relu, and tanh ✓ Return sequences True or False ✓ Learning rate among 0.001, 0.0001, 0.00001

3.4. Experimental Procedure

We used cross-validation on a rolling basis [23], also known as prequential or back testing, to evaluate the models’ performance and make comparisons between them. Figure 1 shows the general process. In the case of the machine learning and deep learning models, we split the training part into train and validation for the tuning. The performance metric MAPE was computed from the test sample in each of the ten replicas and then averaged to make the comparisons between models. In every replica, the test sample was the same as the model forecast horizon, either 3 or 12.

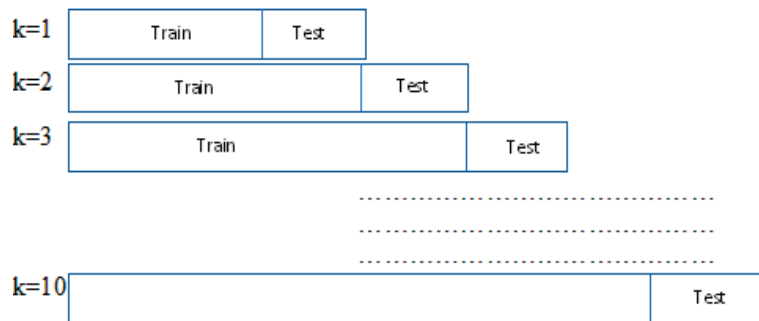


Figure 1. Cross-validation on a rolling basis for the model’s evaluation.

4. Results

Tables 2 and 3 show the average MAPE according to the model type for a horizon output of 3 and 12, respectively. The last two columns of each table present the best metric average of the transfer learning models and the best metric average for the rest of the models. The sky-blue color indicates which model has the lowest average metric, and the gray color indicates that there is no statistical difference between the type of models that the cells represent and the best model, using the Wilcoxon Sign Test with a *p*-value of 0.05.

Table 2. Average MAPE for each kind of model when output horizon = 3.

Time Serie	Type of Model					Best DI	Best Rest
	St	MI	DI/wh	DI/t	DI/t_M4		
BU_CPI	0.009	0.026	0.04	0.026	0.026	lstm_M4 = 0.01	arima = 0.007
BU_ER	0.02	0.018	0.042	0.021	0.022	lstm = 0.017	arima = 0.017
BU_IPI	0.052	0.062	0.067	0.074	0.079	tcn = 0.069	ets = 0.05
BU_PPI	0.022	0.066	0.052	0.028	0.029	lstm_M4 = 0.025	ets = 0.02
BU_TVE	0.074	0.149	0.104	0.077	0.076	tcn_M4 = 0.075	ets = 0.069
CR_CPI	0.004	0.01	0.049	0.023	0.023	lstm_M4 = 0.004	ets = 0.004
CR_ER	0.009	0.068	0.039	0.027	0.027	lstm_M4 = 0.011	theta = 0.008
CR_PPI	0.008	0.011	0.028	0.015	0.015	tcn = 0.011	ets = 0.007
CR_IPI	0.031	0.043	0.059	0.044	0.044	tcn_M4 = 0.041	ets = 0.029
CR_TVE	0.073	0.163	0.129	0.078	0.078	tcn_M4 = 0.075	arima = 0.071
KR_CPI	0.004	0.01	0.023	0.017	0.017	lstm_M4 = 0.005	theta = 0.003
KR_ER	0.02	0.023	0.035	0.02	0.02	tcn = 0.019	arima = 0.017
KR_PPI	0.007	0.013	0.023	0.011	0.011	lstm_M4 = 0.008	arima = 0.007
KR_IPI	0.036	0.051	0.063	0.045	0.049	lstm = 0.041	arima = 0.036
KR_TVE	0.055	0.083	0.089	0.066	0.066	tcn_M4 = 0.063	arima = 0.053
US_CPI	0.006	0.018	0.028	0.015	0.015	lstm_M4 = 0.007	ets = 0.006
UK_CPI	0.004	0.013	0.023	0.014	0.015	lstm = 0.004	ets = 0.003
UK_ER	0.024	0.022	0.032	0.02	0.02	cnn = 0.019	XGB = 0.018
UK_PPI	0.007	0.01	0.025	0.013	0.012	tcn = 0.007	arima = 0.007
UK_IPI	0.015	0.019	0.028	0.024	0.027	tcn = 0.018	arima = 0.013
UE_PPI	0.007	0.019	0.031	0.014	0.015	lstm_M4 = 0.009	ets = 0.006
UE_IPI	0.039	0.036	0.041	0.038	0.038	cnn = 0.036	lstm_wot = 0.033
UK_TVE	0.104	0.112	0.113	0.1	0.102	tcn = 0.097	RF = 0.099
UE_TVE	0.078	0.104	0.104	0.092	0.092	tcn = 0.09	theta = 0.076

Note: The name of the time series in the first column is composed of two parts separated by an underscore. The first part is the country's name, and the second is the macroeconomic variable's name. The acronyms meaning are in the macroeconomic time series. The blue color means that it is the best metric value, and the gray color appears when there is no statistically significant difference with the best metric, according to the Wilcoxon test with *p*-values of 0.05. Best DI is the best deep learning–transfer learning model, and Best Rest = Best of the rest models.

We compare the deep learning models created from the M4 subsample dataset with the models created from the macroeconomic dataset to determine if the usage of a dataset more oriented to the domain of the target time series generates the best results. However, the performance metrics between both types of models were similar. Tables 2 and 3 show that the MAPE was almost identical for most time series. Therefore, the M4 time series can be as valuable as the macroeconomic time series in generating pre-trained models for the variables analyzed in this study.

The results show that the deep learning models with transfer learning tended to perform better than deep learning models without transfer learning and other machine learning models for an output of 3 or 12. Recent studies have trained deep learning models directly on the economic monthly target time series e.g., [1,10,15]; however, our results suggest that there is a possibility that transfer learning could give the best performance.

The statistical models performed better in most time series than deep learning models with transfer learning, although the difference is not statistically significant for many time series. The differences between statistical and transfer learning models were smaller when the output was 12. For example, in 11 time series, there was a significant difference in the metric MAPE when the output was 3 (Table 2), and there was a significant difference in only four time series (Table 3) when the output was 12. Statistical models were not the best in all time series; there were time series for which the transfer learning models obtained the best results.

On the other hand, the results of the best specific models indicate that the best transfer learning model is mainly based on lstm for an output horizon of 3 and tcn for an output horizon of 12. The best of the other models was usually a statistical model for an output of 3 and 12. Although, in most time series, the statistical model had a slightly better value in terms of performance metric, the deep learning model has the advantage that it is simpler to make forecasts because only one model is required to predict the five variables in any country.

Table 3. Average MAPE for each kind of model when output horizon = 12.

Time Serie	Type of Model					Best DI	Best Rest
	St	MI	DI/wh	DI/t	DI/t_M4		
BU_CPI	0.019	0.043	0.064	0.017	0.019	tcn = 0.013	ets = 0.012
BU_ER	0.047	0.052	0.087	0.045	0.046	cnn = 0.04	arima = 0.041
BU_IPI	0.036	0.053	0.063	0.042	0.042	cnn_M4 = 0.037	arima = 0.035
BU_PPI	0.033	0.073	0.089	0.032	0.033	lstm = 0.031	ets = 0.026
BU_TVE	0.09	0.166	0.164	0.09	0.091	cnn = 0.089	theta = 0.08
CR_CPI	0.01	0.059	0.06	0.016	0.015	tcn_M4 = 0.01	arima = 0.009
CR_ER	0.025	0.12	0.059	0.027	0.027	tcn_M4 = 0.024	theta = 0.023
CR_PPI	0.019	0.047	0.06	0.022	0.02	tcn_M4 = 0.016	theta = 0.017
CR_IPI	0.027	0.116	0.064	0.033	0.034	cnn_M4 = 0.033	ets = 0.025
CR_TVE	0.074	0.118	0.131	0.065	0.066	lstm = 0.061	arima = 0.068
KR_CPI	0.007	0.039	0.039	0.009	0.008	lstm_M4 = 0.005	theta = 0.004
KR_ER	0.04	0.041	0.052	0.035	0.038	tcn = 0.031	arima = 0.033
KR_PPI	0.016	0.032	0.043	0.016	0.016	lstm_M4 = 0.015	ets = 0.013
KR_IPI	0.03	0.127	0.07	0.033	0.033	cnn = 0.032	theta = 0.025
KR_TVE	0.091	0.1	0.116	0.082	0.082	tcn = 0.079	RF = 0.079
US_CPI	0.008	0.055	0.032	0.01	0.011	tcn_M4 = 0.009	arima = 0.008
UK_CPI	0.006	0.044	0.034	0.009	0.009	tcn_M4 = 0.007	theta = 0.006
UK_ER	0.037	0.056	0.064	0.037	0.037	cnn = 0.035	arima = 0.035
UK_PPI	0.027	0.058	0.065	0.028	0.027	lstm_M4 = 0.025	ets = 0.026
UK_IPI	0.022	0.024	0.036	0.027	0.028	lstm = 0.024	XGB = 0.02
UE_PPI	0.012	0.065	0.038	0.014	0.014	lstm_M4 = 0.011	theta = 0.01
UE_IPI	0.023	0.031	0.036	0.026	0.025	tcn_M4 = 0.022	ets = 0.021
UK_TVE	0.101	0.116	0.122	0.087	0.092	tcn = 0.073	lstm_wot = 0.094
UE_TVE	0.057	0.102	0.109	0.063	0.062	cnn_M4 = 0.061	theta = 0.049

5. Conclusions

To our knowledge, this is the first study that analyses the application of transfer learning to macroeconomic monthly time series. Our conclusion is that transfer learning tends to perform better than the application of deep learning models without transfer learning and other machine learning models. The statistical models still provide better results in most cases, although their performance is mainly similar to that of transfer learning, and even in some series, the transfer learning showed better metrics values at a descriptive level. These findings are promising for applying transfer learning to macroeconomic time series which would simplify the generation of forecasts since, with a single pre-trained model, several economic variables can be predicted for different countries.

Additionally, there are different options to explore the generation of a model that improves the predictions of the statistical models. Future studies should apply transfer learning to other kinds of networks that have provided very positive results in natural language processing, such as transforms and seq2seq models. We only used the lags of the time series as input; however, it is relevant to analyze the performance of deep learning models when the lags of the own series and the information of external variables

are received as input, which can be used for the prediction of specific macroeconomic variables. It could improve the performance of deep learning models that have shown the capacity to extract patterns from wide input arrays. Additionally, it is possible to explore the creation of a model trained with different datasets or create ensembles from different pre-trained models.

Author Contributions: Conceptualization, M.S. and L.-A.C.-V.; methodology, M.S. and L.-A.C.-V.; software, M.S.; validation, M.S.; formal analysis, M.S.; investigation, M.S. and L.-A.C.-V.; resources, M.S.; data curation, M.S.; writing—original draft preparation, M.S.; writing—review and editing, M.S. and L.-A.C.-V.; visualization, M.S.; supervision, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Macroeconomic datasets available at https://github.com/martin12cr/DatasetTL_macroeconomic, accessed on 22 May 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Vargas, A.R. Pronóstico del crecimiento trimestral de costa rica mediante modelos de frecuencia mixta. *Rev. Cienc. Econ.* **2014**, *32*, 189–226. [CrossRef]
- Rodríguez-Vargas, A. Forecasting Costa Rican inflation with machine learning methods. *Lat. Am. J. Central Bank.* **2020**, *1*, 100012. [CrossRef]
- Aastveit, K.A.; McAlinn, K.; Nakajima, J.; West, M. Multivariate Bayesian Predictive Synthesis in Macroeconomic Forecasting. *J. Am. Stat. Assoc.* **2019**, *115*, 1092–1110. [CrossRef]
- Jung, J.-K.; Patnam, M.; Ter-Martirosyan, A. An Algorithmic Crystal Ball: Forecasts-Based on Machine Learning. *SSRN Electron. J.* **2018**, *2018*, 1–35. [CrossRef]
- Kaushik, M.; Giri, A.K. Forecasting Foreign Exchange Rate: A Multivariate Comparative Analysis between Traditional Econometric, Contemporary Machine Learning & Deep Learning Techniques. *arXiv* **2020**, arXiv:2002.10247.
- Mariano-Hernández, D.; Hernández-Callejo, L.; Solís, M.; Zorita-Lamadrid, A.; Duque-Perez, O.; Gonzalez-Morales, L.; Santos-García, F. A Data-Driven Forecasting Strategy to Predict Continuous Hourly Energy Demand in Smart Buildings. *Appl. Sci.* **2021**, *11*, 7886. [CrossRef]
- Hewage, P.; Behera, A.; Trovati, M.; Pereira, E.; Ghahremani, M.; Palmieri, F.; Liu, Y. Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. *Soft Comput.* **2020**, *24*, 16453–16482. [CrossRef]
- Chen, J.; Zhao, C.; Liu, K.; Liang, J.; Wu, H.; Xu, S. Exchange Rate Forecasting Based on Deep Learning and NSGA-II Models. *Comput. Intell. Neurosci.* **2021**, *2021*, 2993870. [CrossRef] [PubMed]
- Nguyen, H.T.; Nguyen, D.T. Transfer Learning for Macroeconomic Forecasting. In Proceedings of the 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), Ho Chi Minh City, Vietnam, 26–27 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 332–337. [CrossRef]
- Pratap, B.; Sengupta, S. Macroeconomic Forecasting in India: Does Machine Learning Hold the Key to Better Forecasts? In *RBI Working Paper Series*; Reserve Bank of India: Mumbai, India, 2019. [CrossRef]
- Kim, S. Macroeconomic and Financial Market Analyses and Predictions through Deep Learning. In Bank of Korea WP 2020-18. Available online: <https://ssrn.com/abstract=3684936> (accessed on 25 November 2021).
- Solís, M.; Calvo-Valverde, L.-A. Performance of Deep Learning models with transfer learning for multiple-step-ahead forecasts in monthly time series. *Intel. Artif.* **2022**, *25*, 110–125. [CrossRef]
- Otović, E.; Njirjak, M.; Jozinović, D.; Mauša, G.; Michelini, A.; Štajduhar, I. Intra-domain and cross-domain transfer learning for time series data—How transferable are the features? *Knowl.-Based Syst.* **2022**, *239*, 107976. [CrossRef]
- Poghosyan, A.; Harutyunyan, A.; Grigoryan, N.; Pang, C.; Oganesyan, G.; Ghazaryan, S.; Hovhannisyann, N. An Enterprise Time Series Forecasting System for Cloud Applications Using Transfer Learning. *Sensors* **2021**, *21*, 1590. [CrossRef] [PubMed]
- Mulaudzi, R.; Ajoodha, R. Application of Deep Learning to Forecast the South African Unemployment Rate: A Multivariate Approach. In Proceedings of the 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 16–18 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6. [CrossRef]
- Longo, L.; Riccaboni, M.; Rungi, A. A neural network ensemble approach for GDP forecasting. *J. Econ. Dyn. Control* **2022**, *134*, 104278. [CrossRef]

17. Yasir, M.; Afzal, S.; Latif, K.; Chaudhary, G.M.; Malik, N.Y.; Shahzad, F.; Song, O.-Y. An Efficient Deep Learning Based Model to Predict Interest Rate Using Twitter Sentiment. *Sustainability* **2020**, *12*, 1660. [CrossRef]
18. Dodevski, A.; Koceska, N.; Koceski, S. Forecasting exchange rate between macedonian denar and euro using deep learning. *J. Appl. Econ. Bus.* **2018**, *6*, 50–61.
19. Cao, W.; Zhu, W.; Wang, W.; Demazeau, Y.; Zhang, C. A Deep Coupled LSTM Approach for USD/CNY Exchange Rate Forecasting. *IEEE Intell. Syst.* **2020**, *35*, 43–53. [CrossRef]
20. Sun, S.; Wang, S.; Wei, Y. A new ensemble deep learning approach for exchange rates forecasting and trading. *Adv. Eng. Informatics* **2020**, *46*, 101160. [CrossRef]
21. Cook, T.R.; Hall, A.S. Macroeconomic Indicator Forecasting with Deep Neural Networks. In Proceedings of the CARMA 2018—2nd International Conference on Advanced Research Methods and Analytics, Valencia, Spain, 12–13 July 2018. [CrossRef]
22. M4 Team. M4 Competitor's Guide: Prizes and Rules. 2018. Available online: <https://www.m4.unic.ac.cy/wpcontent/uploads/2018/03/M4-CompetitorsGuide.pdf> (accessed on 25 November 2021).
23. Hu, M.Y.; Zhang, G.; Jiang, C.X.; Patuwo, B.E. A Cross-Validation Analysis of Neural Network Out-of-Sample Performance in Exchange Rate Forecasting. *Decis. Sci.* **1999**, *30*, 197–216. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

A Simulation Package in VBA to Support Finance Students for Constructing Optimal Portfolios [†]

Abdulnasser Hatemi-J ^{1,*} and Alan Mustafa ²

¹ Department of Economics and Finance, College of Business and Economics, UAE University, Al Ain P.O. Box 15551, United Arab Emirates

² Institute of Electrical and Electronics Engineers—IEEE, London W6 7DS, UK; alan.mustafa@ieee.org

* Correspondence: ahatemi@uaeu.ac.ae

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: This paper introduces a software component created in Visual Basic for Applications (VBA) that can be applied for creating an optimal portfolio using two different methods. The first method is the seminal approach of Markowitz and is based on finding budget shares via the minimization of the variance of the underlying portfolio. The second method, developed by Hatemi-J and El-Khatib, combines risk and return directly in the optimization problem and yields budget shares that lead to maximizing the risk-adjusted return of the portfolio. This approach is consistent with the expectation of rational investors since these investors consider both risk and return as the fundamental basis for the selection of the investment assets. Our package offers another advantage that is usually neglected in the literature, which is the number of assets that should be included in the portfolio. The common practice is to assume that the number of assets is given exogenously when the portfolio is constructed. However, the current software component constructs all possible combinations and thus the investor can figure out empirically which portfolio is the best one among all portfolios considered. The software is consumer-friendly via a graphical user interface. An application is also provided to demonstrate how the software can be used using real-time series data for several assets.

Keywords: VBA; time series data; portfolio diversification; optimization; risk and return

Citation: Hatemi-J, A.; Mustafa, A. A Simulation Package in VBA to Support Finance Students for Constructing Optimal Portfolios. *Eng. Proc.* **2023**, *39*, 59. <https://doi.org/10.3390/engproc2023039059>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 5 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Knowledge delivery as a method of continuing humanity's mind to the next generation has employed forms of tacit, explicit, and implicit knowledge [1,2]. One major method that these deliveries are shown through is the teaching and training environment. However, to ensure the delivery of knowledge, the learned materials must be assessed. Continuous investigations by researchers are taking place to seek different methods of compatibility regarding teaching materials being in line with the learner's learning style. Research has indicated that, by personalizing teaching materials to suit the specific needs of a learner, the learning performance shows improvement [3,4]. Furthermore, the practice on learned materials emphasizes that deep learning and its effect can stay with the learner for a longer time depending on the sessions of practice.

It is widely agreed that not all theories can be directly put into practice, such as aviation (new trainee pilots require many hours of flights before being called a pilot), medicine (medical students require long hours of practice in surgery before being allowed to perform independent surgery), mathematics (mathematicians and economists require an environment where they can apply theoretical concepts in practice before they become reality), and manufacturing (which requires a tremendous amount of resources through the planning, designing, and implementation of technologies before a tangible product gets produced), and many more industries can be testimony to the value of simulation-based learning with regard to saving resources and money. In addition, detailed instruction

on the process of solving a problem, including giving immediate feedback, can enhance learning [5–8].

Hence, in this paper, simulation software that shows how a decision can be improved before an actual event, such as a decision on portfolio diversification, was developed. The following sections describe the logic behind portfolio diversification, mathematical derivation and the rationale behind each method, and a chosen group of algorithms for the decision maker unit (DMU) in both simulation software, and, at the end, we discuss our findings and conclude our work. The software verbiage is available from the authors upon request.

The rest of the paper is as follows. Section 2 describes the two alternative methods that can be used for constructing a portfolio. This section also illustrates how the dimension of a portfolio can be determined endogenously. Section 3 presents our software component and describes how it can be used. Section 4 presents the finding of an empirical application. The last section provides conclusions. In Appendix A, a schematic representation of both modules that are created in this paper is provided (Figures A1 and A2).

2. Methodology

In this section, we describe the alternative approaches for constructing financial portfolios. Another issue that is usually neglected in the literature is the dimension of the portfolio. It is a common practice in literature to assume that the number of assets included in the portfolio is given *a priori*. However, this does not need to be the case in real markets. For many investors, the selection of assets is also an endogenous question. Our software takes this issue into account by constructing all possible combinations and providing the portfolio that is optimal even with regard to the number of assets also. This approach is described in the section’s sub-section.

2.1. Portfolio Construction

The seminal method for portfolio diversification was established by Markowitz [9], and leads to obtaining budget shares via minimizing the variance of the selected portfolio with regard to the budget restriction. Let us assume that r_i represents the rate of return for asset i , which has a normal as $r_i \sim \Phi(\bar{r}_i, \sigma_i^2)$. The variance and covariance matrix for the assets included in the portfolio (denoted by n) is expressed as $\Omega = (\sigma_{i,j})_{1 \leq i,j \leq n}$; here, σ_{ij} is the covariance measure between the returns of the two assets i and j . Let us also define w_i as the weight for asset i . Therefore, the average return of the portfolio is defined as $F(w) = \sum_{i=1}^n \bar{r}_i w_i$ and its variance as a measure of risk is $V(w) = w' \Omega w$. Hence, the optimization objective of Markowitz [9] is the following minimization problem:

$$\text{Minimize } V(w) = w' \Omega w \tag{1}$$

Bounded by the budget limitation expressed in Equation (2):

$$D(w) = \sum_{i=1}^n w_i = 1 \tag{2}$$

The solution for each w_i of this optimization problem is obtained as the following, assuming that there are four assets in the portfolio (i.e., $n = 4$):

$$w_1 = \frac{\begin{vmatrix} B_{1,2} & B_{1,3} & B_{1,4} \\ B_{2,2} & B_{2,3} & B_{2,4} \\ B_{3,2} & B_{3,3} & B_{3,4} \end{vmatrix}}{|E|} \tag{3}$$

$$w_2 = \frac{\begin{vmatrix} B_{1,1} & B_{1,3} & B_{1,4} \\ B_{2,1} & B_{2,3} & B_{2,4} \\ B_{3,1} & B_{3,3} & B_{3,4} \end{vmatrix}}{|E|} \tag{4}$$

$$w_3 = \frac{\begin{vmatrix} B_{1,1} & B_{1,2} & B_{1,4} \\ B_{2,1} & B_{2,2} & B_{2,4} \\ B_{3,1} & B_{3,2} & B_{3,4} \end{vmatrix}}{|E|} \tag{5}$$

$$w_4 = \frac{\begin{vmatrix} B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,1} & B_{3,2} & B_{3,3} \end{vmatrix}}{|E|} \tag{6}$$

where

$$E = \begin{pmatrix} B_{1,1} & B_{1,2} & B_{1,3} & B_{1,4} \\ B_{2,1} & B_{2,2} & B_{2,3} & B_{2,4} \\ B_{3,1} & B_{3,2} & B_{3,3} & B_{3,4} \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Observe that $|Y|$ signifies the determinant of the matrix Y . In addition, notice that B is an $n \times n$ matrix that has the following attributes:

$$B_{i,j} = (\sigma_{i+1,j} + \sigma_{j,i+1}) - (\sigma_{i,j} + \sigma_{j,i}), \forall 1 \leq i \leq n - 1 \text{ and } \forall 1 \leq j \leq n.$$

The Markowitz approach, which is commonly utilized by investors, constructs a portfolio that has the smallest possible risk. Nonetheless, it is broadly agreed that rational investors pay attention to both risk and return when investment decisions are made. Consequently, Hatemi-J and El-Khatib [10] proposed optimizing the portfolio diversification problem, which combines risk and return directly when the portfolio is created. Specifically, the objective function in the optimization problem is the following as per the authors:

$$\text{Maximize } \frac{F(w)}{\sqrt{V(w)}} = \frac{F(w)}{\sqrt{w' \Omega w}} \tag{7}$$

subject to

$$D(w) = \sum_{i=1}^n w_i = 1 \tag{8}$$

Via the application of Theorem 1 in the study by Hatemi-J, Hajji, and El-Khatib [11], the solutions for the optimal budget shares within this setting are provided in the following equations, when $n = 4$:

$$w_1 = \frac{\begin{vmatrix} G_{1,2} & G_{1,3} & G_{1,4} \\ G_{2,2} & G_{2,3} & G_{2,4} \\ G_{3,2} & G_{3,3} & G_{3,4} \end{vmatrix}}{|K|} \tag{9}$$

$$w_2 = \frac{\begin{vmatrix} G_{1,1} & G_{1,3} & G_{1,4} \\ G_{2,1} & G_{2,3} & G_{2,4} \\ G_{3,1} & G_{3,3} & G_{3,4} \end{vmatrix}}{|K|} \tag{10}$$

$$w_3 = \frac{\begin{vmatrix} G_{1,1} & G_{1,2} & G_{1,4} \\ G_{2,1} & G_{2,2} & G_{2,4} \\ G_{3,1} & G_{3,2} & G_{3,4} \end{vmatrix}}{|K|} \tag{11}$$

$$w_4 = \frac{\begin{vmatrix} G_{1,1} & G_{1,2} & G_{1,3} \\ G_{2,1} & G_{2,2} & G_{2,3} \\ G_{3,1} & G_{3,2} & G_{3,3} \end{vmatrix}}{|K|} \tag{12}$$

where

$$K = \begin{pmatrix} G_{1,1} & G_{1,2} & G_{1,3} & G_{1,4} \\ G_{2,1} & G_{2,2} & G_{2,3} & G_{2,4} \\ G_{3,1} & G_{3,2} & G_{3,3} & G_{3,4} \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Observe that G is an $n \times n$ matrix that has the following definition:

$$G_{i,j} = \bar{r}_i(\sigma_{i+1,j} + \sigma_{j,i+1}) - \bar{r}_{i+1}(\sigma_{i,j} + \sigma_{j,i}), \forall 1 \leq i \leq n - 1 \text{ and } \forall 1 \leq j \leq n.$$

Accordingly, this new method merges risk and return in the optimization problem, which accords well with reality. This is the case because rational investors consider both risk and return when they make any investment decision.

2.2. The Dimension of a Portfolio

Prior to finding the budget shares, the investor must choose the assets to include in the portfolio. This is a crucial matter. The way that the literature deals with this issue is to assume that the number of assets included in the portfolio is provided exogenously. Nevertheless, this is not the way that the investors approach this issue in real markets. The selection of assets for inclusion in the portfolio is better dealt with as an endogenous question according to Hatemi-J and Hajji [12]. The authors suggested a solution that is based on selecting the maximum number of assets that the investor might be interested in based on his/her subjective preferences. Subsequently, a series of portfolios containing different permutations of these assets can be created. Suppose that n is the maximum number of assets considered by the investor for potential inclusion in the portfolio. Thus, the number of combinations (denoted by P) needed to be built is the following according to Hatemi-J and Hajji [12]:

$$P = \sum_{l=0}^{n-2} C(n, n-l) = \sum_{l=0}^{n-2} \frac{n!}{(n-l)! \times l!} \tag{13}$$

Therefore, P is the total number of permutations that are accessible to the investor as portfolios for a given n set of underlying assets. By creating all these P portfolios, the investor should calculate the risk-adjusted return for each portfolio in this set. For instance, when n is 4, P is equal to 11 portfolios based on Equation (13) as is the case in our application. Via the risk-adjusted returns for these P portfolios, the investor can retrieve the best portfolio, the second-best, the third-best, etc. This approach makes it operational to obtain the portfolio amongst these 11 portfolios that produces the highest magnitude of return for each unit of risk; that is, the best portfolio (BP) among the P combinations is acquired as

$$BP = \text{Max}[RAR_k, \dots, RAR_P] \tag{14}$$

where

$$RAR_k = \frac{E[R_{pk}(w)]}{\sqrt{V[R_{pk}(w)]}} \tag{15}$$

The denotation $E[R_{pk}(w)]$ represents the expected return of portfolio k (for $k = 1, \dots, P$) for the given vector for optimal budget shares (i.e., w). $V[R_{pk}(w)]$ represents the variance of the same portfolio and RAR_k denotes the risk-adjusted return of portfolio k . The needed optimal budget shares for each portfolio might be acquired via minimizing the variance of the portfolio as established by Markowitz [9]. Nonetheless, it is also feasible to find the optimal budget shares via the method introduced by Hatemi-J and El-Khatib [10] and generalized by Hatemi-J, Hajji, and El-Khatib [11]. These methods are described by Equations (1)–(12) above.

3. Experimental Design (Designing a New Tool)

In this section, the design of a new tool to present the process of simplifying and solving a complicated process that usually takes a long time using either pen and paper, calculators, or a simple spreadsheet for manually performing the calculations, is recommended. The authors used the power of Microsoft Visual Basic for Applications (VBA) in Microsoft Excel to create this module to automate lengthy processes of creating multiple portfolios and their comparison to select the best possible choice of portfolio from a list of diversified portfolios, where the efficiency of the work is, of course, incomparable, since the use of VBA as a tool to automate complex calculations in the industry has become the norm. Kalwar et. al. [13] gave a comprehensive list of VBA applications in the industry that clearly backs this statement. Blayney et al. [14] also presented the capabilities and use of VBA in conjunction with MS Excel to conduct preliminary analysis in big data research.

As an example, one of the commonly used methods in finance is portfolio diversification. A personal investor or financial organization's task is to carry out investment on a series of instruments. These assets can be chosen from any of commodities, indices, forex, metal, energy, and stocks, to name a few. The issue here is determining what the best combination of those assets for the investment would be based on their historical market price and by minimizing the risk involved in trading those markets.

Markowitz [9], in his paper, recommended a solution for finding the optimal selection of the best combination of assets for investment. The approach was mainly based on the weights as budget shares that minimized the variance of the underlying portfolio. In this approach, however, the risk on the amount of return was not considered. Hatemi-J and El-Khatib [10] devised a new method based on the effects of valuing risks on the selection of assets so as to maximize the return. The method is named 'maximizing the risk-adjusted return of the portfolio'. It combines risk and returns when the optimal budget shares are searched for. Hence, two applications are presented in this section. The basis of the design for the first application is a set of a predetermined list of assets (denoted by PD-RAR, which stands for 'portfolio diversification with risk-adjusted return'), and the second one presents a comparison between portfolios based on a different number of assets (denoted PD-RAR-Comb). This last design is aimed at helping the investor to endogenize the number of assets in the portfolio by considering all possible combinations. Equations (1)–(14) are used for this purpose.

4. Development of the Tool

Since there are complex calculations involved in calculating portfolios with the best performance in both methods of Markowitz [9] and Hatemi-J and El-Khatib [10], there is the need to develop a module that performs all the required calculations efficiently via graphical user interfaces (GUIs). The aim of this research is to fill this gap in the existing literature. Schematics of these two designs as presented in the previous section are given in Appendix A.

The portfolio diversification with risk-adjusted return (PD-RAR) creates a portfolio for a set of data inputs via its dashboard panel (Figure 1). There are two methods available for entering data: (i) either through copy and paste functions to paste the set of data on the sheet named "Data", or (ii) the option named "Data as Parameters". The use of the first option is straightforward, and the application is ready to process data. The second option provides an extra option for entering the input data in the form of the available number of assets, calculated expected values, and covariance of the set of data, which could all have come from another application software (Figure 2).

Steps to Calculate the Estimation of Budget Share

Step 1: Select either of 'Load Sample Data', 'Just an Empty Sheet for My Data', or 'Data as Parameters' button to start the process.

Step 2: If the second option is selected, paste values of variables into the "Data" sheet starting from the cell A1. Make sure that each variable has its title.

Step 3: Click on the button "Construct the Portfolios" to start the calculations. Detailed calculation process is shown in the "Portfolio" sheet and summary of results is presented in a table accessible in the "Summary of Estimation" sheet.

Step 4: Records in the "Summary of Estimation" sheet can be copied or saved and moved to another resource (spreadsheet, as '*.csv' file, etc).

Step 5: As a final step you could either start over by clicking the "Start Over - Clear this workbook" button or leave the application by using "Exit" button.

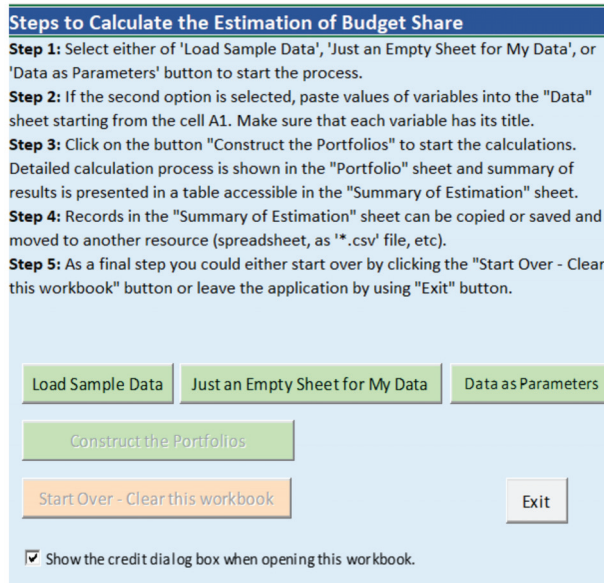


Figure 1. The main dashboard panel.

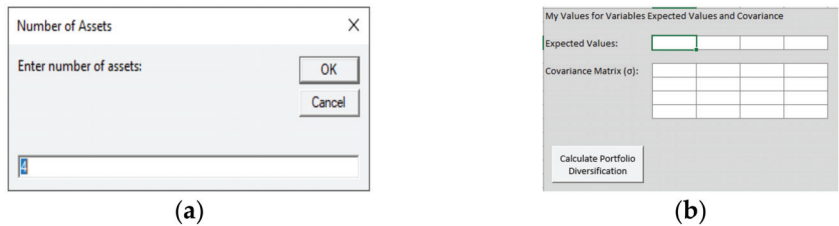


Figure 2. By following the option of “Data as Parameters”, two dialog boxes are presented; (a) entering number of assets, (b) a dialog box ready to enter values for average returns, and the variance-covariance.

After processing data based on Equations (1)–(12), the number of portfolios is given (Figure 3). This is followed by creating two types of output: (i) detailed calculations of applied algorithms (Figure 4), and (ii) the “Estimated Results” (Tables 1 and 2).

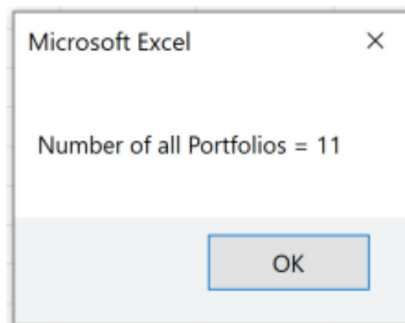


Figure 3. The notification dialog box shows the possible number of portfolios created.

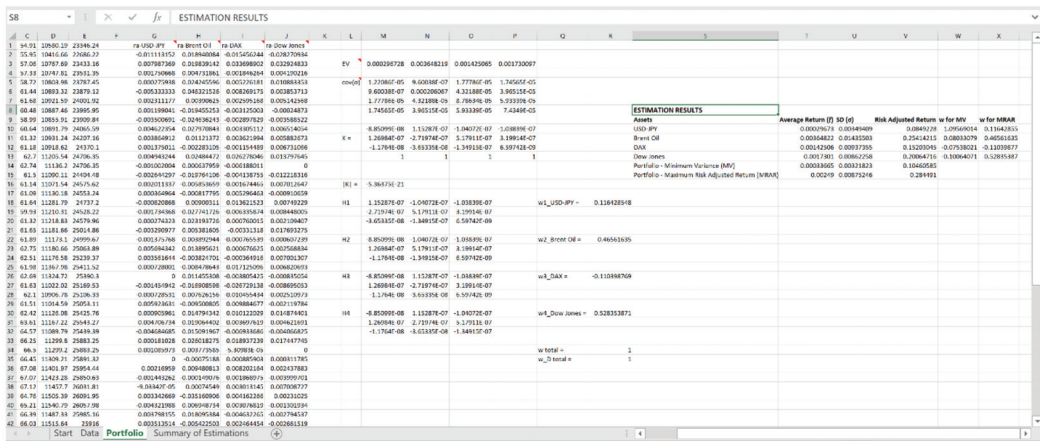


Figure 4. Sample of detailed calculations based on Equations (1)–(12).

Table 1. Summary of estimation results table for portfolio number 10 with best value based on minimum variance approach (MV), which includes two assets.

Assets	Average Return (\bar{r})	SD (σ)	Risk Adjusted Return	w for MV	w for MRAR
Brent Oil	0.00364822	0.01435503	0.25414215	0.17252711	0.4888804
Dow Jones	0.0017301	0.00862258	0.20064716	0.82747289	0.5111196
Portfolio-Minimum Variance (MV)	0.00206103	0.00826817	0.24927208		
Portfolio-Maximum Risk Adjusted Return (MRAR)	0.00266783	0.00940691	0.28360313		

Table 2. Summary of estimation results table for portfolio number 1 with the best value based on maximum risk-adjusted return (MRAR), which includes all assets.

Assets	Average Return (\bar{r})	SD (σ)	Risk Adjusted Return	w for MV	w for MRAR
USD-JPY	0.00029673	0.00349409	0.0849228	1.09569014	0.11642855
Brent Oil	0.00364822	0.01435503	0.25414215	0.08033079	0.46561635
DAX	0.00142506	0.00937355	0.15203045	-0.07538021	-0.11039877
Dow Jones	0.0017301	0.00862258	0.20064716	-0.10064071	0.52835387
Portfolio-Minimum Variance (MV)	0.00033665	0.00321823	0.10460585		
Portfolio-Maximum Risk Adjusted Return (MRAR)	0.00249	0.00875246	0.284491		

The calculation results are provided in Table 3 for the best portfolio that is created based on the maximum risk-adjusted return.

Table 3. List of portfolios with the minimum variance (MV) and maximum risk-adjusted return (MRAR) with the highest selection.

Portfolio	MV	MRAR	Portfolio with the Highest RAR ¹
Portfolio-1	0.10460585	0.284491	MV MRAR
Portfolio-2	0.12222364	0.26876603	0.24927208 0.284491
Portfolio-3	0.10252419	0.28364519	Portfolio 10 Portfolio 1
Portfolio-4	0.0426772	0.20488306	
Portfolio-5	0.23869796	0.28430838	
Portfolio-6	0.13814161	0.26645813	

Table 3. *Cont.*

Portfolio	MV	MRAR	Portfolio with the Highest RAR ¹
Portfolio-7	0.0582005	0.15205724	
Portfolio-8	0.04421402	0.20429712	
Portfolio-9	0.21516302	0.26479205	
Portfolio-10	0.24927208	0.28360313	
Portfolio-11	0.19537426	0.20076821	

¹ Portfolio construction methods: MV = minimum variance approach; MRAR = maximum risk-adjusted return approach.

PD-RAR-Comb

In this sub-section, the results for all eleven combinations are briefly presented in Table 3. These results are obtained by using the PD-RAR-Comb module. This module provides a list of all possible combinations of portfolios to be created for all assets with the addition of presenting a comparison between both algorithms used in creating those portfolios [9,10] [Appendix A part A.2]. The first phase of this process is to create a list of possible combinations of assets, and then to create a portfolio for each combination and list them in a sheet arranged in descending order from the maximum number of assets to the minimum number of assets in combination. The next step is to find the maximum value for both used algorithms of minimum variance (MV) and maximum risk-adjusted return (MRAR). The outcome of this process is shown in Table 3.

5. Findings

In this section, the findings for executing both modules are discussed. The performance of calculations mainly depends on the type of data processing (i.e., the option of “with or without detailed presentation of step-by-step calculations”) and on the size of the dataset (i.e., the number of assets and recorded closing prices for each asset). For example, by running a set of 10 assets with 65 records (which results in 1013 different portfolios), it takes around 100 s to process the data using the option “without details”. Comparatively, it takes more than 15 min to implement the same calculations with the option “with details”. The reason for this is due to the interaction with an individual worksheet (i.e., reading and writing data from and into a worksheet). It should be mentioned that the main purpose of using the module with detailed steps is for educational purposes, which gives the outcome of step-by-step calculations in the process of creating portfolios.

Note that portfolio construction based on the method by Hatemi-J and El-Khatib [4] clearly shows a better outcome if the goal is finding a portfolio that provides the highest possible return per unit of risk. However, the portfolio that is constructed by Markowitz’s [9] method results in the lowest possible risk. By using this module, it is also possible to directly enter the parameters that are necessary inputs for portfolio diversification (such as the average returns and the variance–covariance values) instead of importing the time-series data of the prices (see Figure 2).

6. Conclusions

Constructing an optimal portfolio is an important issue for investors and financial institutions. There are several methods available in the literature for this purpose. The seminal approach provided by Markowitz [9] yields an optimal portfolio that results in the minimum possible risk. This approach is widely applied by practitioners. An alternative approach that was developed by Hatemi-J and El-Khatib [10] produces a conditional optimal portfolio that gives the maximum return per unit of risk. The aim of this work was to provide a VBA module that can construct portfolios using both methods.

A pertinent issue within this context that is usually neglected in the literature is the dimension of the portfolio; that is, the number of assets included in the portfolio is assumed to be exogenous. However, it is rational to deal with this issue endogenously. The approach

that was suggested by Hatemi-J and Hajji [12] for this purpose is to estimate all possible combinations of portfolios and estimate the risk-adjusted return for each. The portfolio that gives the highest risk-adjusted return among all possible ones is the one that should be selected. Our module also provides this possibility. It constructs all possible portfolios that the investor might be interested in and indicates the optimal one using both portfolio diversification methods. An example of four assets was provided to demonstrate how the module operates. However, the results can be generalized in future applications. The module is very consumer-friendly. The software verbiage of the module is accessible from the authors upon demand.

Author Contributions: Conceptualization, A.H.-J.; methodology, A.H.-J.; software, A.M. and A.H.-J.; validation, A.H.-J. and A.M.; formal analysis, A.M.; investigation, A.H.-J. and A.M.; writing—review and editing, A.H.-J. and A.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Dataflow and schematics for system design of both recommended modules of PD-RAR and PR-RAR-Comb

A.1: Figure A1 presents a data processing mechanism for calculating the PD-RAR. Detailed mathematical algorithm for this process is given in Equations (1)–(12) in Section 2.1, and a screenshot of its output is given in Tables 1 and 2.

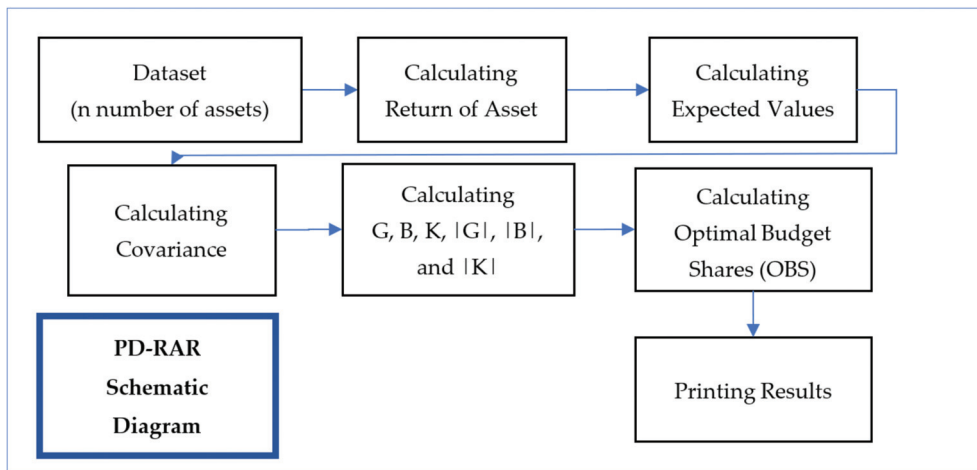


Figure A1. Schematic diagram of portfolio diversification for PD-RAR model.

A.2: Figure A2 gives a list of combinations of portfolios developed based on (i) maximum values of minimum variance (MV) and (ii) maximum risk-adjusted return approaches (MRAR). The process uses the same mechanism of calculating portfolios with the addition of looping through portfolios. The outcome is given in Table 3.

The calculations are based on Equations (1)–(15) of Sections 2.1 and 2.2 for the implementation of the module.

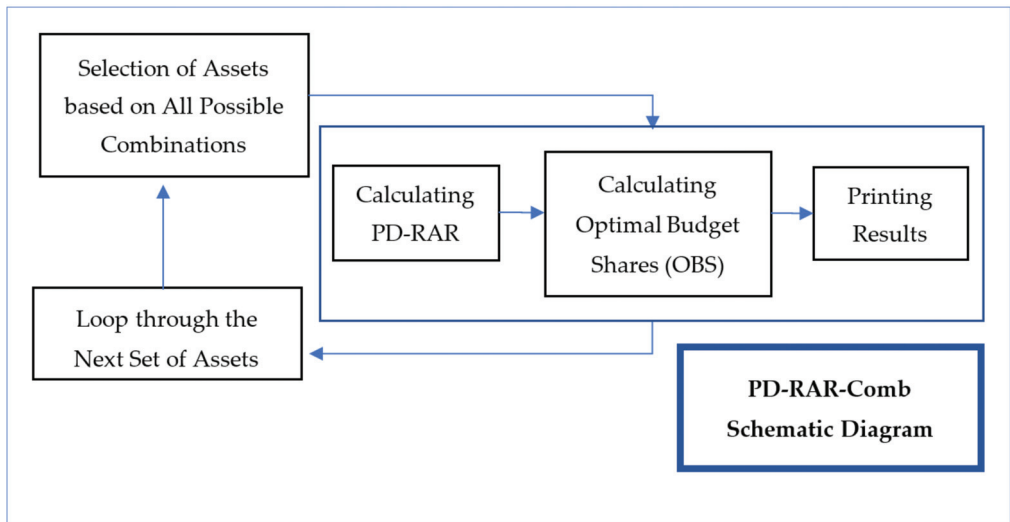


Figure A2. Schematic diagram of portfolio diversification for PD-RAR-comb model.

References

1. Mustafa, A. Impact of Learner Control on Learning in Adaptable and Personalised e-Learning Environments. Ph.D. Thesis, University of Greenwich, London, UK, 2011. Available online: <https://gala.gre.ac.uk/id/eprint/7143/> (accessed on 1 June 2023).
2. Mustafa, A. The personalization of e-learning systems with the contrast of strategic knowledge and learner's learning preferences: An investigatory analysis. *Appl. Comput. Inform.* **2021**, *17*, 153–167. Available online: <https://www.emerald.com/insight/content/doi/10.1016/j.aci.2018.08.001/full/html> (accessed on 1 June 2023). [CrossRef]
3. Mustafa, A. Effects of types of assessment questions on learning performance of two types of e-learning systems, adaptable and personalised ELs. In Proceedings of the International Engineering, Science and Education Conference, INESEC, Diyarbakir, Turkey, 1–3 December 2016. Available online: <http://www.inesegconferences.org/wp-content/uploads/2019/11/Education-Proceeding-Book-V1.pdf> (accessed on 20 April 2020).
4. Bartolomé, A.; Castañeda, L.; Adell, J. Personalisation in educational technology: The absence of underlying pedagogies. *Int. J. Educ. Technol. High. Educ.* **2018**, *15*, 14. [CrossRef]
5. Lateef, F. Simulation-based learning: Just like the real thing. *J. Emergencies Trauma Shock.* **2010**, *3*, 348–352. Available online: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2966567/> (accessed on 1 June 2023). [CrossRef]
6. ElearningInfographics [Weblog Post]. Available online: <https://elearninginfographics.com/4-examples-of-simulation-based-learning/> (accessed on 20 April 2020).
7. EtCourse [Weblog Post]. How Simulation Tools Are Transforming Education and Training. Available online: <https://www.etcourse.com/simulation-tools-transform-education-and-training.html> (accessed on 20 April 2020).
8. Dinsmore, A. Association for Talent Development. Available online: <https://www.td.org/insights/research-backs-benefits-of-vr-training> (accessed on 20 April 2020).
9. Markowitz, H. Portfolio selection. *J. Financ.* **1952**, *7*, 77–91.
10. Hatemi-J, A.; El-Khatib, Y. Portfolio selection: An alternative approach. *Econ. Lett.* **2015**, *135*, 141–143. [CrossRef]
11. Hatemi-J, A.; Hajji, M.A.; El-Khatib, Y. Exact Solution for the Portfolio Diversification Problem Based on Maximizing the Risk-Adjusted Return. *Res. Int. Bus. Financ.* **2022**, *59*, 101548. [CrossRef]
12. Hatemi-J, A.; Hajji, M.A. The Dimension of a Portfolio. 2022; unpublished work.
13. Kalwar, M.A.; Shaikh, A.S.; Khan, M.A. Optimization of Vendor Rate Analysis Report by Visual Basic for Applications (VBA): A Case Study of Footwear Industry. In Proceedings of the International Conference on Industrial & Mechanical Engineering and Operations Management, Dhaka, Bangladesh, 26–27 December 2020. Available online: <http://www.ieomsociety.org/imeom/228.pdf> (accessed on 1 June 2023).
14. Blayney, P.J.; Sun, Z. Using Excel and Excel VBA for Preliminary Analysis in Big Data Research. In *Managerial Perspectives on Intelligent Big Data Analytics*; IGI Global: Hershey, PA, USA, 2019; pp. 110–136. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Multivariable NARX Based Neural Networks Models for Short-Term Water Level Forecasting [†]

Jackson B. Renteria-Mena ¹, Douglas Plaza ² and Eduardo Giraldo ^{3,*}

¹ Facultad de Ingeniería, Universidad Tecnológica Del Chocó, Quibdó 270001, Colombia; jackson.renteria@utp.edu.co

² Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral, Guayaquil 090902, Ecuador; douplaza@espol.edu.ec

³ Research Group in Automatic Control, Electrical Engineering Department, Universidad Tecnológica de Pereira, Pereira 660003, Colombia

* Correspondence: egiraldo@utp.edu.co

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: In this work a novel application for multivariable forecasting is presented, applied to hydrological variables and based on a multivariable NARX model. The proposed approach is designed for two hydrological stations located at the Atrato River in Colombia where the variables of water level, water flow and water precipitation are correlated by using the NARX model based on a neural network structure. The structure of the NARX-based neural network is designed in order to consider the complex dynamics of hydrological variables and their corresponding cross-correlations. A short-term water level forecasting is designed based on the NARX model, to be used as an early warning flood system. The validation of the proposed approach is performed by comparing the estimation error with an ARX dynamic model. As a result, it is shown that a NARX model structure is more suitable for water level forecasting than simplified structures.

Keywords: forecast; water level; neural network

1. Introduction

The accurate modeling of hydrological variables is crucial for effective flood forecasting and the design and operation of water resource systems, as stated in the reference [1]. To achieve this, two types of techniques are commonly used: white-box algorithms, which rely on mathematical modeling, and black-box algorithms, which employ non-linear neural network techniques based on artificial intelligence. In the context of early warning systems, the latter technique has been used to great effect, as highlighted in [2], which discussed the use of artificial neural networks (ANNs) for prediction and forecasting. Furthermore, the combination of ANNs with the Soil and Water Assessment Tool (SWAT) has been applied for runoff prediction and water management resources, as described in [3].

Hydrological models have been developed to support flood early warning systems through the use of estimation and prediction algorithms. In [4], the authors focused on developing a flood forecasting system (FFS) capable of providing early warning to UDS managers of potential flooding, using a nonlinear autoregressive neural network with exogenous inputs (NARX) to predict the impact of a storm. Meanwhile, in [5], a neural network short-term memory model (LSTM) was proposed for flood forecasting, using daily flow and rainfall as input data, and analyzing features that may affect the model's performance. In [6], the authors presented flood early warning systems that used machine learning (ML) techniques, comparing the performance of five ML classification techniques for short-duration flood forecasting. Lastly, Bande and Shete [7] described an IoT-based flood monitoring and flood prediction system based on artificial neural networks (ANN),

Citation: Renteria-Mena, J.B.; Plaza, D.; Giraldo, E. Multivariable NARX Based Neural Networks Models for Short-Term Water Level Forecasting. *Eng. Proc.* **2023**, *39*, 60. <https://doi.org/10.3390/engproc2023039060>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 5 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

which aimed to monitor humidity, temperature, pressure, rainfall, and water level of rivers and analyze their temporal correlation for flood prediction. The system was designed to improve the scalability and reliability of the flood management system.

Several studies related to flood forecasting and monitoring using various ANN techniques have also been developed. For example, in [8], the authors evaluated the bias correction of real-time precipitation data and the improvement of hydrological models using the ANN bias correction method for real-time flood forecasting. In [9], the authors focused on developing five different ANN models for flood forecasting and compared their performance. In [10], the authors use a multilayer perceptron to design a flood prediction model with flow as input-output variables, and the proposed model's effectiveness was demonstrated through intensive experiments. In [11], the authors designed a flood monitoring system that integrated flow and water level sensors and used a two-class neural network to predict flood status from data stored in the database. Finally, in [12], the authors employed a convolutional neural network (CNN) to predict time series variables such as water level in a flood model, despite CNNs typically being used for two-dimensional image classification with transfer learning.

In addition, in [13], a flood forecasting model that predicted future flood occurrence was designed and evaluated by constructing a hybrid deep learning algorithm called ConvLSTM, which integrated the predictive merits of CNN and a long-term memory network (LSTM). In [14], a fuzzy neural network that used fuzzy numbers to account for uncertainty in the results and model parameters was proposed to predict the peak flow in an urban river. In [15], the potential of the AI computational paradigm for modeling streamflow was explored by developing nine different flood prediction models using all available training algorithms of ANN, fuzzy logic, and adaptive neuro-fuzzy inference systems (ANFIS) algorithms. Lastly, in [16], a deep neural network was used to predict floods as a function of temperature and rainfall intensity, and its accuracy and error were compared with other machine learning models, such as the support vector machine (SVM), K-nearest neighbor (KNN), and Naïve Bayes.

The use of real-time methods based on ANN was also proposed based on neural networks. For example, in [17], a system for predicting flood levels was developed based on real-time sensor data. The system used a multi-layer artificial neural network model created with MATLAB to predict flood levels in advance using data collected from sensors in a real-time monitoring system. In [18], a hybrid river flood forecasting model was presented using time series analysis and artificial neural networks to explain and forecast daily water discharge of the Mohawk River in New York. Multiple linear regression (MLR) models and an ANN model were used to describe each component for predicting the water discharge time series. In [19], five alternative machine learning techniques were used to improve the hydrological model, including linear regression, neural network regression, Bayesian linear regression, and reinforced decision tree regression, with the MIKE-11 hydrologic forecasting model used as a test system. In [20], a machine learning method was presented that uses historical typhoon paths to predict flood hydrographs of a Taiwanese watershed. Finally, in [21], a general framework for probabilistic flood forecasting was introduced, which uses an unaccented Kalman filter (UKF) postprocessing technique to model point forecasts made with a recurrent neural network and their corresponding observations. The methodology was tested using a 6 h long-term time series of the Three Gorges reservoir in China.

In this work, the application of a multivariable NARX model based on neural network for short-term water level forecasting along the Atrato river in Colombia is presented and evaluated. To this end, the present work uses data from two hydrological stations located in the Atrato river, which are monitored by the Institute of Hydrology, Meteorology and Environmental Studies (IDEAM). The data includes measurements of flow, precipitation, and water levels sampled every 12 h over a period of 789 days. The multivariable NARX model is trained to predict the water levels of each station based on the inputs of water level, water flow and water precipitation by considering the inherent dynamic and correlation of

the process. The performance of the models is evaluated based on the mean square error of the estimated outputs compared to the actual data. The performance of the proposed approach is compared to a multivariable ARX system. The main contribution of this paper is a general design of a multivariable NARX model structure based on neural networks for short-term water level forecasting. This work is organized as follows; in Section 2 the theoretical framework where the hydrological variables and the NARX model for their corresponding dynamic approximation is proposed. In Section 3 the experimental setup and the estimation results are shown, and finally in Section 4 the conclusions and final remarks are presented.

2. Theoretical Framework

2.1. Hydrological Variables

In order to perform water level forecasting based on the dynamic of a river, two hydrological stations are located on a river in two different positions. In order to consider the correlation among all the variables of the system and their corresponding nonlinearities, a nonlinear dynamical model is proposed. In (1) the inputs and outputs of the proposed model are shown.

$$y[k] = \begin{bmatrix} y_{L_1}[k] \\ y_{L_2}[k] \end{bmatrix}, u[k] = \begin{bmatrix} u_{F_1}[k] \\ u_{PT_1}[k] \\ u_{F_2}[k] \\ u_{PT_2}[k] \end{bmatrix} \quad (1)$$

where $y_{L_1}[k]$, $y_{L_2}[k]$ correspond to the two outputs of the level variable of the two stations, $u_{F_1}[k]$, $u_{F_2}[k]$, $u_{PT_1}[k]$, $u_{PT_2}[k]$ are the four inputs of the neural network system corresponding to the two stations of the multivariable system, i.e., the $u_{F_j}[k]$ represents the j -th two inputs of the flow variable of the two stations and $u_{PT_j}[k]$ represents j -th; two more inputs of the rainfall variable of the two stations already mentioned, thus obtaining a multivariable system with four inputs and two outputs.

The dynamic of the hydrological variables is defined by considering a Nonlinear function with an Auto-Regressive and exogenous inputs (NARX) as follows:

$$y[k] = f(y[k-1], \dots, y[k-n], u[k-1], \dots, u[k-n]) + \eta[k] \quad (2)$$

being n the order of the NARX model and $f(\cdot)$ the nonlinear function, and $\eta[k]$ the additive noise at time instant k .

2.2. Narx Based Neural Network Structure

In order to consider the NARX model of (2), the inputs are selected as $u[k-j]$ and $y[k-j]$, with $j = 1, \dots, n$, which correspond to a n -th order model. In this work a 4-th order model ($n = 4$) is considered according to [22] where an analysis of the order selection is performed and the lowest estimations error is obtained for the 3-rd order model or higher. Therefore, by considering the variables described in (1), the proposed NARX model consists of 24 inputs and 2 outputs.

In order to approximate the nonlinear function of (2), the nonlinear function $f(\cdot)$ is approximated by using a neural network structure $f^*(\cdot)$, as depicted in Figure 1.

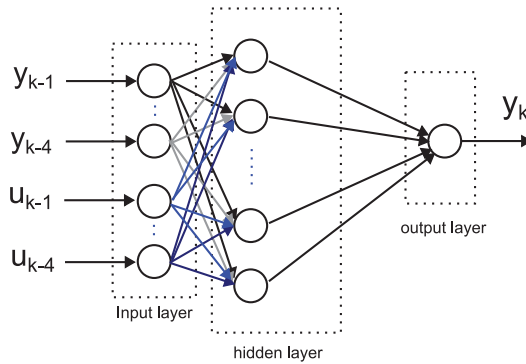


Figure 1. NARX based Neural Networks Structure.

where the NARX model can be defined as follows:

$$y[k] = f^*(y[k-1], \dots, y[k-n], u[k-1], \dots, u[k-n]) + \eta[k] \quad (3)$$

To this end, 24 input activation functions with one hidden layer and 2 outputs are considered. A feed-forward network is selected as a candidate for the NARX model in order to speed-up the training process. The training of the NARX model is performed offline by considering the data sample.

A linear ARX structure can be obtained by neglecting the hidden layer as depicted in Figure 2.

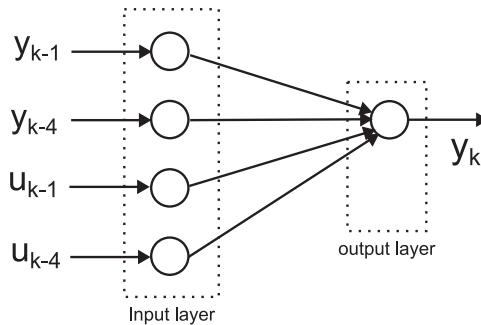


Figure 2. ARX Structure.

where the ARX model can be defined as follows:

$$y[k] = - \sum_{j=1}^4 a_j y[k-j] + \sum_{j=1}^4 b_j u[k-j] \quad (4)$$

where $A_j \in \mathbb{R}^{m \times m}$ and $B_j \in \mathbb{R}^{m \times m}$ are the parameters of the model matrix, where y are the outputs and u are the inputs; with $j = 1$, being p the order of the system, $e[k]$ the noise with m , the number of outputs and inputs of the system, $y[k] \in \mathbb{R}^{m \times 1}$ and $u[k] \in \mathbb{R}^{m \times 1}$.

3. Results

3.1. Experimental Setup

In order to validate the proposed approach, real measurements from two hydrological stations located at the Atrato river are considered. The measured variables are: water level, water flow, and water precipitation. The Atrato river is located in Colombia (South America) and has a total length of 750 km with a variable width in a range of 150 m to

500 m. In addition, the depth of the river has a variability in the range of 31m to 38 m. A total amount of 789 days of data with a sample time of 12 h is considered, with initial date 1 January 2021.

In Table 1 the geographical positions of the hydrological stations are shown.

Table 1. Location of the hydrological stations.

	Station 1 (E1)	Station 2 (E2)
Longitude	76°40'10.75" W	76°39'44.13" W
Latitude	5°45'53.38" N	5°41'52.77" N
Altitude	20.579 MASL	20.83 MASL
City	Belén de Bajirá	Quibdó

In addition, it is worth noting that the distance in kilometers between the two stations E1 and E2 is 447.1 km.

In order to validate the proposed approach, a comparison analysis of the proposed NARX approach based on neural networks (3) is performed with a multivariable ARX model (4). A visual comparison of the real and estimated signals is presented for the ARX and NARX methods and also a quantitative evaluation based on the mean squared error is performed. Both systems (ARX and NARX) are trained offline by considering the measurement data. An evaluation in terms of the training error is also presented. The feed-forward network structure for the NARX approach considers one hidden layer with 256 units. The Rectified Linear Unit (ReLU) Activation Function is selected for the proposed approach, where the ReLU is a piecewise linear function that will output the input directly if it is positive, or zero otherwise.

The implementation of the proposed NARX model based on neural networks and also the ARX model is performed in Python by using Tensorflow, which is an open-source machine learning library developed by Google.

3.2. Estimation Results

In this subsection the forecasting results for the two considered methods are presented: The proposed multivariable NARX approach, and the multivariable ARX.

In Figure 3 the estimation results for the ARX method for each of the two water level outputs are presented. In Figure 3a the short-term estimation of the first water level output and also the real measurements are shown. In Figure 3b the short-term water level forecasting of the second output as well as the real corresponding measurements are depicted.

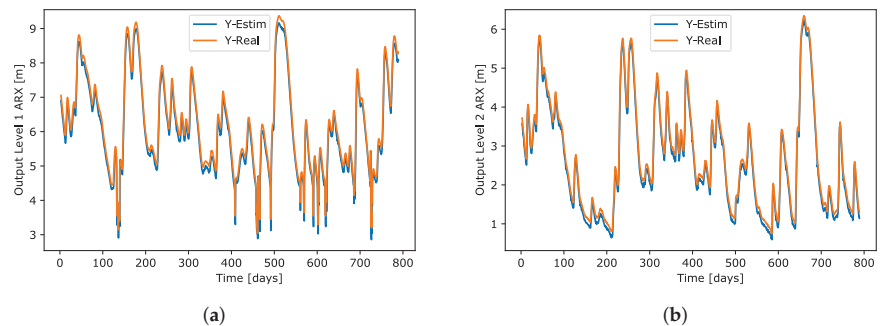


Figure 3. Multivariable ARX short-term forecasting results for two water level outputs. (a) First water level output and short-term estimation based on a multivariable ARX model. (b) Second water level output and short-term estimation based on a multivariable ARX model.

In Figure 3a,b shows that the real measurements are adequately estimated by using the multivariable ARX model.

In Table 2 an analysis for the nonlinear neural network NARX in terms of the number of nodes in the hidden layer and their corresponding mean squared estimation error is shown.

From Table 2 it can be seen that the total estimation error is reduced by increasing the number of nodes in the hidden layer. It is noticeable that there is no significant reduction in the total estimation error between 256 and 512 nodes. Therefore, in this work, 256 nodes in the hidden layer are used for evaluation of the NARX models.

In Figure 4 the estimation results for the proposed multivariable NARX method for each of the two water level outputs are presented, by using 256 nodes at the hidden layer according to the results presented in Table 2. In Figure 4a the short-term estimation of the first water level output and also the real measurements are presented. In Figure 4b the short-term water level forecasting of the second output as well as the real corresponding measurements are presented.

Table 2. Mean squared Estimation error for several nodes configurations.

NARX Hidden Layer Nodes	Level 1	Level 2	Total
2	1.6867	1.6859	3.3726
4	0.0266	0.0275	0.0541
8	0.0254	0.0272	0.0526
16	0.0227	0.0257	0.0484
32	0.0232	0.0201	0.0433
64	0.0235	0.0180	0.0415
128	0.0176	0.0133	0.0309
256	0.0085	0.0108	0.0193
512	0.0078	0.0101	0.0179

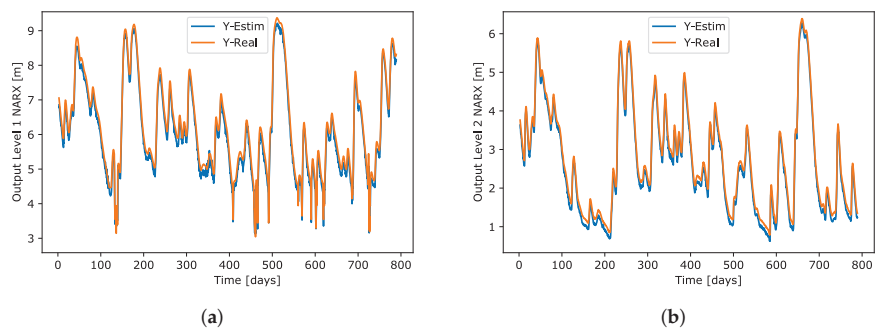


Figure 4. Multivariable NARX short-term forecasting results for two water level outputs. (a) First water level output and short-term estimation based on a multivariable NARX model. (b) Second water level output and short-term estimation based on a multivariable NARX model.

It is worth noting that in Figure 4a,b it is shown that also for the multivariable NARX model, the real measurements are adequately estimated.

By considering the forecasting results presented in Figures 3 and 4, the proposed multivariable NARX and ARX approaches show an adequate performance by visual inspection. In order to determine which approach tracks the dynamics of the hydrology variables more adequately, a quantitative evaluation is performed. To this end, the mean squared error is computed in order to compare the real measurements and their corresponding forecasting for each of the considered methods. As a result, in Table 3 the mean squared error for

the proposed multivariable NARX approach and also the multivariable ARX method are presented. It can be seen that the estimation error for the proposed NARX model is lower than the ARX model. It is worth noting that the reduction of estimation error for the NARX approach in comparison to the ARX approach is over the 50% for all variables.

In Figure 5 the estimation errors during training for each of the considered methods are shown. It can be seen that the multivariable ARX approach converge faster than the proposed multivariable NARX approach. However, the training error was lower for the NARX approach in comparison to the ARX approach. This behaviour validated the fact that there are nonlinear dynamics inherent to the measured hydrological variables and therefore the proposed NARX model forecast the data behaviour more adequately.

Table 3. Mean squared Estimation error.

Neural Network Model	Level 1	Level 2	Total
ARX	0.0280	0.0263	0.0543
NARX	0.0085	0.0108	0.0193

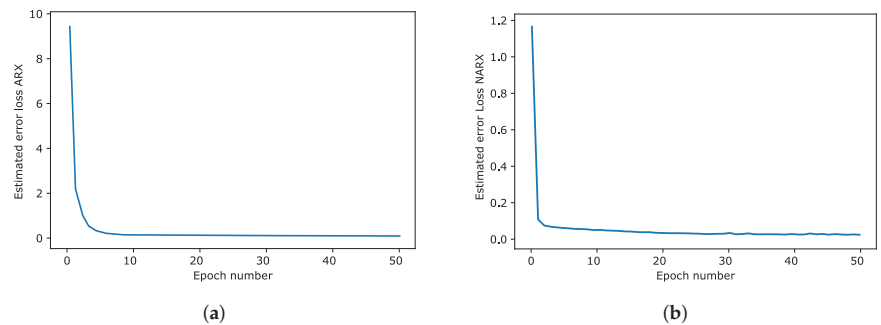


Figure 5. Training error. (a) Training error with the ARX multivariable model. (b) Training error with the NARX model.

4. Discussions and Conclusions

In this work a novel application for multivariable forecasting method for hydrological variables based on multivariable NARX model is presented. To this end, the nonlinear function of the NARX model has been approximated by using neural networks. The proposed multivariable NARX approach is compared to a multivariable ARX approach, where the proposed NARX approach shows a lower estimation error (a reduction over a 50% error as shown in Table 3). By considering the training errors, it can be seen that the training error is lower for the NARX approach in comparison to the ARX approach due to the nonlinear dynamics inherent to the measured hydrological variables. In summary, the proposed multivariable NARX model based on neural networks is an effective tool for water level forecasting by considering the correlation among several hydrological variables and several stations. It is worth noting that the main contribution of the proposed approach is the design of a general structure for modeling that can be extended to several hydrological systems with more stations and variables. In future work, online training, white-box or more complex nonlinear structures can be used to describe the nonlinear behavior of the system.

Author Contributions: All authors contributed to the conception and design of the analysis. Data acquisition was performed by J.B.R.-M. The manuscript was written by J.B.R.-M., and all authors reviewed, commented, and edited the draft version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by project no. 6-22-8 entitled “Identificación y control de sistemas multivariables interconectados a gran escala” by Universidad Tecnológica de Pereira, Pereira, Colombia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All relevant data has been made available in the paper.

Conflicts of Interest: The authors declare that there are no conflict of interest regarding the publication of this paper.

References

1. Bras, R.L.; Rodriguez-Iturbe, I. *Random Functions and Hydrology*; Courier Corporation: Chelmsford, MA, USA, 1993.
2. Palchevsky, E.; Antonov, V.; Enikeev, R.; Breikin, T. A system based on an artificial neural network of the second generation for decision support in especially significant situations. *J. Hydrol.* **2023**, *616*, 128844. [CrossRef]
3. Lv, Z.; Zuo, J.; Rodriguez, D. Predicting of Runoff Using an Optimized SWAT-ANN: A Case Study. *J. Hydrol. Reg. Stud.* **2020**, *29*, 100688. [CrossRef]
4. Abou Rjeily, Y.; Abbas, O.; Sadek, M.; Shahrou, I.; Hage Chehade, F. Flood forecasting within urban drainage systems using NARX neural network. *Water Sci. Technol.* **2017**, *76*, 2401–2412. [CrossRef] [PubMed]
5. Le, X.H.; Ho, H.V.; Lee, G.; Jung, S. Application of long short-term memory (LSTM) neural network for flood forecasting. *Water* **2019**, *11*, 1387. [CrossRef]
6. Muñoz, P.; Orellana-Alvear, J.; Bendix, J.; Feyen, J.; Céleri, R. Flood Early Warning Systems Using Machine Learning Techniques: The Case of the Tomebamba Catchment at the Southern Andes of Ecuador. *Hydrology* **2021**, *8*, 183. [CrossRef]
7. Bande, S.; Shete, V.V. Smart flood disaster prediction system using IoT & neural networks. In Proceedings of the 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), Bengaluru, India, 17–19 August 2017; pp. 189–194.
8. Jabbari, A.; Bae, D.H. Application of Artificial Neural Networks for accuracy enhancements of real-time flood forecasting in the Imjin basin. *Water* **2018**, *10*, 1626. [CrossRef]
9. Tabbussum, R.; Dar, A.Q. Comparative analysis of neural network training algorithms for the flood forecast modelling of an alluvial Himalayan river. *J. Flood Risk Manag.* **2020**, *13*, e12656. [CrossRef]
10. Dtissibe, F.Y.; Ari, A.A.A.; Titouna, C.; Thiare, O.; Gueroui, A.M. Flood forecasting based on an artificial neural network scheme. *Nat. Hazards* **2020**, *104*, 1211–1237. [CrossRef]
11. Abdullahi, S.I.; Habaebi, M.H.; Malik, N.A. Flood disaster warning system on the go. In Proceedings of the 2018 7th International Conference on Computer and Communication Engineering (ICCCCE), Kuala Lumpur, Malaysia, 19–20 September 2018; pp. 258–263.
12. Kimura, N.; Yoshinaga, I.; Sekijima, K.; Azechi, I.; Baba, D. Convolutional neural network coupled with a transfer-learning approach for time-series flood predictions. *Water* **2019**, *12*, 96. [CrossRef]
13. Moishin, M.; Deo, R.C.; Prasad, R.; Raj, N.; Abdulla, S. Designing deep-based learning flood forecast model with ConvLSTM hybrid algorithm. *IEEE Access* **2021**, *9*, 50982–50993. [CrossRef]
14. Khan, U.T.; He, J.; Valeo, C. River flood prediction using fuzzy neural networks: An investigation on automated network architecture. *Water Sci. Technol.* **2018**, *2017*, 238–247. [CrossRef] [PubMed]
15. Tabbussum, R.; Dar, A.Q. Performance evaluation of artificial intelligence paradigms—Artificial neural networks, fuzzy logic, and adaptive neuro-fuzzy inference system for flood prediction. *Environ. Sci. Pollut. Res.* **2021**, *28*, 25265–25282. [CrossRef] [PubMed]
16. Sankaranarayanan, S.; Prabhakar, M.; Satish, S.; Jain, P.; Ramprasad, A.; Krishnan, A. Flood prediction based on weather parameters using deep learning. *J. Water Clim. Chang.* **2020**, *11*, 1766–1783. [CrossRef]
17. Cruz, F.R.G.; Binag, M.G.; Ga, M.R.G.; Uy, F.A.A. Flood prediction using multi-layer artificial neural network in monitoring system with rain gauge, water level, soil moisture sensors. In Proceedings of the TENCON 2018–2018 IEEE Region 10 Conference, Jeju, Republic of Korea, 28–31 October 2018; pp. 2499–2503.
18. Tsakiri, K.; Marsellos, A.; Kapetanakis, S. Artificial neural network and multiple linear regression for flood prediction in Mohawk River, New York. *Water* **2018**, *10*, 1158. [CrossRef]
19. Noymanee, J.; Theeramunkong, T. Flood forecasting with machine learning technique on hydrological modeling. *Procedia Comput. Sci.* **2019**, *156*, 377–386. [CrossRef]
20. Chang, L.C.; Chang, F.J.; Yang, S.N.; Tsai, F.H.; Chang, T.H.; Herricks, E.E. Self-organizing maps of typhoon tracks allow for flood forecasts up to two days in advance. *Nat. Commun.* **2020**, *11*, 1983. [CrossRef] [PubMed]

21. Zhou, Y.; Guo, S.; Xu, C.Y.; Chang, F.J.; Yin, J. Improving the reliability of probabilistic multi-step-ahead flood forecasting by fusing unscented Kalman filter with recurrent neural network. *Water* **2020**, *12*, 578. [CrossRef]
22. Renteria-Mena, J.B.; Giraldo, E. Multivariable AR Data Assimilation for Water Level, Flow, and Precipitation Data. *IAENG Int. J. Comput. Sci.* **2023**, *50*, 263–273.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Enhancement of Consumption Forecasting by Customers' Behavioral Predictability Segregation [†]

Maria Koshkareva ^{1,*},[‡] and Anton Kovantsev ^{2,*},[‡]¹ Department of Digital Transformation, ITMO University, St. Petersburg 197101, Russia² National Center for Cognitive Research, ITMO University, St. Petersburg 197101, Russia

* Correspondence: koshk.mp@gmail.com (M.K.); ankovantsev@itmo.ru (A.K.)

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

‡ These authors contributed equally to this work.

Abstract: The easiest approach to customer activity forecasting involves using the whole available and applicable population of customers that a certain data set contains. The drawback of this simple technique is twofold: the set could be too big, and it could contain customers of very different peculiarities, which means that customers whose previous behavior is helpful for the forecast and whose one is not are mixed, and while the first performs a good-quality prediction, the second spoils it by adding noise. Hence, if we could choose the customers with good predictability and put aside the others “as a shepherd divideth his sheep from the goats” (Matthew 25:32), we would solve both problems: less data volume and less noise; the principle is like ancient “*divide et impera*”. In our research, we developed the method of customers separation by predictability and its dynamics with the help of LSTM models. Our research shows that (1) customer separation helps to improve the forecasting quality of the whole population due to the decomposition of all clients' time series, and (2) environmental instability such as pandemics or military action can be leveled out with incremental models.

Keywords: predictability; consumer behavior; incremental learning

Citation: Koshkareva, M.; Kovantsev, A. Enhancement of Consumption Forecasting by Customers' Behavioral Predictability Segregation. *Eng. Proc.* **2023**, *39*, 61. <https://doi.org/10.3390/engproc2023039061>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 5 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The finance sector has long adopted machine learning techniques in their client behavior analysis (Source code can be found in ref. [1]) to evaluate credit scoring [2], predict customer churn [3], detect fraudulent transactions [4], recommend personalized entities [5], and predict next purchases or trips abroad [6]. Apart from making predictions based on customer behavior, it is wise to detect clients who would be more likely to make such predictions come true, in other words, agents with high predictability. This wisdom is explained by fewer risks when trying out new features and better feedback after targeted recommendations.

Additionally, it may be useful to discard clients with low predictability from available population in quest of higher prediction accuracy. Following the assumption that unpredictable agents add nothing except noise, making predictions for all clients and considering only the ones with high predictability may result in more accurate forecasts. Taking into account that well predicted agents may as well be the first to react to changes and/or have fewer fluctuations in their time series, there is useful information to be extracted from their behavior.

Once highly predictable, an actor may lose this status if their transactional behavior changes due to military-political events, pandemic-related restrictions, or even less influential circumstances. To overcome this, clients' predictability should be evaluated incrementally for shift detection over time.

The primary objective of this paper is to improve forecast quality through financial actor segmentation. In our research, we attempted to do so by extending to the whole population the forecast made for a subset of customers with predictable behavior. Moreover, our incremental learning approach helps to reduce forecast errors caused by environment instability such as pandemics or military action. Source code can be found in ref. [1].

2. Related Works

2.1. Predictability Dynamics

The first mention of dividing financial clients by their predictability can be found in ref. [6], where the authors describe the method of binary client classification based on the predictability of a certain event; clients were divided based on a dataset's median quality metric. The main idea of the method is to perform client segmentation without using a prediction model beforehand. However, a client's predictability is bound to change, which was taken into account in ref. [7]. Not only did the authors use incremental learning techniques for dynamic classification, but they also described the procedure for classifying actors into 32 classes according to the predictability of five chosen transactions. The predictability dynamic can be seen as a transition among classes over time.

Another view on the dynamics of predictability was described in ref. [8], where the dynamic is shown from a predictability quality and model sensitivity perspective. As opposed to previously mentioned works, classification of entities was conducted according to earlier chosen quality metric thresholds, which resulted in five groups.

In this paper, we use the same general idea as in refs. [6,7], including the application of an LSTM model (a recurrent neural network with long short-term memory [9]), but with several alterations: actors are classified based on a forecast quality threshold (similar to [8]) of all transactions; predictions are calculated on all levels, not just the micro- one; the classes of actors are utilized to lower the forecast uncertainty of all clients.

2.2. Incremental Learning

As stated in ref. [10], incremental learning is a learning system that can continuously learn new knowledge and maintain most of the previously learned ones. The main scenarios of incremental learning and its problems can be found in refs. [10,11]. We would do the most basic scenario, also described as fine-tuning to showcase that even simple models can improve upon non-incremental ones when the goal is to overcome a concept drift [12] once critical events occur; the concept drift in our research may appear in a binary classification task on the micro-level when a client stops making transactions for a long time or suddenly starts making them.

According to the first classification of incremental learning scenarios [13], our scenario is domain-incremental learning, according to the second one [14]—"new instances" scenario, because binary classes on the micro-level stay the same, only new samples arrive. The authors in ref. [15] have the same domain scenario but for human state monitoring; they showed that recently developed incremental models have trouble accumulating new knowledge, as opposed to simple models, mainly the ones with replay and cumulative strategies. The authors of another work with the same scenario [16] describe new models' inability to prevent catastrophic forgetting; they perform worse than a replay model. As a result, our most simple model should be able to overcome the concept drift as well as accumulate new information, which is the goal because the model has to adapt and show good forecast quality even after critical events.

3. Materials and Methods

3.1. Dataset Description and Preprocessing

The dataset contains transactional data from our industrial partner: 19,262,668 transactions from 10,000 clients from 1 January 2018 to 15 August 2022. Each transaction is described by a client id, their debit card id, date of a transaction, spent amount, and merchant category code (MCC).

As issues of customers' behavior are of concern in our research, we aggregate payments distributed by MCC to the most obvious groups of consumption interest, which are 'food', 'outfit', 'dwelling', 'health', 'beauty', 'travel', 'kids', 'nonfood', 'telecom', 'fun', and 'money'. The last group is the amount of cash received from ATMs. Moreover, we gather all groups except the one named 'money' into three basic values called 'survival', 'socialisation' and 'self-realisation'. The intuition behind not considering group 'money' is explained by the unpredictable nature of clients' intents once they decide to withdraw money.

The time series for total spending on each of the values is presented in Figure 1. The series were smoothed by a moving average with a 7-day window.

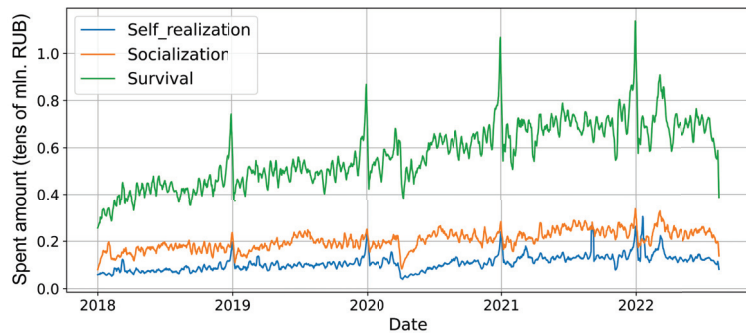


Figure 1. Time series for total expenses for all customers by basic values.

Additionally, to lower the computational cost of the experiment, 3000 of the most active clients were chosen from the available 10,000, where “most active” means the ones with the highest number of transactions through the whole time period.

3.2. Measurement Model for Micro-Level

Before a thorough description, it is worth noting that the idea of predicting the transactional behavior of just one client is usually discarded on the basis of a nearly random series of actions. Consequently, the goal of micro-level forecasting in this research is not going to be the highest possible prediction accuracy. However, the model can find, utilize, and measure recurring patterns in a client behavior sequence, if there are any, which is good enough for the objective of the paper.

The model on the micro-level is used for one client and one transaction group at a time to predict whether any transactions were made in the following days using retrospective transactional data. We consider transactions on a certain day to have occurred if the spent amount is equal to or greater than 10 money units; otherwise, this is a day without any activity. The model calculates predictions for 7 days at once, so it was decided to use the week-based pattern while preparing input data as well, which resulted in 28 days being chosen as input, each with 5 features: whether there were transactions for a given basic value, the sine and cosine of a day of the week number as Equation (1) shows, and the sine and cosine of a month number in a year (2).

$$f_1 = \sin \frac{2\pi D}{7}; f_2 = \cos \frac{2\pi D}{7}; \tag{1}$$

where f_1 and f_2 are the day of week features and D is the number of the day in a week (0 being Monday and so on).

$$f_3 = \sin \frac{2\pi M}{12}; f_4 = \cos \frac{2\pi M}{12}; \tag{2}$$

where f_3 and f_4 are the month features and M is the number of the month in a year (0 being January).

The model graph can be seen in Figure 2, where LSTM is constructed of two layers with dropout of 20% in between, dropout between dense layers is 10%, and the activation function is leaky rectified linear unit (Leaky ReLU). The graph was drawn by hand to be similar to Keras 'plot_model' function's output because the model was created in PyTorch.

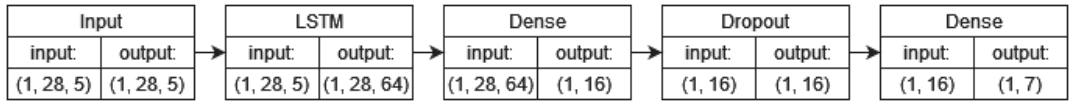


Figure 2. Architecture of the predictability measuring neural network.

The loss function is binary cross entropy, optimizer— Adam with a learning rate 5×10^{-4} . Training was conducted with a batch size of 128 for 1000 epochs with early stopping: if a test error is not improving after 10 epochs, the learning rate is lowered by 20%; no improvement after 100 epochs—stop training. Almost every client has a class imbalance: a number of days with completed transactions and days without any. To deal with this problem, the loss function was constructed with a weight parameter, calculated as a ratio of negative instances to positive ones.

3.3. Model for Time Series Forecasting on Meso- and Macro-Level

The model on the meso- and macro-level is used to predict the amount of spent money for a certain class of clients for the next day; therefore, the final layer consists of just one neuron. The model graph can be seen in Figure 3, where LSTM is stateful and is constructed of one layer, dropout between layers is 10%, the activation function is rectified linear unit (ReLU). The graph was created with the use of Keras 'plot_model' function.

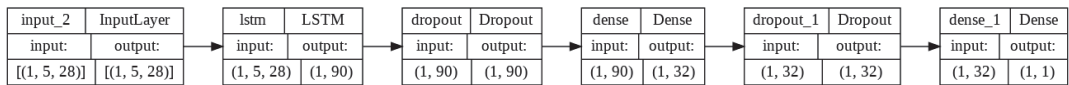


Figure 3. Architecture of neural network model for time series forecasting.

Due to the use of stateful LSTM, the batch sizes for training and validation have to match; therefore, its size was decided to be 1. The loss function is mean squared error (MSE), optimizer—Adam with a learning rate 5×10^{-5} . The model was trained for 10 epochs.

These parameters were chosen as optimal after a series of tests. The same was achieved with the measuring neural network on the micro-level.

4. Experiments

4.1. Micro-Level Predictability Measuring

As stated earlier, the objective of the measurement model on the micro-level is to predict whether any transactions will be completed on the following days or not. After encoding time features with sine/cosine transformation, each day is described by 7 features: whether there were transactions in the self-realization group, in the socialization group, in the survival group, sine of a day of the week, corresponding cosine, sine of a month, corresponding cosine.

Available data has critical time periods that should be left out of a training step, which is why the training set consists of a year's worth of days from 1 January 2018 to 31 December 2018, whereas the validation set has days from 1 January 2019 to 15 July 2019.

For further adaptation to client behavior, the model with incremental learning was trained on top of the base one: given 28 days, the model predicts the next seven and then trains on this very data with a higher learning rate (10^{-3}) and without class imbalance techniques to adapt to changes faster.

After training the base model and the incremental one for one basic value group for one client (6 models total for each client), predictions for their transactional behavior are made from 1 January 2019 to 8 August 2022. These forecasts consist of 1 value per day; therefore, the days were merged into weeks with further averaging of the results. The chosen quality metric is the F_1 -score (harmonic average of precision and recall); three values for each basic value group were used to calculate a euclidean norm of the three-component vector divided by a square root of three to keep the values in the $[0, 1]$ range.

The forecast quality for an incremental model is used to segment actors into two classes of predictability: clients with a high predictability (F_1 -score higher than 0.7) and those with a low one.

4.2. Meso- and Macro-Level Forecasting

As mentioned above, the model at these levels is going to predict the amount of spent money, not just whether there were any transactions on a given day, which is why it is necessary to prepare the data once again: for each class of clients, their sum of spent money per day over the whole time period is smoothed with a moving median with a 21-day window. Only transactions in a group of basic values called “survival” are considered during training and prediction.

The experiment was divided into four parts with two differences: a base model or an incremental one; finding class clients every week; or choosing class clients once. To provide more details regarding the second difference: the first idea consists of choosing corresponding client classes every week, whereas the second one is for working with client classes identified once for the 52nd week because it is one of the last weeks in the relatively calm year 2019. The course of action with the first idea is as follows: find clients in the given class for the current week, predict the spent amount for the following week 1 day at a time, and repeat for each of the available weeks. The plan for the second one differs just in the first step: certain class clients should not be found every week as they are already known and will not change.

All the following experiments were conducted with the same model architecture and hyperparameters to check the hypothesis that it is easier to work with highly predictable clients because they have less noise and fewer unpredictable patterns. Each model was trained in series from the beginning of 2018 until the end of 2019.

Following the objective of the paper, which was to lower the forecast error by client segmentation, the original time series of all actors were compared with the predicted amounts for actors inside a certain class. Taking into account the difference in the number of clients within these classes, the resulting series of forecasts is multiplied by the ratio of median to median, where the first median describes spent money for all clients and the second one for clients in a given class. The chosen quality metric for all forecasts on these levels is mean absolute percentage error (MAPE).

5. Results and Discussion

5.1. Client Segmentation and Incremental Learning on Micro-Level

The quality of predictions given by the base and incremental models on the micro-level can be seen in Figure 4; these are the average results for all clients and all 3 groups of basic values. The first dotted line depicts the week when COVID-19 lockdown was imposed (31 March–4 April 2020), and the second one shows the week when Russia invaded Ukraine (22 February–28 February 2022). It is clear that the incremental model consistently shows higher forecast quality and even has an upward trend in F_1 -score.

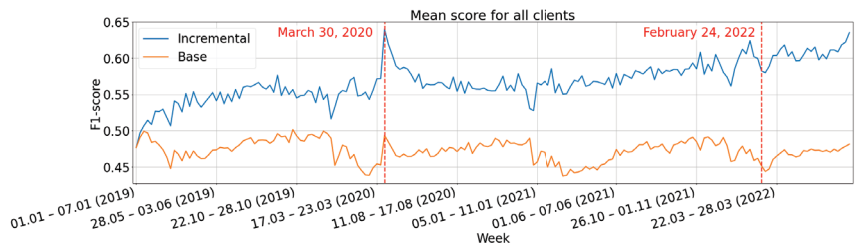


Figure 4. Incremental and base models forecast quality.

Several weeks before the lockdown began, the predictability plummeted, which can be explained by haphazard client behavior in response to the first news regarding deaths from COVID-19 in Russia: some may have begun stocking up on provisions, travelling, and visiting places of interest while it was still possible. The forecast quality for the week of 31 March is the highest, and it is important to remember that the incremental model has high forecast quality when the current week does not differ significantly from previous ones because it adapts to changes and learns from previous instances. The majority of people have been staying at home for several weeks already, so there were few if any transactions made; the incremental model has no problem following such a pattern. The base model does not have methods to adapt to changes, but perhaps some actors returned to their usual behavior seen in the training set after all unusual preparations were completed.

The week when the war began is depicted with decreased forecast quality because the news was less expected than the lockdown imposition, so there was no time to prepare; client behavior became unusual only at the start of the critical period. The given graph also shows reoccurring drops in prediction quality every New Year (1 January), which are explained by expected stress and behavior change due to big holidays.

5.2. Forecast Errors for the Same Class on Meso- and Macro-Level

The first experiment on the meso- and macro-level consisted of measuring forecast quality for the same predictability class in four variations: the trained model was used to predict spent amounts for a given class of clients, which were recalculated each week; the previous model was incrementally trained further; the trained model was used for a given class of clients found on week 52; the previous model was incrementally trained further. The results can be seen in Table 1, where the forecast quality was measured with the help of MAPE and hit probability, which were proposed in ref. [17] and measure the fraction of values in a given range. For easier interpretation, the forecasting error distribution for all experiments is presented in Figure 5.

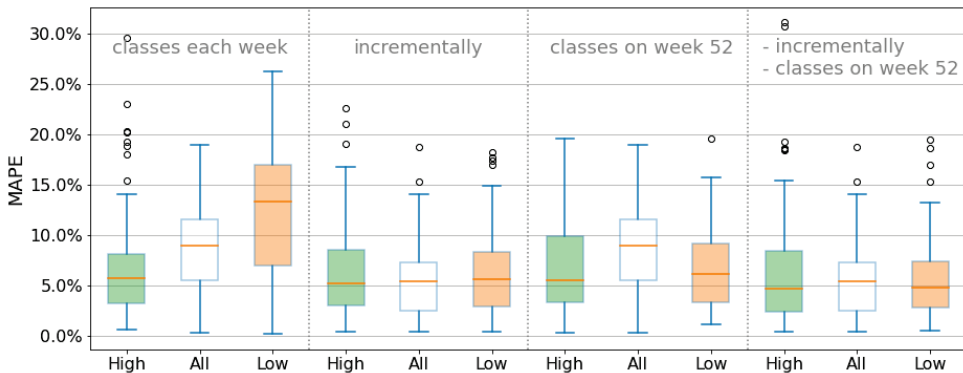


Figure 5. Forecast quality for the same class, where ‘High’ and ‘Low’ correspond to clients with high or low predictability.

Table 1. Forecast quality measured for the same class.

Chosen Client Class	Median MAPE	HP: MAPE ≤ 5%
Clients in the class were chosen every week		
Clients with high predictability	5.74%	0.422
All clients	8.97%	0.222
Clients with low predictability	13.32%	0.178
With incremental learning; classes every week		
Clients with high predictability	5.18%	0.481
All clients	5.38%	0.444
Clients with low predictability	5.62%	0.437
Clients in the class were chosen once on week 52		
Clients with high predictability	5.53%	0.393
All clients	8.97%	0.222
Clients with low predictability	6.17%	0.333
With incremental learning; classes on week 52		
Clients with high predictability	4.71%	0.548
All clients	5.38%	0.444
Clients with low predictability	4.82%	0.518

As the results show, clients with high predictability have both the smallest MAPE and the highest hit probability for all cases, which suggests that the model has an easy time training on time series belonging to well predicted agents. The first two cases, where given class clients were found every week, support the hypothesis that clients with low predictability add noise, which results in unpredictable clients having the worst results in these cases. Experiment cases where the class clients were chosen once have a different picture, which can be explained by the following suggestion: the forecast was evaluated across the test period from 1 January 2020 until 15 August 2022; therefore, the clients who were deemed unpredictable on week 52 (from the year 2019) can have predictable patterns in the following years. This assumption supports our idea to use incremental models for classifying clients based on predictability because client’s predictability is not constant.

5.3. All Clients Forecast Enhancement on Meso- and Macro-Level

The second experiment’s aim was to figure out which class of clients, if any, makes forecasting all clients’ time series more accurate. The predictions for given class clients are multiplied by a ratio described in Section 4.2 and compared with the real-time series of all clients with the help of MAPE and hit probability. The predictions are depicted in Figure 6. For easier interpretation, the boxplots for all cases are presented in Figure 7, whereas the numerical results are shown in Table 2.

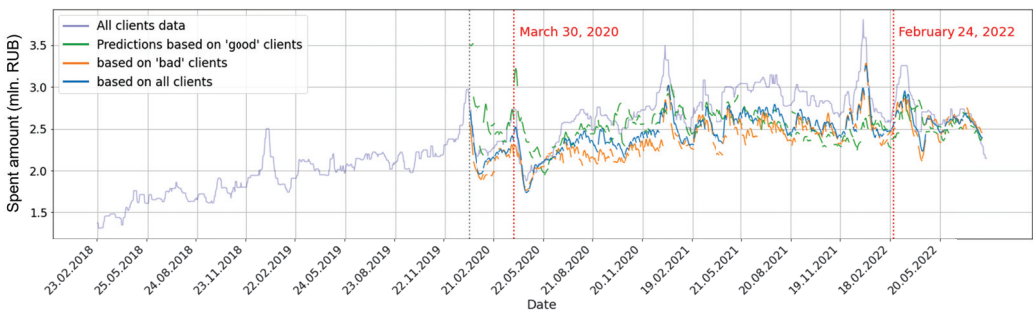


Figure 6. Predictions for all clients.

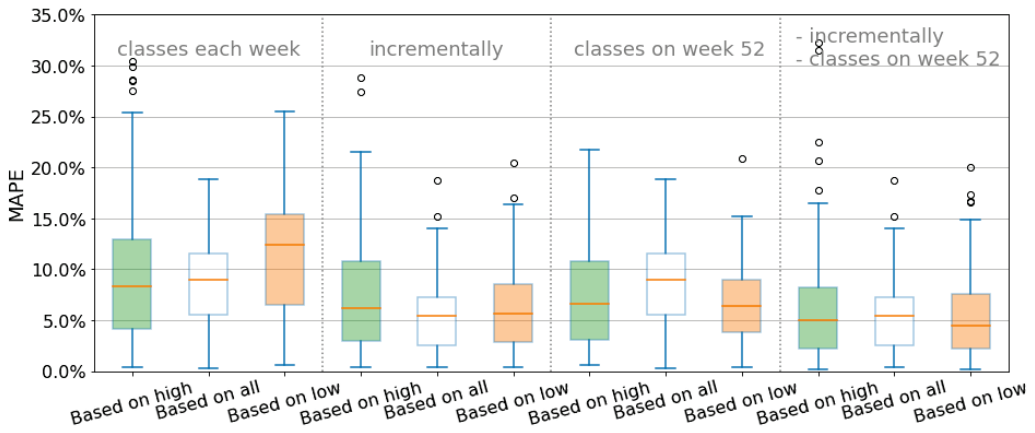


Figure 7. Forecast quality for all clients, where “Based on high” means based on clients with high predictability.

Table 2. Forecast quality measured for all clients.

Chosen Client Class	Median MAPE	HP: MAPE ≤ 5%
Clients in the class were chosen every week		
Based on clients with high predictability	8.34%	0.304
Based on all clients	8.97%	0.222
Based on clients with low predictability	12.39%	0.148
With incremental learning; classes every week		
Based on clients with high predictability	6.21%	0.370
Based on all clients	5.38%	0.444
Based on clients with low predictability	5.67%	0.459
Clients in the class were chosen once on week 52		
Based on clients with high predictability	6.65%	0.385
Based on all clients	8.97%	0.222
Based on clients with low predictability	6.35%	0.370
With incremental learning; classes on week 52		
Based on clients with high predictability	4.95%	0.508
Based on all clients	5.38%	0.444
Based on clients with low predictability	4.44%	0.548

The first experiment case shows the perfect picture: predictions based on well-predicted clients are the most accurate, with predictions based on all clients in second place. The surprises arise when either incremental learning is used or clients in classes are chosen once.

The addition of incremental learning in the second experiment case results in a seemingly strange order, where clients with low predictability deliver better values than predictions based on well predicted clients. We suppose that because unpredictable clients’ series have more noise, their time series are more similar to the ones of all clients. In addition, once the incremental model trains on examples of well predicted clients on week n , well predicted clients on week $n + 1$ may be different, and their time series may be different as well, which makes their scaled predictions look less smooth; unpredictable clients, on the other hand, have a higher chance to have a bigger overlap between subsequent weeks because the number of unpredictable clients on any week is always greater.

Another point to consider is that the client segmentation on the micro-level was completed based on binary forecast quality: whether there were any transactions in a

given day. The model on the meso- and macro-levels predicts the amount of spent money; therefore, even though the client is considered highly predictable on the micro-level (their sequence is full of *True*), they might have spent 100 money units on one day and 100,000 on the other, which makes the client a lot less predictable on other levels.

The last two cases show that predictions based on all clients perform poorly in comparison to predictions based on certain client classes on week 52. These outputs may be due to the decomposition of all clients' time series based on client predictability; forecasting distinct components is easier than the original series. Another interesting outcome is that the number of predictable clients on week 52 is 796 people out of 3000, so on the third experiment, 26.6% of the population can give more accurate predictions than when all clients are considered.

The difference between MAPE values for customers with high and low predictability for the first experiment case can be seen in Figure 8. Once the red dotted line goes under 0% MAPE, the predictions based on unpredictable clients perform better than the ones based on predictable clients. It can be observed that this happens during critical times: the lockdown announcement, the start of the war, and the time around New Year. The clients with high predictability show better results in quieter periods.

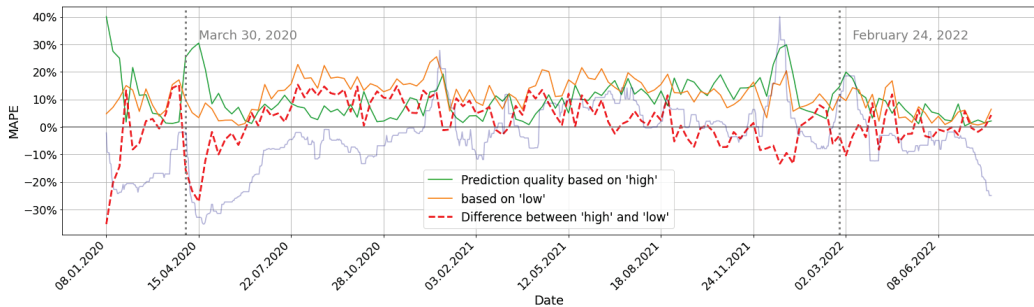


Figure 8. Difference between forecast quality based on clients with high and low predictability, where the pale blue line in the background depicts spent amount for all clients.

6. Conclusions

In our research, we developed the method for financial actors separation by predictability with the help of incremental learning on the micro-level, demonstrated its benefits and drawbacks for all clients forecasting on the macro-level, and showed how different predictability classes influence the model's forecast quality on the meso-level.

Our experiments show that the model with incremental learning was able to perform better on both the micro-levels in terms of F_1 -score and meso- and macro-levels in terms of MAPE and hit probability, which supports our idea that incremental learning is useful for financial clients' behavior analysis considering the fact that clients predictability changes due to a lot of factors and events. Our research also shows that customer separation by predictability of their consumption behavior helps to improve the forecasting quality of the whole population's consumption due to the decomposition of all clients' time series, even though the best results are delivered by using the class of clients with low predictability. Further experiments with the same forecast objective for all levels are to be held in the future to ensure the same predictability class across levels.

The results are of practical value to companies operating in the finance sector that need to analyze their customers' behavior. Our research suggests that environmental instability such as pandemics or military action can be leveled out with incremental models, whereas the problem of huge data volumes and noise can be solved by client segmentation.

Author Contributions: Data curation, M.K. and A.K.; formal analysis, M.K. and A.K.; investigation, M.K.; methodology, A.K.; project administration, A.K.; software, M.K.; supervision, A.K.; validation, M.K.; visualization, M.K. and A.K.; writing—original draft preparation, M.K. and A.K.; writing—review and editing, M.K. and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research is financially supported by the Russian Science Foundation, Agreement 17-71-30029 (<https://rscf.ru/en/project/17-71-30029/>, accessed on 27 June 2023), with co-financing of Bank Saint Petersburg.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Bank Saint Petersburg and are available at <https://cloud.bspb.ru/> (accessed on 27 June 2023) with the permission of Bank Saint Petersburg.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Koshkareva, M. Available online: https://github.com/Mpkosh/Enhancement_of_Consumption_Forecasting (accessed on 27 June 2023).
2. Dumitrescu, E.; Hué, S.; Hurlin, C.; Tokpavi, S. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *Eur. J. Oper. Res.* **2022**, *297*, 1178–1192. [CrossRef]
3. Karvana, K.G.M.; Yazid, S.; Syalim, A.; Mursanto, P. Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry. In Proceedings of the International Workshop on Big Data and Information Security (IW BIS), Bali, Indonesia, 11 October 2019; pp. 33–38. [CrossRef]
4. Al-Hashedi, K.G.; Magalingam, P. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Comput. Sci. Rev.* **2021**, *40*, 100402. [CrossRef]
5. Hernández-Nieves, E.; Hernández, G.; Gil-González, A.B.; Rodríguez-González, S.; Corchado, J.M. Fog computing architecture for personalized recommendation of banking products. *Expert Syst. Appl.* **2020**, *140*, 112900. [CrossRef]
6. Stavina, E.; Bochenina, K.; Chunaev, P. *Predictability Classes for Forecasting Clients Behavior by Transactional Data*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 187–199. [CrossRef]
7. Bezbochina, A.; Stavina, E.; Kovantsev, A.; Chunaev, P. Dynamic Classification of Bank Clients by the Predictability of Their Transactional Behavior. *Lect. Notes Comput. Sci.* **2022**, *13350*, 502–515.
8. Feng, Q.; Sun, X.; Hao, J.; Li, J. Predictability dynamics of multifactor-influenced installed capacity: A perspective of country clustering. *Energy* **2021**, *214*, 118831. [CrossRef]
9. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
10. Luo, Y.; Yin, L.; Bai, W.; Mao, K. An Appraisal of Incremental Learning Methods. *Entropy* **2020**, *22*, 1190. [CrossRef] [PubMed]
11. Gepperth, A.; Hammer, B. Incremental Learning Algorithms and Applications, In Proceedings of European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, 27–29 April 2016.
12. Lu, J.; Liu, A.; Dong, F.; Gu, F.; Gama, J.; Zhang, G. Learning under Concept Drift: A Review. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 2346–2363. [CrossRef]
13. van de Ven, G.M.; Tolias, A.S. Three scenarios for continual learning. *arXiv* **2019**, arXiv:1904.07734.
14. Lomonaco, V.; Maltoni, D. CORE50: A New Dataset and Benchmark for Continuous Object Recognition. *arXiv* **2017**, arXiv:1705.03550.
15. Matteoni, F.; Cossu, A.; Gallicchio, C.; Lomonaco, V.; Bacciu, D. Continual Learning for Human State Monitoring. *arXiv* **2022**, arXiv:2207.00010.
16. Rahman, M.S.; Wright, M.; Mandiant, S.E.C. On the Limitations of Continual Learning for Malware Classification. *arXiv* **2022**, arXiv:2208.06568.
17. Kovantsev, A. Probabilistic criteria for time-series predictability estimation. *Sci. Tech. J. Inf. Technol. Mech. Opt.* **2023**, 1–8. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Improving the Accuracy of Firm Failure Forecasting Using Non-Financial Variables: The Case of Croatian SME[†]

Tamara Kuvek¹, Ivica Pervan^{2,*} and Maja Pervan²¹ Erste&Steiermärkische Bank d.d., I. Lucica 2, 10 000 Zagreb, Croatia; tamara.pds@gmail.com² Faculty of Economics, Business and Tourism, University of Split, Cvite Fiskovica 5, 21 000 Split, Croatia; mpervan@efst.hr

* Correspondence: pervan@efst.hr; Tel.: +385-214430639

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Empirical findings based on a bivariate logistic regression model with two SME categories (successful and failed) indicate that by adding non-financial indicators to the model based on financial variables, the accuracy of forecasting increases significantly. Namely, the total classification error decreases by an average of 26.99%, while the AUROC value increases by an average of 7.33%. In the additional model, with three firm categories (successful, sensitive, and failed), the findings reveal that one financial variable (self-financing) and three non-financial variables (orderly settlement of obligations, export, and age) significantly explain the occurrence of the early stage of SME failure.

Keywords: SME; firm failure; non-financial variables

1. Introduction

Firm failure modeling has been an important research topic for many years, for both academia and practitioners in banks, investment funds, and other institutions. Firm failure often has a wide range of negative effects on numerous subjects, especially for employees, investors, creditors, and suppliers. Every new economic crisis, such as Global Financial Crisis (2007–2008), Great Recession (2008–2012), or the recent COVID-19-caused economic crisis (2020), brings this issue into the spotlight again.

The problem of firm failure has been an intriguing issue in Croatia for many years, primarily due to the large number of insolvent companies. According to official statistical data (www.dzs.hr), there were 137,664 companies in Croatia at the end of 2022, while current data (February 2023) from the state agency FINAs database (www.infobiz.hr, accessed on 5 February 2023.) reveal that 13,901 companies were insolvent because of blocked accounts (EUR 406.58 million). In other words, the Croatian business environment is quite risky since 10.5% of companies have problems meeting their due liabilities. The riskiness of doing business in Croatia was confirmed by the World Bank's Doing Business data in 2020 (<https://www.worldbank.org/en/home>, accessed on 14 January 2023.), as only 35.2% of receivables were collected in insolvency proceedings. For comparison, the percentage of receivables collection in insolvency proceedings is 90% in Slovenia, 67.5% in the Czech Republic, and 79.8% in Germany, while the average in OECD countries is 70.2%. The reason for a low percentage of claims collection in Croatian insolvency proceedings is the late opening of proceedings and the fact that many companies enter bankruptcy procedures with negative equity. In such a business environment, predicting legal failure, i.e., bankruptcy, is not very useful, but it is much more useful to create a model for predicting firm insolvency and the early stages of firm failure [1]. This research emphasizes the modeling of firm failure in the SME segment due to the large number of such companies in Croatia and their relative importance in the national economy. According to the 2021

Citation: Kuvek, T.; Pervan, I.; Pervan, M. Improving the Accuracy of Firm Failure Forecasting Using Non-Financial Variables: The Case of Croatian SME. *Eng. Proc.* **2023**, *39*, 62. <https://doi.org/10.3390/engproc2023039062>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 5 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

aggregated data retrieved from the FINAs database (www.infobiz.hr), the SME sector in Croatia comprises 55.4% of total assets and generates 58.3% of revenue.

Our study adds to the existing literature in several ways. Firstly, we developed a unique set of non-financial variables to explore how much these variables can improve firm failure forecasting. Secondly, we developed a model for the prediction of the early stage of the firm failure process, which enables timely decision making to avoid credit losses. The estimated multinomial logistic regression model indicates that one financial variable (self-financing) and three non-financial variables (settlement of obligations, export, and firm age) significantly contribute to explaining the early stage of SME failure. Finally, we conducted research for the sample of Croatian SMEs for which this kind of modeling is almost nonexistent. In addition to confirming the theoretical assumptions about the usefulness of non-financial variables, the designed model also has the possibility of practical use, especially in commercial banks.

2. Literature Review

There is a large body of literature dealing with firm failure from different perspectives; however, the main goal of almost all papers is to design a prediction model with the lowest possible forecasting error. Early studies [2–5] put focus on the use of financial indicators in the prediction of firm failure. Given that financial indicators are based on financial statements, such studies explore the usefulness of accounting information in the context of crediting decisions and firm failure modeling. As a general conclusion of the mentioned early studies, as well as many recent studies [6–9], one can point out the finding that financial indicators are useful in predicting business failure. However, studies that analyzed the predictive power of financial indicators over time showed that as the accounting data age ($t-2$, $t-3$. . .), the predictive power of financial indicators declines sharply. The accuracy of forecasting over a long period directly depends on the stationarity of the data, which implies a stable correlation between the variables in the forecast period. Empirical research has shown that this is difficult to achieve, which is emphasized by Du Jardin and Severin [10], who analyzed 34 studies and determined that the accuracy of the model decreases by 15% in 3 years before the bankruptcy. Pervan et al. [1] report similar findings in a more recent study.

Over the years, one direction of firm failure research focuses on SMEs. Namely, the modeling of firm failure for large listed companies is not identical to modeling for SME failure. The first such study was published in the US by Edminister [11], followed by numerous recent studies for SME samples. Edminister designed the model with seven financial ratios (different from Altman's Z score ratios) with a classification accuracy of 93%. Altman and Sabato [12] developed the SME failure model and compared it with Altman's Z'' (model for unlisted firms). A comparison of the SME failure prediction model and the Z'' model indicated that the SME model outperformed Altman's Z'' model by 30%. Similar research focused on SMEs can be found for SMEs from Portugal [13], Russia [14], Belgium [15], Estonia [16], etc.

To improve the predictive power of forecasting models, authors such as Gudmundson [17], Grunert et al. [18], Altman et al. [19], Pervan and Kuvek [20], Laitinen [21], Habachi and Benbachir [22], and Altman et al. [23] use non-financial variables. The general finding from most of the mentioned studies is that the inclusion of non-financial indicators in addition to financial indicators improves the accuracy of predicting firm failure. This can be explained by the characteristics of qualitative variables that do not change over time (or only partially change) and achieve more stable correlations as compared to financial variables. Previous papers often use firm age, firm size, industry, and region as a set of non-financial variables since these data are publicly available.

3. Research Design

The research sample included 4639 SME clients of a commercial bank, while the dataset incorporated data from the period 2011–2015. An important element in firm

failure modeling is the definition of the dependent variable, i.e., the firm failure variable. In countries such as Croatia, where bankruptcies are opened at the very late stage of the failure process and where the percentage of receivables collection in bankruptcy is quite low, it is much more useful to predict the early stage of firm failure than legal failure—bankruptcy. Therefore, the total sample of SMEs was divided into three categories (successful, sensitive, and failed) depending on the bank's internal credit rating and regularity in the settlement of due obligations (Table 1). The group of successful firms includes only those firms that have an intact high credit rating and that have not had any delays in settling their obligations. A firm entered the sensitive category (early stage of firm failure) if it had a reduced credit rating and a delay in meeting obligations for a duration between 30 and 90 days. Finally, the firm was classified as failed if it had the lowest credit rating with delays in the settlement of obligations longer than 90 days, accompanied by a recorded amount of loss for the bank.

Table 1. Sample structure.

SME Category	Number of Observations
Successful	3046
Sensitive	779
Failed	814
Total	4639

Following the example of similar studies, this research also uses accounting information and financial ratios as influential variables for predicting SME failure. Modern accounting frameworks (IFRS, FASB, etc.) point out that accounting information should be useful for investing and crediting decisions. Previous studies generally confirmed that accounting information and resulting financial ratios are useful as independent variables for firm failure modeling. However, some papers such as [1,3,10] point out that older accounting information results in lower prediction accuracy. In the segment of financial variables, 14 financial ratios were used, which were calculated as shown in Table 2:

Table 2. Financial variables.

Financial Variable	Acronym	Description
Return on equity	ROE	Net earnings/Equity
Return on assets	ROA	Net earnings/Assets
Operating margin	OM	Operating earnings/Sales
EBITDA to assets	EBITDAA	EBITDA/Assets
Sales to equity	SE	Sales/Equity
Operating cash flow to assets	OCFA	Net operating cash flow/Assets
Working capital	WC	Working capital/Assets
Current ratio	CR	Current assets/Current liabilities
Quick ratio	QR	Current assets-Stock/Current liabilities
Debt to assets	DA	Total debt/Assets
Self-financing	SF	Equity/Assets
Short-term debt to assets	STDA	Short-term debt/Assets
Debt to EBITDA	DEBITDA	Total debt/EBITDA
Operating cash flow to debt	OCFD	Net operating cash flow/ Total debt

To improve forecasting accuracy, further modeling of SME failure includes non-financial variables. The starting assumption is that non-financial variables (due to their characteristics) only partially change over time, which enables them to be more stable failure predictors in comparison with financial variables. A unique dataset obtained from a Croatian commercial bank enabled the development of a complex prediction model, which combines financial and non-financial variables. Therefore, this research is one of the few whose modeling includes a battery of non-financial variables, as described in Table 3.

Table 3. Non-financial variables.

Non-Financial Variable	Acronym	Description
Managerial experience	ME	Three groups (<5 years, 5–10 years, >10 years)
Business diversification	BD	Three groups (one business, two or more businesses within one industry, businesses in different industries)
Settlement of obligations	SO	Four groups (late payment up to 30 days, late payment from 30 to 60 days, late payment from 60 to 90 days, late payment for more than 90 days)
Size	S	Ln of assets
County	C	One of 21 counties in Croatia
Export	EX	Four groups (export sales 0%, up to 30% export sales, export sales from 30% to 60%, export sales more than 60%)
Age	A	Three groups (<5 years, 5–10 years, >10 years)

Regarding the use of statistical methods, a review of previous studies indicates that many papers often followed Altman [3] and used multiple discriminant analysis (MDA). Here, it is important to point out that MDA has very strict requirements (normality of explanatory factors, equal variance–covariance matrices, prior groups' probabilities) which often are not met by data. After Ohlson's [24] seminal study, the majority of later studies started to use logit/probit/logistic regression since this method is much more robust. Therefore, for this study, we employed binary logit regression and multinomial logit regression.

4. Research Results

The first logit model (Table 4) includes only financial variables, and given a large number of financial variables, it was important to control for the potential problem of multicollinearity. Due to the high correlation ($r > 0.8$) with other variables, two variables (STDA and DA) were omitted from further analysis. In this model, the dependent variable, SME failure, can take only one of two values (failed—1; successful—0). The application of the Prabhakaran algorithm [25] in the R application resulted in the following final model with financial variables.

Table 4. Bivariate logit model with only financial variables (FVs).

Variable	Estimate	St. Error	Z Value
Const.	0.2150	0.2289	0.939
WC	−2.0607 ****	0.5271	−3.909
SF	−5.4357 ****	0.7314	−7.431
OM	−2.8503 ***	0.8898	−3.203
ROE	−0.3980	0.2327	−1.710

Significances: **** $p \approx 0$; *** $p < 0.001$.

Three statistically significant financial variables (WC, SF, and OM) had a negative sign, which, under theoretical expectations, indicates that greater liquidity, self-financing, and profitability reduce the probability of failure. However, a model based only on financial variables shows the instability of predictions because model error increased over time (from 7.91% in 2015 to 13.27% in 2011). The same conclusion can be drawn for the AUROC value, which decreased over time (from 89.34% in 2015 to 86.36% in 2011).

To improve prediction accuracy and reduce the model instability, in the next step, we added non-financial variables from Table 3. Non-financial variables (except for the size variable) were first transformed into multi-level factor variables [26] with the initial category dropped from the regression (base category). The Prabhakaran algorithm and

R application estimated model were used with financial and non-financial variables, as presented in Table 5.

Table 5. Bivariate logit model with financial and non-financial variables (F&NFV).

Variable	Estimate	St. Error	Z Value
Const.	2.1976	1.8158	1.210
WC	−1.8168 **	0.8411	−2.160
SF	−4.0941 ****	0.9895	−4.138
OM	−3.8662 ***	1.4867	−2.600
S	−0.3061	0.2678	−1.143
A 5–10 y	−2.0328 ****	0.6100	−3.333
A > 10 y	0.3079	0.8719	0.353
ME 5–10 y	−1.5885 **	0.7128	−2.228
ME > 10 y	−1.8246 **	0.8042	−2.269
SO 30–60 d	−0.0903	1.0309	−0.088
SO 60–90 d	0.9045	0.9897	0.917
SO > 90 d	3.7638 ****	0.6429	5.854

Significances: **** $p \approx 0$; *** $p < 0.001$; ** $p < 0.01$.

Of seven non-financial variables included in the modeling, three were found to be statistically significant (age, management experience, and obligation settlement). As expected, the aging of SMEs (5–10 years) and longer management experience (>5 years) reduce the probability of SME failure. In addition, late obligation payments for more than 90 days significantly explain SME failure. Empirical findings based on a bivariate logit model (successful and failed firms) indicate that by adding non-financial indicators into the model based on financial variables, the accuracy of forecasting increases significantly (Table 6). In particular, the total classification error decreases by an average of 26.99%, while the AUROC value increases by an average of 7.33%.

Table 6. Comparison of model error and AUROC.

Year	Model Error (%)		AUROC (%)	
	FV	F&NFV	FV	F&NFV
2011	7.91	5.04	89.34	97.20
2012	7.00	4.74	90.99	96.65
2013	9.27	6.36	89.20	96.61
2014	11.21	7.65	86.87	95.04
2015	13.27	12.84	86.36	89.73

In the additional model, the dependent variable, SME failure, was grouped into three categories: successful (0Y), sensitive (1Y), and failed firms (2Y). The test for combining dependent categories [27] starts from the null hypothesis H_0 , which asserts that no independent variable significantly predicts the m category of the dependent variable in relation to the n category of the dependent variable, and that categories m and n cannot be distinguished from each other in relation to the variables in the model. All combinations of the categories of the dependent variable (Table 7) in the estimation sample have statistically significant Chi2 ($p < 0.05$) values, which indicates that the categories of the dependent variable cannot be combined, as they are mutually independent, and according to the test of combining dependent variables, the conditions are met for the application of the multinomial approach.

Table 7. Test for combining dependent categories.

	Chi2	df	p > Chi2
Successful and sensitive	2113.08	10	0.0001
Successful and failed firms	3720.04	10	0.0001
Sensitive and failed firms	837.44	10	0.0001

Particular interest was in the sensitive firms' category (1Y) because it is interesting to investigate whether entering the early stage of firm failure prediction can be forecasted with the proposed set of financial and non-financial variables. The estimated multinomial logit regression model (Table 8) indicates that one financial variable (self-financing) and three non-financial variables (orderly payment of obligations, export, and age of the company) significantly explain the occurrence of the early stage of firm failure. The direction of the influence of quantitative and qualitative variables on the probability of the occurrence of the early stage of failure (1Y) concerning the successful category (0Y) is in line with theoretical expectations. The regression coefficients of self-financing (SF) and the qualitative variables' regularity of settlement of obligations (SO) and export (EX) have a negative sign, which indicates that the probability of the early stage failure is higher in SMEs that have a smaller share of self-financing, which are not exporters and which are late in settling their due obligations. The positive sign with the qualitative variable age (A) suggests that SMEs that have been present on the market for more than 5 years are less likely to enter the early stage of failure.

Table 8. Multinomial panel with financial and non-financial variables.

	Coefficient	St. Error	Z Value	p
0Y	Base Outcome			
1Y				
SF	−0.6096	0.2238	−2.72	0.006
OCFD	−0.0469	0.0412	−1.14	0.254
BD-MBI	−0.7155	0.2413	−0.30	0.767
BD-MBMI	−0.5302	0.3135	−1.69	0.091
SO 30–60 d	4.4176	0.2054	21.51	0.000
SO 60–90 d	4.8246	0.2226	21.67	0.000
SO > 90 d	6.5753	0.5190	12.67	0.000
EX < 30%	−0.7165	0.2003	−3.58	0.000
EX 30–60%	−0.8018	0.3646	−2.20	0.028
EX > 60%	−0.3717	0.3107	−1.20	0.232
A 5–10 y	−0.6534	0.2221	−2.94	0.003
A > 10 y	−0.9005	0.3541	−2.54	0.011
Const	−1.9046	0.1574	−12.10	0.000
2Y				
SF	−0.6185	0.2228	−2.71	0.007
OCFD	−0.2855	0.0907	−3.15	0.002
BD-MBI	−0.8810	0.2807	−3.14	0.002
BD-MBMI	−1.6507	0.4590	−3.60	0.000
SO 30–60 d	3.6686	0.5651	6.49	0.000
SO 60–90 d	6.0407	0.3984	15.16	0.000
SO > 90 d	10.3727	0.5884	17.63	0.000
EX < 30%	−1.1166	0.3018	−3.70	0.000
EX 30–60%	−1.0661	0.5791	−1.84	0.066
EX > 60%	−0.4927	0.5031	−0.98	0.327
A 5–10 y	−1.0914	0.2620	−4.17	0.000
A > 10 y	−0.5193	0.4147	−1.25	0.210
Const	−3.7462	0.3066	−12.22	0.000

Log likelihood = −1591.53; N = 4639.

The highest classification power, exp (b), in predicting the sensitive SME (1Y) category has the variable regularity of settlement of obligations (SO), while the exp (b) values of

the other variables are much smaller (age, export, and self-financing). For example, the probability of the sensitive SME status (1Y) compared to the successful SME status (0Y) is 717.1 times higher if the delay in the settlement of obligations increases from “SO < 30 d” (base category) to “SO > 90 d”.

5. Conclusions

The results of this research confirm that the inclusion of non-financial variables in addition to financial variables into SME failure modeling improves prediction accuracy. By adding non-financial variables, total classification error decreases by an average of 26.99%, while the AUROC value increases by an average of 7.33%. The evaluated model revealed that the most important financial variables are working capital, self-financing, and operating margin. The signs for all three financial variables were negative, which, in accordance with theoretical expectations, indicates that greater liquidity, self-financing, and profitability reduce the probability of SME failure. Of all the non-financial variables tested, only age, management experience, and obligation settlement were found to be statistically significant. The aging of SMEs (5–10 years) and longer management experience (>5 years) reduce the probability of firm failure. According to theoretical expectations, a lower degree of regularity in settling obligations, i.e., late obligation payment for more than 90 days, significantly contributes to SME failure. Additional modeling, based on a multinomial logit model and three SME categories (successful, sensitive, and failed), revealed that the self-financing variable and three non-financial variables (settlement of obligations, export, and age of the company) significantly explain the occurrence of the early stage of firm failure. The findings of this research confirm the theoretical viewpoints on the usefulness of non-financial indicators in predicting SME failure and can serve as guidelines for commercial banks when developing models for assessing the credit risk of SME clients.

Author Contributions: T.K., I.P. and M.P. authors contributed equally. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is not available without special permission from the commercial bank that owns the data of its corporate clients.

Conflicts of Interest: The authors declare no conflict of interest.

Disclaimer: Findings from this study are the author’s views and not the Erste&Steiermärkische Bank d.d. views on the investigated issue of SME failure.

References

1. Pervan, I.; Pervan, M.; Kuvck, T. Firm Failure Prediction: Financial Distress Model vs. Traditional Models. *Croat. Oper. Res. Rev.* **2018**, *9*, 269–279. Available online: <https://hrcak.srce.hr/file/310571> (accessed on 18 December 2022). [CrossRef]
2. Beaver, W. Financial ratios as predictor of failure, empirical research in accounting: Selected studies 1966. *J. Account. Res.* **1967**, *4*, 71–111. [CrossRef]
3. Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 589–609. [CrossRef]
4. Deakin, E.B. A Discriminant Analysis of Predictors of Business Failure. *J. Account. Res.* **1972**, *10*, 167–179. [CrossRef]
5. Taffler, R.J. Forecasting Company Failure in the UK using Discriminant Analysis and Financial Ratio Data. *J. R. Soc.* **1982**, *3*, 342–358.
6. Lukason, O.; Laitinen, E.K. Firm failure processes and components of failure risk: An analysis of European bankrupt firms. *J. Bus. Res.* **2019**, *98*, 380–390. Available online: <https://www.sciencedirect.com/science/article/abs/pii/S0148296318303126> (accessed on 22 October 2022). [CrossRef]

7. Smiti, S.; Soui, M. Bankruptcy Prediction Using Deep Learning Approach Based on Borderline SMOTE. *Inf. Syst. Front.* **2020**, *22*, 1067–1083. Available online: <https://link.springer.com/article/10.1007/s10796-020-10031-6> (accessed on 22 October 2022). [CrossRef]
8. Tong, Y.; Serrasqueiro, Z. Predictions of failure and financial distress: A study on Portuguese high and medium-high technology small and mid-sized enterprises. *J. Int. Stud.* **2021**, *14*, 9–25. Available online: <https://www.proquest.com/docview/2546876504?pq-origsite=gscholar&fromopenview=true> (accessed on 5 December 2022). [CrossRef]
9. Crespi-Cladera, R.; Martín-Oliver, A.; Pascual-Fuster, B. Financial distress in the hospitality industry during the COVID-19 disaster. *Tour. Manag.* **2021**, *85*, 104301. [CrossRef]
10. Du Jardin, P.; Severin, E. Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of a financial failure model. *Decis. Support Syst.* **2011**, *51*, 701–711. [CrossRef]
11. Edminster, R.O. An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction. *J. Financ. Quant. Anal.* **1972**, *7*, 1477–1493. [CrossRef]
12. Altman, E.I.; Sabato, G. Modeling Credit Risk for SMEs: Evidence from U.S. Market. *Abacus* **2007**, *43*, 332–357. Available online: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6281.2007.00234.x> (accessed on 24 February 2023). [CrossRef]
13. Pindado, J.; Rodrigues, L.F. Parsimonious models of financial insolvency in small companies. *Small Bus. Econ.* **2004**, *22*, 51–66. [CrossRef]
14. Lugovskaya, L. Predicting default of Russian SMEs on the basis of financial and non-financial variables. *J. Financ. Serv. Mark.* **2009**, *14*, 301–313. [CrossRef]
15. Cultrera, L.; Bredart, X. Bankruptcy prediction: The case of Belgian SMEs. *Rev. Account. Financ.* **2016**, *15*, 101–119. Available online: <https://www.emerald.com/insight/content/doi/10.1108/RAF-06-2014-0059/full/html> (accessed on 2 May 2022). [CrossRef]
16. Susi, V.; Lukason, O. Corporate governance and failure risk: Evidence from Estonian SME population. *Manag. Res. Rev.* **2019**, *42*, 703–720. [CrossRef]
17. Gudmundsson, S.V. Airline distress prediction using non-financial indicators. *J. Air Transp.* **2002**, *7*, 4–24.
18. Grunert, J.; Norden, L.; Weber, M. The role of non-financial factors in internal credit ratings. *J. Bank. Financ.* **2005**, *29*, 509–531. [CrossRef]
19. Altman, E.I.; Sabato, G.; Wilson, N. The Value of Non-Financial Information in SME Risk Management. 2008. Available online: <https://ssrn.com/abstract=1320612> (accessed on 18 December 2022).
20. Pervan, I.; Kuvek, T. The relative importance of financial ratios and nonfinancial variables in predicting of insolvency. *Croat. Oper. Res. Rev.* **2013**, *13*, 187–197. Available online: <https://hrcak.srce.hr/97397> (accessed on 18 December 2022).
21. Laitinen, E.K. Financial and non-financial variables in prediction failure of small business reorganization. *Int. J. Account. Financ.* **2013**, *4*, 1–34. Available online: <https://www.inderscienceonline.com/doi/epdf/10.1504/IJAF.2013.053111> (accessed on 4 December 2022). [CrossRef]
22. Habachi, M.; Benbachir, S. Combination of linear discriminant analysis and expert opinion for the construction of credit rating models: The case of SMEs. *Cogent Bus. Manag.* **2019**, *6*, 1685926. [CrossRef]
23. Altman, E.I.; Iwanicz-Drozdzowska, M.; Laitinen, E.K.; Suvas, A. A Race for Long Horizon Bankruptcy Prediction. *Appl. Econ.* **2020**, *52*, 4092–4111. [CrossRef]
24. Ohlson, J.A. Financial ratios and the probabilistic prediction of bankruptcy. *J. Account. Res.* **1980**, *18*, 109–131. [CrossRef]
25. Prabhakaran, S. Model Selection Approaches. Available online: <http://r-statistics.co/Model-Selection-in-R.html> (accessed on 3 April 2023).
26. UCLA, Factor Variables. R Learning Modules. Available online: <https://stats.idre.ucla.edu/r/modules/factor-variables/> (accessed on 3 April 2023).
27. Williams, R. Post-Estimation Commands for MLogit. Available online: <https://www3.nd.edu/~rwilliam/stats3/Mlogit2.pdf> (accessed on 3 April 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Hyperautomation in Super Shop Using Machine Learning [†]

Shuvro Ahmed ^{*‡}, Joy Karmoker [‡], Rajesh Mojumder [‡], Md. Mahmudur Rahman [‡],
Md. Golam Rabiul Alam and Md Tanzim Reza

Department of Computer Science and Engineering, School of Data and Sciences, Brac University, 66 Mohakhali, Dhaka 1212, Bangladesh; joy.karmoker@g.bracu.ac.bd (J.K.); rajesh.mojumder@g.bracu.ac.bd (R.M.); md.mahmudur.rahman2@g.bracu.ac.bd (M.M.R.); rabiul.alam@bracu.ac.bd (M.G.R.A.); tanzim.reza@bracu.ac.bd (M.T.R.)

* Correspondence: ahmedshuvro01@gmail.com

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

‡ These authors contributed equally to this work.

Abstract: The purpose of this research was to determine how we can optimize both customer and seller experiences in a super shop using hyperautomation technology. Here, a smart bot was employed to speed up responses of simple consumer queries by utilizing natural language processing in real time. We also used machine learning frameworks, such as XGBoost, linear regression, random forest, and hybrid models together, to predict future product demand. In addition, data mining methods, such as the Apriori algorithm, FP growth algorithm, and GSP algorithm, were used to find out which algorithm can be used to determine the right way to place a product to increase the super shop sale.

Keywords: hyperautomation; data mining; machine learning; NLP; voice bot; time-series analysis; hybrid model

1. Introduction

Hyperautomation is a business-driven automation process that combines artificial intelligence, machine learning, and robotic process automation, which can solve repetitive task patterns efficiently. In this research, we used this technology to improve a super shop in terms of service to the customers and for the internal improvement of customer-to-seller communications. Mainly three methodologies were used. For RPA, a voice bot was used, and for AI and ML data mining algorithms, prediction and forecasting were used.

The key contributions of this research are:

- (1) The voice bot and the product placement will help the customer to find their desired product very easily in an efficient way;
- (2) Product sales forecasting will help the super shop to maintain proper stock levels of products under high demand according to the market need.

2. Related Works

2.1. Chatbot and Voice Recognition Systems

The chatbot, Doly, uses NLP to converse with users and its accuracy increases with user inputs [1]. Chatbots can handle any format and generally provide accurate responses [2]. Python is needed to create BLTK tools, and adapters can employ techniques such as the dynamic programming method's edit distance and naive Bayes classifier. Chatbots can reduce effort and response times, but they are sometimes not well known and can be erroneous, causing communication gaps and cost difficulties. Chatbot performance can be improved with conversation success measures.

Citation: Ahmed, S.; Karmoker, J.; Mojumder, R.; Rahman, M.M.; Alam, M.G.R.; Reza, M.T. Hyperautomation in Super Shop Using Machine Learning. *Eng. Proc.* **2023**, *39*, 63. <https://doi.org/10.3390/engproc2023039063>

Published: 6 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

2.2. Product Placement

Store managers usually put the most profitable goods on the top. Chen, M. found the contrary [3], i.e., customer attention is focused laterally and vertically in stores. According to Y. Ito and S. Kato [4], recognizing product connection in order picking can improve the shopping timing, and the order picking travel time is very high if the products are not placed wisely. Xiang used GSP to predict enterprise dynamic costs. In a changing market, it is crucial to know how to informatize to meet consumers' brand preferences [5].

2.3. Product Sales Forecasting

Linear regression, a basic yet famous forecasting technique, was used to forecast the sales of a big superstore with an accuracy rate of 84% [6]. According to Ramachandra [7], when the dataset was balanced and using random forest regressor, it let them anticipate nonlinear trends and estimate black Friday sales with an 83.6% accuracy. XGBoost, another nonlinear algorithm, forecasts the short-term power load in [8]. To do so, a combination of K-means clustering, CART, and XGBoost with temperature and date factors were used.

3. Methodology

The use case diagram as shown in Figure 1 is our overall system.

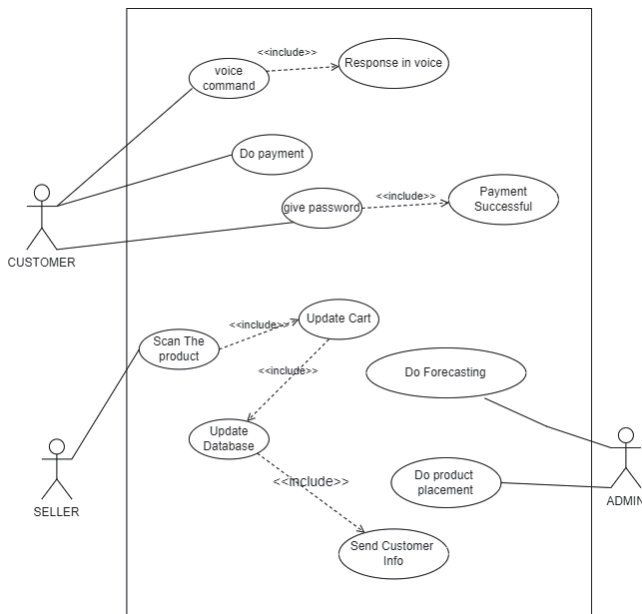


Figure 1. Super shop management system diagram.

3.1. Data Description

The real-world data source was from the Pran group “Daily shopping megastore”. It contains information from the shop’s sales from January 2022 through to December 2022 and contains 158,293 rows and 10 columns in total.

3.2. Data Preprocessing

3.2.1. Product Placement

In the dataset, we did not come across any null values. Later on, we simplified the presentation of our complete dataset by converting the data from the column that we were using into a list and applied one hot encoding to it.

3.2.2. Product Sales Forecasting (Daily)

To create a time-series data frame, we performed null checking and removed the irrelevant column. Then, we transformed the date column to a date time datatype with a day as the period. Next, we summed up the daily total quantity sold with regard to the date column and transformed the resultant column into integer type. The Dickey–Fuller test revealed data stationarity with a p -value of less than 0.05. A time plot supports the claim in Figure 2.

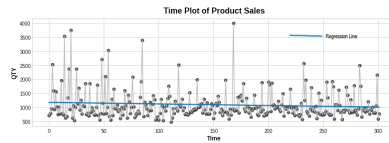


Figure 2. Product sales time plotting.

Time steps and delays are important features of time-series forecasting. We created a supervised dataset using the shift function to retrieve daily sales delayed numbers.

3.2.3. Hybrid Preprocessing

Figure 3 below is a moving average graph, which we used to try to figure out the dataset’s overall trend.

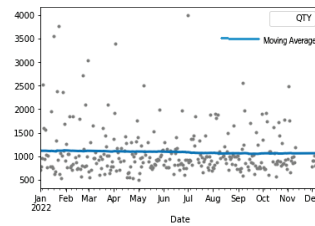


Figure 3. Moving average graph of the entire dataset.

A linear trend, which is steady because data are stationary, is analogous to the current trend. As a result, we implemented a deterministic process of order = 1. Since the trend formed below is now analogous to the one generated above, we can deduce that a linear trend might be an asset to the hybridization method. We utilized training data from the previous $(301 - 90) = 211$ days and test data from the previous 90 days in both the standalone and hybrid implementations. The generated trend is shown in Figure 4.



Figure 4. Linear trend generation.

3.3. Model Specification

Since the voice chatbot is a well-known AI-based software used by many successful software companies, we decided to include it in this hyperautomation project, where it largely worked based on two essential concepts: to participate in conversation with our customers and to answer their questions. Voice bots follow spoken commands. The voice bot technology recognizes and transcribes the input voice. The voice bot then responds to requests by text and converts them to voice. In The Figure 5, concepts of a voice bot has been explained in workflow. of voice chat bot.

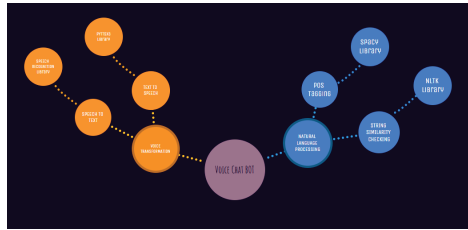


Figure 5. The concepts of a voice bot.

Workflow: PyAudio, SpeechRecognition, and pandas must be installed to create an ideal voice chatbot. Spacy beats NLTK was used for word tokenization and POS tagging. POS tagging extracts relevant text and stores it in variables. NLTK’s ‘bleu’ function compares extracted data to dictionary data. The voice bot responds with the closest comparison. Lastly, the pyttsx3 library speaks the responded text.

3.3.1. Product Placement

Market basket analysis, a data mining method, is used in retail to identify purchase trends. We used the Apriori algorithm, FP growth algorithm, and GSP algorithm.

Implementation: We used all our three algorithms in our dataset, which contains the daily sales information of a super shop. By going through the data of first 10 months of sales, we tried to establish the relations between different products and product categories. Several metrics, including support, confidence, and lift, are utilized by data mining algorithms to extract these rules.

Support refers to the frequency of an item set in the transaction dataset.

$$Support\ of\ (A \rightarrow B) = \frac{(A \cup B)}{n} \tag{1}$$

Confidence indicates how often a rule appears to be true.

$$Confidence(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A)} \tag{2}$$

Lift is a measure that tells us the probability of consequent increases or decreases given the purchase of the antecedent.

$$Lift\ of\ (A \rightarrow B) = \frac{Support(A \cup B)}{Support(A) \times Support(B)} \tag{3}$$

We accepted rules that meet these measures’ minimal thresholds. We tested which method found all the rules faster using different minimum support and minimum confidence levels.

3.3.2. Product Sales Forecasting

In this research, three algorithms were chosen to perform forecasting: linear regression, random forest, and XGBoost.

Implementation: After importing the libraries and preprocessing, we implemented linear regression. As mentioned before, simple linear regression is

$$y = (weights * features) + bias \tag{4}$$

The algorithm learns the weight of each feature and picks the weight and bias depending on the best fit goal during training. After training, we fitted and predicted the model to obtain an approximation forecast to test using the test dataset. Next, we used the decision tree-based random forest. This renowned classification model worked effectively with our dataset to average the tree output. This training used bagging. This minimized overfitting.

The Gini index determined this algorithm's root node. This showed dataset impurity. The formula for this is

$$1 - \sum_{i=1}^n (P_i)^2 \quad (5)$$

Then, we calculated the weighted Gini index, which is the total Gini index of a particular branch. The feature with lowest Gini index is chosen as the root node. Entropy can be used to calculate impurity. After setting up n estimators and the max depth parameter, which are the number of decision trees and their depth, we applied the model. Trees produced better quality but took longer. The gradient-boosted decision tree method XGBoost followed. Decision trees were used to determine this. Unlike random forest, XGBoost may change a leaf node into an if condition if it helps the model, as judged by the loss function. After the max depth, this method prunes backward. The loss function is as follows:

$$\sum_{i=1}^n l(y_i + \hat{y}_i f_t(x_i)) + \Omega(f_t) \quad (6)$$

As a result, this improved the efficiency on the whole. Next, we attempted to see whether we could improve the performance by combining linear-random forest and linear-XGBoost in a hybrid model. In summary, linear regression was used initially for both training and prediction. We then used linear regression to make forecasts about the X train-1. Then, we used the residual series to fit a second model, which was the following: train the model of target series—the predicted series from the first model. Then, we used this information to forecast using the second model that we fitted with the additional feature values (X Train-2). At this point, we combined the two forecasts to form a unified one. Here, as said above, we obtained a linear trend; thus, we trained it using linear regression, and the overall trend for the out-of-bounds sample is shown in Figure 6.

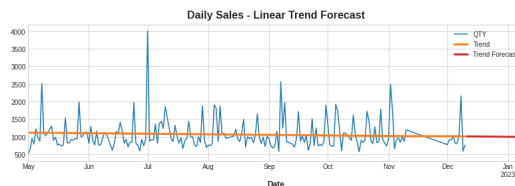


Figure 6. Linear trend forecast.

Therefore, we extrapolated the trend and then removed it by transforming y and applying the next model on the error series.

4. Results

4.1. Voice Chat Bot

Text-to-voice was performed with pyttsx3. GTTS, IBM Watson Text to Speech, and Amazon Polly are online libraries that convert text to voice. With a local speech engine, pyttsx3 can work offline. Hence, our voice chatbot can speak without the internet. The Spacy library was used for part-of-speech tagging because it tokenizes words quickly and accurately. Nltk's bleu function is more accurate than Spacy's similarity function since it compares the voice input text to the dataset's reference data.

4.2. Product Placement

The FP growth algorithm outperformed Apriori and GSP. In Tables 1 and 2, it is clear that, while verifying with different minimum support values, the Apriori and FP growth algorithms took almost the same amount of time. However, FP growth produced more rules. The Apriori algorithm runtime rose exponentially with transactions. The FP growth algorithm's runtime exponentially grew with transactions. The FP growth method

generated rules faster than the GSP and Apriori algorithms since it only iterates the dataset twice, while the other two algorithms iterate the dataset multiple times to generate rules.

Table 1. Product category.

Algorithm	Minimum Support	Time
Apriori	0.03	1.126 s
	0.05	0.413 s
FP growth	0.03	0.473 s
	0.02	0.493 s
GSP	0.03	26.418 s
	0.02	86.762 s

Table 2. Product name.

Algorithm	Minimum Support	Time
Apriori	0.002	3.493 s
	0.005	0.428 s
FP growth	0.002	2.316 s
	0.005	1.927 s
GSP	0.002	10,530.413 s
	0.005	4255.696 s

As the FP growth algorithm in Figures 7 and 8 performed better in comparison to the other two, we suggest using this algorithm for finding product placement rules.

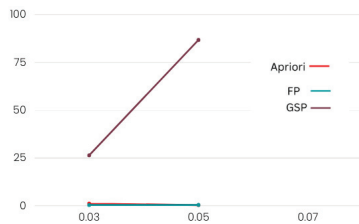


Figure 7. Product category performance.

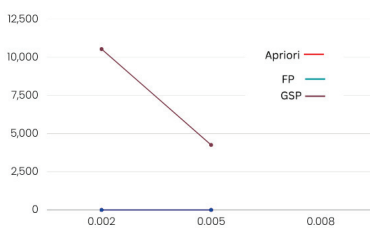


Figure 8. Product name performance.

4.3. Product Sales Forecasting

We evaluated the algorithms using mean absolute error, root mean squared error, and mean absolute percentage error. All three individual algorithms exhibited bad performance. Hybridization improved both models to a 90% accuracy. This is because linear regression assists XGBoost and random forest in learning how to extrapolate trends beyond the training data. Mean absolute error measures the forecast-to-actual difference, but unfavorable

outcomes are possible. Mean squared error may be used to calculate distance, although the unit is squared. Root mean squared error removes this. Finally, we calculated the total error using the mean absolute percentage error to interpret the forecast.

$$MAPE = 1 \div n(\sum_{i=1}^n |(At - Ft) \div At|) \tag{7}$$

where At is the actual value, Ft is the forecasted value, and n is the number of summation iterations.

$$MAE = (\sum_{i=1}^n |yi - xi|) \div n \tag{8}$$

$$RMSE = \sqrt{\sum_{i=1}^n (xi - \bar{xi})^2 \div n} \tag{9}$$

where xi is observations from the time series, xi bar is the estimated time series, and n is the number of nonmissing data points.

In the following, we verify the performance of the models using these metrics.

In Table 3, the first three rows show the results of algorithms individually, and the last two show the hybrid models. Hybrid linear regression–XGBoost performed better as shown in Figure 9, while the hybridization of linear–random forest regression produced a slightly lower accuracy, as shown in Figure 10. However, as seen in Figures 11–13, the individual models performed very poorly. Thus, hybrid linear regression–XGBoost was selected as the basis for our model and the current predictions.

Table 3. Performance of different machine learning models.

Algorithm	MAE	RMSE	MAPE
Linear Regression	343.52	427.80	0.33
Random Forest	363.71	513.90	0.34
XGBoost	346.66	498.37	0.32
Linear Regression-Random Forest	82.71	93.58	0.077
Linear Regression-XGBoost	68.94	77.48	0.09

For a further comparison with R-Squared metrics, which is a statistical fit metric that quantifies the proportion of a dependent variable’s variance that can be accounted for by the independent variable(s) in a regression, the linear regression-XGBoost result was 0.963 and linear regression-random forest was 0.943.

$$R^2 = 1 - RSS \div TSS \tag{10}$$

where RSS is the sum of the square of the residuals and TSS is the total sum of the squares.

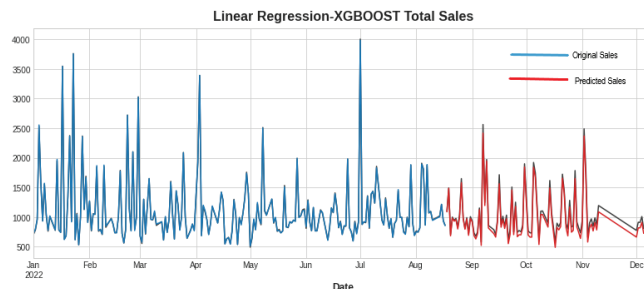


Figure 9. Forecasting using linear regression–XGBoost.

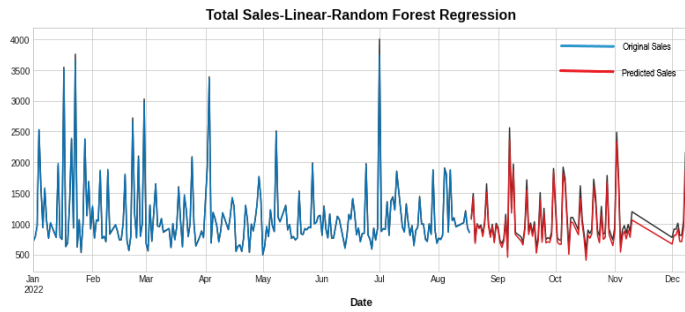


Figure 10. Forecasting using linear regression–random forest.

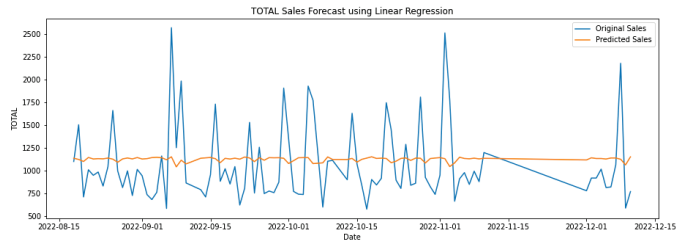


Figure 11. Forecasting using linear regression.

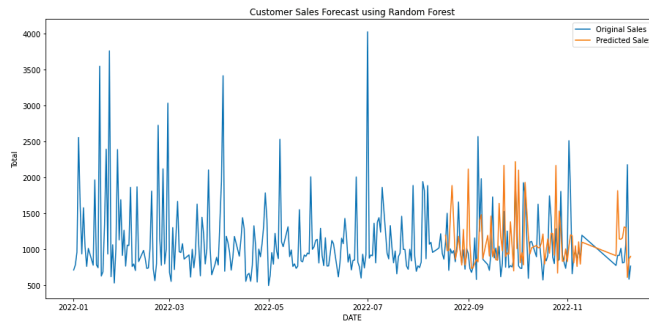


Figure 12. Forecasting using random forest.

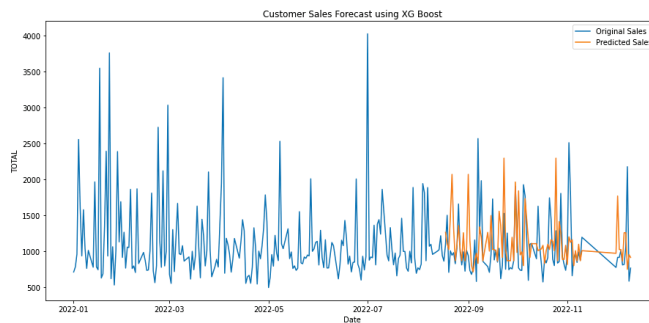


Figure 13. Forecasting using XGBoost.

5. Conclusions

In the modern technological era, hyperautomation is having a revolutionary impact in the relevant fields. In our research, we show its positive impact in the supermarket

using several methods, as discussed above. First of all, the system features a sophisticated AI-powered voice chatbot that effectively comprehends customer inquiries through advanced speech recognition and natural language processing (NLP) techniques. This was designed to provide accurate responses to customer queries using machine learning (ML), and it operates seamlessly even without an internet connection. In addition, the FP growth algorithm performed best among all the algorithms used in the product placement methodology. Using this algorithm, shopkeepers will be able to place products according to the customer's choice and it will help them to grow their business. Moreover, they will not have to worry about how to place their products. Lastly, the hybrid linear regression–XGBoost outperformed every single algorithm in product sales forecasting. Thus, it was chosen to be the basis for our custom model. This ensures business owners can obtain a complete picture of future product sales. In the future, our research will focus on working on hyperautomation features more.

Author Contributions: Conceptualization, S.A. and J.K.; methodology, S.A., J.K. and M.M.R.; software, S.A., J.K., M.M.R. and R.M.; validation, S.A., J.K. and M.M.R.; formal analysis, S.A.; investigation, M.M.R. and R.M.; resources, S.A.; data curation, S.A. and J.K.; writing—original draft preparation, S.A. and J.K.; writing—review and editing, S.A., J.K., M.M.R. and R.M.; visualization, S.A.; supervision, M.G.R.A. and M.T.R.; project administration, M.T.R.; funding acquisition, S.A., J.K., M.M.R. and R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are grateful to the faculties of BRAC University for administrative and technical support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kowsher, M.; Tithi, F.S.; Ashraful Alam, M.; Huda, M.N.; Md Moheuddin, M.; Rosul, M.G. Doly: Bengali Chatbot for Bengali Education. In Proceedings of the 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, 3–5 May 2019; pp. 1–6. [CrossRef]
2. Tiwari, A.; Talekar, R.; Patil, S.M. College information chat bot system. *Int. J. Eng. Res. Gen. Sci.* **2017**, *5*, 131–137.
3. Chen, M.; Burke, R.R.; Hui, S.K.; Leykin, A. Understanding Lateral and Vertical Biases in Consumer Attention: An In-Store Ambulatory Eye-Tracking Study. *J. Mark. Res.* **2021**, *58*, 002224372199837. [CrossRef]
4. Ito, Y.; Kato, S. Dynamic Product Placement Method in Order Picking Using Correlation between Products. In Proceedings of the 2016 IEEE 5th Global Conference on Consumer Electronics, Kyoto, Japan, 11–14 October 2016; pp. 1–3. [CrossRef]
5. Xiang, C.; Xiong, S. The GSP algorithm in dynamic cost prediction of enterprise. In Proceedings of the 2011 Seventh International Conference on Natural Computation, Shanghai, China, 26–28 July 2011; Volume 4, pp. 2309–2312.
6. Gopalakrishnan, T.; Choudhary, R.; Prasad, S. Prediction of Sales Value in Online Shopping Using Linear Regression. In Proceedings of the 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 14–15 December 2018; pp. 1–6. [CrossRef]
7. Ramachandra, H.V.; Balaraju, G.; Rajashekar, A.; Patil, H. Machine Learning Application for Black Friday Sales Prediction Framework. In Proceedings of the International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 5–7 March 2021; pp. 57–61. [CrossRef]
8. Liu, Y.; Luo, H.; Zhao, B.; Zhao, X.; Han, Z. Short-Term Power Load Forecasting Based on Clustering and XGBoost Method. In Proceedings of the IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 23–25 November 2018; pp. 536–539. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Measuring Extremal Clustering in Time Series [†]

Marta Ferreira

Centro de Matemática, Universidade do Minho, 4710-057 Braga, Portugal; msferreira@math.uminho.pt

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: The propensity of data to cluster at extreme values is important for risk assessment. For example, heavy rain over time leads to catastrophic floods. The extremal index is a measure of Extreme Values Theory that allows measurement of the degree of high-value clustering in a time series. Inference about the extremal index requires a prior choice of values for tuning parameters, which impacts the efficiency of existing estimators. In this work, we propose an algorithm that avoids these constraints. Performance is evaluated based on simulations. We also illustrate with real data.

Keywords: extreme values theory; stationary sequences; extremal index

1. Introduction

The occurrence of extreme values can lead to risky situations. Climate change, the global economic and financial crisis resulting from the COVID-19 pandemic situation, and the war in Ukraine have contributed to continuously growing attention from analysts, namely, to the risk of extreme phenomena. The duration of extreme values in time means the generation of clusters, the extension of which can aggravate the phenomenon. Extreme Values Theory (EVT) presents a set of adequate tools in this context. The extremal index is a measure of serial dependence assessing the propensity of data for the occurrence of clusters of extreme values. Figure 1 shows the maximum of sea-surge heights, where clusters of high values are visible.

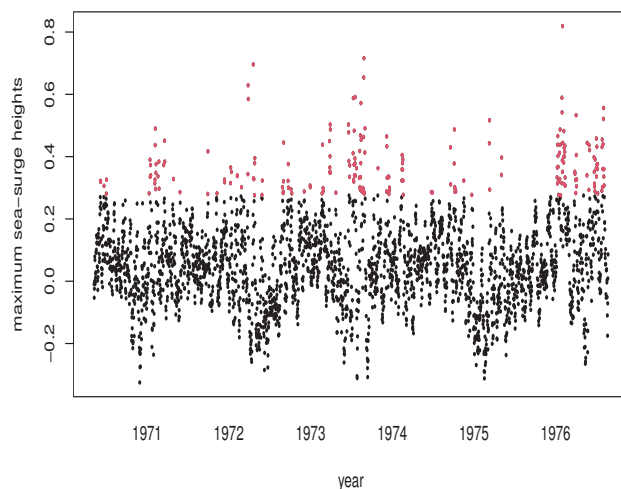


Figure 1. Maximum hourly sea-surge heights (over contiguous 15-h time periods) in years 1971–1976 at the Newlyn Coast, Cornwall, UK.

More precisely, considering $\mathbf{X} = \{X_n\}_{n \geq 1}$ as a stationary sequence of random variables (r.v.) with a common marginal distribution function (d.f.) F and denoting $M_n =$

Citation: Ferreira, M. Measuring Extremal Clustering in Time Series. *Eng. Proc.* **2023**, *39*, 64. <https://doi.org/10.3390/engproc2023039064>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 6 July 2023



Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

$\max(X_1, \dots, X_n)$, then \mathbf{X} has extremal index $\theta \in (0, 1]$ if for each real $\tau > 0$ there exists a sequence of normalized levels u_n , i.e., satisfying $n(1 - F(u_n)) \rightarrow \tau$, as $n \rightarrow \infty$, such that $P(M_n \leq u_n) \rightarrow \exp(-\theta\tau)$. In the independent and identically distributed (i.i.d.) case, we have $P(M_n \leq u_n) \rightarrow \exp(-\tau)$ and thus $\theta = 1$. On the other hand, if $\theta = 1$, then the tail behavior of \mathbf{X} resembles an i.i.d. sequence. Clustering of extreme values takes place whenever $\theta < 1$, and the smaller the θ is, the larger is the propensity for clusters to appear. Under some dependence conditions, θ is stated as the arithmetic inverse of the mean cluster size (Hsing et al. [1] 1988).

Assuming F is continuous, we have $U_i = F(X_i)$, $i = 1, \dots, n$ standard uniform r.v. and $P(-n \log(F(M_n)) \geq \tau) \approx P(n(1 - F(M_n)) \geq \tau) = P(M_n \leq u_n) \rightarrow \exp(-\theta\tau)$, with $F(M_n) = \max(U_1, \dots, U_n)$. Thus, $Y_n = -n \log(F(M_n))$ and $Z_n = n(1 - F(M_n))$ follow asymptotically an exponential distribution with parameter θ . The maximum likelihood estimator was considered by Northrop ([2] 2015) based on Y_n . More precisely, dividing the time series X_1, \dots, X_n into k_n blocks of length b_n , with $n = b_n k_n$, and considering $M_{ni} = M_{((i-1)b_n+1):(ib_n)} = \max(X_{(i-1)b_n+1}, \dots, X_{ib_n})$, $i = 1, \dots, k_n$, the maximum of the i -th block in the disjoint blocks case, and $M_{ni} = M_{((i-1)b_n):(i+b_n-1)} = \max(X_{i-1}, \dots, X_{i+b_n-1})$, $i = 1, \dots, n - b_n + 1$, the maximum of the i -th block in the sliding blocks case, the Northrop estimator is given by

$$\hat{\theta}^N = \left(\frac{1}{t_n} \sum_{i=1}^{t_n} \hat{Y}_{ni} \right)^{-1}, \tag{1}$$

where $\hat{Y}_{ni} = -b_n \log(\hat{F}(M_{ni}))$ and \hat{F} denotes the empirical d.f. estimating the usually unknown F , with $t_n = k_n$ or $t_n = n - b_n + 1$ depending on whether we are using disjoint or sliding blocks, respectively. Berghaus and Bücher ([3] 2018) considered

$$\hat{\theta}^B = \left(\frac{1}{t_n} \sum_{i=1}^{t_n} \hat{Z}_{ni} \right)^{-1}, \tag{2}$$

with $Z_{ni} = b_n(1 - \hat{F}(M_{ni}))$, a more amenable formulation to derive the asymptotic properties. Here, we consider the Berghaus and Bücher estimator with bias adjustment given by

$$\hat{\theta} = \hat{\theta}^B - 1/b_n. \tag{3}$$

We also consider the sliding blocks version since it usually performs better (Northrop [2] 2015, Berghaus and Bücher [3] 2018).

Observe that the estimators above only depend on a tuning parameter: the block length $b \equiv b_n$. This is an advantage of these methods since most estimators of θ presented in the literature have two sources of uncertainty and thus two parameters to be defined in advance: the clustering generation of high values and the choice of a high threshold above which the clusters occur. To mention the best known ones, there are the Nandagopalan ([4] 1990), Runs and Blocks (Weissman and Novak, [5] 1998 and references there in), K -gaps (Süveges and Davison, [6] 2010), censored/truncated (Holševský and Fusek, [7,8] 2020/22), and cycles estimator (Ferreira and Ferreira, [9] 2018). We also refer to other estimators that require a single tuning parameter, such as the intervals estimator, which needs to fix a high threshold (Ferro and Segers, [10] 2003), and, similar to the Northrop estimator above, where we only choose the block length for maxima, we cite Gomes ([11] 1993), Ancona-Navarrete and Tawn ([12] 2000), and Ferreira and Ferreira ([13] 2022).

As already highlighted in the literature, there is no simple optimal methodology for the best choice of block length and a single estimate for θ . In EVT, we have a typical bias-variance trade-off observed in sample path estimates of rare event parameters. For block estimators, the bias decreases with b while the variance increases. A recurrent method is to plot the estimates obtained for successive block size values and visually identify case-by-case plateau zones of these estimates. The stability around a value is an indicator

of a reasonable estimate, and this stability region, in general, should have neither too small nor too large a value of b due to the trade-off between bias and variance already mentioned. Figure 2 is a plot of the trajectory of estimates (full line) along with 95% confidence intervals (CI) (dashed line) obtained for each block length b from 1 to 100 in a random sample of dimension 1000 generated from a moving maximum model with standard Fréchet margins. We can see a plateau region in the estimates around the true value (horizontal line) $\theta = 0.5$ for the block sizes between 25 and 45. Observe the large variability occurring for large values of b and the higher bias for small values of b .

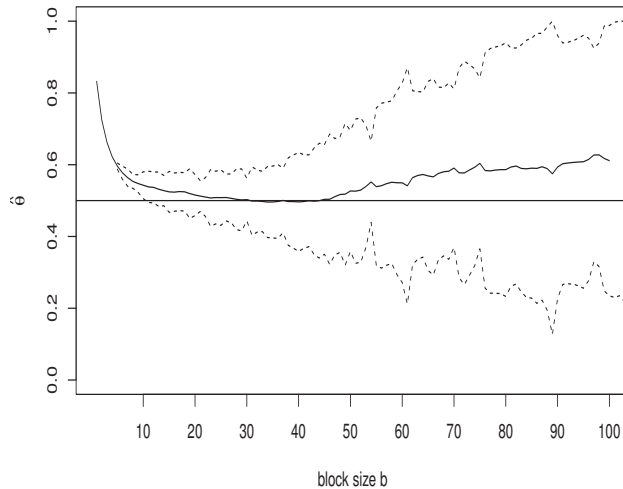


Figure 2. Estimates of $\hat{\theta}$ given in (3) for successive values of block size $b = 1, \dots, 100$ (full line) obtained for a sample simulated from a moving maxima Fréchet model with $\theta = 0.5$ (horizontal line). The dashed lines correspond to 95% CI.

Some methods have been proposed in the literature to help in the choice of tuning parameters based on the stability regions of the estimates graph: see, e.g., Frahm et al. ([14] 2005), Gomes and Neves ([15] 2020), and their references. In particular, the algorithm proposed in Frahm et al. ([14] 2005) was implemented in the context of estimating the bivariate tail dependence, and in Ferreira ([16] 2018), it was applied to extremal index estimators requiring the choice of a high threshold. In this work, our objective is to propose an adaptation of the algorithm developed in Frahm et al. ([14] 2005) applied to estimator (3) in order to find a suitable plateau of estimates taking into account the bias–variance trade-off. As a byproduct, this will allow us to circumvent the unique tuning parameter selection corresponding to the block size of where the sequence of maximums will be extracted, as described above. The method will be detailed in Section 2 and analyzed through simulation in Section 3. We end with an application to real data.

2. Estimation Method

Our proposed estimation of θ is based on the bias-corrected estimator $\hat{\theta}$ in (3) by considering sliding blocks and on the heuristic plateau-finding algorithm of Frahm et al. ([14] 2005).

The algorithm is described in the following steps:

- Step 1. Calculate estimates $\hat{\theta}_b$ from estimator (3) for $1 \leq b \leq t < n$;
- Step 2. Smooth the results of the previous step by taking means of $2w + 1$ successive estimates; we consider bandwidth $w = \lfloor 0.02t \rfloor$;
- Step 3. Define plateaus of length $m = \lfloor \sqrt{t - 2w} \rfloor$, i.e., $p_j = (\hat{\theta}_j, \dots, \hat{\theta}_{j+m-1})$, $j = 1, \dots, t - 2w - m + 1$;

- Step 4. Compute the standard deviation s of $\bar{\theta}_1, \dots, \bar{\theta}_{i-2w}$ and choose the first plateau p_j satisfying $\sum_{i=j+1}^{j+m-1} |\bar{\theta}_i - \bar{\theta}_j| \leq 2s$;
- Step 5. The extremal index is estimated through $\frac{1}{m} \sum_{i=1}^m \bar{\theta}_{j+i-1}$, i.e., taking the average of the estimates that constitute the plateau chosen in the previous step. This is denoted the plateau estimator.

The estimators (1), (2), and (3) are already implemented in package *exdex* of software R (Northrop and Christodoulides [17] 2019) with the respective CIs. We use package *exdex* to compute estimator (3) under sliding blocks and the respective upper and lower 95% CI bounds. We also apply Steps 1, 2, and 3 to the lower and upper bounds of the CIs. Once the plateau of *theta* estimates is chosen in Step 4, we pick the corresponding plateau in the CI limits, and in Step 5, we apply the average of the plateau values of the lower limit of the CI as well as the average of the plateau values of the upper limit of the CI.

We are going to analyze the estimation method described above through simulation. The models that will be used are the following:

- First-order auto-regressive model with Cauchy standard marginals (ARC), $X_i = \rho X_{i-1} + \epsilon_i$, $\{\epsilon_i\}$ i.i.d. having Cauchy d.f. with mean 0 and scale $1 - |\rho|$ and $\theta = 1 - \rho$ if $\rho > 0$ (Chernick et al. [18], 1991); we consider $\rho = 0.9$ and $\theta = 0.1$;
- An m -dependent model (MMU), $X_i = \max(U_i, U_{i+1}, \dots, U_{i+m-1})$, $i \geq 1$, where $\{U_i\}$ is an i.i.d. sequence of r.v. (Newell [19] 1964) with $\theta = 1/m$; we consider U_i , $i \geq 1$, standard uniform r.v., and $m = 3$, and thus, $\theta = 1/3$;
- Moving maxima Fréchet model (MMF), $X_i = \max_{j=0, \dots, d} a_j Z_{i-j}$ with $a_j \geq 0$, $\sum_{j=0}^d a_j = 1$ and $\{Z_i\}$ i.i.d. standard Fréchet where $\theta = \max_{j=0, \dots, d} a_j$ (Weissman and Cohen [20] 1995); we consider $d = 2$ and parameters $a_0 = 1/3$, $a_1 = 1/6$, and $a_2 = 1/2$, and thus, $\theta = 1/2$;
- ARCH(1) process, $X_i = (\beta + \alpha X_{i-1}^2)^{1/2} \epsilon_i$, with i.i.d. Gaussian innovations $\{\epsilon_i\}$, $\alpha = 0.7$, and $\beta = 2 \cdot 10^{-5}$, where $\theta = 0.721$ (Cai, [21] 2019);
- First-order max auto-regressive (MAR), $X_i = \max(\phi X_{i-1}, \epsilon_i)$, $i \geq 1$, $X_0 = \epsilon_1 / (1 - \phi)$, $\{\epsilon_i\}$ i.i.d. with standard Fréchet marginals and $\theta = 1 - \phi$ (Davis and Resnick [22] 1989); we consider $\phi = 0.1$ and $\theta = 0.9$;
- An i.i.d. sequence (Ind) of Fréchet r.v. where $\theta = 1$.

3. Simulation Study and Application

The simulation study is based on random generation of samples with size 1000 replicated 1000 times within each of the models described above. We consider different models with different values of θ . We apply the estimation plateau method of Section 2 both to estimate θ and the respective 95% CI lower and upper bounds. Table 1 contains the estimation global results of the plateau method. See also the simulation results of $\hat{\theta}$ given in (3) for each block size b in Figure 3 as well as the results of the plateau method. We can observe in each model that the plateau estimate (thicker gray horizontal full line) is located in a plateau zone of the sample path of estimates plotted as a function of block size b (full black line), as expected. We can also see that the plateau estimate is close to the true value (blue horizontal full line). In all cases, it is verified that the limits of the 95% CIs estimated by the plateau method (thicker gray horizontal dotted–dashed lines) include the true value of θ . In the ARCH case, the estimates closest to the true value of θ occur for large values of b where the variability is very high, which makes it difficult to apply the plateau methodology. Even so, the root mean squared error (rmse) of 0.1126 is not very expressive. The independent model (Ind) has $\theta = 1$ and, therefore, constitutes a frontier value of the parameter support, which typically leads to difficulties in statistical estimation. Still, the plateau method shows relatively low bias and rmse. Observe also that in the MAR model, we have $\theta = 0.9$, which is quite near to the boundary value of 1, and the plateau method does a very good job.

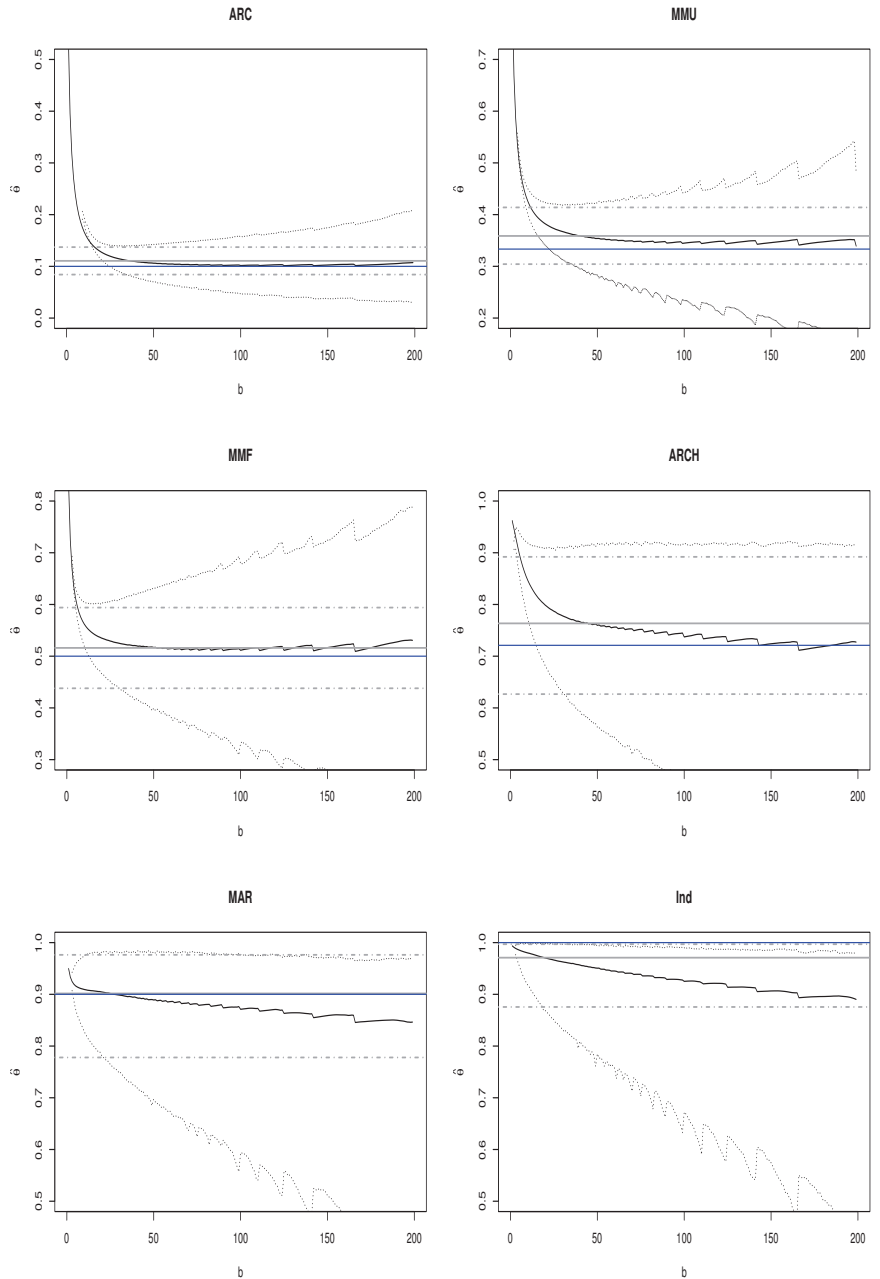


Figure 3. Simulation results: average of estimates of θ for each block size $b = 2, \dots, 200$ using $\hat{\theta}$ in (3) (full black line) and average of respective 95% CI upper and lower bounds (dotted lines); plateau estimation of θ (thicker gray horizontal full line) and respective plateau estimates of 95% CI upper and lower bounds (thicker gray horizontal dotted–dashed lines). The true value of θ corresponds to the blue horizontal full line.

Table 1. Simulation results of plateau method: average of θ estimates (mean), average of lower and upper 95% CI bound estimates, bias, root mean squared error (rmse), and standard deviation of θ estimates (sd).

	mean	lower	upper	bias	rmse	sd
ARC ($\theta = 0.1$)	0.1106	0.0841	0.1372	0.0106	0.0218	0.0190
MMU ($\theta = 1/3$)	0.3587	0.3042	0.4139	0.0254	0.0494	0.0424
MMF ($\theta = 0.5$)	0.5160	0.4379	0.5940	0.0160	0.0636	0.0616
ARCH ($\theta = 0.721$)	0.7634	0.6267	0.8920	0.0424	0.1126	0.1044
MAR ($\theta = 0.9$)	0.9017	0.7779	0.9763	0.0017	0.0827	0.0827
Ind ($\theta = 1$)	0.9709	0.8756	0.9969	-0.0291	0.0643	0.0573

Application to Real Data

We illustrate the method with the real data *newlyn* available in the R package *exdex* consisting of 2894 sea-surge heights measured at the coast of Newlyn, Cornwall, UK, over years 1971–1976. The observations correspond to the maximum hourly surge heights during periods of 15 h. See the left plot in Figure 4. Previous analysis of this data can be seen in Northrop ([2] 2015) and references therein. The sample path of estimates from (3) as a function of block size b and respective 95% confidence limits are plotted on the right graph of Figure 4. The horizontal full line corresponds to the plateau estimate of θ given by 0.2577, and the horizontal dotted–dashed lines correspond to the plateau 95% CI estimate (0.2206, 0.2948).

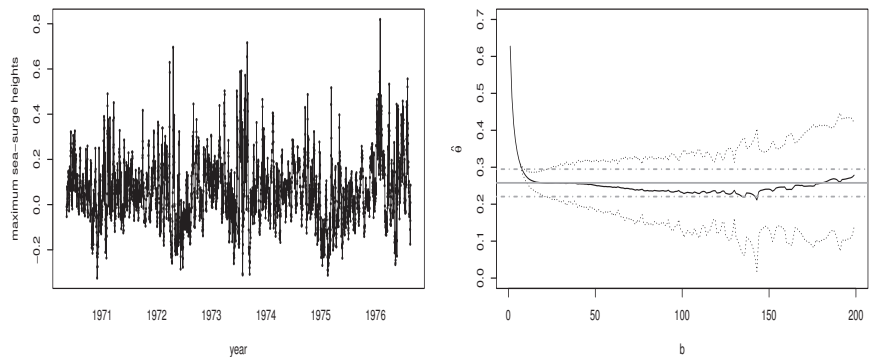


Figure 4. (Left) Maximum hourly (within successive 15-hour periods) surge height time series at Newlyn Coast, Cornwall, UK, in years 1971–1976; (Right) Sample path estimates obtained from estimator in (3) (full line) and respective 95% CI limits (dotted lines) for successive values of block size b , plateau estimate of θ (horizontal full line), and respective 95% CI plateau estimate limits (horizontal dotted–dashed lines).

4. Conclusions

This work addresses the estimation of the extremal index θ . This is an important measure in time series, namely in assessing risky phenomena, as it measures the propensity for the occurrence of clusters of extreme values. The estimation of θ requires a prior setting of tuning parameter values that impacts the precision of estimates. In this work, we presented an algorithm that allows estimation of θ free of tuning parameters. We applied this methodology to diverse models, and the results are encouraging in several cases. In

the future, it is intended to continue the study of this methodology and develop it in order to improve its applicability to different types of models.

Funding: The research at CMAT was partially financed by Portuguese Funds through FCT - Fundação para a Ciência e a Tecnologia within the Projects UIDB/00013/2020 and UIDP/00013/2020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Hsing, T.; Hüsler, J.; Leadbetter, M.R. On the exceedance point process for a stationary sequence. *Probab. Theory Relat. Fields* **1988**, *78*, 97–112. [CrossRef]
2. Northrop, P.J. An efficient semiparametric maxima estimator of the extremal index. *Extremes* **2015**, *18*, 585–603. [CrossRef]
3. Berghaus, B.; Bücher, A. Weak convergence of a pseudo maximum likelihood estimator for the extremal index. *Ann. Stat.* **2018**, *46*, 2307–2335. [CrossRef]
4. Nandagopalan, S. Multivariate Extremes and Estimation of the Extremal Index. Ph.D. Thesis, University of North Carolina, Chapel Hill, NC, USA, 1990.
5. Weissman, I.; Novak, S.Y. On blocks and runs estimators of the extremal index. *J. Stat. Plan. Inference* **1998**, *66*, 281–288. [CrossRef]
6. Süveges, M.; Davison, A.C. Model misspecification in peaks over threshold analysis. *Ann. Appl. Stat.* **2010**, *4*, 203–221. [CrossRef]
7. Holěšovský, J.; Fusek, M. Estimation of the extremal index using censored distributions. *Extremes* **2020**, *23*, 197–213. [CrossRef]
8. Holěšovský, J.; Fusek, M. Improved interexceedance-times-based estimator of the extremal index using truncated distribution. *Extremes* **2022**, *25*, 695–720. [CrossRef]
9. Ferreira, H.; Ferreira, M. Estimating the extremal index through local dependence. *Ann. L'Institut Henri-Poincaré-Probab. Stat.* **2018**, *54*, 587–605. [CrossRef]
10. Ferro, C.A.T.; Segers, J. Inference for clusters of extreme values. *J. R. Stat. Soc. Ser. B* **2003**, *65*, 545–556. [CrossRef]
11. Gomes, M. On the estimation of parameters of rare events in environmental time series. In *Statistics for the Environment 2: Water Related Issues*; Barnett, V., Turkman, K., Eds.; Wiley: Hoboken, NJ, USA, 1993; pp. 225–241.
12. Ancona-Navarrete, M.A.; Tawn, J.A. A comparison of methods for estimating the extremal index. *Extremes* **2000**, *3*, 5–38. [CrossRef]
13. Ferreira, H.; Ferreira, M. A new blocks estimator for the extremal index. *Commun.-Stat.-Theory Methods* **2022**, *in press*. [CrossRef]
14. Frahm, G.; Junker, M.; Schmidt, R. Estimating the tail-dependence coefficient: Properties and pitfalls. *Insur. Math. Econ.* **2005**, *37*, 80–100. [CrossRef]
15. Gomes, D.P.; Neves, M.M. Extremal index blocks estimator: The threshold and the block size choice. *J. Appl. Stat.* **2020**, *47*, 2846–2861. [CrossRef] [PubMed]
16. Ferreira, M. Heuristic Tools for the Estimation of The Extremal Index: A Comparison of Methods. *Revstat-Stat. J.* **2018**, *16*, 115–136.
17. Northrop, P.J.; Christodoulides, C. Exdex: Estimation of the Extremal Index. R Package Version 1.0.1. 2019. Available online: <https://CRAN.R-project.org/package=exdex> (accessed on 10 January 2023).
18. Chernick, M.R.; Hsing, T.; McCormick, W.P. Calculating the extremal index for a class of stationary sequences. *Adv. Appl. Probab.* **1991**, *23*, 835–850. [CrossRef]
19. Newell, G.F. Asymptotic Extremes for m -Dependent Random Variables. *Ann. Math. Stat.* **1964**, *35*, 1322–1325. [CrossRef]
20. Weissman, I.; Cohen, U. The extremal index and clustering of high values for derived stationary sequences. *J. Appl. Prob.* **1995**, *32*, 972–981. [CrossRef]
21. Cai, J.J. Statistical inference on $D^{(d)}(u_n)$ condition and estimation of the Extremal Index. *arXiv* **2019**, arXiv:1911.06674.
22. Davis, R.; Resnick, S. Basic properties and prediction of max-ARMA processes. *Adv. Appl. Probab.* **1989**, *21*, 781–803. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Automata Based Multivariate Time Series Analysis for Anomaly Detection over Sliding Time Windows [†]

Arnold Hien ^{1,*}, Nicolas Beldiceanu ^{1,*}, Claude-Guy Quimper ² and María-I. Restrepo ¹

¹ Department of Automation, Production and Computer Sciences, IMT Atlantique, 44300 Nantes, France; maria-isabel.restrepo-ruiiz@imt-atlantique.fr

² Computer Science Department, Laval University, Quebec City, QC G1V 0A6, Canada; claude-guy.quimper@ift.ulaval.ca

* Correspondence: arnold.hien@imt-atlantique.fr (A.H.); nicolas.beldiceanu@imt-atlantique.fr (N.B.)

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: We describe an optimal linear time complexity method for extracting patterns from sliding windows of multivariate time series that depends only on the length of the time series. The method is implemented as an open-source Java library and is used to detect anomalies in multivariate time series.

Keywords: multivariate time series; transducers; sliding windows; anomaly detection

1. Introduction

Multivariate time series [1,2] are sequences or streams of more than one time-dependent variable corresponding to the simultaneous evolution of several variables over time. They can be observed in many areas and can thus be used to describe the evolution of key indicators.

Context. The analysis of time series makes it possible to extract certain behaviours that can be described by patterns [3]. These patterns inform us about the evolution of variables and provide trends observed in the time series. Patterns describing abnormal situations can be captured by regular expressions. The analysis of the time series consists of first identifying pattern occurrences in the time series, then associating a numerical value with each occurrence through the computation of a *feature value*. Anomaly detection then performs according to the following steps:

- Symbolically describe abnormal behaviours through patterns;
- Find the occurrences of these patterns in the time series;
- Identify the occurrences of those patterns whose numerical characteristics are deviant.

To identify these patterns, Beldiceanu et al. [3,4] used transducers, i.e., finite-state automata producing an output, which made it possible to efficiently identify pattern occurrences and calculate the corresponding feature value. This work and that of Arafailova [5] laid the necessary foundations for the development of our tool for detecting anomalies in time series.

Question addressed by this paper. The challenge is to design an efficient algorithm capable of identifying a succession of pattern occurrences denoting anomalies within the sliding time windows of a multivariate time series, where the patterns are described generically.

Our contribution. Given a multivariate time series with measurements over n instants and all sliding time windows over m consecutive instants, we describe an optimal time complexity algorithm in $\Theta(n)$ to identify all time windows containing occurrences of patterns corresponding to anomalies. A parameterised version [6] of this algorithm handling a variety of patterns was implemented as a Java library.

Citation: Hien, A.; Beldiceanu, N.; Quimper, C.-G.; Restrepo, M.-I. Automata Based Multivariate Time Series Analysis for Anomaly Detection over Sliding Time Windows. *Eng. Proc.* **2023**, *39*, 65. <https://doi.org/10.3390/engproc2023039065>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 6 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Paper organisation. In Section 2, we present the required background, such as patterns, features, and transducers. Then, in Section 3, we define the extraction of patterns occurrences on sliding windows; we present how patterns are evaluated both qualitatively and quantitatively using regular expressions and features. In Section 4, we present our anomaly detection tool and illustrate its use in Section 4.2 on environmental sensor data [7].

2. Background on Multivariate Time Series

A multivariate time series is obtained by observing the evolution of d measures over regular periods [8]. It is denoted as a n -dimensional array $\mathcal{X} = \langle X_1, X_2, \dots, X_n \rangle$, where n is the length of the time series, d is the number of measures, $X_i \in \mathbb{R}^d$ is the i -th vector of measures, and X_i^j is the j -th component of vector X_i . As a stream is unbounded, searching anomalies on a full stream does not make sense as data is generated continuously and sent in multiple data records; we rather want to identify anomalies on sliding windows of the stream [9]. Each window is a subsequence denoted by $X_{i,j}$ (with $i < j$) whose measures are defined from instant i to instant j . The next section shows how to describe conditions between two consecutive measures of a multivariate time series.

2.1. Alphabet as a Mean to Describe Conditions between Adjacent Measures

To specify patterns on a multivariate time series, the first step is to describe the basic elements of a pattern, namely a finite set of conditions between p consecutive measures of the time series. Each condition is interpreted as the letter of the alphabet Σ that we now introduce.

Definition 1 (alphabet). *Given p consecutive measures $X_i, X_{i+1}, \dots, X_{i+p-1}$, an alphabet Σ is defined as a set of mutually exclusive conditions $\{C_1, C_2, \dots, C_k\}$ such that $C_1 \vee C_2 \vee \dots \vee C_k$ is true, where each condition C_ℓ (with $\ell \in [1, k]$) compares the components of $X_i, X_{i+1}, \dots, X_{i+p-1}$ using the operators $<, =, \text{ or } >$. Each condition C_ℓ of Σ must have its mirror condition C_ℓ^{mir} in Σ , where C_ℓ^{mir} is obtained by flipping the comparison operators $<$ and $>$ in C_ℓ . Each of the conditions C_1, C_2, \dots, C_k will be called a symbolic letter [10].*

2.2. Signature of the Multivariate Time Series

The first step to analyse a multivariate time series \mathcal{X} is to generate the sequence \mathcal{S} of symbolic letters S_i (with $i \in [1, n - p + 1]$) associated with p consecutive measures of \mathcal{X} . This leads to the notion of *signature* \mathcal{S} .

Definition 2 (Signature, arity). *Consider a sequence of n measures \mathcal{X} and a function $\mathcal{F} : \mathbb{R}^p \rightarrow \Sigma$, where Σ is a finite set denoting an alphabet. Then, the signature of \mathcal{X} is a sequence of symbolic letters $\mathcal{S} = \langle S_1, S_2, \dots, S_{n-p+1} \rangle$ where each S_i equals $\mathcal{F}(X_i, \dots, X_{i+p-1})$.*

The alphabet Σ is used to define regular expressions to symbolically characterise the occurrences of anomalies in \mathcal{S} . For this, we use patterns and features.

2.3. Pattern and Feature as Qualitative and Quantitative Aspects of Anomalies

The qualitative aspect of anomalies is described as the words of the language \mathcal{L}_σ associated with the regular expression σ defined over the alphabet Σ [11].

Definition 3 (Patterns [3]). *A pattern σ over the alphabet Σ is a triple $\langle \text{reg}, b, a \rangle$, where reg is a regular expression over Σ that is only matched by non-empty words, while b and a are two non-negative integers, whose role is to delete parts of the pattern that are used to detect the start and end of a pattern.*

Definition 4 (Pattern reverse [4]). *Two patterns $\sigma = \langle \text{reg}, b, a \rangle$ and $\sigma^r = \langle \text{reg}^r, b^r, a^r \rangle$ are the reverse of each other if $w_1 w_2 \dots w_k \in \mathcal{L}_\sigma \Leftrightarrow w_k^{mir} w_{k-1}^{mir} \dots w_1^{mir} \in \mathcal{L}_{\sigma^r}$, $a = b^r$, $b = a^r$.*

A list of 22 patterns can be found in [4,12].

Features. After identifying a pattern occurrence in a time series, it is possible to characterise it with a numerical value. For this, we use *features*, which are functions allowing us to compute certain characteristics of a pattern occurrence, such as the min/max value. In [4], Beldiceanu et al. used five features for the quantitative evaluation of patterns in the context of sliding windows: ONE, WIDTH, SURFACE, MIN, and MAX.

Aggregators. Sometimes, several occurrences of a pattern are identified in a sliding window. To obtain a unique result for the whole window, we use *aggregators*, which are functions that aggregate the features values on the different occurrences of the pattern. In [3,4], three aggregation functions are proposed: MIN, MAX, and SUM. In this paper, we only use the SUMaggregator. To identify patterns occurrences in a time series, we use transducers.

2.4. Seed Transducers

Identifying pattern occurrences is achieved by using *seed transducers* [3]. We use deterministic finite transducers [13,14], which are automata \mathcal{M} that generate an output sequence over the alphabet Σ' from an input sequence over the alphabet Σ . To identify the occurrences of a pattern σ , our transducer reads one by one the symbolic letters S_i in Σ and triggers a transition from state q_{i-1} to q_i to produce a *semantic letter* τ_i in Σ' associated with S_i . Each semantic letter designates a phase in the recognition of an occurrence of the pattern, e.g., when an occurrence of σ is found, the semantic letter FOUND is generated. The semantic letter MAYBE_b means that the transducer has found the first letters of a potential occurrence of σ but needs to read more letters to confirm it. The output alphabet $\Sigma' = \{\text{OUT}, \text{MAYBE}_b, \text{OUT}_r, \text{FOUND}, \text{FOUND}_e, \text{IN}, \text{MAYBE}_a, \text{OUT}_a\}$ of a seed transducer is called the *semantic alphabet*. More details about their meaning can be found in [3].

Example 1. Let us consider a temperature and humidity measuring device that allows one measurement every hour. Our multivariate time series \mathcal{X} is given in Table 1. Assume we want to identify the situation where, for two consecutive measures, i.e., $p = 2$, both the temperature and the humidity increase. For this purpose, we define the alphabet $\Sigma = \{<, \leq, =, \geq, >\}$ as:

$$\left\{ \begin{array}{l} <: \text{ if } X_i^1 < X_{i+1}^1 \wedge X_i^2 < X_{i+1}^2 \\ \leq: \text{ if } (X_i^1 < X_{i+1}^1 \wedge X_i^2 = X_{i+1}^2) \vee (X_i^1 = X_{i+1}^1 \wedge X_i^2 < X_{i+1}^2) \\ =: \text{ if } X_i^1 = X_{i+1}^1 \wedge X_i^2 = X_{i+1}^2 \\ \geq: \text{ if } (X_i^1 > X_{i+1}^1 \wedge X_i^2 = X_{i+1}^2) \vee (X_i^1 = X_{i+1}^1 \wedge X_i^2 > X_{i+1}^2) \\ >: \text{ if } X_i^1 > X_{i+1}^1 \wedge X_i^2 > X_{i+1}^2 \\ \geq: \text{ if } (X_i^1 > X_{i+1}^1 \wedge X_i^2 < X_{i+1}^2) \vee (X_i^1 < X_{i+1}^1 \wedge X_i^2 > X_{i+1}^2) \end{array} \right.$$

We then define two patterns using the following observation. Normally, when the temperature increases, the humidity decreases and vice versa. Thus, when both metrics change in the same way (increasing or decreasing), it may be a sign of an anomaly. These problematic changes are captured by the patterns σ_{\searrow} and σ_{\nearrow} , respectively, corresponding to $> | > (> | \geq)^* >$ and $< | < (< | \leq)^* <$, where σ_{\searrow} describes a simultaneous decrease in both temperature and humidity, and σ_{\nearrow} an increase. Figure 1A shows two maximal occurrences of σ_{\nearrow} in the multivariate time series \mathcal{X} . Using the WIDTH feature, we obtain $f_1 = 2$ and $f_2 = 4$ as the lengths of the two occurrences. Using the SUMaggregator, we obtain a total length $g = 6$. These values are computed using the transducer given in Figure 1B, which describes the transitions from the initial state s .

Table 1. Multivariate time series \mathcal{X} : temperature and humidity level evolution over 17 h.

Time	1 am	2 am	3 am	4 am	5 am	6 am	7 am	8 am	9 am	10 am	11 am	12 pm	1 pm	2 pm	3 pm	4 pm	5 pm
Temp. (C)	19.3	21.5	19.2	21.4	23.6	22.8	22.8	20.1	20.9	21.5	22.7	23.6	23.6	19.2	21.5	21.5	21.5
Hum. (%)	74.9	52.2	74.8	52.1	73.2	72.3	65.7	55.9	52.1	64.5	64.5	72.7	62.4	59.8	52.1	55.2	55.2

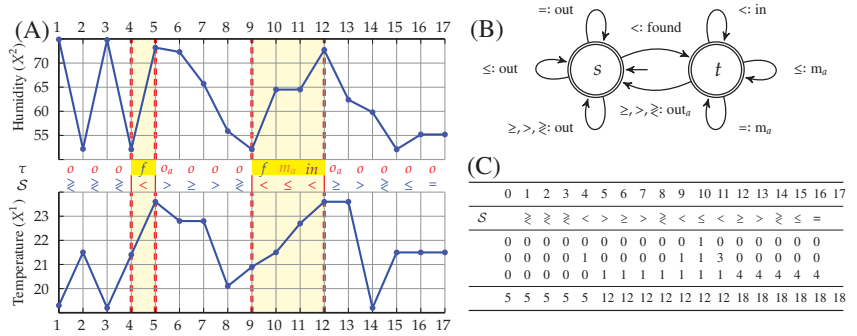


Figure 1. (A) Occurrences of pattern σ_f in a multivariate time series, (B) Transducer of pattern σ_f , (C) Accumulators updates.

3. Optimal Patterns Extraction from Sliding Windows

As explained in Section 2, the analysis of time series makes it possible to characterise them qualitatively with patterns, and quantitatively with features. The sum of the feature values of all pattern occurrences in a time series is called its *contribution*. We describe an optimal time-complexity algorithm for computing such contribution. This algorithm is used both when a multivariate time series corresponds to a single finite sequence of timed data, or when we have a data stream consisting of successive subsequences of timed data. Without loss of generality, we focus on a single finite sequence and show how to generalise it to a stream at the end of this section.

3.1. Register-Based Features Evaluation on a Time Series

Consider a multivariate time series $\mathcal{X} = \langle X_1, X_2, \dots, X_n \rangle$, a pattern σ and a feature f . To obtain the contribution of σ on \mathcal{X} we associate three accumulators R, C and D to the transducer \mathcal{M} of σ . We obtain a register automaton [3] in which each accumulator is updated as \mathcal{X} is read:

- R gradually records the sum of the feature values of f on each completely terminated found occurrence of σ (i.e., $\tau_i \in \{\text{OUT}_a, \text{FOUND}_e\}$);
- C stores the feature value of the current occurrence for which we did not yet reach the end (i.e., $\tau_i \in \{\text{FOUND}, \text{IN}\}$);
- D contains the feature value of the current potential part of an occurrence ($\tau_i \in \{\text{MAYBE}_b, \text{MAYBE}_a\}$).

Accumulators R, C, and D are updated according to the semantic letter τ_i returned by \mathcal{M} . Details of this evaluation can be found in [3,12].

Example 2 (Continuation of Example 1). Reading $S_9 = '<'$ leads to $\tau_9 = \text{FOUND}$. As shown in Table C of Figure 1, we then compute $C \leftarrow D + 1$ (i.e., $C \leftarrow 1$), meaning that the length of the current occurrence of σ_f is 1. Similarly, $\tau_{10} = m_a$ means that we obtain a potential extra part of the already found occurrence of σ_f . We then compute its length with $D \leftarrow D + 1$. $\tau_{11} = \text{in}$ means that we are still inside an occurrence of σ_f . It then confirms the membership of the encountered extra parts. Thus, we compute $C \leftarrow C + D + 1$. Finally, $\tau_{12} = o_a$ means that we are no longer in an occurrence of σ_f . We then compute $R \leftarrow R + C$ to integrate C in R.

3.2. Register-Based Features Evaluation on Sliding Windows

The contribution of a pattern on a sliding window $X_{i,j} = X_i, X_{i+1}, \dots, X_j$ [15,16] is computed using Equation (1).

$$f_\sigma(X_{i,j}) = \begin{cases} 0 & \text{if there is no occurrence of } \sigma \\ f_\sigma(X_{1,j}) + f_\sigma(X_{n,i}) - f_\sigma(X_{1,n}) & \text{otherwise.} \end{cases} \quad (1)$$

Computing $f_\sigma(X_{i,j})$ involves different steps. The first step consists of checking the presence of an occurrence of σ in $X_{i,j}$, and the second step computes $f_\sigma(X_{1,j})$, $f_{\sigma^r}(X_{n,i})$, and $f_\sigma(X_{1,n})$. In this section, we first show how to compute the contribution of σ on $X_{1,j}$, then describe our method of identifying occurrences of σ on sliding windows.

Computing the Contribution of σ on a Sliding Window

In Equation (1), $f_\sigma(X_{1,n})$ corresponds to the final value of R after reading \mathcal{X} and $f_\sigma(X_{1,j})$ to its value after reading the subsequence $X_{1,j}$. Similarly, $f_{\sigma^r}(X_{n,i})$ corresponds to the value of R after reading the reverse sequence $X_{n,i}^r$ using the transducer of σ^r . To compute $f_\sigma(X_{i,j})$, we first have to know the values of R, C, and D associated with each semantic letter returned. A first step is, therefore, performed to acquire the needed values exploited to optimally compute $f_\sigma(X_{i,j})$.

Pattern Occurrences Checker in Slidings Windows

To obtain an optimal time complexity algorithm, we also need to check whether each sliding window contains at least one pattern occurrence, i.e., see the first case of Equation (1). A naïve approach would be to check whether there is an occurrence of σ in each window independently. Thus, considering a window size of m , the occurrence check of σ on all sliding windows would lead to a time complexity of $O(m \cdot n)$ [4].

To obtain an optimal time complexity of $\Theta(n)$, we create a new array, denoted as E, which provides for each position in the time series, the *end of the next occurrence of pattern in \mathcal{X}* . Indeed, if there is an occurrence of σ in $X_{i,j}$, then this occurrence will be defined between positions u and v , with $i \leq u \leq v \leq j$. The accumulator E will indicate that an occurrence of σ ends at v . Similarly, given that σ^r and \mathcal{X}^r are, respectively, the reverse of σ and \mathcal{X} , then the end of an occurrence of σ^r in \mathcal{X}^r matches the start of an occurrence of σ in \mathcal{X} [4]. This makes it possible to say that an occurrence of σ begins at u . The new accumulator E records at position k the end of the next occurrence of σ from X_k . Table C of Figure 1 gives the values of E indicating the end of the next occurrences of $\sigma_{\mathcal{X}}$ in the multivariate time series \mathcal{X} of Example 1.

Computing the End of the Next Pattern Occurrence from the Pattern Transducer

Depending on the presence of FOUND or FOUND _{ϵ} in the transducer \mathcal{M} , two cases must be distinguished:

- When FOUND _{ϵ} $\in \mathcal{M}$, E is updated according to lines 3–9 of Algorithm 1;
- When FOUND $\in \mathcal{M}$, E is updated according to lines 10–20 of Algorithm 1.

In Algorithm 1, we use two types of assignments: *value assignment*, denoted ' \leftarrow ', and *variable linkage*, denoted '='. For the first one, a value is directly assigned to a variable. For the second one, two variables are made equal using a linked list; when one of these variables is assigned, this assignment is automatically propagated to all the linked variables.

Linking two consecutive subsequences of a data stream. To find a pattern occurrence located across consecutive subsequences of a data stream, we use a buffer that records the last $m - 1$ measures. Each new received sequence $\langle X_1, X_2, \dots, X_k \rangle$ then integrates these past measurements as follows: $\mathcal{X} = \langle X_{-m+1}, X_{-m+2}, \dots, X_0, X_1, X_2, \dots, X_k \rangle$.

Algorithm 1: Computing the end of the next occurrence of pattern for each position.

```

1  Input  $S[1..n-1]$ : time series signature;  $\sigma$ : pattern;  $\mathcal{M}$ : transducer of  $\sigma$ ; Output  $E[0..n]$ : next pattern occurrence
   end;
2  begin
3    If  $\text{FOUND}_e \in \mathcal{M}$  then
4      state  $\leftarrow$  init_state;  $E[0] \leftarrow 0$ ;
5      For each  $k \in 1, \dots, n-1$  do
6         $\tau \leftarrow \mathcal{M}(\sigma, \text{state}, S[k])$ ;
7        If  $\tau \in \{\text{OUT}, \text{OUT}_r, \text{MAYBE}_b\}$  then  $E[k] = E[k+1]$ ;
8        else if  $\tau = \text{FOUND}_e$  then  $E[k] \leftarrow k+1$ ;
9       $E[n] \leftarrow n+1$ ; return E;
10   else
11      $I[0..n]$ : accumulator array;  $MA[0..n]$ : accumulator array;
12     state  $\leftarrow$  init_state;  $I[0] \leftarrow 0$ ;  $I[n] \leftarrow 0$ ;  $MA[0] \leftarrow 0$ ;  $MA[n] \leftarrow n+1$ ;  $MA[n-1] = E[n-1]$ ;
13     For each  $k \in 1, \dots, n-1$  do
14        $\tau \leftarrow \mathcal{M}(\sigma, \text{state}, S[k])$ ;
15       If  $\tau \in \{\text{OUT}, \text{OUT}_r, \text{MAYBE}_b\}$  then  $I[k] \leftarrow 0$ ;  $MA[k] \leftarrow 0$ ;  $E[k-1] = E[k]$ ;
16       else if  $\tau = \text{FOUND}$  then  $E[k-1] = E[k]$ ;  $E[k] = MA[k]$ ;  $I[k] \leftarrow 1$ ;
17       else if  $\tau = \text{IN}$  then  $E[k-1] = E[k]$ ;  $E[k] = MA[k]$ ;  $MA[k-1] = MA[k]$ ;  $I[k] \leftarrow 1$ ;
18       else if  $\tau = \text{MAYBE}_a$  then  $E[k] = E[k+1]$ ;  $MA[k-1] = MA[k]$ ;  $I[k] \leftarrow I[k-1] + 1$ ;
19       else if  $\tau = \text{OUT}_a$  then  $MA[k-1] \leftarrow k+1 - I[k-1]$ ;  $MA[k] \leftarrow 0$ ;  $I[k] \leftarrow 0$ ;
20      $E[n] \leftarrow n+1$ ;  $E[n-1] \leftarrow n+1 - I[n-1]$ ; return E;

```

4. Anomaly Detection Tool

In this section, we describe an anomaly detection tool that exploits the efficient evaluation of patterns contributions on sliding windows. First, we give the key parameters of the tool. Then we present some experiments carried out.

4.1. Parameters

Anomaly detection is used to identify suspicious behaviour as data evolve. We use three parameters, namely: (i) the pattern σ we are looking for, (ii) the feature f we consider, and (iii) the window size m . Anomalies occur when there are unusual values and when the sum of them exceeds a given threshold. We add two parameters to adjust the sensitivity of our tool to small variations in consecutive measures, and to multiple occurrences of unusual values:

- The *minimum difference threshold* δ_X is used to determine the minimum variation for two consecutive measures to be considered as different.
- The *occupation percentage threshold* ϵ is the minimum percentage of the window occupation by the pattern wrt its contribution within the window. Thus, an anomaly is detected when the occupation percentage exceeds ϵ .

4.2. Experiments

We have implemented our anomaly detection tools using Java 17. For the experiments, we analysed data from an environmental sensor [7]. These data show the evolution of temperature and humidity measurements over time, as shown in Figure 2. A visual analysis of Figure 2A highlights the existence of strong variations in the dataset with temperature or humidity, often dropping sharply to 0. A similar phenomenon can be observed with temperature increases of more than three degrees. Figure 2B gives a zoom-in and more detailed view of these variations. Each of these variations are potential anomalies that the tool identifies.

For our analysis, we used combinations of values of the previous parameters of Figure 2C. For all the combinations of values, we followed the following protocol: first, we identify problematic windows; second, we colour them in red and plot them; then we analyse the effects of each parameter variations. For space reasons, we will only show the results of two combinations of parameters, one for each of pattern $\sigma_{\mathcal{F}}$ and $\sigma_{\mathcal{X}}$. The analysis of the results shown in Figure 3 then allows us to conclude that our tool allows one to efficiently identify anomalies occurrences in windows. The addition of parameters δ_X and

ϵ , and the possibility of choosing the pattern to identify makes it possible to characterise the anomalies and to adjust their detection in a better way.

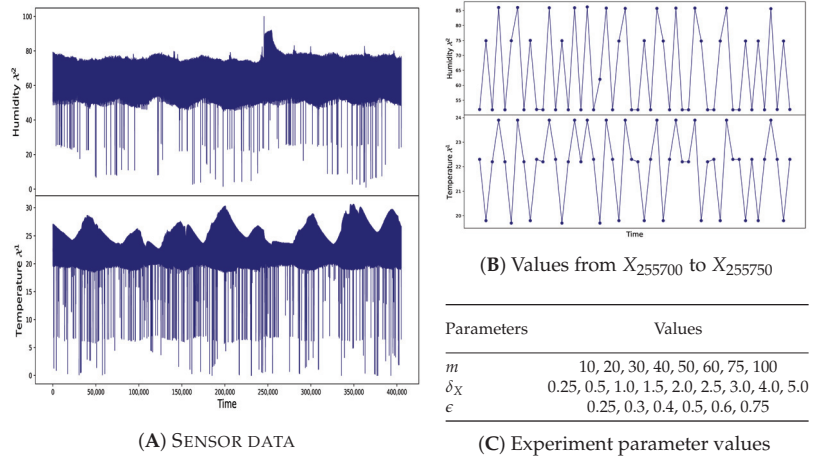


Figure 2. Evolution of the values of the analysed dataset and summary of the values of the parameters used in our experiments.

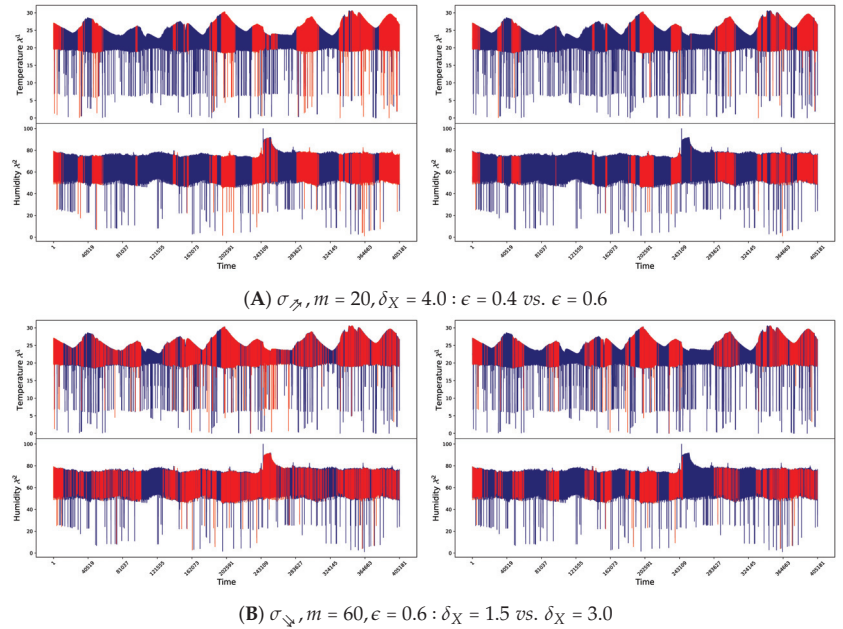


Figure 3. Problematic windows identified when using patterns σ_T and σ_S , and varying the values of δ_X and ϵ . These problematic windows are plotted in red, the non-problematic windows remain in blue.

Effects of δ_X Variation

When analysing the effect of δ_X on the results, we notice that, as expected, small values of δ_X lead to the detection of more problematic windows. Indeed, large values of δ_X make it possible to ignore the small variations in the values of $X_k \in \mathcal{X}$ to consider only the large variations. Therefore, many, probably non-problematic occurrences of patterns are ignored.

Conversely, with small values of δ_X , these occurrences will be considered problematic and lead to more anomalies being detected. This behaviour is maintained whatever the pattern, the dataset or the values of m and ϵ .

Effects of m and ϵ Variation

The analysis of the effects of m and ϵ shows that the bigger m is, the smaller must ϵ be (and vice versa), if we want to catch a maximum number of problematic windows. Indeed, a large window size m may make it unlikely to find a high number of occurrences of σ . Therefore, the values of these two parameters should be adjusted inversely. This behaviour is maintained whatever the pattern, the dataset, or the values of δ_X .

5. Conclusions

In this paper, we have proposed an efficient method for multivariate time series analysis. This transducer-based approach makes it possible to extract occurrences of patterns on sliding windows and to characterise them quantitatively with an optimal time complexity. We used the method for detecting anomalies and obtained a parameterised detection tool. The experiments we conducted show the ability of our approach to efficiently identify inconsistencies in data. In the future, we may consider other uses such as the automatic annotation of multivariate time series or the generation of time series.

Author Contributions: Conceptualization, A.H. and N.B.; methodology, A.H. and N.B.; software, A.H. and N.B.; validation, A.H. and N.B.; formal analysis, A.H. and N.B.; investigation, A.H. and N.B.; resources, A.H. and N.B.; data curation, A.H. and N.B.; writing—original draft preparation, A.H. and N.B.; writing—review and editing, A.H., N.B., C.-G.Q., and M.-I.R.; visualization, A.H., N.B., C.-G.Q., and M.-I.R.; supervision, N.B.; project administration, N.B.; funding acquisition, N.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the EU-funded ASSISTANT project no. 101000165.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs (<https://gitlab.com/postdochien/atisad> (accessed on 5 July 2023)). Publicly available datasets were analyzed in this study. This data can be found here: <https://www.kaggle.com/datasets/garystafford/environmental-sensor-data-132k> (accessed on 5 July 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Audibert, J. Unsupervised Anomaly Detection in Time-Series. (Détection Non Supervisée des Anomalies Dans Les Séries Temporelles). Ph.D Thesis, Sorbonne University, Paris, France, 2021.
2. Fawaz, H.I.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* **2019**, *33*, 917–963. [CrossRef]
3. Beldiceanu, N.; Carlsson, M.; Douence, R.; Simonis, H. Using finite transducers for describing and synthesising structural time-series constraints. *Constraints* **2016**, *21*, 22–40. [CrossRef]
4. Beldiceanu, N.; Carlsson, M.; Quimper, C.; Restrepo-Ruiz, M. Classifying Pattern and Feature Properties to Get a $\Theta(n)$ Checker and Reformulation for Sliding Time-Series Constraints. *CoRR* **2019**, abs/1912.01532. Available online: <https://arxiv.org/abs/1912.01532> (accessed on 5 July 2023).
5. Arafailova, E. Functional Description of Sequence Constraints and Synthesis of Combinatorial Objects. Ph.D. Thesis, IMT Atlantique, Nantes, France, 2018.
6. Hien, A.; Beldiceanu, N.; Quimper, C.; Restrepo-Ruiz, M. Code and Supplementary Material. 2023. Available online: <https://gitlab.com/postdochien/atisad> (accessed on 5 July 2023).
7. Stafford, G. Environmental Sensor Telemetry Data. 2020. Available online: <https://www.kaggle.com/datasets/garystafford/environmental-sensor-data-132k> (accessed on 5 July 2023).
8. Morrill, J.; Fermanian, A.; Kidger, P.; Lyons, T.J. A Generalised Signature Method for Time Series. *CoRR* **2020**, abs/2006.00873. Available online: <https://arxiv.org/abs/2006.00873> (accessed on 5 July 2023).

9. Keogh, E.; Chu, S.; Hart, D.; Pazzani, M. Segmenting Time Series: A Survey and Novel Approach. In *Data Mining in Time Series Databases*; World Scientific: Singapore, 2004; Volume 57, pp. 1–21. [CrossRef]
10. Veanes, M.; Hooimeijer, P.; Livshits, B.; Molnar, D.; Bjørner, N.S. Symbolic finite state transducers: Algorithms and applications. In *Proceedings of the 39th ACM SIGPLAN-SIGACT*, Philadelphia, PA, USA, 25–27 January 2012; pp. 137–150. [CrossRef]
11. Crochemore, M.; Hancart, C.; Lecroq, T. *Algorithms on Strings*; Cambridge University Press: Cambridge, MA, USA, 2007.
12. Arafailova, E.; Beldiceanu, N.; Douence, R.; Carlsson, M.; Flener, P.; Rodríguez, M.A.F.; Pearson, J.; Simonis, H. Global Constraint Catalog, Volume II, Time-Series Constraints. *CoRR* **2016**, abs/1609.08925. Available online: <https://arxiv.org/abs/1609.08925> (accessed on 5 July 2023).
13. Sakarovitch, J. *Elements of Automata Theory*; Cambridge University Press: Cambridge, MA, USA, 2009.
14. Hopcroft, J.E.; Motwani, R.; Ullman, J.D. *Introduction to Automata Theory, Languages, and Computation*, 3rd ed.; Pearson International Edition: London, UK, 2006.
15. Kolev, B.; Akbarinia, R.; Jiménez-Peris, R.; Levchenko, O.; Masegla, F.; Patiño, M.; Valduriez, P. Parallel Streaming Implementation of Online Time Series Correlation Discovery on Sliding Windows with Regression Capabilities. In *Proceedings of the 9th International Conference on Cloud Computing and Services Science*, Heraklion, Crete, Greece, 2–4 May 2019; SciTePress: Setúbal, Portugal, 2019; Volume 1, pp. 681–687.
16. Kontaki, M.; Papadopoulos, A.N.; Manolopoulos, Y. Adaptive similarity search in streaming time series with sliding windows. *Data Knowl. Eng.* **2007**, *63*, pp. 478–502. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Growth Curves Modelling and Its Application [†]

Ana García-Burgos ^{1,*‡}, Beatriz González-Alzaga ², María José Giménez-Asensio ², Marina Lacasaña ²,
Nuria Rico-Castro ^{1,‡} and Desirée Romero-Molina ^{1,‡}

¹ Department of Statistics and Operational Research, University of Granada, 18071 Granada, Spain; nrico@ugr.es (N.R.-C.); deromero@ugr.es (D.R.-M.)

² Andalusian School of Public Health, 18011 Granada, Spain; beatriz.gonzalez.easp@juntadeandalucia.es (B.G.-A.); mariajosesases@hotmail.com (M.J.G.-A.); marina.lacasa.easp@juntadeandalucia.es (M.L.)

* Correspondence: agburgos@ugr.es

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

‡ These authors contributed equally to this work.

Abstract: In this article, we compare two ways of modelling measures of fetal growth. The goal is to impute the missing information for certain ultrasound measurements that are observed at different times and with different numbers of observations. To analyze the effect that other variables have, such as environmental exposure to certain substances or diet, on fetal growth based on these data, we need to handle the information measured at the same instant of time for all the individuals under study, preferably in three time windows of pregnancy (first trimester, week 12; second trimester, week 20; third trimester, week 34). For this, data at these chosen times, in case they are not available, must be imputed from the available information using an appropriate statistical model. One option is to use a linear model, specifically a generalized least squares model that is fitted to the features shown in the data. The other option is to use diffusion processes, estimating their parameters based on the available information. In both options, missing data can be estimated with the unconditional fitted model, conditional on the previous available measurement, or conditional to the closest measurement.

Keywords: growth curves; diffusion processes; linear models

Citation: García-Burgos, A.; González-Alzaga, B.; Giménez-Asensio, M.J.; Lacasaña, M.; Rico-Castro, N.; Romero-Molina, D. Growth Curves Modelling and Its Application. *Eng. Proc.* **2023**, *39*, 66. <https://doi.org/10.3390/engproc2023039066>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 6 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Growth Curves Modelling and Its Application

The aim of this work is to compare different methods for statistical modelling growth curves. We compare two different methodologies in the study of a dataset from the GENEIDA (Genetics, Early life environmental Exposures and Infant Development in Andalusia) <https://www.easp.es/web/geneida/> (accessed on 27 June 2023) project. This project details a cohort born in 2014, made up of 800 mother–child pairs. They are followed up during pregnancy, birth and childhood. One of the objectives of the project is to understand how diet and exposure to environmental substances of the pregnant mothers affect fetal growth. To this end, we have some ultrasound measurements performed during pregnancy, which are as follows:

- Biparietal diameter (BPD): distance in millimetres between both parietal bones of the baby's head.
- Abdominal circumference (AC): distance in millimetres around the abdomen.
- Head circumference (HC): distance in millimetres around the head measured above the eyebrows and ears.
- Femur length (FL): length in millimetres of baby's femur.
- Estimated fetal weight (EFW): we estimate the fetal weight using Hadlock's formula (Hadlock et al. [1]).

The data obtained in this study have the following characteristics: the echographic information is measured at different instants of time and the number of ultrasounds is different for each mother. We can see these two characteristics in Figure 1, since the three individuals represented have seven, four and five ultrasound measurements, respectively, at different gestational ages.

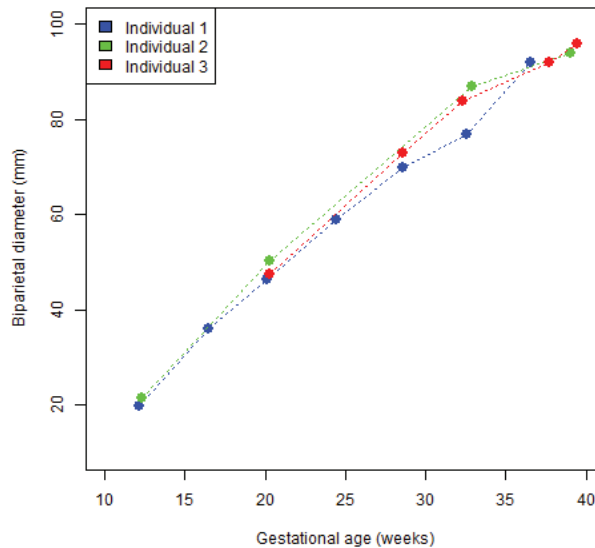


Figure 1. Biparietal diameter of three individuals at different gestational ages.

However, in order to analyze the effect of certain variables on fetal growth based on these data, we need them to be measured at the same instant of time for all individuals. To do this, these data must be imputed based on the available information, preferably using an appropriate statistical model. There are different methodologies to find such a model that fits a set of observations of a quantitative variable that evolves, presenting a growth throughout the time. The curves that model this type of data are called growth curves and their use homogenizes the information from the ultrasound measurements.

In this work, we propose a comparison of two different methodologies in growth modelling to impute the ultrasound measurements at the desired instant of time: linear models and diffusion processes.

Firstly, we approach the problem using linear models. This consists of considering that the observations are a function of a variable, time, and there is a linear relationship that relates them. Following Iñiguez et al. [2], we created models to predict the five fetal measurements at 12, 20 and 34 weeks of gestation. Initially, we tried to use the generalized linear model, but some of the hypotheses failed. In particular, these two factors stood out: heteroskedasticity of the residuals and autocorrelation. Therefore, the generalized least squares model has been used to obtain the predictions, since this model is less restrictive in terms of assumptions (Kariya and Kurata [3]).

Secondly we solve the problem through diffusion processes. In this case, the observed variable $X(t)$ evolves over time t and at each instant there is a probability distribution for $X(t)$ that depends both on time and on the values observed at previous times $X(0), X(1), \dots, X(t-1)$. Diffusion processes are useful for modelling time-dependent variables that increase, usually with an exponential or sigmoidal trend (Baudoin [4]). We choose the type of process based on the characteristics of the observed sample paths. In this case, as the data have a sigmoidal or exponential trend, as we can see in Figure 2, we set the mixed Gompertz–lognormal process. In particular cases, it includes the lognormal

process, associated with an exponential curve, and the Gompertz-type process, related to a sigmoidal curve (Romero et al. [5]).

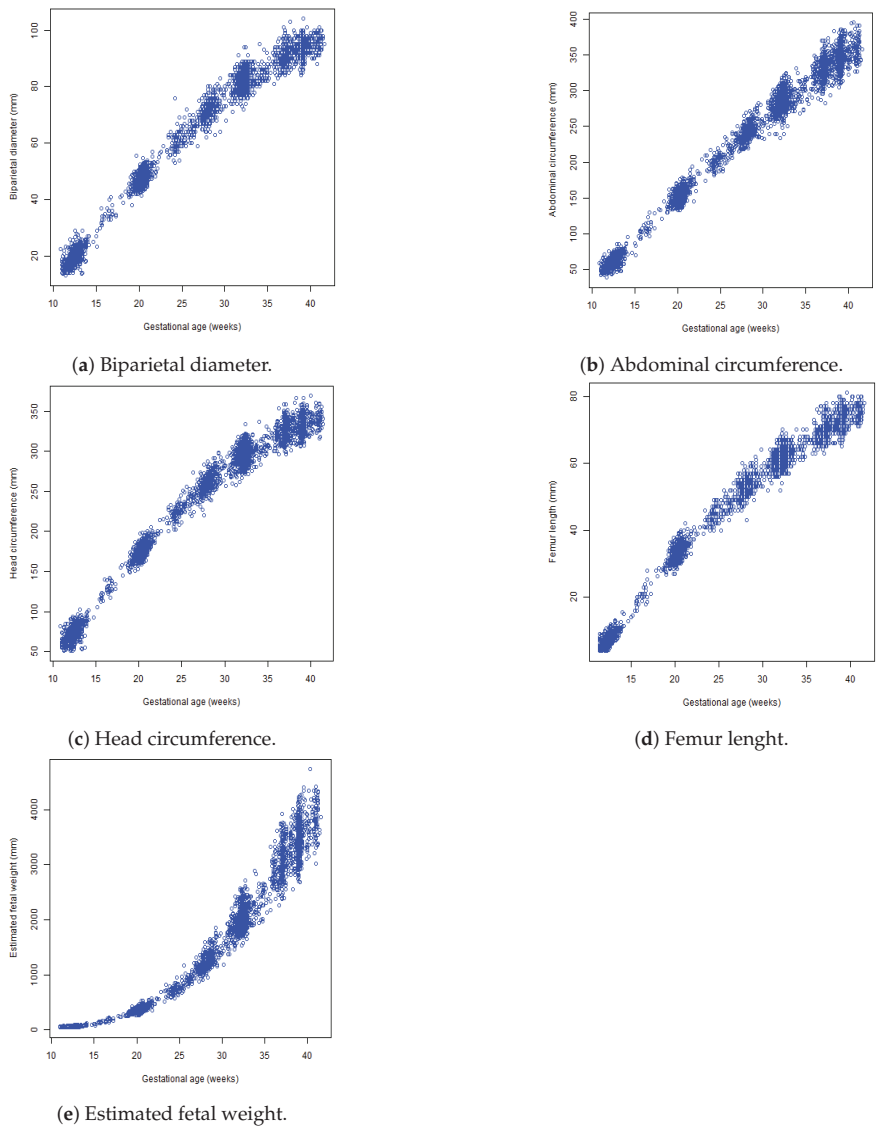


Figure 2. Observed sample paths.

Before fitting the growth curves, we carried out an exhaustive clean up of the data by studying the outliers and eliminating defective data. In the case of linear models, we fit a model for each measure with their respective confounders to later review the influential data and recalculate the model. In the case of diffusion processes, we performed a weighted cluster analysis to group the data depending on the result of the analysis and we fit a process for each cluster in each measurement. Finally, we obtained the data of the measurement in the desired gestational age using an unconditional model, one conditioned to the previous data and another conditioned to the closest available data.

After making the adjustments using linear models and diffusion processes, we compared the results of the two methodologies to find out which best imputes the data. To do this, we used different measures to study the error made in the data imputations (Shcherbakov et al. [6]).

Author Contributions: Conceptualization, A.G.-B., N.R.-C. and D.R.-M.; methodology, A.G.-B., N.R.-C. and D.R.-M.; software, A.G.-B., N.R.-C. and D.R.-M.; validation, N.R.-C. and D.R.-M.; formal analysis, A.G.-B., N.R.-C. and D.R.-M.; investigation, A.G.-B., B.G.-A., M.L., N.R.-C. and D.R.-M.; resources, B.G.-A., M.J.G.-A. and M.L.; data curation, B.G.-A., M.J.G.-A. and M.L.; writing—original draft preparation, A.G.-B., N.R.-C. and D.R.-M.; writing—review and editing, A.G.-B., N.R.-C. and D.R.-M.; visualization, A.G.-B.; supervision, N.R.-C. and D.R.-M.; project administration, M.L.; funding acquisition, M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and the protocol was approved by the Hospital Ethics Committee, the Ethics Committee of “Consejería de Salud y Familias, Junta de Andalucía” (PI-0405-2014) and “Consejería de Igualdad, Salud y Políticas Sociales, Junta de Andalucía” (PI13/01559). We follow the standards described in Andalusian and Spanish laws of personal data protection and biomedical research for the treatment of information and biological samples of human origin.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data available on request due to restrictions eg privacy or ethical. The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the informed consent, obtained from all participants or the legally responsible before they participated in the study, does not establish the transfer to third parties or make it public.

Acknowledgments: FQM147-Análisis estadístico de datos multivariantes y procesos estocásticos.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GENEIDA	Genetics, Early life environmental Exposures and Infant Development in Andalusia
BPD	Biparietal Diameter
AC	Abdominal circumference
HC	Head circumference
FL	Femur length
EFW	Estimated fetal weight

References

- Hadlock, F.P.; Harrist, R.B.; Sharman, R.S.; Deter, R.L.; Park, S.K. Estimation of fetal weight with the use of head, body, and femur measurements—A prospective study. *Am. J. Obstet. Gynecol.* **1985**, *151*, 333–337. [CrossRef] [PubMed]
- Iñiguez, C.; Esplugues, A.; Sunyer, J.; Basterrechea, M.; Fernández-Somoano, A.; Costa, O.; Estarlich, M.; Aguilera, I.; Lertxundi, A.; Tardón, A.; et al. INMA Project.: Prenatal Exposure to NO₂ and Ultrasound Measures of Fetal Growth in the Spanish INMA Cohort. *Environ. Health Perspect.* **2016**, *124*, 235–242. [CrossRef] [PubMed]
- Kariya, T.; Kurata, H. *Generalized Least Squares*; Wiley Series in Probability and Statistics; John Wiley & Sons: Hoboken, NJ, USA, 2004.
- Baudoin, F. *Diffusion Processes and Stochastic Calculus*; European Mathematical Society / American Mathematical Society: West Lafayette, IN, USA, 2014.
- Romero, D.; Rico, N.; Arenas, M. A new diffusion process to epidemic data. *Lect. Notes Comput. Sci.* **2013**, *8111*, 69–76.
- Shcherbakov, M.V.; Brebels, A.; Shcherbakova, N.L.; Tyukov, A.; Janovsky, T.; Kamaev, V.A. A Survey of Forecast Error Measures. *World Appl. Sci. J.* **2013**, *24*, 171–176.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Slope Entropy Characterisation: Adding Another Interval Parameter to the Original Method [†]

Mahdy Kouka and David Cuesta-Frau *

Department of System Informatics and Computers, Universitat Politècnica de València, 03801 Alcoy, Spain; mahdykouka@gmail.com

* Correspondence: dcuesta@disca.upv.es

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Slope Entropy (SlpEn) is a recently proposed time series entropy estimation method for classification. This method has yielded better results than other similar methods in all the published studies so far. It is based on a signal-gradient thresholding scheme using two parameters, δ and γ , in addition to the usual embedded dimension parameter m . In this work, we investigated the possibility of adding one thresholding parameter more, termed θ , and we compared the original method to the new one. The experiment results showed a small improvement using the new method in terms of classification accuracy. However, the temporal cost increased significantly and therefore we concluded it is not worth the extra effort unless maximum accuracy is of utmost importance.

Keywords: slope entropy; time series classification; parameter optimisation

1. Introduction

Entropy estimation methods are very popular among scientists for extracting part of the possible hidden information present in a time series. These methods calculate the relative frequency of a set of numerical or symbolic subsequences. Many scientific fields have benefited from the high segmentation power of these methods. For example, they have been widely used in biomedicine to classify electroencephalograms, time series of electrocardiogram-RR, body temperature, and actigraph records, among many others. Each of the current entropy calculation methods has its strengths and weaknesses.

In this work, we investigated the effect of adding more gradient quantisation intervals to the recently proposed Slope Entropy (SlpEn) method on signal classification accuracy [1,2]. This method is based on assigning symbols to intervals of slopes between consecutive samples of time series [2].

In the general method, the δ and γ thresholds are responsible for labelling a slope (difference between two time series consecutive samples) as low, high, or flat (tie). If it is below δ , it is classified as tie. If it is between δ and γ , the slope is considered low. Otherwise, it is high.

The analysis was carried out as a comparative study. Many datasets with different signal types were employed to understand the impact of using a new additional gradient parameter in SlpEn. A grid search assessed the behaviour of all the datasets with different values of the input parameters to optimise them, see $\delta < \gamma$ and $\gamma < \theta$ in the new SlpEn variation.

The results obtained confirmed that adding a new parameter resulted in a small improvement in the classification accuracy. Specifically, the highest increment achieved using the new variation was 3% higher, at most. However, the execution time was a lot longer than for the original SlpEn method due to the nested resulting additional combinations of δ , γ , and θ values in the grid search.

Citation: Kouka, M.; Cuesta-Frau, D. Slope Entropy Characterisation: Adding Another Interval Parameter to the Original Method. *Eng. Proc.* **2023**, *39*, 67. <https://doi.org/10.3390/engproc2023039067>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 7 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

The structure of the paper is as follows. In Section 2, we present the datasets used in the experiments, a review of SlpEn, the proposed variation method, and the classification process. In Section 3, we report all the results. In Section 4, we provide an interpretation and analysis of all the results. Finally, we summarise our conclusions in the last section.

2. Methods

2.1. Datasets

The experimental dataset comprises several types of time series with different characteristics in terms of bandwidth, length, and regularity. All of them are publicly available, and many of the databases from which they have been extracted have already been used in similar works, serving as a reference for result comparison. The datasets are (two classes are used from each one):

- The Bern–Barcelona database [3]: A set of electroencephalographic records.
- The Fantasia database [4]: A set of electrocardiographic records of R-R intervals.
- The Ford A dataset [5]: A set of records obtained from industrial processes.
- The House Twenty dataset [6]: A set of records obtained from the electricity consumption of 20 households in the UK.
- The PAF prediction dataset [7]: A set of electrocardiographic records of R-R intervals.
- The Worms two class dataset [8,9]: A set of records obtained from the movement of genetically modified worms.
- The Bonn EEG dataset [10]: A set of electroencephalographic records.

2.2. SlpEn

SlpEn applies the general expression of Shannon entropy to the estimated probabilities of a set of symbols. These symbols are assigned based on a range of differences between consecutive samples of subsequences extracted from a time series, $X = \{x_0, x_1, x_2, \dots, x_{N-1}\}$. These symbols are generically obtained from $x_i - x_{i-1}$, with the thresholds defined by the two parameters mentioned above: δ and γ [2]. Typically, δ is assigned a value of 0.001.

In the standard method, symbols +2, +1, 0, -1, and -2 are assigned according to the range in which the differences are located. This process is graphically represented in Figure 1.

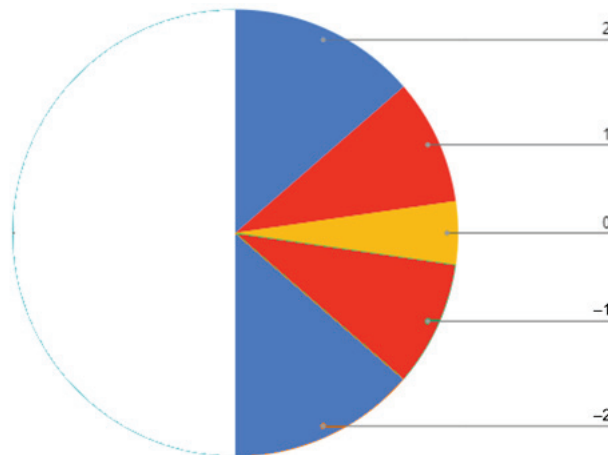


Figure 1. Graphical representation of the calculation of symbols used in SlpEn based on the thresholds γ and δ .

For each subsequence of length m , the corresponding symbol string is generated, and a histogram is constructed with the number of occurrences of each pattern. Finally, Shannon entropy is calculated on this histogram, as previously discussed.

2.3. Modified SlpEn Using an Additional Gradient Interval

In the original method, symbols are assigned based on the difference between two consecutive values. If the value $x_i - x_{i-1} < \delta$, the symbol 0 is assigned and the slope is considered a tie. If the value $x_i - x_{i-1} > \delta$ and $x_i - x_{i-1} > -\gamma$ or $x_i - x_{i-1} < -\delta$ and $x_i - x_{i-1} < \gamma$, the symbol 1 or -1 is assigned, and the slope is considered low. The last symbols assigned are 2 and -2, respectively, when the values $x_i - x_{i-1} > \gamma$ or $x_i - x_{i-1} < -\gamma$, indicating that the slope is high.

The proposed modified SlpEn splits the symbols into three levels instead, including ties. Therefore, the assignment of symbols is now as follows.

- If $x_i > x_{i-1} + \theta$ (maximum difference with respect to the parameter θ), the symbol assigned is +3, indicating a large positive slope.
- If $x_i > x_{i-1} + \gamma$ and $x_i \leq x_{i-1} + \theta$ indicating a medium positive slope, the symbol assigned is +2.
- If $x_i > x_{i-1} + \delta$ and $x_i \leq x_{i-1} + \gamma$ (below γ), an area that can be considered low from the point of view of positive slopes, the symbol assigned is +1.
- In the region close to a gradient or slope of 0, when $|x_i - x_{i-1}| \leq \gamma$, the symbol assigned is 0. This area represents ties or equal values, which can create ambiguities in other metrics.
- If $x_i < x_{i-1} - \delta$ and $x_i \geq x_{i-1} - \gamma$ (above the -45° angle when $\gamma = 1$ and below the 0 slope zone), the resulting symbol is -1. SlpEn uses a symmetric quantization, but an asymmetric one could be used in future studies.
- If $x_i < x_{i-1} - \gamma$ and $x_i \geq x_{i-1} - \theta$ is assigned as symbol -2, representing the average negative value.
- Finally, if $x_i < x_{i-1} - \theta$ (maximum negative difference with respect to the parameter θ), the symbol assigned is -3, indicating a large negative slope.

So, instead of having -2, -1, 0, 1, and 2, we now have -3, -2, -1, 0, 1, 2, and 3, as shown in Figure 2.

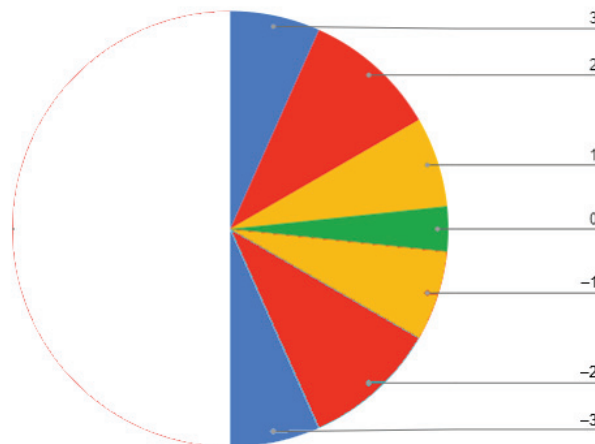


Figure 2. Graphical representation of the calculation of symbols used in SlpEn based on the thresholds γ , δ , and θ .

2.4. Classification Scheme

Using the experimental datasets described earlier, the optimal value of SlpEn that maximised the accuracy of classifying records was calculated using the symmetric strategy represented in Figures 1 and 2. Classification accuracy was defined as the percentage or ratio of time series correctly classified with respect to the total number of series in an experimental dataset.

This process was repeated using a three-parameter distribution of regions as in Figure 2. Now, in addition to having a specific value of γ , a higher value of θ was required. On the negative slopes region, θ is lower than γ , following the relationship $\gamma < \theta$ and $-\gamma > -\theta$.

A time series classification analysis was carried out, comparing accuracy between the original SIpEn and the new proposed SIpEn variation. A grid search was conducted using the described databases in Section 2.1 to find the optimal input parameter combination that yielded maximum accuracy in each case.

For the baseline SIpEn method, we varied the parameter m within the range 3 to 9, the δ parameter from 0 to δ , and γ from δ to 1.5. When using the additional parameter, θ varied from γ to 1.5. The threshold used for classification was obtained from the ROC curve of the process [11]. Specifically, the point on the curve closest to (1, 0) was used.

3. Experiments and Results

The experiments results showed a small improvement using the newly proposed method. Specifically, the proposed SIpEn variation exhibited small improvements of around 3% in classification accuracy after using a grid search. Table 1 presents a report of the highest values of accuracy obtained with both SIpEn methods. However, the modified SIpEn is far more time consuming than the original SIpEn.

Table 1. A comparative study between original SIpEn and modified SIpEn.

Datasets	Classification Accuracy	
	Original SIpEn	Modified SIpEn
The Bern–Barcelona	79%	81%
The Fantasia	86%	89%
The Ford A	94%	94%
The House Twenty	97%	97%
The PAF prediction	81%	83%
The Worms two class	72%	72%
The Bonn EEG dataset	95%	95%

4. Discussion

The highest reported accuracy was for Fantasia, which improved by 3% from 86% to 89%. PAF prediction and Bern–Barcelona both increased by 2%, from 79% to 81% and from 81% to 83%, respectively. Ford A, House Twenty, Worms two class, and Bonn EEG datasets maintained the same accuracy, at 94%, 97%, 72%, and 95%, respectively.

Dividing the gradient into three or five levels does not seem to have a clear impact on classification performance. Therefore, adding more parameters to SIpEn is not advisable considering the amount of time consumed to achieve the small accuracy gains.

5. Conclusions

In this work, we presented a comparative study using different time series datasets to understand the impact of adding a new thresholding parameter to SIpEn. We introduced the parameter θ , and added it to δ and γ , expanding the symbolic intervals from $-2, -1, 0, 1$, and 2 to $-3, -2, -1, 0, 1, 2$, and 3 . The results confirmed that the new method achieved a minor improvement of 3%, but at the expense of a significant processing time increase. Therefore, we do not recommend adding a new thresholding parameter due to the diminishing return achievable, unless a minor classification improvement is critical (for instance, in medical diagnosis applications).

Author Contributions: M.K. implemented the algorithms and carried out the experiments, and also wrote the initial version of the paper and prepared the presentation. D.C.-F. devised the idea and objectives of the study, the methodology, and reviewed the paper and the presentation. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data used in the experiments is publicly available at the sites included in the bibliographic references.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kouka, M.; Cuesta-Frau, D. Slope Entropy Characterisation: The Role of the δ Parameter. *Entropy* **2022**, *24*, 1456. [CrossRef]
2. Cuesta-Frau, D. Slope entropy: A new time series complexity estimator based on both symbolic patterns and amplitude information. *Entropy* **2019**, *21*, 1167. [CrossRef]
3. Andrzejak, R.G.; Schindler, K.; Rummel, C. Nonrandomness, nonlinear dependence, and nonstationarity of electroencephalographic recordings from epilepsy patients. *Phys. Rev. E* **2012**, *86*, 046206. [CrossRef] [PubMed]
4. Iyengar, N.; Peng, C.K.; Morin, R.; Goldberger, A.L.; Lipsitz, L.A. Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **1996**, *271*, R1078–R1084. [CrossRef] [PubMed]
5. FordA Description. Available online: <http://www.timeseriesclassification.com/description.php?Dataset=FordA> (accessed on 7 March 2023).
6. HouseTwenty. Available online: <http://www.timeseriesclassification.com/description.php?Dataset=HouseTwenty> (accessed on 7 March 2023).
7. Dean, M.E. *Prefiltering for Improved Unknown and Known Source Correlation Detection of Broadband Oscillatory Transients and Predicting the Onset of Paroxysmal Atrial Fibrillation Using Feature Extraction and a Hamming Neural Network*; University of New Orleans: New Orleans, LA, USA, 2003.
8. WormsTwoClass. Available online: <https://www.timeseriesclassification.com/description.php?Dataset=WormsTwoClass> (accessed on 7 March 2023).
9. Yemini, E.; Jucikas, T.; Grundy, L.J.; Brown, A.E.; Schafer, W.R. A database of caenorhabditis elegans behavioral phenotypes. *Nat. Methods* **2013**, *10*, 877–879. [CrossRef] [PubMed]
10. Tsipouras, M.G. Spectral information of EEG signals with respect to epilepsy classification. *Eurasip J. Adv. Signal Process.* **2019**, *2019*, 10. [CrossRef]
11. Hand, D.J.; Till, R.J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach. Learn.* **2001**, *45*, 171–186. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Recurrent Forecasting in Singular Spectrum Decomposition [†]

Maryam Movahedifar ^{1,*}, Hossein Hassani ^{2,‡} and Mahdi Kalantari ^{3,‡}

¹ Institute for Statistics, University of Bremen, 28359 Bremen, Germany

² Research Institute of Energy Management and Planning, University of Tehran, Tehran 1417466191, Iran; hosseinhassani57@webster.edu

³ Department of Statistics, Payame Noor University, Tehran 19395-4697, Iran; kalantarimahdi@pnu.ac.ir

* Correspondence: movahedm@uni-bremen.de

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

[‡] These authors contributed equally to this work.

Abstract: In this paper, the Recurrent Singular Spectrum Decomposition (R-SSD) algorithm is proposed as an improvement over the Recurrent Singular Spectrum Analysis (R-SSA) algorithm for forecasting non-linear and non-stationary narrowband time series. R-SSD modifies the embedding step of the basic SSA method to reduce energy residuals. This paper conducts simulations and real-case studies to investigate the properties of the R-SSD method and compare its performance with R-SSA. The results show that R-SSD yields more accurate forecasts in terms of ratio root mean squared errors (RRMSEs) and ratio mean absolute errors (RMAEs) criteria. Additionally, the Kolmogorov–Smirnov Predictive Accuracy (KSPA) test indicates significant accuracy gains with R-SSD over R-SSA, as it measures the maximum distance between the empirical cumulative distribution functions of recurrent prediction errors and determines whether a lower error leads to stochastically less error. Finally, the non-parametric Wilcoxon test confirms that R-SSD outperforms R-SSA in filtering and forecasting new data points.

Keywords: Singular Spectrum Analysis; signal extraction; recurrent forecasting; Kolmogorov–Smirnov

1. Introduction

Singular Spectrum Analysis (SSA) is a widely used tool for time series analysis and signal processing, first introduced by Broomhead and King [1] in 1986. Over the years, several studies, including [2–9], have attempted to improve the decomposition, reconstruction, and forecasting capabilities of SSA in various fields. The method breaks down a time series into a few principal components that are used to reconstruct the original series, making it an efficient analysis tool that focuses on the most relevant features of the data. Moreover, SSA does not rely on statistical assumptions such as linearity or stationarity, which are often unrealistic in real-world scenarios. Both univariate and multivariate time series data can be analyzed using SSA, with the former examining a single time series variable and the latter studying multiple time series variables simultaneously, for more details see [9–17]. Singular Spectrum Analysis (SSA) can be utilized for forecasting future trends. The first step in applying SSA to forecasting is to decompose the time series into its trend, seasonal, and noise components. Once these components have been identified, they can be extrapolated into the future using various methods, such as Vector SSA (V-SSA) and Recurrent SSA (R-SSA) [17]; while V-SSA has proven effective in many instances, there is still room for improvement in the R-SSA forecasting approach. This paper proposes an innovative recurrent forecasting algorithm called R-SSD, which is expected to generate more accurate results. The R-SSD method generates its coefficients from a modified trajectory matrix based on the new Singular Spectrum Decomposition (SSD) method over time–frequency datasets, see [18]. SSD is an iterative approach that is based on the SSA decomposition method and chooses the embedding dimension and principal components for the reconstruction and forecasting of a

Citation: Movahedifar, M.; Hassani, H.; Kalantari, M. Recurrent Forecasting in Singular Spectrum Decomposition. *Eng. Proc.* **2023**, *39*, 68. <https://doi.org/10.3390/engproc2023039068>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 7 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

specific component series in a fully data-driven manner. In the Singular Spectrum Analysis (SSA) method, the number of observations needed to construct the trajectory matrix is not fixed and can vary. On the other hand, the Singular Spectrum Decomposition (SSD) method requires a fixed number of repetitions of observations to construct the trajectory matrix. The window length, denoted as L , determines the number of rows in the trajectory matrix in both methods. A larger window length is preferred if the goal is to retain more information, while a smaller window length is better for achieving statistical confidence, for more details see [19,20]. When addressing time series that exhibit different frequency domains, such as those with harmonic patterns where, for example, the first half of the signal has low-frequency and the second half has high-frequency oscillations, extracting the oscillatory components using the SSA method can be challenging as it requires setting an appropriate window length at each step. However, the SSD method overcomes this limitation by setting the embedding dimension or window length ($L \leq N/2$) as a linear function of the inverse of the dominant frequency of the data, denoted as $1/f_{max}$. This adaptive approach ensures that SSD is a flexible decomposition method that can increase its ability to capture oscillatory components while reducing residual energy, as detailed in Appendix A.1 of [21]. As a result, it can be expected that SSD, being an improved version of SSA, can provide more accurate predictions for new data points in time series with different frequency domains.

The structure of this paper is as follows. In Section 2, we provide an introduction to the methodology of the basic SSA method and the recurrent forecasting algorithm. In Section 3, we present the methodology of the novel R-SSD forecasting approach. The results of a simulation study, evaluating the properties and performance of the proposed R-SSD method and comparing it to the established R-SSA approach, as well as the analysis of real data, are reported in Section 4. All calculations were performed using R software, specifically the `Rssa` package. Finally, in Section 5, we provide concluding remarks and highlight the key findings of our study.

2. Singular Spectrum Analysis (SSA)

Singular Spectrum Analysis (SSA) is an effective nonparametric technique for analyzing data. It can decompose a series into multiple components and make predictions based on them. The method comprises two distinct stages: decomposition and reconstruction, each of which involves two separate steps. To perform the SSA method, Algorithm 1 outlines the general process, and we primarily rely on the guidelines presented in [22,23].

Algorithm 1: Singular Spectrum Analysis (SSA).

Input: Time series $\mathbb{Y} = (y_1, \dots, y_N)$, $N > 2$, embedding window length L , and number of eigentriples r

Output: Underlying components of the time series

1. Embedding: Construct the trajectory matrix \mathbf{X} by taking time-lagged vectors of length L from the time series.
 2. Singular value decomposition (SVD): Compute the SVD of the trajectory matrix \mathbf{X} to obtain the singular vectors and singular values.
 3. Grouping: Select the first r eigentriples based on the characteristics of the singular values.
 4. Reconstruction: Reconstruct the underlying components of the time series by multiplying the retained eigentriples with the appropriate columns of the trajectory matrix, and summing across these products.
-

R-SSA Forecasting Algorithm

Forecasting with SSA is applicable to time series that approximately satisfy a linear recurrent relation (LRR). The general process for forecasting using the SSA method is outlined by Algorithm 2, also described by Golyandina et al. [24].

Algorithm 2: Recurrent Forecasting in Singular Spectrum Analysis (SSA).

Input: Time series $\mathbb{Y} = (y_1, \dots, y_N), N > 2$, Window Length $L, 1 < L < N$, Linear space $\mathcal{L}_r \subset \mathbb{R}^L$ of dimension $r < L$. {It is assumed that $e_L \notin \mathcal{L}_r$ where $e_L = (0, 0, \dots, 0, 1)^T \in \mathbb{R}^L$, in other terms, \mathcal{L}_r is not a ‘vertical’ space.}

Output: Forecasts for the next h time steps

1. Construct the trajectory matrix $\mathbf{X} = [X_1, \dots, X_K]$ of the time series $\mathbb{Y} = (y_1, \dots, y_N)$.
2. Compute the singular value decomposition (SVD) of \mathbf{X} to obtain the orthonormal basis vectors $U_i (i = 1, \dots, r)$ for the subspace \mathcal{L}_r .
3. Perform the orthogonal projection step by computing the matrix $\hat{\mathbf{X}} = [\hat{X}_1, \dots, \hat{X}_K] = \sum_{i=1}^r U_i U_i^T \mathbf{X}$. The vector \hat{X}_i is the orthogonal projection of X_i onto the subspace \mathcal{L}_r .
4. Construct the matrix $\tilde{\mathbf{X}} = \mathcal{H}\hat{\mathbf{X}} = [\tilde{X}_1 : \dots : \tilde{X}_K]$, which is referred to as the Hankelization step.
5. Set $\nu^2 = \pi_1^2 + \dots + \pi_r^2$, where π_i is the last component of the vector $U_i (i = 1, \dots, r)$. Moreover, assume that $e_L \notin \mathcal{L}_r$. This implies that \mathcal{L}_r is not a vertical space. Therefore, $\nu^2 < 1$.
6. Determine the vector $R = (\alpha_{L-1}, \dots, \alpha_1)^T: R = 1(1 - \nu^2) \sum_{i=1}^r \pi_i U_i^*$, where $U_i^* \in \mathbb{R}^{L-1}$ is the vector consisting of the first $L - 1$ components of the vector U_i . Note that this does not depend on the choice of a basis U_1, \dots, U_r in the linear space \mathcal{L}_r .
7. Define the time series $Y_{N+h} = (y_1, \dots, y_{N+h})$ using the following formula:

$$y_i = \begin{cases} \tilde{y}_i & \text{for } i = 1, \dots, N \\ \sum_{j=1}^{L-1} \alpha_j y_{i-j} & \text{for } i = N + 1, \dots, N + h, \end{cases} \quad (1)$$

where $\tilde{y}_i (i = 1, \dots, N)$ are the reconstructed series. The values y_{N+1}, \dots, y_{N+h} are the h step-ahead recurrent forecasts.

3. Singular Spectrum Decomposition (SSD)

In this section, we will introduce the Singular Spectrum Decomposition (SSD) method and the related recurrent forecasting technique. The SSD method consists of a two-stage approach with two steps in each stage as follows:

Stage 1. Decomposition (Modified Embedding and SVD)

The proposed approach enhances the basic SSA method by using a modified trajectory matrix for a given time series $Y = (y_1, \dots, y_N)$. The trajectory matrix is of size $(L \times N)$, where L is the embedding dimension, and is denoted as \mathbf{X}_{SSD} . It can be expressed as

$$\mathbf{X}_{SSD} = \left(\begin{array}{cccc|ccc} y_1 & y_2 & \cdots & y_K & y_{K+1} & \cdots & y_N \\ y_2 & y_3 & \cdots & y_{K+1} & y_{K+2} & \cdots & y_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & \cdots & y_N & y_1 & \cdots & y_{L-1} \end{array} \right) = [\mathbf{X} \mid \mathbf{A}]. \quad (2)$$

Compared to the basic SSA method, the trajectory matrix in the SSD method includes an additional block \mathbf{A} , which leads to the incorporation of different permutations of the total time series vector in each row of the modified trajectory matrix denoted as \mathbf{X}_{SSD} . Further details can be found in references [18,21].

Stage 2. Reconstruction (Grouping and Diagonal Averaging)

Similar to the grouping step in basic SSA (Section 2), a group of l eigentriples is selected in the SSD method. In the diagonal averaging step, a matrix denoted as \tilde{X}_{SSD} is computed as an approximation of X_{SSD} . This is achieved by computing the sum of l matrices, each obtained by taking the outer product of the corresponding eigenvectors. Mathematically, $\tilde{X}_{SSD} = \sum_{k=1}^l \tilde{U}_k \tilde{U}_k^T X_{SSD}$, where \tilde{U}_k s are the corresponding eigenvectors. The transition to a one-dimensional time series can be achieved as follows:

$$i + j = \begin{cases} k + 1 \text{ and } k + 1 + N & \text{when } i + j < N \\ k + 1 & \text{when } i + j \geq N. \end{cases} \tag{3}$$

3.1. Choice of the Embedding Dimension

The choice of the embedding dimension in Singular Spectrum Analysis (SSA) is crucial for accurately capturing the underlying structure of a time series. The embedding dimension determines the number of time-lagged vectors used to construct the trajectory matrix, affecting the amount of information retained in the decomposition. The embedding dimension should be chosen large enough to capture all relevant information, but not too large so as to include noise or irrelevant information, which can lead to an inaccurate decomposition and overfitting. A common rule of thumb for choosing the embedding dimension L in SSA is $L \leq N/2$, see [25]. Furthermore, Vautard [26] proposed a criterion for determining the appropriate window length in Singular Spectrum Analysis (SSA) when analyzing time series with intermittent oscillations. According to this criterion, SSA can isolate intermittent oscillations correctly if the inverse of the maximum spectral density of the time series, denoted as f_{max} , is less than or equal to the window length L . In other words, L should be chosen such that $1/f_{max} \leq L$. However, for time series with varying frequency domains, extracting oscillatory components using the SSA method can be challenging due to the need to set an appropriate window length at each step, while in the SSD method, the window length L is selected as a linear function of $1/f_{max}$ and should be less than $N/2$, where N is the length of the time series. This approach captures local structures in the time series while minimizing noise inclusion.

3.2. R-SSD Forecasting Algorithm

Let $\tilde{\lambda}_1, \dots, \tilde{\lambda}_L$ be the eigenvalues of $X_{SSD} X_{SSD}^T$, and $\tilde{U}_1, \dots, \tilde{U}_L$ be the corresponding eigenvectors for the trajectory matrix X_{SSD} . Then, the new R-SSD coefficients can be computed as $\tilde{R} = (\tilde{\alpha}_{L-1}, \dots, \tilde{\alpha}_1) = 1/(1 - \tilde{v}^2) \sum_{i=1}^L \tilde{\pi}_i \tilde{U}_i^T$, where \tilde{U}_i^T is the vector consisting of the first $L - 1$ components of the vector \tilde{U}_i , $\tilde{\pi}_i$ is the last component of the vector \tilde{U}_i and $\tilde{v}^2 = \sum_{i=1}^L \tilde{\pi}_i^2$. Finally, to obtain the forecasting algorithm of R-SSD, we replace the α_j values in Equation (1) with $\tilde{\alpha}_j$ values, where \tilde{y}_i s ($i = 1, \dots, N$) are the reconstructed series obtained using the SSD method.

4. Empirical Results

We assess the performance of the R-SSA and R-SSD forecasting methods on real and simulated time series in this section. A portion of the data is used for training, while the remaining data are reserved for testing. We evaluate the accuracy of forecasting using the root mean squared error (RMSE) and mean absolute error (MAE) criteria and compare the results using the ratios defined in Equations (4) and (5).

$$RRMSE_h = \frac{\sqrt{\sum_{t=m}^{m+n-h} (y_{t+h} - \hat{y}_{t+h|t})^2}}{\sqrt{\sum_{t=m}^{m+n-h} (y_{t+h} - \hat{y}_{t+h|t})^2}} \tag{4}$$

$$RMAE_h = \frac{\sum_{t=m}^{m+n-h} |y_{t+h} - \hat{y}_{t+h|t}|}{\sum_{t=m}^{m+n-h} |y_{t+h} - \hat{y}_{t+h|t}|} \tag{5}$$

where the lengths of the training sample, test sample, and forecast horizon are denoted by m , n and h , respectively. On the other hand, $\hat{y}_{t+h|t}$ denote the h -step ahead forecast obtained via the new R-SSD forecasting method and $\hat{y}_{t+h|t}$ denote the h -step ahead forecast obtained via the R-SSA forecasting method. If the ratio of the average RMSE values obtained by R-SSD and R-SSA, denoted as $RRMSE$, is less than 1 at a given forecasting horizon h , denoted as $RRMSE_h < 1$, then the R-SSD procedure is more accurate than R-SSA at horizon h . Alternatively, when $RRMSE_h > 1$, it can be inferred that the accuracy of the R-SSD procedure is less than R-SSA. The same inference can be made using the ratio of the average MAE values obtained by R-SSD and R-SSA, denoted as $RMAE$. Additionally, to compare the accuracy of two sets of forecasts, the Kolmogorov–Smirnov Predictive Accuracy (KSPA) test is considered, as proposed in [27]. The KSPA test has two objectives: firstly, to determine if there is a significant statistical difference between the distribution of predictive errors by testing if the empirical cumulative distribution functions F_{SSD} and F_{SSA} for the forecast errors of the two methods are significantly different. The two-sided KSPA test evaluates this difference with $H_0 : F_{SSD}(z) = F_{SSA}(z)$ and $H_1 : F_{SSD}(z) \neq F_{SSA}(z)$. The second objective of the KSPA test is to determine if the method with the lowest error based on a given loss function also exhibits a statistically significantly smaller error than the corresponding method. The one-sided KSPA test is formulated as $H_0 : F_{SSD}(z) \leq F_{SSA}(z)$ and $H_1 : F_{SSD}(z) > F_{SSA}(z)$. Rejecting the null hypothesis indicates that the cumulative distribution function (c.d.f.) of forecast errors obtained from the SSD model is shifted toward the left and above the c.d.f. of forecast errors obtained from the SSA model, suggesting that the SSD method has a smaller stochastic error compared to the SSA method, for more details see [27].

In the following, two simulated time series with a length of 200 are generated, with the first 140 observations being designated as the training sample ($m = 140$) and the remaining data as the test sample ($n = 60$). The number of leading eigenvalues (r) for reconstructing and forecasting the time series is selected based on the rank of the corresponding trajectory matrix. This simulation is repeated 1000 times, and the mean of RRMSEs and RMAEs are calculated.

4.1. Simulated Examples

Example 1. In the first example, we examine a sine series that encompasses two distinct frequencies, as illustrated below:

$$y_t = \begin{cases} \sin(2\pi t) + \varepsilon_t, & 1 \leq t \leq 100 \\ \sin(5\pi t) + \varepsilon_t, & 101 \leq t \leq 200 \end{cases}$$

where the noise term ε_t is generated from a normal distribution at varying levels of signal-to-noise ratio (SNR). In this example, both basic SSA and SSD methods are compared using a rank value of $r = 5$ for forecasting horizons of $h = 1, 3, 6, 12$, and 24 steps ahead. The R-SSD method outperforms the basic R-SSA method in terms of forecasting accuracy across all window lengths (L) and SNR levels tested, as shown in Figures 1 and 2. For nearly all forecast horizons h , the values of RRMSE and RMAE are less than 1, indicating that the R-SSD method provides more accurate predictions than the R-SSA method. The accuracy is consistent across different metrics, with the lowest RRMSE and RMAE values occurring at the lowest window length level ($L = 6$) for all SNR levels when $h = 12$ and 24. Overall, the results suggest that the R-SSD method is superior to the R-SSA method in providing accurate predictions.

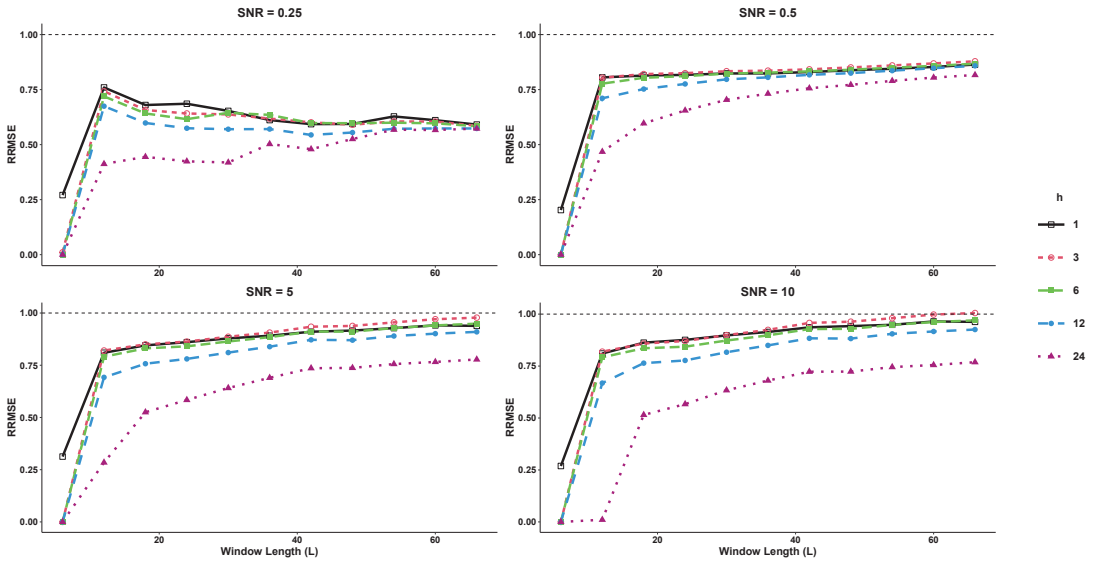


Figure 1. The RRMSE results for different forecast horizons ($h = 1, 3, 6, 12,$ and 24) in Example 1.

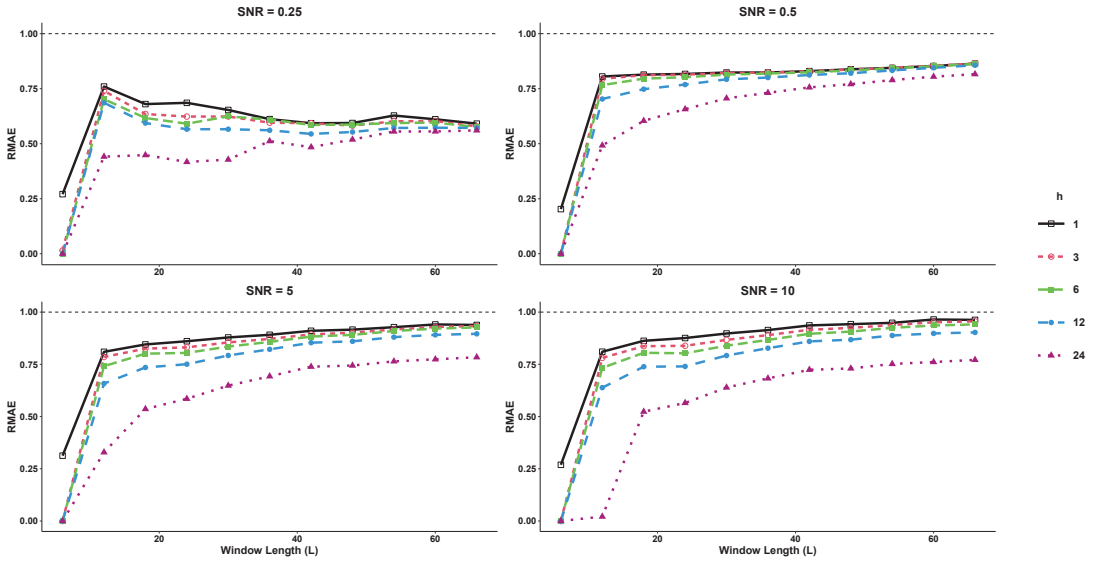


Figure 2. The RMAE results for different forecast horizons ($h = 1, 3, 6, 12,$ and 24) in Example 1.

Example 2. Example 2 involves an exponential series with two different frequencies as follows:

$$y_t = \begin{cases} \exp(\alpha_0 + \alpha_1 t) + \cos(2\pi t/6) + \varepsilon_t, & 1 \leq t \leq 100 \\ \exp(\alpha_0 + \alpha_1 t) + \cos(5\pi t/6) + \varepsilon_t, & 101 \leq t \leq 200 \end{cases}$$

where the term ε_t represents the noise generated from a normal distribution at various levels of signal-to-noise ratio (SNR). In this study, both basic SSA and SSD methods use a rank of 25 for the trajectory matrix of the time series, with $\alpha_0 = 0$ and $\alpha_1 = 0.01$. RRMSE and RMAE are computed for various forecast horizons ($h = 1, 3, 6, 12,$ and 24) and SNR levels. As shown in Figures 3 and 4, RRMSE and RMAE increase as the forecast horizon decreases, but decrease significantly when

$h = 24$. The results indicate that the R-SSD method performs better as the value of L decreases. However, for higher SNR levels, the R-SSA method outperforms the R-SSD method for larger values of L . The lowest RRMSE and RMAE values are achieved when the window length and SNR are at their lowest values.

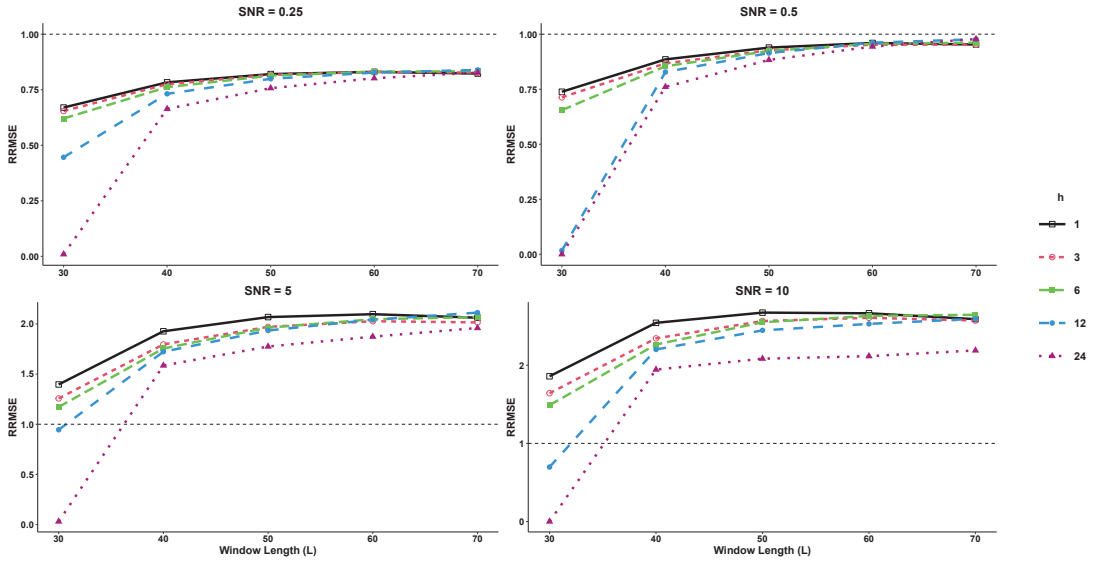


Figure 3. The RRMSE results for different forecast horizons ($h = 1, 3, 6, 12,$ and 24) in Example 2.

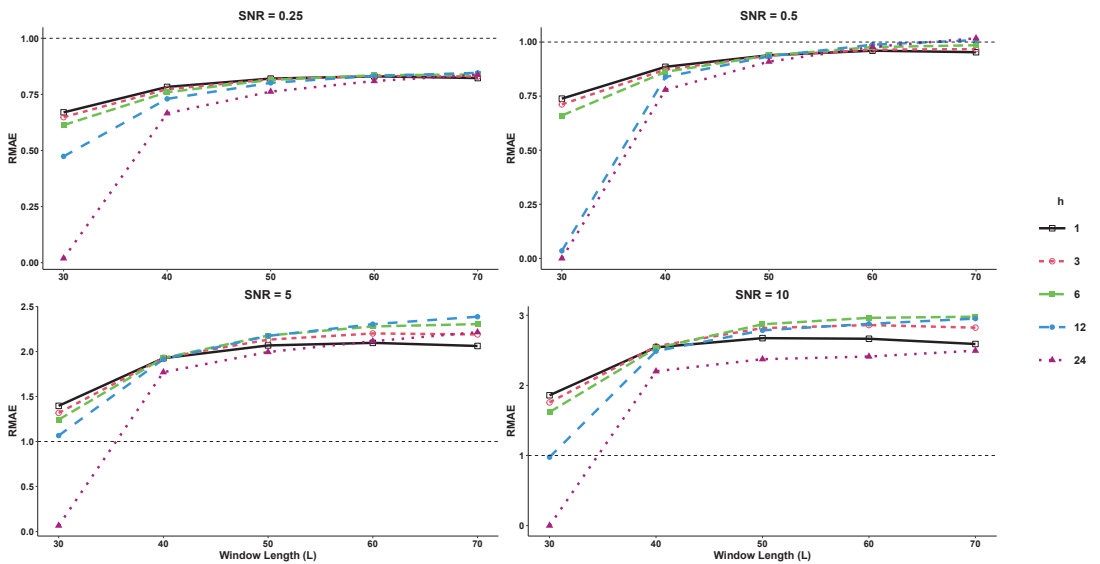


Figure 4. The RMAE results for different forecast horizons ($h = 1, 3, 6, 12,$ and 24) in Example 2.

4.2. Real Data Analysis

In this section, we compare the forecasting performance of the proposed R-SSD method with the basic R-SSA method using real data from fruit fly (*Drosophila melanogaster*) embryos. The caudal protein in fruit fly embryos plays a crucial role in tail formation,

acting as a transcription factor that regulates the expression of other genes by binding to specific DNA sequences. The caudal protein is expressed in the cells of the “tail bud”, which gives rise to the tail, and its activation triggers a gene expression cascade that controls cell division, differentiation, and migration, ultimately leading to tail formation. Mutations in the caudal gene can result in a loss of function of the protein, leading to defects in tail formation such as a short or absent tail, as well as other developmental defects related to segmentation. However, it is important to note that causality detection techniques, as demonstrated by previous studies, can be sensitive to noise [23,28–30]. Here, we analyze four gene expression profiles with varying lengths and dominant frequencies to demonstrate the importance of utilizing an accurate noise filtering method, such as SSD, for conducting reliable causality studies. To compare the forecasting performance of the R-SSA and R-SSD approaches, we provide Tables 1–4 to summarize the obtained results for four different time classes: ab2, ab18, be11, ad14. These tables display the respective forecasting metrics, including RRMSE and RMAE, for each time class, enabling a comprehensive comparison between the two methods. For each dataset, we considered the first 80% of observations as the training sample and the remaining 20% as the test sample. The number of leading eigenvalues (r) for reconstructing and forecasting the time series was selected based on the rank of the corresponding trajectory matrix. Additionally, the dominant frequency of the data ($1/f_{max}$) was calculated for each dataset, and the window length was chosen as a multiple of $1/f_{max}$ and less than $N/2$. After selecting the appropriate L and r , we utilized the observations from the training set to forecast the test sample data and calculate the RRMSE and RMAE criteria for different h step-ahead recurrent forecasts, using Equations (4) and (5).

Table 1. RRMSE and RMAE analysis of Cad Profile ab2, $1/f_{max} = 2$ with $r = 16$.

Horizon	$L = 20$		$L = 30$		$L = 50$	
	RRMSE	RMAE	RRMSE	RMAE	RRMSE	RMAE
1	0.96	0.96	1.02	1.02	1.12	1.12
3	0.86	0.91	1.03	1.05	1.13	1.16
6	0.78	0.80	0.96	1.02	1.09	1.19
12	0.59	0.62	0.86	0.92	1.11	1.22
24	0.17	0.21	0.59	0.67	1.00	1.09

Based on the results presented in Table 1, it is evident that there is a discernible difference in the RRMSE and RMAE values obtained using the R-SSA and R-SSD methods for $L = 20, 30$, and 50 . The performance metrics show contrasting outcomes for these window lengths, indicating that the choice of method can significantly impact the forecasting accuracy.

Table 2. RRMSE and RMAE analysis for Cad profile ab18, $1/f_{max} = 2$, with $r = 14$.

Horizon	$L = 20$		$L = 50$		$L = 80$	
	RRMSE	RMAE	RRMSE	RMAE	RRMSE	RMAE
1	0.82	0.82	1.27	1.27	1.14	1.14
3	0.88	0.86	1.15	1.21	1.01	1.07
6	0.91	0.92	1.17	1.26	1.03	1.10
12	0.90	0.91	1.23	1.31	1.08	1.17
24	0.84	0.86	1.25	1.36	1.16	1.31

Table 2 shows the RRMSE and RMAE values obtained by each model for the cad profile ab18. As indicated in the table, the R-SSD method achieves a significant reduction in both RRMSE and RMAE values for $L = 20$, which suggests that it generally provides better signal extraction and forecast results compared to the R-SSA model for this window length.

Additionally, for $L = 50$ and 80 , the accuracy of the two methods is similar, indicating that the R-SSD method can be preferable for smaller window lengths.

Table 3. RRMSE and RMAE analysis of Cad Profile be11, $1/f_{max} = 3$ with 3.

Horizon	$L = 6$		$L = 18$		$L = 30$	
	RRMSE	RMAE	RRMSE	RMAE	RRMSE	RMAE
1	1.16	1.16	1.21	1.21	1.15	1.15
3	1.16	1.17	1.07	1.11	1.13	1.17
6	1.22	1.20	1.05	1.09	1.11	1.16
12	1.35	1.38	0.80	0.85	0.95	0.98
24	1.89	2.03	0.21	0.25	0.43	0.44

Table 3 summarizes the results of RRMSE and RMAE for the cad profile be11. The findings indicate that the R-SSD method outperforms the R-SSA method, particularly for $L = 18$ and 30 and horizons $h = 12$ and 24 . Furthermore, a closer examination of the table reveals that the highest accuracy is obtained when $L = 18$ and $h = 24$, as evidenced by the greatest reduction in both RRMSE and RMAE values.

Table 4. RRMSE and RMAE analysis of Cad Profile ad14, $1/f_{max} = 11$ with 5.

Horizon	$L = 11$		$L = 88$		$L = 110$	
	RRMSE	RMAE	RRMSE	RMAE	RRMSE	RMAE
1	1.26	1.26	1.28	1.28	1.15	1.15
3	1.22	1.30	1.23	1.30	1.17	1.17
6	1.12	1.25	1.27	1.33	1.16	1.20
12	0.96	1.03	1.37	1.43	1.31	1.32
24	0.73	0.78	1.50	1.59	1.47	1.53

Additionally, the forecasting methods R-SSD and R-SSA were evaluated for statistical significance using the non-parametric two-sample Wilcoxon test and the Kolmogorov–Smirnov Predictive Accuracy (KSPA) test. The results show a statistically significant difference between the two methods, with R-SSD forecasts having smaller errors than R-SSA forecasts with 95% confidence based on the one-sided KSPA test. The two-sided KSPA test further supports the significant differences between the two methods with 95% confidence, and these findings are consistent across different embryos and L values, especially for $h = 24$. These results demonstrate the superior accuracy of the R-SSD method and highlight the importance of utilizing an accurate noise filtering method such as SSD for precise causality studies. The Wilcoxon test also confirms the significant differences between the two methods for all tested embryos and L values, with p -values less than 0.05.

5. Discussion

In this paper, we introduced a new forecasting method called Recurrent Singular Spectrum Decomposition (R-SSD), which improves upon the standard R-SSA method by enhancing the identification of fluctuation content and enabling a fully data-driven selection of window length and principal components for reconstructing component series based on dominant frequency periods. The results were evaluated using the non-parametric two-sample Wilcoxon test and RRMSE/RMAE criteria, which demonstrated the superiority of R-SSD over basic R-SSA in the majority of cases for various window lengths and forecasting horizons. KSPA tests confirmed the ability of R-SSD to obtain significant components for accurate forecasting of new data points. In summary, the proposed R-SSD method with its improved trajectory matrix definition and window length selection shows promising results in time series forecasting. Overall, the R-SSD method offers a viable alternative to the standard R-SSA method and could lead to improved forecasting accuracy in a wide range of applications. Further research and investigation into the R-SSD method’s performance

under different scenarios and datasets may be valuable for its continued development and potential adoption in practical settings.

Author Contributions: All authors contributed equally to the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available, upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Broomhead, D.S.; King, G. Extracting qualitative dynamics from experimental data. *Phys. D Nonlinear Phenom.* **1989**, *20*, 217–236. [CrossRef]
2. Hassani, H.; Ghodsi, Z.; Silva, E.; Heravi, S. From nature to maths: Improving forecasting performance in subspace-based methods using genetics Colonial Theory. *Digit. Signal Process.* **2016**, *51*, 101–109. [CrossRef]
3. Gong, Y.; Song, Z.; He, M.; Gong, W.; Ren, F. Precursory waves and eigenfrequencies identified from acoustic emission data based on Singular Spectrum Analysis and laboratory rock-burst experiments. *Int. J. Rock Mech. Min. Sci.* **2017**, *91*, 155–169. [CrossRef]
4. Yu, C.; Li, Y.; Zhang, M. An improved Wavelet Transform using Singular Spectrum Analysis for wind speed forecasting based on Elman Neural Network. *Energy Convers. Manag.* **2017**, *148*, 895–904. [CrossRef]
5. Rahman Khan, M.R.; Poskitt, D.S. Forecasting stochastic processes using singular spectrum analysis: Aspects of the theory and application. *Int. J. Forecast.* **2017**, *33*, 199–213. [CrossRef]
6. Heravi, S.; Osborn, D.R.; Birchenhall, C.R. Linear versus neural network forecasts for European industrial production series. *Int. J. Forecast.* **2004**, *20*, 435–446. [CrossRef]
7. Lai, L.; Guo, K. The performance of one belt and one road exchange rate: Based on improved singular spectrum analysis. *Phys. A Stat. Mech. Its Appl.* **2017**, *483*, 299–308. [CrossRef]
8. Hassani, H.; Yeganegi, M.; Khan, A.; Silva, E. The Effect of Data Transformation on Singular Spectrum Analysis for Forecasting. *Signals* **2020**, *1*, 4–25. [CrossRef]
9. Hassani, H.; Silva, E.S.; Antonakakis, N.; Filis, G.; Gupta, R. Forecasting accuracy evaluation of tourist arrivals. *Ann. Tour. Res.* **2017**, *63*, 112–127. [CrossRef]
10. Movahedifar, M.; Hassani, H.; Yarmohammadi, M.; Kalantari, M.; Gupta, R. A robust approach for outlier imputation: Singular spectrum decomposition. *Commun. Stat. Case Stud. Data Anal. Appl.* **2021**, *8*, 234–250. [CrossRef]
11. Chao, S.; Loh, C. Application of singular spectrum analysis to structural monitoring and damage diagnosis of bridges. *Struct. Infrastruct. Eng.* **2014**, *10*, 708–727. [CrossRef]
12. Chen, Q.; van Dam, T.; Sneeuw, N.; Collilieux, X.; Weigelt, M.; Rebeschung, P. Singular spectrum analysis for modeling seasonal signals from GPS time series. *J. Geodyn.* **2013**, *72*, 25–35. [CrossRef]
13. Hassani, H.; Webster, A.; Silva, E.; Heravi, S. Forecasting U.S. Tourist arrivals using optimal Singular Spectrum Analysis. *Tour. Manag.* **2015**, *46*, 322–335. [CrossRef]
14. Hutny, A.; Warzecha, M.; Derda, W.; Wiecezorek, P. Segregation of Elements in Continuous Cast Carbon Steel Billets Designated for Long Products. *Arch. Metall. Mater.* **2016**, *61*, 2037–2042. [CrossRef]
15. Liu, K.; Law, S.; Xia, Y.; Zhu, X.Q. Singular spectrum analysis for enhancing the sensitivity in structural damage detection. *J. Sound Vib.* **2014**, *333*, 392–417. [CrossRef]
16. Muruganatham, B.; Sanjith, M.A.; Kumar, B.; Murty, S.A.V.; Swaminathan, P. Roller element bearing fault diagnosis using singular spectrum analysis. *Mech. Syst. Signal Process.* **2013**, *35*, 150–166. [CrossRef]
17. Sanei, S.; Hassani, H. *Singular Spectrum Analysis of Biomedical Signals*; CRC Press: Boca Raton, FL, USA, 2015.
18. Movahedifar, M.; Yarmohammadi, M.; Hassani, H. Bicoid signal extraction: Another powerful approach. *Math. Biosci.* **2018**, *303*, 52–61. [CrossRef]
19. Hiemstra, C.; Jones, J.D. Testing for Linear and Nonlinear Granger Causality in the Stock Price- Volume Relation. *J. Financ.* **1994**, *49*, 1639–1664.
20. Ancona, N.; Marinazzo, D.; Stramaglia, S. Radial basis function approach to nonlinear Granger causality of time series. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **2004**, *70*, 056221. [CrossRef]
21. Bonizzi, P.; Karel, J.; Meste, O.; Peeters, R. Singular Spectrum Decomposition: A new method for time series decomposition. *Adv. Adapt. Data Anal.* **2014**, *6*, 1450011. [CrossRef]
22. Hassani, H. Singular Spectrum Analysis: Methodology and Comparison. *J. Data Sci.* **2007**, *5*, 239–257. [CrossRef] [PubMed]

23. Golyandina, N.; Korobeynikov, A.; Zhigljavsky, A. *Singular Spectrum Analysis with R*; Springer: Berlin/Heidelberg, Germany, 2018. [CrossRef]
24. Golyandina, N.; Nekrutkin, V.; Zhigljavsky, A.A. *Analysis of Time Series Structure: SSA and Related Techniques*, 1st ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2001. [CrossRef]
25. Golyandina, N.; Zhigljavsky, A. *Singular Spectrum Analysis for Time Series*; Springer: Berlin/Heidelberg, Germany, 2013.
26. Vautard, R.; Yiou, P.; Ghil, M. Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Phys. D Nonlinear Phenom.* **1992**, *58*, 95–126. [CrossRef]
27. Hassani, H.; Silva, E. A Kolmogorov–Smirnov Based Test for Comparing the Predictive Accuracy of Two Sets of Forecasts. *Econometrics* **2015**, *3*, 590–609. [CrossRef]
28. Vautard, R.; Ghil, M. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Phys. D Nonlinear Phenom.* **1989**, *35*, 395–424. [CrossRef]
29. Zou, C.; Feng, J. Granger causality vs. dynamic Bayesian network inference: A comparative study. *BMC Bioinform.* **2009**, *10*, 122.
30. Golyandina, N.E.; Holloway, D.M.; Lopes, F.J.; Spirov, A.V.; Spirova, E.N.; Usevich, K.D. Measuring gene expression noise in early *Drosophila* embryos: Nucleus-to-nucleus variability. *Procedia Comput. Sci.* **2012**, *9*, 373–382. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Sim-to-Real Transfer in Deep Learning for Agitation Evaluation of Biogas Power Plants [†]

Andreas Heller ^{1,*}, Peter Glösekötter ¹, Lukas Buntkiel ², Sebastian Reinecke ² and Sven Annas ¹

¹ Fachhochschule Münster, Stegerwaldstr. 39, 48565 Steinfurt, Germany;

peter.gloeseikoetter@fh-muenster.de (P.G.); s.annas@fh-muenster.de (S.A.)

² Helmholtzinstitut Dresden-Rossendorf, Bautzner Landstraße 400, 01328 Dresden, Germany;

l.buntkiel@hzdr.de (L.B.); s.reinecke@hzdr.de (S.R.)

* Correspondence: andreas.heller@fh-muenster.de

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Biogas is an important driver in carbon-neutral energy sources. Many biogas digester setups, however, are not well optimized and waste energy or fail to maximize their gas output potential. To optimize these systems, a framework was developed to measure and predict digester systems' efficiencies by closely monitoring fluid movements. This framework includes a numerical calculation of fluid behavior (Computational Fluid Dynamics (CFD)), and Deep Learning to estimate the fluid shear-rates introduced by the agitator's action. Additionally, a novel measurement system is presented that can measure the same metrics, as simulated, in real-world environments. Lastly, an outlook is given that presents the options and extensions of the presented setup to reduce prediction error, minimize measuring efforts further, and recommend optimization approaches to the operator.

Keywords: ANN; artificial neural networks; CNN; convolutional neural networks; deep learning; CFD; computational fluid dynamics; agitation performance prediction; shear-rate

1. Introduction

Carbon-neutral energy sources play a crucial role in mitigating climate change. Statistics show that the energy generated from biomass in Germany has been continuously rising since 1991 [1]. These systems, however, are often built up and operated without an exact setup or analysis of the maximum efficiency. Thus, these systems often do not reach their full potential [2]. To produce energy, these systems agitate a fluid that consists of animal waste, energy crops, organic waste, and other materials, in varying amounts [3]. The agitation action keeps the fluid fermenting, thus producing a valuable biogas that consists mainly of methane (CH₄) and carbon dioxide (CO₂), and small amounts of other gases. Biogas is used in many different applications, such as heating or locomotion [4]. Effective agitation is essential during the fermentation process [5]. On the one hand, over-agitation wastes energy; on the other hand, under-agitation risks the formation of solid or foamy swimming layers that can result in the fermentation process stopping completely. In addition to the high variance in biogas fluid characteristics, these systems are set up with many varying factors, such as the type of rotor used, the size of the rotor, the height and diameter of the agitation vessel, and if the vessel is equipped with features to aid agitation by increasing turbulence. Having one setup for every biogas digester is impossible, individual analyses will probably increase the efficiency of these systems. This work approaches the optimization of process by deploying deep learning. The amount of data required to properly train a neural network far outpaces what is collectible in a reasonable timeline. For this, many different systems will have to be located and measured for extended periods. This problem can be overcome by computationally generating the required data. Computational Fluid Dynamics is a common practice to understand the flow of fluids in any configuration. This

Citation: Heller, A.; Glösekötter, P.; Buntkiel, L.; Reinecke, S.; Annas, S. Sim-to-Real Transfer in Deep Learning for Agitation Evaluation of Biogas Power Plants. *Eng. Proc.* **2023**, *39*, 69. <https://doi.org/10.3390/engproc2023039069>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 10 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

approach can be used to generate knowledge about the flow behavior in biogas power plants. A sim-to-real transfer is achieved by tweaking the pre-trained Artificial Neural Network (ANN) with real-world measurements.

This work presents the methodology used to design, simulate, measure, and predict agitation efficiencies using deep learning. First, the methodology of this approach is presented. The Computer-Aided Design and Computational Fluid Dynamics setup are outlined to simulate systems and generate data for the deep learning phase. Next, post-processing steps are implemented to convert simulation results to data that can also be obtained with a real-world measuring system. Then, the Neural Network Architecture is presented, where three commonly used approaches are implemented and compared. The Section 2 is concluded with a detailed look at the three Neural Network Architectures and their performances in the presented problem. After the Section 2, a Wireless Sensor Network (WSN) to measure real-world systems is outlined. Its purpose is to measure real-world systems, and to refine the neural network predictions. In the outlook, future work is described that will further improve the presented system by minimizing measurement effort and helping the operator to create an optimal agitation setup.

2. Methodology

In this section, the methodology used to set up simulation environments, numerical simulation, post-processing, the Neural Network Architectures, and deep learning is explained. The first part outlines the development of the 3D models that are to be used in the numerical analysis. Next, the numerical analysis in OpenFOAM is explained, including the setup of fluid properties, mesh-movements, mesh-boundary conditions, and information about the overall analysis process. The subsequent section outlines further processing of the results of numerical analysis using ParaView. These operations are crucial to align the data formats of real-world measurements and simulation. After the CFD post-processing, three artificial-neural-network architectures are outlined. Lastly, the performance of these networks on training and test data is presented.

2.1. CAD

This section outlines the Computer-Aided Design (CAD) of the biogas plant models used for this work. Two different models were implemented, from which three different simulation cases were derived.

Figure 1 shows the different models and their stirrer setups. A 3D model of the Landia-POP-I Slurry Mixer was implemented in a top model. The same mixer model was modified to allow for more degrees of freedom in regard to the rotors' position. For this, the body of the rotor was removed and all open faces were patched. This resulted in an abstract floating stirrer that can be placed anywhere in the vessel, without a disruption of the flow caused by the mixers body. Surrounding the stirrer are the vessel walls, and top and bottom lids. Depending on the type of biogas plant, these can be rounded (Figure 1 top) or fitted with plates (Figure 1 bottom). Not shown in the figure above are the top lid, which acts as the vessel walls, and the bottom lid, which keeps fluid from spilling out.

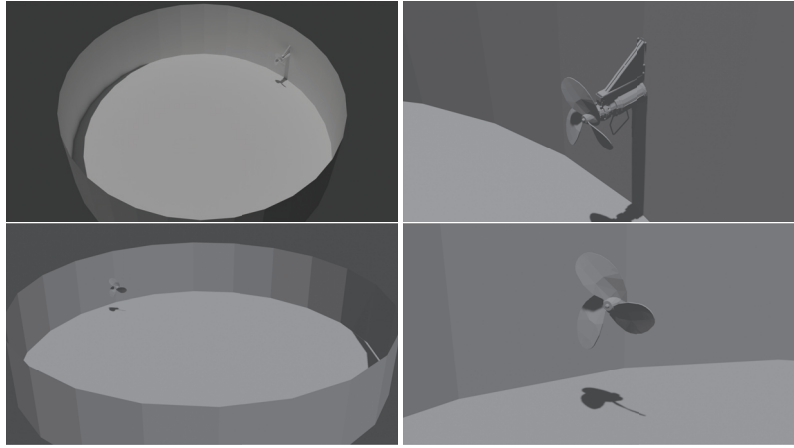


Figure 1. Renders of two different plant model setups. The top lids were removed for this visualization. **(Top)** shows a standard fermenter with a 28 m vessel diameter and filling height of 7 m **(left)** and a Landia POP inclined blade stirrer [6] that launches the fluid perpendicular from the vessel wall **(right)**. **(Bottom)** shows an agitation setup that is based on a real-world system where fluid-tracking measurements were taken in the past [7]. The vessel diameter is 11 m, vessel fill height is 2 m **(left)**, and a reduced Landia POP inclined blade stirrer is used to launch the fluid in a parallel motion to the vessel walls.

2.2. Computational Fluid Dynamics

For this work, three simulation cases were developed, where one is based on a previously measured system [7]. One challenge when simulating biogas plants is the non-Newtonian fluid's behavior. The *shear-thinning* fluid decreases its viscosity when a force is applied to it, varying its behavior in comparison to most fluids, like water. Additionally, these fluids exact properties that vary from their compositions and other variables, e.g., animal feed. To approximate the fluid's shear-thinning viscosity behavior, the approach of Fosca Conti et al. [8] was followed. Thus, for this simulation, the power-law model for the viscosity of Oswald–deWaele was chosen [9].

$$\nu = k\dot{\gamma}^{n-1}, \nu_{min} \leq \nu \leq \nu_{max} \quad (1)$$

with the kinematic viscosity ν and a consistency factor $K = 16.8\text{Pas}^m$, which equals a kinematic consistency factor. (The kinematic consistency factor k is the quotient of the consistency factor K and the fluid's density ρ ; $k = \frac{K}{\rho}$). of $k = 15.4 \cdot 10^{-3} \text{ m}^2\text{s}^{-1}$, a power-law index of $m = 0.35$, where $m < 1$, describes the fluid as shear-thinning. The fluid's kinematic viscosity limits are set to $\nu_{min} = 10^{-6} \text{ m}^2\text{s}^{-1}$ and $\nu_{max} = 10^{-3} \text{ m}^2\text{s}^{-1}$.

This simulation aims to compute field velocities v , and, with a custom post-processing-module, field shear-rates $\dot{\gamma}$. After the definition of fluid behaviors, the CAD models of all parts involved in fluid movements are implemented in a simulation case.

Figure 2 shows a CFD setup used for one of three case setups that have been created for this work. Additionally, the meshing of all the parts' features is highlighted.

For two cases, the Landia POP-I Slurry Mixer [6] was designed as per manufacturer's specifications, to describe rotating speeds of 150, and 300 RPM, depending on the chosen gearing. In a third case, a faulty agitator setup is implemented. Here, the rotor spins in the wrong direction with a very low rotating speed of 75 RPM. For all three cases, Arbitrary Mesh Interfaces (AMI) were implemented to cause rotor blades to spin and inflict movement on the fluid. Other case properties include:

- Turbulence model: Only the overall resulting fluid flow is of interest. The turbulence behavior is set to laminar.

- **Transient/steady state:** The start-up behavior of the system is of interest; a transient solver has been chosen.
- **Fluid compressibility:** With the overall slow fluid velocities, we can safely assume an incompressible fluid.

With the presented case characteristics, OpenFOAM's pimpleFoam-solver was chosen [10]. The CFD Simulation was carried on a Workstation PC, equipped with an Intel Core i5-8600k hexa-core processor [11]. Each case took five days of continuous computation to complete 100 s of simulation time. Data points of were captured in at least one-second time intervals.

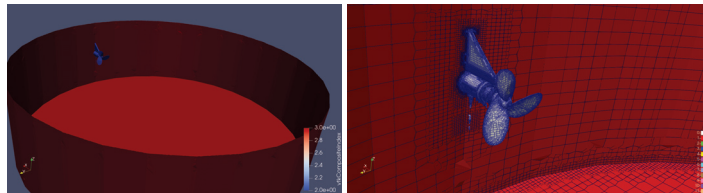


Figure 2. ParaView capture of a simulation case. **Left** shows a case setup of a previously simulated agitation system. The displayed vessel's radius is 14.25 m and its height is 7.8 m. **right** side shows a magnification view of the setup's agitator, with a 1.35 m rotor radius and rated rotational speed configured to 150 RPM.

2.3. CFD Post-Processing

After computing the required metrics, namely velocity v and shear rate $\dot{\gamma}$, virtual mass- and volume-less nodes are generated within the fluid using ParaView's *Particle Tracer* filter. These nodes follow the local velocity fields of the fluid and closely replicate the behavior of a real-world flow follower. The real-world measurement technique is outlined in Section 3.

Figure 3 shows the startup of an agitation system and the effects on the mass-less particles (blue nodes) floating in the fluid (hidden). The presented setup was developed in ParaView v5.10 [12], utilizing the engine's Python scripting interface. Numerous equidistant lines are created inside the vessel, which are used as seed sources for the particle tracer. Nodes created like this can then be used for further processing in Deep-Learning Applications.

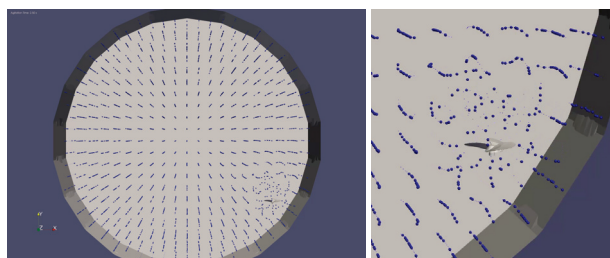


Figure 3. ParaView setup for CFD post-processing with particle tracer nodes at simulation 1.5 s after system start-up. **Left** shows a top-down view of an agitation setup, with 5.5 m vessel radius and 2 m height. The fluid is hidden. Mass-less are shown in blue. **Right** figure shows a magnified view of the rotor (0.5 m radius, 150 RPM), which is pointed north in this case. These figures show the system in startup, where particles around the rotor have already been affected by its agitation motion.

Figure 4 shows two perspectives of a trajectory plot with 100 random particles. Since all lines in this figure span the same time interval, longer lines denote particles with a greater average velocity. With this information, the rotor's location can be gauged. To decouple absolute node positions and related velocities, as seen in Figure 4, incremental

position information is derived from absolute positions. This step effectively transforms the position data into local particle velocities u .

$$u = \frac{\Delta p_{x,y,z}}{\Delta t} \tag{2}$$

where $\Delta p_{x,y,z}$ represents the particles position in a three-dimensional space, and Δt is the time interval between two position measurements. For this work, $\Delta t = 1$ s.

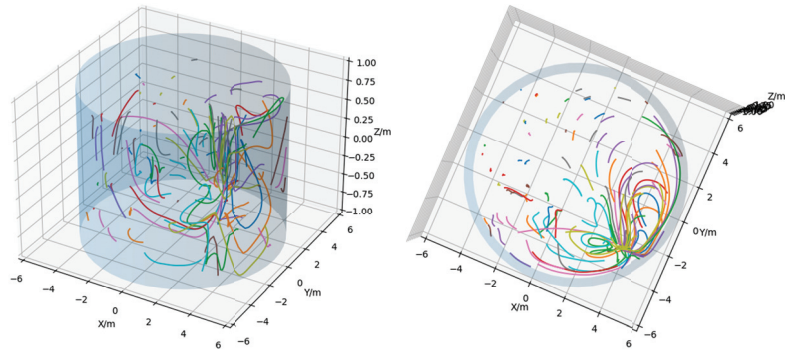


Figure 4. 3D Plot of 100 randomly chosen particle trajectories from the system seen in Figure 3. The cylinder approximates vessel walls and the colored lines denote the particles trajectory over the whole 100 s simulation time. **Right** shows the same 3D plot viewed from a top-down angle.

2.4. Neural Network Architecture

In this work, widely used regression network architectures are compared to estimate fluid shear-rates $\dot{\gamma}$. Fully Connected Neural Networks (FCNN) are simple setups, where the raw input (described in previous Section 2.3) is flattened and fed into the first dense layer of the FCNN. Subsequently, three hidden layers follow, with thirty-two, sixteen, and five neurons, before the network terminates with a single neuron that reflects its prediction of the mean trajectory shear-rate $\dot{\gamma}$.

The Fully Connected Neural Network architecture is shown in Figure 5.

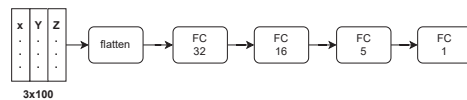


Figure 5. Fully Connected Neural Network (FCNN) architecture, with three densely connected hidden layers of thirty-two, sixteen, and five neurons. The activation function of the hidden layers is the Rectified Linear Unit (ReLU). The output layer consists of one single neuron with linear activation.

Another widely used network architecture is the Convolutional Neural Network (CNN), which adds various convolution- and max-pooling layers, preceding densely connected layers, to enhance features. Two different architectures for this type of network were developed.

The first CNN will be referred to as *1D-CNN*, because, like the FCNN, it starts with a flattening layer, resulting in Rank-1 tensor operations. Next, two 1D convolution layers, with kernel sizes of ten and seven, follow. After convolution, global max-pooling is applied, after which the network proceeds with the same three dense layers as presented for the FCNN (32, 16, 5, and 1 neurons).

Figure 6 shows the architecture of the 1D Convolutional Neural Network.

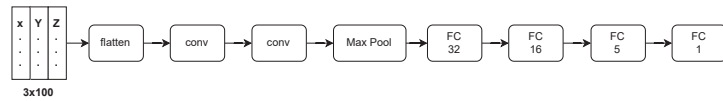


Figure 6. Convolutional Neural Network (CNN) with flattening operation in the first layer, referred to as *1D-CNN* in this work. The flattening follows the Rank-1 Tensor operations convolution with a size of 10 kernels, convolution with a size of 7 kernels, global max-pooling and the same densely connected layers as described in Figure 5. All hidden layers are configured with Rectified Linear Unit (ReLU) activation; the single neuron output layer has linear activation.

In contrast, a Rank-2 Tensor based CNN was implemented. This architecture follows the same outline as *1D-CNN*, using 3×3 convolution kernels. The main difference is that flattening is applied right before the first dense layer.

The *2D-CNN* architecture is shown in Figure 7.

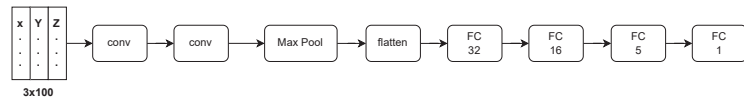


Figure 7. Convolutional Neural Network (CNN) with flattening operation in the last layer before the dense layers, referred to as *2D-CNN* in this work. Input data are fed into two Rank-2 Tensor convolution layers with 3×3 kernels, followed by a global max-pooling layer, before feeding these into the same densely connected layers as described in Figure 5. All hidden layers are configured with Rectified Linear Unit (ReLU) activation; the single neuron output layer undergoes linear activation.

2.5. Deep Learning

All networks outlined in the last section (Section 2.4) were trained over 150 epochs. The Mean-Squared Error (MSE) loss function was chosen for these regression architectures. With the framework explained in Sections 2.1–2.3, 25,977 datasets of complete length (100 s simulated time span, and 1 Hz update rate) were generated. Datasets were split into 80% training data, and 20% test data, which are only used for validation tests.

Figure 8 show all three network performances throughout the 150 epochs of training. The FCNN shows no meaningful improvements after around 20 training epochs, where the loss stays around 0.4 MSE. Additionally, the FCNN shows signs of overtraining after around 60 training epochs: the loss in the validation set increased, where as the loss in the training set continued to decrease. The *1D-CNN* performed very similarly to the FCNN, although it needs around 40 training epochs to achieve the same performance of around 0.4 MSE. After around 100 training epochs, this network shows sings of overtraining, with the MSE not deviating much from the FCNN's. The FCNN and the *1D-CNN* never reached validation losses close to the ones reached on the training-sets; the difference in the validation set is always bigger than 0.3 MSE. Lastly, the *2D-CNN* performs the best for the presented problem. After around 50 training epochs, the network reaches its best performance with the validation loss function decreasing to around 0.1 MSE, where the loss functions on the training and validation sets align. Like the other two, this network starts showing sings of overtraining after around 70 training epochs. All three networks, especially the *2D-CNN*, show very high variability in their losses throughout the training. This could originate from a lack of training data.

Since the single difference between the *1D-* and *2D-CNN* is the position of the flattening layers, this comparison shows how valuable the convolution and pooling operations on the 3D space data in the presented problem are.

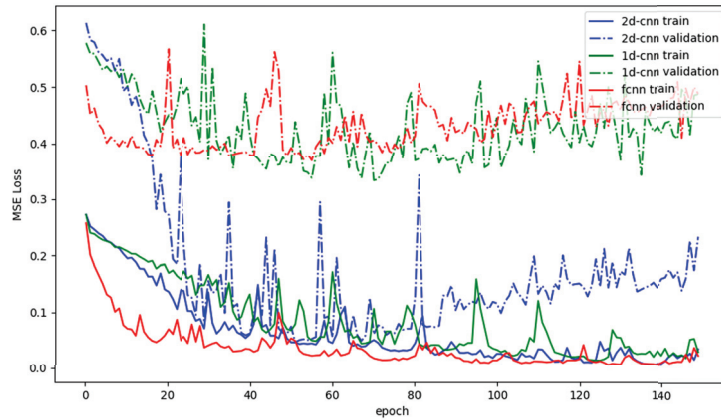


Figure 8. Mean-Squared-Error losses of the presented deep learning architectures. **Solid lines** denote the losses on the training set of the network, whereas the **dashed lines** denote the losses on a test set. **Red lines** show losses of the FCNN; **green lines** refer to the 1D-CNN; lastly, **blue lines** show the losses of the 2D-CNN.

3. Real-World Measurements

After focusing on the computational methods for data collection, this chapter focuses on real-world measurements and post-processing to refine the same deep-learning models as explained in Section 2.4.

3.1. Real-World Measuring Setup

Figure 9 left shows a system evaluation made in the past [7]. This system was able to follow the behavior of the fluid on its surface, although it offered no insight into sub-surface flow characteristics. To achieve this, the measurement system was extended in the NeoBio research project [13]. This updated node is shown in Figure 9 right. It introduces important features to extend the existing functionality. This version can vary its volume to mass proportions by moving a flexible membrane, enabling it to sink, come up to the fluid surface, and, in a *neutral setting*, follow the surrounding fluid's motion. Additionally, the fluid's conductivity is measured, which may provide further insights into fluid compositions and help specify fluid properties for computational analysis and simulation.

For the complete measuring setup, the node seen in Figure 9 right, is accompanied by anchor nodes mounted on the vessel side walls and a central data collection system, called *backbone*. The anchor nodes are shown in Figure 9 left.

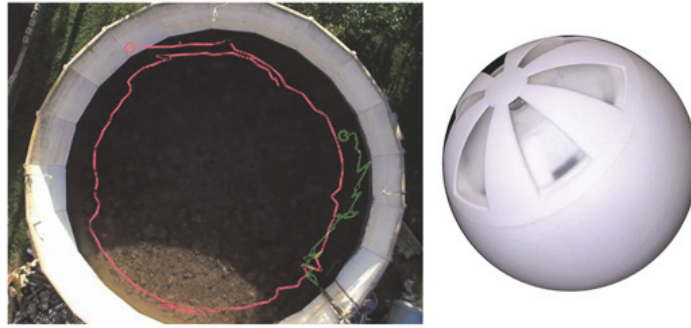


Figure 9. Left shows an overlay of a past measurement with a revision 1 of the measurement system on a top-down photo of the measurement in progress. Anchor nodes are shown on the vessel’s west-, north- and east-facing rims. Right shows an updated version of the sensor particle that can measure fluid behavior below the fluid’s surface.

Figure 10 outlines the real-world data collection methodology. The sensor follows the fluid movements by setting the sensors’ buoyancy to neutral, so it follows the surrounding fluid’s flow. While the sensor is submerged, the movement is tracked using an inertial measurement unit (IMU) that integrates accelerations into an absolute trajectory [14]. Since this method of movement tracking is only accurate for a short time, the sensor node will frequently increase its buoyancy, letting itself rise to the substrate surface. From there, the WSN locates its absolute position in the fermenter using ultra-wideband (UWB) localization [15]. Next, all gathered data from previous *dive* are sent to the backbone, again using the UWB interface. After all data are off-loaded to the backbone, the sensor node will return to measuring mode, restarting the cycle.

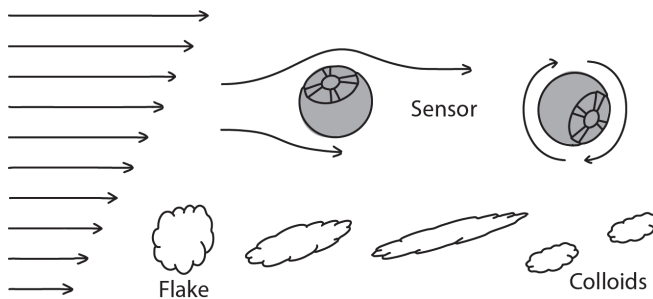


Figure 10. Shear-rate measurement concept utilizing the sensors’ inertial measurement unit and gyroscope, and the same shear-rate effect on biogas substrate flakes, which are forced to break up into colloids [13].

3.2. Real-World Measurement Post-Processing

To use gathered datasets for deep-learning (see Section 2.4), they have to be converted into the same format, as explained in Section 2.2. Most importantly, this includes calculating shear-rates along the sensor’s trajectory. In Section 2.2, the simulated measuring nodes were defined with no volume and mass, and thus did not have inertia. Since this cannot be achieved in the real world, it has to be taken into account.

Figure 11 top outlines the effect of fluid shear-rates on a real-world sensor flowing with a liquid stream. Arrows on the left denote liquid flow velocities, which are increasing from bottom to top. The difference in the fluid speeds on the sensor’s surface will cause it to spin. This spin can be detected by the sensor’s IMU, and gyroscope. Figure 10 bottom shows the effect of the same change in fluid velocities on flakes that is found in a biogas

substrate. These flakes are not rigid, like the sensor, and thus are elongated by the different forces applied to it until one flake breaks up into smaller colloids.

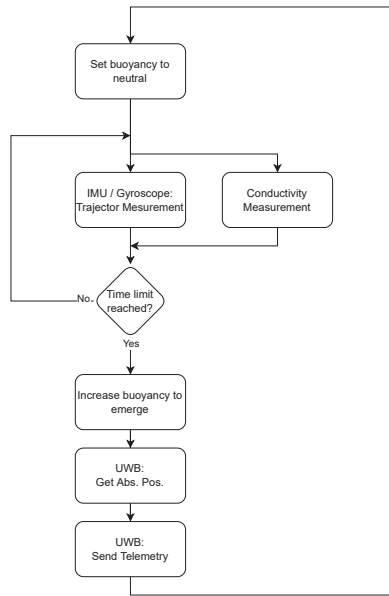


Figure 11. Real-World Measurement methodology to track fluid movement and substrate conductivity.

4. Conclusions and Outlook

This work presented a framework for setting up, extracting and pre-processing data from simulated as well as real-world biogas plants. In addition, three deep-learning models were presented that, based on the generated data, predict a biogas plant’s agitation efficiency via its shear-rate. The presented 2D-CNN is capable of predicting shear rates with a Mean-Squared Error (MSE) of less than 0.1 MSE, although all three models show signs of overtraining after 80 epochs. To gauge the accuracy of these systems in the real world, a framework for measuring these metrics in real-world environments was developed, although this could not be tested due to semiconductor shortages and the COVID-19 pandemic. Since the full range of biogas characteristics heavily depend on numerous factors, it is hard to specify these in a mathematical model for a Computational Fluid Dynamics (CFD) simulation. More research and a standardized model for a wide range of biogas plant setups will help to achieve results that can mimic real-world systems more closely.

Physics-Informed Neural Networks (PINNs) are a novel type of Deep-Learning setup that is specifically designed to predict physics problems and may increase the performance of the models presented here. In addition to PINNs, other features will be implemented using other deep-learning techniques. To further reduce the number of measurements required for each system, and increase the energy-efficiency of the real-world measuring system, *Time Series Forecasting* will be utilized to predict node trajectories.

Another problem to solve is the actual optimization of a real-world setup. Since the framework presented here only shows how well a system is performing, the optimization process is left to the user. A *Recommender System* will be implemented to solve this problem. This system will provide approaches to optimize agitation by, for example, suggesting optimal agitator settings or positions.

Author Contributions: Conceptualization, A.H., P.G., S.A., L.B. and S.R.; methodology, A.H.; software, A.H.; validation, A.H. and P.G.; formal analysis, A.H.; investigation, A.H.; resources, A.H.; data curation, A.H.; writing—original draft preparation, A.H.; writing—review and editing, A.H.; visualization, A.H.; supervision, P.G.; project administration, A.H.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Resources presented in this work are available at <https://git.fh-muenster.de/ah160996/sim-to-real-transfer-in-deep-learning-for-agitation-evaluation> (accessed on 8 May 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. BDEW. Share of Biomass in Gross Electricity Generation in Germany from 1991 to 2022. 2022. Available online: <https://de-statista-com.ezproxy.fh-muenster.de/statistik/daten/studie/251214/umfrage/anteil-der-biomasse-an-der-stromerzeugung-in-deutschland/> (accessed on 9 May 2023).
2. Annas, S.; Czajka, H.; Jantzen, H.; Janoske, U. Experimentelle und numerische Untersuchung der Strömungsvorgänge in einer Biogasanlage mit Paddelrührwerk. In Proceedings of the 7. Wissenschaftskongress Abfall- und Ressourcenwirtschaft, Tagungsbandbeitrag, Aachen, 16–17 March 2017; pp. 67–71.
3. Statista GmbH. Substrate Composition in Biogas Plants in Germany from 2010 to 2019. 2020. Available online: <https://de-statista-com.ezproxy.fh-muenster.de/statistik/daten/studie/198554/umfrage/anteil-des-substrateinsatzes-in-biogasanlagen/> (accessed on 13 May 2023).
4. FNR. Basisdaten Bioenergie Deutschland 2022. 2021. Available online: https://www.fnr.de/fileadmin/Projekte/2022/Mediathek/broschuere_basisdaten_bioenergie_2022_06_web.pdf (accessed on 3 July 2023).
5. Wang, B.; Björn, A.; Strömberg, S.; Nges, I.A.; Nistor, M.; Liu, J. Evaluating the influences of mixing strategies on the Biochemical Methane Potential test. *J. Environ. Manag.* **2017**, *185*, 54–59. ISSN 0301-4797. [CrossRef] [PubMed]
6. Landia POP Slurry Mixer: Landia a/s. Available online: https://www.landia.de/Files/Images/landia/dataark/Landia_Datenblatt_POP-I.pdf (accessed on 8 May 2023).
7. Heller, A.; Horsthemke, L.; Glösekötter, P. Design, Implementation, and Evaluation of a Real Time Localization System for the Optimization of Agitation Processes. In Proceedings of the 6th IFIP TC 10 International Embedded Systems Symposium (IESS 2019), Friedrichshafen, Germany, 9–11 September 2019; pp 39–50.
8. Conti, F.; Saidi, A.; Goldbrunner, M. Numeric Simulation-Based Analysis of the Mixing Process in Anaerobic Digesters of Biogas Plants. *Bioenergy X-Factor* **2022**, *43*, 1522–1529. [CrossRef]
9. OpenFOAM: Transport/Rheology Models. Available online: <https://doc.cfd.direct/openfoam/user-guide-v10/transport-rheology> (accessed on 6 February 2023).
10. OpenFOAM pimpleFoam Solver. Available online: <https://www.openfoam.com/documentation/guides/latest/doc/guide-applications-solvers-incompressible-pimpleFoam.html> (accessed on 9 May 2023).
11. Intel Core i5-8600k Hexacore Workstation Processor. Available online: <https://www.intel.com/content/www/us/en/products/sku/126685/intel-core-i58600k-processor-9m-cache-up-to-4-30-ghz/specifications.html> (accessed on 9 May 2023).
12. ParaView Post-Processing Visualization Engine. Available online: <https://www.paraview.org/> (accessed on 9 May 2023).
13. FH Münster laboratory for fluid dynamics, FH Münster laboratory for Semiconductors, FH Münster Laboratory for environmental engineering, HZDR innovation GmbH, Budelmann, Verbundvorhaben: Neue Entwicklungswerkzeuge zur Optimierung der Mischregime in Bioreaktoren; Teilvorhaben 2: Qualifizierung eines autonomen Sensorsystems zur Strömungs- und Mischcharakterisierung-Akronym: NEOBIO, Funding ID: 22032618. 2019. Available online: <https://www.fnr.de/index.php?id=11150&fkz=22032618> (accessed on 6 February 2023).
14. Buntkiel, L.; Reinecke, S.; Hampel, U. Richtungsaufgelöste Messung von Beschleunigungen mit Sensorpartikeln in industriellen Prozessbehältern. In Proceedings of the Dresdner Sensor-Symposium 2022, Dresden, Germany, 5–7 December 2022.
15. Buntkiel L.; Heller, A.; Budelmann, C.; Hampel, U. Mit UWB-Lokalisierung gekoppelte inertielle Lage- und Bewegungsverfolgung für instrumentierte Strömungsfolger. In Proceedings of the Dresdner Sensor-Symposium 2021, Dresden, Germany, 6–8 December 2021; pp. 22–27.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Modeling Contagion of Financial Markets: A GARCH-EVT Copula Approach [†]

Gueï Cyrille Okou ¹ and Amine Amar ^{2,*}

¹ LSTE Environmental Sciences and Technologies Laboratory, Jean Lorougnon Guede University, Daloa BP 150, Côte d'Ivoire; okou.guei.cyrille@gmail.com

² School of Science and Engineering, Al Akhawyn University, Ifrane 53000, Morocco

* Correspondence: a.amar@aui.ma

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: To better assess the financial contagion through the VaR, several recent studies used copula models. In the same context, this paper addresses the inefficiency of the classical approach such as a normal distribution in modeling the tail risk, by using the conditional Extreme Value Theory (GARCH-EVT), in order to assess extreme risks with contagion effect. The GARCH-EVT approach is a two-stage hybrid method that combines a Generalized Autoregressive Conditional Heteroskedasticity (GARCH) filter with the Extreme Value Theory (EVT). To implement our approach, we use macroeconomic time series from Morocco, Spain, France, and the USA.

Keywords: contagion effects; extreme value theory; GARCH-EVT; optimal tail selection; value at risk

1. Introduction

Financial resilience in banking is considered a key pillar when discussing the strength of the international financial system and the world economy as a whole. Indeed, financial resilience becomes more puzzling and worrying in the context of increasingly frequent, significant, and complex events. These extreme events include the stock market crash of 1929, the stock market crash of 1987, the sudden devaluation of the Mexican peso against the U.S. dollar in December 1994, the 1997 Asian financial crisis, and the global financial crisis between mid-2007 and early 2009. All these crashes are characterized by a subsequent rapid spread, significant severe losses incurred by financial institutions, spillovers, and high contagion risks. These events revealed substantial weaknesses in the banking system and the prudential framework and thus, motivated many of the managers and researchers, to recover existing tools and to implement new management strategies that offer significant improvement, by taking into consideration the increased severity, the high frequency of extreme events and spillover effects.

One important suggestion is to reconsider the Value-at-Risk (VaR); the widely used risk management tool, in the context of extreme events and contagion effects, which are nonlinear, time-varying, and dependent in nature.

The VaR can be defined as the maximum potential change in the value of a portfolio of financial instruments with a given probability over a certain horizon. There are several approaches for the estimation of VaR, such as historical simulation, variance-covariance, and the Monte Carlo approaches. In addition, contagion can be empirically identified through the propagation of extreme negative returns, the increase in interdependence compared to normal times, and the distinction from common shocks [1]. The literature contains various definitions of financial contagion [2]. However, financial contagion is present if a statistically significant increase is observed in cross-market correlation after the occurrence of extreme shocks [3].

Citation: Okou, G.C.; Amar, A. Modeling Contagion of Financial Markets: A GARCH-EVT Copula Approach. *Eng. Proc.* **2023**, *39*, 70. <https://doi.org/10.3390/engproc2023039070>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 10 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

To better assess the financial contagion through the VaR, several recent studies used copula models to describe the multivariate dependence structure between financial markets, estimate the return period, and assess the corresponding losses. In the same context, this paper addresses the inefficiency of the classical approach such as a normal distribution in modeling the tail risk, by using the conditional Extreme Value Theory (GARCH-EVT), in order to assess extreme risks with contagion effect. The GARCH-EVT approach is a two-stage hybrid method that combines a Generalized Autoregressive Conditional Heteroskedasticity (GARCH) filter with the Extreme Value Theory (EVT). The Peaks-Over-Threshold approach will be used for the pre-specification of the threshold that separates distribution tails from its middle part.

To implement our approach, we use time series retrieved by assessing the open-source records available on an international website. All statistical analyses were performed using R packages and our results provide important insights on risk management.

The remaining parts of the study are laid out as follows: Section 2 covers the research methodology and design, and results and findings are delineated in Section 3. Section 4 concludes the paper.

2. Materials and Methods

Our methodology is based on four main stages (Figure 1):

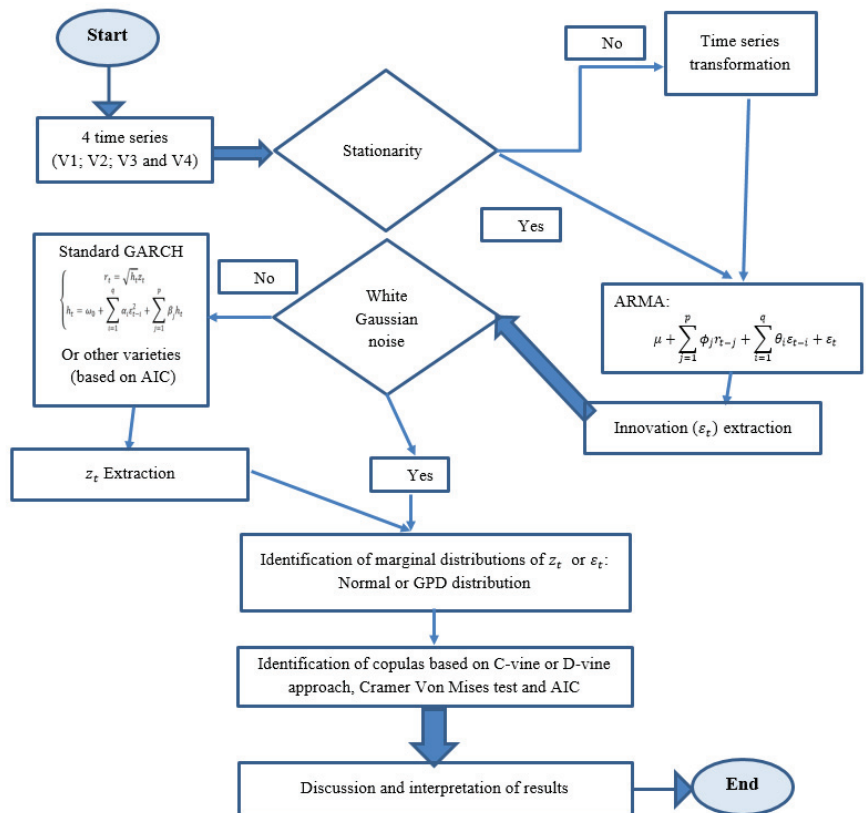


Figure 1. Proposed methodology flowchart.

Stage 1: Modeling time series with ARMA models. We first check whether the used times series are stationary or not, using visualization or analytical approach such as the Augmented Dickey-Fuller test (ADF). When the stationarity is not accepted, transformation

is needed and thus, an ARMA model is identified [4]. A mixed autoregressive moving average process of order (p, q) process is a stationary process $\{Y_t\}$ which satisfies the relation:

$$r_t = \mu + \sum_{j=1}^p \phi_j r_{t-j} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t, \tag{1}$$

where $\phi_j, j = 1, 2, \dots, p, \theta_i, i = 1, 2, \dots, q$ are parameters of the ARMA model to be estimated.

Stage 2: Innovations extraction and Gaussian white noise assumption checking. Innovations ε_t are extracted from the ARMA model and the Ljung and Box portmanteau test is used to examine if ε_t can be considered Gaussian White Noise.

Stage 3: Use of GARCH and identification of marginal distributions. If the assumption of the Gaussian White Noise is not validated, a standard sGARCH model or other GARCH varieties, such as GJR-GARCH [5], are identified, and then, z_t are extracted. The dynamics of the conditional volatility of the GARCH(p, q) model are given by:

$$\begin{cases} r_t = \sqrt{h_t} z_t \\ h_t = \omega_0 + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^p \beta_j h_{t-j} \end{cases} \tag{2}$$

The GJR-GARCH is given by

$$\begin{cases} r_t = \sqrt{h_t} z_t \\ h_t = \omega_0 + \sum_{i=1}^q (\alpha_i + \chi_i I(\varepsilon_{t-i} < 0)) \varepsilon_{t-i}^2 + \sum_{j=1}^p \beta_j h_{t-j} \end{cases} \tag{3}$$

where z_t is normalized white noise and h_t is the conditional variance of the innovation ε_t , ω is the intercept, and the parameters α_i, χ_i and β_j are the autoregressive coefficients of the variance. Marginal distribution is identified for ε_t and z_t .

Stage 4: Copulas fitting based on C-Vine and D-vine approach. The main idea of vine copulas is the modeling of copulas in high dimensions, based on a structure of interconnected trees of bivariate copula. This construction approach makes it possible to model complex dependencies in high dimensions by bivariate copulas [6].

The main issue in the field of financial contagion is to analyze the underlying process and to emphasize the main variables that could indicate a financial crisis in a country. This has led researchers to broaden the scope of the investigation and thus, several categories such as common shocks, trade spillovers, and financial linkages are identified. In this context, the empirical studies suggested by Račickas and Vasiliauskaitė [7] present a number of determinants of a financial crisis. It is worth noting that some explanatory variables are exclusive for currency crises, banking crises, or debt crises; others are informative for more than one type of crisis.

To implement our approach, we identify four time-series associated with financial contagion from the World Development Indicators (WDI) website. More details are provided in the following table (Table 1).

Table 1. Used data.

	Variables	Unit	Countries	Covered Period
V1	GDP per capita growth	Annual %	France	1970–2021
V2	Trade	% of GDP	Morocco	
V3	Inflation, consumer prices	Annual %	Spain	
V4	Exports as a capacity to import	Constant LCU	US	

3. Results and Discussions

A descriptive analysis shows that there is a strong interconnection between Morocco and Spain. These two neighboring partners are linked by more than 16 billion euros of

trade and; Morocco is the third economic partner of Spain outside the EU. In addition, Spanish exports to Morocco have increased by 29% in 2020/2021, 17,000 Spanish companies have trade relations with Morocco and 700 are established in the neighboring country. It is also worth noting the increase in the range of Moroccan exports to Spain in recent years, reflecting the modernization of the national productive fabric.

France remains one of Morocco’s leading economic partners, despite growing competition in the areas of trade and investment. The relationship between the two countries makes France the first partner of Morocco on the level of commercial exchange, tourist arrivals, and direct investments.

A detailed analysis, from Figure 2, shows a simultaneous trend in terms of inflation, exports, trade, and inflation in Morocco, France, Spain, and the US. In addition, the trade and the exports time series exhibit an upward trend, while the GDP time series are more or less stable, with the exception of Morocco for which more fluctuations are noted.

The analysis of inflation leads to discerning two distinguishable periods. The first one, before 1985, was characterized by high levels of inflation (% annual) with a maximum of 17.6%, 13.6%, 13.5%, and 24.5% for Morocco, France, the US, and Spain, respectively. The second period is characterized by controlled inflation.

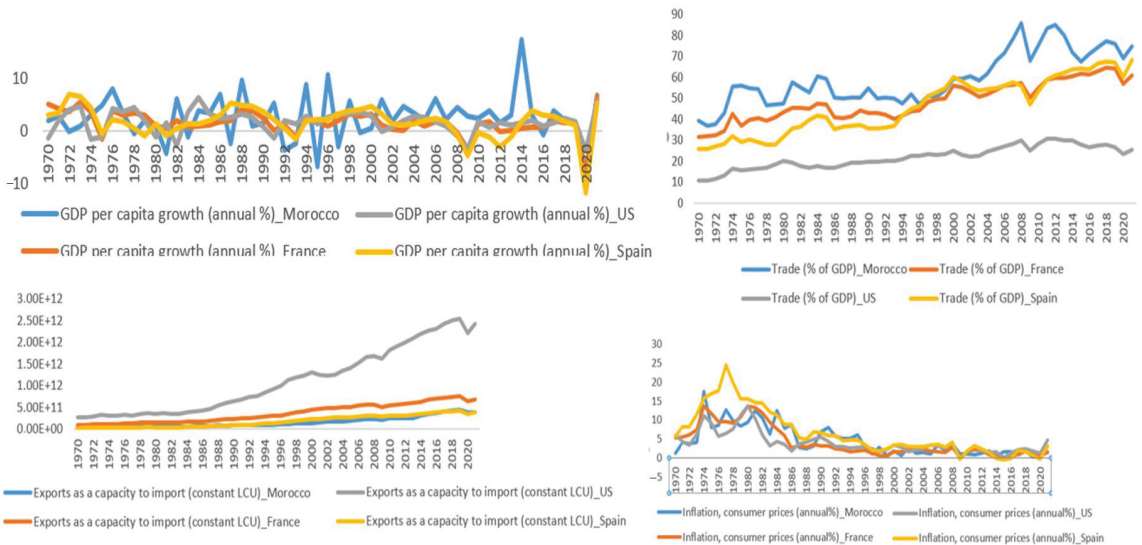


Figure 2. Trends comparison in terms of inflation, exports, trade, and GDP.

The analysis of the inflation leads to discerning two distinguishable periods. The first one, before 1985, was characterized by high levels of inflation (% annual) with a maximum of 17.6%, 13.6%, 13.5%, and 24.5% for Morocco, France, the US, and Spain, respectively. The second period is characterized by controlled inflation.

It is worth noting that all computations are performed using R software. To implement our methodology, we transform the raw data to have stationary time series. Results show also that most fitted models are ARMA(1,1) which is characterized by autocorrelation functions that decline dramatically and ARMA(1,0) which predicts the present value of a time series, using the immediately prior value in time.

Results (Table 2) indicate that most of the analyzed time-series in this study have a non-homogenous variance, so there is a GARCH effect. Except for the US time series, the assumption of Gaussian white noise is not satisfied, thus SGARCH or GJR-GARCH are estimated for those time-series.

Table 2. ARMA and GARCH modeling.

Indicators		Time Series	Model
V1	France	Raw	ARMA (1,1) $\mu = 0; \varphi = -0.03; \theta = -0.99$
		Residuals	GJR-GARCH(1,1): $\alpha_0 = 0.31; \alpha = 0.43; \beta = 0.94; \gamma = 0.76$
	Morocco	Raw	ARMA (1,0) $\mu = 0; \varphi = -0.72$
		Residuals	BB Gaussian
Spain	Raw	ARMA (1,1) $\mu = 0; \varphi = 0.45; \theta = -1$	
	Residuals	sGARCH(1,1): $\alpha_0 = 0; \alpha = 0.31; \beta = 0.68$	
	US	Raw	ARMA (1,1) $\mu = -0.01; \varphi = 0.11; \theta = -0.99$
	Residuals	BB Gaussian	
V2	France	Raw	ARMA (1,1) $\mu = 4.53; \varphi = -0.17; \theta = -0.99$
		Residuals	GJR-GARCH(1,1): $\alpha_0 = 1.67; \alpha = 0.017; \beta = 0.99; \gamma = 0.18$
	Morocco	Raw	ARMA (1,1) $\mu = 1.19; \varphi = 0.74; \theta = -0.99$
		Residuals	BB Gaussian
Spain	Raw	$\sqrt{h_t}z_t : \mu = 0; \varphi = 0; \theta = 0$	
	Residuals	sGARCH(1,1): $\alpha_0 = 0; \alpha = 0; \beta = 0.888$	
	US	Raw	ARMA (1,0) $\mu = 1.69; \varphi = 0.05$
	Residuals	BB Gaussian	
V3	France	Raw	ARMA (1,0) $\mu = -41.13; \varphi = -0.22$
		Residuals	GJR-GARCH(1,1): $\alpha_0 = 932.71; \alpha = 0.98; \beta = 0; \gamma = 0.04$
	Morocco	Raw	ARMA (1,0) $\mu = -1.14; \varphi = -0.56$
		Residuals	BB Gaussian
Spain	Raw	ARMA (1,1) $\mu = -2.77; \varphi = -0.90; \theta = 1$	
	Residuals	sGARCH(1,1): $\alpha_0 = 0; \alpha = 0; \beta = 0.83$	
	US	Raw	ARMA (1,1) $\mu = -0.01; \varphi = -0.44; \theta = 0.76$
	Residuals	BB Gaussian	
V4	France	Raw	ARMA (1,1) $\mu = 11.73; \varphi = 0.99; \theta = -0.94$
		Residuals	GJR-GARCH(1,1): $\alpha_0 = 2.93; \alpha = 0; \beta = 0.99; \gamma = 0.26$
	Morocco	Raw	ARMA (1,1) $\mu = 5.22; \varphi = 0.56; \theta = -0.99$
		Residuals	BB Gaussian
Spain	Raw	$\mu + \sqrt{h_t}z_t : \mu = 13.78; \varphi = 0; \theta = 0$	
	Residuals	GJR-GARCH(1,1): $\alpha_0 = 3.54; \alpha = 0.21; \beta = 0.62; \gamma = 0.32$	
	US	Raw	ARMA (1,1) $\mu = 10.57; \varphi = -0.79; \theta = 1$
	Residuals	GJR-GARCH(1,1): $\alpha_0 = 2.66; \alpha = 0; \beta = 1; \gamma = 0.26$	

Once the innovations are extracted, marginal distributions are identified and copulas are fitted. It is worth noting that Copulas are fitted with two marginal; the Normal and the Generalized Pareto distribution (GPD) and by using C-Vines and D-vines approaches.

Tables 3 and 4 present the multivariate dependencies between the different retained indicators (Trade, GDP, Inflation, and Export) among countries (Morocco, Spain, France, and the US), using D-Vines and C-Vines copulas. From these results, we can have a clear idea about the different structures of the dependencies, and thus contagion mechanisms, such as the dependence between exportation in Morocco and exportation in the US given information on exportations in Spain and France (Copula $(U_{\text{Export Morocco}}; U_{\text{Export US}}/U_{\text{Export France}}; U_{\text{Export Spain}})$). For this example, the identified Copula is the survival of Clayton Copula. The structure of dependence between the exportations in France and the exportations in Morocco, given the information on the exportations in Spain (Copula $(U_{\text{Export France}}; U_{\text{Export Morocco}}/U_{\text{Export Spain}})$) is as Gumbel Copulas. Both Gumbel and Survival Clayton are considered as extreme value Copulas. The consequence of these findings is that the impact

of the contagion is remarkable at the extremes, characterized by subsequent rapid spread, significant severe losses, spillovers, and high contagion risks.

Table 3. Copulas fitting (marginals are considered Normal distributions).

Copulas	Designation (D-Vines)	Copulas Family	Designation (D-Vines)	Copulas Family
c ₃₂	$U_{\text{Export Spain}}; U_{\text{Export Morocco}}$	t	$U_{\text{GDP Spain}}; U_{\text{GDP Morocco}}$	t
c ₁₃	$U_{\text{Export France}}; U_{\text{Export Spain}}$	t	$U_{\text{GDP France}}; U_{\text{GDP Spain}}$	t
c ₄₁	$U_{\text{Export US}}; U_{\text{Export France}}$	Normal	$U_{\text{GDP US}}; U_{\text{GDP France}}$	Survival Gumbel
c _{12,3}	$U_{\text{Export France}}; U_{\text{Export Morocco}} / U_{\text{Export Spain}}$	Gumbel	$U_{\text{GDP Morocco}}; U_{\text{GDP Spain}}$	Rotated Gumbel 90 degrees
c _{43,1}	$U_{\text{Export US}}; U_{\text{Export Spain}} / U_{\text{Export France}}$	Rotated Clayton 270 degrees	$U_{\text{GDP France}}; U_{\text{GDP US}}; U_{\text{GDP Spain}} / U_{\text{GDP France}}$	Joe
c _{24,13}	$U_{\text{Export Morocco}}; U_{\text{Export US}} / U_{\text{Export France}}; U_{\text{Export Spain}}$	Survival Clayton	$U_{\text{GDP Morocco}}; U_{\text{GDP US}} / U_{\text{GDP France}}; U_{\text{GDP Spain}}$	Rotated Clayton 270 degree
Copulas	Designation (D-vines)	Copulas Family	Designation (C-vines)	Copulas Family
c ₃₂	$U_{\text{Inflation Spain}}; U_{\text{Inflation Morocco}}$	Normal	c ₁₂ : $U_{\text{trade France}}; U_{\text{trade Morocco}}$	Normal
c ₁₃	$U_{\text{Inflation France}}; U_{\text{Inflation Spain}}$	Frank	c ₁₃ : $U_{\text{trade France}}; U_{\text{trade Spain}}$	Survival Gumbel
c ₄₁	$U_{\text{Inflation US}}; U_{\text{Inflation France}}$	Gaussian	c ₁₄ : $U_{\text{trade France}}; U_{\text{trade US}}$	Survival BB7
c _{12,3}	$U_{\text{Inflation France}}; U_{\text{Inflation Morocco}} / U_{\text{Inflation Spain}}$	Frank	c _{24,1} : $U_{\text{trade Morocco}}; U_{\text{trade US}} / U_{\text{trade France}}$	Survival Clayton
c _{43,1}	$U_{\text{Inflation US}}; U_{\text{Inflation Spain}} / U_{\text{Inflation France}}$	t	c _{34,1} : $U_{\text{trade Spain}}; U_{\text{trade US}} / U_{\text{trade France}}$	Survival Clayton
c _{24,13}	$U_{\text{Inflation France}}; U_{\text{Inflation Morocco}}; U_{\text{Inflation US}} / U_{\text{Inflation Spain}}$	Frank	c _{23,14} : $U_{\text{trade Morocco}}; U_{\text{trade Spain}} / U_{\text{trade France}}; U_{\text{trade US}}$	Rotated Joe 90 degree

Table 4. Copulas fitting (marginals are considered GPD).

Copulas	Designation (C-Vines)	Copulas Family	Designation (C-Vines)	Copulas Family
c ₃₁	$U_{\text{Export Morocco}}; U_{\text{Export France}}$	t	c ₂₁ : $U_{\text{Inflation Spain}}; U_{\text{Inflation France}}$	Joe
c ₃₂	$U_{\text{Export Morocco}}; U_{\text{Export Spain}}$	t	c ₃₂ : $U_{\text{Inflation US}}; U_{\text{Inflation Spain}}$	Joe
c _{21,3}	$U_{\text{Export Spain}}; U_{\text{Export France}} / U_{\text{Export Morocco}}$	Survival Joe	c _{31,2} : $U_{\text{Inflation US}}; U_{\text{Inflation France}} / U_{\text{Inflation Spain}}$	Survival Gumbel
c _{2,1}	$U_{\text{GDP Spain}}; U_{\text{GDP France}}$	Survival BB7	c _{1,2} : $U_{\text{trade France}}; U_{\text{trade Spain}}$	Joe
c _{3,2}	$U_{\text{GDP US}}; U_{\text{GDP Spain}}$	Joe	c _{3,1} : $U_{\text{trade US}}; U_{\text{trade France}}$	Survival BB7
c _{31,2}	$U_{\text{GDP US}}; U_{\text{GDP France}} / U_{\text{GDP Spain}}$	Joe	c _{32,1} : $U_{\text{trade US}}; U_{\text{trade Spain}} / U_{\text{trade France}}$	Joe

For Inflation, it is worth noting that most fitted Copulas are Gaussian Copulas, which means that the structure of the dependence is not strong at the extremes such as different crises. This is because the inflation is controlled by central banks that have different visions and implement different adequate policies to maintain the inflation controlled.

4. Conclusions

The dependence on financial markets during a period of extreme fluctuations has received considerable attention within the literature. In the same context, the main contribution of this work is to understand the structure of dependence between different pertinent variables that can be used, to explain the contagion of financial Markets. Financial contagion can be defined as the spread of an economic crisis from one market or region to another so, events in one market can affect other markets.

In this study, the used methodology is based on the GARCH-EVT Copula approach. Copula modeling is a popular tool for analyzing the dependencies between variables.

It allows the investigation of tail dependencies and the specification of models for the marginal distributions separately from the dependence structure and more specifically, the Copula co-movements capture how shocks in a particular market may transcend to other currency markets. These implications are of particular interest in risk, survival applications, and prediction of financial contagion. More recently, there are different empirical applications of Copula-based methods in economics, due to the flexibility of the approach and the gain in terms of the computational complexity of estimation. Our findings starkly highlight the adequacy of two copulas. The first one; the Normal Copula, is appropriate for the inflation while the second is suitable for trade, exportations, and the GDP. It is worth noting that the normal copula provides a general linear form of the dependence and captures a general form of the dependence, while the survival Clayton Copula characterizes the dependence in the extremes.

Author Contributions: G.C.O.: Conceptualization, methodology, software, visualization, investigation, writing—review and editing; A.A.: Conceptualization, validation, formal analysis, data curation, writing—original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data supporting reported results can be found: WDI: <https://databank.worldbank.org/source/world-development-indicators> (accessed on 1 January 2023); UNDP: http://hdr.undp.org/sites/default/files/mpi2022_technical_notes.pdf (accessed on 1 January 2023); OPHI: <https://ophi.org.uk/publications/mpi-methodological-notes/> for details on how the Multidimensional Poverty Index is calculated (accessed on 1 January 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. *Financial Stability Review Report*; European Central Bank: Frankfurt, Germany, 2005.
2. Davidson, S.N. Interdependence or contagion: A model switching approach with a focus on Latin America. *Econ. Model.* **2020**, *85*, 166–197. [CrossRef]
3. Forbes, K.J.; Rigobon, R. No contagion, only interdependence: Measuring stock market co-movements. *J. Finance* **2002**, *57*, 2223–2261. [CrossRef]
4. Harvey, A.C. *Forecasting, Structural Time Series Models and the Kalman Filter*; Cambridge University Press: Cambridge, UK, 1990. [CrossRef]
5. Glosten, L.R.; Ravi, J.; David, E.R. On the relation between the expected value and the volatility of the nominal excess return on stocks. *J. Finance* **1993**, *48*, 1779–1801. [CrossRef]
6. Aas, K.; Claudia, C.; Arnoldo, F.; Henrik, B. Pair-copula constructions of multiple dependence. *Insur. Math. Econ.* **2009**, *44*, 182–198. [CrossRef]
7. Račickas, E.; Asta, V. Classification of financial crises their occurrence frequency in global financial markets. *Soc. Tyrim.* **2012**, *4*, 32–44.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Evaluation of Heuristics for Taken's Theorem Hyper-Parameters Optimization in Time Series Forecasting Tasks †

Rodrigo Hernandez-Mazariegos ¹, Jose Ortiz-Bejar ^{2,*} and Jesus Ortiz-Bejar ¹

¹ Facultad de Ciencias Físico-Matemáticas “Mat. Luis Manuel Rivera Gutiérrez”, UMSNH, Avenida Universidad 100, Villa Universidad, 58060 Morelia, Michoacán, Mexico; 1301441a@umich.mx (R.H.-M.); jesus.ortiz@umich.mx (J.O.-B.)

² División de Estudios de Posgrado de la Facultad de Ingeniería Eléctrica, UMSNH, Building “Ω2” Ciudad Universitaria, Francisco J. Múgica S/N, 58030 Morelia, Michoacán, Mexico

* Correspondence: jose.ortiz@umich.mx

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: This study compares three methods for optimizing the hyper-parameters m (embedding dimension) and τ (time delay) from Taken's Theorem for time-series forecasting to train a Support Vector Regression system (SVR). Firstly, we use a method which utilizes Mutual Information for optimizing τ and a technique referred to as “Dimension Congruence” to optimize m . Secondly, we employ a grid search and random search, combined with a cross-validation scheme, to optimize m and τ hyper-parameters. Lastly, various real-world time series are used to analyze the three proposed strategies.

Keywords: Taken's Theorem; time-series; SVR forecasting; mutual information; dimension congruence; random search; grid search

1. Introduction

Several complex phenomena are often modeled as a sequence of states. This sequence is known as the phase space. A time series is a finite sequence of states in a dynamical system measured directly or indirectly. A relevant approach to perform time series analysis is Taken's embedding theorem [1], which states that, from a sequence of states $S = \{y_{t_1}, y_{t_2}, \dots, y_{t_n}\}$ (i.e., time series) in a dynamical system, it is possible to generate all the system's phase space U . More specifically, for a sequence of observations x of dimension m (embedding dimension) and a constant τ (time delay), there exists a function f such as:

$$y(t) = f(x) = f[y(t - \tau), y(t - 2\tau), \dots, y(t - (m - 1)\tau)] \quad (1)$$

From Equation (1) it can be inferred that, given a time series S , it is possible to predict the state at time t (hereafter, y_t) by using m previous observations sampled at frequency τ . The two problems, and their solutions, are the following: (1) the function f is often too complex to be found analytically, which is when machine learning algorithms comes into play with the objective of using a supervised learning algorithm to learn f ; (2) it is necessary to find the correct modeling for the time series, i.e., the optimal values for m and τ , for which Random search, Grid search, and Mutual information + Dimension Congruence can be used.

2. Theoretical Background

Given Equation (1), the first task is to find the optimal value for the time delay τ and embedding dimension m .

Citation: Hernández-Mazariegos, R.; Ortiz-Bejar, J.; Ortiz-Bejar, J. Evaluation of Heuristics for Taken's Theorem Hyper-Parameters Optimization in Time Series Forecasting Tasks. *Eng. Proc.* **2023**, *39*, 71. <https://doi.org/10.3390/engproc2023039071>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 10 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

2.1. Mutual Information

Regarding the τ , in [2] Cao, L. proposes using mutual information. The process relies on making $y(t)$ and $y(t - \tau)$ as independent as possible to maximize the information obtained from each variable in the reconstruction of the phase space. To achieve this, the *mutual information function* (2) can be applied:

$$I_\tau = \sum_{\Omega} P(N_{i+\tau}|N_i) \ln \left(\frac{P(N_{i+\tau}|N_i)}{P(N_{i+\tau})P(N_i)} \right) \tag{2}$$

Note the similarity of this with entropy, i.e., this function measures how surprising it is that $N_{i+\tau}$ results, given that N_i resulted, i.e., when $N_{i+\tau}$ and N_i are very independent, then $I_\tau \approx 0$. To find the τ it is, therefore, enough to minimize the function (2).

Once the τ is fixed, it is necessary to find the embedding dimension m . The latter is achieved by using the false neighbors [2] to determine the dimension congruence.

2.2. Dimension Congruence

The aim of this procedure is for the distances between neighbors (data close to each another) on dimension m of Equation (1) to be constant. To this end, firstly, the distance $E(i, j, m)$ between $y(t)$ and $y(t')$ in the dimension m is defined as the maximum between the differences of their components, as in Equation (3):

$$E(t, t', m) = \max_{(k,l) \in [0, m-1]} |y(t - k\tau) - y(t' - l\tau)| \tag{3}$$

Now, we can say that the nearest neighbor of $y(t)$ is $y(t')$ if t' is satisfied such that:

$$E(t, t', m) = \min_{t'' \in [0, n-m\tau], t'' \neq t} E(t, t'', m) \tag{4}$$

where n is the sample size, it is worth mentioning that t' depends on t so we call it $t'(t)$ and then we define the “nearest neighbor” congruence of $y(t)$ in m as:

$$F(t, m) = \frac{E(t, t'(t), m)}{E(t, t'(t), m + 1)} \tag{5}$$

Note that in $F(t, m) \approx 1$ if $y(t'(t))$ is sufficiently congruent being the nearest neighbor of $y(t)$ in m , there is the possibility to define the “dimension congruence” of m as follows:

$$G(m) = \frac{1}{n - m\tau} \sum_{t \in [0, n-m\tau]} F(t, m) \tag{6}$$

In summary, the dimension congruence measures how true it is that the nearest neighbors continue to be nearest neighbors as the dimension increases, which is useful, given the assumption that there is an attractor in the system under study [2].

In this work, m was selected as lower m satisfying $G(m) > 0.95$. As alternative strategies to find m and τ , Evolutionary computation algorithms, Random Search(RS) and Grid Search (GS) can be used. In this paper we focus on comparing Mutual Information + Dimension Congruence with RS and GS, given [3], which states that random search is good enough for parameter optimization.

2.3. Random Search

Let f be a model that depends on a parameter λ . The random search method [3] involves defining a range (a_0, a_1) for λ , a probability distribution $g : (0, 1) \rightarrow (a_0, a_1)$, and the number of values to be tested. Then, n parameters $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n$ are drawn from the distribution g , and the behavior of each of the corresponding models $f_{\lambda_1}, f_{\lambda_2}, f_{\lambda_3}, \dots, f_{\lambda_n}$

is evaluated by computing a fitness function. The best-performing model, f_{λ_i} , is selected based on the fitness value.

2.4. Grid Search

In contrast with random search, the grid search method [4] involves sampling λ values equally spaced in the range (a_0, a_1) , specifically, $\lambda_1 = a_0 + \frac{1}{n}$, $\lambda_2 = a_0 + \frac{2}{n}$, $\lambda_3 = a_0 + \frac{3}{n}, \dots, \lambda_n = a_0 + \frac{n}{n} = a_1$.

From Figure 1 we can appreciate the differences between random search and grid search.

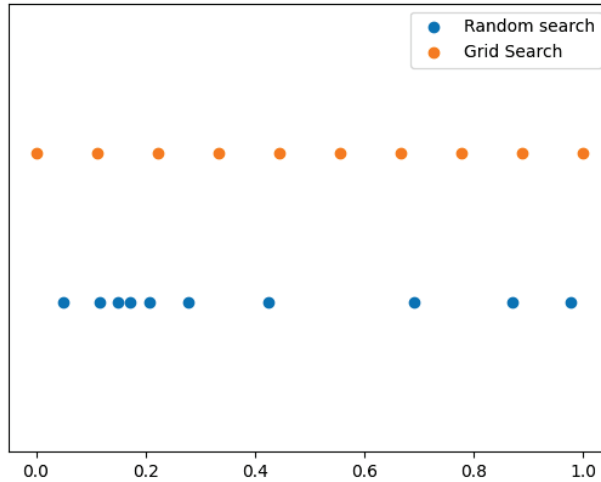


Figure 1. Differences between Random Search and Grid Search.

Having described the procedures for finding m and τ , it is time to describe the fitness measures used.

2.5. Fitness Function

The Mean Absolute Percentage Error (MAPE) is the fitness function determining the optimal value for parameters m and τ . Additionally, the Mean Squared Error (MSE) and the Coefficient of Determination (R^2) were used to compare the three optimization procedures with MAPE. For completeness, a brief description of each one is provided.

2.5.1. MAPE

MAPE [5] is a widely used measure in time series forecasting and seems to yield good results. Note that in Equation (7) MAPE takes the average of the absolute value of the errors expressed as a percentage of the actual value, and if it approaches 0 the better the fit, while if it approaches ∞ the worse the fit.

$$MAPE = \frac{1}{n - e} \sum_{i=e+1}^n \left| \frac{N_i - \hat{N}_i}{N_i} \right| \tag{7}$$

2.5.2. MSE

Ref. [6] is the average of the squared errors. If the model fits perfectly then $MSE = 0$. The closer it is to ∞ the worse it is. It is computed by using Equation (8):

$$MSE = \frac{1}{n - e} \sum_{i=e+1}^n (N_i - \hat{N}_i)^2 \tag{8}$$

2.5.3. R^2

Ref. [7] calculates the ratio between the model’s variance and the actual data’s variance. In other words, it ascertains how similar the predicted and actual data variances are. If they are equal, R^2 is equal to 1, which means the model fits perfectly. The worse value for R^2 is $-\infty$. To find R^2 Equation (9) is used:

$$R^2 = 1 - \frac{\sum_{i=e+1}^n (N_i - \hat{N}_i)^2}{\sum_{i=e+1}^n (N_i - \bar{N})^2}, \quad \bar{N} = \frac{1}{n - e} \sum_{i=e+1}^n N_i \quad (9)$$

Finally, in the next section we describe the models we used for the experiments.

2.6. Support Vector Regression Algorithm (SVR)

The SVR algorithm is based on Support Vector Machine (SVM) Algorithm [8]. SVM is an algorithm for separating samples depending on the class they belong to. The algorithm works by increasing the size of the sample space via a kernel, and in that larger size, three parallel hyperplanes are constructed, separated by an ϵ distance each. The main idea is to optimize the kernel and the hyperplanes so that only a small number of samples, controlled by the parameter ζ , are outside the region to which they belong, i.e., the samples of one class belong to one side of the hypertube and those of the other class belong to the other side of the hypertube.

On the other hand, SVR aims for all the samples to be inside the hypertube and only a small amount of samples to be outside the hypertube, controlled by the parameter ζ , and, thus, uses the image of the central hyperplane projected in the original space to predict future values of the time series.

- \mathbb{S} , \mathbb{H}_ϵ and $\mathbb{H}_{-\epsilon}$ are the parallel hyperplanes.
- \mathbb{H}_ϵ is located at a distance ϵ above \mathbb{S} .
- $\mathbb{H}_{-\epsilon}$ is located at a distance ϵ below \mathbb{S} .
- Then \mathbb{H}_ϵ and $\mathbb{H}_{-\epsilon}$ form the hypertube
- The quantity $\frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^n \|\vec{\zeta}_i\|$ is the minimum possible, subject to $|N'_i - \vec{w} \cdot \vec{x}_i| < \|\vec{\epsilon}\| + \|\vec{\zeta}_i\|$

Where \vec{w} is the SVR kernel weight and $\vec{\zeta}_i$ is the distance that the i -th data moves away from the hypertube, while C is a regularization parameter.

Note how the larger c is, the less freedom the data have to move out of the hypertube. The idea is to find a hypertube that approximates the data.

3. Experiments and Results

The study focused on six real-world time series, each representing a measurement of a real-world phenomenon. The aim was to examine more complex time series than artificially generated ones. The selected time series displayed a wide range of characteristics, including exponential and moderate growth patterns, general trends, and horizontal patterns. The goal was to evaluate the generalization ability of the proposed methodologies by considering time series with diverse characteristics. Below is a brief description of each of the time series used.

3.1. SARS-CoV-2 in Mexico (COV)

The time series data for this study was obtained from the General Direction of Epidemiology (<https://www.gob.mx/salud/documentos/datos-abiertos-152127> (accessed on 19 December 2022)). It consists of the confirmed and suspected COVID-19 cases in Mexico. The data spans 1025 days, and the number of confirmed cases per laboratory ranges from 0 to 9800. The data were normalized such that the number of cases fell from 0 to 1. For the purposes of this study, this time series is referred to as *COV*. Figure 2a depicts the evolution of the *COV* time series.

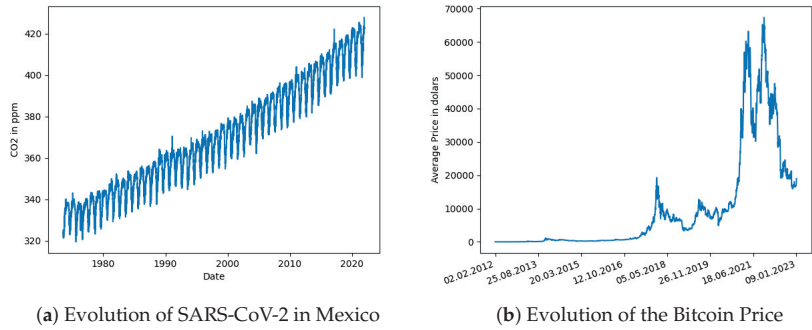


Figure 2. Covid and Bitcoin.

3.1.1. Bitcoin Price on Bitfinex (BIT)

This time series comprises daily variations in the price of Bitcoin in dollars, recorded on the Bitfinex platform between February 2012 and January 2023 (The data is available at <https://www.investing.com/crypto/bitcoin/btc-usd-historical-data> (accessed on 10 February 2023)). The dataset includes the daily high and low prices. The average price is computed as $(minimumPrice + maximumPrice)/2$. The time series were normalized between 0 and 1 for consistent analysis with the other time series used in this study. Figure 2b displays the evolution of the BIT time series.

3.1.2. Air Temperature in Acuitzio del Canje (TEM)

This time series consists of temperature data recorded by the MXN00016001 weather station located in Acuitzio del Canje between 2004 and 2007 (The data was obtained from <https://www.nci.noaa.gov/> (accessed on 15 January 2023)). The dataset comprises 1401 data points of daily minimum and maximum temperatures. The average temperature is calculated as $(T_{max} + T_{min})/2$. The data is recorded in degrees Fahrenheit and was subsequently normalized between 0 and 1 for comparison with other time series in this study. The evolution of the TEM time series is depicted in Figure 3a.

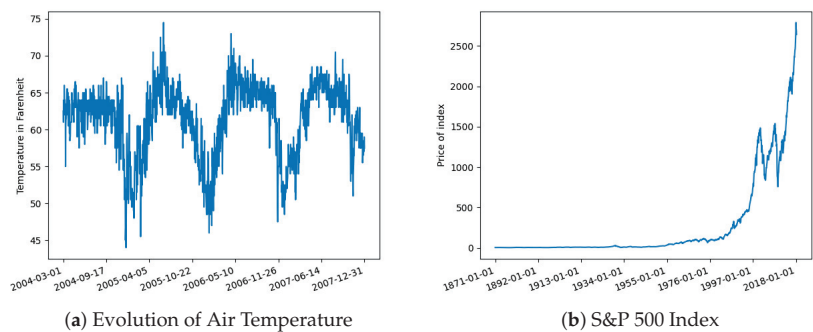


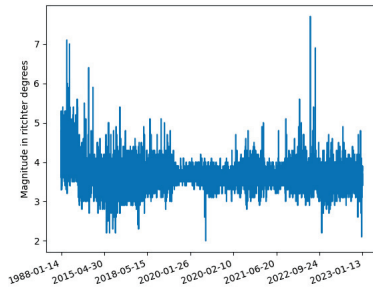
Figure 3. Temperature and S&P.

3.1.3. S&P500 Index

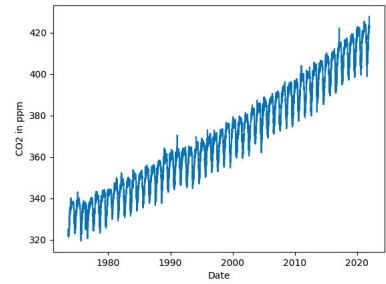
The S&P500 index series (Data was sourced from: <https://datahub.io/core/s-and-p-500> (accessed on 11 February 2023)) is a monthly measurement of the value of the S&P500 stock index, which represents the 500 most valuable companies in the United States. It consists of 1768 monthly value data points calculated from 1871 to 2018. The data was normalized for analysis between the range of (0, 1). Figure 3b shows the evolution of the S&P500 index graphically.

3.1.4. Seismic Activity in Michoacán

This series comprises seismic activity recorded by the National Seismological System in Michoacán (The time series is available at the following URL: <http://www2.ssn.unam.mx:8080/catalogo/> (accessed on 22 January 2023)). The values cover the period from 1988 to 2023. This time series was of interest as the data was not evenly spaced. One possibility was to summarize the data to create an indicator to identify “how active each month was”. However, for our study, the original sampling frequency was maintained. Each event is a numerical value representing its magnitude in Richter scale degrees and consists of 17,500 data points, which were normalized between (0, 1). Figure 4a graphically depicts these data.



(a) Seismic Activity in Michoacán



(b) Carbon Dioxide Concentration.

Figure 4. Seismicity and CO₂.

3.1.5. Atmospheric Carbon Dioxide Concentration

This is a series of daily atmospheric carbon dioxide (CO₂) concentration measurements taken at the Barrow Atmospheric Baseline Observatory (Data obtained from <https://www.co2.earth/daily-co2> (accessed on 9 February 2023)) in the United States. The CO₂ concentrations are reported in parts per million (ppm) and cover the period from 1973 to 2021. To facilitate the analysis and interpretation of the data, all the values were normalized to the range of (0, 1).

Figure 4b shows this time series.

3.2. Experimental Setup

For each of the analyzed time series, the parameters m, τ , and the parameters C and ϵ for the SVR were optimized. Three strategies were used: mutual information + congruence, grid search, and random search. The flow diagram in Figure 5 summarizes the process.

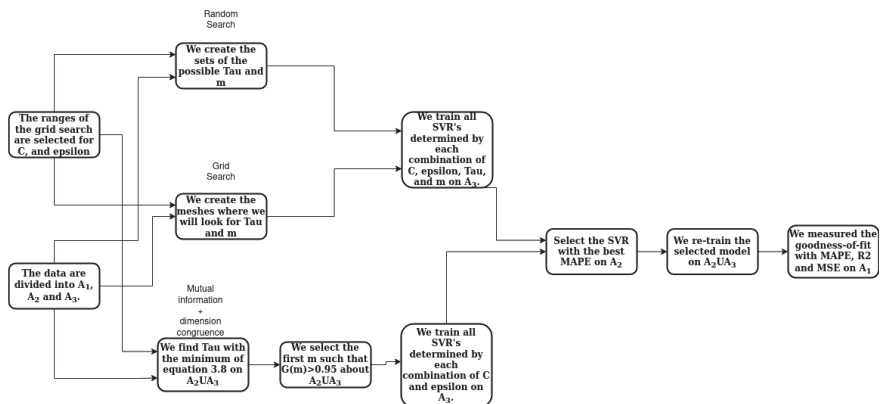


Figure 5. Flow diagram of each procedure applied to each time series.

The diagram in Figure 5 illustrates the general overview of the three procedures applied to each time series. It is noteworthy that the final outcome for each time series were nine goodness of fit measures. These measures were then used to compare the procedures. Before diving into the specifics of each process, it is essential to consider a few key points.

The data were divided into three sets:

- Set \mathbb{A}_1 contained the last 5% of the data to be used for testing and calculating the model's fitness.
- Set \mathbb{A}_2 contained the last 5% of the data once set \mathbb{A}_1 had been removed, to be used for hyper-parameter tuning.
- Set \mathbb{A}_3 consisted of the remaining data to be used for training the parameters.

All of the models used were Support Vector Regression (SVR), and for the SVR C and ϵ hyper-parameters, the following applied:

- Grid search was used to find the C and ϵ for all SVR models.
- The grid for C values was in $\mathbb{C} = [0.1, 1, 10, 100]$.
- The grid for ϵ values was $\mathbb{E} = [0.001, 0.01, 0.1, 1]$.

For RS and GS of τ and m , the following conditions were met:

- The sets $\mathbb{T}_i, \mathbb{M}_i$ contained all possible values of τ and m for each time series, and each procedure (Random Search and Grid Search) had 20 elements (for computational capacity reasons).
- The infimum of these sets was always 2.
- The supremum was always $\text{int}(\sqrt{|\Omega|} - 10)$ (so that $m * \tau < |\Omega|$).
- The distribution used for the random search was always uniform.

With this in mind, the procedures used to search m (dimension of the reconstructed phase space) and τ (delay) were:

- Mutual information + dimension congruence:
 1. Find τ using the mutual information function from Equation (2) on $\mathbb{A}_2 \cup \mathbb{A}_3$, and take the minimum.
 2. Find the embedding dimension by selecting the first m that satisfies $G(m) > 0.95$ in Equation (6) with the obtained τ on $\mathbb{A}_2 \cup \mathbb{A}_3$.
 3. Train all possible SVRs determined by the elements of $\mathbb{C} \times \mathbb{E}$ on \mathbb{A}_3 .
 4. Select the model having MAPE on \mathbb{A}_2 which is the minimum.
 5. Measure the goodness of the selected model using MAPE on \mathbb{A}_1 .
- Random search and grid search:
 1. Use each element of $\mathbb{C} \times \mathbb{E} \times \mathbb{T}_i \times \mathbb{M}_i$ to train $|\mathbb{C}| |\mathbb{E}| |\mathbb{T}_i| |\mathbb{M}_i| = 4 * 4 * 20 * 20 = 6400$ models on \mathbb{A}_3 .
 2. Select the model having MAPE on \mathbb{A}_2 which is the minimum.
 3. Measure the goodness of the selected model using MAPE on \mathbb{A}_1 .

It is worth mentioning that the parameter space in both the grid search and the random search was not very large due to the lack of hardware. It is to be expected that enlarging the size of these spaces would improve the results.

Upon completion of the procedures, a comparison was made by evaluating the distributions generated by each of the fitness measures obtained by each proposed method.

3.3. Results

Table 1 shows the results for the three metrics. The MAPE, R^2 evaluation, and MSE, best results are boldfaced. For instance, in the series "BIT" GS was the best procedure with respect to MAPE, but also with respect to R^2 and with respect to MSE. As you can see, there was no procedure that was always better than another. There were some series where RS and IC were better than RS. However, it is essential to clarify that even though GS had better results, it is a brute force algorithm, in that, although it has better optimizations, the computational cost is too high (RS and GS take in the order of hours, while IC takes in

the order of minutes, on a i9 7th generation). From the results, it is recommended to work with IC to optimize the τ and m parameters, and, for the regression system parameters, to use RS. It is relevant to point out that IC is the fastest, while it provides a competitive prediction performance.

Table 1. Quality measurements for each time series made with each one of the proposed optimization strategies.

Series	MAPE-RS	MAPE-GS	MAPE-IC	R ² -RS	R ² -GS	R ² -IC	MSE-RS	MSE-GS	MSE-IC
COV	3.6714	8.9901	5.6877	-3.4295	-41.4618	-11.4179	0.0005	0.00511	0.0014
BIT	0.228	0.1566	0.2656	-3.016	-1.026	-4.383	0.0046	0.0023	0.0061
TEM	0.265	0.1973	0.02007	-2.537	-0.3787	-0.4936	0.0283	0.011	0.0119
S&P	0.525	0.508	0.5368	-5.642	-5.336	-5.851	0.162	0.152	0.165
TEL	0.1447	0.1395	0.14087	-0.1595	-0.1061	-0.1011	0.001895	0.001808	0.001800
CO2	0.0526	0.0491	0.0619	0.2866	0.3706	-0.0578	0.0027	0.0024	0.004

Figure 6a–c suggests that GS had better results, both in its mean and dispersion. However, if we look only at IC and RS we observe that when one of the two had a better mean, it would also have worse dispersion, which indicates that some series work very well with IC and others with RS, but, in general, it is a good idea to try both methods.

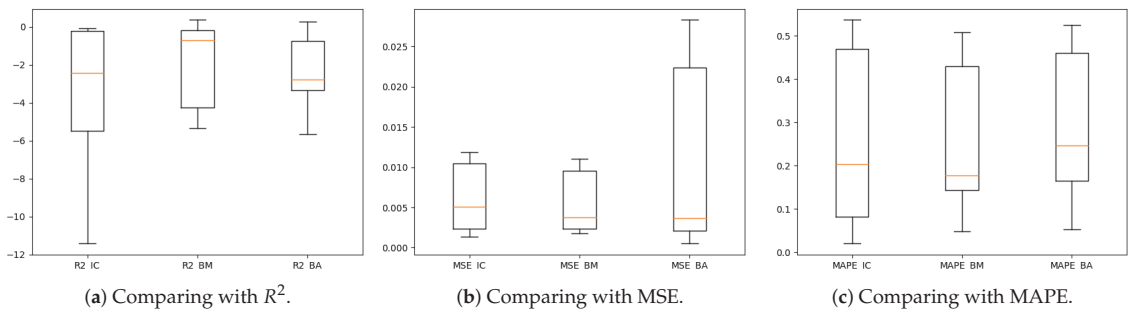


Figure 6. Boxplots comparing each procedure with different goodness-of-fit measures.

3.4. Future Work

It remains for future work to evaluate the procedures with additional quality measurements. Selecting the model with Mean Squared Logarithmic Error (MSLE) could improve the predictions. Including the regression system in the optimization could improve the prediction performance, by, for instance, using Naïve Bayes and K-Nearest Neighbor systems.

Author Contributions: R.H.-M.: Conceptualization, formal analysis, investigation writing original draft preparation; J.O.-B. (Jose Ortiz-Bejar): Conceptualization, supervision, writing and proofreading; J.O.-B. (Jesus Ortiz-Bejar): Review, formal analysis and proofreading. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found at the URLs mentioned in Section 3 of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Takens, F. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*, Warwick; Springer: Berlin/Heidelberg, Germany, 1980; pp. 366–381.
2. Cao, L. Practical method for determining the minimum embedding dimension of a scalar time series. *Phys. D Nonlinear Phenom.* **1997**, *110*, 43–50. [CrossRef]
3. Bergstra, J.; Bengio, Y. *Random Search for Hyper-Parameter Optimization*; Universite de Montreal: Montreal, QC, Canada, 2012. Available online: <https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf> (accessed on 10 May 2023).
4. Yu, S.; Pritchard, M.; Ma, P.L.; Singh, B.; Silva, S. Two-step hyperparameter optimization method: Accelerating hyperparameter search by using a fraction of a training dataset. *arXiv* **2023**, arXiv:2302.03845.
5. De Myttenaere, A.; Golden, B.; Le Gr, B.; Rossi, F. Mean Absolute Percentage Error for regression models. *Neurocomputing* **2016**, *192*, 38–48. [CrossRef]
6. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*, 2nd ed.; Springer: New York, NY, USA, 1998; ISBN 978-0-387-98502-2.
7. Yin, P.; Fan, X. Estimating R^2 Shrinkage in Multiple Regression: A Comparison of Different Analytical Methods. *J. Exp. Educ.* **2001**, *69*, 203–224. [CrossRef]
8. Rivas-Perea, P.; Cota-Ruiz, J.; Chaparro, D.G.; Venzor, J.A.P.; Carreón, A.Q.; Rosiles, J.G. Support Vector Machines for Regression: A Succinct Review of Large-Scale and Linear Programming Formulations. *Int. J. Intell. Sci.* **2013**, *3*, 5–14. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Simulation of the Queuing Situation of Patients at a Health Center [†]

Kalle Saastamoinen ^{*}, Antti Rissanen, Juho Suni, Juho Hyttinen, Petteri Paakkunainen and Aaro Liakka

Department of Military Technology, National Defence University, P.O. Box 7, 00861 Helsinki, Finland; antti.rissanen@mil.fi (A.R.); juho.suni@mil.fi (J.S.); juho.hyttinen@mil.fi (J.H.); petteri.paakkunainen@mil.fi (P.P.); aaro.liakka@mil.fi (A.L.)

^{*} Correspondence: kalle.saastamoinen@mil.fi

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: At the starting point of this case study, a garrison hospital performed an assessment of the need for treatment when the number of conscripts queuing at reception is at its highest level. The research aims to find out the reasons for conscripts' perceived long waiting times, which causes absence from the conscripts' training. According to the predictions made by the queuing simulation, the hospital's staff are able to receive patients arriving at reception in the morning without the queue time causing undue harm to training. However, during large congestion peaks, the waiting times may become unreasonable, which would require an increase in human resources. Peaks of congestion usually occur at the beginning of the week, as well as on days with heavy military training.

Keywords: hospital; conscript; queue; simulation

1. Introduction

The modelling of health care resources is especially important after a time of crisis, which is the case for example after a pandemic. Indeed, COVID-19 has had a significant impact on health care in Finland. According to the Finnish Institute for Health and Welfare (THL), the number of hospital visits due to COVID-19 increased a lot between March and April 2020 [1]. In addition, there was an increase in waiting times for specialized health care ([2], Figure 7). The pandemic also put a strain on personnel resources within the health care system, with many of the staff reallocated from other departments or working long hours to cope with demand.

In the defense forces, the general service regulations (YLPALVO) define the following for those performing health care services: "The defense forces are responsible for organizing the health care of conscripts in service, women performing voluntary military service and students being trained for military service, as well as those participating in voluntary exercises and supervised shootings of the defense forces and training ordered from the National Defense Training Association. The health care of the Defense Forces is open health care by nature. Nurse and doctor appointments are arranged according to local special conditions, needs, and resources. The goal is to use health care methods to secure the service safety of the persons under the care of the Defense Forces. A conscript in service has the right to medical examination by a healthcare professional without unnecessary delay. Dental care is organized in accordance with local needs and conditions, either as the Defense Forces' own activity or as a purchased service." ([3], pp. 42–43).

Here, we use a simulation to show the effects of alternating queuing conditions for health center operational status. In order to establish our simulation model, we used queueing theory, which is the mathematical study of queues [4]. This study is also part of operations research since we use the simulation results to make decisions about the resources needed to provide a service.

Citation: Saastamoinen, K.; Rissanen, A.; Suni, J.; Hyttinen, J.; Paakkunainen, P.; Liakka, A. Simulation of the Queuing Situation of Patients at a Health Center. *Eng. Proc.* **2023**, *39*, 72. <https://doi.org/10.3390/engproc2023039072>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 11 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Queueing theory is important because it helps to identify and analyze the queues that form in many different systems. It provides insights into how to manage queues, as well as how to allocate resources so that serving customers is effective with minimal waiting times. Queueing theory has applications in many areas such as healthcare, telecommunications networks, airport check-in, computer processing, traffic control systems, manufacturing processes, etc. If one understands the principles of queueing theory, one can make better decisions when it comes to resource allocation and customer service management. Earlier, we used queueing theory in the Finnish Defence Forces (FDF), for example, in the modelling of custom inspections in [5] and in the tactical warfare simulator SANDIS, which uses Markov chains [6].

The main factors of queueing theory are [7]:

1. Arrival rate—the rate at which customers or jobs enter the system.
2. Number of service channels—the number of servers or machines available to process requests from customers in the queue.
3. Queue discipline—the serving order of customers, such as first come, first served (FIFO) or last in, first out (LIFO).
4. Service rate—the service rate of customers or jobs.
5. Queue length—the number of customers waiting in line to be served by the service facility.
6. Utilization—a measure of how busy a resource is, typically expressed as a percentage of its capacity (e.g., 80% utilization).
7. Waiting time—the amount of time that elapses between when a customer arrives and when he/she is served by the service facility.
8. Throughput—the average number of customers or jobs processed per unit of time (e.g., per hour).

In medical services, we use queue theory to help optimize the efficiency of patient care. To achieve this, we manage waiting times, prioritize patients based on their need for treatment, and analyze how changes in staffing or scheduling affect the overall efficiency of a facility [8]. Queue theory helps healthcare providers balance demand with resources by helping them understand how many beds they need, when they should schedule staff shifts, and where they should allocate additional resources. Queue theory can also forecast future patient loads so that facilities can plan accordingly.

This work deals with the use of the health care services of a garrison health center by those performing their military service from the point of view of the time spent waiting in line. We investigate the capacity of the health center and the effect of “peak congestion” on service times. We find out what would be the optimal number of reception points, without the need for resources changing significantly. The aim is to model the current situation as accurately as possible, from the perspective of how long a conscript waits for access to receive treatment in normal situations and during peak traffic.

2. Problem Description

We investigated the ability of the Finnish garrison hospital of the Kainuu brigade to perform an assessment of the need for medical treatment when the number of conscripts (max. 600) queuing at reception is at its highest level and the ratio of patients to nurses is at its maximum level. The aim of the study is to illustrate with which nurse capacity the duration of treatment queues could be at such a level that all conscripts queuing at reception in the morning would be able to return to their basic unit service during the morning. The starting point of the review is how to serve conscripts registered for the morning health check-up within the 120-min time window. The purpose of the work is to investigate the capacity of the health center’s queuing services and the effect of “traffic peaks” on service times. We are looking for an answer to the question of what the optimal number of reception points would be without the need for resources changing significantly. We model how long a conscript waits for access to the evaluation of the need for treatment in normal situations and during peak traffic.

Here, we studied patients from the moment they sign up for the morning reception. The observed queue begins to form from the moment when the conscript reports their need for treatment and reaches the nurse's preliminary check-up and the doctor's check-up if it is necessary. The reviewed queue ends when a medical professional has assessed the person's state of health. The research aims to find ways to reduce the time spent in the queue. The morning reception starts at 07:00. In terms of the smooth running of conscript training, persons capable of service must return to service within two hours. We are investigating whether it is possible to serve all conscripts within 120 min: how many conscripts can we examine and what number of nursing staff does this require?

3. Model and Results

The queuing mechanics of the garrison hospital can be thought of as a single queue with several servants and stages. The behavior of the queue can be described using certain expectation values. The parameters needed for the calculation are the number of customers joining the queue per time unit and the average number of customers served per time unit. The subject under review is the Kainuu brigade and the healthcare services provided by its health center. The service regulations of the Kainuu brigade dictate the following when registering for the morning reception of basic units:

- To register for the morning reception, the conscript reports to the unit's duty officer after waking up.
- The health center calls through the units and collects the preliminary strengths of those who come to the reception by unit.
- Conscripts registered for the morning reception do not take part in the day's service and wait in their rooms until the unit's personnel are ordered to the health center for their turn.
- Those who have signed up for the reception of the first three basic units of the breakfast shift of the Kainuu brigade will go directly from breakfast to the reception.
- Other units are called to the reception systematically, and the health center schedules their arrival.

The queue model is defined here by the following five rules:

- A: How to join the queue? Since all conscripts aiming for the morning reception start queuing when they report to the duty officer of their unit immediately after being woken up, in queue modeling, everyone starts queuing at the same moment. The method of joining the queue does not involve a statistical distribution and therefore does not affect the modeling values. In the modeling, it has been assumed that the customers are in a queue together and they are served at the next free service point in order of arrival. There is therefore no queue for each service point.
- B: How to exit the queue? You leave the queue after the treatment time is over. In the modeling of the queue, the treatment duration is normally distributed so that the average is 2.0 min and the standard deviation is 0.5 min.
- C: Number of service channels. This study examines the garrison hospital's ability to process patients in a given time, where the analysis involves changing the number of service places to achieve the most optimal outcome.
- D: Maximum queue length. The maximum length of the queue has been defined as 600 people because according to experts, this is the longest queue formed at the KAIPR garrison hospital.
- E: Queuing principle. The queue at the garrison hospital works with the first in, first out (FIFO) method, i.e., the first customer in the queue is also the first to leave the service point. The model does not take into account the fact that the customer could leave the queue in the middle of queuing, as this is practically not possible for on-duty personnel. It is also not possible to skip the queue, and acute cases do not even enter this queue.

Here, it is not meaningful to calculate the load factor of the queue because every customer starts queuing at the same time. After the customer joins the queue, they are taken care of anyway. The purpose of the modeling is to find out by what means the total duration of the queue can be made as short as possible.

In the simulation model in Figure 1, the “waiting time” column shows when the conscript can get to reception. The model takes into account the fact that the first conscripts get to the nurse immediately, without queuing. Half a minute has been added to the waiting time after each person on duty so that the nurse has time to invite a new patient in before the start of the treatment period. In addition, the breaks taken by the nurses have been taken into account by adding 15 min at 60, 120, and 180 min for each service point. At present, a maximum of 15 nurses can be provided to the health center, and in the worst case, all 500 conscripts would feel sick.

Garrison hospital queue simulation model

New simulation is calculated when you change the table or press the F9 key

Service duration in minutes (normal distribution)	
Average	4
Deviation	1

Summary (600 patients)	
How many have to wait	585
Average waiting time	106.26
The longest wait	223.0
Had to wait more than a minut	585

Instruction
 When adding or removing reception points, the waiting time must be 0.5 min added to the waiting time, because the service does not start immediately when the reception point is free.

Conscript	Waiting time	Treatment duration	Ready	Service point is free																
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
1	0.0	2.2	2.2	2.2	0.0															
2	0.0	4.6	4.6	2.2	4.6	0.0														
3	0.0	2.7	2.7	2.2	4.6	2.7	0.0													
4	0.0	3.7	3.7	2.2	4.6	2.7	3.7	0.0												
5	0.0	3.4	3.4	2.2	4.6	2.7	3.7	3.4	0.0											
6	0.0	4.4	4.4	2.2	4.6	2.7	3.7	3.4	4.4	0.0										
7	0.0	4.7	4.7	2.2	4.6	2.7	3.7	3.4	4.4	4.7	0.0									
8	0.0	4.6	4.6	2.2	4.6	2.7	3.7	3.4	4.4	4.7	4.6	0.0								
9	0.0	3.4	3.4	2.2	4.6	2.7	3.7	3.4	4.4	4.7	4.6	3.4	0.0							
10	0.0	4.4	4.4	2.2	4.6	2.7	3.7	3.4	4.4	4.7	4.6	3.4	4.4	0.0						
11	0.0	4.1	4.1	2.2	4.6	2.7	3.7	3.4	4.4	4.7	4.6	3.4	4.4	4.1	0.0					
12	0.0	4.2	4.2	2.2	4.6	2.7	3.7	3.4	4.4	4.7	4.6	3.4	4.4	4.1	4.2	0.0				
13	0.0	3.2	3.2	2.2	4.6	2.7	3.7	3.4	4.4	4.7	4.6	3.4	4.4	4.1	4.2	3.2	0.0			
14	0.0	4.8	4.8	2.2	4.6	2.7	3.7	3.4	4.4	4.7	4.6	3.4	4.4	4.1	4.2	3.2	4.8	0.0		
15	0.0	1.5	1.5	2.2	4.6	2.7	3.7	3.4	4.4	4.7	4.6	3.4	4.4	4.1	4.2	3.2	4.8	1.5		
16	2.0	3.7	5.8	2.2	4.6	2.7	3.7	3.4	4.4	4.7	4.6	3.4	4.4	4.1	4.2	3.2	4.8	5.8		
17	2.7	3.2	5.9	5.9	4.6	2.7	3.7	3.4	4.4	4.7	4.6	3.4	4.4	4.1	4.2	3.2	4.8	5.8		
18	3.2	4.6	7.8	5.9	4.6	7.8	3.7	3.4	4.4	4.7	4.6	3.4	4.4	4.1	4.2	3.2	4.8	5.8		
19	3.7	3.7	7.4	5.9	4.6	7.8	3.7	3.4	4.4	4.7	4.6	3.4	4.4	4.1	4.2	7.4	4.8	5.8		
20	3.9	3.4	7.3	5.9	4.6	7.8	3.7	3.4	4.4	4.7	4.6	7.3	4.4	4.1	4.2	7.4	4.8	5.8		
21	3.9	3.5	7.4	5.9	4.6	7.8	3.7	7.4	4.4	4.7	4.6	7.3	4.4	4.1	4.2	7.4	4.8	5.8		

Figure 1. An Illustration of the simulation model of the garrison hospital.

4. Results

Table 1 reflects the current efficiency and personnel situation of the garrison hospital. The research wanted to examine the actual maximum strengths, so field nurses have also been included in the nurse strength from the enhanced strength onwards.

Table 1. The largest possible number of patients with the current nursing staff.

	Normal Strength (3–5 Nurses)	Enhanced Strength (10 Nurses)	Maximum Strength (15 Nurses)
Max patients (2 h)	70–115	230	350
Max patients (4 h)	130–220	430	645

Table 2 shows the number of nurses needed at different times. If the need for treatment is assessed in two hours, the conscript still has time to participate in afternoon training. If the assessment takes more than four hours, the conscript does not have time to participate in training for the whole day.

Table 2. The required number of nurses in different situations.

	Minimum Patients (10)	Normal Patients (50)	Large Patients (200)	Maximum Patients (550)
Required nurses (2 h)	1	3	9	24
Required nurses (4 h)	1	2	5	13

5. Discussion and Future

Most of the garrisons in Finland, like the Kajaani brigade considered here, are large entities comprising several basic units centrally served by the garrison health station. Thus, we can conclude that the modeling carried out here gives a sufficiently accurate overview for identifying problem points. According to the results of the research, the hospital's staff are able to receive patients arriving for the morning reception without the waiting time causing undue harm to training. This means an absence of less than two hours from training due to queuing. This is realized when the total number of arriving patients is a maximum of approximately 250 patients. At peaks of congestion, where the number of patients can rise to more than 500 patients, the capacity of the garrison health center is not sufficient to keep the waiting times reasonable. This would require an increase in human resources. Congestion peaks occur especially after returning from holidays and on days with heavy outdoor training. The research aimed to observe the reasons for conscripts' perceived long waiting times, which partly cause unnecessary absences from training. Based on the research, it is possible to examine the current system and evaluate its ability to cope with the number of customers. However, the research does not provide practical answers as to how to correct the observed problems. This is because health care is very carefully regulated both by law and by the Defence Forces' own regulations. Consequently, the effects of all changes must be evaluated from the perspective of the patient's legal protection, for example. This modeling clearly supports the existence of the problem. Solutions to the problem should be explored with further research. This new research should focus on how congestion peaks in particular could be smoothed out by referring non-urgent patients to other days based on advance reservations.

Author Contributions: Conceptualization, K.S.; methodology, K.S.; software, K.S.; validation, J.S., J.H., P.P. and A.L.; formal analysis, J.S., J.H., P.P. and A.L.; investigation, J.S., J.H., P.P. and A.L.; resources, J.S., J.H., P.P. and A.L.; data curation, J.S., J.H., P.P. and A.L.; writing—original draft preparation, K.S.; writing—review and editing, K.S. and A.R.; visualization, K.S.; supervision, K.S.; project administration, K.S.; funding acquisition, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: We have used only anonymous data during this research.

Data Availability Statement: Data is available through Kalle Saastamoinen, kalle.saastamoinen@mil.fi.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Open Data-THL. Finnish Institute for Health and Welfare. Available online: <https://thl.fi/en/web/thlfi-en/statistics-and-data/data-and-services/open-data> (accessed on 9 February 2023).
2. Tiirinki, H.; Tynkkynen, L.K.; Sovala, M.; Atkins, S.; Koivusalo, M.; Rautiainen, P.; Jormanainen, V.; Keskimäki, I. COVID-19 pandemic in Finland—Preliminary analysis on health system response and economic consequences. *Health Policy Technol.* **2020**, *9*, 649–662. [CrossRef] [PubMed]
3. Finnish Defence Forces. *Yleinen Palvelusohjesääntö (YLPALVO)*; Halonen, P., Karvinen, I., Eds.; Puolustusvoimat; 2017; Available online: https://puolustusvoimat.fi/documents/1948673/2258487/PEVIESTOS_YLPALVO+2017/3684dac2-c7ac-4d93-b792-34649f6e2f5d/PEVIESTOS_YLPALVO+2017.pdf (accessed on 13 February 2023).
4. Sundarapandian, V. *Probability, Statistics and Queuing Theory*; PHI Learning Pvt. Ltd.: New Delhi, India, 2009.

5. Saastamoinen, K.; Mattila, P.; Rissanen, A. A Simulation of a Custom Inspection in the Airport. In *Theory and Applications of Time Series Analysis: Selected Contributions from ITISE 2018*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; Volume 5, pp. 319–330. [CrossRef]
6. Lappi, E. Computational methods for tactical simulations. Ph.D. Thesis, Maanpuolustuskorkeakoulu, Helsinki, Finland, 2012.
7. Gross, D. *Fundamentals of Queueing Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
8. Fomundam, S.; Herrmann, J.W. A Survey of Queuing Theory Applications in Healthcare. In *Institute for Systems Research Technical Reports*; A. James Clark School of Engineering at the University of Maryland: College Park, MD, USA, 2007.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Development of Methodology for the Evaluation of Solar Energy through Hybrid Models for the Energy Sector [†]

Georgina González-González ^{1,2,*}, Jesús Cerezo-Román ^{1,*} and Guillermo Satamaria-Bonfil ³

¹ Center for Engineering and Applied Sciences, Autonomous University of State Morelos, Cuernavaca 62209, Mexico

² Laboratory Technician School, Autonomous University of State Morelos, Cuernavaca 62209, Mexico

³ Data Portfolio Manager Department, Unique Experience and Data General Directorate, Banco Bilbao Vizcaya Argentaria, Mexico City 06600, Mexico; guillermo.santamaria@bbva.com

* Correspondence: georgina.glgz@uaem.edu.mx (G.G.-G.); jesus.cerezo@uaem.mx (J.C.-R.)

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: The forecast of the generation of electrical energy from the solar resource is associated with its uncertainty due to the meteorological variations that it presents. Solar power generation forecasts are important for the efficient operation of solar plants. This article shows a methodology entailing a multilayer neural network with backpropagation and input data from a model with time lag coordinates for a horizon of 24 h and beyond. The neural network model was compared with statistical and prediction models numerical time, resulting in a MAPE of 0.57% and a MAE of 69.29 W.

Keywords: forecasting; power energy; neural network

1. Introduction

Currently, economic development has increased disproportionately, causing greater energy demand throughout the world, putting the supply and demand of it at risk. To satisfy the need for energy, sources of conventional origin have been exploited; however, these compromise the health of living beings and the environment, which is why the use of sources of renewable origin with a low carbon ratio has been proposed [1].

Photovoltaics is an affordable, free, and easily accessible energy type that has proven to be a clean renewable source and is found in abundance almost everywhere in the world. Its use has increased in recent years, being incorporated into the energy repertoire in different parts of the world [2]. In 2021, fifty countries generated a tenth of their electricity from renewable sources, with photovoltaic energy standing out. In 2020 there were only 43 countries and in 2019 there were 36 [3], which indicates that more and more countries are betting on the development of research in the use of photovoltaic energy; however, this brings with it particular challenges posed by the intermittent origin of such renewable energies, such as intermittency depending on their availability and variability [4].

The use of photovoltaic energy has been one of the topics of interest as a research objective in recent years. This is due to the growth of the clean energy industry and the commitments obtained at the United Nations Conference on Climate Change, the latter seeking the use of energy with a low carbon ratio [5], in addition to the increasing meteorological events that have directly affected the generation of electric power [6].

Large-scale photovoltaic power plants present difficulties in the management of the solar resource due to their intermittency, affecting the system of connection to the network, storage, and distribution; it is, therefore, necessary to protect the system from such adversities, which is why the search for more accurate and precise forecasts of photovoltaic energy is the development area of this work [7].

Citation: González-González, G.; Cerezo-Román, J.; Satamaria-Bonfil, G. Development of Methodology for the Evaluation of Solar Energy through Hybrid Models for the Energy Sector. *Eng. Proc.* **2023**, *39*, 73. <https://doi.org/10.3390/engproc2023039073>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 11 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

There are different methods to develop the prediction of electrical energy from renewables, such as: statistical models, Numerical Weather Prediction (NWP), Artificial Intelligence (AI), and hybrid models [8–10]. Each of the previous models has their best use and their respective areas for improvement.

Statistical models are based on the history of the data, that is, from past observations characteristics are obtained that help predict future data through the minimization of errors [11,12]. This approach depends on the quality of the data and their pre-processing; among the most used are: Autoregressive Models (AR), exponential smoothing model, Autoregressive Models and Moving Averages (ARMA), Autoregressive Models Integrated with Moving Averages (ARIMA), Autoregressive Models Integrated with Moving Averages with Seasonality (SARIMA) [13–15]. The models based on NWP are based on the physical-mathematical phenomena of meteorological and geological origin through atmospheric parameters, with their use enabling understanding of the current state of the atmosphere. The data are extracted through satellite stations and soil measurement devices, with the measurement instruments requiring constant monitoring and calibration [16,17]. On the other hand, models based on Artificial Intelligence have been widely used in recent years, since they allow the prediction of stochastic data so that photovoltaic and wind energy present such behavior in their observations [18], thus becoming a method that allows its development and improves its performance. In the case of hybrid methods, they are used to obtain the best qualities of each of the previously described methods and improve their performance [19,20].

Photovoltaic (PV) power forecasting is characterized by two types of models according to the time scale: ultrashort for data from seconds to hours and short term for next day observations [21]. The first model is used in real time, while the second is for planning the next day. Currently there are investigations that have been developed with the aim of providing a good prediction of photovoltaic energy, models have been proposed based on statistical methodologies, such as the case of models that use a SARIMA technique to generate information from their data. Past observations are later incorporated into a multilayer neural network with a backward propagation algorithm, where the selection of the parameters will achieve a prediction to an ultra-short horizon. Neural networks with short-term memory are a widely used technique due to their performance capacity; however, this type of network is improved with the contribution of other techniques. This is the case of models that use a convolutional neural network as a base [22], which is a type of classifying network. On the other hand, within the prediction models of photovoltaic energy, artificial intelligence techniques are used, as is the case of supervised learning machines that pre-treat their input elements and incorporate a linear regression for the correlation of their data [23]. There is a classification according to the efficiency of photovoltaic energy prediction models in which, according to the mean absolute percentage error, it establishes that the models that present a value less than 10% are accurate and reliable, a value between 10–20% indicates a good prediction, 20–50% means a reasonable prediction, and more than 50% indicates an inaccurate model [24].

The prediction models of photovoltaic energy are important and fundamental to avoid possible penalties to the operators of the photovoltaic power generation plants, reduce the risks of their connection to the electrical grid, and specify the use of energies with a low carbon ratio [25]. Its study is necessary for the development and fulfillment of the goals established to reduce climate change, as there are still areas of opportunity that must be explored to improve the performance of forecast models. Hybrid models have been shown to be capable of improving PV power prediction performance; however, they are not yet fully explored for development in research. Given the technological advances that have been developed in recent decades, this paper shows a methodology of a hybrid method that establishes:

- A technique that combines PV power prediction methods for a short-term scale for large amounts of data.
- The development of a model for the prediction of photovoltaic energy through neural networks that will have as input information the data of an embedding model with delay coordinates and will be compared with a clear sky model and a SARIMA.
- Finally, the proposed model will be validated with real data from a photovoltaic plant.

2. Materials and Methods

2.1. Data Acquisition

The acquisition of the database was obtained through the Solar Radiation Monitoring Laboratory of the University of Oregon, a free source that allows visualization of experimental data from its research projects. Figure 1 shows a satellite image of the photovoltaic array installation used as the objective of this investigation, which corresponds to daily observations of a system [26].

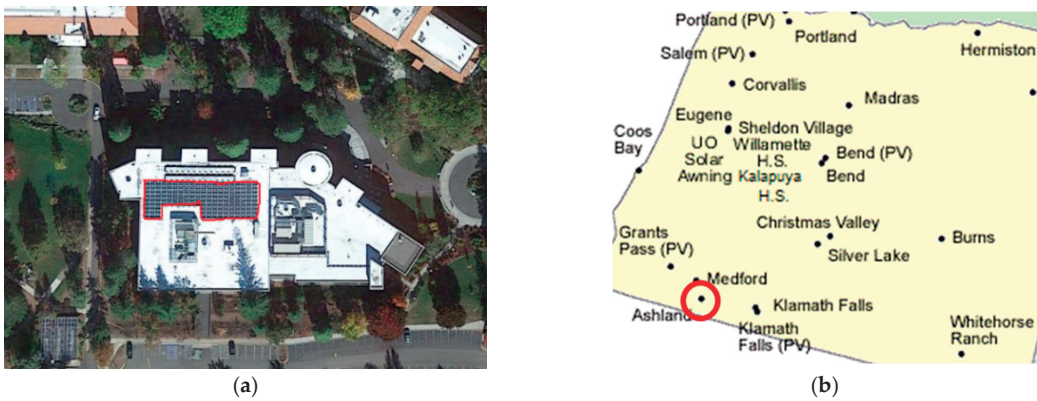


Figure 1. Area of the study experiment. (a) The red line shows the array of a photovoltaic system in Ashland; (b) the red circle shows the location of the photovoltaic array has a latitude of 42.19 and a longitude of 122.70 at an altitude of 595 m.

The database has 315,648 observations with a horizon resolution of every five minutes, the information period of the observations is from 1 January 2018 00:00 to 30 November 2021 23:55. Table 1 presents six variables of the photovoltaic array that were used to carry out the present experiments.

Table 1. Variables extracted by Solar Radiation Monitoring Laboratory of the University of Oregon.

Variables	Units
Global radiation	Wh/m ²
Direct radiation	Wh/m ²
Diffuse radiation	Wh/m ²
Power	W
Wind speed	m/s
Temperature	°C

Figure 2 shows the time series of the first 2000 observation of the power variable. The behavior of the series is cyclical.

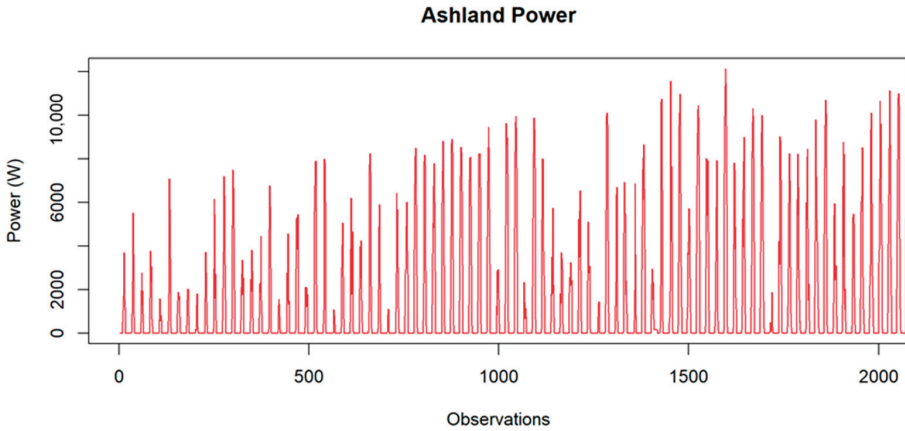


Figure 2. Time series of the power a photovoltaic system.

The time series has a total of 1283 missing data, which corresponds to less than 10% of the total data, so a data imputation was applied by taking the last observation into consideration for the periods of missing hours. Table 2 shows the percentage of missing data according to the database variables.

Table 2. Missing time series data.

Global Radiation	Power	Wind Speed	Temperature
0.28%	0.39%	0.23%	0.13%

Once the time series was completed, the following experiments were carried out in a sequential form:

- Clear Sky Model
- SARIMA Model
- Lag Coordinate Embedding Model
- Multilayer neural networks

2.2. Clear Sky Model

A clear sky model is based on the calculation of solar radiation transfer through algorithms designed for the simulation of the wavelength in the physical interactions between solar radiation and atmospheric particles. Equation (1) shows the calculation of global solar radiation:

$$G = G_{CS} \times \tau_c \tag{1}$$

where G is the global solar irradiance ($\frac{W}{m^2}$), G_{CS} is the global irradiance of the clear sky ($\frac{W}{m^2}$), and τ_c is the transmissivity of the clouds that model the system. To carry out the clear sky model, the apparent instantaneous movement of the sun was calculated using the equation of Cooper: the angle of inclination δ establishes the amount of solar radiation that reaches the earth, which is inversely proportional to the square of the distance from the sun [26]. Equation (2) shows the magnitudes to be considered in the angle of inclination according to the Cooper equation.

$$\delta = 23.45 \sin \left[\frac{360}{365} (d_n + 284) \right] \tag{2}$$

where d_n is an arbitrary day of the year. To calculate the apparent movement of the sun, the latitude of Ashland was incorporated as reference data, which resulted in the global irradiance; later, based on the information from the photovoltaic array, an adjustment was

made to the IV curve of the solar panel according to the characteristics of the manufacturer, in such a way that the output power of the system was calculated according to each observation. The clear sky model does not consider specific characteristics of the system or sudden changes of physical origin, such as: system damage, system maintenance, or a physical phenomenon with little anticipation, among others, so one of its main disadvantages is its characterization of ideal conditions of the environment and the system.

2.3. Autoregressive Model of Order Moving Averages with Seasonality (P, D, Q)s

The Integrated Autoregressive Model of Moving Averages of Order with Seasonality (P, D, Q)s is a combination of the autoregressive models AR (p) and moving averages MA (q), with seasonality of order p, d, q, with the particularity of including a restoration process called differences. In addition, it incorporates seasonality as a component for the forecast calculation of a variable, leaving the following order (P, D, Q)s. It is a model that works with past observations and has the ability to identify seasonal behavior in a time series. Equation (3) shows the variables considered for calculating the model:

$$Y_t = \varphi_1 Y_{t-1} + \dots + \varphi_p Y_{t-p} + \varepsilon_t - \theta_q \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q} \tag{3}$$

where Y_t is the instantaneous moment of the forecast, φ is the autoregressive coefficient together with Y_{t-p} , i.e., the normalized record of the time series to be modeled, θ is the moving average coefficient with its respective error term of each record, i.e., ε_{t-q} . Calculation of the model commences with the selection of 80% of its observations as training and the other remaining for testing, followed by a Dickey-Fuller test to identify if the time series is stationary; if it is not, the differences are calculated for its transformation.

The model obtained an array (4, 0, 3) (0, 1, 0) (288) which indicates that it considers four autoregressive values and three moving averages of past observations with no difference in a one-day seasonality, corresponding to 288 observations every five minutes.

2.4. Time Delay Coordinate Embedding Model

A Time Delay Coordinate Embedding model (TDC) consists of mapping the observations in different phases of space. The TDC model is useful for discovering effective coordinate systems to represent the dynamics of physical systems. Recently, models identified by dynamic mode decomposition into time lag coordinates have been shown to provide linear representations of strongly nonlinear systems. The use of significant models of complex non-linear systems from measurement data aims to potentiate and improve the characterization, prediction, and control of observations.

Takens and Sauer [27] established that if the sequence really consists of scalar measurements of the state of a dynamical system, then, under certain assumptions, the time delay embedding provides a one-to-one picture of the original ensemble, described by the following equation:

$$s_{n-h} = f(x) - (m - 1)\tau, s_n - (m - 2)\tau, \dots, s_n \tag{4}$$

where (s_{n-h}) is the time series observed at regular intervals, $f(x)$ is the length of the time series, (τ) is the time lag, (m) is the number of dimensions in which to embed $(\tau)s_n$, meaning that the time lag of the time series is large enough to provide information for the next instant in time. Figure 3 shows the structure of the TDC model. The column headers locate the dimensions of the experiments in the time series dynamically as $X_{-t_2}, X_{-t_1}, X_{-t_0}$; the observation in the current time written as X_{-t_0} and two previous ones, for the forecast of the following observation, are based on these three observations resulting in h_1 , yielding the observation X_4 ; once the predicted value is established, the following observations are embedded successively within the same matrix.

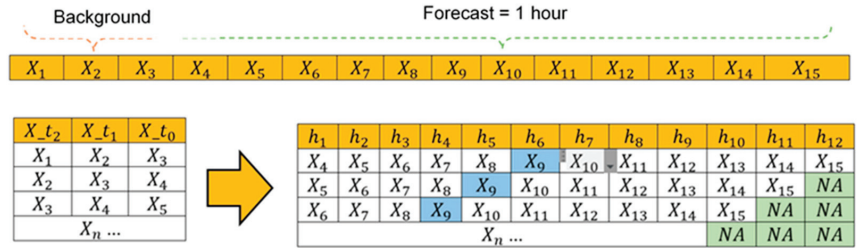


Figure 3. The Time Delay Coordinate model (past observations forecasting) Neural Networks.

Figure 4 also shows that, as the displacement of the dimensions given by is advanced, i.e., X_{t-2} , X_{t-1} , X_{t_0} , the moment will come when there will be no observations that can continue to return results for the last forecasted observations and will be marked as NA (absence of data).

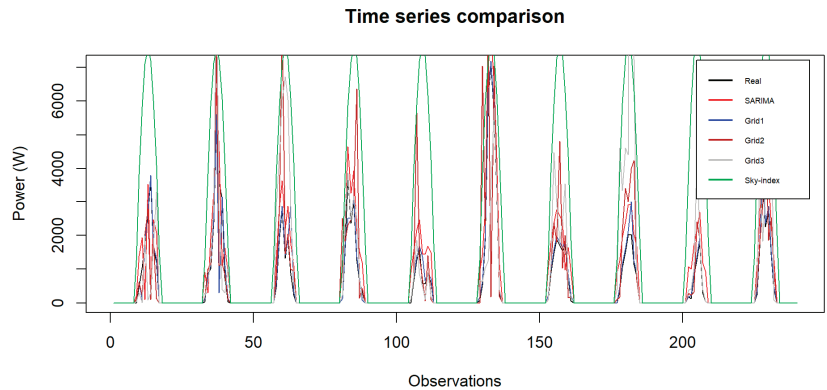


Figure 4. Comparison of the time series of the models.

2.5. Artificial Neural Networks

Artificial Neural Networks (ANN) are mathematical models that try to reproduce the functioning of the nervous system, made up of a set of units called neurons. The functioning of a neural network depends on the structure selected for its performance. In the development of the neural network model, it was decided to use a multilayer-type network, and the information resulting from the TDC model was used as input data in order to provide more information to the network for its training. In the structure of the network, different parameters were tested in order to obtain an accurate forecast. The multilayer neural network had 80% of the training information and the rest was used for validation. Table 3 shows the structures that had a positive degree of forecast accuracy.

Table 3. Multilayer network structure.

Structure ANN	Parameters
Network 1.	
Hidden layers	(6, 123, 10)
Activation function	Hyperbolic tangent
Error threshold	0.01
Algorithm	Back propagation
Epoch	100

Table 3. Cont.

Structure ANN	Parameters
Network 2.	
Hidden layers	(3, 143, 7)
Activation function	Sigmoid
Error threshold	0.01
Algorithm	Back propagation
Epoch	100
Network 3.	
Hidden layers	(7, 128, 12)
Activation function	Hyperbolic tangent
Error threshold	0.01
Algorithm	Back propagation
Epoch	100

3. Results and Discussion

Figure 4 shows a comparison of the models of neural networks with different architectures in their configuration, SARIMA, and clear sky index with respect to the actual observations of the photovoltaic array. The observations estimated with each model were obtained as a product time series with the same cyclical pattern that corresponds to the generation of electrical energy from the solar resource. The time of the clear sky index model is the one with the greatest variation compared to the actual observations. Figure 4 shows that the network 1, which has a network architecture and configuration, presents a fluctuation closer to reality.

The models with different error metrics were evaluated to determine their reliability, including mean absolute percentage error, mean absolute error (MAE), and coefficient of determination (R^2). Table 4 shows that in the calculation of the MAPE, the model that had the lowest degree of error was network 1, giving a value of 0.57% compared to the clear sky index model that had an error of 38.6%, the latter due to the model assuming that, at all times, the meteorological conditions are stable and there are no technical failures of the photovoltaic system. In the case of the values obtained for network 2 and 3, there were variations that depend on the architecture of the grid from the hidden layers and the activation function. The largest value of the MAE was obtained by the clear sky index with 1096.34 W of deviation compared to the research models; the value of the lowest deviation was obtained by network 1 with 69.29 W. In the case of the coefficient of determination, the model that had the best approximation of the estimate with respect to the real value was network 1 with a value of 0.97, while other models presented greater variation.

Table 4. Forecast error metrics.

Models	MAPE	MAE	MSE	R^2
SARIMA	9.06%	302.91	313,756.96	0.87
Network 1	0.57%	69.29	82,826.71	0.97
Network 2	1.57%	335.17	1,089,554.92	0.93
Network 3	4.05%	521.73	4,505,871.20	0.91
Sky Index	38.6%	1096.34	28,563,065.34	0.51

The neural network model 1 presented a MAPE value of 0.57%, which indicates that the performance of the model has a good reliability since it belongs to the range of 0–10%. [24].

4. Conclusions

This paper proposes supplying input data to a back-propagated multilayer neural network from output data of a time delay coordinate embedding model and comparing the results with statistical and numerical weather prediction models, as well as different

architectures of the neural network. The model of network 1 obtained a MAPE of 0.57% and an R^2 of 0.97, indicating that the model based on multilayer neural networks presents a good performance in the forecast of solar power.

Author Contributions: All authors contributed to the study conception and design. Material preparation, data and analysis were performed by G.G.-G., G.S.-B. and J.C.-R. The first draft of the manuscript was written by G.G.-G. and all authors commented on previous versions of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: G.G.-G. would like to acknowledge grant given to by Conahcyt, México, grant number 827363.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data was obtained from the Solar Radiation Monitoring Laboratory of the University of Oregon.

Conflicts of Interest: The authors declare no conflict of interest. The author G.S.-B. wants to clarify that the presented is all his own opinion and research lines, and not necessarily the opinion of BBVA México.

References

1. IEA. Renewable Electricity. Available online: <https://www.iea.org/fuels-and-technologies/electricity> (accessed on 3 March 2023).
2. International Renewable Energy Agency. Available online: <https://www.irena.org/Energy-Transition/Technology/Solar-energy> (accessed on 1 April 2023).
3. IEA. Solar. Available online: <https://www.iea.org/fuels-and-technologies/solar> (accessed on 3 March 2023).
4. Gürel, A.E.; Agbulut, Ü.; Bakır, H.; Ergün, A.; Yıldız, G. A state of art review on estimation of solar radiation with various models. *Heliyon* **2023**, *9*, e13167. [CrossRef]
5. UNFCCC. UN Climate Change Quarterly. Available online: <https://unfccc.int/documents?f%5B0%5D=conference%3A4526> (accessed on 2 April 2023).
6. Parasad, A.A.; Kay, M. Evaluation of simulated solar irradiance on days of high intermittency using WRF-Solar. *Sol. Energy* **2020**, *13*, 2200–2217.
7. Antonanzas, J.; Osorio, N.; Escobar, R.; Urraca, R.; Pison, F.J.M.-D.; Antonanzas-Torres, F. Review of photovoltaic power forecasting. *Sol. Energy* **2016**, *136*, 78–111. [CrossRef]
8. Khalid, M.; Savkin, A. A method for short-term wind power prediction with multiple observation points. *IEEE Trans. Power Syst.* **2012**, *27*, 579–586. [CrossRef]
9. Chen, N.; Qian, Z.; Nabney, I.; Meng, X. Wind power forecasts using Gaussian processes and numerical weather prediction. *IEEE Trans. Power Syst.* **2014**, *29*, 656–665. [CrossRef]
10. Dowell, J.; Pinson, P. Very short-term probabilistic wind power forecasts by sparse vector autoregression. *IEEE Trans. Smart Grid.* **2016**, *7*, 763–770. [CrossRef]
11. Zhao, X.; Bai, M.; Yang, X.; Liu, J.; Yu, D.; Chang, J. Short-term probabilistic predictions of wind multi-parameter based on one-dimensional convolutional neural network with attention mechanism and multivariate copula distribution estimation. *Energy* **2021**, *234*, 121306. [CrossRef]
12. Mellit, A.; Pavan, A.M.; Ogliaari, E.; Leva, S.; Lughi, V. Advanced methods for photovoltaic output power forecasting. *Appl. Sci.* **2020**, *10*, 487. [CrossRef]
13. Karner, O. ARIMA representation for daily solar irradiance and surface air temperature time series. *J. Atmos. Sol.-Terr.* **2009**, *71*, 841–847. [CrossRef]
14. Inman, R.H.; Pedro, H.T.; Coimbra, C.F. Solar forecasting methods for renewable energy integration. *Energy Combust.* **2013**, *39*, 535–576. [CrossRef]
15. Alsharif, M.H.; Younes, M.K.; Kim, J. Time series ARIMA model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea. *Symmetry* **2019**, *11*, 240. [CrossRef]
16. Arbizu-Barrena, C.; Ruiz-Arias, J.A.; Rodríguez-Benitez, F.J.; Pozo-Vázquez, D.; Tovar-Pescador, J. Short-term solar radiation forecast using advection and diffusion of the MSG cloud index. *Energy Sol.* **2017**, *155*, 1092–1103. [CrossRef]
17. Diagne, M.; David, M.; Boland, J.; Schmutz, N.; Lauret, P. Post processing of solar irradiance forecasts from the WRF model in Island. *Sol. Energy* **2014**, *105*, 99–108. [CrossRef]
18. Amit Kumar Yadav, S.S. Chandel, Solar radiation prediction using Artificial Neural Network techniques: A review. *Renew. Sustain. Energy Rev.* **2014**, *33*, 772–781. [CrossRef]
19. López, G.; Batlles, F.J.; Tovar-Pescador, J. Selection of input parameters to model direct solar irradiance by using artificial neural networks. *Energy* **2005**, *30*, 1675–1684. [CrossRef]
20. Wanga, Z.; Wanga, F.; Sub, S. Solar irradiance short-term prediction model based on BP neural network. *Energy Procedia* **2011**, *12*, 488–494. [CrossRef]

21. Ahmad, M.J.; Tiwari, G.N. Solar radiation models—Review. *Int. J. Energy Environ.* **2010**, *1*, 513–532. [CrossRef]
22. Huang, X.; Liu, J.; Xu, S.; Li, C.; Li, Q.; Tai, Y. A 3D ConvLSTM-CNN network based on multi-channel color extraction for ultrashort-term solar irradiance forecasting. *Energy* **2023**, *272*, 127140. [CrossRef]
23. Osah, S.; Acheampong, A.A.; Fosu, C.; Dadzie, I. Deep learning model for predicting daily IGS zenith tropospheric delays in West Africa using TensorFlow and Keras. *Adv. Space Res.* **2021**, *68*, 1243–1262. [CrossRef]
24. Lewis, C.D. *International and Business Forecasting Methods*; Butter Worths: London, UK, 1982.
25. University of Oregon Solar Radiation Monitoring Laboratory. Available online: <http://solardata.uoregon.edu/cgi-bin/ShowArchivalFiles.cgi> (accessed on 1 April 2023).
26. Yang, L.; Gao, X.; Hua, J.; Wang, L. Intra-day global horizontal irradiance forecast using FY-4A clear sky index. *Sustain. Energy Technol. Assess.* **2022**, *50*, 101816. [CrossRef]
27. Yagasaki, K.; Uozumi, T.G. Controlling chaos using nonlinear approximations and delay coordinate embedding. *Phys. Lett. A* **1998**, *247*, 129–139. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Proceeding Paper

Analysis of the Application of Different Forecasting Methods for Time Series in the Context of the Aeronautical Industry [†]

Antônio Augusto Rodrigues de Camargo ^{1,2,*} and Mauri Aparecido de Oliveira ^{1,2,‡}

¹ Department of Management and Decision Support, São José dos Campos, Aeronautics Institute of Technology, São Paulo 12228-900, SP, Brazil; mauri@ita.br

² Institute of Science and Technology, São José dos Campos, Federal University of São Paulo, São Paulo 12247-014, SP, Brazil

* Correspondence: antonio.camargo@ga.ita.br

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

[‡] These authors contributed equally to this work.

Abstract: The aeronautical sector is a vital part of the Brazilian industrial landscape, contributing to the development of new technologies and production techniques with potential applications in other industries. However, due to its restricted nature, there are limited studies on implementing improvements in its systems, highlighting the need for attention in specific subareas of companies in this sector. One such area is the production planning department, especially the forecasting techniques applied in the supply chain, which play a crucial role in the operations of any company and are a determining factor in decision making. The objective of this research is to compare the effectiveness of various time-series forecasting methods, including classical statistical methods and neural networks. The study employs a real-time series that depicts the consumption of a specific material extensively used in the production line of a major Brazilian aircraft manufacturer. The proposed forecasting methods are applied, and the results are compared using three different evaluation metrics. The objective is to emphasize the significance of optimizing strategic planning within the industry and the potential savings that can be achieved by selecting the best forecast. In conclusion, the findings of this study can be used to enhance the efficiency of the supply chain and operations of companies in the aeronautical sector.

Keywords: forecasting; time series; aeronautical industry; supply chain; statistical methods

Citation: de Camargo, A.A.R.; de Oliveira, M.A. Analysis of the Application of Different Forecasting Methods for Time Series in the Context of the Aeronautical Industry. *Eng. Proc.* **2023**, *39*, 74. <https://doi.org/10.3390/engproc2023039074>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 11 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The constant evolution of human needs has always necessitated new technologies and improved processes to meet the growing demand. Throughout history, industries have undergone significant transformations as new mechanisms were developed. Today, we are witnessing the emergence of Industry 4.0, a new revolution that is transforming manufacturing into a more connected and automated environment. This technological era emphasizes the integration of systems, both vertically and horizontally, to facilitate decision making within the production chain [1].

Enterprise Resource Planning (ERP) systems facilitate integration and enable companies to manage all aspects of planning and raw material procurement. However, in practice, various planning techniques are utilized based on the physical, chemical, or commercial characteristics of the materials. For example, in the aeronautical industry, the average monthly consumption method is commonly used for materials with a low unit cost and high turnover rate in the production line. This technique involves configuring a periodic numeric parameter within the ERP system to predict the monthly demand for a product over a set number of months; then, purchase orders are generated based on this forecast.

The average monthly consumption technique is often controlled and implemented in the production line using the Kanban system. As described by Slack et al. [2], the Kanban system is a method of operationalizing pull planning and control, where the customer stage signals its supplier stage to provide the necessary supply. The system aims to efficiently control the production stages and simplify administrative mechanisms, allowing users to liquidate a larger amount of stock in the system at once instead of piece by piece [3].

The most common family of inputs for the type of planning discussed previously is known as hardware, which refers to the physical equipment made of metal, such as screws, nuts, rivets, rods, collars, and so on. In the aeronautical industry, the term 'hardware' is further categorized, and this study focuses on the analysis of a specific category known as fasteners, which are devices used to assemble various structures. Based on the complex nature of the manufacturing process in this sector, planning and purchasing for these materials are often performed individually, using nonstandard means, such as historical consumption averages or future demands based on product structure, which can be highly inaccurate due to the possibility of using optional and alternative materials that are not linked to the bill of materials. However, such arrangements are highly susceptible to errors that can cause both excessive purchases, leading to storage overload and raw material obsolescence, as well as material shortages in stock, potentially resulting in production line stoppages. As a result, inventory planning and management must be carefully managed to ensure the smooth operation of the production line.

Given the large number of parts and components in an aircraft, it can be challenging to manage all sectors and necessary inputs accurately. This can result in the procurement of raw materials in erroneous quantities, which can directly affect specific stages of production. Therefore, there is a need for forecasting techniques that provide more accurate projections of the real demand within the supply chain.

2. Case Study and Data

The main objective of this article is to compare different time-series forecasting methods applied to a real database. The database consists of 42 monthly consumption values for a specific category of raw material utilized in aircraft construction by a prominent Brazilian aeronautical manufacturer. The material under study is a flat steel washer used in various types of aircraft within the company for assembling internal structural parts in a variety of areas such as panels, supports, windows, seats, air conditioning and refrigeration systems, landing gear, cabling, electrical systems, doors, equipment, tubes, and more. To ensure business confidentiality, it will be referred to as "Material 1" rather than by its real market identification (part number).

In this way, the proposal is to compare the efficiency of these methods by presenting some error metrics. The following nine methods are discussed in this work:

1. Simple Exponential Smoothing;
2. Holt;
3. Holt–Winters Additive;
4. Holt–Winters Multiplicative;
5. ETS (M,N,A);
6. Naïve;
7. ARMA (2,1);
8. AR (2);
9. Neural Network.

3. Materials and Methods

The first stage of the process involved obtaining the database by extracting it from the company's Enterprise Resource Planning (ERP) system. The original database is a Microsoft Excel spreadsheet, where each row represents a different material, and each column represents different months/years of consumption in the production line, along with other information that is irrelevant for this study.

The next step of the methodology involved the selection of the specific material to be studied. In order to ensure the feasibility and accuracy of the results, it is preferable to choose a material that does not exhibit extreme variations or a large number of null values, as these can make it difficult to execute the proposed forecasting methods effectively. Once a suitable material was identified, it was important to carefully clean and preprocess the data in preparation for the subsequent analysis. This typically involves removing other materials not chosen and removing columns that contain information not useful for this study, for example, the location of the manufacturing plant, the specific company code/identification, lead time, transport time, and other irrelevant details. Finally, to facilitate reading, the columns were inverted by the lines of this worksheet, so that the identification of the material was represented by the column, and the values of the monthly consumption were represented by the lines, remaining vertical.

Thus, the treated database was imported into the integrated development environment (IDE) RStudio (2022.07.1 version), where through R language the time-series forecasting methods were applied. After the import, the first step in analyzing the time series was to examine the behavior of the data and assess whether they exhibited seasonality and stationarity. This involved generating a line chart of the complete series, calculating the descriptive statistics, creating a histogram, a box plot, and decomposing the data. In order to develop a forecast model, a training dataset was created using 36 of the 42 months of consumption data, covering the period from March 2019 to February 2022 (approximately 85.7% of the original dataset). The remaining six months of data were used for final analysis and comparisons with the results of the forecast models.

The final stage of the study involved applying all the proposed forecasting methods and measuring their respective metrics: the Symmetric Mean Absolute Percentage Error (sMAPE), Theil's U Index of Inequality, and the Root Mean Square Error (RMSE). This allowed for an evaluation of which models were the best fit for the analyzed data. It is important to note that the efficiency of the methods was evaluated by determining the accuracy of each procedure, resulting in a comprehensive and effective comparative analysis.

4. Results

To begin with, it is important to observe the complete time series for the studied material in this work. Figure 1 shows the line chart for all the consumption data (from March 2019 to August 2022), which indicated a significant reduction in consumption towards the end of 2019, followed by a gradual increase from the beginning of 2022. While there could be several hypotheses to explain this phenomenon, such as a change in the product structure via a study of the company's engineering, this study only focuses on the mathematical analysis and does not delve into any managerial aspects.

To understand the behavior of the series, we analyzed some of the data obtained from descriptive statistics. As shown in the previous Figure, the series did not display any apparent seasonality. Table 1 presents the descriptive statistics for this series.

Upon analyzing Table 1, it becomes apparent that Material 1 had a slightly positive skewness, indicating that the right tail of the distribution was slightly longer than the left tail. This was further confirmed by the histogram shown in Figure 2, although it was not easily noticeable by visual inspection. However, the kurtosis value was positive, indicating that the distribution had heavier tails than a normal distribution, which characterizes the flattening or lengthening of the curve. Additionally, Figure 2 highlights that there was a significant concentration of consumption values in the range of 20,000 to 30,000 units.

Figure 3 displays a boxplot that can help identify any outliers in the data, which are observations that deviate significantly from the rest of the time series values. It is evident from the plot that the series did not contain any outliers.

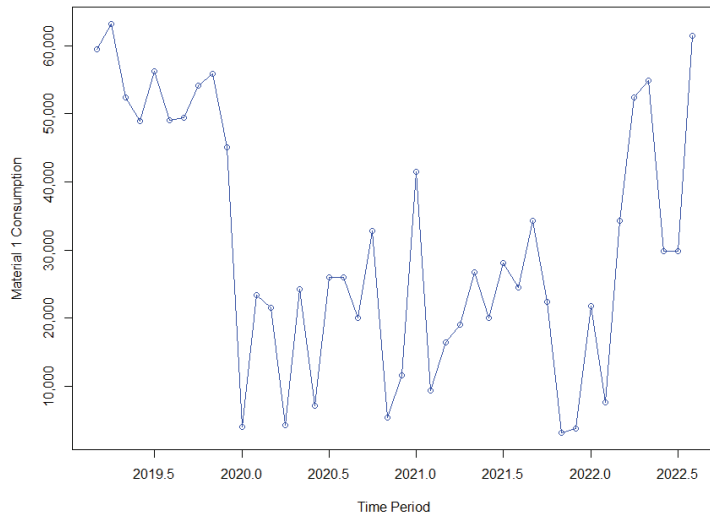


Figure 1. Material 1 time series.

Table 1. Descriptive statistics for Material 1.

Statistics	Value
Minimum	3108
First Quartile	19,261
Median	26,367
Mean	30,507
Third Quartile	49,003
Maximum	63,268
Variance	342,361,647
Standard Deviation	18,503.02
Skewness	0.2
Kurtosis	1.84

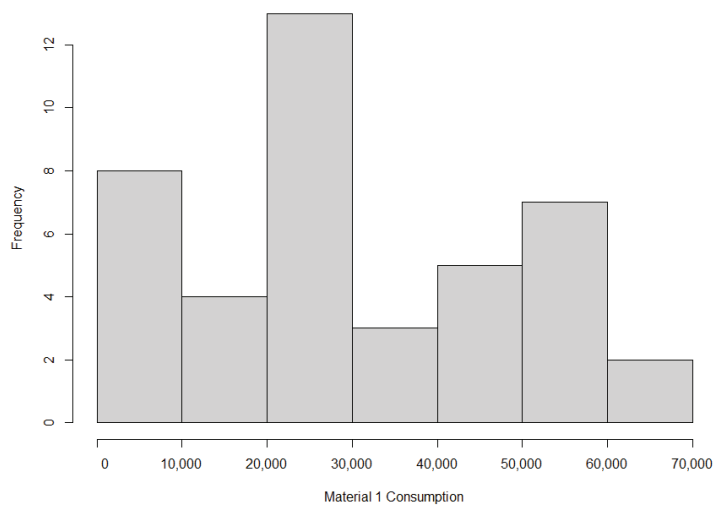


Figure 2. Material 1 histogram.

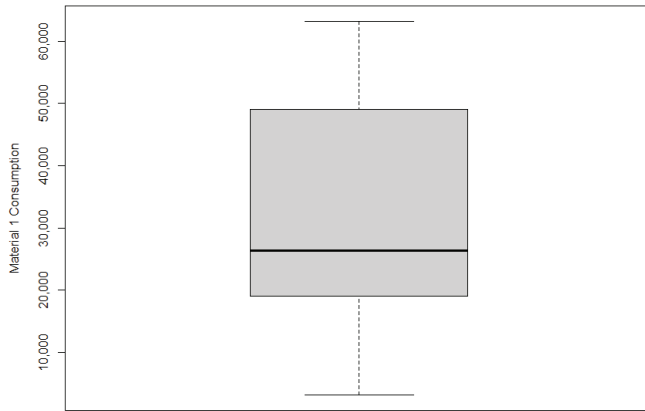


Figure 3. Material 1 boxplot.

The time series decomposition shown in Figure 4 provides valuable insights into the behavior of the data, where it revealed the absence of seasonality in the series. Furthermore, regarding the trend, as mentioned before, there was a considerable reduction from 2019 onwards, which remained practically stable and only showed an upward trend again from the beginning of 2022, forming an approximate drawing of a negative parabola.

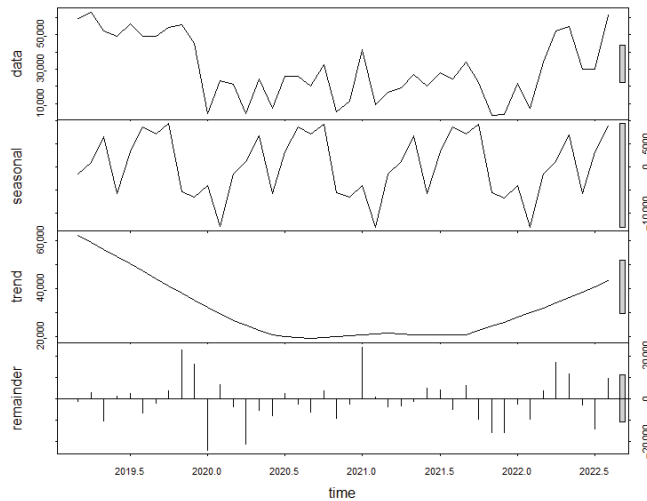


Figure 4. Decomposition of Material 1.

4.1. Forecasting Methods

4.1.1. Simple Exponential Smoothing

The simple exponential smoothing model is a widely used method in demand forecasting and can be used when the sample size is small. The technique is built through a weighted average of past and present values, where exponential weighting assigns greater weights to more recent data and smaller weights to more distant observations [4].

The result of this technique are always constant; in other words, all the forecasts assume the same value, equal to the last level component. This implies that it is appropriate

only when the time series does not have a trend or seasonal component [5]. The results of applying this method to the database can be observed in Figure 5.

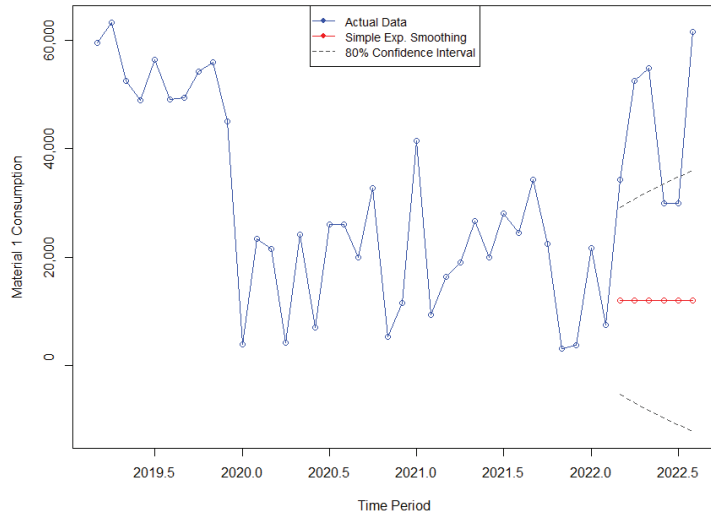


Figure 5. Graph with forecast confidence intervals for Material 1 using Simple Exponential Smoothing.

According to Figure 5, a significant disparity was observed between the actual values of the material 1 time series and the forecasted values for the same period. Hence, the consistent outcomes of the technique were inadequate for the data of this study, as it failed to predict the consumption peak that commenced in March 2022.

4.1.2. Holt

The Holt method, proposed by Holt [6], extends simple exponential smoothing to enable the forecasting of data with a trend. As a result, the forecast values generated by this method are not constant but exhibit a consistent trend (either increasing or decreasing) that extends indefinitely into the future.

The results of applying this method can be observed in Figure 6 below.

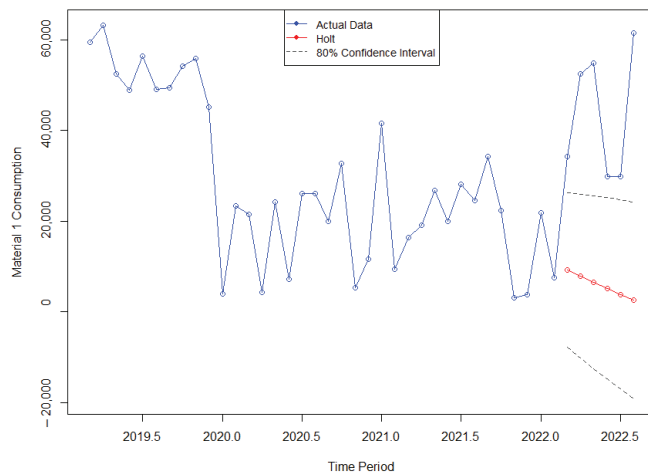


Figure 6. Graph with forecast confidence intervals for Material 1 using Holt's method.

According to Figure 6, there was a significant disparity between the predicted data and the actual consumption values for Material 1. The expected trend should be upward, but the forecast depicted a distinct downward trend. Consequently, this error was substantial enough to conclude that this method was not suitable for this type of time series.

4.1.3. Holt–Winters

The Holt–Winters method is a refined extension of the exponential smoothing approach, where the smoothing procedure provides an overall impression. This method also allows for studying future trends by generating medium and long-term forecasts.

Holt [6] and Winters [7] extended the Holt method to capture the seasonality of a series by proposing two variations that differ in the nature of the seasonal component: additive and multiplicative. Hyndman and Athanasopoulos [5] demonstrated that the additive method is suitable when seasonal variations are relatively constant throughout the series. In this case, the seasonal component is expressed in absolute terms on the scale of the observed series, and in the level equation, the series is seasonally adjusted by subtracting the seasonal component, resulting in an approximately zero sum within each year. On the other hand, the multiplicative method is advised when seasonal variations change proportionally with the level of the series. In this case, the seasonal component is expressed in relative terms (percentages), and the series is seasonally adjusted by dividing it by this seasonal component.

Therefore, beginning with the additive method, the outcomes of applying the Holt–Winters to the Material 1 series are illustrated in Figure 7 below:

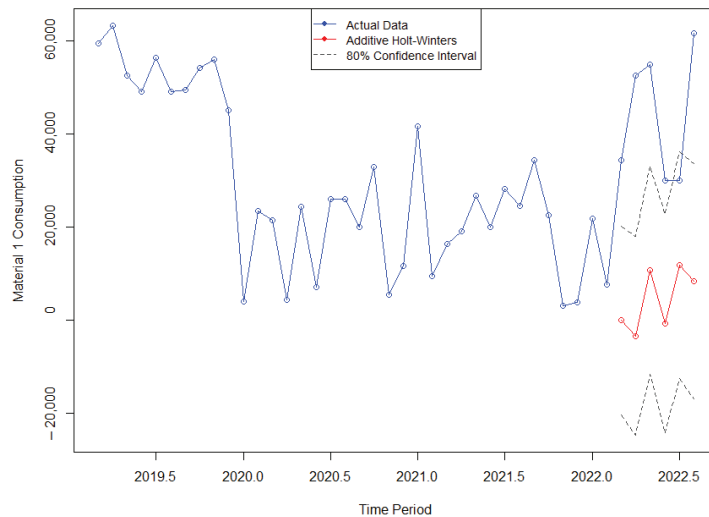


Figure 7. Graph with forecast confidence intervals for Material 1 using the Additive Holt–Winters.

By examining the preceding figure, it becomes evident that the method predicted three negative values for the months of March, April, and June 2022. However, such negative values were not feasible in this application. This study employed a real-time series that represented the consumption of a raw material in a production line, and given this context, consumption below zero was not possible. Therefore, it can be concluded that the method was not suitable for the Material 1 series.

Regarding the multiplicative method, while there were no negative values in the forecast for this six-month period, the forecasted trend ended up showing a negative tilt, which contradicted the actual data that exhibited consumption peaks starting from March 2022. This discrepancy in the value relationship can be observed in Figure 8, where due

to the multiplication equation employed by the method, the confidence intervals were significantly larger, resulting in a noticeable change in the scale of the line graph.

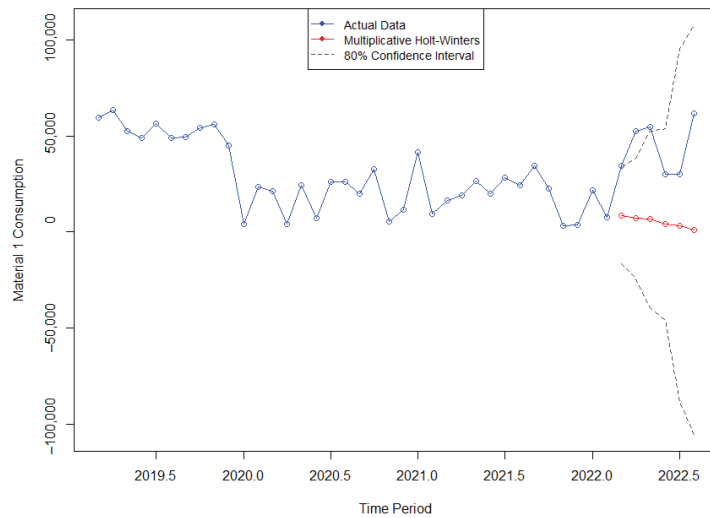


Figure 8. Graph with forecast confidence intervals for Material 1 using the Multiplicative Holt–Winters.

4.1.4. ETS

Considering the variations in the combinations of trend and seasonality components in the previously mentioned exponential smoothing method, it is possible to use ten new techniques. Each one is labeled by a pair of letters (T and S) that define the type of trend (T) and seasonality (S) components. This classification was first proposed by Pegels [8], who also included a method with a multiplicative trend. It was later extended by Gardner [9] to include methods with an additive damped trend and by Taylor [10] to include methods with a multiplicative damped trend.

The point forecasts generated by the models are identical when the same smoothing parameter values are used. However, they produce different prediction intervals. Additionally, for each method, there can be two models: one with additive errors and another with multiplicative errors. According to Hyndman and Athanasopoulos [5], to differentiate between these two models, a third letter is introduced, denoting the error term. Consequently, each state space model is labeled as ETS (*,*,*) representing (error, trend, seasonality), and this labeling convention can also be interpreted as exponential smoothing. Each combination of components has its own set of equations, and the possibilities for each component are as follows: Error = A, M; Trend = N, A, Ad; and Seasonality = N, A, M. In this context, A represents additive, M represents multiplicative, N represents none, and Ad represents additive damped.

In this study, all the possible label combinations were tested, and the root mean square error (RMSE) was measured for each combination by comparing the predicted data with the actual data. The model with the lowest RMSE was selected, which happened to be the ETS (M,N,A): multiplicative error, no trend, and additive seasonality. The computed results of this method can be observed in Figure 9.

Upon comparing the predicted data generated by the ETS (M,N,A) method with the actual consumption values for the corresponding period, it is apparent that the technique accurately forecasted a positive trend, distinguishing itself from certain previous methods. Nevertheless, there remained a notable disparity in the magnitude of the peaks.

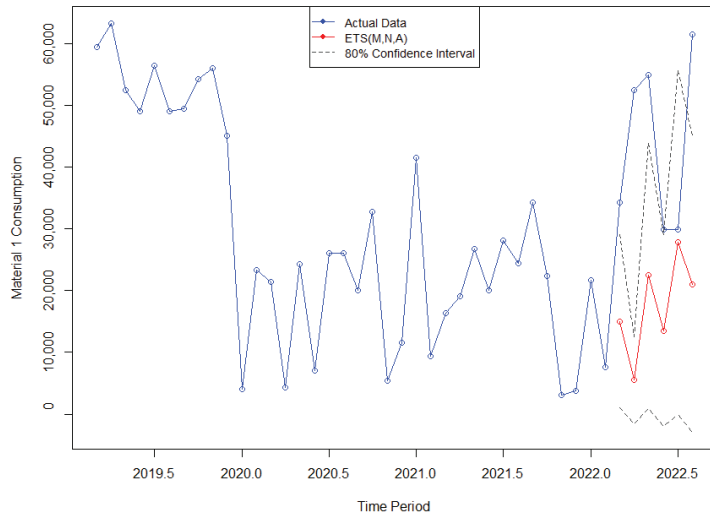


Figure 9. Graph with forecast confidence intervals for Material 1 using the ETS (M,N,A).

4.1.5. Naïve

The Naïve model is one of the simplest methods for time series forecasting and works well for many economic sectors and financial time series. The Naïve Simple technique involves using the exact value of the last observation in the time series as the forecast, but some variations take into consideration the seasonality and are referred to as Seasonal Naïve. In this case, the forecast is based on the same observed value from a previous point in the same season, such as the value from the same month but in the previous year [5].

The model used in this work is the Seasonal Naïve method, considering the forecast value of the same month from the previous year in the time series. Figure 10 depicts the graph of the results obtained by applying this model.

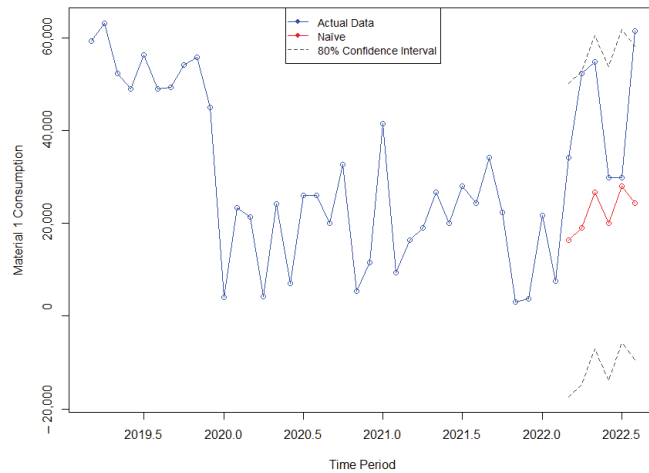


Figure 10. Graph with forecast confidence intervals for Material 1 using Naïve.

Upon examining the forecast graph generated by the Naïve method and comparing it with the actual data, several noteworthy observations come to light. While this method demonstrated limitations in accurately predicting the extreme peaks observed in four

particular months, it showcased exceptional accuracy during the months of June and July, closely aligning with the actual data and capturing the upward trend exhibited by the data. These findings suggest that the Naïve method exhibits potential for capturing seasonality in specific months, albeit with limitations in predicting extreme fluctuations.

4.1.6. ARIMA

The designation of the ARIMA model stands for Autoregressive Integrated Moving Average and refers to a type of self-regressive model that allows for predicting the values of a variable based on its previous values without the need for other auxiliary information or related variables [11]. The generic name ARIMA for these models refers to their three main components: Autoregressive (AR), Integrated (I), and Moving Average (MA). In these models, the aim is to describe autocorrelations in the data, where each observation of a variable at a given time is modeled based on previous values over time for the same variable.

In this approach, the modeling process involves deriving an ARIMA model that fits the given dataset, which requires analyzing the essential characteristics of the time series, such as trend, seasonality, cyclical variations, autocorrelation functions, and residuals [5]. Another point is that for the application of the model, the time series must necessarily be stationary, meaning that their statistical properties remain constant over time. If they are not stationary, it will be necessary to differentiate the data until they become stationary.

The initial step of the ARIMA model involved applying a logarithmic transformation to the data and subsequently differencing them to achieve stationarity. In the case of the time series of Material 1, it was only necessary to difference it once to achieve stationarity. To confirm this, the Dickey–Fuller unit root test was used.

Consequently, to proceed with the application of the method, it was necessary to identify the model using the autocorrelation function (ACF) for the “MA” term and the partial autocorrelation function (PACF) for the “AR” term. Both were applied to the differenced time series and can be analyzed in Figure 11.

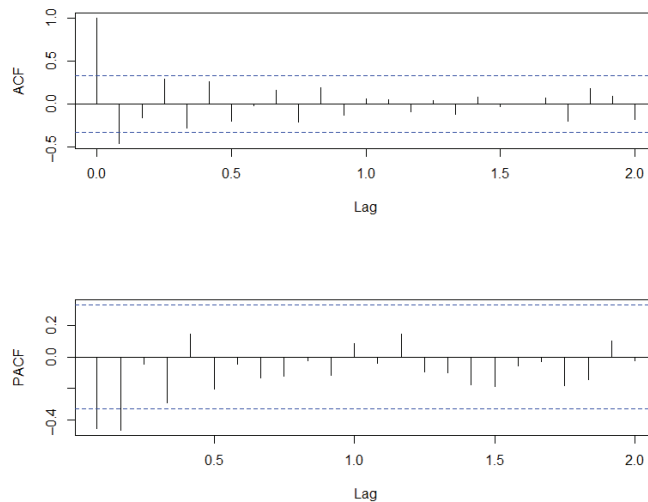


Figure 11. Graph of the cumulative sum of the autocorrelation function (ACF) and partial autocorrelation function (PACF) for the differenced time series of Material 1.

Based on these analyses, several combinations for the method can be considered. The chosen models were the ARMA($p = 2, d = 0, q = 1$) and the AR($p = 2, d = 0, q = 0$). Subsequently, the forecasts generated by both the ARMA(2,1) and AR(2) models for the Material 1 time series did not capture a significant peak in March 2022, in contrast to the actual value of the series, which was considerably higher. However, the confidence intervals

were able to approximate the substantial increase in values from that period onwards, which some of the previously mentioned methods failed to capture. A visual comparison between the actual data and the generated forecasts can be observed in Figures 12 and 13.

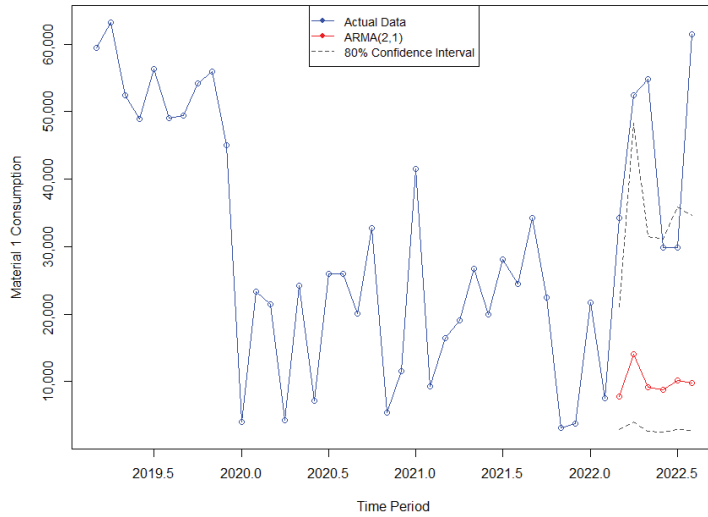


Figure 12. Graph with forecast confidence intervals for Material 1 using ARMA(2,1).

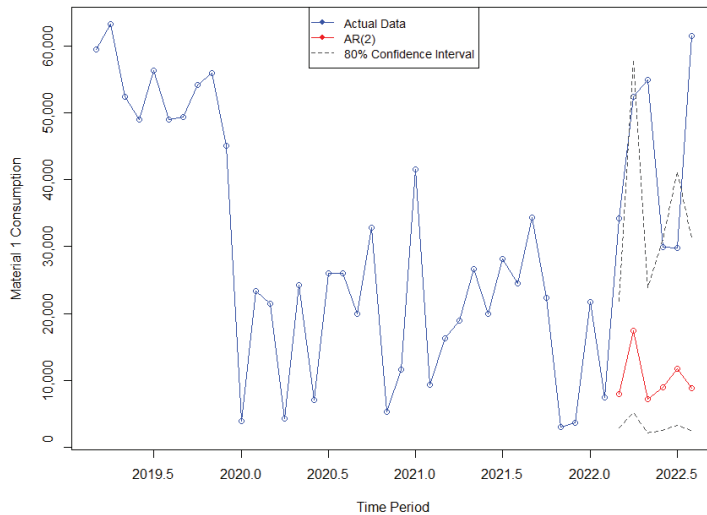


Figure 13. Graph with forecast confidence intervals for Material 1 using AR(2).

4.1.7. Neural Network

The final technique applied in this study was the neural network, based on the autoregression with neural networks (AR-NN) approach that combines autoregression (AR) and neural networks (NN) techniques to model time series. The results obtained with this technique are presented in Figure 14.

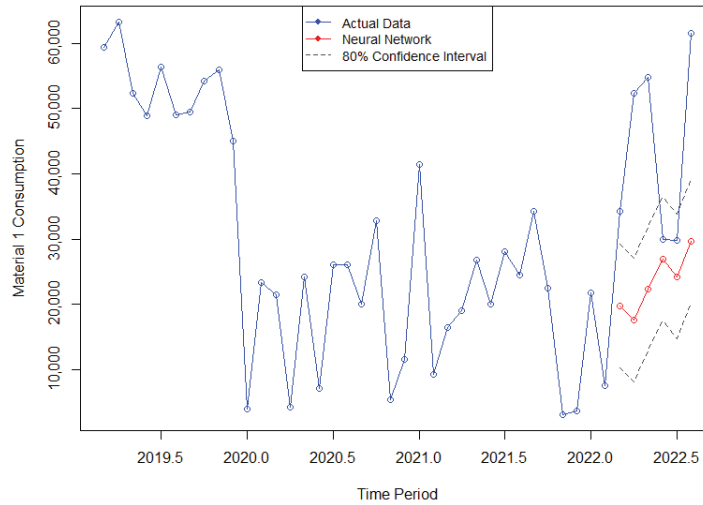


Figure 14. Graph with forecast confidence intervals for Material 1 using the Neural Network.

The analysis of the forecast graph generated by the neural network reveals several noteworthy observations. Firstly, the method successfully captured and predicted a positive trend that aligned reasonably well with the actual data. However, it showed limitations in accurately predicting the highest peaks of the data, which suggests potential challenges in capturing extreme fluctuations. Despite this, it is important to highlight the remarkable performance of the forecast during the months of June and July, where during these months, the predicted data closely aligned with the actual data, indicating a high level of accuracy and precision during that specific period. These findings demonstrate the model's ability to capture and replicate patterns effectively, particularly during months characterized by more stable and predictable trends.

4.2. Evaluation Metrics

To determine the best-performing forecasting method among those presented, it was necessary to measure the errors by comparing the actual data with the predicted data. Therefore, this study employed three different metrics to analyze the effectiveness of the techniques: the Symmetric Mean Absolute Percentage Error (sMAPE), Theil's U Index of Inequality, and the Root Mean Square Error (RMSE).

The Symmetric Mean Absolute Percentage Error (sMAPE) was proposed by Makridakis [12] in order to correct some disadvantages; that is, a modified Mean Absolute Percentage Error (MAPE) has a heavier penalty for forecasts that exceed the actual than those that are less than the actual. So, this metric is a modified MAPE, in which the divisor is half of the sum of the actual and forecast values.

The Theil's U Index of Inequality is an accuracy measure often cited in the literature, and according to Bliemel [13], there is confusion about this index, which may result from the fact that Theil [14] proposed two distinct formulas, but with the same name. The first proposal is bounded between 0 and 1, and this metric is used in this study. In the second proposal, the upper limit is infinite. This metric analyzes the quality of forecasts, and the closer it is to zero, the lower the prediction error generated by a specific model. In other words, it indicates that a forecast is better than the trivial forecast [15].

The Root Mean Square Error (RMSE) is calculated as the square root of the mean of the square of all of the error. It is widely used and considered an excellent general-purpose error metric for numerical predictions. The RMSE provides a reliable measure of accuracy, particularly when comparing forecasting errors among different models or

model configurations for a specific variable. However, it should be noted that the RMSE is scale-dependent and cannot be directly compared between variables [16].

In Table 2, it is possible to observe the three mentioned metrics for each of the nine forecasting methods.

Table 2. Evaluation metrics for Material 1.

Forecasting Method	sMAPE	Theil's U Index of Inequality	RMSE
Simple Exponential Smoothing	1.094186	0.7519919	34,314.91
Holt	1.500099	0.8816619	40,232.01
Holt–Winters Additive	1.623714	0.9098164	41,516.76
Holt–Winters Multiplicative	1.559045	0.8978318	40,969.88
ETS(M,N,A)	0.8413673	0.6653724	30,362.28
Naïve	0.6081907	0.5440008	24,823.86
ARMA(2,1)	1.230298	0.7892819	36,016.53
AR(2)	1.203849	0.7858103	35,858.11
Neural Network	0.5643272	0.5319228	24,272.71

After examining the table above, it becomes evident that the Holt–Winters methods, both additive and multiplicative, yielded the highest error measurements. This indicates that these particular methods were less effective in accurately forecasting the given time series. Similarly, Holt's method and the combinations of the ARIMA, such as the ARMA(2,1) and AR(2), exhibited high error metrics, further suggesting their inefficiency in this context. Surprisingly, the simple exponential smoothing method, despite its simplicity and constant forecasting values, outperformed more complex approaches such as the ARIMA models. The top three performing methods, ranked in order, were the Neural Network, Naïve, and ETS (M,N,A). These findings highlight the importance of selecting the appropriate forecasting techniques tailored to the characteristics of the specific time series at hand. In the following section, conclusions are drawn based on these results and potential avenues for future research are discussed.

5. Discussion

The obtained results provide valuable insights into the performance of different forecasting methods in the context of the analyzed time series. The observed high error measurements for the Holt–Winters methods, both additive and multiplicative, suggest that these approaches may not be well-suited for capturing the underlying patterns and dynamics of the given time series. Similarly, the relatively high error metrics observed for Holt's method, ARMA(2,1), and AR(2) indicate their suboptimal performance in capturing the complexities of the analyzed time series. These methods, although widely used, rely on assumptions that might not hold true for every type of time series. Consequently, alternative approaches should be considered for improved forecasting accuracy in similar contexts.

Remarkably, the simple exponential smoothing method exhibited better performance compared to the more complex models. Despite its straightforward nature and constant forecasting values, it demonstrated competitive accuracy in predicting the examined time series. This finding aligns with Makridakis et al.'s [17] study that emphasized the effectiveness of simple methods, which often produced more accurate forecasts compared to complex approaches like ARIMA models.

The top three performing methods, namely the Neural Network, Naïve, and ETS (M,N,A), merit further attention. The ETS ((M,N,A), based on exponential smoothing, incorporates multiple components such as error, trend, and seasonality, and has been successfully applied to various time series forecasting problems. Naïve forecasting, although simplistic in its approach, often serves as a benchmark against which more sophisticated methods are

evaluated. Its competitive performance in this study suggests that even basic forecasting strategies can yield accurate results under certain conditions. Finally, the Neural Network approach, known for its ability to capture nonlinear relationships and complex patterns, displayed promising results, indicating its potential for accurate time series forecasting.

From a broader perspective, these findings underscore the significance of comprehending the characteristics and dynamics of the specific time series when choosing an appropriate forecasting method. In the context of this study, the time series exhibited high volatility, posing challenges for accurate forecasting. Consequently, no single method emerged as universally superior in all scenarios, highlighting the imperative nature of meticulous evaluation and comparison of diverse techniques.

Future research directions in time series forecasting may include investigating the effectiveness methods that combine the strengths of multiple forecasting techniques, exploring hybrid approaches that integrate machine learning and statistical modeling, and considering also the impact of external factors on the forecasting accuracy.

In conclusion, this study offers valuable insights into the performance of various forecasting methods, with implications for practitioners and researchers in the field of time series analysis, particularly in the context of the aeronautical industry, where raw materials play a vital role. The findings highlight the significance of selecting the most suitable method, as even a slight difference in forecasting error can lead to substantial cost savings when planning and procuring essential inputs. By leveraging these findings and considering the suggested future research directions, it becomes feasible to enhance forecasting capabilities and make significant contributions to the advancement of the field of time series forecasting.

Author Contributions: Conceptualization, A.A.R.d.C. and M.A.d.O.; methodology, A.A.R.d.C. and M.A.d.O.; software, A.A.R.d.C. and M.A.d.O.; validation, A.A.R.d.C. and M.A.d.O.; formal analysis, A.A.R.d.C. and M.A.d.O.; investigation, A.A.R.d.C. and M.A.d.O.; resources, A.A.R.d.C. and M.A.d.O.; data curation, A.A.R.d.C. and M.A.d.O.; writing—original draft preparation, A.A.R.d.C. and M.A.d.O.; writing—review and editing, A.A.R.d.C.; visualization, A.A.R.d.C. and M.A.d.O.; supervision, M.A.d.O.; project administration, A.A.R.d.C. and M.A.d.O.; funding acquisition, A.A.R.d.C. and M.A.d.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Embraer and are available from the corresponding author with the permission of Embraer.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Rüßmann, M.; Lorenz, M.; Gerbert, P.; Waldner, M.; Justus, J.; Engel, P.; Harnisch, M. Industry 4.0: The Future of Productivity and Growth in Manufacturing Industries. *BCG—Boston Consulting Group* **2015**, *9*, 2–6.
- Slack, N.; Chambers, S.; Johnston, R. *Operations Management*, 1st ed.; Atlas: Sao Paulo, Brazil, 1996; p. 368.
- Ohno, T. *The Toyota Production System: Beyond Large-Scale Production*, 1st ed.; Bookman: Porto Alegre, Brazil, 1997.
- Montini, A.d.Á.; Fornazza, J.R.; Oliveira, M.A. *Strategic Management of Stocks and Demand*, 1st ed.; IESDE Brasil: Curitiba, Brazil, 2009; p. 161.
- Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 1st ed.; OTexts: Melbourne, VIC, Australia, 2018.
- Holt, C.E. *Forecasting Seasonals and Trends by Exponentially Weighted Averages*, 1st ed.; O.N.R. Memorandum No. 52; Carnegie Institute of Technology: Pittsburgh, PA, USA, 1957.
- Winters, P.R. Forecasting sales by exponentially weighted moving averages. *Manag. Sci.* **1960**, *6*, 324–342. [CrossRef]
- Pegels, C.C. Exponential forecasting: Some new variations. *Manag. Sci.* **1969**, *15*, 311–315.
- Gardner, E.S. Exponential smoothing: The state of the art. *J. Forecast.* **1985**, *4*, 1–28. [CrossRef]
- Taylor, J.W. Exponential smoothing with a damped multiplicative trend. *Int. J. Forecast.* **2003**, *19*, 715–725. [CrossRef]
- Fattah, J.; Enzzine, L.; Aman, Z. Forecasting of demand using ARIMA model. *Int. J. Eng. Bus. Manag.* **2018**, *10*, 1847979018808673. [CrossRef]

12. Makridakis, S.G. Accuracy Measures: Theoretical and Practical Concerns. *Int. J. Forecast.* **1993**, *9*, 527–529. [CrossRef]
13. Bliemel, F. Theil's Forecast Accuracy Coefficient: A Clarification. *J. Mark. Res.* **1973**, *10*, 444–446. [CrossRef]
14. Theil, H. Applied Economic Forecasting. *J. Bus.* **1966**, *39*, 242–253.
15. Makridakis, S.; Wheelwright, S.C.; Hyndman, R.J. *Forecasting: Methods and Applications*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 1998.
16. Christie, D.; Neill, S.P. Measuring and Observing the Ocean Renewable Energy Resource. In *Comprehensive Renewable Energy*; Letcher, T.M., Ed.; Elsevier: Amsterdam, The Netherlands, 2022; pp. 149–175.
17. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 competition: Results, findings, conclusion and way forward. *Int. J. Forecast.* **2018**, *34*, 802–808. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Proceeding Paper

Forecasting System for Inbound Logistics Material Flows at an International Automotive Company[†]

John Anderson Torres Mosquera^{1,*}, Carlos Julio Vidal Holguín¹, Alexander Kressner²
and Edwin Loaiza Acuña³

¹ Faculty of Industrial Engineering, Universidad del Valle, Cali 760032, Valle del Cauca, Colombia; carlos.vidal@correounivalle.edu.co

² Logistics and Supply Chain Management, Duale Hochschule Baden-Württemberg, 70178 Stuttgart, Germany; alexander.kressner@dhbw-stuttgart.de

³ Faculty of Exact Sciences, Universidad del Valle, Cali 760032, Valle del Cauca, Colombia; edwin.loaiza@correounivalle.edu.co

* Correspondence: john.anderson.torres@correounivalle.edu.co

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: This paper analyzes how a robust and dynamic forecasting system was designed and implemented to predict material volumes for the inbound logistics network of an international automotive company. The system aims to reduce transportation logistics costs and improve demand capacity planning for freight forwarders. The forecasting horizon is set for 4 months and 12 months ahead in the future. To solve this problem, a time series modeling approach was carried out by using different time series forecasting methods like ARIMA, Neural Networks, Exponential Smoothing, Prophet, Automated Simple Moving Average, Multivariate Time Series, and Ensemble Forecast. Additionally, important data preprocessing methods and a robust model selection framework were used to train the models and select the best-performing one. This is known as Forward Chaining Nested Cross Validation with origin recalibration. The system performance was assessed using the Symmetric Mean Absolute Error (SMAPE). The final version of the forecasting system can deliver 4-month-ahead forecasts with a SMAPE lower than 10% for 86% of all material flow connections. The system's forecast output is updated on a monthly basis and was integrated into the inbound logistics network system of the company.

Keywords: forecasting system; time series; automated model selection; inbound transportation logistics

Citation: Torres Mosquera, J.A.; Vidal Holguín, C.J.; Kressner, A.; Loaiza Acuña, E. Forecasting System for Inbound Logistics Material Flows at an International Automotive Company. *Eng. Proc.* **2023**, *39*, 75. <https://doi.org/10.3390/engproc2023039075>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 12 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

There are many business factors influencing a company's performance. Among these, accurate forecasts have the greatest impact on an organization's ability to satisfy customers and manage resources cost-effectively [1]. A forecast is not simply a projection of future business; it is a request for products and resources that ultimately impacts almost every business decision the company makes across sales, finance, production, management, logistics, and marketing [2]. An improvement in forecast accuracy, even just one percent, can have a ripple effect across the business, including reducing inventory buffers, obsolete products, expedited shipments, distribution center space, and non-value-added work [2].

Forecasting multiple time series can be challenging since every individual time series can display different properties. Some data might be trended, others might show seasonality. In other cases, data might just have random variations with underlying patterns which are hard to predict. Since there are different models whose properties better match up to particular time series characteristics [2–4], a common approach is to select the most effective and flexible models, blend their best features, and shift between them as needed

to optimize forecast accuracy. Hence, enabling a forecasting system to automatically choose the best forecasting method over time is the best approach [2].

The current research considers an International Automotive Company that produces vehicles in more than 20 assembly plants around the world. The company currently has more than 1000 suppliers worldwide and more than 30 freight forwarders, which deliver different vehicle parts, components, and finished goods to their corresponding consolidation centers in the forwarding areas. To be precise, the company has an Area Forwarding-Based Inbound Logistics Network [5].

The increasing complexity in the Inbound Logistics Network, with regards to the production capacities from suppliers, the transportation availability from freight forwarders, and the changing materials demand from the assembly plants, have increased the need for reliable mid- and long-term capacity planning, especially for the freight forwarders. These are normally the actors in the logistics network with the lowest capacity and the less flexibility to abrupt planning changes. Therefore, a forecasting system focused on freight forwarders' needs was set into place to predict inbound material transportation volumes from suppliers to plants.

Before the implementation of the forecasting system, there was a lack of synchronization between suppliers and freight forwarders, causing over- or under-capacity planning whenever a plant's material demands change abruptly, leading to higher logistics transportation costs. Consequently, forecasting planning values delivered neither in the granularity nor in the frequency the freight forwarders expected.

The forecasting system has evolved since its creation in 2016. There have been three main Versions. Version 1.0 in 2016 which implemented five forecasting methods. Version 2.0 in 2018 implemented two additional forecasting methods, namely [6] and Multivariate Timeseries [7], as well as an automated outlier detection process [8] and a linear interpolation methodology [9]. Finally, Version 3.0 after the Coronavirus pandemic implemented further features related to improving production planning accuracy.

The system performance was assessed using the Symmetric Mean Absolute Error (SMAPE). Comparing the first and third versions, the system improved from 4-month-ahead to month-ahead forecasts with a SMAPE lower than 10% for 18% of all material flow connections, to a SMAPE lower than 10% for 86% of all material flow connections.

This paper is structured as follows: Section 2 points out the forecasting problem and explains the different strategies used in order to develop the forecasting system. Section 3 explains the forecasting system performance along the three versions. Limitations and further development are addressed in Section 4, and, finally, the conclusions are summarized in Section 5.

2. Materials and Methods

2.1. Business Problem Description

The realization of the forecasting system started with building an appropriate problem understanding together with the subject matter experts. Therefore, meetings with them were held in order to learn the most relevant characteristics of the logistic network, the data quality and availability, as well as the specific expectation of a forecasting system designed for the freight forwarders. From a methodological point of view, a time series approach was the most suitable tool since the problem comprises multiple hundreds of monthly time-related data. The system was developed using the programming language R.

The problem concerns the supply chain of an international automotive company with an Area Forwarding Inbound Logistics Network. This network consists of three major participants, (1) the assembly plant which has to be supplied with goods; (2) the suppliers which produce the material required by the plants. The suppliers are classified into groups, most likely regarding their geographical location. Such a group is called area forwarding. (3) Furthermore, the freight forwarder organizes the transportation of materials between the suppliers and the plants. In different areas, different freight forwarders can be hired by the same supplier. The freight forwarder operates a consolidation center within the

area, collects goods from the suppliers, and gathers them in their consolidation center. This action is limited to cross-docking, i.e., there is no warehousing in the consolidation center. The pre-leg or first leg is the transportation step from the supplier to the consolidation center. At this point on, the goods from different suppliers in the area forwarding can be consolidated together. The transportation from the consolidation center to the assembly plants is called main leg transport. If the load in the pre-leg exceeds the volume of one vehicle, the materials are transported directly to the plants. This transportation type is called full truckload [5]. This network structure can be seen in Figure 1.

Syntetos' Supply Chain Structure Framework is well known in the literature to help outline the components of a Logistics Supply Chain when it comes to forecasting [10]. Based on this framework, the company's Inbound Logistics Network can be described as follows [1], (1) at the product dimension level the forecast regards all material components aggregated as tons; (2) at the location dimension it concerns all main leg material flows from the inbound material forwarding areas to the assembly plants; (3) at the time dimension forecasts are generated on a monthly basis for the next following 4 months and 12 months; and finally (4) at the echelon dimension the supply chain level corresponds to the material flows connections among the consolidation centers and the assembly plants.

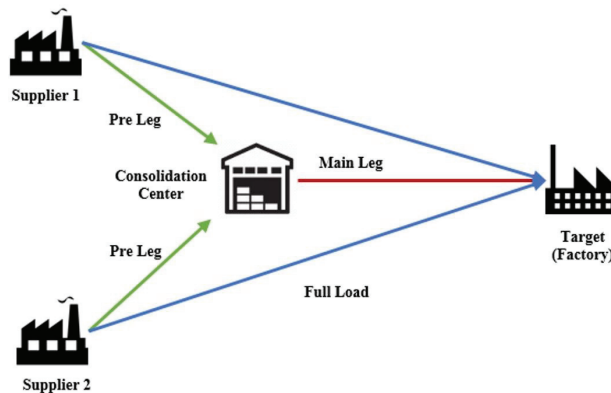


Figure 1. Area Forwarding-based Inbound Logistics Network.

2.2. Time Series Analysis

The project consisted of generating an adaptive and automated forecasting system for more than 400 main-leg material flows. Data were available on a monthly basis since 2014. The material flows display almost all possible demand patterns, i.e., positive and negative trends, seasonality, and irregular demand except for intermittent demand. Additionally, most of the time series contain outliers or missing values.

It is important to point out that a high correlation between the monthly production planning units and the monthly delivered material to the plants can be observed in the data. Another noticeable fact is that the monthly production planning forecasts are available up to 24 months into the future for every plant, giving an idea about how many vehicles are expected to be produced and how much material is expected to be delivered. Therefore, in order to make use of these data, rather than forecasting the monthly material volumes directly, a forecast of the ratio of material volume, and the production units (tons/vehicle units) is carried out:

$$\alpha = \frac{\text{material (tons)}}{\text{production (vehicle units)}} \tag{1}$$

This new time series is then referred to as α time series (1). This is a smoother time series that is able to correct for outliers or extreme events in the material volumes. Finally, the business forecast in material tons is then given by the vehicle's production forecast

multiplied by the \hat{a} time series forecast. It is important to consider that since the vehicle's production forecast is itself uncertain, its error is further propagated through the material volume forecast. This issue is addressed in version 3.0 of the forecasting system [11].

Regarding the error measures, the Mean Squared Error (MSE) was used to choose the best forecasting method in the model selection framework for a given material flow time series [4]. On the other hand, the Symmetric Mean Absolute Percentage Error (SMAPE) was used to evaluate the forecasts from the business perspective. Additionally, the SMAPE is a better estimator of the error than the MAPE when the true value of the forecast is close or equals zero since those tend to generate extremely large errors or infinite values [12]. Time series, with zero transported material volume, are common in this logistics network, and during the coronavirus crisis, it was even more likely to appear. When evaluating forecast accuracy, it is better to have different forecast error measures which can be then compared [4,13]. Therefore, the MSE was used to select the best-performing model; however, the interpretation from the business perspective and, therefore, the impact of the models will be analyzed using the SMAPE.

2.3. Model Selection Framework

According to the Logility Consulting Group [2], for many supply chain scenarios, it is best to employ a variety of methods to achieve optimal forecasts. Ideally, supply chain planners should take advantage of different methods and build them into the foundation of the forecast. The best practice is to use automated methods which switch to accommodate the selection and deployment of the most appropriate forecast method for optimal results. Henceforth, due to the nature of the problem, multiple models are evaluated and then the best-performing model on each material flow connection is selected to generate the monthly forecasts. To be precise, the Forward Chaining Nested Cross Validation with origin recalibration [4,14–17] method was implemented to carry out the model training and testing so that the best model can be chosen to generate the monthly forecasts. This method, explained in Figure 2, is able to replicate the data generation process so that the forecasting system learns in every iteration to select the best-performing forecasting algorithm; in consequence, it can dynamically adapt to short-term changes.

A Nested Cross Validation approach provides an almost unbiased estimate of the true error of a model [17]. This refers to having two for loops in the train–test process; the inner for loop finds the best parameters estimates in the training set, then the outer for loop validates the true accuracy of the model using a rolling test window. Specifically, every time series, of n values, is split up into two sections, the training set and test set. Then, every model is fit using the training set and the best parameters are selected (inner for loop), then the model uses the best parameters to generate a forecast for the test window, and the MSE is calculated. Afterward, the training dataset increases by 1 value, and the test window is also moved 1 position into the future and the process is carried out repeatedly until no more test windows can be generated; this is called the literature forecast origin recalibration [16].

Due to the time dependency between the out-of-sample error measures of the cross-validation tests, a simple average of the resulting errors generates a biased indicator for choosing the best-performing model. Therefore, an exponential weighting approach can be applied to circumvent this problem [14]. The Exponentially Weighted Moving Average (EWMA) is a weighted average of all current and previous forecast errors, whose weights decrease geometrically with the “age” of the forecast error [4]. Therefore, the lowest resulting EWMA MSE is then used to select the best-performing model.

The resulting performance metrics are then stored in a database so that every month only the newest performance metrics for every material flow time series are added. This method enables the reduction in computing time and the forecast output for all the time series can be calculated in less than 30 min using a computer with 32 GB RAM and an Intel Core i7 Processor.

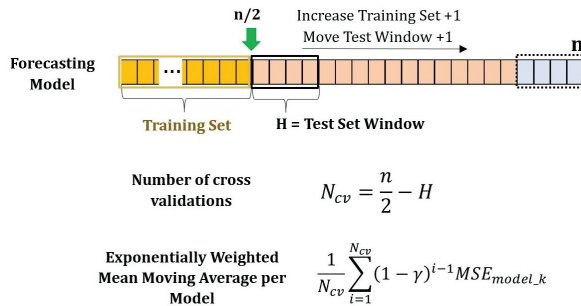


Figure 2. Diagram of the Forward Chaining Nested Cross Validation with rolling origin recalibration described in [4,14–17].

2.4. Forecasting System Version 1.0: First System Implementation

To create a forecasting system for monthly inbound material flows, first of all, meetings with the subject matter experts were held. From which the most relevant results were (1) the definition of the target variable, the monthly freight volume in tons; (2) the scope of the inbound logistics network, main legs; and (3) the forecast horizons, 4 months for mid-term and 12 months for long-term scenarios.

The subject matter experts pointed out the importance of including the production volume forecast as a feature in the forecasting process. For this, the α time series transformation was implemented (see Equation (1)). This accounts for modeling the relationship between the monthly material flows in tons and the vehicles produced by each assembly plant and also the adjustment of extreme values in the times series. This is important since when outliers and missing values are incorrectly handled, they can certainly reduce the forecast accuracy [8,18].

Version 1.0 of the forecasting system was developed using the programming language R. The system focused on forecasting more than 400 main leg material flows within Europe. Furthermore, the model selection framework explained in Section 2 was also set up. This initial framework included the forecasting methods of Naive, ARIMA, Neural Network, Exponential Smoothing, and Ensemble Forecast. The last one refers to the average of the forecasts delivered by the other methods [19].

2.5. Forecasting System Version 2.0: New Forecasting Methods

Version 2.0 of the forecasting system implemented three additional forecasting methods to improve the forecast accuracy, namely Prophet Algorithm, Automated Simple Moving Average, and Multivariate Timeseries Method: Vector Autoregression.

The Prophet Algorithm from Facebook displays two main features, (1) parameters can easily accommodate seasonality with multiple periods and let the analyst make different assumptions about trends, (2) as opposed to ARIMA models, the measurements do not need to be regularly spaced, and missing values do not need to be interpolated, e.g., from removing outliers [6].

On the other hand, there are also important features that are left out when only using univariate methods. For this, Multivariate Methods are able to consider lag-cross correlations among different time series [7]. This cross-correlation feature, along with the historical data, considers the influence of past values of a time series A on the future value of a time series B and vice versa. Since there are multiple suppliers delivering to the same plants, the material quantity delivered from one supplier is highly correlated with the material delivered by other suppliers. This means a relevant cross-correlation between these material flows connections exist and can be exploited by this method.

Furthermore, a simple but useful method still not considered is the Simple Moving Average. The Simple Moving Average is the best model for products whose demand histories have random variations, including no seasonality or trend, or fairly flat demand [2].

However, finding the optimal parameters can be time-consuming. Therefore, using the R package `smooth` can help automate this process.

Additionally, the Ensemble Forecast method, which considers the simple linear combination (simple average) of the forecast values from the other methods, can be also extended, i.e., the Prophet Algorithm, Simple Moving Average, and Multivariate Time Series can also be included in the linear combination so that the likelihood of better forecasts accuracy increases [20].

One additional issue was the elimination of some past values due to a database update to the main Enterprise Resource System (ERP) Database. This leads to incomplete time series. Enabling a linear interpolation algorithm to find the missing values instead of using the mean of the observations can also improve forecasting accuracy. Linear interpolation is easy to implement [18], this enables us to find missing values for the time series in short computing run time. This method is efficient and most of the time is better than non-linear interpolations for predicting missing values [9].

Furthermore, an automated outlier detection and cleaning method was added. A common approach to deal with outliers in a time series is to identify the locations and the types of outliers and then use intervention models [21]. There are some main important issues caused by outliers, i.e., (a) the presence of outliers might result in an inappropriate model, (b) even if the model is appropriately specified, outliers in a time series might still produce bias in parameter estimates and, therefore, might affect the efficiency of outlier detection. A typical problem found in this approach is that both the types and locations of outliers may change at different iterations of model estimation, and (c) some outliers may not be identified due to a masking effect. For problems (b) and (c), Chen and Liu [8] designed a procedure that is less prone to the spurious and masking effects during outlier detection and is able to jointly estimate the model parameters and outlier effects. The approach is to classify an outlier impact into four types, an innovational outlier (IO), an additive outlier (AO), a level shift (LS), and a temporary change (TC). This method can be easily implemented using the R package `tsoutliers`. The process starts with setting SARIMA models to the time series, then the automated outlier detection method is applied to these ARIMA models, which delivers the outliers and their corresponding adjusted value. These adjusted values are then used instead of the outliers and a newly adjusted time series is generated, which can be later used for model training.

2.6. Forecasting System Version 3.0: Production Accuracy Improvement

Version 3.0 of the forecasting system focused on reducing the impact of the coronavirus crisis and the chip crisis by means of handling the increased volatility of both the material flows and the production planning so that reliable forecast values can still be delivered.

According to (Gultekin et. al, 2022), one of the most important freight forwarders' risk areas, caused by the COVID-19 pandemic, was demand fluctuation. The pandemic increased the volatility in supply chain demand planning, making it even harder to generate accurate forecasts [22]. In total, 68% of the respondents on a 1000-company survey made by Capgemini 2020 stated that they experienced difficulties in demand planning due to a lack of data on fluctuating demand [23]. Furthermore, the current chip crisis is also one of the most relevant disruptive factors in the automotive supply chain. Opposed to forecasts, vehicle sales quickly rebounded within just a few months after the pandemic. Henceforth, imperfect inventory planning caused chip shortages and unprecedented halted production cycles [24].

Due to the heavy increase in the demand planning variability [23] post-COVID-19 outbreak, the production forecast has become less reliable. As explained in Section 2, when calculating the material volume forecast, the approach is to multiply the production demand planning by the α time-series forecast. This leads, however, to error propagation since the production demand planning is itself a forecast. Therefore, in order to reduce this effect in the monthly material forecasting system, an additional data preprocessing approach was implemented.

This approach is called Production Planning Error Deviation Adjustment and helps to reduce the error propagation [11,25,26]. Since the production planning forecast is updated on a monthly basis, a database of monthly historical production plans was created. In other words, not only the actual number of vehicles to be produced are available but also the planning values in the previous months. The database consists of the monthly production plans since February 2019.

The approach is quite straightforward. The idea is to track the deviation of an actual production quantity from the forecast values in the past. For example, a plant produced 1000 in December 2020. If the planning value for this particular month is traced back to the previous months, then in some months the planning value would be over 1000 and in others under 1000, due to the variability in demand planning, as well as other internal and external factors. In consequence, using historical data, the relative error deviation to the actual produced cars can be calculated. The month distance from the production planning month to the actual production month will be called lag or planning horizon. Henceforth, let the Relative Planning Error Deviation for a lag l be

$$RPE_l = \frac{\hat{v}_l}{v} \tag{2}$$

where RPE_l = Relative Planning Error for lag l , \hat{v} corresponds to the planning vehicles for lag l , and v to the actual produced vehicles.

This metric can be interpreted as follows. A value greater than 1 indicates that the planning demand was higher than the actual demand, therefore it was a planning overestimation. The opposite is a lower planning value than the actual demand, which is considered a planning underestimation.

Thence, to track the most recent changes in production planning the RPE is calculated for every actual month available for the planning horizons 1 to 12 for every plant. To that end, the mean RPE for lag l for every plant can be computed as

$$\overline{RPE}_l = \frac{1}{n_l} \sum_{i=1}^{n_l} \frac{\hat{v}_l}{v} \tag{3}$$

where n_l is the number of RPE values which could be calculated for the lag l with the available planning data. If the \overline{RPE}_l for a given plant for $lag = 1$ is 1.09, it means that historically the production planning overestimates on average about 9% the number of vehicles to be produced. Henceforth, the planning value can be adjusted by this amount.

Using the properties of the expected value of RPE_l , assuming that the realizations of these errors are independent and identically distributed, an adjusted value of the vehicle's demand planning \hat{v} can be computed. Since the actual number of produced vehicles in a month v is a constant and the expected value of \hat{v} is estimated with the most recent planning value, then the actual number of produced vehicles can be estimated as:

$$E[RPE_l] = E[\hat{v}_l] E\left[\frac{1}{v}\right] \tag{4}$$

$$v \cong \frac{E[\hat{v}_l]}{E[\overline{RPE}_l]} \tag{5}$$

This formula enables us to estimate the true value of v , which the resulting demand planning values are now used to calculate the future monthly volume forecasts.

Figure 3 shows the Relative Mean Error Deviation (Production Forecast/Production Actual) for 4 plants. As expected, the further the forecasting horizon, the lower the quality of the forecasting values. Therefore, using the proposed method helps adjust the production planning data quality, and the forecast error propagation in the forecasting system is reduced therewith.

This was added as a data preprocessing step in the Forecasting System, enabling the future production demand planning values to be adjusted up to 12 months in the future.

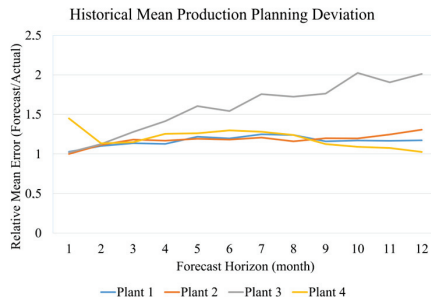


Figure 3. Historical mean production planning deviation.

3. Results

As pointed out in Section 2, the MSE was used to select the best-performing model. To analyze the business impact of the models, we used the SMAPE to compare model performance and improvement, since this enables a more straightforward interpretation of the error deviation. This section analyzes the performance of the forecasting system versions using the historically exponentially weighted moving average (EWMA) of the rolling 4-month-ahead out-of-sample SMAPEs. When comparing version A vs. version B, both systems’ versions were used to generate the forecasts for multiple months in the past using the same amount of data, then the EWMA is then calculated. With this, the EWMA-SMAPE Category distribution on the Material flow connections and the Cumulative EWMA-SMAPE Distribution is generated, as explained in the following sections.

3.1. Comparison Performance Version 1.0 vs. 2.0

The performance of the Forecast System Version 1.0 vs. Version 2.0 can be seen in Table 1.

Table 1. EWMA-SMAPE Category Distribution on Material Flow Connections for Version 1.0 and 2.0.

EWMA-SMAPE Category	Version 1.0	Version 2.0
lower than 10%	18.0%	43.6%
between 10% and 20%	48.7%	36.5%
between 20% and 30%	20.2%	11.5%
between 30% and 40%	7.6%	3.9%
higher than 40%	5.6%	4.5%

To be precise, Table 1 shows the distribution of the EWMA SMAPE in groups. It is worth noting that the number of material flow connections with an EWMA SMAPE of less than 10% increased from 18.0% to 43.6%, i.e., about 25.6 pp (percentage points). Moreover, the number of material flow connections with an EWMA SMAPE higher than 40% decreased from 5.6% to 4.5%. Quantitatively, 80.1% of all material flows had an EWMA SMAPE of less than or equal to 20%, in comparison with the 66.7% of all material flows which had the same behavior when using version 1.0 of the forecasting system.

Regarding the different improvement steps carried out in version 2.0, Table 2 summarizes the results. The first column "Algorithms" presents the approaches used, whether it was a single forecasting method, a combination of multiple methods, or the implementation of data pre-processing steps. All tests were elaborated with the most recent data available at that moment. Furthermore, the column "Averaged EWMA SMAPE Improvement" shows the average percentage change for the EWMA SMAPE in the corresponding material flow connections selecting the new approach, which is found in the column "Material Flow Timeseries".

Table 2. Improvement Results Forecasting System Versions 2.0 vs. 1.0.

Algorithms	Average EWMA-SMAPE Improvement	Material Flow Timeseries
Prophet Algorithm	8.3%	9.0%
Automated Simple Moving Average	3.8%	17.4%
Multivariate Time Series	23.8%	6.8%
Prophet + SMA + MVTS	7.8%	42.8%
P + SMA + MVTS + Outliers detection	22.3%	66.4%
P + SMA + MVTS + Outliers detection + Interpolation	24.8%	67.6%

First, the prophet algorithm was implemented; using the R package prophet [6]. This forecasting method showed an improvement of 9% in the material flows for monthly forecasts, with an average EWMA SMAPE improvement of 8.31%. Later, the algorithm Simple Moving Average (SMA) was introduced into the system. The function SMA from the R-Package smooth applies the Simple Moving Average method on a time series vector [27,28]. The SMA order was set to be chosen automatically by the function, which chooses the optimal one. In total, 24% of the material flows chose the SMA instead of the old methods. The average EWMA-SMAPE improvement was 3.78%.

For the implementation of the Multivariate Time Series, a vector autoregressive method was used. There are a couple of things which must be considered in advance. First, The Ljung-Box test is used to test the lag-cross correlation along n time series. The time series are divided into groups that are more likely to have the highest lag-cross correlation coefficient, namely, all the material flows coming to a single plant. Secondly, the automated vector autoregressive method might break down if too many time series with too few values are calculated. Explicitly, the algorithm takes up a large amount of memory and long runtime to calculate all the parameters involved in the matrices. Moreover, the number of lags consider to fit the model also affects the algorithm performance, which is why a 1-lagged automated vector autoregressive model was implemented in this case. Therefore, another routine was implemented to eliminate the parameters with a significance level lower than 5%. This step improved the model accuracy, as well as the final forecast errors. For this approach, 6.8% of the material flows realized a lower EWMA MSE cross-validation accuracy rather than using the old methods, that is an averaged EWMA SMAPE improvement of 23.78%.

Henceforth, the three new forecasting methods were tested together for all the material flow connections, for which, 42.8% of the connections displayed higher performance when choosing the new methods. This performance was translated into a 23.78% averaged EWMA SMAPE improvement in all four-step-ahead out-of-sample tests for data available.

Afterward, the data preprocessing methods were introduced. First, the automated outlier detection method together with the new forecasting algorithms was tested. These steps provided an improvement of 66.4% of the material flows with an average improvement for the EWMA SMAPE of 22.25%.

Later, the interpolation method for missing values was included delivering that 67.6% of the material flows chose the three new forecasting methods with an average improvement of 24.84% on the EWMA SMAPE. This can be seen in Figure 4. When plotting the cumulative EWMA SMAPE for all the material flow connections, the new forecasting system's version shows a curve laying higher and more to the left of the graphic than version 1.0. This can be interpreted as more material flow connection forecasts displaying lower forecast errors.

Finally, from the company's point of view, the SMAPE improvement has a greater impact on cost and demand capacity planning reduction when this is lower for material

flow connections which transport on average more than 1000 tons monthly. Therefore, the logistics planners can better assess the forecasting system performance using a plot that shows the relationship between SMAPE performance and monthly average transported volume for every material flow connection. Figure 5 plots the average monthly material volume in tons and the EWMA SMAPE; every dot represents a material flow connection. From this plot, it can be implied that version 2.0 of the forecasting system delivered better results than version 1.0, in which almost all blue dots realized a lower four-month-ahead out-of-sample EWMA SMAPE than the red dots.

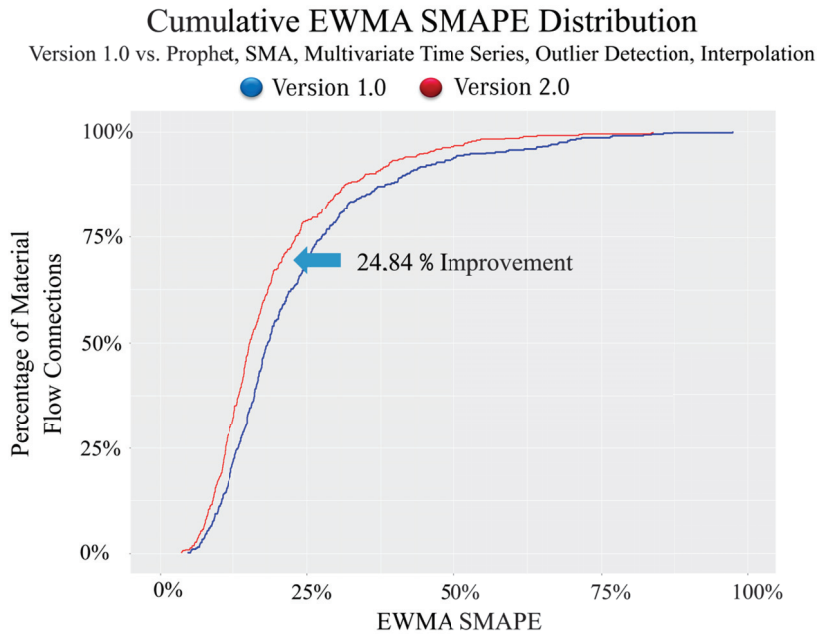


Figure 4. Phase 1: EWMA MAPE comparison original forecasting system vs. improvements—Stand 2018.

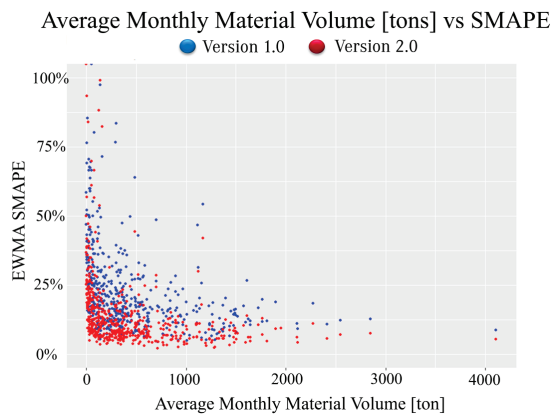


Figure 5. Plot for Average monthly volume vs. EWMA SMAPE for Version 1.0 and 2.0.

3.2. Comparison Performance Version 2.0 and 3.0

The comparison of the Version 2.0 and Version 3.0 was carried out after the coronavirus crisis, at this time the structure of the inbound logistics network changed substantially, which is why the number of total material flows changed. The two versions of the system

were assessed using new data, which is why the performance of Version 2.0 differs from that in the previous section.

The production accuracy improvement approach helped further improve the forecasting system. Considering that this approach traces the planning relative error deviation, future planning values are better estimated; thus, the monthly forecasts are more accurate. Table 3 summarizes the results after applying this methodology. When comparing the out-of-sample EWMA SMAPE values the new adjustment shows an improvement of 25.4% for material flow connections with an SMAPE lower than 10%. Furthermore, the number of material flow connections with a SMAPE greater than 40% was reduced from 2.7% to 0.2%. The major reason behind these improvements is due to two main factors, (1) the data input to the models uses the α time series, which considers the historical production volume as an influencing factor; and (2) the final forecast is given by the production planning forecast times the α time series forecast; the error propagation caused by the production planning is then highly reduced when applying the production accuracy improvement approach.

Finally, Figure 6 shows the performance of the forecasting system before and after applying this new approach, regarding the average monthly material volume and the SMAPE. It can be stated that the red dots representing the new forecasting system Version 3.0 realized a lower EWMA SMAPE over most material flow connections. Furthermore, material flow connections with more than 1000 tons on average also reached better performance, which can be directly translated into better performance planning in the inbound logistic network reducing logistic costs and capacity planning efforts.

Table 3. EWMA-SMAPE Category Distribution on Material Flow Connections for Version 2.0 and 3.0.

EWMA-SMAPE Category	Version 2.0	Version 3.0
lower than 10%	60.3%	85.7%
between 10% and 20%	29.8%	11.7%
between 20% and 30%	5.8%	1.8%
between 30% and 40%	1.4%	0.6%
higher than 40%	2.7%	0.2%

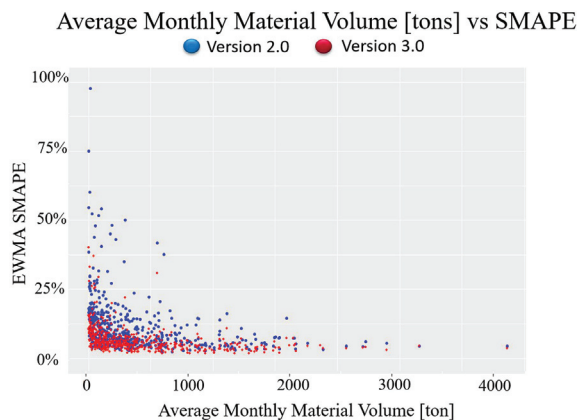


Figure 6. Plot for Average monthly volume vs. EWMA SMAPE for Version 2.0 and 3.0.

4. Discussion

It was observed that after COVID-19, the prophet algorithm was the less recommended algorithm by the forecasting system; even in some months no material flow time series was predicted by it; i.e., the prophet algorithm was not capable of automatically adapting to the abrupt changes in the time series. This can be due to the decomposable time series model [6] in which the Facebook prophet is based, which resembles more than a mere curve fitting that does not take into account the conditional dependency of past realizations [29].

Further development of the forecasting system can be achieved by exploding the graph dependency structure of the data. Since capacity and production restrictions affect all plants, that means that future production capacity reductions in one plant can affect the production planning in another one. The use of most modern algorithms can help improve the forecasting accuracy and flexibility of the whole system. One proposed architecture is the use of Graph Neural Networks, which has been proven to successfully model graph-structured data [30].

5. Conclusions

The current research developed, designed, and implemented a monthly material flows forecasting system for the inbound logistics network of an international automotive company using multiple forecasting algorithms and robust data preprocessing routines. The system was also improved along with the changes in the market and adjusted to the company's needs and challenges to deliver the highest forecasting accuracy, which was assessed using the Symmetric Mean Absolute Error (SMAPE). The output of the forecasting system was integrated into the inbound logistics system of the company delivering newly forecast values for the freight carriers on a monthly basis. This enabled the freight forwarders to better plan their capacities in the mid- and long-term (4-month-ahead and 12-month-ahead forecasts are delivered on a monthly basis) scenarios. Furthermore, the system supports the company by reducing logistics transportation costs and improving demand capacity planning since the material planning volume better meets the freight forwarder's capacity.

Regarding the performance of the Forecasting System in the different versions; for 4-month-ahead forecast values, it can be seen that the number of material flows with an average EWMA-SMAPE of less than 10% increased through the different versions. From Version 1.0 to Version 2.0 the number of material flows with this performance increased from 18% to 43.6% (25.6 pp), whereas from Version 2.0 to Version 3.0 it increased from 60.3% to 85.7% (25.4 pp). This impact can be assessed using the Dupont Equation, which states that a 10% increase in forecasting accuracy can be translated into a return of shareholder value between 39% and 47% [2].

Furthermore, in the forecasting system's versions 2.0 and 3.0, the methods having the highest impact on forecasting performance are those related to improving the data quality, i.e., the automated outlier detection procedure and the data interpolation method, which helped increase the impact of the three new algorithms from an average EWMA-SMAPE improvement of 7.8% of all material flow connections to 22.3%. In addition, the approach regarding production accuracy improvement helped increase the number of materials flows with an EWMA SMAPE of less than 10%. This result proves that when it comes to forecasting even the simplest method can deliver high performance if the quality of the input data is high enough.

The former assertion implies that the success of the forecasting system was not focused on the forecasting methods themselves but rather on the problem understanding, the data modeling, and the data preprocessing steps. Among these, we can highlight the use of the α time series, the automatic outlier detection methods, and the error propagation correction for the production volume planning data. Enabling the system to robustly model the problem and adapt flexibly to upcoming supply chain disruption, not only from a mathematical but also from a business perspective, is the key to creating a highly performing forecasting system.

Author Contributions: Conceptualization, J.A.T.M. and A.K.; methodology, J.A.T.M., A.K., C.J.V.H. and E.L.A.; software, J.A.T.M. and A.K.; validation, J.A.T.M., A.K., C.J.V.H. and E.L.A.; formal analysis, J.A.T.M. and A.K.; investigation, J.A.T.M. and A.K.; resources, J.A.T.M.; data curation, J.A.T.M. and A.K.; writing—original draft preparation, J.A.T.M.; writing—review and editing, A.K., C.J.V.H. and E.L.A.; visualization, J.A.T.M.; supervision, A.K., C.J.V.H. and E.L.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Syntetos, A.A.; Babai, Z.; Boylan, J.E.; Kolassa, S.; Nikolopoulos, K. Supply chain forecasting: Theory, practice, their gap and the future. *Eur. J. Oper. Res.* **2016**, *252*, 1–26. [CrossRef]
2. Logility. Eight Methods That Improve Forecasting Accuracy Eight Methods that Improve Forecasting Accuracy. Available online: https://www.logility.com/wp-content/uploads/dlm_uploads/2018/09/Eight-Methods-to-Improve-Forecasting-Accuracy-in-2019-Logility2019.pdf (accessed on 6 July 2023).
3. Hyndman, R.J.; Khandakar, Y. Automatic Time Series for Forecasting: The Forecast Package for R. *J. Stat. Softw.* **2008**, *27*, 1–22. [CrossRef]
4. Montgomery, D.; Jennings, C.; Kulahci, M. *Introduction to Time Series Analysis and Forecasting*; John Wiley & Sons: Hoboken, NJ, USA, 2008; p. 472.
5. Schöneberg, T.; Koberstein, A.; Suhl, L. An optimization model for automated selection of economic and ecologic delivery profiles in area forwarding based inbound logistics networks. *Flex. Serv. Manuf. J.* **2010**, *22*, 214–235. [CrossRef]
6. Taylor, S.J.; Letham, B. Forecasting at Scale. *PeerJ* **2017**, *5*, e3190v2. [CrossRef]
7. Tsay, R.S. *Multivariate Time Series Analysis with R and Financial Applications*; John Wiley and Sons: Hoboken, NJ, USA, 2014.
8. Chen, C.; Liu, L.M. Joint Estimation of Model Parameters and Outlier Effects in Time Series. *JASA J. Am. Stat. Assoc.* **1993**, *88*, 284–297. [CrossRef]
9. Gnauck, A. Interpolation and approximation of water quality time series and process identification. *Anal. Bioanal. Chem.* **2004**, *380*, 484–492. [CrossRef]
10. Punia, S.; Singh, S.P.; Madaan, J.K. A cross-temporal hierarchical framework and deep learning for supply chain forecasting. *Comput. Ind. Eng.* **2020**, *149*, 106796. [CrossRef]
11. Lorenzato de Oliveira, J.F.; Pacífico, L.D.S.; Gomes de Mattos Neto, P.S.; Barreiros, E.F.S.; Rodrigues, C.M.d.O.; Filho, A.T.d.A. A hybrid optimized error correction system for time series forecasting. *Appl. Soft Comput.* **2020**, *87*, 105970. [CrossRef]
12. Vladik, K.; Hung, T.; Nguyen, R.O. *How to Estimate Forecasting Quality: A System-Motivated Derivation of Symmetric Mean Absolute Percentage Error (SMAPE) and Other Similar Characteristics*; Technical Report: UTEP-CS-14-53; University of Texas at El Paso: El Paso, TX, USA, July 2014.
13. Fildes, R.; Goodwin, P. Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces* **2007**, *37*, 570–576. [CrossRef]
14. Hjorth, U. Model Selection and Forward Validation. *Scand. J. Stat.* **1982**, *9*, 95–105.
15. Bergmeir, C.; Benítez, J.M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **2012**, *191*, 192–213. [CrossRef]
16. Tashman, L.J. Out-of-sample tests of forecasting accuracy: An analysis and review. *Int. J. Forecast.* **2000**, *16*, 437–450. [CrossRef]
17. Varma, S.; Simon, R. Bias in Error Estimation When Using Cross-Validation for Model Selection. *BMC Bioinform.* **2006**, *7*, 91. [CrossRef] [PubMed]
18. Lepot, M.; Aubin, J.-B.; Clemens, F.H.L.R. Interpolation in Time Series: An Introductory Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment. *Water* **2017**, *9*, 796. [CrossRef]
19. Claeskens, G.; Magnus, J.R.; Vasnev, A.L.; Wang, W. The forecast combination puzzle: A simple theoretical explanation. *Int. J. Forecast.* **2016**, *32*, 754–762. [CrossRef]
20. Smith, J.; Wallis, K.F. A simple explanation of the forecast combination puzzle. *Oxf. Bull. Econ. Stat.* **2009**, *71*, 331–355. [CrossRef]
21. Box, G.; Tiao, G.C. Intervention Analysis with Applications to Economic and Environmental Problems. *J. Am. Stat. Assoc.* **1975**, *70*, 70–79. [CrossRef]
22. Gultekin, B.; Demir, S.; Gunduz, M.A.; Cura, F.; Ozer, L. The logistics service providers during the COVID-19 pandemic: The prominence and the cause-effect structure of uncertainties and risks. *Comput. Ind. Eng.* **2022**, *165*, 107950. [CrossRef] [PubMed]
23. Gya, R.; Lago, C.; Becker, M.; Junghanns, J.; Petit, J.-P.; Perea, L.; Schneider-Maul, R.; Dahlmeier, S.C.; Kumar, V.; Penka, A. et al. Fast Forward Rethinking Supply Chain Resilience for a Post-COVID-19 World. Available online: https://www.capgemini.com/wp-content/uploads/2020/11/Fast-forward_Report.pdf (accessed on 6 July 2023).
24. Alam, S.F.; Crean, S.; LeBlanc, J.; Naik, V. The Long View of the Chip Shortage: Building Resiliency in Semiconductor Supply Chains. Available online: https://www.accenture.com/_acnmedia/PDF-159/Accenture-The-Long-View-Of-The-Chip-Shortage.pdf (accessed on 6 July 2023).
25. Liu, H.; Chen, C. Data processing strategies in wind energy forecasting models and applications: A comprehensive review. *Appl. Energy* **2019**, *249*, 392–408. [CrossRef]
26. De Mattos Neto, P.; Cavalcanti, G.; Madeiro, F. Nonlinear Combination Method of Forecasters applied to PM Time Series. *Pattern Recognit. Lett.* **2017**, *95*, 65–72. [CrossRef]
27. Svetunkov, I. *Statistical Models UNDERLYING Functions of ‘Smooth’ Package for R*; Workingpaper; Lancaster University Management School: Lancaster, UK, 2017.

28. Svetunkov, I.; Petropoulos, F. Old dog, new tricks: A modelling view of simple moving averages. *Int. J. Prod. Res.* **2017**, *56*, 1–14. [CrossRef]
29. Seitz, S. Online: Facebook Prophet, COVID and Why I Do not Trust the Prophet. 2022. Available online: <https://www.sarem-seitz-com.cdn.ampproject.org/c/s/www.sarem-seitz.com/facebook-prophet-covid-and-why-i-dont-trust-the-prophet/amp/> (accessed on 6 July 2023).
30. Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Zhang, C. Graph WaveNet for Deep Spatial-Temporal Graph Modeling. *arXiv* **2019**, arXiv:1906.00121. <https://doi.org/10.48550/arXiv.1906.00121>

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Integrating Seasonal Adjustment Approaches of Official Surveys on Labor Supply and Demand [†]

Cinzia Graziani, Annalisa Lucarelli *, Maurizio Lucarelli, Emilia Matera and Andrea Spizzichino

Istat, Via Cesare Balbo 16, 00182 Rome, Italy; cingraziani@istat.it (C.G.); maurizio.lucarelli@istat.it (M.L.); ematera@istat.it (E.M.); spizzich@istat.it (A.S.)

* Correspondence: anlucare@istat.it

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: This paper illustrates the application of the indirect seasonal adjustment approach to the index series of hours worked per capita from the Istat VELA survey, which is currently seasonally adjusted using the direct approach instead. The experience already gained during the Istat LFS allowed us to test the reliability of the indirect approach on the VELA series. In this case, the use of the indirect approach was twofold: firstly, the seasonally adjusted index series was obtained by seasonally adjusting the series of hours and the series of the employed separately and then relating them. Secondly, for the numerator, as well as for the denominator, the different series disaggregated by the variables of interest were seasonally adjusted separately.

Keywords: seasonal adjustment; direct approach; indirect approach; per capita hours worked indicator

1. Introduction

The quarterly Istat official survey on job vacancies and hours worked (VELA) collects information on hours worked and Short-Time Working Allowance (Cig in Italian). Two main indices on labor input can be derived from this survey: the number of hours worked and hours worked per capita. These indices are on a fixed basis and can be obtained for each economic activity sector as a ratio between the value of the indicator in the reference quarter and the average value of the base year (currently 2015). The number of hours worked is the sum of the hours worked by employees (ordinary and extraordinary). The hours worked per capita are obtained by dividing the total hours worked with the average number of employee positions occupied in the quarter.

Following the Eurostat guidelines, the direct approach is advisable when the component series have similar characteristics (Eurostat, 2015). Conversely, in cases where the series characteristics are very different, it is preferable to use an indirect approach, recommended when the seasonally adjusted aggregate also contains component series that show a weak seasonality that is difficult to identify.

The use of one approach or the other when the series are similar leads to results with negligible discrepancies. Otherwise, when series differ, the discrepancies reflected in the seasonally adjusted aggregate series may be significant: it is often the case that a relevant number of inconsistencies (also known as out-of-range data) are observed between the quarter-on-quarter changes in the aggregated activity sectors and those in each of their component sections.

The impact of the COVID emergency, which affected the economic activity section in various significant ways, revealed some of the limits of the direct approach. For the number of hours worked, switching from the direct to the indirect approach is relatively simple, while the implementation of the indirect approach for per capita hours worked is more complex and is the object of this work.

The experience gained from the seasonal adjustment of a per capita variable in the ISTAT Labor Force Survey (LFS) has been very useful. In particular, the indirect seasonal

Citation: Graziani, C.; Lucarelli, A.; Lucarelli, M.; Matera, E.; Spizzichino, A. Integrating Seasonal Adjustment Approaches of Official Surveys on Labor Supply and Demand. *Eng. Proc.* **2023**, *39*, 76. <https://doi.org/10.3390/engproc2023039076>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 12 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

adjustment method adopted is based on the seasonal adjustment of each single series separately (for the numerator as well as for the denominator), and the series can then be re-aggregated according to the area of interest. For example, the seasonally adjusted (SA) number of hours worked per capita can be obtained as the ratio of the SA hours worked to the SA number of employed persons.

This paper describes the integration experience—and the resulting synergies—of the seasonal adjustment method between two official surveys on labor supply (LFS) and demand (VELA).

In particular, the next section deals with a comparison of when the application of the direct and indirect seasonal adjustment approaches are most recommended.

Section 3 describes in detail the indirect method, adopted by the LFS, for seasonally adjusting the hours worked per capita series and its application to the VELA ones.

In the Section 4, the main revision measures are considered in order to assess the impact of the transition to the indirect seasonal approach on the VELA series.

Section 5 shows the main results so far. Conclusion are drawn in Section 6.

2. Seasonal Adjustment Approaches: Direct Versus Indirect

Official economic indicators are represented within a system characterized by elementary series (or components), the aggregation of which results in sub-total or total (marginal) series. For example, economic series are disaggregated according to the NACE Rev.2 classification of economic activities by the specific sector of economic activity (the economic activity sections) and their aggregations by macro sector.

Aggregated, or marginal, time series can be treated for seasonal effects using two main approaches, with different properties and results [1,2] (Eurostat 2015, *Metodi e Norme*, Oros 2019):

- The direct approach consists of individually seasonally adjusting all series, both elementary and aggregates. A possible indicator would first be calculated by aggregating the raw data of its component series and then the data would be seasonally adjusted.
- The indirect approach involves combining two or more seasonally adjusted series. The indicator in this case would be calculated by combining the component series after they have been seasonally adjusted separately.

In practice, it is also possible to use a combination of the two approaches: for example, when the quality of seasonal adjustment cannot be guaranteed at the lowest level of detail, one can consider using a direct approach up to a certain level and an indirect approach for higher levels of aggregation.

These two methods do not lead to the same results and are not equivalent. With the direct approach, which is easier to implement, seasonal adjustment is applied directly to the series of interest. If, for example, the raw series are not additive, the direct method is simpler and more transparent. If the raw series are additive, the indirect approach guarantees by construction that the sum of the seasonally adjusted components is equal to the seasonally adjusted aggregate, since the aggregate is obtained by summing the components. Moreover, seasonal adjustment with the direct approach can lead to inconsistency problems between aggregate and disaggregated data, generating inconsistencies (also known as out-of-range data), although these can be overcome by applying appropriate reconciliation techniques. On the contrary, as mentioned above, with the indirect approach, the internal consistency between aggregate and component series is always respected by construction.

The decision to apply one approach or the other must be made while taking into account the characteristics of the raw series and the consistency between aggregates at different levels. The choice of one approach or the other is an open question: there is no theoretical or empirical evidence in favor of one or the other. For each case, different assessments have to be made according to statistical and other considerations, empirical rules and criteria, and certain properties to be obtained a priori. According to Eurostat, if the seasonally adjusted component series have similar trends, the direct approach is

preferred; if, on the other hand, they have very different characteristics and vary in weight over time, the indirect approach is preferred.

3. Integration of the Indirect Approach on Hours Worked: The Labor Force Survey Experience Applied to the Vacancies and Hours Worked One

With a view to the integration and harmonization of the methodologies applied to different surveys in Istat, it was possible to share with the quarterly survey on vacancies and hours worked (VELA) the experience gained within the framework of the Labor Force Survey (LFS) on the seasonal adjustment of the variable hours worked.

The Labor Force Survey makes available comparable time series of data from 2004 onwards, both monthly and quarterly, for the main aggregates associated with the labor market. In addition, the number of hours usually worked and the number of hours actually worked by the employed, detailed with respect to their main characteristics, are collected.

3.1. The Indirect Approach in the Seasonal Adjustment of Hours Worked per Capita in the LFS

During the period of the COVID pandemic, it was realized that information on employment trends alone was not exhaustive for analyzing the labor market and that the analysis of hours worked could be useful. In fact, along with employed persons, those affected by lay-offs (in Cig) and those who had stopped working, either partially or completely, were counted. On the other hand, the availability of the series on actual hours worked, in particular by referring to the hours per capita, provided a timely, dynamic, and easily interpretable reading of the labor input.

Therefore, in order to assess the impact on the productivity of the employed, the monthly information on the number of employed people was supplemented by information on the number of hours actually worked.

However, in order to properly analyze the trend, it was necessary to adjust the series of hours worked to account for seasonal effects. The production of monthly, seasonally adjusted data on hours worked was made possible by the use of internationally established seasonal adjustment techniques.

For the seasonal adjustment of the indicator on hours per capita, the indirect approach was used. The decision to use the indirect method arose from the strong difference between the seasonality that characterizes the series of the total hours in the numerator and that of the employed in the denominator. In these cases, international best practices on seasonal adjustment recommend that, instead of proceeding directly on the aggregate series, we proceed separately on the components and then aggregate them to obtain the aggregate seasonally adjusted series. In the case of hours per capita, the use of the indirect approach was twofold: firstly, the seasonally adjusted index series was not obtained by seasonally adjusting the raw index series (i.e., the series of ratios between the raw series of hours at the numerator, and the raw series of employed persons at the denominator—direct method). It was obtained instead by seasonally adjusting the series of hours and the series of the employed separately and then relating them. Secondly, for the numerator, as well as for the denominator, the different series disaggregated by the variables of interest were seasonally adjusted separately.

The formulae below summarize the two different approaches. Let X denote the series of total hours and Y that of the employed, disaggregated according to one or more variables of interest. Let M then be a mode of a variable of interest, expressible as the sum of n subcategories $\{M_1, \dots, M_i, \dots, M_n\}$, for which:

$$X_M = \sum_{i=1}^n x_{M_i} \text{ and } Y_M = \sum_{i=1}^n y_{M_i}$$

If by $SA(I)$ we denote the seasonally adjusted I index of hours per capita, the direct approach can be written as:

$$SA(I_M) = SA\left(\frac{X_M}{Y_M}\right), \quad (1)$$

while the indirect approach can be expressed as:

$$SA(I_M) = \frac{\sum_{i=1}^n SA(x_{M_i})}{\sum_{i=1}^n SA(y_{M_i})} \tag{2}$$

The breakdown variables chosen are those most associated with the indicator of interest; in particular, for the employed and hours worked, the use of a logistic model identified the variables relating to gender, professional position (employee or self-employed), and type of working time (part-time and full-time) as most explanatory.

By way of example, Figure 1 shows the difference in the seasonal patterns of the hours worked by male and female part-time employees (a vs. b) and of the number of part-time employees who were male and female (c vs. d). At the same time, it is possible to highlight the different seasonal patterns between the number of hours worked and the number of part-time employees who were male (a vs. c) and female (b vs. d).

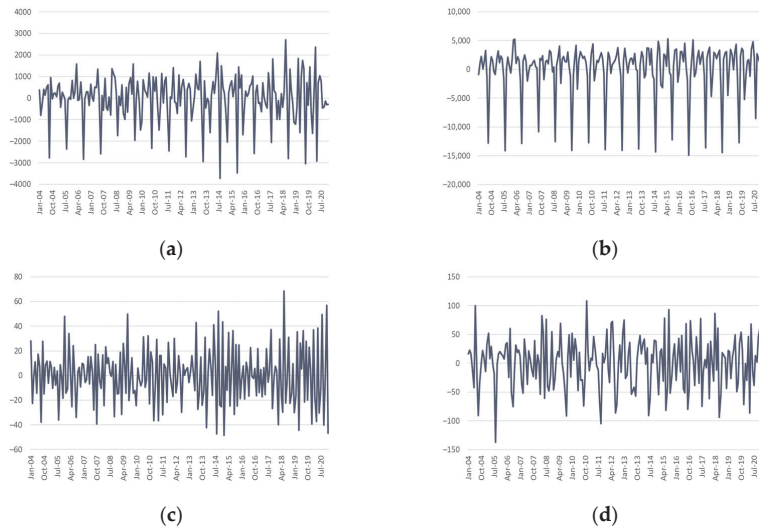


Figure 1. Seasonal pattern of part-time employees: hours worked by males (a) and females (b); number of part time employees who were male (c) and female (d). Monthly series, Jan 2004–Dec 2020 (in thousands).

3.2. The Indirect Approach in VELA

Different seasonal patterns are observed also in VELA data series (Figure 2); in particular, the seasonality of the number of hours (a and b) and employees (c and d) in the economic activity sectors H (a and c) and I (b and d) are shown here. Sector H (transport and storage) and sector I (accommodation and food service activities) are two of the main sectors contributing to the market services aggregate (sectors G to N), with a weight of between 10 and 15%. The difference in seasonality emerges both in the hours worked in the two sectors (a vs. b), with sector H (a) having almost no seasonality, and in the number of employees employed in them (c vs. d). The comparison of hours worked and employees employed within the same sector (a vs. c for sector H and b vs. d for sector I) also reveals strong seasonal differences that justify the application of the indirect approach.

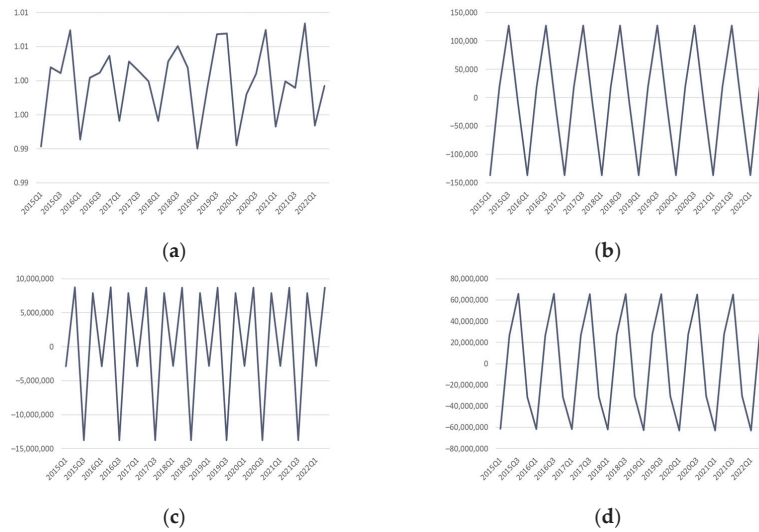


Figure 2. Seasonal patterns in sector H and I: hours worked (a,b); number of employees (c,d). Quarterly series, 2015Q1–2022Q2.

The use of the indirect method also makes it possible to re-aggregate the number of seasonally adjusted hours and number of employees according to the detail of interest, and to obtain, through a ratio of these figures, the number of hours actually worked per capita, adjusted for seasonal effects.

The quality of this methodology was assessed using all of the indicators defined in the literature. In particular, in the seasonal adjustment phase of the single series, the guidelines established by Eurostat [1] for model definition, outlier detection, the use of calendar effects where present, etc. were respected, while in the evaluation phase of the results, checks on residual seasonality were carried out, as well as a revision analysis.

For the seasonal adjustment of each series, the TRAMO/SEATS algorithm [3] implemented in JDemetra+ was used.

This algorithm, adopted by leading European statistical agencies and central banks, allows each series to be decomposed into its stochastic, not directly observable components: the trend cycle, the seasonality, and the erratic component. The decomposition of the series is achieved via a parametric ARIMA model-based procedure. The first stage of the method (TRAMO) allows for the pre-treatment of the series: calendar-related systematic components and possible outliers are identified. It also performs the automatic selection of the ARIMA model and estimation of its parameters, as well as regression coefficients related to outliers and calendar-related systematic effects. In the second step (SEATS), the linearized series obtained from the pre-treatment is then decomposed into its cyclo-trend, seasonal, and irregular components.

In the case of the Labor Force Survey, and for the following application to the VELA series, the presence of calendar effects was checked only for the series relating to hours actually worked, for which it is plausible that the presence of holidays results in fewer hours worked. In contrast, for the denominator series relating to the employed, the absence of calendar effects was imposed because, by definition, an individual's employment status is not affected by the number of working days but depends on having worked at least one hour in the reference week.

TRAMO's automatic identification procedure was relied upon to identify outliers, subject to ex-post evaluation of their significance and eligibility.

As mentioned above, once the seasonally adjusted series have been obtained, the application of the indirect method allows for the aggregation of the series of total hours

with that of employment according to the detail of interest and the construction of the index of hours actually worked per capita. Then, the series obtained indirectly by aggregation or by ratios are again processed in TRAMO/SEATS to check the absence of residual seasonality. The possible presence of residual seasonality would require a new seasonal adjustment of the series in search of the optimal decomposition, with negative seasonality tests at the aggregation stage.

All of the above regarding the methodology adopted in the Labor Force Survey for the production of the hours actually worked per capita was applied to the VELA survey series.

The quarterly Istat official survey on job vacancies and hours worked (VELA) collects information on hours worked and Cig hours. Two indices regarding labor input are derived from this survey: the number of hours worked and hours worked per capita. The indices are on a fixed basis and are obtained for each economic activity section as a ratio of the value of the indicator in the reference quarter to the average value of the base year (2015). The number of hours worked is the sum of the hours worked by employees (ordinary and extraordinary). The hours worked per capita are obtained by dividing the total hours worked by the average number of employee positions occupied in the quarter.

A direct approach is currently used in the VELA Survey to produce a seasonally adjusted index of hours worked per capita.

The exercise on VELA involved adopting the indirect approach but using disaggregated numerator and denominator series for the economic activity sector only. Specifically, 17 numerator and denominator series corresponding to 17 macro-sectors of economic activity identified according to the NACE Rev.2 classification were seasonally adjusted.

In this case, at the aggregation stage, the absence of residual seasonality was checked both in the aggregate series at the sector level, separately for total hours and number of employees, and in the index series given by the ratio of total hours actually worked to the number of employees. The aggregation covers, in detail, macro-sectors B to E; B to F; B to N; B to S, G, H, and I; G to N; G to S; M and N, L, M, and N; and P to S.

Once the indicator series had been obtained via the indirect method, an initial validation was carried out by comparing them with the same ones produced by the direct method and then also with those derived from other closely related surveys, in order to assess their consistency (for more details, see Section 5). The comparison was carried out on the series of cyclical changes by using indicators usually applied in the revisions analysis that allowed us to assess the magnitude and significance of the occurred differences between different releases of the same series.

In the following sections (see Section 5), the indicators resulting from the comparisons will be detailed. Through an analysis of them, it will be possible to give an assessment of the eligibility of the indirect method in the seasonal adjustment of hours actually worked per capita derived from the VELA survey.

4. Revision Measures Applied

The quality of the new VELA seasonally adjusted series has been assessed by means of the main revision standard measures [4,5].

In particular, to assess the average magnitude of revision, the following were considered: the mean absolute revision (MAR), which provides a measure of the revision adjusted for the offsetting effect due to opposite revisions, expressed by Formula (3) as

$$MAR = \frac{1}{n} \sum_{t=1}^n |L_t - P_t| \quad (3)$$

where L_t represents the value of the cyclical change at time t relative to the series calculated by the indirect method and P_t that of the series obtained by the precedent method; and the relative mean absolute revision (RMAR), or RMA normalization, illustrated by Formula (4):

$$RMAR = \frac{\sum_{t=1}^n |L_t - P_t|}{\sum_{t=1}^n |P_t|} \quad (4)$$

This was to take into account the fact that, in periods characterized by very large fluctuations, revisions may be larger and to compare series for different economic sectors or time periods with each other.

For an assessment of the direction of the revisions (underestimation/overestimation), the mean revision (MR), of which only the sign and not the numerical value was considered (Formula (5)), was considered, accompanied by the corresponding test on the hypotheses that it is or is not significantly different from zero:

$$MR = \frac{1}{n} \sum_{t=1}^n (L_t - P_t) \tag{5}$$

Therefore, a positive (negative) mean revision with significant test denotes a systematic underestimation (overestimation) by the source series (P) compared with the L series defined by the indirect method.

5. Main Results

A comparison has been made between the use of the indirect and direct approach when applied to the seasonally adjusted VELA worked hours per capita indices. In particular, Figures 3–5 show the hours worked per capita and their quarterly-on-quarterly changes in the two seasonal adjustment approaches, for the total economy (economic activity sections from B to S of the classification NACE Rev. 2), industry (sections from B to F), and services (G–S), separately.

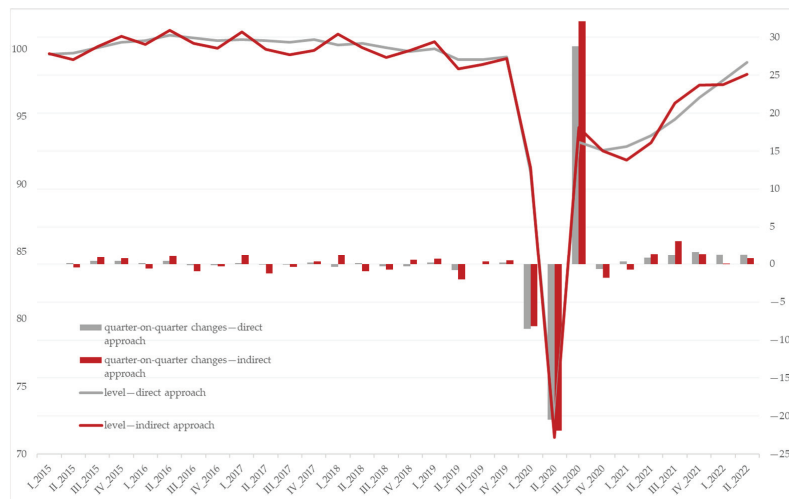


Figure 3. VELA worked hours per capita indices, indirect vs. direct approach. Total economy (B–S)—I quarter 2015–II quarter 2022. Level (left scale) and quarter-on-quarter changes (right scale). Seasonally adjusted indices.



Figure 4. VELA worked hours per capita indices, indirect vs. direct approach. Industry (B–F)—I quarter 2015—II quarter 2022. Level (left scale) and quarter-on-quarter changes (right scale). Seasonally adjusted indices.

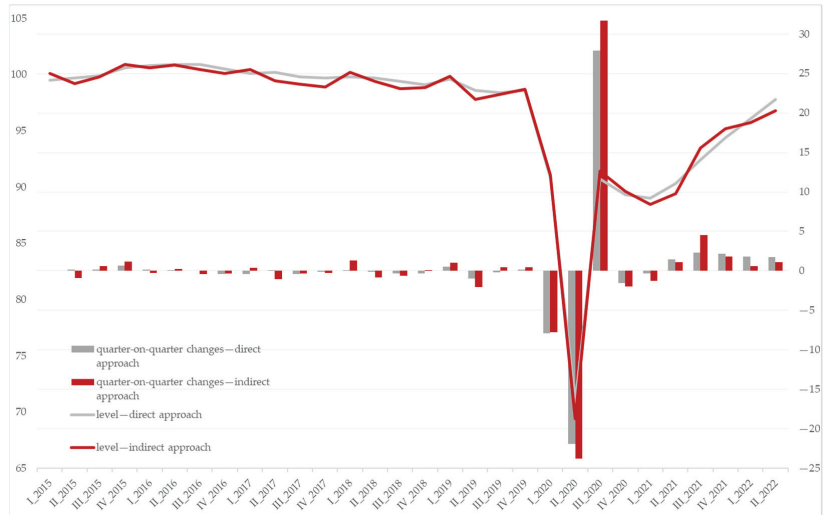


Figure 5. VELA worked hours per capita indices, indirect vs. direct approach. Services (G–S)—I quarter 2015—II quarter 2022. Level (left scale) and quarter-on-quarter changes (right scale). Seasonally adjusted indices.

As the figures show, the series seasonally adjusted by the indirect method show cyclical changes that are slightly larger—mainly in the aggregated industry sector—than those obtained by the direct method, and of the same sign in most quarters (around 80%). This was somewhat expected as the indirect method allows the seasonally adjusted aggregate series to better represent the different characteristics and behavior of each component series than the direct approach.

Moreover, the MAR and RMAR revision measures were calculated based on the quarter-on-quarter changes and taking into account all the economic activity sections (Table 1).

Table 1. MAR and RMAR revision measures by NACE Rev. 2 economic activity section—average values, I quarter 2015–II quarter 2022.

Economic Activity Section	MAR	RMAR
Mining and Quarrying—B	1.2	0.8
Manufacturing—C	1.2	0.5
Electricity, gas, steam, and air conditioning supply—D	1.0	0.8
Water supply: sewerage, waste management, and remediation activities—E	0.8	0.6
Construction—F	0.5	0.2
Wholesale and retail trade; repairs—G	1.1	0.4
Transportation and storage—H	0.3	0.1
Accommodation and food service activities—I	3.2	0.3
Information and communication—J	1.1	0.6
Financial and insurance activities—K	1.6	1.4
Real estate activities—L	0.8	0.2
Professional, scientific, and technical activities—M	0.8	0.3
Administrative and support service activities—N	1.1	0.5
Education—P	1.8	0.3
Human health and social work Activities—Q	0.2	0.1
Arts, entertainment, and recreation—R	6.3	0.5
Other service activities—S	1.1	0.3
Industry B–F	0.9	0.4
Services G–S	0.8	0.3
Total Economy B–S	0.8	0.4

The absolute differences account for less than two percentage points, with the exception of sectors I (Accommodation and food services) and R (Arts, entertainment, and recreation). These two sectors were the most affected by the COVID emergency in the period between Q3 2020 and Q3 2021. Therefore, they were characterized by higher quarterly changes during this period. In this case, the RMAR measure is a more reliable revision measure for understanding the actual impact of the new seasonal adjustment method. For these two sections, the RMAR measure shows no significant differences compared to the other sections.

In addition to this, the signs of quarter-on-quarter changes in the new VELA series were compared with those of other Istat macroeconomic indicators, related to hours worked: namely, the total employee jobs from the OROS (employment, wages and salaries, and social charges) survey, the Industrial Production Index, the Construction Production Index, and the Turnover in Services Index. The signs are concordant between 60% and 90% of the total quarters, depending on the economic activity section. Moreover, the average difference between quarter-on-quarter changes in the new VELA series and those of the above-mentioned indicators does not exceed 3%, on the total quarters considered, varying according to the economic activity section.

After these preliminary analyses of the performance of the new VELA series compared to those currently in use, the impact of the new method on reducing the number of outliers and the size of the residual outlier was evaluated. No out-of-range data were observed with the new method, whereas the current procedure based on the direct method produced around 20 out-of-range data points in the period under observation, mainly concentrated in the industry sector. As mentioned above, in this sector, the new indirect approach seems to have modified the original series more than in the service sector.

6. Conclusions

The application of the indirect approach to the seasonal adjustment to the VELA hours worked per capita series, as produced by the LFS, has led to important results both in terms of quality and as an example of the synergies developed by integrating the methodologies followed by different surveys.

The impact of the COVID emergency, which affected the economic activity sectors in various significant ways, highlighted the presence of a non-negligible number of in-

consistencies (also called out-of-range data) between the quarter-on-quarter changes of aggregated activity sectors and those of each component section.

Using the method adopted by the LFS, the number of out-of-range data points was greatly reduced, with the advantage that the VELA series did not demonstrate significant revisions when switching from the direct to the indirect method.

This work represents a positive example of the replication of methodologies among surveys with different characteristics. The use of a common methodology is also an important step in the direction of greater comparability between data from different sources.

The encouraging results shown so far by the application of the LFS seasonally adjusted method to the VELA indicator on hours worked per capita encourage us to continue in this direction.

Author Contributions: C.G. wrote Section 4; A.L. wrote Section 5; M.L. wrote Section 3.2; E.M. wrote Section 2; A.S. wrote main text of Sections 3 and 3.1. All authors wrote the Introduction and Conclusion sections. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing is not applicable to this article. The data presented in this study are still experimental, as they are the result of a preliminary exercise under development.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eurostat. ESS guidelines on seasonal adjustment. In *Eurostat Manuals and Guidelines*; Eurostat: Luxembourg, 2015.
2. Istat. La rilevazione trimestrale OROS su occupazione e costo del lavoro: Indicatori e metodologie. In *Metodi e Norme*; Istat: Rome, Italy, 2019.
3. Gomez, V.; Maravall, A. Programs TRAMO and SEATS: Instructions for the user. *Mimeo Banco de España* **1997**, 1–129.
4. Di Fonzo, T. The OECD project on revisions analysis: First elements for discussion. In *Proceedings of the OECD STESEG Meeting*, Paris, France, 27–28 June 2005.
5. McKenzie, R.; Gamba, M. Interpreting the results of Revision Analyses: Recommended Summary Statistics. Contribution to the OECD/Eurostat Task Force on “Performing Revisions Analysis for Sub-Annual Economic Statistics”. 2008. Available online: <https://www.oecd.org/sdd/40315546.pdf> (accessed on 27 June 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Online Pentane Concentration Prediction System Based on Machine Learning Techniques [†]

Diana Manjarrés ^{1,*}, Erik Maqueda ¹ and Itziar Landa-Torres ²

¹ TECNALIA, Basque Research & Technology Alliance (BRTA), Technological Park of Bizkaia, 48160 Derio, Spain; erik.maqueda@tecnalia.com

² Petronor Innovación S.L, 48550 Muskiz, Spain; itziar.landa@repsol.com

* Correspondence: diana.manjarres@tecnalia.com

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Industry 4.0 has emerged together with relevant technological tools that have enabled the rise of this new industrial paradigm. One of the main employed tools is Machine Learning techniques, which allow us to extract knowledge from raw data and, therefore, devise intelligent strategies or systems to improve actual industrial processes. In this regard, this paper focuses on the development of a prediction system based on Random Forest (RF) to estimate Pentane concentration in advance. The proposed system is validated offline with more than a year of data and is also tested online in an Energy plant of the Basque Country. Validation results show acceptable outcomes for supporting the operator's decision-making with a tool that infers Pentane concentration in Butane 400 min in advance and, therefore, the quality of the obtained product.

Keywords: random forest; pentane concentration prediction; refineries; machine learning; artificial intelligence

1. Introduction

The fourth industrial revolution, coined as Industry 4.0, is characterized by the integration of advanced digital technologies such as Internet of Things (IoT), Artificial Intelligence (AI), Robotics, Cloud Computing, Big Data and Cybersecurity into the industrial process. These technologies enable factories and supply chains to become more efficient, productive and adaptable to changing market demands. In this context, many industries have monitored their processes and units with the aim of optimizing their operational conditions and, thus, improving the quality of the final products [1,2]. Regarding the Energy Industry, in [3], a method for estimation the oxygen content in a coke furnace is presented. Similarly, in [4], a soft-sensor for the prediction of MAE and SWA acid gases is shown.

A common and relevant fact that encompasses these kind of problems is the need to build intelligent strategies that extract valuable insights from the available data. In this context, Feature Selection (FS) and Feature Weighting (FW) techniques along with Machine Learning (ML) models that enable the construction of automated decision support systems based on data are a hot research topic nowadays. In this sense, several works apply FS and FW strategies to problems related to the Energy sector, such as [5,6]. In [6], a Butane concentration estimator at the bottom of the debutanizer column with an FW strategy is presented. Similarly, authors in [5] propose an autoML approach that considers feature preprocessing and selects the best algorithm configuration for developing a soft-sensor for Pentane concentration prediction at the end of a debutanizer column.

This paper focuses on this last open challenge, i.e., to predict approximately 400 min in advance the percentage of Pentane concentration in Butane at the end of a debutanizer column. In contrast to [5], a regression model based on a Random Forest (RF) technique is proposed. Thus, it is possible to assess the trend of Pentane concentration prediction

Citation: Manjarrés, D.; Maqueda, E.; Landa-Torres, I. Online Pentane Concentration Prediction System Based on Machine Learning Techniques. *Eng. Proc.* **2023**, *39*, 77. <https://doi.org/10.3390/engproc2023039077>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 12 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

through this implementation. Furthermore, online results obtained by applying this proposal in an Energy plant are presented.

The remainder of the paper is structured as follows: Section 2 depicts the real industrial use case. Section 3 presents the Random Forest technique and the offline and online validation results. Finally, Section 4 shows the conclusions of the work.

2. Industrial Use Case Description

The Industrial use case focuses on a specific line for the production of Butane—this being a product of great added value. In order to be marketed, it must meet a series of requirements and specifications. One of them is the proportion of Pentanes in the Butane itself, where the maximum admissible threshold is 1.5%.

Figure 1 shows the Butane production scheme.

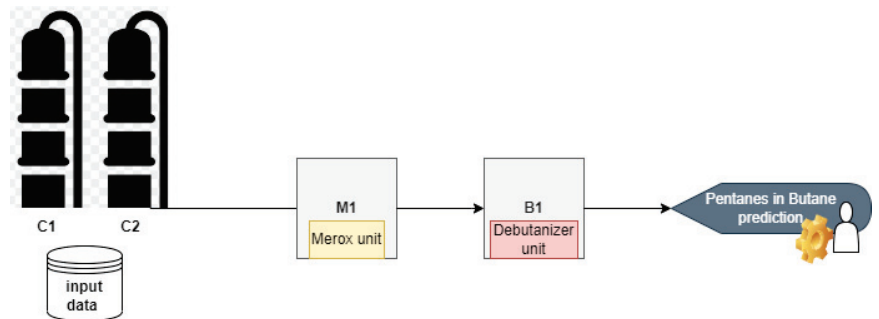


Figure 1. Schema of the industrial use case.

In Figure 1, the main line for the blending of Butanes can be observed. It consists of the head of the naphtha stabilizers (C1, C2), the Merox unit (M1) and ends in the first Butanes unit (B1). The data collected come from columns C1 and C2 wherein information about flow, temperature and pressure is gathered. The aim is to predict the percentage of Pentanes in Butane that will be at the end of the debutanizer column but approximately 400 min in advance. The process variables information are obtained every 10 min from October 2017 to February 2019.

3. Percentage of Pentanes in Butane Prediction System

Firstly, a feature importance analysis of the process variables is conducted and the most relevant in terms of Pentane production are selected as input to the prediction system. The feature importance methods used to perform the study are Pearson correlation, Random Forest, ANOVA and Mutual Information. For each of these methods, the 30 most influential variables are selected and those that appear as relevant in three of the four methods with a correlation >0.9 between them are finally chosen.

After this first analysis, a Random Forest regressor model [7] is implemented with all the available variables and with the most relevant ones. Two different methods for training the model are tested: (1) to train and validate the model using the cumulative learning method and (2) using the sliding window method. By means of employing the cumulative learning method, the model is trained with the first 14 months and tested with the last three months. The absolute mean error obtained is 0.58(%). Moreover, it is observed that the prediction error increases as the test data move away in time from the training data—that is, for the first hours of the test, the error is low, but as the hours pass the error increases. On the other hand, the model is trained using the sliding window method with a window of one month, i.e., training with the data of one month and testing with the next value, and so on, sliding the window until all the months of the sample are covered. In this way, a mean absolute error in the prediction of 0.21(%) is obtained, which is significantly lower than that obtained by the cumulative training method. In addition, it is observed that the importance

of the variables also varies over time. As a conclusion, the model with a sliding window of one month is chosen for the construction of the final prediction system. The fact of obtaining a lower error through the sliding window suggests that there is seasonality—that is, the relationship between the process variables and the Pentane concentration varies over time.

As commented in the previous sections, the main objective is to develop a model that predicts in advance a peak in the Pentane concentration—that is, when it exceeds the 1.5% threshold. Therefore, the developed RF regressor model is used as a core part solution to develop a decision support system that generates an alarm when the prediction exceeds the 1.5% threshold.

In Figure 2, an interval of the real signal in green associated with the proportion of Pentanes in the production of Butane is presented. The red vertical line indicates the first point where the proportion has exceeded the limit set at 1.5% (black horizontal line). Finally, the cyan line corresponds to all points where the real signal exceeds said limit.

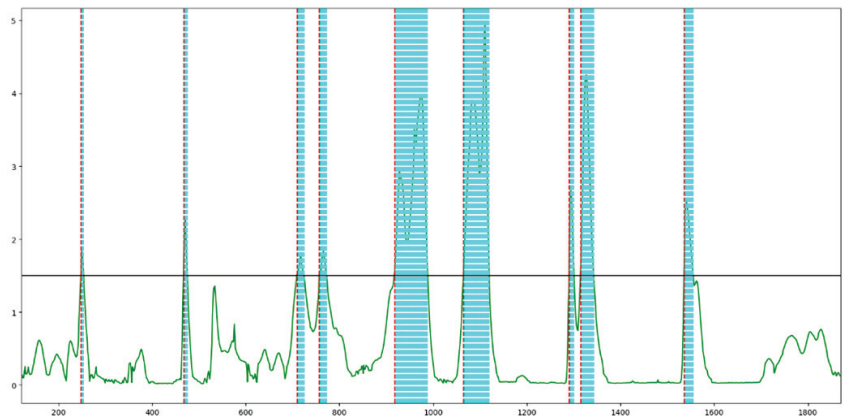


Figure 2. Real signal associated with the proportion of Pentanes in the production of Butane. Black horizontal line indicates the 1.5% limit. Cyan line: points that exceed the limit. Red line: first point that exceeds the limit per section.

In order to evaluate the results provided by the percentage of Pentanes in Butane prediction system, a set of well-known metrics, slightly modified for the problem at hand, are used: True Positives (TP), False Positives (FP) and False Negatives (FN). For the entire period studied (October 2017–February 2019), a total of 185 points were identified that exceeded the limit of 1.5%. It should be noticed that the real process has an average offset of around 400 min, which, as contrasted with the domain experts, may vary over time. This fact is considered for calculating the evaluation metrics, named True Positives and False Positives, as follows:

- True positives (TP) and timeTP1: Analyzing the real signal, when it exceeds the limit of 1.5%, the time that the prediction takes to exceed that value is calculated (timeTP1time). If after 400 min the prediction does not exceed it, it means that the rise in Pentanes has not been detected sufficiently in advance and it is counted as FN.
- True positives (TP), timeTP2 and timeFP: Analyzing the real signal, when it exceeds the limit of 1.5%, the time that the prediction is ahead in predicting the rise in Pentanes (timeTP2) is calculated. If it exceeds the maximum margin, it is computed as FP and timeFP is calculated, and if it is not exceeded, it is computed as TP and timeTP2 is allocated. When establishing this maximum time margin for timeTP2, it was agreed with the domain experts to consider 460 min (400 + 60).

Table 1 shows the results obtained by the application of the RF algorithm and the RF followed by a Savitzky–Golay filter [8] for smoothing the prediction outcome and, thus, reducing the FPs. Note that by increasing the window size, the FPs are reduced at the cost of also reducing the TPs.

Table 1. Obtained results by employing RF and RF plus Savitzky–Golay filter (SG) with windows sizes $w = \{3, 7, 21\}$.

	TPs	FPs	FP/TP	timeTP1 (min) min/mean/max—std	timeTP2 (min) min/mean/max—std
RF	93	145	1.55	20/274/390—99	400/445/450—12
RF + SG $w = 3$	83	96	1.15	20/275/380—93	400/444/450—13
RF + SG $w = 7$	70	68	0.97	20/253/370—94	380/425/430—10
RF + SG $w = 21$	53	44	0.83	20/183/280—75	310/354/360—14

With the aim of investigating new alternatives that could improve the RF prediction system, a detailed analysis of the data and results was performed and the following conclusions were obtained: On the one hand, the limitation of imposing a constant offset of 400 min for all variables is too strict. As verified during the validation, there is an average offset of 400 min. Although for most of the peaks the offset is between 350 to 450 min, it is not always 400 min. On the other hand, during the analysis, it is observed that the concentration of Pentane at 400 min seems to be influenced by the previous values of Pentane concentration. Therefore, it seems reasonable that if the value of Pentane concentration at the instant of the prediction is incorporated, the results could be improved. As a result of these conclusions, the following two new implementations are tested in order to see if they improve the results of FP/TP ratio:

- RF model implementation 1: introducing the previous values of Pentane concentration.
- RF model implementation 2: introducing different offsets in the process variables (offsets from -450 to -350 min) together with the previous values of Pentane concentration.

These two RF model implementations are validated for the month of January 2020. In order to compare the results with the previous ones, the FP/TP ratio is used.

Tables 2 and 3 present the results obtained by RF model implementations 1 and 2.

Table 2. Obtained results by employing RF model implementation 1 and RF model implementation 1 followed by a Savitzky–Golay filter (SG) with windows sizes $w = \{3, 5, 7, 9, 11, 21\}$.

	TPs	FPs	FP/TP	timeTP1 (min) min/mean/max—std	timeTP2 (min) min/mean/max—std
RF	7	12	1.71	250/318/380—43	450/450/450—0
RF + SG $w = 3$	7	10	1.42	140/293/370—73	440/440/440—0
RF + SG $w = 5$	7	9	1.28	110/280/360—80	430/430/430—0
RF + SG $w = 7$	7	7	1	100/270/350—80	410/410/410—0
RF + SG $w = 9$	6	4	0.66	80/244/310—83	400/400/400—0
RF + SG $w = 11$	5	3	0.6	70/230/290—80	-
RF + SG $w = 21$	1	2	2	230/230/230—0	-

Table 3. Obtained results by employing RF model implementation 2 and RF model implementation 2 followed by a Savitzky–Golay filter (SG) with windows sizes [3, 5, 7, 9, 11, 21].

	TPs	FPs	FP/TP	timeTP1 (min)		timeTP2 (min)	
				min/mean/max—std	min/mean/max—std		
RF	7	8	1.14	190/270/330—51	360/380/400—20		
RF + SG w = 3	6	5	0.83	240/285/320—35	360/380/400—20		
RF + SG w = 5	6	4	0.66	230/265/310—30	340/365/390—25		
RF + SG w = 7	6	4	0.66	210/255/300—36	330/355/380—25		
RF + SG w = 9	6	4	0.66	190/237/280—35	320/354/370—25		
RF + SG w = 11	4	3	0.75	210/240/270—30	310/330/350—20		
RF + SG w = 21	3	2	0.66	180/180/180—0	280/290/300—10		

After analyzing the results for both RF model implementations, the following conclusions are obtained:

- RF model implementation 1 reduces the number of FPs up to window 11.
- RF model implementation 2 reduces the number of FPs up to window 9.
- The best approximation is that of RF model implementation 2 with SG of window 5 as it provides the best FP/TP ratio and timeTP.
- The best solutions for both approaches allow capturing 5–6 TPs out of a total of 8, generating 3–4 FPs.
- The FP/TP ratio is still quite high, around 0.66. That is to say, for approximately every 3 TPs we capture, 2 FPs are generated.

Finally, RF model implementation 2 is set online with SG of window 5 from 1 December 2020 to 26 January 2021. Figure 3 and Table 4 depict the obtained results.

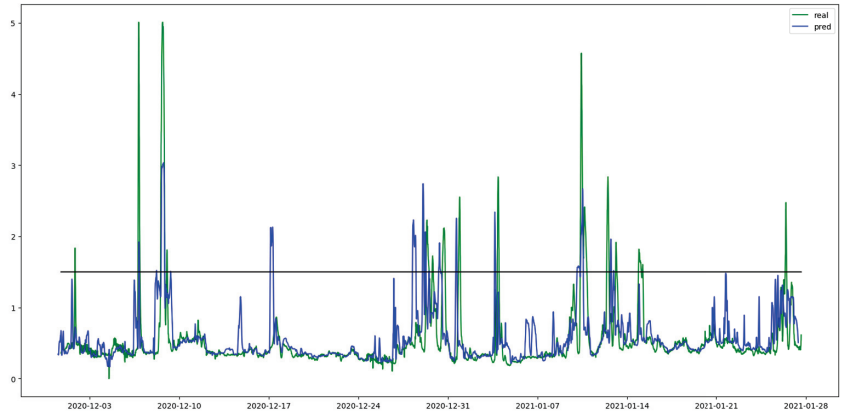


Figure 3. Online validation from 1 December 2020 to 26 January 2021. The real Pentane concentration signal is in green color and the predicted one in blue color.

Table 4. Obtained online validation results by employing RF model implementation 2 and RF model implementation 2 plus Savitzky–Golay filter (SG) with window size w = 5.

	TPs	FPs	FP/TP	timeTP1 (min) min/mean/max—std
RF model implementation 2 online	7	4	0.57	220/335/390—57

Figure 4 shows four examples of detection of the peak of Pentane concentration. As observed, the peak is detected in advance; so, the operators can take proper actions to minimize the consequences of that Pentane concentration peak.

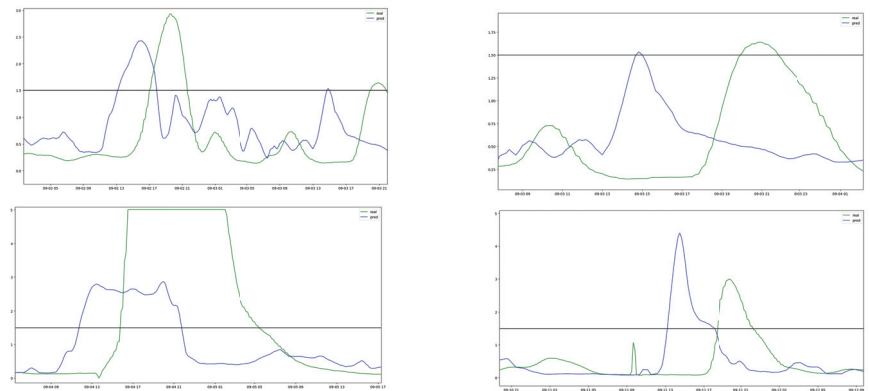


Figure 4. Examples of Pentane concentration peak detection. Green signal is the real one and blue signal is the predicted one.

4. Conclusions

This paper proposes a Pentane concentration prediction system based on ML techniques capable of detecting the quality of the Butane at the end of the debutanizer column 400 min in advance. Specifically, a Random Forest (RF) regressor followed by a Savitzky–Golay filter is proposed. The prediction system is validated offline with data from October 2017 to February 2019 employing a sliding window training strategy; it has also been tested online, providing acceptable results. Obtained results show that the proposed system is able to predict Pentane concentration peaks that occur in recent similar behaviors. However, when new behaviors suddenly appear, the system is not able to learn those behaviors fast enough and predict the peaks in advance.

In order to face this situation, future steps will be devoted to collaborating with the process operators and analyzing the possibility of eliminating some false positives with some extra process information, such as the crude composition.

Author Contributions: Conceptualization, I.L.-T.; methodology, D.M., I.L.-T. and E.M.; software, D.M. and E.M.; validation, D.M., I.L.-T. and E.M.; formal analysis, D.M. and E.M.; investigation, D.M. and E.M.; resources, I.L.-T.; data curation, D.M.; writing—original draft preparation, D.M.; writing—review and editing, D.M., I.L.-T. and E.M.; visualization, D.M. and E.M.; supervision, I.L.-T.; project administration, I.L.-T.; funding acquisition, I.L.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Petronor Innovación S.L.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Köksal, G.; Batmaz, İ.; Caner, M.; Testik, F. A review of data mining applications for quality improvement in manufacturing industry. *Expert Syst. Appl.* **2011**, *38*, 13448–13467. [CrossRef]
2. Saihi, A.; Awad, M.; Ben-Daya, M. Quality 4.0: Leveraging Industry 4.0 technologies to improve quality management practices—A systematic review. *Int. J. Qual. Reliab. Manag.* **2023**, *40*, 628–650. [CrossRef]
3. Zhang, R.; Jin, Q. Design and Implementation of hybrid modeling and PFC for oxygen content regulation in a coke furnace. *IEEE Trans. Ind. Inform.* **2018**, *14*, 2335–2342. [CrossRef]

4. Wang, K.; Shang, C.; Yang, F.; Jiang, Y.; Huang, D. Automatic hyper-parameter tuning for soft sensor modeling based on dynamic deep neural network. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017; pp. 989–994.
5. Niño-Adan, I.; Landa-Torres, I.; Manjarres, D.; Portillo, E.; Orbe, L. Soft-Sensor for Class Prediction of the Percentage of Pentanes in Butane at a Debutanizer Column. *Sensors* **2021**, *21*, 3991. [CrossRef] [PubMed]
6. Yuan, X.; Huang, B.; Wang, Y.; Yang, C.; Gui, W. Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE. *IEEE Trans. Ind. Inform.* **2018**, *14*, 3235–3243. [CrossRef]
7. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
8. Savitzky, A.; Golay, M.J.E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36*, 1627–1639. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Non-Invasive Arterial Blood Pressure Estimation from Electrocardiogram and Photoplethysmography Signals Using a Conv1D-BiLSTM Neural Network[†]

Federico Delrio^{1,*}, Vincenzo Randazzo^{1,*}, Giansalvo Cirrincione^{2,3} and Eros Pasero¹

¹ Department of Electronics and Telecommunications, Politecnico Di Torino, 10129 Turin, Italy; eros.pasero@polito.it

² Lab. LTI, Université de Picardie Jules Verne, 80039 Amiens, France; exin@u-picardie.fr

³ SEP, University of South Pacific, Suva 1168, Fiji

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: This paper presents a neural network model to estimate arterial blood pressure (ABP) waveforms using electrocardiogram (ECG) and photoplethysmography (PPG) signals and its first two order mathematical derivatives (PPG', PPG''). In order to achieve this objective, a lightweight and optimized neural network architecture has been proposed, made of Conv1D and BiLSTM layers. To train the network, the UCI Database "Cuff-Less Blood Pressure Estimation Data Set" has been used, which contains ECG and PPG signals together with the corresponding ABP waveform data; then the first two PPG derivatives have been computed. Four different configurations and parameter sets have been tested to choose the best structure and set of parameters. Additionally, various batch sizes, numbers of BiLSTM layers, and the presence of a maximum pooling layer have been tested. The best performing model achieves a mean absolute error of around 2.97, which is comparable to the state-of-the-art methods. Results prove deep learning techniques can be effectively used for non-invasive cuffless arterial blood pressure estimation. The lightweight and optimized model can be effectively used for continuous monitoring of blood pressure, which has significant clinical implications. Further research can focus on integrating the proposed model with wearable devices for real-time blood pressure monitoring in daily life.

Citation: Delrio, F.; Randazzo, V.;

Cirrincione, G.; Pasero E.

Non-Invasive Arterial Blood Pressure Estimation from Electrocardiogram and Photoplethysmography Signals Using a Conv1D-BiLSTM Neural Network. *Eng. Proc.* **2023**, *39*, 78.

<https://doi.org/10.3390/engproc2023039078>

<https://doi.org/10.3390/engproc2023039078>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 12 July 2023



Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: neural networks; arterial blood pressure; ECG; PPG

1. Introduction

Arterial blood pressure (ABP) is a crucial indicator of an individual's health. It measures the amount of pressure that blood exerts against the walls of arteries during circulation. Accurate measurement of ABP is crucial in diagnosing and timely managing cardiovascular diseases, such as hypertension [1]. However, conventional methods for measuring ABP are either invasive, requiring the insertion of a catheter into an artery, or need a cuff to be inflated around the arm, which can lead to patient discomfort [2]. Consequently, cuffless non-invasive methods based on ABP estimation from electrocardiogram (ECG) and/or photoplethysmogram (PPG) signals have gained popularity due to their ease of use and safety.

Recent studies indicate that deep learning techniques can accurately predict arterial blood pressure from ECG and PPG signals [3–5]. However, these methods tend to be computationally intensive and time-consuming due to their complexity and requirement of large datasets. The proposed model employs a simpler architecture that combines a Conv1D neural network and a bidirectional long short-term memory (BiLSTM) network to capture the temporal and spectral features of the ECG and PPG signals. Additionally, the first two derivatives of PPG are incorporated to capture the dynamic changes in ABP over time.

The proposed approach achieved promising results in predicting arterial blood pressure from ECG and PPG signals, with an overall mean absolute error (MAE) of only 2.97 mmHg on the test set. It is computationally efficient and requires less memory than state-of-the-art methods, making it a practical and effective solution for non-invasive arterial blood pressure regression and in principle could be more easy-to-transfer on wearable/portable devices, such as [6].

1.1. Related Work and Previous Studies

There are two main routes to take to predict blood pressure using artificial neural networks [7]: regression problem, which aims to predict the entire ABP signal waveform; and the direct systolic blood pressure (SBP), diastolic blood pressure (DBP) prediction as the maximum and minimum of ABP signals.

In [3], several deep learning techniques are compared to infer ABP, starting from photoplethysmogram and electrocardiogram signals. The ABP is first predicted using only PPG and, then, by using both PPG and ECG. Both convolutional neural networks (ResNet and WaveNet) and recurrent neural networks (LSTM) are compared and analyzed for the regression task. The results show that the use of the ECG have resulted in improved performance for every proposed configuration.

In [8], a U-net deep learning architecture that uses fingertip PPG signal as input to estimate ABP waveform non-invasively is proposed. From this waveform, they have also measured SBP, DBP, and the mean arterial pressure.

In [9], a deep learning model is presented, named ABP-Net, to transform photoplethysmogram signals into ABP waveforms that contain vital physiological information related to cardiovascular systems.

In [5], the applicability of autoencoders in predicting BP from PPG and ECG signals was explored.

These works demonstrate the potential of deep learning techniques in predicting blood pressure using non-invasive signals. They also highlight the importance of using ECG signals in combination with PPG ones to improve prediction performance. However, further research is needed to establish the accuracy and generalization capability of these models in predicting blood pressure in different populations and settings. Furthermore, many studies rely on massive neural networks, some with as many as 60 million parameters, like [5].

1.2. State of the art limitations

While the studies on non-invasive estimation of arterial blood pressure using ECG and PPG signals have shown promising results, there are some limitations to consider.

Firstly, the studies typically evaluate the performance of the proposed methods on small- to medium-sized datasets, which may not be representative of the wider population. Therefore, further validation on larger datasets is required to assess the generalizability of these methods.

Secondly, the studies often focus on predicting the systolic and diastolic blood pressure values separately, rather than predicting the full arterial blood pressure waveform. This limits the ability to capture the complex variations in blood pressure over time.

Thirdly, some studies may not consider the influence of various factors such as age, gender, and underlying medical conditions that may affect blood pressure, which can impact the accuracy of the predictions.

Finally, the use of non-invasive methods to estimate blood pressure may not be suitable for all individuals, such as those with certain medical conditions or those who are critically ill. In these cases, invasive methods may still be necessary to obtain accurate blood pressure measurements.

1.3. Potential Advantages of Proposed Model

The architecture presented in this paper, which is the logical continuation of the work by Paviglianiti [3,10] and Mahmud [5], aims to be as compact as possible without sacrificing

accuracy. Also, it provides better predictions as a result of the addition of the two PPG derivatives. Since the model is lightweight when compared to many others, it can be embedded into wearable or portable devices, or, more generally it can be applied to edge computing.

2. Dataset and Methods

The pulsatile nature of the cardiac output results in the pulse pressure waveform. The interaction of the heart's stroke volume, the arterial system's compliance (ability to expand), which is primarily due to the aorta and large elastic arteries, and the arterial tree's resistance to flow determines the magnitude of the pulse pressure. Systolic blood pressure (SBP) is defined as the peak of the ABP pulsewave in an ABP signal, see *orange stars* in Figure 1a. The minimum of ABP pulses is known as the diastolic blood pressure (DBP), as shown in Figure 1a, *green stars*. In our case, we have an entire waveform lasting 8 s rather than just a single pulse. Since the waveform is varying in time and peaks and minimums are not constant, for each sample, the mean of all the peaks and minimums was used to calculate the SBP and DBP, respectively (see Figure 1b).

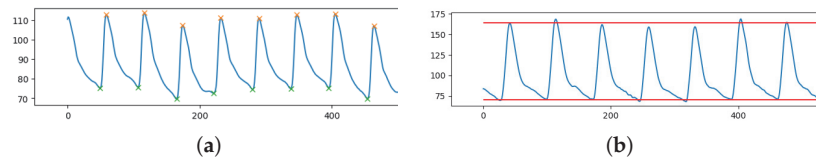


Figure 1. Fraction of ABP signal (*y*-axis) over time (*x*-axis) with: (a) highlighted peaks (SBP, *orange stars*) and minimum points (DBP, *green stars*); (b) extracted SBP (mean of peaks) and DBP (mean of minimums).

Using Python's ready-to-use function "`scipy.signal.find_peaks`", peak detection was carried out. In order for the peaks to be detected, it is necessary to define a "prominence" (a sort of threshold for the height of the peaks). Prominence must be assessed on a case-by-case basis because the ABP range is flexible. This parameter was empirically chosen to be the difference between each ABP signal median and minimum value.

2.1. Database

The UCI dataset, also known as the Cuff-Less Blood Pressure Estimation Dataset, was used in this study due to its simplicity and readiness [11,12]. It was sourced from the MIMIC-II Waveform database, which tracks physiological measurements such as ABP and PPG [13]. The UCI dataset consists of 12,000 instances of simultaneous PPG, ABP, and ECG data from 942 patients, and was pre-processed by Kachuee et al. to smooth signals, eliminate unacceptable values, and autocorrelate PPG signals [11]. Pre-processing of the entire dataset was necessary before model training.

2.2. Preprocessing

Inspired by [3], the PPG recordings were filtered using a band-pass 4th order Butterworth filter with a bandwidth of 0.5 Hz to 8 Hz to exclude frequencies responsible for baseline wandering and high frequency noise. Moreover, in order to prevent motion artifacts and powerline artifacts, the ECG signal was filtered with an 8th order passband Chebyshev type 1 filter with a cut-off frequency of 2 Hz and 59 Hz. Then, the inputs were standardized instance-wise using a min-max normalization between 0 and 1.

It is crucial to emphasize that ABP was not altered in any manner in order to preserve pressure information. The SBP, DBP pressure readings and forecasts would have been inaccurate and information would have been lost if filters or pre-processing were applied.

2.3. Data Selection and Training Set Creation

Data are sampled at 125 Hz. Since the maximum duration of an instance is 10 min, each instance of ECG, PPG and ABP will have at most 75,000 data points. To have adequate information to forecast the overall trend and the impact of the ECG and PPG on ABP, all

the instances were divided in segments of length of 8 s (1000 points), comparable to [5] (1024 points).

The signal segments extracted from the UCI dataset contain many highly distorted signals that prevented the deep learning model from properly mapping input signals to the corresponding ABP waveform and, thus, hinder correct SBP and DBP estimation. Experimentally, it was found that highly distorted signals typically lie on one of the following categories: SBP below 80 mmHg or over 190 mmHg; DBP below 50 mmHg and above 120 mmHg; blood pressure ranges ($|SBP - DBP|$) below 20 mmHg or above 120 mmHg. As a result, these ABP samples, together with their corresponding ECG and PPG signals, were removed from the dataset.

Afterwards, since the peaks of ECG, PPG, and ABP signals are frequently non-uniform, e.g., due to patient movements during acquisition, an additional data selection step has been performed, based on the standard deviation of peak heights and peak distances within each extracted signal. Here, some maximum values have been fixed to filter out noisy signals; Table 1 summarizes these thresholds for PPG, ECG, and ABP, respectively.

Table 1. Pruning thresholds for the standard deviation σ of PPG, ECG, and ABP.

	PPG	ECG	ABP
σ of peak distances	>6	>6	>6
σ of peak heights	>0.1	>0.1	>6

2.4. Input and Output of the Network

The PPG, ECG, $VPPG = \frac{d(PPG)}{dt}$, and $APPG = \frac{d(VPPG)}{dt}$, each with a length of 1000 (8 s) time instants, has been fed, in this order, to a different network channel to make the input tensor with shape (1000, 4). On the other side, an ABP waveform lasting 8 s will be the network target, which the model must forecast. Figure 2 shows an example of input tensor (Figure 2a–d) and the relative ABP output to be predicted (Figure 2e).

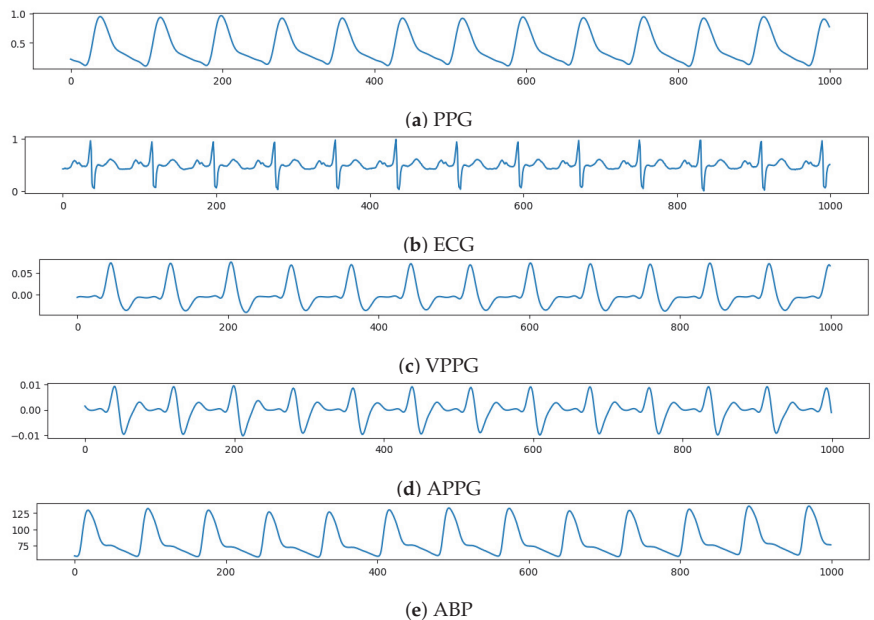


Figure 2. Example of input vector over time (x -axis) composed, from top to bottom (y -axis), by: PPG, ECG, VPPG, APPG; and then the relative ABP to be predicted.

2.5. Extracted Signals Analysis

After the previously described preprocessing, thresholding and cleaning phases, 192,661 input tensors (each of 8 s) were obtained, for a total of 428.13 h of training data in total. It can be helpful to examine the SBP and DBP distribution of the extracted examples shown in Figure 3. Due to the thresholds defined during the extraction and selection of signal steps, it follows that the distributions are truncated where the upper and lower boundaries have been set. It can also be observed DBP is generally less dispersed than SBP (see Figure 3 right).

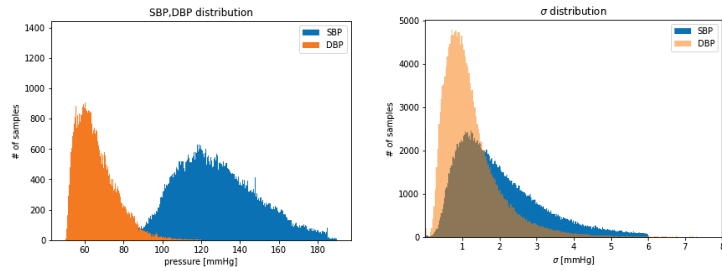


Figure 3. Distribution of: SBP and DBP extracted values (left); σ of extracted values of SBP, DBP (right).

3. Network Architecture

Inspired by the great results of [3] and [5] our first idea was to use a “mixture” of the two models. However, since the architecture of [5] was huge (about 120 million of parameters), it is evident that this approach could not be the optimal one for the ABP prediction, especially looking at the network of [3] (just around 2 million parameters). Therefore, it was opted to use a series of Conv1D (with 128 filters and kernel size equal to 3), and BiLSTM layers.

Conv1D [14,15] are commonly used in time-series analysis because they can effectively extract temporal features from the data. This is particularly useful when dealing with noisy or variable data, where traditional statistical methods may not be effective.

Bidirectional Long Short-Term Memory (BiLSTM) [3] networks are also commonly used in time-series analysis because they can effectively capture both past and future information in the time-series.

The idea behind the model is to use a sort of “encoder–decoder” structure, based on the Conv1D layers, with the BiLSTM as a backbone, instead of a simple Multi Layer Perceptron as [5]. Also, a skip connection between and after the backbone to prevent the vanishing gradient was used. The resulting structure can be seen in detail in Figure 4.

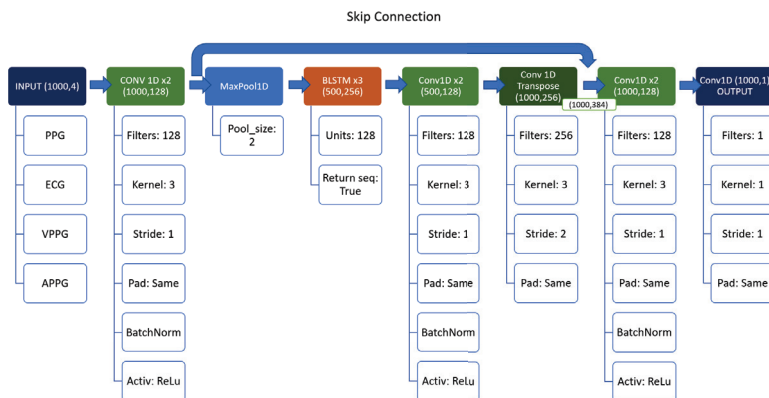


Figure 4. Structure of the first network.

4. Experiments and Results

All of the examples were randomly shuffled before the network training began (using the same seed for each experiment, for sake of comparability), and the training set was made up of 80% of the data, while validation and testing subsets received 10% each, respectively.

All the experiments were performed using Adam optimizer Keras default configuration [16,17] and 30 epochs. Due to the presence of the MaxPool layer, the stride of the Conv1DTranspose layer must be set to 2 to match the input/output dimensions, and keep the same number of parameters in the network.

To stick with the state of the art, two metrics—Mean Absolute Error (MAE) as the observed metric and Mean Square Error (MSE) as the loss function for the model—were used in all the experiments. Overall, while other loss functions may be used for time-series regression, MSE is an effective choice due to its simplicity, interpretability, and effectiveness in capturing the error between predicted and actual values over time.

The first experiment was made using the architecture of Figure 4, which employed 64 input batches, a Max Pool layer, and 3 BiLSTM layers. Results in terms of MAE and MSE, for both training and validation, are shown in Figure 5.

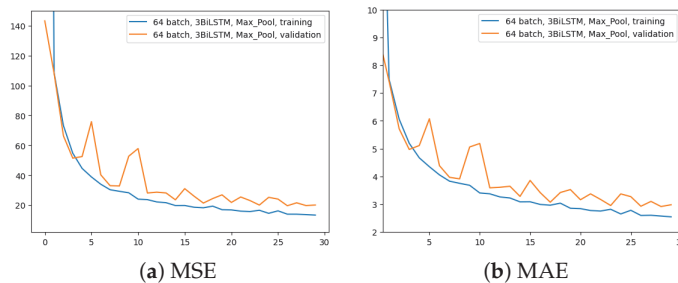


Figure 5. Error (*y-axis*) over epochs (*x-axis*) for the first experiment (batch size = 64, 3 BiLSTM and MaxPool layer): (a) MSE on training (*blue line*) and validation (*red line*) sets; (b) MAE on training (*blue line*) and validation (*red line*) sets.

The second experiment aimed to comprehend the importance of removing the Max-Pool layer. Here, 3 BiLSTM and 64 input batches were still used. The time for the computation increased from 2.5 h to 4 h. The outcomes are displayed in Figure 6.

Figure 7 shows a comparison among these two network configurations with regard to the MAE metric. It can be seen that removing the layer has almost no impact on the regression performance. With a slightly more pronounced difference at early epochs, the two configurations appear to converge in the same way at the latest epochs. Since removing the MaxPool layer does not seem to bring any noticeable improvements, in all future experiments this layer will be used due to its much faster network training time.

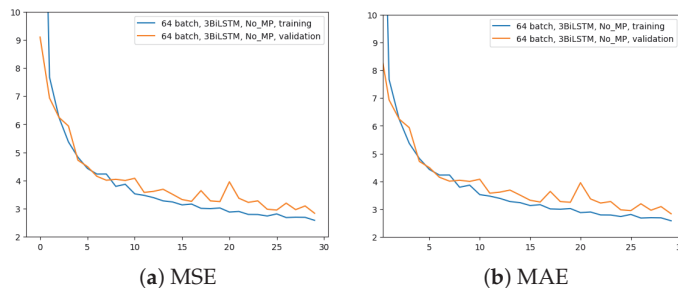


Figure 6. Error (*y-axis*) over epochs (*x-axis*) for the second experiment (batch size = 64, 3 BiLSTM and no MaxPool layer): (a) MSE on training (*blue line*) and validation (*red line*) sets; (b) MAE on training (*blue line*) and validation (*red line*) sets.

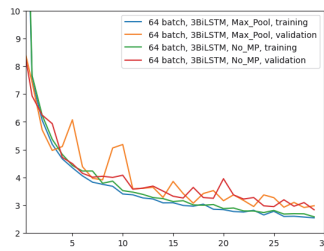


Figure 7. Comparison of MAE (*y-axis*) over epochs (*x-axis*) between Figures 5b and 6b experiments.

In the third experiment, just two BiLSTM were implemented to assess the impact of different number of BiLSTM layers; the rest remains as per the first experiment (i.e., 64 input batches and Max Pool layer). Figure 8 displays the outcome in comparison to the other two experiments for the two metrics. It is evident that using only two BiLSTM reduces performance across the board. The validation MAE never achieves the outcomes of the first two experiments.

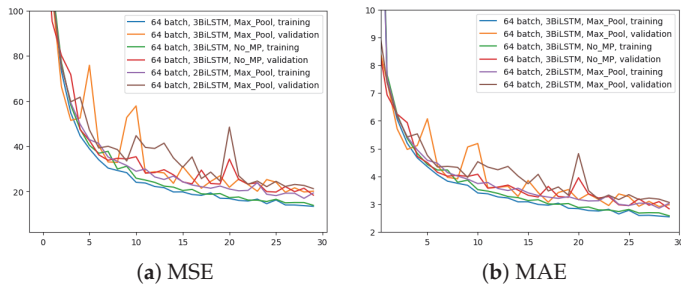


Figure 8. Error (*y-axis*) over epochs (*x-axis*) for the third experiment (batch size = 64, 2 BiLSTM and MaxPool layer) compared to the previous two networks: (a) MSE on training and validation sets; (b) MAE on training and validation sets.

The last experiment aimed to highlight the effects of increasing the batch size to 256 with MaxPool and 3 BiLSTM layers. Figure 9 yields the results. Even though the network is still learning, it does not appear to converge as quickly as it did in the previous experiments, as it is evident just by looking at Figure 9a. However, at higher epochs, the MAE training trend appears to converge to similar values. The validation MAE, however, never falls below the other model with a 64 batch size, as can be seen by looking at Figure 9b.

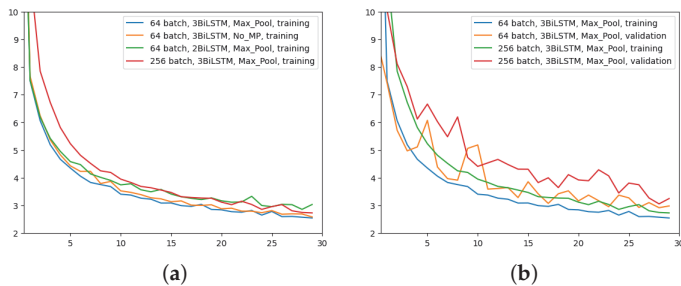


Figure 9. Comparison of MAE (*y-axis*) over epochs (*x-axis*): (a) All networks, only training; (b) first and last network configurations, training and validation.

In conclusion, the ablation study that performed pruning on some parts of the initial network proved that the architecture shown in Figure 4 is the best performing one. Indeed, it reaches lower MAE values than the third and fourth configurations and requires a

quite shorter training time than the second network, as previously stated. For sake of completeness, after choosing the ideal structure and parameters, it is possible to see some examples of the model predictions using random inputs from the test set, in Figure 10.

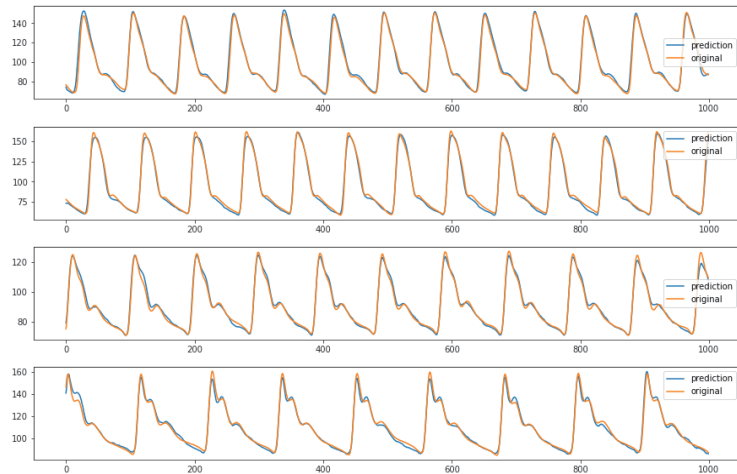


Figure 10. Examples of predicted ABP (*y*-axis) over time (*x*-axis) from the first network configuration (64 batch size, MaxPool and 3 BiLSTM layers). The predictions are relative to random inputs from the test set, to show the network capability of reproducing different kind of waveform.

5. Conclusions

This paper aimed to build a lightweight neural network model to predict 8 s of ABP signal, using the “Cuff-Less Blood Pressure Estimation Dataset”. At this purpose, a novel network, based on Conv1D as encoder/decoder block and BiLSTM as backbone, was proposed. The initial architecture was designed using 64 input batch size, MaxPool and 3 BiLSTM layer. Then, three additional network configurations were presented by means of an ablation strategy, and their performances were compared in terms of MAE and MSE metrics. The best performing model was the initial one, which achieved, on test data, a MAE of around 2.97 mmHg.

The main contribution of this paper is to have given a simple way of predicting ABP and lay the foundations for a transfer of research results on possible portable/wearable medical devices. One of the next steps will be to apply this method to real-world scenarios, where data are often more irregular and noisy. Furthermore, it may be interesting to use this method to derive SBP and DBP directly without predicting the entire ABP waveform. Also, the use of a larger database, such as the new MIMIC IV, could also improve performance and generalization.

Author Contributions: Conceptualization, V.R. and E.P.; methodology, F.D., G.C. and V.R.; software, F.D.; validation, F.D., V.R. and G.C.; formal analysis, F.D. and G.C.; investigation, F.D.; resources, E.P. and V.R.; data curation, F.D.; writing—original draft preparation, F.D. and V.R.; writing—review and editing, V.R.; visualization, F.D.; supervision, V.R., E.P. and G.C.; project administration, E.P. and V.R.; funding acquisition, E.P. and V.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been partly funded by the BrS-AI-ECG project of the Italian Ministry of Foreign Affairs and International Cooperation. Dr. Randazzo also acknowledges funding from the research contract no. 32-G-13427-2 (DM 1062/2021) funded within the Programma Operativo Nazionale (PON) Ricerca e Innovazione of the Italian Ministry of University and Research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in The UCI dataset, also known as the Cuff-Less Blood Pressure Estimation Dataset [11,12].

Acknowledgments: Delrio acknowledges Eurecom for support in his PhD research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Parati, G.; Stergiou, G.S.; Dolan, E.; Bilo, G. Blood pressure variability: Clinical relevance and application. *J. Clin. Hypertens.* **2018**, *20*, 1133–1137. [CrossRef] [PubMed]
2. Ilies, C.; Bauer, M.; Berg, P.; Rosenberg, J.; Hedderich, J.; Bein, B.; Hinz, J.; Hanss, R. Investigation of the agreement of a continuous non-invasive arterial pressure device in comparison with invasive radial artery measurement. *Br. J. Anaesth.* **2012**, *108*, 202–210. [CrossRef] [PubMed]
3. Paviglianiti, A.; Randazzo, V.; Villata, S.; Cirrincione, G.; Pasero, E. A Comparison of Deep Learning Techniques for Arterial Blood Pressure Prediction. *Cogn. Comput.* **2021**, *14*, 1689–1710. [CrossRef] [PubMed]
4. Ibtihaz, N.; Mahmud, S.; Chowdhury, M.E.H.; Khandakar, A.; Salman Khan, M.; Ayari, M.A.; Tahir, A.M.; Rahman, M.S. PPG2ABP: Translating Photoplethysmogram (PPG) Signals to Arterial Blood Pressure (ABP) Waveforms. *Bioengineering* **2022**, *9*, 692. [CrossRef] [PubMed]
5. Mahmud, S.; Ibtihaz, N.; Khandakar, A.; Tahir, A.M.; Rahman, T.; Islam, K.R.; Hossain, M.S.; Rahman, M.S.; Musharavati, F.; Ayari, M.A.; et al. A Shallow U-Net Architecture for Reliably Predicting Blood Pressure (BP) from Photoplethysmogram (PPG) and Electrocardiogram (ECG) Signals. *Sensors* **2022**, *22*, 919. [CrossRef] [PubMed]
6. Randazzo, V.; Ferretti, J.; Pasero, E. Anytime ECG Monitoring through the Use of a Low-Cost, User-Friendly, Wearable Device. *Sensors* **2021**, *21*, 6036. [CrossRef] [PubMed]
7. Cirrincione, G.; Randazzo, V.; Pasero, E. A neural based comparative analysis for feature extraction from ECG signals. In *Neural Approaches to Dynamics of Signal Exchanges*; Springer: Singapore, 2020; pp. 247–256.
8. Athaya, T.; Choi, S. An Estimation Method of Continuous Non-Invasive Arterial Blood Pressure Waveform Using Photoplethysmography: A U-Net Architecture-Based Approach. *Sensors* **2021**, *21*, 1867. [CrossRef] [PubMed]
9. Cheng, J.; Xu, Y.; Song, R.; Liu, Y.; Li, C.; Chen, X. Prediction of arterial blood pressure waveforms from photoplethysmogram signals via fully convolutional neural networks. *Comput. Biol. Med.* **2021**, *138*, 104877. [CrossRef] [PubMed]
10. Paviglianiti, A.; Randazzo, V.; Cirrincione, G.; Pasero, E. Neural recurrent approaches to noninvasive blood pressure estimation. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–7.
11. Kachuee, M.; Kiani, M.M.; Mohammadzade, H.; Shabany, M. Cuffless Blood Pressure Estimation Algorithms for Continuous Health-Care Monitoring. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 859–869. [CrossRef] [PubMed]
12. Kachuee, M.; Kiani, M.; Mohammadzade, H.; Shabany, M. Cuff-Less Blood Pressure Estimation. UCI Machine Learning Repository; 2015. Available online: <https://archive.ics.uci.edu/dataset/340/cuff+less+blood+pressure+estimation> (accessed on 30 June 2023).
13. Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **2000**, *101*, e215–e220. [CrossRef] [PubMed]
14. Ferretti, J.; Barbiero, P.; Randazzo, V.; Cirrincione, G.; Pasero, E. Towards uncovering feature extraction from temporal signals in deep CNN: The ECG case study. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–7.
15. Ferretti, J.; Randazzo, V.; Cirrincione, G.; Pasero, E. 1-D convolutional neural network for ECG arrhythmia classification. In *Progresses in Artificial Intelligence and Neural Systems*; Springer: Singapore, 2021; pp. 269–279.
16. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. arXiv **2016**, arXiv:1603.04467. [CrossRef]
17. tf.keras.optimizers.Adam | TensorFlow v2.11.0. Available online: <https://www.tensorflow.org/> (accessed on 30 June 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Financial Time Series Models—Comprehensive Review of Deep Learning Approaches and Practical Recommendations †

Mateusz Buczyński ^{1,2,*}, Marcin Chlebus ¹, Katarzyna Kopczewska ¹ and Marcin Zajenkowski ³

¹ Faculty of Economic Sciences, University of Warsaw, 00-241 Warsaw, Poland; mchlebus@wne.uw.edu.pl (M.C.); kkopczewska@wne.uw.edu.pl (K.K.)

² Interdisciplinary Doctoral School, University of Warsaw, 00-312 Warsaw, Poland

³ Faculty of Psychology, University of Warsaw, 00-183 Warsaw, Poland; zajenkowski@psych.uw.edu.pl

* Correspondence: mp.buczynski2@uw.edu.pl

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: There have been numerous advances in financial time series forecasting in recent years. Most of them use deep learning techniques. We identified 15 outstanding papers that have been published in the last seven years and have tried to prove the superiority of their approach to forecasting one-dimensional financial time series using deep learning techniques. In order to objectively compare these approaches, we analysed the proposed statistical models and then reviewed and reproduced them. The models were trained to predict, one day in advance, the value of 29 indices and the stock and commodity prices over five different time periods (from 2007 to 2022), with 4 in-sample years and 1 out-of-sample year. Our findings indicated that, first of all, most of these approaches do not beat the naive approach, and only some barely beat it. Most of the researchers did not provide enough data necessary to fully replicate the approach, not to mention the codes. We provide a set of practical recommendations of when to use which models based on the data sample that we provide.

Keywords: financial forecast; deep learning; reproducibility; forecast comparison

Citation: Buczyński, M.; Chlebus, M.; Kopczewska, K.; Zajenkowski, M. Financial Time Series Models—Comprehensive Review of Deep Learning Approaches and Practical Recommendations. *Eng. Proc.* **2023**, *39*, 79. <https://doi.org/10.3390/engproc2023039079>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 12 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Many researchers have been struggling for decades to understand how the markets behave [1–3]. Some argue that the markets are unpredictable due to the Efficient Market Hypothesis (EMH), stating that, in the short term, financial time series follow a random walk. In contrast, there is a large number of behavioural economists that do not agree with such a statement, believing that investors do not always behave rationally [4,5]. They suggest that the market “can be beaten”, because cognitive biases, such as overconfidence and herd behaviour or risk aversion, exist. One is certain and empirically confirmed: investors are winning on the market—mostly because they are ahead of their “brothers in arms”.

Generally, there are two main approaches that are used to predict the financial markets: technical and fundamental analysis. Technical analysis approaches focus widely on building the predictions based on the past movements or changes of the stock market [6,7]. On the other hand, fundamental analysis considers the information about the economic status of the company underlying the asset, news, social media, financial reports, etc. Lately, the most emphasis is put on employing machine or deep learning methods to combine these tasks, due to their ability to find and quantify nonlinear correlations very easily [8–12], but researchers are still struggling to provide an objective way to compare the results. There has been a massive progress in artificial intelligence approaches implemented in the financial area, mainly portfolio optimisation, time series prediction, agent-based modelling, etc. A large number of scientists also agree that the origin of successful prediction lies not only in the data related to the predicted object, but also in finding additional data sources [13–15].

Systematic reviews [16,17] show that there have been more than 125 new approaches to time series prediction in the past few years. Many researchers claim to provide better and better-performing models; however, no consensus exists on the best approach yet. The availability of numerous model options in the market without a clear indication of their costs can result in a dilemma known as “choice overload” [18]. This can lead to a situation where one may end up making no choice at all.

When it comes to financial time series, there are also many different areas that can be covered by deep learning. We can see different feature sets being used in the model, either univariate modelling or enriching the data with additional supporting data sources. There is also emerging work using text mining, sentiment analysis, or social media analysis in feature sets. The target variable can also change: it can predict stock prices, indices, commodities, or cryptocurrencies. Some researchers are also looking at volatility and trend forecasting. As far as models are concerned, the horizon is even broader: from simple neural networks, to long-term memory (LSTM) architectures, to sophisticated state-of-the-art approaches such as graph networks or generative adversarial networks. Finally, the prediction horizon is also a point of contention: Are predictions longer than one time step forward of good quality, and what should it be—a regression task or a classification task?

This abundance of different possibilities and options generates a large grid of approaches that cannot be compared with each other solely on the basis of the article provided. There are three main problems when it comes to comparing approaches to predicting financial time series in a deep learning setting:

- Different data, timespans, and metrics used in every experiment;
- Lack of publicly available codes supporting the experiment’s execution;
- Lack of a detailed architecture and hyperparameters that are necessary for the experiment’s reproduction.

The first problem stems from the lack of a single stock framework, indexes, or any other data samples to objectively test the effectiveness of the models. The usual duo is the S&P 500 and the SSE Composite, according to [17], accounting for 80% of the papers they reviewed. However, a large subset of researchers use single stock quotes or commodity prices. When models are supplemented with additional data, e.g., enriched with text-mining techniques, such datasets are not publicly available (only 10% of the reviewed papers in [17]). In time series problems, it is not only about the data, but also about the time sample to be used: different results will be obtained for models trained in 2019 and 2020, or a different training sample or lookback time horizon: different results will be obtained for models trained in 1 year and 5 years. In terms of metrics, there is some minor consensus there: for regression, the common metric is the mean absolute error; for classification, it is the accuracy. However, this consensus does not mean that every researcher provides a grid of “must have” metrics, but rather selects a few of the most-common ones.

The second problem stems from the reluctance of researchers to publish the code they used to train the models. Only three of the papers reviewed in this thesis were taken from publicly available sources such as GitHub. The lack of code reduces the usefulness of the work, as the cost of selection is increased by the time it takes to implement. Another disadvantage is the discrepancy between the implemented solution and the one presented in the original work (e.g., different versions of the base packages).

A final problem is the poor description of the approach along with the hyperparameters that are used to train the model. Typically, in time series deep learning, we can expect the following hyperparameters (depending on the approach implemented): number of training epochs, learning rate, optimiser type, batch size, and number of backward steps (number of time series lags). However, in the works mentioned by [16,17], there are huge gaps in the description of the training approach. Most of the approaches lack a concrete specification of the architecture used (number of neurons, number of layers, activation functions, etc.) or lack parameters. Researchers usually stop at explaining that the architecture used is LSTM or NN. Additionally, only one paper mentions the random seed value, which is also necessary to fully reproduce the model weights.

To overcome these problems, we decided to carry out an extensive practical reproduction of fifteen papers that are listed in [16,17]. We rebuilt each approach that was reported in a duplicate article, taking into account the hyperparameters that were reported. Where important parameters were missing (such as the number of neurons, the optimiser, the learning rate, or the number of training epochs), we supplemented with the average of the remaining, non-missing articles. We compared these models with simple statistical approaches—naive forecasting, ARIMA, and exponential smoothing. The result was 18 forecasts over five different time periods, resulting in 90 forecasts, run for 29 different types of financial data, including indices, equities, and commodities. We compared models using the mean absolute percentage error (MAPE), mean-squared error (MSE), mean absolute error (MAE), and mean absolute error compared in the first time step only. We propose a data, time, and model framework to run when performing time series problems with deep learning.

2. Methods

We reviewed 15 different deep learning models that have been mentioned in recent literature reviews on financial time series prediction. We focused on selecting the broadest possible sample of different deep learning models. In this section, we describe them briefly. To limit the size of the article, we refer the reader to the original papers for more details on the individual models:

1. **A hybrid attention-based EMD-LSTM model [19]:**
The paper proposes a two-stage model for time series prediction, combining empirical mode decomposition (EMD) and attention-based long short-term memory (LSTM-ATTE). EMD was used to decompose the time series into a few inherent mode functions (IMFs), which were then taken as the input to LSTM-ATTE for prediction. They used the SSE Composite index to run the predictions. The attention mechanism was used to extract the input features of the IMF and improve the accuracy of the prediction. Researchers have evaluated the model's predictive quality using linear regression analysis of the stock market index and compared it to other models, showing better prediction accuracy.
2. **Empirical. mode decomposition factorisation neural network (EMD2FNN) model [20]:**
A simpler approach proposed by [20], includes feeding the IMFs of some time series into a factorisation neural network, concatenating all the IMFs into a single vector. The data used for the experiment were: the SSE Composite, NASDAQ, and S&P 500. The authors performed a thorough comparison between the proposed method and other neural network models, comparing the mean absolute error (MAE) and root-mean-squared error (RMSE).
3. **Neural network ensemble [21]:**
The paper describes a deep neural network ensemble that aims to predict the SSE Composite and SZSE (Shenzhen) Component. The model consists of a set of neural networks that were trained using open, high, low, close (OHLC) data. Every neural network takes the last few days of such data, flattened to a vector form. Later, bagging is used to combine these networks and reduce the generalisation error.
4. **Wavelet denoising long short-term memory model [22]:**
The proposed model in this paper is a combination of real-time wavelet denoising and the LSTM neural network. The wavelet denoising was used to separate signals from noise in the stock data and was then taken as the input to the LSTM model. The authors conducted an experiment on several indexes, including the SSE, SZSE, and NIKKEI, using the mean absolute percentage error (MAPE) as a metric.
5. **Dual-stage attention-based recurrent neural network [23]:**
This paper proposes a two-stage attention-based recurrent neural network (DA-RNN) model for time series prediction. The DA-RNN model uses an input attention mechanism in the first stage to extract the relevant driving series at each time step based on the previous hidden state of the encoder. In the second stage, the temporal attention

mechanism is used to select the relevant hidden encoder states at all time steps. The experiment was conducted on the SML 2010 and NASDAQ datasets and showed that the model outperformed state-of-the-art time series prediction methods. The metrics used were the MAE, MAPE, and RMSE.

6. **Bidirectional LSTM [24]:**

This paper compared the performance of the bidirectional LSTM (BiLSTM) and unidirectional LSTM models. BiLSTM is able to traverse the input data twice (left to right and right to left) and, thus, has additional training capabilities. The study showed that BiLSTM-based modelling offers better predictions than regular LSTM-based models and outperformed the ARIMA and LSTM models. However, BiLSTM models reach equilibrium much slower than LSTM-based models. The experiment was carried out on several indices and stocks, including the Nikkei and NASDAQ, as well as the daily IBM share price and compared using RMSE.

7. **Multi-scale, recurrent convolutional neural network [25]:**

The proposed method is a multi-scale temporal dependent recurrent convolutional neural network (MSTD-RCNN). The method utilises convolutional units to extract features on different time scales (daily, weekly, monthly) and a recurrent neural network (RNN) to capture the temporal dependency (TD) and complementarity across different scales of financial time series. The proposed method was evaluated on three financial time series datasets from the Chinese stock market and achieved state-of-the-art performance in trend classification and simulated trading compared to other baseline models.

8. **Time-weighted, LSTM [26]:**

This paper proposes a novel approach to predicting stock market trends by adding a time attribute to stock market data to improve prediction accuracy. The approach involves assigning weights to the data according to their temporal proximity and using formal stock market trend definitions. The approach also uses a custom long short-term memory (LSTM) network to discover temporal relationships in the data. The results showed that the proposed approach outperformed other models and can be generalised to other stock indices, achieving 83.91% accuracy in a test with the CSI 300 index.

9. **ModAugNet [27]:**

The paper proposes a data augmentation approach for stock market index forecasting through the ModAugNet framework, which consists of a fitting-prevention LSTM module and a prediction LSTM module. The prediction module is a simple LSTM network that is fit based only on the historical data on the index realised prices. The prevention module builds on that by adding a set of regressors that are other indexes, highly correlated with the predicted one. Using the MSE, MAE, and MAPE on the S&P500 and KOSPI200, the authors proved the validity of their solution.

10. **State frequency memory (SFM) [28]:**

The state frequency memory (SFM) model is the twin of the LSTM model. The SFM model was inspired by the discrete Fourier transform (DFT) and was designed to capture multi-frequency trading patterns from past market data to make long- and short-term predictions over time. The model decomposes the latent states of memory cells into multiple frequency components, where each component models a specific frequency of the latent trading pattern underlying stock price fluctuations. The model then predicts future share prices by combining these frequency components. The authors tested their solution of 50 different stocks in 10 industries using the MSE.

11. **Convolutional neural-network-enhanced support vector machine [29]:**

The proposed model in this text is a convolutional neural network (CNN), which is supposed to discover features in the data, which are later passed into the support vector machine (SVM) model. The text then discusses the influence of the model parameters on the prediction results. The model was evaluated empirically on the

Hong Kong Hang Seng Index using the RMSE, and the results showed that both models are feasible and effective.

12. **Generative adversarial network [30]:**
The generative adversarial network (GAN) in this paper consists of two main components: a discriminator and a generator. The discriminator was designed using a simple feed-forward neural network and is responsible for distinguishing real stock market data from generated data. The generator, on the other hand, was built using an LSTM and is responsible for generating data with the same distribution as the actual stock market data. The model was trained on daily data from the S&P500 index and several other stocks for a wide range of trading days. The LSTM generator learns the distribution of the stock data and generates new data, which are then fed to the MLP discriminator. The discriminator learns to distinguish between the actual stock data and the data generated by the generator. The authors tested their model on several time series, including the S&P 500 and stocks such as IBM or MSFT.
13. **Long short-term memory and gated recurrent unit models [31]:**
The paper proposes a hybrid model that combines the long short-term memory (LSTM) and gated recurrent unit (GRU) networks. The authors used the S&P 500 historical time series data and evaluated the model using metrics such as the MSE and MAPE on the S&P500.
14. **CNN and bi-directional LSTM model [32]:**
The paper proposes a model combining multiple pipelines of convolutional neural network (CNN) and bidirectional long short-term memory (LSTM) units. The model improved the prediction performance by 9% compared to a single pipelined deep learning model and by more than six-times compared to a support vector machine regressor model on the S&P 500. The paper also illustrates the improvement in the prediction accuracy while minimising overfitting by presenting several variants of multi- and single-pipelined deep learning models based on different CNN kernel sizes and number of bidirectional LSTM units.
15. **Tim convolution (TC) LSTM model [33]:**
The authors of this paper propose time convolution long short-term memory (TC-LSTM), employing convolutional neural networks (CNNs) to capture long-term fluctuation features in the stock prices and combining this with LSTM. This combination allows the model to capture both the long-term dependencies of stock prices, as well as the overall change pattern. The authors compared the performance of their TC-LSTM model to three baseline models on 50 stocks from the SSE 50, as well as the index itself. They showed that their model outperformed the others in terms of the mean-squared error.

The proposed architectures were build from scratch in pytorch [34], based on the explanation provided in the article itself. In addition to these models, to provide a more thorough comparison, we also utilised the ARIMA model (tuned, best parameters on the training sample), the naive approach (prediction as: $Y_t = Y_{t-1}$), and exponential smoothing.

The hyperparameters derived based on the text or the publicly available codes are presented in Table A1. All missing parameters were filled in with either the mean of other parameters or the mode.

3. Data and Methodology

We provide a more comprehensive background of the developed models we propose to broaden the range of the time series on which the model is tested. The data used in this study were taken from the following financial types: indexes, currency pairs, stocks, cryptocurrencies, and commodities. The purpose of including a comprehensive range of financial types was to provide a comprehensive comparison of the models' performance.

The following time series were included in this study for analysis:

- Indexes: WIG20 (PL), S&P 500 (US), NASDAQ (US), Dow Jones Industrial (US), FTSE 250 (UK), Nikkei 225 (JP), DJI (USA), KOSPI 50 (KR), SSE Composite (CN), DAX 40 (DE), CAC40 (FR);
- Currency pairs: EURPLN, PLNGBP, USDPLN, EURUSD, EURGBP, USDGBP, CHFGBP, CHFUSD, EURCHF, PLNCHF;
- Stocks: AAPL, META, AMZN, TSLA, GOOG, NFLX;
- Cryptocurrencies: BTCUSD;
- Commodities: XAUUSD.

For each of these time series, we identified five periods in which we made predictions:

- 2016–2020;
- 2013–2017;
- 2007–2011;
- 2009–2013;
- 2018–2022.

Each period consisted of 4 years of training and 1 year of day-ahead predictions (ca. 250 testing time steps) without re-training the model. The periods differed significantly between each other due to the different levels of variability between the training and test trials.

The data were preprocessed by performing normalisation on the input features using the MinMaxScaler function from the Scikit-learn library [35]. The normalised data were then split into training and test sets with a ratio of 4 years:1 year, respectively. The model was then trained using only the training sample, and predictions were made for every time step in the testing sample.

The evaluation metrics were used to compare the performance of each model across the different financial types and time series. The best-performing model was selected based on the lowest values of:

- $MSE = \frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_t)^2;$
- $RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_t)^2};$
- $MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{Y_t - \hat{Y}_t}{Y_t} \right| \times 100\%.$

Y_t is the actual value at time t ; \hat{Y}_t is the predicted value at time t ; n is the total number of time periods.

Finally, the results were analysed to identify patterns and insights that could help improve the accuracy of predictions in future studies. We also provide the MAPEs at the first time step of testing (i.e., calculated for $t = 1$) to provide a quality metric for the trained model with the full set of information, i.e., to provide a metric that would allow the quality of the model to be calculated in the short term.

4. Results

The results are presented in Table 1. To be concise, we only report the MAPE in this paper (for the MAE, MSE, or detailed predictions, please contact the authors directly). In bold, we can notice the best (lowest) MAPE for every financial time series.

We can see that the best model was the naive approach. This was mainly due to the fact that the quality of models tends to deteriorate if they are not retrained after a certain period of time; however, we wanted to keep the reproduction of the models as close to the original as possible. Furthermore, the researchers in their original work also did not retrain the model in the test sample (or mention doing so in any other way), nor did they compare to statistical approaches.

When we leave out the statistical approaches, we can see a few approaches that stand out (5, 6, 7, 9, 12, and 13). These models generally have MAPEs lower than five percent. What these models have in common is either simple LSTM/RNN architectures or sophisticated CNN operations, which increase the range of features. The best model that achieved an average MAPE of 1.79% was the multi-scale recurrent convolutional

neural network (Model 7). We believe that the surplus in prediction quality came from the additional operations performed on the data (multi-scale CNN), which improved the information processing. The second-best model was ModAugNet (Model 9), which achieved a MAPE of 2.07%. This model relies heavily on the additional data sources that are provided in training for a given time series. On the other hand, the worst model was Model 11 using SVM after preprocessing the data from the SVM—52.24% MAPE.

From another perspective, the worst-performing financial time series was BTCUSD, followed by TSLA, AAPL, and NFLX. These all achieved high returns over the time periods studied, so what we believe is the reason for the deterioration was the inability of all models to correctly identify and predict rapid price increases. We also observed that indices have a roughly similar MAPE (5–6%), as do currency pairs (1–2%). Stocks, on the other hand, have the highest MAPE of all financial time series (>10%).

Since we can clearly see that the models performed worse due to the lack of hyper-tuning, we propose to run this procedure for each model training. However, such an experiment will be computationally exhaustive (this experiment already consisted of 2610 model trainings), so some restrictions should be introduced. Furthermore, the testing procedure was detached from how these models are used in reality. Time series models should be trained daily to provide the best-possible fit based on the set of information available at the time of prediction. In this case, the information set becomes smaller and smaller with each prediction.

As a proof of this statement, we provide in Table 2 the MAPE results calculated for the first testing time step. This metric allowed us to confirm that, in the short term, these models are correct and better than the naive approach for each stock. Only for five series was the statistical approach (exponential smoothing) better. Model 7 was still the best in 7 of the remaining 24 cases. Several other models came in first (Model 1 and Model 13) proving that the performance of the models deteriorates over time.

Table 1. An average of the mean absolute percentage error for every model and every financial time series calculated over five different timespans.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	ARIMA naive	ExpSmooth	
AAPL	10.33	19.98	63.42	10.77	5.55	5.53	2.48	7.12	1.69	7.32	56.94	4.47	4.32	9.74	8.36	1.7	1.42	1.57
AMZN	14.57	19.21	64.15	12.36	6.78	7.01	2.94	12.11	2.82	6.08	31.67	4.98	3.83	14.8	14.73	1.88	1.61	1.62
BTUSD	19.74	41.24	72.07	22.48	17.23	25.6	5.38	24.5	11.3	19.8	1110.76	12.46	17.05	36.32	32.05	3.81	3.17	3.37
CHFGBP	3.03	5.16	18.39	1.73	0.99	1.02	0.7	2.26	0.58	4.48	6.44	0.67	0.91	2.82	1.05	0.49	0.46	0.46
CHFUSD	1.7	4.69	9.97	2.9	1.42	1.07	0.56	1.13	0.79	1.23	7.74	0.73	0.79	1.96	1.24	0.51	0.44	0.45
EURCHF	1.78	2.76	6.66	1.49	1.2	0.84	0.57	1.15	0.39	1.47	4.6	0.55	0.57	3.21	2.2	0.35	0.31	0.32
EURGBP	1.28	2.42	11.31	1.02	0.83	0.58	0.52	0.64	0.39	0.6	1.86	0.45	0.68	1.61	1.0	0.42	0.38	0.39
EURPLN	0.85	2.01	8.39	1.12	0.65	0.6	0.47	0.85	0.39	1.04	1.74	0.41	0.49	1.41	0.65	0.41	0.34	0.36
EURUSD	1.35	2.95	6.55	1.86	1.18	1.03	0.67	1.07	1.24	1.13	2.28	0.68	0.83	2.27	1.82	0.49	0.43	0.43
GOOG	10.0	9.53	44.2	7.98	4.64	4.42	2.91	15.73	2.28	5.85	9.43	4.04	6.99	7.53	9.93	2.08	1.77	1.97
META	23.0	18.65	34.98	14.39	7.06	6.01	3.19	14.99	2.48	8.14	15.44	5.51	7.69	15.65	17.61	2.98	2.41	1.90
NFLX	18.02	29.01	61.32	16.19	8.92	7.86	5.33	28.51	3.46	9.9	33.45	5.55	5.73	21.48	21.07	2.86	2.26	2.28
PLNCHF	1.86	5.54	6.31	2.61	1.92	2.0	0.78	2.3	0.56	2.9	6.93	1.0	1.14	3.94	2.73	0.54	0.51	0.69
PLNGBP	0.61	2.04	10.86	1.27	1.01	0.73	0.7	1.28	0.55	0.78	2.28	0.61	0.83	2.11	1.38	0.54	0.51	0.56
TSLA	17.76	33.73	71.83	26.83	14.56	23.22	4.03	27.29	14.78	21.81	65.77	16.47	13.34	34.94	23.61	3.47	2.98	2.48
USDGBP	2.3	3.47	13.73	1.81	0.89	0.88	0.71	1.18	0.47	0.73	2.0	0.56	0.97	1.81	1.04	0.48	0.45	0.46
USDPLN	1.82	5.01	16.22	2.96	1.64	1.72	0.77	1.59	0.7	2.09	3.46	0.84	1.26	3.12	1.51	0.73	0.61	0.59
XAUUSD	5.76	8.55	28.57	5.53	3.14	3.55	1.42	2.43	1.35	4.67	13.33	1.74	1.74	6.74	3.45	0.94	0.79	0.90
CAC	2.41	7.19	26.6	2.51	2.18	1.93	1.52	3.03	1.0	2.11	6.19	1.27	2.08	4.75	3.49	1.21	0.99	0.92
DAX	3.66	8.89	34.69	3.77	2.7	2.71	1.62	5.21	1.07	2.14	6.01	1.64	2.5	4.93	4.32	1.2	0.98	0.92
DJC	4.64	8.2	38.77	4.07	3.52	2.67	1.28	3.52	1.16	2.27	11.45	2.1	1.92	6.74	5.92	1.0	0.86	1.22
DJI	5.85	9.01	38.47	3.09	2.51	2.21	1.48	4.65	1.12	2.42	15.79	2.01	1.79	4.15	5.91	0.97	0.83	1.20
FTM	4.55	9.58	33.35	3.47	3.42	3.01	1.38	3.92	1.04	3.38	16.33	2.05	2.04	5.57	3.49	0.98	0.84	0.96
KOSPI	3.54	8.65	31.47	4.7	2.06	2.83	1.75	2.85	1.15	2.28	5.59	1.41	2.36	4.28	4.09	1.03	0.88	1.11
NDX	7.77	11.68	49.29	7.26	3.99	4.65	2.2	7.07	1.47	4.84	34.68	3.04	3.47	8.89	7.94	1.25	1.07	24.63
NKX	3.75	10.5	29.97	4.8	3.05	4.19	2.14	3.69	2.59	3.06	10.39	1.55	2.97	8.44	4.03	1.18	1.01	0.58
SHC	1.71	7.52	19.43	2.26	1.76	1.24	1.15	3.39	0.86	1.33	6.3	0.97	1.59	3.38	2.21	0.94	0.77	0.78
SPX	4.86	8.05	40.44	4.91	3.44	2.95	1.63	6.66	1.21	4.66	18.57	2.07	2.38	7.43	5.64	1.04	0.89	1.21
WIG20	3.32	8.54	25.5	3.67	2.82	1.91	1.52	3.1	1.18	1.77	7.59	1.54	2.16	4.76	4.53	1.33	1.11	1.27

GOOG and META calculated for two forecasting horizons; TSLA and BTCUSD calculated for three forecasting horizons. Numbers in bold indicate minimum for a given timeseries.

Table 2. An average of the mean absolute percentage error at the first predicted time step for every model and every financial time series calculated over five different timespans.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	ARIMA naive	ExpSmooth	
AAPL	6.64	16.68	62.55	3.97	3.31	2.91	1.65	2.96	2.04	1.69	10.48	2.99	1.98	6.37	5.80	2.16	2.03	1.90
AMZN	5.23	11.70	63.84	1.86	1.88	1.89	1.54	7.67	2.01	2.03	7.86	2.24	1.89	3.56	1.73	2.15	2.02	1.81
BTCUSD	15.14	26.29	73.84	8.68	9.68	11.37	4.49	10.38	3.34	2.91	9.83	4.65	7.26	15.16	10.13	4.11	3.29	3.21
CHFGBP	0.88	3.50	16.07	0.78	0.71	0.80	0.36	1.88	0.44	1.04	3.31	0.58	0.75	2.37	1.40	0.43	0.46	0.48
CHFUSD	0.80	4.36	9.62	2.53	0.37	0.51	0.44	0.79	0.47	0.58	3.42	0.53	0.28	0.91	0.33	0.40	0.41	0.42
EURCHF	0.57	2.76	4.44	0.51	0.54	0.25	0.28	0.58	0.12	0.26	3.48	0.20	0.18	1.92	1.61	0.16	0.14	0.17
EURGBP	0.85	2.61	9.11	0.96	0.79	0.21	0.26	0.62	0.38	0.42	2.27	0.24	0.47	0.98	0.50	0.47	0.34	0.34
EURPLN	0.44	1.61	6.60	0.50	0.41	0.31	0.56	0.64	0.19	0.34	1.29	0.33	0.41	0.71	0.54	0.31	0.27	0.25
EURUSD	0.27	2.23	4.90	1.01	0.53	0.57	0.49	1.02	0.40	0.27	2.10	0.40	0.82	0.96	1.37	0.31	0.40	0.38
GOOG	3.11	7.19	47.74	1.07	0.56	1.83	1.55	19.93	1.24	2.76	9.02	1.57	6.62	1.65	2.13	1.20	1.25	1.38
META	2.03	12.28	47.44	2.60	2.07	2.14	0.60	11.48	1.34	1.28	3.07	2.00	1.16	3.68	3.84	1.49	1.40	1.39
NFLX	9.50	13.23	55.41	3.70	1.59	1.39	2.59	19.01	1.66	4.36	9.67	1.26	4.00	4.35	3.36	1.66	1.56	1.26
PLNGHF	0.32	4.85	3.52	0.86	0.71	1.19	0.51	0.96	0.27	0.33	4.12	0.46	0.25	1.45	2.17	0.38	0.39	0.37
PLNGBP	0.22	1.20	10.07	0.59	0.74	0.28	0.23	1.20	0.41	0.32	2.83	0.23	0.22	0.97	0.53	0.39	0.38	0.36
TSLA	11.77	21.97	65.74	11.01	9.36	7.85	4.86	9.54	5.78	5.78	4.09	4.94	10.58	11.63	6.78	4.68	5.40	5.56
USDGBP	1.30	3.17	12.62	0.40	0.52	0.37	0.31	1.75	0.51	0.51	2.04	0.19	1.02	1.53	1.02	0.53	0.51	0.47
USDPLN	0.76	3.28	15.68	1.35	0.83	0.72	0.33	0.75	0.40	0.34	1.98	0.53	1.15	1.95	1.15	0.31	0.33	0.30
XAUUSD	1.20	6.98	26.94	1.21	1.50	0.88	0.53	1.27	0.76	1.00	5.12	0.75	1.17	1.15	1.09	1.06	0.80	0.79
CAC	0.79	7.72	30.38	1.52	1.57	1.54	1.08	2.04	1.63	1.30	4.87	1.42	1.49	3.59	2.66	1.61	1.46	1.46
DAX	1.63	8.79	36.95	2.12	2.28	1.64	0.76	4.24	1.27	1.15	4.66	1.22	1.61	3.17	3.66	1.61	1.22	1.28
DJC	2.72	6.95	38.85	1.31	1.88	1.34	1.07	0.59	0.77	1.73	6.72	1.21	0.70	3.51	2.61	1.03	0.80	0.77
DJI	2.61	7.17	38.40	1.68	1.10	1.39	1.01	1.99	1.01	1.11	7.44	1.35	1.73	1.88	3.38	1.23	1.10	1.08
FTM	1.62	9.71	37.59	1.83	3.22	2.25	0.93	2.69	1.09	1.05	5.82	1.69	1.24	3.77	3.02	1.45	1.28	1.64
KOSPI	1.26	6.42	31.89	2.01	1.20	1.33	1.30	2.18	0.81	0.52	3.18	1.29	2.62	1.18	1.25	1.21	0.80	0.77
NIDX	2.96	8.35	48.43	3.06	2.31	2.57	1.66	3.18	1.56	1.59	6.95	1.86	2.21	3.56	3.59	1.68	1.67	1.60
NKX	1.74	6.67	30.44	2.75	2.46	1.96	2.32	3.71	1.87	2.18	5.78	2.19	2.81	2.55	2.90	2.68	2.10	2.04
SHC	1.24	4.93	22.28	1.77	1.98	1.70	1.22	2.91	0.93	1.29	5.30	1.39	0.79	2.60	2.32	1.52	0.86	1.00
SPX	2.10	7.42	40.64	2.86	1.43	2.20	1.31	5.93	1.11	3.78	7.87	1.48	1.86	4.81	2.50	1.30	1.18	1.10
WIG20	0.56	5.00	30.25	3.92	2.92	1.53	1.17	3.51	1.25	1.82	7.55	1.37	2.26	3.80	3.44	1.71	1.32	1.34

GOOG and META calculated for two forecasting horizons. TSLA and BTCUSD calculated for three forecasting horizons. Numbers in bold indicate minimum for a given timeseries.

5. Conclusions

The experiment presented in this paper aimed to compare the predictive performance of various deep learning models on different financial time series. The experiment was conducted using the data of daily prices for 29 financial time series, including stocks, indexes, and currency pairs, over a period of 15 years (2007–2022).

The models used in the experiment included classical statistical approaches such as exponential smoothing and ARIMA, as well as deep learning models such as NN, LSTM, CNN, or GAN. The models were trained using a sliding window approach and evaluated using the mean absolute percentage error, mean squared error, and mean absolute error, as well as the mean absolute percentage error at the first time step.

The results of the experiment showed that the best model was the naive approach, but when disregarding the statistical approaches, several deep learning models showed promising results. In particular, the multi-scale recurrent convolutional neural network (Model 7) achieved the best MAPE of 1.79% on average, while ModAugNet (Model 9) achieved a MAPE of 2.07%. The worst-performing model was Model 11, which utilised SVM after data preprocessing with the CNN.

Based on the results presented in this study, it can be concluded that simple time series models, even naive approach, can perform relatively well against more-complex deep learning models in forecasting financial time series, notably in the long run. However, deep learning models, in particular those using LSTM/RNN architectures or complex CNN functions, have the potential to outperform statistical models in the short term, provided they are regularly retrained and properly tuned.

It has also been observed that the quality of models tends to deteriorate if they are not retrained after a certain period of time. This highlights the importance of regular retraining of time series models to ensure the best-possible fit based on all the information available at the time of forecasting. In addition, it was noted that stocks tend to have a higher MAPE than indices or currency pairs, which may be due to their higher volatility and the need for more sophisticated modelling techniques.

Author Contributions: Conceptualization: M.B. and M.C.; methodology and software: M.B.; writing—original draft preparation: M.B.; validation: M.C.; formal analysis: all authors; writing—review and editing: M.C., K.K. and M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Program of Integrated Activities for the Development of the University of Warsaw (ZIP Program), co-financed by the European Social Fund under the Knowledge Education Development Operational Program 2014-2020, Path 3.5.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analysed in this study. These data can be found here: <https://stooq.com/> (accessed on 12 June 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ARIMA	Autoregressive integrated moving average
BiLSTM	Bidirectional LSTM
CNN	Convolutional neural network
DFT	Discrete Fourier transform
EMD	Empirical mode decomposition
EMH	Efficient market hypothesis
GAN	Generative adversarial network
GRU	Gated recurrent unit
IMF	Inherent mode function

LSTM	Long short-term memory
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MSE	Mean-squared error
NN	Neural network
OHLC	Open, high, low, close
RNN	Recurrent neural network
SFM	State frequency memory

Appendix A. Hyperparameters Used for Training

Table A1. List of hyperparameters used for training.

Model No.	NN Architecture	Epochs	Learning Rate	Optimiser	Batch Size	Steps Back	Data Sample	Performance Metrics
1	-	-	-	Adam	-	20	3407 (Jan 2004–Jan 2018)	MAE, RMSE, MAPE, R ²
2	-	-	-	SGD	-	3,4,5	Jan 2012–Dec 2016/Jan 2007–Dec2011	RMSE, MAE, MAPE
3	randomly selected number of layers (1–6) ensembled 10 times	200,000	0.0001	Adam	-	20	-	relative error
4	2 layer LSTM, with 1/2 and 1/3 input neurons	-	-	-	-	2, 4, 8, 16, 32, 64, 128, 256, 512	Jan 2010–Dec 2016 (testing last year)	MAPE
5	grid search over layer sizes (16, 32, 64, 128, 256)	-	0.001 (decreasing)	Adam	128	3, 5, 10, 15, 25	Jul 2016 - Dec 2016 minutely data	RMSE, MAE, MAPE
6	one layer, 4 neurons differently (3)	1 or 2	-	Adam	-	-	Jan 1985–Aug 2018	RMSE
7	scaled time series -> CNN (16 filters)-> GRU (16 × 3)	100	0.0005	Adam	32	30	Jan 2016–Dec 2016	accuracy
8	320 neurons LSTM x3	4500	0.0024	-	-	20	Jan 2002–Dec 2017	accuracy
9	1 LSTM 2 layers 5 and 3 neurons, 2LSTM: 4 and 2	200	0.00005	Adam	32	20	Jan 2000–July 2017	MSE MAE MAPE
10	-	4000	0.01	RMSProp	-	3, 5, 10, 15, 20	2007–2014	MSE MAE MAPE
11	G: LSTM -> 7 neuron FC D: FC NN with 3 layers (72, 100, 10 neurons)	-	-	-	-	5	last 20 years	MAE MSE MAPE
12	2–4 CNN layers	-	-	-	-	30, 40, 50, 60	1990–2014	MSE
13	-	20	0.001	Adam	-	-	1950–2016	MAE MSE MAPE
14	CNN -> MaxPooling -> LSTM -> Dense	-	-	AdaDelta	-	50	2008–2018	MSE
15	-	-	-	-	-	100	2008–2017	MSE

References

1. Ang, A.; Bekaert, G. Stock Return Predictability: Is It There? *Rev. Financ. Stud.* **2007**, *20*, 651–707. [CrossRef]
2. Campbell, J.Y.; Hamao, Y. Predictable Stock Returns in the United States and Japan: A Study of Long-Term Capital Market Integration. *J. Financ.* **1992**, *47*, 43–69. [CrossRef]
3. Granger, C.W.J.; Morgenstern, O. *Predictability of Stock Market Prices*, 1st ed.; Heath Lexington Books: Lexington, MA, USA, 1970.
4. Bollerslev, T.; Marrone, J.; Xu, L.; Zhou, H. Stock Return Predictability and Variance Risk Premia: Statistical Inference and International Evidence. *J. Financ. Quant. Anal.* **2014**, *49*, 633–661. [CrossRef]
5. Phan, D.H.B.; Sharma, S.S.; Narayan, P.K. Stock Return Forecasting: Some New Evidence. *Int. Rev. Financ. Anal.* **2015**, *40*, 38–51. [CrossRef]
6. Campbell, J.Y.; Thompson, S.B. Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *Rev. Financ. Stud.* **2008**, *21*, 1509–1531. [CrossRef]
7. Agrawal, J.; Chourasia, V.; Mittra, A. State-of-the-Art in Stock Prediction Techniques. *Int. J. Adv. Res. Electr. Electron. Instrum. Energy* **2013**, *2*, 1360–1366.
8. Yim, J. A Comparison of Neural Networks with Time Series Models for Forecasting Returns on a Stock Market Index. In *Developments in Applied Artificial Intelligence; Lecture Notes in Computer Science*; Hendtlass, T., Ali, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2002; pp. 25–35. [CrossRef]
9. Bao, W.; Yue, J.; Rao, Y. A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and Long-Short Term Memory. *PLoS ONE* **2017**, *12*, e0180944. [CrossRef]
10. Lahmiri, S.; Bekiros, S. Cryptocurrency Forecasting with Deep Learning Chaotic Neural Networks. *Chaos Solitons Fractals* **2019**, *118*, 35–40. [CrossRef]
11. Long, W.; Lu, Z.; Cui, L. Deep Learning-Based Feature Engineering for Stock Price Movement Prediction. *Knowl.-Based Syst.* **2018**, *164*, 163–173. [CrossRef]
12. Chong, E.; Han, C.; Park, F.C. Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies. *Expert Syst. Appl.* **2017**, *83*, 187–205. [CrossRef]
13. Salinas, D.; Flunkert, V.; Gasthaus, J.; Januschowski, T. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. *Int. J. Forecast.* **2020**, *36*, 1181–1191. [CrossRef]
14. Oreshkin, B.; Carpo, D.; Chapados, N.; Bengio, Y. N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting. *arXiv* **2019**, arXiv:1905.10437.
15. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 Competition: 100,000 Time Series and 61 Forecasting Methods. *Int. J. Forecast.* **2020**, *36*, 54–74. [CrossRef]
16. Sezer, O.; Gudelek, U.; Ozbayoglu, M. Financial Time Series Forecasting with Deep Learning: A Systematic Literature Review: 2005–2019. *Appl. Soft Comput.* **2020**, *90*, 106181. [CrossRef]
17. Jiang, W. Applications of Deep Learning in Stock Market Prediction: Recent Progress. *Expert Syst. Appl.* **2021**, *184*, 115537. [CrossRef]
18. Reutskaja, E.; Lindner, A.; Nagel, R.; Andersen, R.A.; Camerer, C.F. Choice Overload Reduces Neural Signatures of Choice Set Value in Dorsal Striatum and Anterior Cingulate Cortex. *Nat. Hum. Behav.* **2018**, *2*, 925–935. [CrossRef]
19. Chen, L.; Chi, Y.; Guan, Y.; Fan, J. A Hybrid Attention-Based EMD-LSTM Model for Financial Time Series Prediction. In Proceedings of the 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 25–28 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 113–118. [CrossRef]
20. Zhou, F.; Zhou, H.; Yang, Z.; Yang, L. EMD2FNN: A Strategy Combining Empirical Mode Decomposition and Factorization Machine Based Neural Network for Stock Market Trend Prediction. *Expert Syst. Appl.* **2018**, *115*, 136–151. [CrossRef]
21. Yang, B.; Gong, Z.J.; Yang, W. Stock Market Index Prediction Using Deep Neural Network Ensemble. In Proceedings of the 2017 36th Chinese Control Conference (CCC), Dalian, China, 11 September 2017; pp. 3882–3887. [CrossRef]
22. Li, Z.; Tam, V. Combining the Real-Time Wavelet Denoising and Long-Short-Term-Memory Neural Network for Predicting Stock Indexes. In Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, USA, 27 November–1 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–8. [CrossRef]
23. Qin, Y.; Song, D.; Chen, H.; Cheng, W.; Jiang, G.; Cottrell, G. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. *arXiv* **2017**, arXiv:1704.02971.
24. Siami-Namini, S.; Tavakoli, N.; Namin, A.S. A Comparative Analysis of Forecasting Financial Time Series Using ARIMA, LSTM, and BiLSTM. *arXiv* **2019**, arXiv:1911.09512.
25. Guang, L.; Xiaojie, W.; Ruifan, L. Multi-Scale RCNN Model for Financial Time-series Classification. *arXiv* **2019**, arXiv:1911.09359.
26. Zhao, Z.; Rao, R.; Tu, S.; Shi, J. Time-Weighted LSTM Model with Redefined Labeling for Stock Trend Prediction. In Proceedings of the 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), Boston, MA, USA, 6–8 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1210–1217. [CrossRef]
27. Baek, Y.; Kim, H.Y. ModAugNet: A New Forecasting Framework for Stock Market Index Value with an Overfitting Prevention LSTM Module and a Prediction LSTM Module. *Expert Syst. Appl.* **2018**, *113*, 457–480. [CrossRef]
28. Zhang, L.; Aggarwal, C.; Qi, G.J. Stock Price Prediction via Discovering Multi-Frequency Trading Patterns. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; ACM: New York, NY, USA, 2017; pp. 2141–2149. [CrossRef]

29. Cao, J.; Wang, J. Stock Price Forecasting Model Based on Modified Convolution Neural Network and Financial Time Series Analysis. *Int. J. Commun. Syst.* **2019**, *32*, e3987. [CrossRef]
30. Zhang, K.; Zhong, G.; Dong, J.; Wang, S.; Wang, Y. Stock Market Prediction Based on Generative Adversarial Network. *Procedia Comput. Sci.* **2019**, *147*, 400–406. [CrossRef]
31. Hossain, M.A.; Karim, R.; Thulasiram, R.; Bruce, N.D.B.; Wang, Y. Hybrid Deep Learning Model for Stock Price Prediction. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1837–1844. [CrossRef]
32. Eapen, J.; Bein, D.; Verma, A. Novel Deep Learning Model with CNN and Bi-Directional LSTM for Improved Stock Market Index Prediction. In Proceedings of the 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 7–9 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 264–270. [CrossRef]
33. Zhan, X.; Li, Y.; Li, R.; Gu, X.; Habimana, O.; Wang, H. Stock Price Prediction Using Time Convolution Long Short-Term Memory Network. In *Knowledge Science, Engineering and Management; Lecture Notes in Computer Science*; Liu, W., Giunchiglia, F., Yang, B., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 461–468. [CrossRef]
34. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
35. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Proceeding Paper

Urban Heat Island Intensity Prediction in the Context of Heat Waves: An Evaluation of Model Performance [†]

Aner Martinez-Soto ^{1,2,*}, Johannes Fürle ^{2*} and Alexander Zipf ^{2,3*}

¹ Department of Civil Engineering, Faculty of Engineering and Science, Universidad de La Frontera, Temuco 4780000, Chile

² GIScience, Institute of Geography, University of Heidelberg, 69120 Heidelberg, Germany; johannes.fuerle@uni-heidelberg.de (J.F.); zipf@uni-heidelberg.de (A.Z.)

³ Heidelberg Institute for Geoinformation Technology gGmbH, 69118 Heidelberg, Germany

* Correspondence: aner.martinez@ufrontera.cl; Tel.: +56-45-2596816

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Urban heat islands, characterized by higher temperatures in cities compared to surrounding areas, have been studied using various techniques. However, during heat waves, existing models often underestimate the intensity of these heat islands compared to empirical measurements. To address this, an hourly time-series-based model for predicting heat island intensity during heat wave conditions is proposed. The model was developed and validated using empirical data from the National Monitoring Network in Temuco, Chile. Results indicate a strong correlation ($r > 0.98$) between the model's predictions and actual monitoring data. Additionally, the study emphasizes the importance of considering the unique microclimatic characteristics and built environment of each city when modelling urban heat islands. Factors such as urban morphology, land cover, and anthropogenic heat emissions interact in complex ways, necessitating tailored modelling approaches for the accurate representation of heat island phenomena.

Keywords: urban heat islands; heat waves; prediction model

Citation: Martinez-Soto, A.; Fürle, J.; Zipf, A. Urban Heat Island Intensity Prediction in the Context of Heat Waves: An Evaluation of Model Performance. *Eng. Proc.* **2023**, *39*, 80. <https://doi.org/10.3390/engproc2023039080>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 12 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

An urban heat island (UHI) is defined as the temperature difference observed between urban areas and the surrounding rural regions [1]. UHIs can occur in any season of the year and any time of day [2]. However, their effects are more noticeable during periods of temperature increase (e.g., heatwaves in summer). The increase in global temperatures due to global warming intensifies the effect of urban heat islands [3]. As the ambient temperature rises, urban areas experience even higher temperatures [4].

High temperatures in heat islands lead to the need for air conditioning and cooling in buildings, which increases energy consumption and results in higher greenhouse gas emissions, further contributing to global warming [5,6]. Furthermore, high temperatures can have adverse health effects on individuals, such as heat strokes, dehydration, and respiratory problems. For example, during the summer of 2003 in Europe, more than 70,000 additional deaths were attributed to heat waves [7]. Elderly individuals and households without access to air conditioning systems are identified as the first at-risk group. However, this risk level increases in urban heat islands, making the identification of these areas crucial for the development of mitigation measures (e.g., incorporating green spaces or planning open spaces that promote air circulation and shade), as well as for the protection of people.

Various techniques are employed to map urban heat islands in cities [8–10]. These include satellite remote sensing for large-scale temperature assessment, ground-based sensors and weather stations for real-time and precise data collection, aerial thermography using infrared cameras mounted on aircraft or drones to obtain detailed thermal images,

on-site temperature measurements using portable thermometers or thermographic devices, and simulation models that incorporate urban geometry, land use, vegetation, and solar radiation to predict and map heat islands [11–19]. A combination of these techniques and data sources is crucial to gain a comprehensive understanding of heat islands, enabling informed decision-making in urban planning, mitigation strategies, and the informed safeguarding of residents' health [13,20].

This study presents a combined technique for locating heat islands in the city of Temuco, Chile, as a case study. Using data from 23 monitoring stations and utilizing QGIS, areas with higher temperatures were mapped. Subsequently, a methodology for predicting heat islands was proposed for days when the external temperature exceeded 30 degrees Celsius for 3 consecutive days (heatwaves). The results are validated by comparing modeled values for specific heat island sectors in the city with actual measurements taken during heatwave days in the summer of 2019. Due to the accuracy of the results ($r > 0.98$), it is concluded that it is possible to predict the location of heat islands during heatwave events using the proposed methodology.

2. Methods

2.1. Case Study

Temuco is located in a valley surrounded by hills and mountains. The city sits at an altitude of approximately 350 m above sea level and is crossed by the Cautín River. The predominant vegetation in the area is the temperate rainforest, characteristic of the southern zone of Chile. Temuco is a relatively large city with a population of around 300,000 inhabitants. It is an urban center that is constantly growing and developing. The city is an important commercial, educational, and cultural hub in the region, offering a wide range of services and activities.

2.2. Measurement of the Temperature and Mapping of the Heat Island

To measure the temperature in different sectors of the city, monitoring stations belonging to the National Monitoring Network (ReNaM) of Chile were used. The network in Temuco consists of 23 weather stations (from Netatmo) represented by black dots in Figure 1, which are installed in private properties across various zones in Temuco.

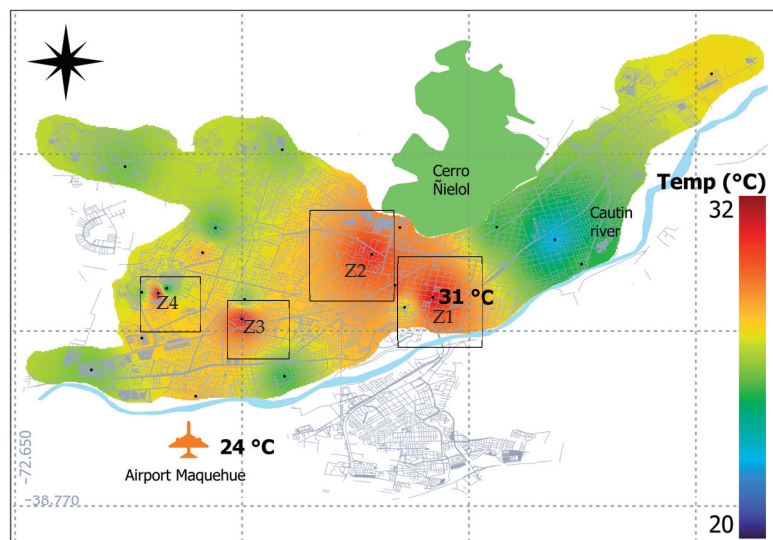


Figure 1. Mapping of the UHI phenomenon in Temuco using fixed station methodology (left) and IDW interpolation in QGIS. Temperatures from 4 December 2019, at 2 pm, are taken into account.

The Netatmo weather stations comprise two devices (indoor and outdoor) made of UV-resistant aluminum, capable of measuring temperatures ranging from $-40\text{ }^{\circ}\text{C}$ to $65\text{ }^{\circ}\text{C}$ with an accuracy of $\pm 0.3\text{ }^{\circ}\text{C}$. The outdoor sensors are shielded from rain and direct sunlight to prevent deterioration and to ensure better data accuracy. Data are captured at thirty-minute intervals, following a fixed schedule to maintain consistency (e.g., 8:00–8:30–9:00, etc.). The sensors underwent calibration and validation by the Ministry of Housing and Urban Development (MINVU) in collaboration with the Chile Foundation.

For the geolocation of heat islands, individual values from each station were used at the same hour (e.g., 2 pm). Subsequently, the temperature values along with the station coordinates were inputted into the QGIS software. The IDW interpolation technique was employed in QGIS to map the heat islands into a continuous space within the city (Figure 1). In Figure 1, the spatial distribution of temperatures measured in the city on 4 December 2019, at 2 pm, is shown. Here, there are four zones in the city (Z1–Z4) where the temperature is higher than the measurement taken in the outer part of the city (Maquehue Airport). In Z1 and Z2 the central part of the city (characterized by higher building density and low vegetation), it is observed that the temperature ($31\text{ }^{\circ}\text{C}$) is $7\text{ }^{\circ}\text{C}$ higher than at the Maquehue Airport station ($24\text{ }^{\circ}\text{C}$).

2.3. Prediction of Heat Islands during Heatwaves

To predict the intensity of heat islands in the city of Temuco, temperature differences were observed/analyzed between the weather station at Maquehue Airport and the stations situated in the four zones within the city that recorded the highest temperatures. The study specifically focused on visualizing the characteristics of heat islands during two heatwave episodes (referred to as HW1 and HW2) that occurred during the summer of 2020. By recording the temperature differences, Equation (1) has been formulated to describe the temperature in the 4 zones in the city (with the highest temperatures Z1–Z4) as a function of the temperature at Maquehue (TR = reference temperature) for each hour. Subsequently, these equations have been used to generate a general 24 h model that predicts the temperature in the four sectors of the city based on the temperature recorded at the Maquehue Airport station.

$$T_{i,z}(TR_i) = a_{i,z}TR_i + b_{i,z} \quad (1)$$

where:

T = temperature;

TR = reference temperature (measured at the Maquehue airport);

i = time of a day in hours (1, 2 . . . 24);

z = zone in Temuco where heat islands are identified (1, 2, 3, 4).

3. Results

3.1. Measurement and Recording of Temperatures during Heatwave Conditions

In Figure 2, a temperature comparison appears for two heat wave episodes (HW1 and HW2) that occurred in February 2020 between the Maquehue station (blue and red line respectively) and the zones with the highest temperature in Temuco (Z-1 to Z-4 in Figure 1). Here, it is observed first that the maximum temperature in Temuco occurs between 3 pm and 5 pm, which represents a behavior that does not follow the common pattern of the UHI phenomenon. The maximum temperature peak in most of the cases studied was reached between 7 pm and 8 pm. However, in the case of Temuco at that time, the temperature in the city dropped, whereas outside the city (reference temperature in Maquehue airport) the temperature reached its daily maximum.

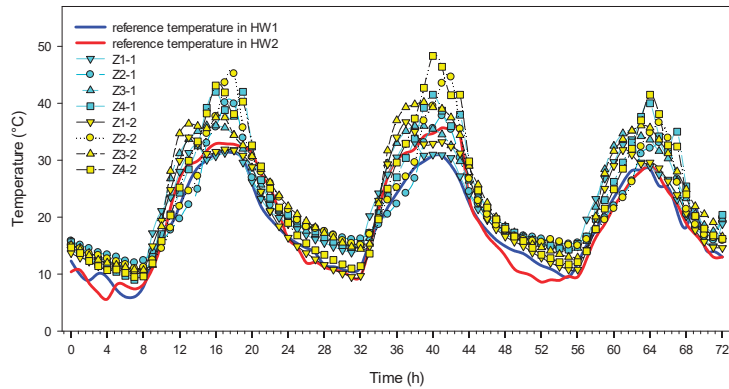


Figure 2. Comparison of temperature for two heat wave episodes between Maquehue station ((blue and red line respectively)) and the zones with the highest temperature in Temuco (Z1, Z2, Z3, Z4). Date for heat wave 1 (HW1): 8–10 February 2020. Date for heat wave 2 (HW2): 20–22 February 2020.

Additionally, the maximum temperatures recorded in the city had an average increase of 26 °C in a short period of time (between 10 am and 3 pm). The same speed of temperature change appeared from 4 pm where it reached on average 5.3 °C per hour. This is higher than the maximum acceptable temperature change rate (3 °C/h), which is set to prevent the human body from suddenly feeling hot or cold. This suggests the need to conduct more detailed studies on the behavior of the phenomenon in Temuco. Analyzing from the materiality of constructions, green areas, etc., several factors may explain more precisely the behavior of the UHI phenomenon in Temuco.

3.2. Prediction of Temperatures during Heatwave Conditions

Using Equation (1), the temperature profiles of the zones with the highest temperatures in the city (Z1, Z2, Z3, and Z4) have been predicted based on the temperature recorded at Maquehue (HW3) during a new heatwave that occurred on March 2020. The predicted temperatures were compared with the actually recorded temperatures during the third heatwave that took place between March 1st and March 3rd in order to validate the model and verify the accuracy of the results (Figure 3). Figure 3 shows a comparison between the temperature modeling for Zone 1 using temperatures from the reference station and the actual measured temperatures.

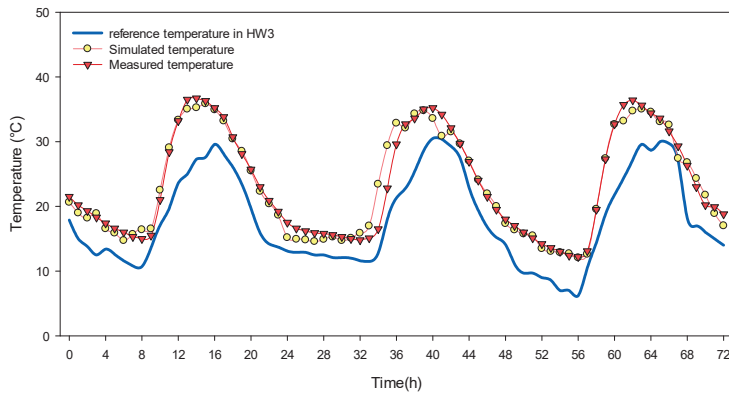


Figure 3. Comparison of real and modeled temperature profiles for zone 1 in Temuco based on the temperatures of the Maquehue station (HW3) located outside the city. Date: 1–3 March 2020.

Figure 3 shows that the values obtained from the modeling are very close to the real temperature profiles that occurred during that same heat wave event. Here, it was determined that the average real and modeled temperature differences in the 72-h period reached 1 °C and the correlation coefficient was 0.98. This strong correlation suggests it is possible to determine the temperature in the different zones of the city from the temperature of the weather station located outside the city (Maquehue). This would also imply that it is possible to know the past profiles and make a prediction for future heat wave events based on climate change scenarios. Nevertheless, additional studies are needed to verify these hypotheses and, in that sense, they represent a continuation of the work presented here.

In this study, different equations were developed to predict temperatures in each zone of the city and for each hour. Since each equation is specific to its respective zone and cannot be transferred to other zones, it follows that the modeling of heat islands cannot be generalized into a single predictive model for heat islands. Instead, each zone must be studied within its unique microclimate.

The study strongly emphasizes the critical importance of considering the distinct microclimatic characteristics and built environment of individual cities when modeling urban heat islands. The complex interplay among various factors, such as urban morphology, land cover properties, and anthropogenic heat emissions, requires the adoption of customized modeling approaches to accurately represent the phenomenon of heat islands.

By acknowledging these factors and employing tailored modeling techniques, a more precise representation of heat island phenomena can be achieved. This highlights the need for site-specific analyses and modeling in order to understand and mitigate the impact of heat islands in urban environments.

4. Conclusions

The results of this study demonstrate the close alignment between the modeled and actual temperature profiles during the heatwave event, highlighting their significance in predicting heat islands. Notably, the average temperature differences between the real and modeled data over the 72 h period were only 1 °C, with a high correlation coefficient of 0.98. This strong correlation suggests the feasibility of estimating temperatures in various city zones based on measurements from the weather station located outside the city (Maquehue). The implications of accurate heat island predictions are substantial, as they directly impact public health and energy consumption. Furthermore, these findings open the possibility of retrospectively analyzing past temperature profiles and forecasting future heatwave events under different climate change scenarios. However, further studies are required to validate these hypotheses, representing a natural continuation of the research presented here.

Author Contributions: Conceptualization and methodology, A.M.-S.; software, A.M.-S.; formal analysis, A.M.-S. and J.F.; investigation, A.M.-S.; resources, A.M.-S.; writing—original draft preparation, A.M.-S. and J.F.; writing—review and editing, A.M.-S. and J.F.; supervision, A.Z.; funding acquisition, A.M.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in FigShare at <https://doi.org/10.6084/m9.figshare.20186816.v1>.

Conflicts of Interest: The authors declare no conflict of interest.

References

- García-Cueto, O.R.; Jáuregui-Ostos, E.; Toudert, D.; Tejada-Martinez, A. Detection of the Urban Heat Island in Mexicali, BC, México and Its Relationship with Land Use. *Atmosfera* **2007**, *20*, 111–131.
- Kuznetsova, I.N.; Brusova, N.E.; Nakhaev, M.I. Moscow Urban Heat Island: Detection, Boundaries, and Variability. *Russ. Meteorol. Hydrol.* **2017**, *42*, 305–313. [CrossRef]
- Feinberg, A. Urban Heat Island Amplification Estimates on Global Warming Using an Albedo Model. *SN Appl. Sci.* **2020**, *2*, 2178. [CrossRef]
- Liu, Z.; Zhan, W.; Bechtel, B.; Voogt, J.; Lai, J.; Chakraborty, T.; Wang, Z.H.; Li, M.; Huang, F.; Lee, X. Surface Warming in Global Cities Is Substantially More Rapid than in Rural Background Areas. *Commun. Earth Environ.* **2022**, *3*, 219. [CrossRef]
- Narumi, D.; Levinson, R.; Shimoda, Y. Effect of Urban Heat Island and Global Warming Countermeasures on Heat Release and Carbon Dioxide Emissions from a Detached House. *Atmosphere* **2021**, *12*, 572. [CrossRef]
- Santamouris, M.; Cartalis, C.; Synnefa, A.; Kolokotsa, D. On the Impact of Urban Heat Island and Global Warming on the Power Demand and Electricity Consumption of Buildings—A Review. *Energy Build.* **2015**, *98*, 119–124. [CrossRef]
- Robine, J.M.; Cheung, S.L.K.; Le Roy, S.; Van Oyen, H.; Griffiths, C.; Michel, J.P.; Herrmann, F.R. Death Toll Exceeded 70,000 in Europe during the Summer of 2003. *Comptes Rendus-Biol.* **2008**, *331*, 171–178. [CrossRef] [PubMed]
- Hashim, N.M.; Ahmad, A.; Abdullah, M. Mapping Urban Heat Island Phenomenon: Remote Sensing Approach. *J.-Inst. Eng.* **2007**, *68*, 25–30.
- Elmarakby, E.; Khalifa, M.; Elshater, A.; Afifi, S. Tailored Methods for Mapping Urban Heat Islands in Greater Cairo Region. *Ain Shams Eng. J.* **2022**, *13*, 101545. [CrossRef]
- Abrar, R.; Sarkar, S.K.; Nishtha, K.T.; Talukdar, S.; Shahfahad; Rahman, A.; Islam, A.R.M.T.; Mosavi, A. Assessing the Spatial Mapping of Heat Vulnerability under Urban Heat Island (UHI) Effect in the Dhaka Metropolitan Area. *Sustainability* **2022**, *14*, 4945. [CrossRef]
- Kopecká, M.; Szatmári, D.; Holec, J.; Feranec, J. Urban Heat Island Modelling Based on MUKLIMO: Examples from Slovakia. *AGILE GIScience Ser.* **2021**, *2*, 5. [CrossRef]
- Hafner, J.; Kidder, S.Q. Urban Heat Island Modeling in Conjunction with Satellite-Derived Surface/Soil Parameters. *J. Appl. Meteorol.* **1999**, *38*, 448–465. [CrossRef]
- Voelkel, J.; Shandas, V. Towards Systematic Prediction of Urban Heat Islands: Grounding Measurements, Assessing Modeling Techniques. *Climate* **2017**, *5*, 41. [CrossRef]
- Wang, K.; Aktas, Y.D.; Stocker, J.; Carruthers, D.; Hunt, J.; Malki-Epshtein, L. Urban Heat Island Modelling of a Tropical City: Case of Kuala Lumpur. *Geosci. Lett.* **2019**, *6*, 4. [CrossRef]
- Dorigon, L.P.; Amorim, M.C.d.C.T. Spatial Modeling of an Urban Brazilian Heat Island in a Tropical Continental Climate. *Urban Clim.* **2019**, *28*, 100461. [CrossRef]
- Xu, M.; Bruelisauer, M.; Berger, M. Development of a New Urban Heat Island Modeling Tool: Kent Vale Case Study. *Procedia Comput. Sci.* **2017**, *108*, 225–234. [CrossRef]
- Kubilay, A.; Allegrini, J.; Strebel, D.; Zhao, Y.; Derome, D.; Carmeliet, J. Advancement in Urban Climate Modelling at Local Scale: Urban Heat Island Mitigation and Building Cooling Demand. *Atmosphere* **2020**, *11*, 1313. [CrossRef]
- Garzón, J.; Molina, I.; Velasco, J.; Calabia, A. A Remote Sensing Approach for Surface Urban Heat Island Modeling in a Tropical Colombian City Using Regression Analysis and Machine Learning Algorithms. *Remote Sens.* **2021**, *13*, 4256. [CrossRef]
- Khan, A.; Chatterjee, S.; Wang, Y. *Urban Heat Island Modeling for Tropical Climates*; Elsevier: Amsterdam, The Netherlands, 2020.
- Kim, S.W.; Brown, R.D. Urban Heat Island (UHI) Intensity and Magnitude Estimations: A Systematic Literature Review. *Sci. Total Environ.* **2021**, *779*, 146389. [CrossRef] [PubMed]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Foreign Exchange Forecasting Models: ARIMA and LSTM Comparison [†]

Fernando García ¹, Francisco Guijarro ¹, Javier Oliver ^{1,*} and Rima Tamošiūnienė ²

¹ Department of Economics and Social Sciences, Universitat Politècnica de València, 46022 Valencia, Spain; fergarga@esp.upv.es (F.G.); fraguima@esp.upv.es (F.G.)

² Department of Financial Engineering, Vilnius Gediminas Technical University, 10223 Vilnius, Lithuania; rima.tamosiuniene@vilniustech.lt

* Correspondence: jaolmun@ade.upv.es

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: The prediction of currency prices is important for investors with foreign currency assets, both for speculation and for hedging the exchange rate risk. Classical time series models such as ARIMA models were relevant until the advent of neural networks. In particular, recurrent neural networks such as long short-term memory (LSTM) are shown to be a good alternative model for the prediction of short-term stock prices. In this paper, we present a comparison between the ARIMA model and LSTM neural network. A hybrid model that combines the two models is also presented. In addition, the effectiveness of this model on Bitcoin's future contract is analysed.

Keywords: ARIMA; LSTM; foreign exchange prediction

1. Introduction

The foreign exchange market moves, on average, more than USD 6 million per day. It is a fundamental market for international transactions of services and goods. Hence, it is important to be able to have efficient models for predicting currency prices, as well as being able to determine their evolution. With this information, different economic agents and companies can establish their foreign currency risk levels and hedging strategies. On the other hand, it is also important for investors and speculators to know and understand the evolution of prices, and it is therefore necessary to improve predictions by reducing prediction errors as much as possible. There is a large number of currencies whose price is traded against the US dollar (USD). Those that are the most traded are considered majors. Others are considered exotic, as they are traded to a lesser extent, even though they are priced against the USD.

Since currencies can be considered time series, it is possible to apply different time series forecasting models such as the classical ARIMA model. There is a lot of applied literature on forecasting using these models for different currency pairs. For example, the author of [1] proposes a model for the prediction of the USD/EUR exchange rate taking into account the purchasing power parity theory. This theory is fundamentally based on the non-existence of arbitrage prices. It takes into account the price level differential between two countries. This model has also been applied to exotic currencies. For example, the author of [2] uses the Box–Jenkins methodology to apply an AR(1) model to predict the NGN to USD exchange rate for the period 1982–2011. On the other hand, ARIMA models are static once estimated [3]. It is necessary to create a dynamic model for long-term as well as short-term price forecasting. The main conclusion derived from this study is that the ARIMA model for short-term forecasting is more effective than for it is for long-term forecasting. Another example of exotic currency forecasting can be found in [4]. In this case, ARIMA is applied to the USD and PR currency pair with daily prices between April

Citation: García, F.; Guijarro, F.; Oliver, J.; Tamošiūnienė, R. Foreign Exchange Forecasting Models: ARIMA and LSTM Comparison. *Eng. Proc.* **2023**, *39*, 81. <https://doi.org/10.3390/engproc2023039081>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 12 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

2014 and May 2019. It highlights the importance of the stationarity of the series by taking first differences. The results obtained indicate a robust model with a difference between estimated prices and actual values of less than 1%. However, some authors have not found significant advantages of using the ARIMA model for the prediction of splits. For example, ref. [5] estimate a USD/EUR model for volatility prediction and not for price prediction. This may lead us to think that the ARIMA model may offer interesting results for the prediction of prices and their evolution, but not for determining their volatilities.

Other studies apply different neural networks for price prediction. For example, the authors of [6] compare the ARIMA model to a Backpropagation neural network for the daily prediction of the NGN/CNY and NGN/USD currency pairs. In this work, it is concluded that the neural network improves the results with a lower prediction error. Other authors incorporate additional variables in the neural network. For example, the authors of [7] add different moving averages as inputs to the network for currency prediction. As a result, the neural network performs better for three different error measures. There is no consensus on the amount and type of inputs to incorporate in different neural networks. This leads to a diversity of results when comparing these models with those of other prediction models [8].

However, recurrent neural networks of the LSTM type generally perform well, even with a single price lag as input. For example, the authors of [9] compare this network with others such as Elman's, concluding on the superiority of LSTM for short-term predictions. It even improves against other econometric models such as the VaR model or support vector machine, as described in [10], where they predict the USD/INR currency exchange with 97.83% accuracy.

In [11], a review of the literature on foreign exchange modelling and forecasting is carried out, in which it is concluded that LSTM networks are one of the best solutions for short-term forecasting. However, there is the possibility of building hybrid models in which the combination of two or more methodologies is intended to improve the results over those of the best single model. Thus, hybrid models allow us to increase the accuracy of predictions by reducing the risk of the inadequate use of a single model [12]. The results obtained by combining a model with at least one neural network are promising [13].

Some authors have carried out important reviews of hybrid models. A growing interest in this type of model has been detected, highlighting the hybridisation with neural networks and ARIMA models. These hybrid models can be combined at the same time or sequentially, and may have benefits in terms of predictive power [14]. For example, the authors of [15] analyse the USD/ALL exchange rate with monthly data from 2000 to 2015. They compared an ARIMA model to a hybrid ARIMA-ANN model sequentially. To carry this out, they initially estimated the ARIMA model using the residuals as inputs for the neural network. They used different performance indicators such as RMSE, MAE, and MAPE. In all of them, the improvement in the prediction of the hybrid model was evident. Therefore, the combination of linear and non-linear models is effective. Based on the same idea, the authors of [16] propose a hybrid multiplicative model for price forecasting in which the prediction of the non-linear components of the data series (obtained through the neural network) are multiplied by the predictions of the linear components obtained through the ARIMA model. This multiplicative model seems to work well except for some short-term forecasts.

This paper compares the ARIMA model with the LSTM recurrent neural network, as well as a hybrid ARIMA-LSTM model. For this purpose, these models are applied to the daily closing price prediction of the currencies AUD/USD, GBP/USD, JPY/USD, NZD/USD, and EUR/USD as well as the cryptocurrency Bitcoin (BTC/USD).

The following section sets out the methodology of the different models applied for the prediction of the selected currencies. Next, the main comparative results between the different models are presented. Finally, the main conclusions and limitations of this work are presented.

2. Methods

This section summarises the two main models used in forecasting the closing prices of the selected currencies. First, the classical ARIMA model, widely used in time series forecasting, is presented. Secondly, the long short-term memory (LSTM) neural network is described, which is a type of recurrent neural network that is very efficient in short-term forecasting. Finally, a hybrid model, ARIMA-LSTM, is established, in which the predictions obtained using the ARIMA model are added as inputs to the network.

2.1. ARIMA Model

The general autoregressive integrated moving average (ARIMA) model introduced by the authors of [17] includes auto-regressive as well as moving average parameters and explicitly includes differencing in the formulation of the model. Specifically, the three types of parameters in the model are the autoregressive parameter (p), the number of differencing passes (d), and moving average parameters (q). In the notation introduced by Box and Jenkins, the models are summarised as ARIMA (p,d,q).

With the ARIMA model, although a non-stationary process exhibiting homogeneity with respect to the class of series can occur in many ways, they could have a non-constant mean at time-varying second moments such as that of constant variance, or have both of these properties.

Models useful for representing such behaviours can be obtained by supposing a suitable difference in the process in order for the series to be stationary.

Having a time series, X_t , where t represents the time index, the ARMA(p,q) model is expressed as follows:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_p X_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (1)$$

where α and θ are estimated coefficients and ε is the residual of forecasts. As can be seen in Equation (1), this model is constructed as a combination of the autoregressive process (AR) with past values of the variable and the moving average (MA) process with past predictions errors. On the other hand, the parameters p and q represent the number of lags selected, while the parameter d indicates the number of integrations (usually differencing or the application of logarithms) on the variable to make the series stationary.

2.2. LSTM Model

The LSTM model is a kind of recurrent neural network, that can capture the nonlinear and complex relationships between variables. This network was proposed by the authors of [18] and has been found to be effective at capturing long-term dependencies in sequential data and time series data.

Deep learning-based models, such as LSTM, have shown promising results in time series forecasting. This network can propagate activations to process different sequences including long distance dependencies [19]. To solve the vanishing problem, the recurrent unit is grouped into blocks with cells and three gates. These gates control the flow of information [20]. The LSTM architecture consists of a memory cell and the three gates (Figure 1). Each cell presents a different state (m_t) as information flows through each neuron. The different gates are activated depending on the previous state of the cell (m_t), the output from the previous neuron (h_{t-1}) and the new information input (x_t). Thus, the forget gate is activated to decide, on basis of the inputs, which part of the information to forget from the internal state of the cell. This gate is therefore used to remove or not remove a neuron. On the other hand, the input gate or relevant gate determines how much information from the past is incorporated into the neuron, i.e., how much information is memorised. Finally, the output gate calculates the output information of the cell taking into account the previous state of the cell and the new information.

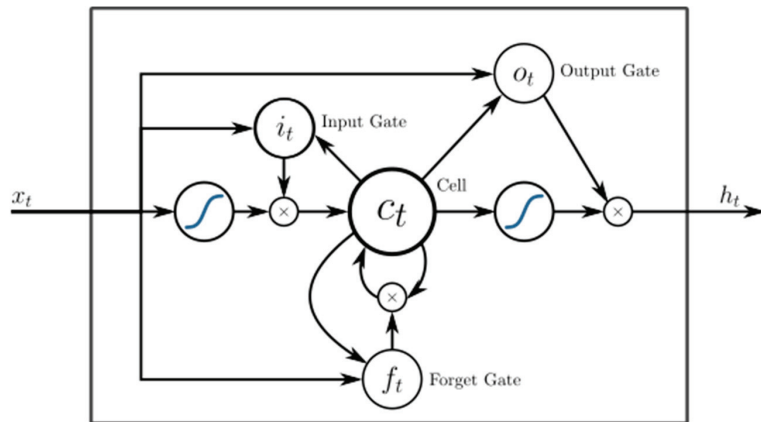


Figure 1. Internal cell structure of an LSTM. Source: “Creative Commons” by Eddie Antonio Santos licensed under BY CC-SA 4.0.

2.3. ARIMA-LSTM Model

As already indicated in the introduction, hybrid models generally show better predictions than models considered individually do. In this case, a two-stage hybrid ARIMA-LSTM model is considered. In the first stage, the ARIMA model is estimated using the Box–Jenkins methodology [17], taking into account the necessary transformations to obtain stationary series. Some authors use the residuals of the estimated model as the only input to the LSTM neural network in order to predict the non-linear patterns of the series. In this way, to obtain the final prediction, they use an additive or multiplicative model, combining in each case the prediction obtained via the ARIMA model (linear) with the non-linear patterns estimated using the LSTM network [21].

In this case, the use of the LSTM neural network as the prediction model is recommended. The first stage is identical to the process described above. That is, the ARIMA model is estimated. Then, in stage 2, two kinds of inputs are incorporated into the neural network. On the one hand, we add the lag closing prices of the time series. On the other hand, we add not the residuals, but the prediction of the prices achieved using the ARIMA model.

3. Results

This section compares forecasts for different currencies using ARIMA, LSTM and ARIMA-LSTM models. To make this comparison, different measures of prediction error are analysed, such as mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean square error (RMSE).

3.1. Database

The following day closing prices have been selected for both majors and exotic currencies: EUR/USD, GBP/USD, JPY/USD, AUD/USD, and NZD/USD. The daily closing price of the Bitcoin cryptocurrency futures contract has also been selected to determine the behaviour of the models in this kind of asset. The database runs from 18 December 2017 to 27 January 2023.

3.2. Data Analysis and Processing

Currencies are time series with a regular presence of kurtosis and skewness. Table 1 shows the main descriptive statistics for EUR/USD as an example.

Table 1. Descriptive statistics for EUR/USD.

	Mean	Median	Sd	Max	Min	Skew.	Kurt.
EUR/USD	1.14	1.13	0.06	1.25	1.25	−0.54	0.02

On the other hand, for the use of ARIMA models, it is required that the series are stationary. The augmented Dickey–Fuller (ADF) test was used to calculate each currencies. An example of the ADF test on EUR/USD is shown in Table 2. As one can see, the p -value is greater than 0.05. Therefore, the null hypothesis of the stationarity of the series was rejected. In all cases, the stationarity hypothesis was rejected.

Table 2. Augmented Dickey–Fuller test for EUR/USD.

	Dickey–Fuller	p -Value
EUR/USD (original series)	−1.5054	0.07877
EUR/USD (log-diff series)	−11.83	0.01

These results imply the need for a transformation of the original series to make them stationary. For this purpose, the logarithm over a difference has been applied, obtaining stationary series (Table 2).

3.3. Model Estimation and Results

Once the stationary series were obtained, the different models described in Section 2 were estimated. The daily estimation and forecasting process for each currency was carried out by starting from the initial observations of the indicated database and adding each day after the forecast to estimate the new model. Different measures were used to compare the prediction error of the three models. Table 3 shows the main results.

Table 3. Measures of model prediction error.

Model	BTC	AUD/USD	GBP/USD	JPY/USD	NZD/USD	EUR/USD
ARIMA						
MAE	665.52	0.00616	0.00295	0.50610	0.00642	0.00394
MAPE	0.03212	0.00387	0.00337	0.00390	0.00377	0.00350
RMSE	1160.09	0.00831	0.00402	0.71309	0.00840	0.00513
LSTM						
MAE	28.81	0.00073	0.00010	0.18372	0.00059	0.00148
MAPE	0.00100	0.00047	0.00011	0.00136	0.00035	0.00137
RMSE	28.87	0.00075	0.00010	0.18429	0.00076	0.00148
ARIMA-LSTM						
MAE	23.57	0.00049	0.00018	0.17750	0.00078	0.00144
MAPE	0.00082	0.00032	0.00022	0.00131	0.00047	0.00133
RMSE	23.57	0.00049	0.00018	0.17750	0.00078	0.00144

Firstly, the advantage of the use of neural networks over the econometric ARIMA model is evident. In all the cases analysed, the LSTM neural network improves the prediction errors, reducing them by high percentages. For example, for EUR/USD, the percentage reduction in the error measures (MAE, MAPE, and RMSE) were, respectively, 62.4%, 60.9% and 71.2%. Similar results were obtained for other “Majors” currencies such as JPY/USD with percentages of 63.7%, 65.1% and 74.2%. However, these percentage reductions in the different measures of the prediction error increased for the “exotic” currencies analysed, including the BTC/USD cryptocurrencies. These percentage reductions reached levels around 90–91% for NZD/USD and between 95–97.5% for BTC/USD. This result may be

related to the higher volatility in these exotic currencies and cryptocurrencies versus that of “Majors” currencies. For example, the volatility of EUR/USD is 0.0051 while for NZD/USD it increases to 0.0085.

On the other hand, the hybrid ARIMA-LSTM model seems to have improved the results obtained using the univariate LSTM model, although these improvements were relatively small. However, this model fails for the GBP/USD and NZD/USD currencies. The reason for the model’s failure is unclear since while the GBP/USD currency is considered to be in the “Majors” category, the NZD/USD currency belongs to the “exotics” category. While GBP/USD has a volatility, measured by the deviation, of 0.004, NZD/USD has a higher volatility of 0.008. However, EUR/USD has a deviation of 0.0051 between the other two currencies. Therefore, it is also not a justification for the failure of the hybrid model. Further analysis in other time periods and an extension to other currencies would be desirable to determine the percentage of currencies where the hybrid ARIMA-LSTM model, as described in this paper, outperforms the univariate LSTM model.

4. Conclusions

In this paper, we have compared between the classical time series model (ARIMA) and the recurrent neural network LSTM. For this purpose, we have modelled and predicted the daily closing prices of different currencies, some of them considered majors and others exotic, as well as the cryptocurrency Bitcoin. The neural network was initially applied as a univariate model in which the input corresponded to a single lag of the closing price. The results suggest that this neural network is very efficient for short-term predictions, i.e., in this case, for the next period. Next, a hybrid ARIMA-LSTM model built in two phases was proposed. The first required the corresponding forecasts to be made using the ARIMA model. In the second phase, these forecasts served as inputs to the network together with a price lag. This approach differs from that of other authors who propose that the input of the neural network should only be the residuals of the ARIMA model, given that these include the non-linear patterns. Finally, either additively or multiplicatively, the non-linear prediction of the network was combined with the linear prediction of the ARIMA model price. The results obtained with the hybrid model suggest a slight improvement with respect to the univariate LSTM neural network and, of course, with respect to the ARIMA model.

One of the limitations of this work was the lack of determination of whether or not this hybrid model also improves the prediction results when the sample data are not daily as in this case, i.e., when there are data with different timeframes (5 min, 15 min, etc.). On the other hand, the hybrid model proposed was not compared with other hybrid models in which the input is the residuals of the ARIMA model prediction, so the advantage of this one over the others is unknown. These analyses and comparisons are left for future work.

Author Contributions: Conceptualisation, F.G. (Fernando García) and F.G. (Francisco Guijarro); literature review, J.O. and R.T.; writing—original draft, J.O., F.G. (Fernando García) and R.T.; writing—review, F.G. (Francisco Guijarro). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analysed in this study. This data can be found here <https://www.visualchart.es>.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ghalayini, L. Modeling and Forecasting the US Dollar/Euro Exchange Rate. *Int. J. Econ. Financ.* **2014**, *6*, 194–207. [CrossRef]
2. Nwankwo, S.C. Autorregressive Inegrated Moving Average (ARIMA) Model for Exchange Rate (Naira to Dollar). *Acad. J. Interdiscip. Stud.* **2014**, *3*, 429–433. [CrossRef]
3. Ufuk Yildiran, C.; Fettahoglu, A. Forecasting USDTRY rate by ARIMA method. *Cogent Econ. Financ.* **2017**, *5*, 1335968. [CrossRef]
4. Asadullah, M. Forecast Foreign Exchange Rate: The Case Study of PKR/USD. *Mediterr. J. Soc. Sci.* **2020**, *11*, 129–137. [CrossRef]
5. Dunis, C.; Huang, X. Forecasting and trading currency volatility: An application of recurrent neural regression model and model combination. *Liverp. Bus. Sch. Working Pap.* **2002**, *21*, 317–354. [CrossRef]
6. Mbagu, Y.V.; Olubusoye, O.E. Foreign Exchange Prediction: A Comparative Analysis of Foreign Exchange Neural Network (FOREXNN) and ARIMA Models. 2014. Available online: <https://www.researchgate.net/publication/280040546> (accessed on 5 April 2023).
7. Kamruzzaman, J.; Sarker, R.A. Comparing ANN Based Models with ARIMA for Prediction of Forex Rates. *ASOR Bull.* **2003**, *22*, 2–11.
8. Huang, W.; Lai, K.K.; Wang, S. Forecasting Foreign Exchange Rates With Artificial Neural Networks: A Review. *Int. J. Inf. Technol. Decis. Mak.* **2004**, *3*, 145–165. [CrossRef]
9. Escudero, P.; Alcocer, W.; Paredes, J. Recurrent Neural Networks and ARIMA Models for Euro/Dollar Exchange Rate Forecasting. *Appl. Sci.* **2021**, *11*, 5658. [CrossRef]
10. Kaushik, M.; Giri, A.K. Forecasting Foreign Exchange Rate: A Multivariate Comparative Analysis between Traditional Econometric, Contemporary Machine Learning & Deep Learning Techniques. *arXiv* **2002**, arXiv:2002.10247. [CrossRef]
11. Islam, M.S.; Hossain, E.; Rahman, A.; Shahadat, M.; Andersson, K. A Review on Recent Advancements in Forex Currency Prediction. *Algorithms* **2020**, *13*, 186. [CrossRef]
12. Hajirahimi, Z.; Khashei, M. Hybrid structures in time series modeling and forecasting: A review. *Eng. Appl. Artif. Intell.* **2019**, *86*, 83–106. [CrossRef]
13. Khashei, M.; Bijari, M. A new class of hybrid models for time series forecasting. *Expert Syst. Appl.* **2012**, *39*, 4344–4357. [CrossRef]
14. Zougagh, N.; Charkaoui, A.; Echchatbi, A. Artificial intelligence hybrid models for improving forecasting accuracy. *Procedia Comput. Sci.* **2021**, *184*, 817–822. [CrossRef]
15. Mucaj, R.; Sinaj, V. Exchange Rate Forecasting using ARIMA, NAR and ARIMA-ANN Hybrid Model. *J. Multidiscip. Eng. Sci. Technol.* **2017**, *4*, 8581–8586.
16. Wang, L.; Zou, H.; Li, L.; Chaudhry, S. An ARIMA-ANN Hybrid Model for Time Series Forecasting. *Syst. Res. Behav. Sci.* **2013**, *30*, 244–259. [CrossRef]
17. Box, G.E.P.; Jenkins, G.M. *Time Series Analysis: Forecasting and Control*; Holden-Day: San Francisco, CA, USA, 1970.
18. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
19. Kelleher, J. *Deep Learning*; The MIT Press: Cambridge, MA, USA, 2019; ISBN 9780262537551.
20. Altché, F.; de La Fortelle, A. An LSTM Network for Highway Trajectory Prediction. In Proceedings of the IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017; pp. 353–359. [CrossRef]
21. Zhang, G.P. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing* **2003**, *50*, 159–175. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Learning Local Patterns of Time Series for Anomaly Detection [†]

Kento Kotera ^{*}, Akihiro Yamaguchi and Ken Ueno

Corporate R&D Center, Toshiba Corporation, Kanagawa 2128582, Japan; akihiro5.yamaguchi@toshiba.co.jp (A.Y.); ken.ueno@toshiba.co.jp (K.U.)

^{*} Correspondence: kento2.kotera@toshiba.co.jp

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: The problem of anomaly detection in time series has recently received much attention, but in most practical applications, labels for normal and anomalous data are not available. Furthermore, reasons for anomalous results must often be determined. In this paper, we propose a new anomaly detection method based on the expectation–maximization algorithm, which learns the probabilistic behavior of local patterns inherent in time series in an unsupervised manner. The proposed method is simple yet enables anomaly detection with accuracy comparable with that of the conventional method. In addition, the representation of local patterns based on probabilistic models provides new insight that can be used to determine reasons for anomaly detection decisions.

Keywords: time series; anomaly detection; subsequence; visualization

1. Introduction

Time series data, including sensor data in factories, are continuously collected in a variety of areas. One of the important applications for analyzing such data is anomaly detection. However, there are two major challenges in the actual implementation of automatic time series anomaly detection systems. First, it is often difficult to obtain labeled data for anomalies, so the data must be treated as unsupervised data, assuming that the majority of the data are normal. The other challenge is that the maintainability and reliability of anomaly detection systems often require a transparent anomaly detection model and interpretability of the output. These requirements make it necessary to use simple models, but this gives rise to a trade-off between model simplicity and anomaly detection performance. Therefore, a new anomaly detection method is needed that is interpretable without compromising anomaly detection performance.

One of the promising methods for solving these problems is OCLTS (One Class Time Seires Shapelets) [1]. OCLTS applies important subsequences in time series data, called shapelets, to enable unsupervised anomaly detection and to provide the specific parts of the time series that are the reason for the anomaly detection. However, OCLTS has several difficulties. First, the anomaly score is based on complex correlations between local patterns, so there is no direct correspondence between the anomaly score and the location in the time series that is the reason for the anomaly. In addition, the shapelets learned by OCLTS tend to take the average shape of similar time series. For example, consider the pattern shown in Figure 1a,b, in which a single concave point appears in a rightward sloping waveform. The position of the concavity is different in Figure 1a,b, but when such a pattern is the learning target, shapelets tends to have an average waveform, as shown in Figure 1c, and the concavity feature becomes unclear. In this case, it is difficult to identify the basis of the anomaly from the anomalous waveform with no concavity.

In this paper, we propose a method for time series representation learning and anomaly detection based on a novel learning procedure inspired by the subsequence-based feature transformation used in OCLTS. In the proposed method, there is one-to-one correspondence

Citation: Kotera, K.; Yamaguchi, A.; Ueno, K. Learning Local Patterns of Time Series for Anomaly Detection. *Eng. Proc.* **2023**, *39*, 82. <https://doi.org/10.3390/engproc2023039082>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 12 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

between the anomaly score and the local pattern representation, and the location of the time series for the cause of the anomaly becomes clearer. By stochastically modeling the subsequences, the proposed method provides another insight into the difference between anomalous and normal patterns, which is different from traditional shapelets-based methods. Despite the simplicity of the proposed method, experiments using public datasets show that it can detect anomalies with accuracy comparable to that of OCLTS.



Figure 1. Examples of similar subsequences (a,b) and the average subsequence of these subsequences (c).

2. Related Work

One of the most successful data mining methods for time series data analysis in recent years is a method based on representations of subsequences called shapelets, which classify time series data according to differences in subsequences rather than the entire time series. The original method [2] searched for the most different subsequences among classes and performed classification based on decision trees, but in [3,4], classification performance was greatly improved by combining advanced machine learning models with feature transformations that treat shapelets as feature transformation parameters. Shapelets can be applied not only to time series classification problems, but also to time series clustering [5,6] and approximation of DTW (Dynamic Time Warping) [7]. OCLTS [1], the method most closely related to this paper, extends shapelets to the anomaly detection problem. In that method, multiple shapelets are used for approximation over the entire time series, and the time series is converted to a vector. A one-class SVM [8] model is defined with transformed vectors as input, and the shapelet shape and one-class SVM model parameters are learned simultaneously using the gradient method. The feature transformation of time series data using shapelets proposed in OCLTS is very promising because of its extensibility in various ways in time series data analysis based on subsequences. In this paper, we propose a simpler algorithm for learning local patterns based on this feature transformation.

A time series data classification method called LOGIC [9] has been studied for the probabilistic representation of local patterns in time series. In LOGIC, local patterns based on multiple subsequences inherent in a set of time series are modeled by Gaussian process regression and mapped to a feature space of dimension equal to the number of models by using the likelihood of each model. By using the mapped features as input data for various machine learning classifiers, such as random forests, time series classification can be performed with accuracy comparable to that of state-of-the-art time series classification methods. This method showed promising results for modeling the probabilistic behavior of local patterns and for potentially representing times series based on the likelihood of features. However, LOGIC targets time series classification, which is a different problem from the one addressed in this paper. In addition, this method learns and evaluates subsequences obtained by fixing the position of time series segmentation, and does not take into account the case where the position of subsequence patterns shifts. In this paper, when learning or evaluating a subsequence, the starting position of the subsequence is searched for according to the input time series data.

3. Preliminaries

3.1. Notation

3.1.1. Time Series Data

Let $T \in \mathbb{R}^{N \times Q}$ denote a set of N normal time series data of length Q . Let $T_n = [t_{n,1}, \dots, t_{n,q}]$ for the n th time series data.

3.1.2. Time Series Subsequence

Let $\tau_{n,j} = [t_{n,j}, \dots, t_{n,j+L-1}]$ of a time series T_n denote the subsequence of length L starting at position j , where $1 \leq j \leq J, J = Q - L + 1$.

3.2. LOPAS Transform

In this section, we describe a method for transforming time series data into feature vectors using local patterns. This was proposed in OCLTS as a feature transform using shapelets. Here we describe a generalized version of the transformation using any local patterns, including shapelets. Because it was not named in [1], we call it LOPAS (Local Patterns-based Similarity) transform.

Let \mathcal{M} be a model for which the similarity to subsequences can be defined. For a set \mathcal{M} containing K models M , denote the k th model by M_k . In [1], the model corresponds to a shapelet, and in our proposed method, described in Section 4, the model corresponds to a multivariate Gaussian distribution. The similarity between a model M and a subsequence τ is denoted by $\Psi(M, \tau)$. In [1], the distance between the shapelet and the subsequence is defined as the dissimilarity; in our proposed method, the log-likelihood is defined as the similarity.

In the LOPAS transform, the K models \mathcal{M} and the n th time series T_n are taken as input. A subsequence $\tau_{n,j}$ on T_n corresponds to any of \mathcal{M} for $k = 1, 2, \dots, K$, based on the similarity $\Psi(M, \tau)$. A K -dimensional vector is output as the feature.

The intuitive explanation is that each model in \mathcal{M} is assumed to represent a subsequence, and the whole time series T_n is approximated by the subsequences. In the approximation, while allowing for overlap, the positions on T_n are slid so that there are no gaps, and the model M_k and its position $j_\omega (1 \leq \omega \leq \Omega_n)$ are searched for the position that best approximates the subsequences on T_n . Here, Ω_n is the number of slides on T_n , in other words, the number of models used to approximate T_n . The model number that maximizes the similarity for the ω th slide and its position (k_ω, j_ω) are given as follows:

$$(k_\omega, j_\omega) = \begin{cases} \underset{1 \leq k \leq K, j=1}{\operatorname{argmax}} \Psi(M_k, \tau_{n,1}) & (\omega = 1) \\ \underset{1 \leq k \leq K, j_{\omega-1} < j \leq j_{\omega-1} + L}{\operatorname{argmax}} \Psi(M_k, \tau_{n,j}) & (1 < \omega < \Omega_n) \\ \underset{1 \leq k \leq K, j=J}{\operatorname{argmax}} \Psi(M_k, \tau_{n,j}) & (\omega = \Omega_n) \end{cases} \quad (1)$$

The set of (k_ω, j_ω) in Equation (1) is denoted by

$$P_n = \{(k_\omega, j_\omega)\}_{\omega=1}^{\Omega_n} \quad (2)$$

Because models are slid with no gaps, j_ω increases between $[1, L]$ as ω increases by 1 for $(k_\omega, j_\omega) \in P_n$. Note that j_ω never exceeds J . The number of slides Ω_n depends on the set of models \mathcal{M} and the time series T_n .

Once P_n is determined, a K -dimensional feature vector Z_n is calculated based on the following equation.

$$Z_{n,k} = \begin{cases} \min_{j \in P_{n,k}} \Psi(M_k, \tau_{n,j}) & \text{if } k \in \{k_\omega\}_{\omega=1}^{\Omega_n} \\ \max_{1 \leq j \leq J} \Psi(M_k, \tau_{n,j}) & \text{if } k \notin \{k_\omega\}_{\omega=1}^{\Omega_n} \end{cases} \quad (3)$$

where $P_{n,k}$ is the set of j_ω in P_n satisfying $k_\omega = k$. That is,

$$P_{n,k} = \{j_\omega | k_\omega = k, 1 \leq \omega \leq \Omega_n\} \tag{4}$$

The LOPAS transform is the above procedure for transforming time series data T_n to features Z_n .

3.3. Assignment Factor

Another way to look at the LOPAS transformation is to consider that the subsequence $\tau_{n,j}$ used in the transformation to the features $Z_{n,k}$ is assigned to the model M_k . Other subsequences are considered not to be assigned to any model. Here, we introduce an assignment factor r and define $r_{k,n,j} = 1$ if a subsequence $\tau_{n,j}$ is assigned to the k th model M_k and $r_{k,n,j} = 0$ otherwise. This is expressed mathematically as follows.

$$r_{k,n,j} = \begin{cases} 1 & \text{if } j = \underset{j^* \in P_{n,k}}{\operatorname{argmin}} \Psi(M_k, \tau_{n,j^*}), k \in \{k_\omega\}_{\omega=1}^{\Omega_n} \\ 1 & \text{if } j = \underset{1 \leq j^* \leq J}{\operatorname{argmax}} \Psi(M_k, \tau_{n,j^*}), k \notin \{k_\omega\}_{\omega=1}^{\Omega_n} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

4. Proposed Method

In this section, we describe time series local patterns for anomaly detection and propose a learning method for them. We named these patterns LOPAD (Local Patterns for Anomaly Detection). We describe the basic idea of local patterns in Section 4.1 and explain how they are learned in Section 4.2. In Section 4.3, we describe how to evaluate anomaly scores of time series data using LOPAD.

4.1. Basic Concept

The proposed method learns a set of multivariate Gaussian distributions that represent a sufficient variety of local patterns in the context of the LOPAS transformation inherent in a set T of normal time series. For diagnostics, the LOPAS transformation is applied to the time series data to be diagnosed using the set of models. The time series data are considered anomalous if any of the subsequences deviates significantly from the normal pattern, and the anomaly is detected by calculating the anomaly score based on the similarity at that time. The anomalous subsequence is output as the reason for the anomaly detection. Furthermore, because the model uses a multivariate Gaussian distribution to model the pattern of subsequences, it can provide the probabilistic pattern with a confidence interval.

Specifically, instead of the shapelets used in OCLTS for the LOPAS transformation, K multivariate Gaussian distributions of dimension L are retained. The mean and covariance matrix parameters of the k th Gaussian distribution M_k are denoted by $M_k = \mathcal{N}(\mu_k, \Sigma_k)$, where μ_k, Σ_k are the parameters of the k th Gaussian distribution. The parameters of M_k are estimated from the set of similar subsequences, and M_k is considered to have a distribution of waveforms with similar patterns. The set of similar subsequences is the set of all local patterns in the normal time series dataset T that satisfy the assignment factor $r_k = 1$ as described in Definition 4. In other words, the set of local patterns defined as

$$\mathcal{T}_k = \{\tau_{n,j} | 1 \leq n \leq N, 1 \leq j \leq J, r_{k,n,j} = 1\} \tag{6}$$

contains the samples for estimating the parameters of the model M_k . Put another way, the local patterns assignment procedure in the LOPAS transformation yields K clusters of patterns. The model M_k represents the probability distribution of the clusters.

Furthermore, the similarity between M_k and the subsequence $\tau_{n,j}$ is defined as the log-likelihood

$$\Psi(M_k, \tau_{n,j}) = \ln p(\tau_{n,j} | \mu_k, \Sigma_k) \tag{7}$$

where

$$p(\tau|\mu, \Sigma) \sim \mathcal{N}(\mu, \Sigma) \tag{8}$$

However, for optimal assignment, an appropriate \mathcal{M} must be obtained but this requires that appropriate clusters be obtained. In the next section, we describe the objective function and the learning method for optimizing these parameters.

4.2. Learning Method

The proposed method aims to obtain multiple multivariate Gaussian models that represent all local patterns inherent in normal time series data. Specifically, the models are obtained by the following two procedures:

- Assign subsequence to clusters by LOPAS transformation.
- Obtain a Gaussian model for each cluster.

These can be formulated as follows:

$$\operatorname{argmax}_{r, \mu, \Sigma} \sum_{n=1}^N \sum_{j=1}^I \sum_{k=1}^K r_{n,j,k} \ln \mathcal{N}(\tau_{n,j} | \mu_k, \Sigma_k) \tag{9}$$

This objective function can be optimized by using the expectation–maximization algorithm, which has long been used in K-means clustering and other methods. First, each parameter in the models $\mathcal{M} \mu_k, \Sigma_k (k = 1, \dots, K)$ is fixed and the assignment factor r is obtained. Second, the parameters of the multivariate Gaussian model corresponding to each cluster are updated by fixing r . By repeating these two procedures until the objective function converges, we obtain a set of multivariate Gaussian models that represent the various subsequence patterns inherent in a set of normal time series data.

Specifically, a model set \mathcal{M} consisting of K Gaussian distributions initialized with appropriate parameters is prepared. Using the procedure in Equation (4), we calculate the assignment factor r for each K and obtain the set of assigned subsequences \mathcal{T}_k . Next, using \mathcal{T}_k as input, we estimate the mean parameter μ_k and the covariance matrix parameter Σ_k of the model M_k for each k . To estimate the mean parameter and covariance matrix parameter of the Gaussian distribution, any method can be used, such as maximum likelihood estimation.

The parameters are updated using the obtained μ_k and Σ_k . As described above, \mathcal{M} representing various subsequence patterns is obtained by repeating the two steps of (1) subsequence assignment and (2) parameter updating until the termination condition is satisfied.

The above algorithm is summarized in Algorithm 1.

Algorithm 1 Algorithm of the proposed method

Require: Time series dataset $T \in \mathbb{R}^{N \times Q}$, number of models K , subsequence length parameter L

Ensure: $\mathcal{M} = \{\mathcal{N}(\mu_k, \Sigma_k)\}_{k=1}^K$

- 1: Initialize $\{\mathcal{N}(\mu_k, \Sigma_k)\}_{k=1}^K$
 - 2: **repeat**
 - 3: **for** $n = 1, \dots, N$ **do**
 - 4: LOPAS transform on T_n using \mathcal{M} to obtain the assignment factor r_n .
 - 5: **end for**
 - 6: **for** $k = 1, \dots, K$ **do**
 - 7: Update μ_k, Σ_k using a set of subsequences \mathcal{T}_k .
 - 8: **end for**
 - 9: **until** Stationarity.
-

4.3. Evaluation Method

Each model in \mathcal{M} obtained by training represents subsequence patterns inherent in normal data. Therefore, when the LOPAS transformation is applied to time series data, the likelihood of each model should be large for normal time series data. Conversely, anomalous data will result in a small likelihood for one or more of the models. Therefore, the smallest of the K -dimensional feature vectors obtained by the LOPAS transformation is adopted as the degree of normality of the time series data. In other words, if Z_{new} represents the K -dimensional features obtained by the LOPAS transformation for time series data T_{new} , the anomaly score $a(T_{new})$ is given by the following formula.

$$a(T_{new}) = - \min_k Z_{new,k} \tag{10}$$

The subsequence with the maximum anomaly can be regarded as the local pattern that most greatly differs from the normal pattern in the time series data. Let M_{k^*} denote the model that has the maximum anomaly score and τ^* denote the corresponding subsequence. Let τ_i^* be the i th value of τ^* . The possible values of τ_i^* are considered to follow a Gaussian distribution $\mathcal{N}(\mu_{i\{1,\dots,L\}\setminus i}, \sigma_{i\{1,\dots,L\}\setminus i}^2)$, conditioned by the points $\{\tau_1^*, \dots, \tau_L^*\} \setminus \tau_i^*$ other than τ_i^* in τ^* in M_{k^*} . Therefore, by obtaining the conditional distribution at all points of i and visualizing the range of

$$\mu_{i\{1,\dots,L\}\setminus i} \pm C\sigma_{i\{1,\dots,L\}\setminus i} \tag{11}$$

for each point, we can analyze the difference between the normal subsequence pattern and the anomalous waveform (where C is an arbitrary constant).

5. Experiment and Evaluation

5.1. Accuracy in the UCR Dataset

We evaluate the proposed method on some data from the UCR dataset [10]. For the experiment, the class with the largest number of data, defined originally as "Train" in the dataset, was assumed as the normal class and was used as training data. The data for the other classes were used as validation data. In the data defined originally as "Test", data with the same class as the training data were taken as the normal class, and data with other classes were taken as the anomalous class. The training data were used for training, and the anomaly score was calculated for each of the normal and anomalous data in the test data, and the area under the curve (AUC) was calculated for evaluation.

The initial parameters of \mathcal{M} were set by estimating the Gaussian mixture model with K mixed models for all subsequences of length L in the time series dataset T , and the parameters of each model were used as initial parameters. Other hyperparameters were set using validation data from among the number of models $K = \{10, 30, 50\}$ and subsequence length $L = \{0.1, 0.2, 0.3\} \times Q$.

The experimental results are shown in Table 1. The proposed method detected anomalies with the same accuracy as OCLTS. Note, however, that OCLTS uses a highly non-linear transformation based on a kernel method to calculate the anomaly, whereas the proposed method uses a simple method to calculate the anomaly.

5.2. Visualization Evaluation Using Current Data

In this experiment, we used a dataset of solenoid current measurements called NASA Shuttle Valve Data [11]. As a preprocessing step, data were sampled every 100 points from the original time series of length 20,000 to make a time series of length 200. The time series is scaled so that the minimum and maximum values are $[0, 1]$. In the dataset, the seven time series shown in Figure 2 are taken as normal data, and the models are trained with the proposed method and OCLTS, respectively. After a certain period of time with noisy steady current in the first half of the time series, the data enters a phase in which the current rises. During the rise phase, the current temporarily drops during the rise, but soon rises and enters the steady current phase, where the current remains high. The current then enters a descending phase, rising temporarily during the descending phase, but eventually

decreasing to near the initial current value. Normal data tend to differ in the timing and magnitude of the temporary drops or rises during the rise and fall phases. For the anomaly data, the time series data shown in Figure 3 are used. These data do not show a temporary drop in the current value during the rise phase. The proposed method and OCLTS were applied to these data to diagnose and visualize the anomaly.

Table 1. Comparison of AUC in UCR time series data.

	LOPAD (Proposed)	OCLTS
Plane	1.000	1.000
Trace	0.985	1.000
SonyAIBORobotSurface1	0.992	0.950
SonyAIBORobotSurface2	0.948	0.914
ECGFivedays	1.000	0.980
ECG200	0.801	0.834
ECG5000	0.932	0.984
MiddlePhalanxTW	0.994	0.991
ProximalPhalanxOutlineAgeGroup	0.899	0.883

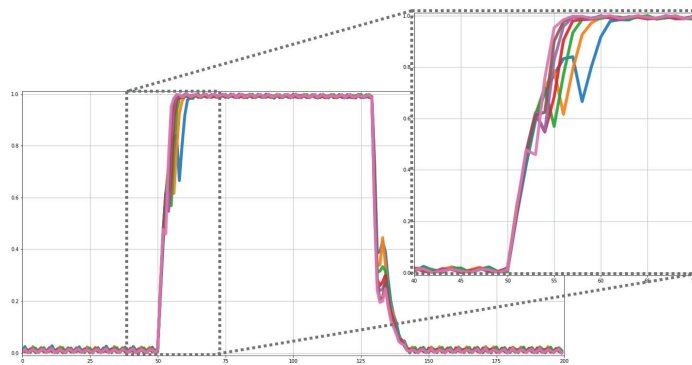


Figure 2. Seven time series of current measurements in NASA Shuttle Valve Data used as training data.

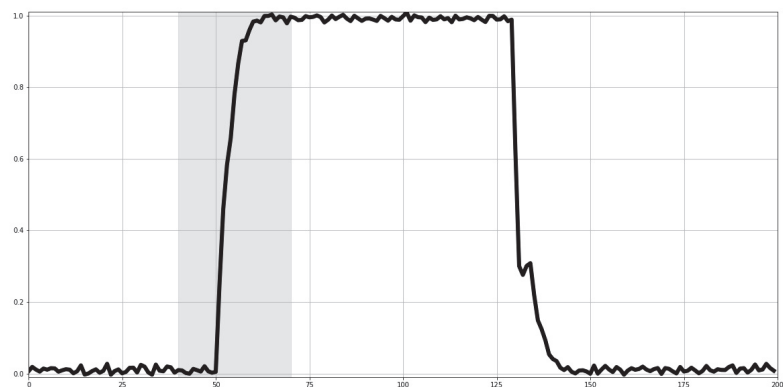


Figure 3. Anomaly data in NASA Shuttle Valve Data. The current rise phase is highlighted, which is different from the pattern of normal data.

First, the results of the OCLTS visualization are shown in Figure 4. The black line is the original time series, and the upward phase showing the anomaly pattern is enlarged. The red line shows the shapelet obtained by training in OCLTS. The learned shapelet appears to be the average pattern of the training data. In the training data, the position of the temporary dip in the ascending phase varies, and as a result of learning the average pattern, the temporary dip disappears in the learned pattern. The visualization results of the proposed method are shown in Figure 5, where the normal pattern region is obtained as $C = 2$ in Formula (11). The normal pattern region obtained using the proposed method shows a distorted region with repeated unevenness, and the original time series (black line) is found to be out of the normal region. This implies that the proposed method's normal pattern contains some kind of uneven waveform in the ascending phase. Thus, the proposed method provides insight into the behavior of patterns based on subsequences, which is difficult to achieve with conventional shapelets.

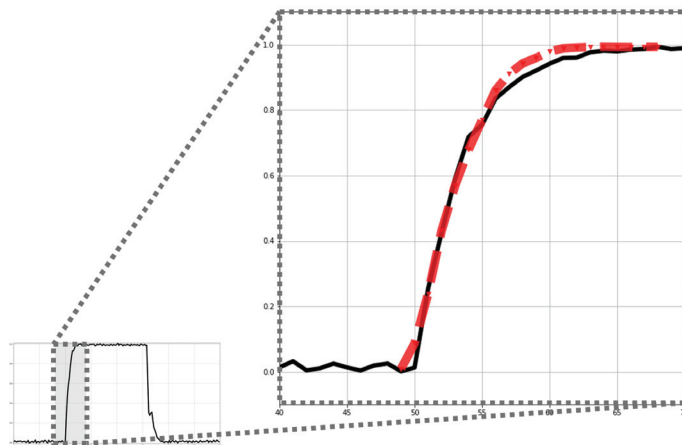


Figure 4. Results of visualization analysis using OCLTS. The black line represents the time series data and the red line represents the shapelet obtained by training.

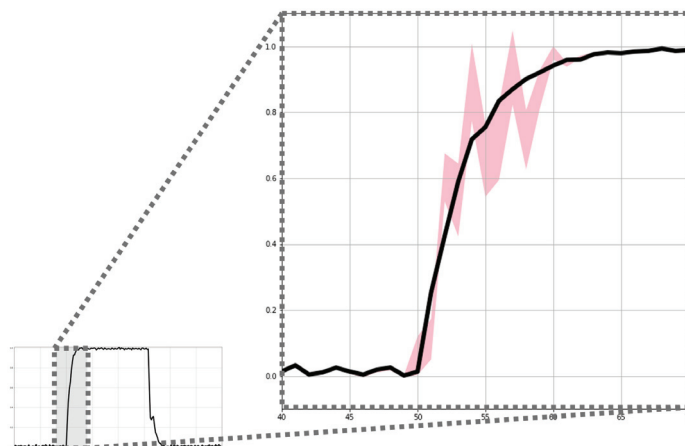


Figure 5. Results of visualization analysis using LOPAD (proposed). The black line is the time series data, and the pink area indicates the normal pattern.

5.3. Limitations of Our Method

The above results demonstrate that our proposed method is effective in anomaly detection in various domains of data. However, we should mention that there may be situations where our proposed method fails. First, our proposed method is designed to identify anomalies primarily based on the most irregular subsequence and is unable to consider correlations between subsequences. In such cases, it is necessary to employ OCLTS, which can capture complex correlations between subsequences that cannot be resolved due to its kernel method. Furthermore, both the proposed method and OCLTS generate features from time series data using the LOPAS transform, they cannot model in which the training data do not have roughly similar shapes.

6. Conclusions

In this paper, we proposed a representation learning method based on the probabilistic behavior of subsequences for anomaly detection in time series. Experiments confirmed that the proposed method has anomaly detection performance comparable to that of the conventional method OCLTS, but with a more transparent anomaly calculation procedure. Furthermore, the probabilistic modeling of subsequence patterns provides insight into the reason why anomaly detection differs from OCLTS.

Author Contributions: Conceptualization, K.K., A.Y. and K.U.; methodology, K.K.; software, K.K. and A.Y.; validation, K.K.; writing—original draft preparation, K.K.; writing—review and editing, A.Y. and K.U.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study applied open access dataset [10,11].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yamaguchi, A.; Nishikawa, T. One-class learning time-series shapelets. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 2365–2372.
2. Ye, L.; Keogh, E. Time series shapelets: A new primitive for data mining. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009; pp. 947–956.
3. Grabocka, J.; Schilling, N.; Wistuba, M.; Schmidt-Thieme, L. Learning time-series shapelets. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014; pp. 392–401.
4. Lines, J.; Davis, L.M.; Hills, J.; Bagnall, A. A shapelet transform for time series classification. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12 August 2012; pp. 289–297.
5. Zakaria, J.; Mueen, A.; Keogh, E. Clustering time series using unsupervised-shapelets. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, 10–13 December 2012; pp. 785–794.
6. Zhang, Q.; Wu, J.; Yang, H.; Tian, Y.; Zhang, C. Unsupervised feature learning from time series. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 2322–2328.
7. Lods, A.; Malinowski, S.; Tavenard, R.; Amsaleg, L. Learning DTW-preserving shapelets. In Proceedings of the Advances in Intelligent Data Analysis XVI: 16th International Symposium, IDA 2017, London, UK, 26–28 October 2017; pp. 198–209.
8. Schölkopf, B.; Platt, J.C.; Shawe-Taylor, J.; Smola, A.J.; Williamson, R.C. Estimating the support of a high-dimensional distribution. *Neural Comput.* **2001**, *13*, 1443–1471. [CrossRef] [PubMed]
9. Berns, F.; Hüwel, J.D.; Beecks, C. LOGIC: Probabilistic Machine Learning for Time Series Classification. In Proceedings of the 2021 IEEE International Conference on Data Mining (ICDM), Auckland, New Zealand, 7–10 December 2021; pp. 1000–1005.

10. Dau, H.A.; Bagnall, A.; Kamgar, K.; Yeh, C.C.M.; Zhu, Y.; Gharghabi, S.; Ratanamahatana, C.A.; Keogh, E. The UCR time series archive. *IEEE/CAA J. Autom. Sin.* **2019**, *6*, 1293–1305. [CrossRef]
11. Ferrell, B.; Santuro, S. NASA Shuttle Valve Data. Available online: <http://www.cs.fit.edu/~pkc/nasa/data/> (accessed on 1 February 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Proceeding Paper

Multi-Output Variational Gaussian Process for Daily Forecasting of Hydrological Resources [†]

Julián David Pastrana-Cortés ^{1,*}, David Augusto Cardenas-Peña ¹, Mauricio Holguín-Londoño ¹, Germán Castellanos-Dominguez ² and Álvaro Angel Orozco-Gutiérrez ¹

¹ Automatic Research Group, Universidad Tecnológica de Pereira, Pereira 660003, Colombia; dcardenas@utp.edu.co (D.A.C.-P.); mau.hol@utp.edu.co (M.H.-L.); aaog@utp.edu.co (A.A.O.-G.)

² Signal Processing and Recognition Group, Universidad Nacional de Colombia, Manizales 170003, Colombia; cgcastellanosd@unal.edu.co

* Correspondence: j.pastrana@utp.edu.co

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Water resource forecasting plays a crucial role in managing hydrological reservoirs, supporting operational decisions ranging from the economy to energy. In recent years, machine learning-based models, including sequential models such as Long Short-Term Memory (LSTM) networks, have been widely employed to address this task. Despite the significant interest in forecasting hydrological series, weather's nonlinear and stochastic nature hampers the development of accurate prediction models. This work proposes a Variational Gaussian Process-based forecasting methodology for multiple outputs, termed MOVGP, that provides a probabilistic framework to capture the prediction uncertainty. The case study focuses on the Useful Volume and the Streamflow Contributions from 23 reservoirs in Colombia. The results demonstrate that MOVGP models outperform classical LSTM and linear models in predicting several horizons, with the added advantage of offering a predictive distribution.

Keywords: streamflow contributions; predictive distribution; forecasting; Gaussian process; useful volume

Citation: Pastrana-Cortés, J.D.;

Cardenas-Peña, D.A.;

Holguín-Londoño, M.;

Castellanos-Dominguez, G.;

Orozco-Gutiérrez, Á.A. Multi-Output

Variational Gaussian Process for

Daily Forecasting of Hydrological

Resources. *Eng. Proc.* **2023**, *39*, 83.

[https://doi.org/10.3390/](https://doi.org/10.3390/engproc2023039083)

[engproc2023039083](https://doi.org/10.3390/engproc2023039083)

Academic Editor: Ignacio Rojas,

Hector Pomares, Luis Javier Herrera,

Fernando Rojas and Olga Valenzuela

Published: 12 July 2023



Copyright: © 2023 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the Creative Commons

Attribution (CC BY) license ([https://](https://creativecommons.org/licenses/by/4.0/)

[creativecommons.org/licenses/by/](https://creativecommons.org/licenses/by/4.0/)

[4.0/](https://creativecommons.org/licenses/by/4.0/)).

1. Introduction

Hydrological forecasting plays a crucial role in planning and operation activities. On a short-term scale, it allows for management of water systems and resources, including irrigation, flood control, and hydropower generation. In the energy sector, hydrological forecasting supports the optimal scheduling of hydroelectric power generation. Accurately predicting hydroelectric plants' water release and reservoir volume enables scheduling optimal thermal plant generation while minimizing fuel costs and improving the energy sector's sustainability [1,2]. For example, Brazil's National Electrical System Operator provides streamflow time series of hydropower plants that supports forecasting research [3]. In Colombia, hydropower plants contribute 97.37% of the renewable energy, including 87.54% from reservoirs [4]. Therefore, there is a great interest in developing accurate hydrological forecasting models to manage and exploit water resources sustainably and effectively.

Forecasting models can be short-, mid-, or long-term, where the former enable dispatch and optimization for power systems. Long-term forecasting supports reliability planning through system expansions and weather analysis at a large scale [5,6]. Hence, the design of forecasting models depends on the prediction horizon, with two primary approaches: physically-driven concept rules and data-driven models which learn from time-series samples. Models in the first category have demonstrated their capability to predict various flooding scenarios. However, physical modeling often requires extensive knowledge

and expertise in hydrological parameters and various datasets, demanding intensive computation which results in unsuitability for short-term prediction [7]. Since data-driven models can learn complex behavior from the data, streamflow forecasting traditionally relies on this second approach. The most extensively used data-driven forecasting model is the Linear AutoRegression (LAR), due to its simplicity and interoperability [8]. However, LAR fails to adequately represent streamflow series because of the complex water resource patterns, such as varying time dependencies, randomness, and nonlinearity [3,6,9].

The most prominent models for dealing with nonlinear trends are neural networks (NNs), known for their flexibility and outperformance of other nonlinear models [10]. For instance, a hybrid model coupling extreme gradient boosting to NNs predicted monthly streamflow at Cuntan and Hankou stations on the Yangtze River, outperforming baseline support vector machines [11]. Additionally, recurrent architectures, such as Long Short-Time Memory (LSTM), have proven to improve the scores of classical NNs in daily streamflow forecasting, given their ability to capture seasonality and stochasticity [5,8]. However, the numerous neural network architectures available make researchers question which one will best fit a given problem, as no single model is universally applicable [7]. Further, the inherent noise present in hydrological time series influences forecasting accuracy [3,10].

Some operational tasks demand uncertainty quantification for the prediction due to the inherent noise. Gaussian Processes (GP) satisfy such a requirement by approximating a predictive distribution. GP-based forecasting has proven remarkable results in streamflow forecasting up to one day and month ahead [12,13]. In addition, a kernel function, combining squared exponential, periodic, and rational quadratic terms, allowed GP models to fit streamflow time series for the Jinsha River [9]. In another approach, a probabilistic LSTM coupled with a heteroscedastic GP produced prediction intervals without any post-processing to manage the daily streamflow time series uncertainty [14]. However, GP-based approaches pose two research gaps [15]. Firstly, the probabilistic couple approach still complicates the model calibration. Secondly, natural probabilistic modes such as GP have only been used to study scalar value signals.

This work develops a forecasting methodology using a GP-based probabilistic approach applied to hydrological resources, supporting multiple output predictions and reducing the model training complexity. The methodology, termed MOVGP, combines the advantages of Multi-Output and Variational GPs for taking advantage of relationships among time series, adapting the individual variability to handle large amounts of samples. The research compares the performance of the MOVGP against an LSTM neural network and a Linear AutoRegression (LAR) model in forecasting two multi-output hydrological time series, namely, Useful Volume and Streamflow Contributions of 23 reservoirs. It is worth noting that the considered time series correspond to actual reservoir data taken into account for hydropower generation in Colombia. Attained results prove the ability of MOVGP to be adapted to varying prediction horizons, generally outperforming contrasted models.

The paper agenda is as follows: Section 2 covers methodologies and theoretical bases used for developing and training Multi-Output Variational GP models; Section 3 validates the MOVGP training and tests the three models in terms of the Mean Square Error (MSE); Final remarks and future work conclude the work in Section 4.

2. Mathematical Framework

2.1. Gaussian Process Modeling Framework

A Gaussian Process (GP) is a collection of random variables related to the infinite-dimensional setting of a joint Gaussian distribution. Consider the dataset of N samples $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$, where $\mathbf{X} \in \mathbb{R}^{L \times N}$ is the design matrix, with columns of vector inputs $\mathbf{x} \in \mathbb{R}^L$ of L features, and $\mathbf{Y} \in \mathbb{R}^{D \times N}$ is the target matrix, with columns of vector outputs $\mathbf{y} \in \mathbb{R}^D$ of D outputs for all N cases. GP framework conditions a subset of observations to create a map that models the relationship between \mathbf{X} to \mathbf{Y} .

Then, the Single-Output GP (SOGP) attempts to represent a scalar-valued function $f : \mathbb{R}^L \rightarrow \mathbb{R}$, i.e., where $D = 1$ with GP framework. This model is completely specified by a mean function $m : \mathbb{R}^L \rightarrow \mathbb{R}$ and covariance (kernel) function $k : \mathbb{R}^L \times \mathbb{R}^L \rightarrow \mathbb{R}$, with vector parameters θ_m and θ_k notated, respectively, as follows:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x} \mid \theta_m), k(\mathbf{x}, \mathbf{x}' \mid \theta_k)) \tag{1}$$

In more realistic scenarios, single-output observation y presents Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma_N^2)$ models as a noise-added version of the function f , such as $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon$. Let $\mathbf{X}_* \in \mathbb{R}^{L \times N'}$ be a test matrix with N' test vector inputs $\mathbf{x}_* \in \mathbb{R}^L$, $\mathbf{f}_* \in \mathbb{R}^{N'}$, in which $\mathbf{m} = [m(\mathbf{x}_i \mid \theta_m)] \in \mathbb{R}^N$ denotes mean train vector, $\mathbf{m}_* = [m(\mathbf{x}_{*i} \mid \theta_m)] \in \mathbb{R}^{N'}$ denotes mean test vector, $\mathbf{K} = [k(\mathbf{x}_i, \mathbf{x}_j \mid \theta_k)] \in \mathbb{R}^{N \times N}$ denotes covariance train matrix, $\mathbf{K}_* = [k(\mathbf{x}_i, \mathbf{x}_{*j} \mid \theta_k)] \in \mathbb{R}^{N \times N'}$ denotes covariance train–test matrix, and $\mathbf{K}_{**} = [k(\mathbf{x}_{*i}, \mathbf{x}_{*j} \mid \theta_k)] \in \mathbb{R}^{N' \times N'}$ denotes covariance test matrix. The joint Gaussian distribution of the observations vector \mathbf{y} and test outputs vector \mathbf{f}_* , named previously, are specified as shown:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m} \\ \mathbf{m}_* \end{bmatrix}, \begin{bmatrix} \mathbf{K} + \sigma_N^2 \mathbf{I}_N & \mathbf{K}_* \\ \mathbf{K}_*^\top & \mathbf{K}_{**} \end{bmatrix}\right) \tag{2}$$

where \mathbf{I}_N is the identity matrix of size N derived in the conditional distribution, calculated as follows:

$$\mathbf{f}_* \mid \mathbf{X}_*, \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*)) \tag{3}$$

with the following definitions:

$$\bar{\mathbf{f}}_* = \mathbf{m}_* + \mathbf{K}_*^\top [\mathbf{K} + \sigma_N^2 \mathbf{I}_N]^{-1} (\mathbf{y} - \mathbf{m}) \tag{4}$$

$$\text{cov}(\mathbf{f}_*) = \mathbf{K}_{**} - \mathbf{K}_*^\top [\mathbf{K} + \sigma_N^2 \mathbf{I}_N]^{-1} \mathbf{K}_* \tag{5}$$

Notice, from Equations (4) and (5), that mapping construction is analytic and, therefore, does not employ an optimization process. Nevertheless, selection of parameters at θ_m and θ_k and observation noise variance σ_N can be estimated using marginal likelihood from Equation (2), $p(\mathbf{y} \mid \mathbf{X}) = \mathcal{N}(\mathbf{m}, \mathbf{K} + \sigma_N^2 \mathbf{I}_N)$, and minimizing negative log marginal likelihood:

$$\min_{\theta_m, \theta_k, \sigma_N} -\ln(p(\mathbf{y} \mid \mathbf{X})) = \frac{1}{2} (\mathbf{y} - \mathbf{m})^\top \mathbf{K}_y^{-1} (\mathbf{y} - \mathbf{m}) + \frac{1}{2} \ln(|\mathbf{K}_y|) + \frac{N}{2} \ln(2\pi) \tag{6}$$

where $\mathbf{K}_y = \mathbf{K} + \sigma_N^2 \mathbf{I}_N$ is the covariance matrix for the noisy observations. Thus, the optimization problem in Equation (6) can be efficiently solved via a gradient-based optimizer [16].

2.2. Multi-Output Gaussian Process (MOGP)

MOGP generalizes SOGP mapping for $D \geq 1$ outputs as $f^D : \mathbb{R}^L \rightarrow \mathbb{R}^D$ with GP framework, where f^D is a vector-valued function. The MOGP model, similar to the SOGP model, is entirely defined by its mean vector function $m^D : \mathbb{R}^L \rightarrow \mathbb{R}^D$ and covariance matrix function $k^D : \mathbb{R}^L \times \mathbb{R}^L \rightarrow \mathbb{R}^{D \times D}$, each with vector parameters θ_m and θ_k , respectively, expressed as follows:

$$f^D(\mathbf{x}) \sim \mathcal{GP}(m^D(\mathbf{x} \mid \theta_m), k^D(\mathbf{x}, \mathbf{x}' \mid \theta_k)) \tag{7}$$

Let $\Sigma_N \in \mathbb{R}^{D \times D}$ be a diagonal matrix such that $\Sigma_N = \text{diag}\{\sigma_{N,d}^2\}_{d=1}^D$, with $\sigma_{N,d}$ being the d^{th} output observation noise variance and $\mathbf{y}^D \in \mathbb{R}^{ND}$ being a ravel version vector of target matrix \mathbf{Y} . Following the procedure established in Equation (3), deriving the MOGP posterior distribution takes place as follows:

$$\mathbf{f}_* \mid \mathbf{X}_*, \mathbf{X}, \mathbf{Y} \sim \mathcal{N}(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*)) \tag{8}$$

with the following definitions:

$$\bar{\mathbf{f}}_*^D = \mathbf{m}_*^D + \mathbf{K}_*^{D\top} [\mathbf{K}_y^D]^{-1} (\mathbf{y}^D - \mathbf{m}^D) \tag{9}$$

$$\text{cov}(\mathbf{f}_*) = \mathbf{K}_{**}^D - \mathbf{K}_*^{D\top} [\mathbf{K}_y^D]^{-1} \mathbf{K}_*^D \tag{10}$$

where $\mathbf{K}_y^D = \mathbf{K}^D + \Sigma_N \otimes \mathbf{I}_N$, \otimes represents the Kronecker product between matrices and upper index D denotes D – dimension version of SOGP quantities.

To deal with developing an admissible correlation between outputs, the Linear Model of Coregionalization takes place, expressing each output of MOGP as a linear combination of Q (known as latent dimension) independent SOGP as follows:

$$f^D(\mathbf{x}) = \sum_{q=1}^Q \mathbf{a}_q g_q(\mathbf{x}) \tag{11}$$

where $\mathbf{a}_q \in \mathbb{R}^D$ is the vector coefficients, with values a_q^d associated with contributions of the q -th independent SOGP $g_q(\mathbf{x})$ at the d^{th} output with kernel function k_q . In this way, the covariance matrix of the MOGP model is given by the following [17,18]:

$$k^D(\mathbf{x}, \mathbf{x}') = \sum_{q=1}^Q \mathbf{B}_q k_q(\mathbf{x}, \mathbf{x}') \tag{12}$$

where $\mathbf{B}_q \in \mathbb{R}^{D \times D} = \mathbf{a}_q \mathbf{a}_q^\top$ is a semi-definite positive matrix known as the coregionalization matrix.

2.3. Variational Gaussian Process (VGP)

The main challenge in implementing MOGP models lies in their complexity $\mathcal{O}(D^3 N^3)$ and storage demand $\mathcal{O}(D^2 N^2)$ becoming intractable for a dataset of a few thousand samples [19], because the need to invert the matrix \mathbf{K}_y^D in Equations (9) and (10) is usually performed by Cholesky decomposition. To overcome the problem of computational complexity, a new set of $M \ll N$ trainable inducing points $\mathbf{Z} \in \mathbb{R}^{L \times M}$ and inducing variables $\mathbf{u} = f^D(\mathbf{Z}) \in \mathbb{R}^{DM}$ augment the output variables $\mathbf{f}^D = f^D(\mathbf{X}) \in \mathbb{R}^{DN}$. The marginal distribution for the output variables is expressed as $p(\mathbf{f}^D | \mathbf{X}) = \int p(\mathbf{f}^D | \mathbf{X}, \mathbf{u}) p(\mathbf{u}) d\mathbf{u}$. The Variational Gaussian Process (VGP) allows approximating $p(\mathbf{f}^D | \mathbf{X})$ with $q(\mathbf{f}^D | \mathbf{X})$ by marginalizing out the set of inducing points [20]:

$$q(\mathbf{f}^D | \mathbf{X}) := \int p(\mathbf{f}^D | \mathbf{X}, \mathbf{u}) q(\mathbf{u}) d\mathbf{u} \tag{13}$$

Since the output distribution comes from a MOGP, $q(\mathbf{u})$ is assumed as Gaussian $\mathcal{N}(\mathbf{u} | \mathbf{m}_z, \mathbf{S})$, with mean $\mathbf{m}_z \in \mathbb{R}^{DM}$ and covariance $\mathbf{S} \in \mathbb{R}^{DM \times DM}$, so that the approximating distribution also becomes Gaussian:

$$q(\mathbf{f}^D | \mathbf{X}) = \mathcal{N}\left(\mathbf{f}^D | \mathbf{A} \mathbf{m}_z, \mathbf{K}^D + \mathbf{A}(\mathbf{S} - \mathbf{K}_{M,M})\mathbf{A}^\top\right) \tag{14}$$

with $\mathbf{A} = \mathbf{K}_M \mathbf{K}_{M,M}^{-1}$, $\mathbf{K}_M \in \mathbb{R}^{DN \times DM}$ as the kernel function in Equation (12) evaluated at all pairs of inducing–training points and $\mathbf{K}_{M,M} \in \mathbb{R}^{DM \times DM}$ the kernel function values between pairs of inducing points. Since optimizing the parameters of $q(\mathbf{u})$ yields a stochastic framework, the cost function in Equation (6) turns into a tractable marginal likelihood bound for the multi-output case:

$$\log p(\mathbf{y}^D | \mathbf{X}) \geq \sum_{n=1}^N \mathbb{E}_{q(f_n^D | \mathbf{x}_n)} [\log p(y_n^D | f^D(\mathbf{x}_n))] - \text{KL}[q(\mathbf{u}) \| p(\mathbf{u})] \tag{15}$$

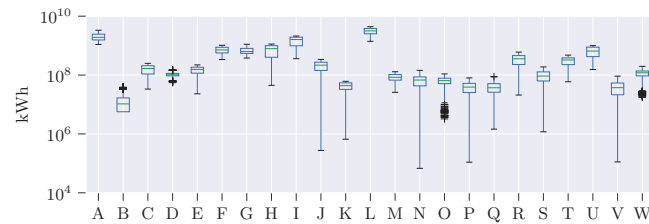
in which $KL[q(\mathbf{u}) \parallel p(\mathbf{u})]$ is the Kullback–Leibler (KL) divergence between $q(\mathbf{u})$ and $p(\mathbf{u})$. This approach offers a notable benefit by diminishing the complexity of the MOGP to $\mathcal{O}(DND^2M^2)$ due to the inversion of the matrix $\mathbf{K}_{M,M}$, which is smaller than \mathbf{K}_y^D . As a result, the model can efficiently handle an increased number of samples N , providing an opportunity to gather more information from the dataset at a reduced cost.

3. Results and Discussions

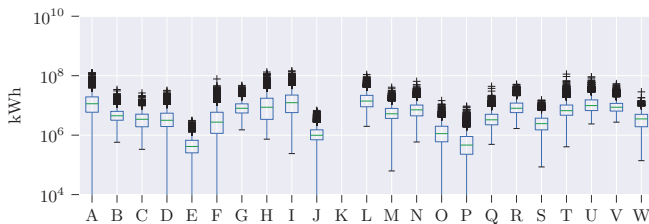
The current section aims to thoroughly communicate the implications of the forecasting results on hydric time series using MOVGP. Firstly, we offer a detailed description of the considered dataset, including information on its sources, characteristics, and preprocessing steps. Further, the manuscript details the hyperparameter tuning experiments and describes the validation strategy. Finally, we examine the forecasting performance and highlight notable trends and observations.

3.1. Dataset Collection

We validate the MOVGP regressor on a hydrological time series forecasting task of the Useful Volume and Contributions from 23 Colombian reservoirs, daily recorded daily from 1 January 2010, to 28 February 2022, yielding 4442 daily measurements. Despite the volumetric nature of the raw data, the hydroelectric power plants report Useful Volumes and Streamflow Contributions as their equivalent in kilowatt-hour (kWh) units, since such a representation is more practical for daily operations. Figure 1 visually describes the statistics for each reservoir. Note that amplitudes vary from millions to billions of kWh among reservoirs, due to each generating at a different capacity. In the case of Useful Volumes, one finds some highly averaging time series with a few variations (see reservoirs A and L), but also cases of low means with a significant variation (as the reservoir B). Notice from Streamflow Contributions that the reservoir K boxplot does not appear, due to zero values reported for all time series. Some reservoirs also present outlier volume reductions (black crosses), contrasting with the outlier increments in streamflow. Two main factors produce the above nonstationary and non-Gaussian behavior: the Colombian weather conditions produce unusual rainy days and long dry seasons, and the operation decisions can impose water saving or generation at total capacity each day.



(a) Useful volume



(b) Streamflow Contributions

Figure 1. Time series distributions of Useful Volume and Streamflow Contribution for each reservoir on the dataset. Amplitude axis is presented in logarithmic mode due to large scale variations.

3.2. MOVGP Setup and Hyperparameter Tuning

Firstly, we define the validation task as predicting the H -th day in the future using the current hydrological measurement in all the reservoirs, yielding $L = 23$ inputs. Prediction horizon H ranges from one to twenty-five days, exploring short- and medium-term hydrological forecasting performance. For proper validation, MOVGP and contrasting approaches were trained on the first 10 years and validated on the data from 28 February 2021 to 28 February 2022, corresponding to 365 testing samples. Figure 2 presents the testing time series for the Useful Volume of three reservoirs. Noting the varying scales that lead to potential bias during the training stage, a preprocessing step normalizes the dataset by centering the time series from each reservoir on zero and scaling it to unit standard deviation. Normalizing means and standard deviations result from the training subset statistics avoids test biasing.

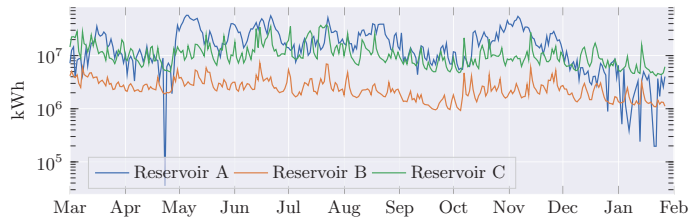


Figure 2. Testing data for Streamflow Contributions of three reservoirs from March 2021 to February 2022, evidencing the within and between time series variability.

Each type of time series, Useful Volume and Streamflow Contribution, considers an individual forecasting model. Hence, the experimental framework trains two independent MOVGPs with $D = 23$ outputs. The proposed methodology considers a constant for the MOVGP mean function, $m^D(\mathbf{x} \mid \theta_m) = \theta_m$, with $\theta_m \in \mathbb{R}^D$ as the single trainable vector parameter. The proposed methodology builds the MOVGP covariance function in Equation (12) from the widely used squared exponential, in Equation (16), allowing a smooth data mapping:

$$k_q(\mathbf{x}, \mathbf{x}' \mid \theta_q) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^\top \Theta_q^{-2}(\mathbf{x} - \mathbf{x}')\right) \quad (16)$$

where the diagonal matrix $\Theta_q = \text{diag}\{\Delta_{ql}\}_{l=1}^L \in \mathbb{R}^{L \times L}$ gathers the length scale factors $\Delta_{ql} \in \mathbb{R}^+$ from each input dimension. The trainable covariance parameters become $\sigma_{N,d}$ and Δ_{lq} from the Q independent SOGP within the MOVGP framework. Then, a 10-fold time series split model selection determines the optimal hyperparameter setting for the forecasting models by searching within the following grid: number of inducing variables $M \in \{4, 8, 16, 32, 64, 128\}$ and latent space dimension $Q \in \{2, 4, 8, 16, 23, 46, 69, 92, 115\}$.

Figure 3 presents the 10-fold-averaged cross-validation mean squared error (MSE) along the grid search while fixing the model horizon to $H = 1$ for both the Useful Volume and Streamflow Contributions. Hyperparameter tuning exhibits that, the larger the Q and M , the smaller the MOVGP error and the slower the improvement. Therefore, the forecast task on a very short horizon yields complex models that hardly overfit. However, the latent dimension influences the performance significantly more than the induced variables, agreeing with the model development: the latent dimension controls the embedding quality, while the induced variables reduce the computational burden without compromising the performance.

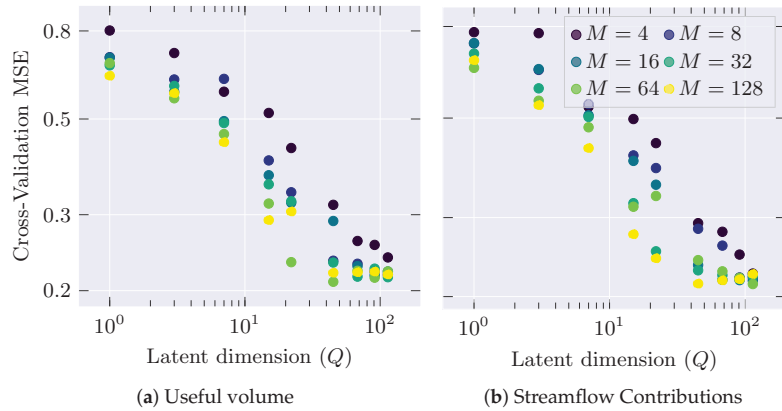


Figure 3. MOVGP hyperparameter tuning for horizon $H = 1$ using a grid search on the latent space dimension L and the number of induced variables M . Testing MSE is computed on a 10-fold cross-validation.

For each considered horizon $H \in \{1, 2, 3, 5, 7, 9, 10, 15, 20, 25\}$, Figure 4 illustrates the hyperparameters which reach the best testing MSE. For the Useful Volume time series, shown in Figure 4a, the number of latent variables Q increases while the number of inducing points M decreases. A Pearson correlation coefficient of -0.82 between the optimal Q and M indicates that the model trades off its complexity between hyperparameters: increasing Q allows a more flexible model, whereas increasing M produces MOVGP models that retain more information about the time series. In turn, the optimal Q for the Streamflow Contribution remains at the highest evaluated value while M decreases for the last horizons (Figure 4b). Such a fact suggests that the latent space is large enough to decode the relationship between past Streamflow Contributions and the farthest horizon. Thus, a flexible model evades a large explicit memory to seize the relevant dynamics, and vice versa.

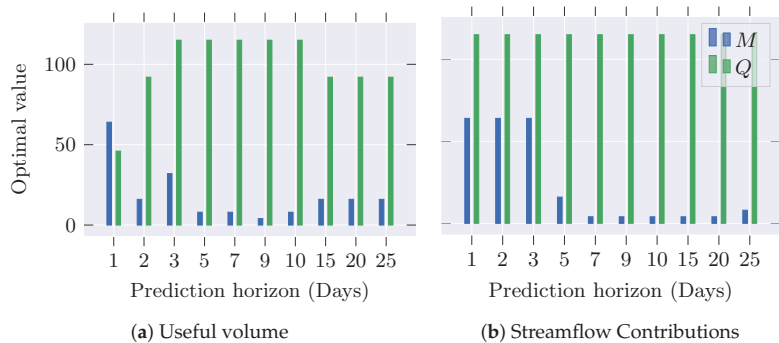


Figure 4. Optimal MOVGP hyperparameters, according to the testing MSE along the prediction horizon H for Useful Volume and Streamflow Contributions.

3.3. Performance Analysis

The performance analysis compares MOVGP against two widely considered hydrological forecasting models: a straightforward Linear AutoRegression (LAR) model as a baseline and a Long Short-Term Memory (LSTM) network with the hidden space dimension and the number of recurrent layers as hyperparameters. Specifically for the LSTM, the same model selection strategy—10-fold time series split—tunes the hyperparameters using the training subset. Figure 5 illustrates the MSE for the 10-fold cross-validation attained by

the three contrasted models along the explored horizons for both time series. In general, error increases with the prediction horizon, due to the forecasting task becoming more complex for far away days. Nonetheless, the autoregressive model outperforms LSTM, closely followed by MOVGP, up to a 15-day horizon, suggesting linearly-captured time dependencies in the short term. In contrast, MOVGP reaches the lowest error for the longest horizons, followed by LSTM, evidencing nonlinear time relationships at medium-term which are profited by more elaborate models. The above results indicate that MOVGP was the most flexible model on average, exploiting the time-varying interactions, competing at short-term, and outperforming at medium-term horizons.

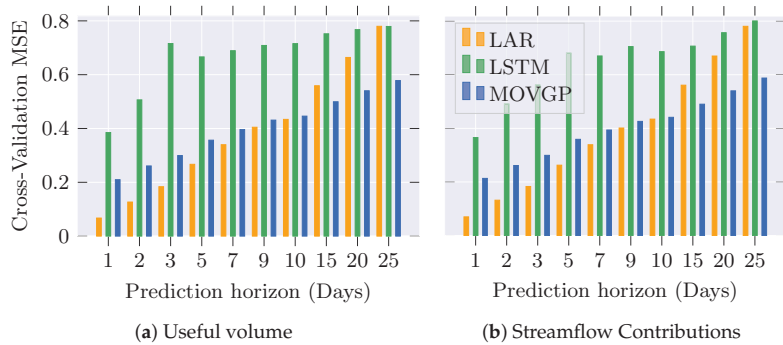


Figure 5. Cross-validation MSE at a 10-fold time series split for Linear AutoRegressive, LSTM, and MOVGP models along the prediction horizon.

At the testing stage, trained LAR, LSTM, and MOVGP models forecast the last 365 days on the dataset for both time series at each horizon considered. Figure 6 depicts reservoir-wise testing MSE boxplots computed over the 10 prediction horizons for the Streamflow Contributions. Note that reservoir T makes the models perform the worst, whereas reservoir D becomes the least challenging. Moreover, the widespread error at reservoir L contrasts with the small dispersion at reservoir D. Therefore, varying boxplots advise the changing forecasting complexity over the horizons and reservoirs despite corresponding to the same hydrological time series. According to the grand averages in dashed lines, the MOVGP model obtains the best average performance, followed by the LAR. Thus, for the Contributions time series, MOGPG offers a better explanation for nonlinearities in the data than LSTM.

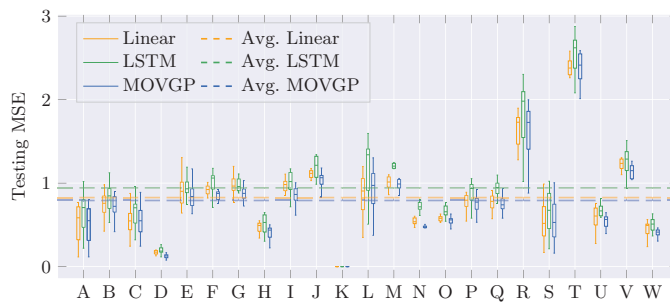


Figure 6. Distribution for the testing MSE of contrasted approaches at each reservoir for the Streamflow Contributions. Statistics are computed over the 10 prediction horizons. The dashed line averages the reservoir-wise MSE.

Figure 7 offers time series plots for Useful Volume and Contributions with their respective one-day horizon predictions from the forecast models for three reservoirs of

interest. Notice, in Figure 7a, that the MOVGP and Linear models reach a well-fit prediction and learn the smoothness of the reservoir data, but the MOVGP model presents the advantage of yielding a predictive distribution and, therefore, a confidence interval. In addition, the LSTM model shows abrupt changes that deviate from the actual behavior of the curve, producing a higher error. For Figure 7c, a narrow confidence interval describes the time series noise. Observe the presence of a peak at day 65 that is out of the confidence interval, possibly an outlier classified as an anomaly by predictive distribution offered by the MOVGP model.

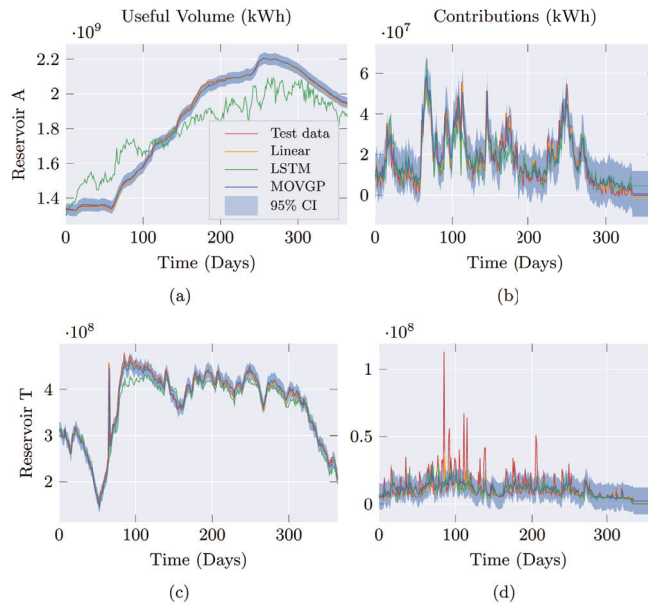


Figure 7. Useful volume (left) and contributions (right) forecasted by the contrasting approaches on one-year test data at reservoirs A, K, and T (top to bottom) and one-day prediction horizon $H = 1$.

In the case of the Streamflow Contributions, the three models closely follow the abrupt curve trend and lie within the confidence intervals for reservoir A in Figure 7b. Lastly, Figure 7d displays peaks in the Streamflow Contributions of reservoir T. Although no model catches the peak’s tendency, MOVGP explains them as outliers because of the predictive distribution. In this way, the MOVGP model is less influenced by anomalies, producing better generalization and, thus, outperforming the other models.

4. Concluding Remarks

This work proposed a forecasting methodology for multiple output prediction of Useful Volume and Streamflow Contributions of Colombian reservoirs using Variational Gaussian Processes. Since the coregionalization of MOVGPs imposes a unique latent process, generating multiple outputs, we devoted a single model for each hydrological variable to minimize overgeneralization issues. The proposed MOVGP was compared against LSTM-based and Linear AutoRegressive models using actual time series. The hyperparameter tuning stage proved that MOVGP suitably adapted to time complexity by optimizing the number of latent variables and inducing points to control model flexibility. The comparison in testing data, shown in Figure 5, revealed that the MOVGP outperformed the others in predicting long-term horizons, particularly when the linear model missed relationships between inputs and outputs. Therefore, MOVGP outperforms hydrological forecasting, providing prediction reliability and outlier detection through the predictive distribution.

For future work, we devise the following research directions. First, we will extend the methodology to support energy-related time series such as daily thermoelectric schedules. Secondly, we will develop deep learning models to learn complex patterns in hydrological time series.

Finally, to overcome the overgeneralization and linear coregionalization restriction issues, we will work on time-variant convolutional kernel integration.

Author Contributions: Conceptualization, J.D.P.-C.; Data Extraction, A.A.O.-G. and M.H.-L.; Validation, D.A.C.-P. and G.C.-D.; Original Draft Preparation, J.D.P.-C., A.A.O.-G. and D.A.C.-P.; Review and Editing, J.D.P.-C. and D.A.C.-P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Minciencias project: “Desarrollo de una herramienta para la planeación a largo plazo de la operación del sistema de transporte de gas natural de Colombia”—código de registro 69982—CONVOCATORIA DE PROYECTOS CONECTANDO CONOCIMIENTO 2019 852-2019.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Thanks to the Maestría en Ingeniería Eléctrica, graduate program of the Universidad Tecnológica de Pereira.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Basu, M. Improved differential evolution for short-term hydrothermal scheduling. *Int. J. Electr. Power Energy Syst.* **2014**, *58*, 91–100. [CrossRef]
- Nazari-Heris, M.; Mohammadi-Ivatloo, B.; Gharehpetian, G.B. Short-term scheduling of hydro-based power plants considering application of heuristic algorithms: A comprehensive review. *Renew. Sustain. Energy Rev.* **2017**, *74*, 116–129. [CrossRef]
- Freire, P.K.D.M.M.; Santos, C.A.G.; Silva, G.B.L.D. Analysis of the use of discrete wavelet transforms coupled with ANN for short-term streamflow forecasting. *Appl. Soft Comput. J.* **2019**, *80*, 494–505. [CrossRef]
- XM. La Generación de Energía en enero fue de 6276.74 gwh. 2022. Available online: <https://www.xm.com.co/noticias/4630-la-generacion-de-energia-en-enero-fue-de-627674-gwh> (accessed on 7 November 2022).
- Cheng, M.; Fang, F.; Kinouchi, T.; Navon, I.; Pain, C. Long lead-time daily and monthly streamflow forecasting using machine learning methods. *J. Hydrol.* **2020**, *590*, 125376. [CrossRef]
- Yaseen, Z.M.; Allawi, M.F.; Yousif, A.A.; Jaafar, O.; Hamzah, F.M.; El-Shafie, A. Non-tuned machine learning approach for hydrological time series forecasting. *Neural Comput. Appl.* **2018**, *30*, 1479–1491. [CrossRef]
- Mosavi, A.; Ozturk, P.; Chau, K.W. Flood prediction using machine learning models: Literature review. *Water* **2018**, *10*, 1536. [CrossRef]
- Apaydin, H.; Feizi, H.; Sattari, M.T.; Colak, M.S.; Shamshirband, S.; Chau, K. Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. *Water* **2020**, *12*, 1500. [CrossRef]
- Zhu, S.; Luo, X.; Xu, Z.; Ye, L. Seasonal streamflow forecasts using mixture-kernel GPR and advanced methods of input variable selection. *Hydrol. Res.* **2019**, *50*, 200–214. [CrossRef]
- Saraiva, S.V.; de Oliveira Carvalho, F.; Santos, C.A.G.; Barreto, L.C.; de Macedo Machado Freire, P.K. Daily streamflow forecasting in Sobradinho Reservoir using machine learning models coupled with wavelet transform and bootstrapping. *Appl. Soft Comput.* **2021**, *102*, 107081. [CrossRef]
- Ni, L.; Wang, D.; Wu, J.; Wang, Y.; Tao, Y.; Zhang, J.; Liu, J. Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. *J. Hydrol.* **2020**, *586*, 124901. [CrossRef]
- Sun, A.Y.; Wang, D.; Xu, X. Monthly streamflow forecasting using Gaussian process regression. *J. Hydrol.* **2014**, *511*, 72–81. [CrossRef]
- Niu, W.; Feng, Z. Evaluating the performances of several artificial intelligence methods in forecasting daily streamflow time series for sustainable water resources management. *Sustain. Cities Soc.* **2021**, *64*, 102562. [CrossRef]
- Zhu, S.; Luo, X.; Yuan, X.; Xu, Z. An improved long short-term memory network for streamflow forecasting in the upper Yangtze River. *Stoch. Environ. Res. Risk Assess.* **2020**, *34*, 1313–1329. [CrossRef]
- Moreno-Muñoz, P.; Artés, A.; Álvarez, M. Heterogeneous Multi-output Gaussian Process Prediction. In *Proceedings of the Advances in Neural Information Processing Systems*; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2018; Volume 31.
- Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; Adaptive Computation and Machine Learning; MIT Press: Cambridge, MA, USA, 2006; pp. 1–248.

17. Liu, H.; Cai, J.; Ong, Y.S. Remarks on multi-output Gaussian process regression. *Knowl.-Based Syst.* **2018**, *144*, 102–121. [CrossRef]
18. Álvarez, M.A.; Rosasco, L.; Lawrence, N.D. Kernels for Vector-Valued Functions: A Review. *Found. Trends[®] Mach. Learn.* **2012**, *4*, 195–266. [CrossRef]
19. Hensman, J.; Fusi, N.; Lawrence, N.D. Gaussian Processes for Big Data. *arXiv* **2013**, arXiv:1309.6835. [CrossRef]
20. Hensman, J.; Matthews, A.; Ghahramani, Z. Scalable Variational Gaussian Process Classification. *arXiv* **2014**, arXiv:1411.2005. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Proceeding Paper

Geodynamic Modeling in Central America Based on GNSS Time Series Analysis—Special Case: The Nicoya Earthquake (Costa Rica, 2012) †

Paola Barba ^{1,*}, Nely Pérez-Méndez ¹, Javier Ramírez-Zelaya ¹, Belén Rosado ¹, Vanessa Jiménez ²
and Manuel Berrocoso ¹

¹ Laboratorio de Astronomía, Geodesia y Cartografía, Departamento de Matemáticas, Facultad de Ciencias, Campus de Puerto Real, Universidad de Cádiz, Puerto Real, 11510 Cádiz, Spain; nelyperez1510@gmail.com (N.P.-M.); javierantonio.ramirez@uca.es (J.R.-Z.); belen.rosado@uca.es (B.R.); manuel.berrocoso@uca.es (M.B.)

² Departamento de Física Teórica y del Cosmos, Facultad de Ciencias (Edificio Mecenas), Campus de Fuentenueva, Universidad de Granada, 18010 Granada, Spain; vanessa.jimenezmorales@hotmail.com

* Correspondence: paola.barba@uca.es

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: GNSS systems allow precise resolution of the geodetic positioning problem through advanced techniques of GNSS observation processing (PPP or relative positioning). Current instrumentation and communications capabilities allow obtaining geocentric and topocentric geodetic high frequency time series, whose analysis provides knowledge of the tectonic or volcanic geodynamic activity of a region. In this work, a GNSS time series study is carried out through the use and adaptation of R packets to determine their behavior, obtaining displacement velocities, noise levels, precursors in the time series, anomalous episodes and their temporal forecast. Statistical and analytical methods are studied; for example, ARMA, ARIMA models, least-squares methods, wavelet functions, Kalman techniques and CATS analysis. To obtain a geodynamic model of the Central American region, the horizontal and vertical velocities obtained by applying the above methods are taken, choosing the velocity with the least margin of error. Significant GNSS time series are obtained in geodynamically active regions (tectonic and/or volcanic).

Keywords: GNSS time series; geodynamic model; Nicoya earthquake

Citation: Barba, P.; Pérez-Méndez, N.; Ramírez-Zelaya, J.; Rosado, B.; Jiménez, V.; Berrocoso, M.

Geodynamic Modeling in Central America Based on GNSS Time Series Analysis—Special Case: The Nicoya Earthquake (Costa Rica, 2012). *Eng. Proc.* **2023**, *39*, 84. <https://doi.org/10.3390/engproc2023039084>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 12 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the area of Central America, one of the most interesting geodynamic zones in the world is observed, with the convergence of five tectonic plates: the South American, North American, Nazca, Cocos, and Caribbean plates. The evolution of these plates is well known, except for the origin of the Caribbean plate, which is still the subject of debate [1–3].

There are two main models for the origin of the Caribbean plate. The first is one that contemplates the origin of the plate “in situ” between the North and South American plates [2]. The second and most commonly accepted model is that the Caribbean plate originated in the Pacific Ocean during the Late Cretaceous, composed primarily of a large igneous province due to hot spot magmatism and normal-thickness oceanic crust that migrated eastward to its present position between the American plates [1,4,5].

The Caribbean plate is an enclosed oceanic basin composed primarily of the Caribbean Greater Igneous Province and large areas of ocean crust of normal thickness within the Venezuelan and Colombian basins [4]. It is bounded to the north and south by two continental plates, the North and South American plates, which are characterized by large strike-slip fault systems, although convergence also occurs on a smaller scale [6]. It is

bounded to the east and west by two main subduction zones, the Lesser Antilles and Central America, respectively, where the oceanic lithosphere of the Atlantic Ocean to the east and the Pacific Ocean to the west subducts beneath the Caribbean plate [7]. In addition, this plate subducts under northwestern South America, reaching a depth of 600 km [4].

At the western edge of the Caribbean plate lies the Nicoya Peninsula, where the Cocos Plate subducts northeastward beneath the Caribbean Plate along the Middle America Trench at about 8 cm/yr, with a range of obliquity of 10° counterclockwise from the normal direction of the trench [8,9]. The movement of the plates along the Middle American Trench is partitioned. In the normal direction of the trench, the subduction velocity is 74–84 mm/yr [10], while in the forward arc the movement is 8–14 mm/yr [8,11,12]. The subducting Cocos plate forms on both the rapidly extending East Pacific Ridge (EPR), with relatively uniform seafloor topography, and the slowly extending Cocos-Nazca Ridge (CNR), with a relatively rugged seabed [13,14].

The Nicoya Peninsula is elongated in a northwest–southeast direction and extends within 60 km of the Trench [15], and is located over the shallow portion of the subduction interface that generates earthquakes, called megathrust [16]. There are only a few subduction zones in the world where there is land access of as close as 50–60 km to a deep trench [17] and where the megathrust seismogenic zone is covered by land rather than ocean [16].

In this area, large megathrust earthquakes with magnitudes greater than 7 have historically occurred with a well-defined seismic cycle of about 50 years, 1853, 1900 and 1950 ($M_w = 7.7$), 2012 ($M_w = 7.6$) [14,16,18], in addition to the record of other smaller nearby events of magnitude ($M_w = 7$) in 1978 and 1990 [14]. Slow slip events (SSEs) are common below the Nicoya Peninsula [10,19]. These events are basically largely aseismic slip that occurs at the plate boundary for weeks to months [20,21] releasing a fraction of the accumulated stress aseismically or weakly seismically, often accompanied by low-frequency earthquakes and non-volcanic seismic tremors [22].

On 5 September 2012, in the North Pacific region of the Nicoya peninsula, a $M_w = 7.6$ earthquake was recorded SW of Sámara, which had a large number of aftershocks in the following months (42 in September to the SSW of the peninsula from Nicoya; 24 in October to the WSW of the Nicoya Peninsula; 10 in November on the coast of the Nicoya Peninsula). Most of these earthquakes have their hypocenters at depths between 15 and 20 km. This earthquake on 5 September and its aftershocks were caused by the subduction process of the Cocos plate under the Caribbean plate, a process that has generated other historic earthquakes in Guanacaste such as the 1950 Nicoya earthquake ($M_s = 7.7$). There was damage to homes and buildings. The solution of the focal mechanism carried out by the USGS for the Samara earthquake shows a pure inverse type mechanism that confirms its relationship with the subduction process of the Cocos plate [23].

Ref. [17] relocated the earthquake of 5 September 2012 using data from the local seismic network. He located the hypocenter at 9.76° N, 85.56° W at 10 km offshore, 13 km deep at megathrust, with seismic moment being 3.5×10^{20} Nm, giving $M_w = 7.6$. The joint finite fault inversion of GPS data, seismometers, accelerometers, teleseismic P-waves, and GPS static offsets revealed that the coseismic rupture propagated downward from the hypocenter with a rupture velocity of 3.0 km/s and a total source duration of 21 s.

2. Data Collection

GNSS technologies and permanent station networks have created a very relevant terrestrial reference framework and tool for the study of deformations of the Earth's crust due to tectonic forces. These technologies are of great interest for geodynamics and deformation studies. Although strain is a more objective indicator than displacement because no frame of reference is required [24], GNSS techniques make it possible to quantify with guarantee the displacements of the stations that occur during earthquakes and relate them to other unaffected areas; as a consequence, the horizontal and vertical movements can be measured in faults and tectonically active regions. The GPS system has proven

to be a very effective tool to carry out deformation studies due to its high precision and accuracy [25].

Nevada Geodetic Laboratory, (<http://geodesy.unr.edu/>, accessed on 2 March 2023) provides all available raw GPS geodetic data from over 17,000 stations around the world; these stations form the MAGNET network. Managing this large amount of data has led to the development of new processing strategies, automated systems, algorithms and robust estimation techniques. From these data, time series of Cartesian coordinates (every 24 h), topocentric coordinates (every 24 h and every 5 min), tropospheric delay and predictions are available. The NGL laboratory provides a multitude of graphs and their corresponding displacement velocities, and indicates with dashed lines the instant of time of the earthquakes that have occurred and the equipment change events near the station.

Plots of the unfiltered and filtered data, the results of fixing the series against an African plate, are provided. The NGL Lab also provides detrending series plots and predictions, filtered against the African and Eurasian Plate [26]. Every week the daily position coordinates of about 10,000 stations are updated. Every day, position coordinates are updated every 5 min for more than 5000 stations. Every hour, we update the 5-min position coordinates for about 2000 stations.

The NGL laboratory routinely updates station velocities, which can be used to image deformation rates of the Earth's surface for a variety of interdisciplinary applications. These velocities are robustly estimated using the asymmetry-adjusted mean interannual difference (MIDAS software), a median-based GPS station velocity estimator that is insensitive to outliers, seasonality, step functions (abrupt changes) arising from earthquakes or equipment changes, and variability of statistical data [27]. Still, at NGL, for cases where an earthquake of magnitude greater than 6.9 has occurred, close enough to the station, we solve an exponential function defined by $A(1 - \exp(-(t - t_0)/\tau))H(t - t_0)$ where t_0 is the earthquake time, τ is the relaxation time, A is the decay amplitude and H is the Heaviside step function. In these cases, we retrend after removing the exponential terms to obtain a self-consistent model for the time series.

Figure 1 shows the vector model developed from the velocities extracted from NGL, where the arrows mark the magnitude and direction in which each permanent GNSS station is moving per year.

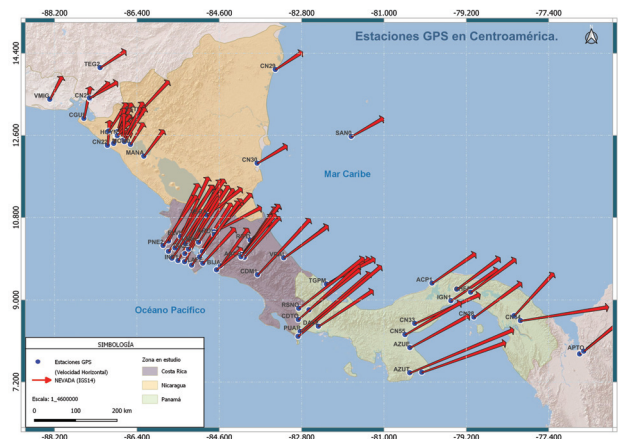


Figure 1. Vector model from the velocities extracted from the NGL laboratory.

3. Methodology

In this paper, the behavior of the time series of the stations near the earthquake that occurred in Nicoya on 5 September 2012 of magnitude 7.6 will be analyzed (see Figure 2).

The topocentric time series provided by the NGL laboratory was extracted. In this work, time series will be distinguished into three phases: preseismic, coseismic and post-

seismic. The preseismic phase is formed by the data prior to the moment of the earthquake; the coseismic phase comprises the data from the occurrence of the earthquake until the earth's crust recovers; and the postseismic phase is characterized by being the data that goes from when the trend becomes linear again (recovers its preseismic behavior) until the end of the available data. Various analytical and statistical techniques have been applied to these data [28]. To obtain the velocities corresponding to the preseismic and postseismic phases, the CATS adjustment will be applied, which provides a model formed by the sine and cosine functions that adapt to the values of the time series; through the given model the displacement velocities are obtained from the series. The linear fit will be used to obtain the velocities of the coseismic series.

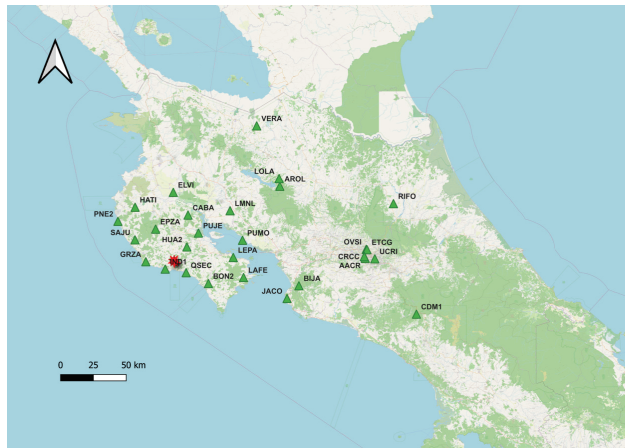


Figure 2. GNSS-GPS stations belonging to the MAGNET network. The coordinates of the earthquake are marked in red.

4. Analysis of the Series Affected by the Earthquake

The series affected by earthquakes have a common behavior between them (see Figure 3):

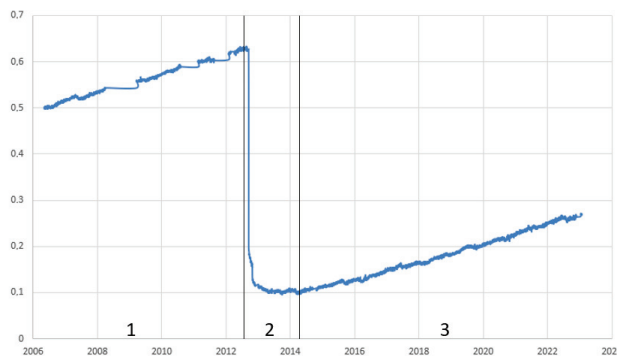


Figure 3. Time series of the GRZA station for the east component, (1) preseismic phase, (2) coseismic phase, (3) postseismic phase.

Following the methodology, the displacement speeds of each of the defined phases are obtained, thus giving three vector models of horizontal displacement. In this way it will be possible to see the magnitude and direction of the displacement of the GNSS-GPS permanent stations during the different phases. In addition, the stress-strain models will be

obtained, thus seeing the zones of maximum geodesic deformation for each of the phases. For this, the “Q-Str2-Models” plugins available in QGIS will be implemented [29].

5. Conclusions

All the time series close to the Nicoya earthquake on 5 September 2012 present a behavior similar to that shown in Figure 3, in which three phases are distinguished: pre-seismic, coseismic and postseismic. The travel speeds were calculated for each station shown in Figure 2, in each of the phases. In this way, the graphs of Figures 4–6 are obtained. In Figure 4a, the vector displacement model can be seen in the pre-seismic phase, seeing that the stations near the coast present a similar behavior. Figure 4b shows the maximum geodesic deformation that occurred in the pre-seismic phase, observing a greater deformation in the area where the SAJU, PNE2 and GRZA stations are located. In Figure 5a we have the vector model for the case of the coseismic phase; it can be seen how the direction and magnitude of the horizontal displacement changes. Figure 5b shows how the geodesic deformation increased compared to the values obtained in Figure 4b. However, said deformation is caused in the same zone by also adding the zones of the HATI, IND1, QSEC and BON2 stations. In Figure 6a you can see the moment in which the Earth’s crust again has a linear trend; even so, it does not present a behavior as homogeneous as that visible in Figure 4a. Figure 6b shows the maximum geodesic deformation of the region; it can be seen how the PNE2, SAJU, GRZA, IND1 and QSEC stations continue to form part of the areas with the greatest deformation; in addition, this deformation can be seen in a downtown station such as CDM1.

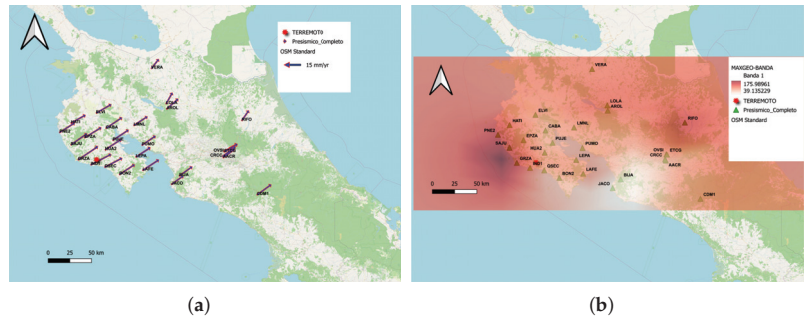


Figure 4. Preseismic phase. (a) vector displacement model in the preseismic phase. (b) representation of the maximum geodetic deformation in the preseismic phase.

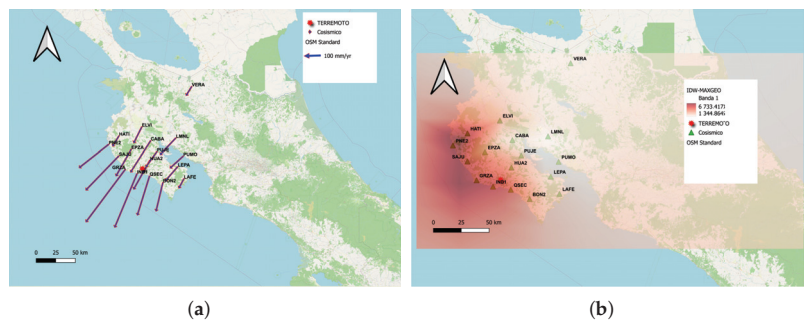


Figure 5. Coseismic phase. (a) vector displacement model in the coseismic phase. (b) representation of the maximum geodetic deformation in the coseismic phase.

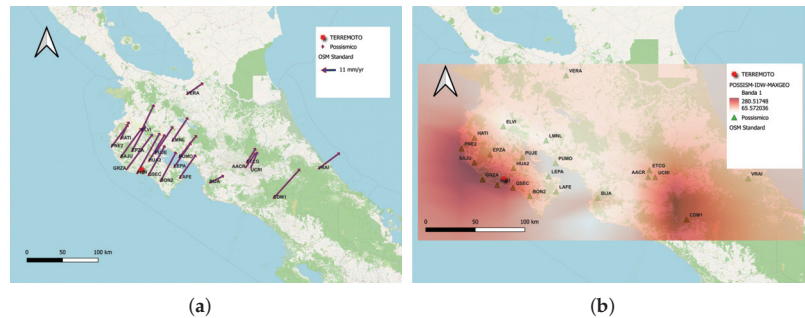


Figure 6. Postseismic phase. (a) vector displacement model in the postseismic phase. (b) representation of the maximum geodetic deformation in the coseismic phase.

Author Contributions: Conceptualization, P.B., V.J. and M.B.; methodology, P.B., N.P.-M. and M.B.; software, P.B., N.P.-M. and J.R.-Z.; validation, P.B., B.R. and M.B.; formal analysis, P.B. and M.B.; investigation, P.B. and N.P.-M.; resources, P.B. and N.P.-M.; data curation, P.B.; writing—original draft preparation, P.B. and V.J.; writing—review and editing, P.B. and M.B.; visualization, P.B., N.P.-M. and M.B.; supervision, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: Funded by the “INICIA-INV” grant from the “Own Plan 2021–2022” from the University of Cádiz.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data corresponding to the time series used are available at <http://geodesy.unr.edu/NGLStationPages/gpsnetmap/GPSNetMap.html>.

Acknowledgments: Thank the University of Cádiz (UCA) for the financial aid “INICIA-INV” of the “Plan Propio 2021–2022”.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Burke, K.; Fox, P.J.; Şengör, A.M.C. Buoyant ocean floor and the evolution of the Caribbean. *J. Geophys. Res.* **1978**, *83*, 3949–3954. [CrossRef]
- James, K.H. Arguments for and against the Pacific origin of the Caribbean Plate: Discussion, finding for an inter-American origin. *Geológica Acta Int. Earth Sci. J.* **2006**, *4*, 279–302. Available online: <https://www.redalyc.org/pdf/505/50540216.pdf> (accessed on 24 April 2023).
- Pindell, J.; Dewey, J.F. Permo-triassic reconstruction of western Pangea and the evolution of the Gulf of Mexico/Caribbean region. *Tectonics* **1982**, *1*, 179–211. [CrossRef]
- Barrera-Lopez, C.V.; Mooney, W.D.; Kaban, M.K. Regional geophysics of the Caribbean and northern South America: Implications for tectonics. *Geochem. Geophys. Geosystems* **2022**, *23*, e2021GC010112. [CrossRef]
- van Benthem, S.; Govers, R.; Spakman, W.; Wortel, R. Tectonic evolution and mantle structure of the Caribbean. *J. Geophys. Res. Solid Earth* **2013**, *118*, 3019–3036. [CrossRef]
- Molnar, P.; Sykes, L.R. Tectonics of the Caribbean and middle America regions from focal mechanisms and seismicity. *Geol. Soc. Am. Bull.* **1969**, *80*, 1639–1684. [CrossRef]
- García-Reyes, A.; Dymment, J. Structure, age, and origin of the Caribbean Plate unraveled. *Earth Planet. Sci. Lett.* **2021**, *571*, 117100. [CrossRef]
- DeMets, C. A new estimate for present-day Cocos-Caribbean plate motion: Implications for slip along the Central American volcanic arc. *Geophys. Res. Lett.* **2001**, *28*, 4043–4046. [CrossRef]
- DeMets, C.; Gordon, R.G.; Argus, D.F. Geologically current plate motions. *Geophys. J. Int.* **2010**, *181*, 1–80. [CrossRef]
- Outerbridge, K.C.; Dixon, T.H.; Schwartz, S.Y.; Walter, J.I.; Protti, M.; Gonzalez, V.; Rabbel, W. A tremor and slip event on the Cocos-Caribbean subduction zone as measured by a global positioning system (GPS) and seismic network on the Nicoya Peninsula, Costa Rica. *J. Geophys. Res. Solid Earth* **2010**, *115*, B10. [CrossRef]
- LaFemina, P.; Dixon, T.H.; Govers, R.; Norabuena, E.; Turner, H.; Saballos, A.; Mattioli, G.; Protti, M.; Strauch, W. Fore-arc motion and Cocos Ridge collision in Central America. *Geochem. Geophys. Geosyst.* **2009**, *10*, Q05S14. [CrossRef]

12. Norabuena, E.; Dixon, T.H.; Schwartz, S.; DeShon, H.; Newman, A.; Protti, M.; Gonzalez, V.; Dorman, L.; Flueh, E.R.; Lundgren, P. Geodetic and seismic constraints on seismogenic zone processes in Costa Rica. *J. Geophys. Res.* **2004**, *109*, B11403. [CrossRef]
13. Barckhausen, U.; Ranero, C.R.; von Huene, R.; Cande, S.C.; Roeser, H.A. Revised tectonic boundaries in the Cocos Plate off Costa Rica: Implications for the segmentation of the convergent margin and for plate tectonic models. *J. Geophys. Res. Solid Earth* **2001**, *106*, 19207–19220. [CrossRef]
14. Protti, M.; McNally, K.; Pacheco, J.; Gonzalez, V.; Montero, C.; Segura, J.; Schillinger, W. The March 25, 1990 (Mw = 7.0, ML = 6.8), earthquake at the entrance of the Nicoya Gulf, Costa Rica: Its prior activity, foreshocks, aftershocks, and triggered seismicity. *J. Geophys. Res. Solid Earth* **1995**, *100*, 20345–20358. [CrossRef]
15. Malservisi, R.; Schwartz, S.Y.; Voss, N.; Protti, M.; Gonzalez, V.; Dixon, T.H.; Vayenko, D. Multiscale postseismic behavior on a megathrust: The 2012 Nicoya earthquake, Costa Rica. *Geochem. Geophys. Geosyst.* **2015**, *16*, 1848–1864. [CrossRef]
16. Protti, M.; González, V.; Newman, A.V.; Dixon, T.H.; Schwartz, S.Y.; Marshall, J.S.; Owen, S.E. Nicoya earthquake rupture anticipated by geodetic measurement of the locked plate interface. *Nat. Geosci.* **2014**, *7*, 117–121. [CrossRef]
17. Yue, H.; Lay, T.; Schwartz, S.Y.; Rivera, L.; Protti, M.; Dixon, T.H.; Newman, A.V. The 5 September 2012 Nicoya, Costa Rica Mw 7.6 earthquake rupture process from joint inversion of high-rate GPS, strong-motion, and teleseismic P wave data and its relationship to adjacent plate boundary interface properties. *J. Geophys. Res. Solid Earth* **2013**, *118*, 5453–5466. [CrossRef]
18. Satake, K. Mechanism of the 1992 Nicaragua tsunami earthquake. *Geophys. Res. Lett.* **1994**, *21*, 2519–2522. [CrossRef]
19. Jiang, Y.; Wdowinski, S.; Dixon, T.H.; Hackl, M.; Protti, M.; Gonzalez, V. Slow slip events in Costa Rica detected by continuous GPS observations, 2002–2011. *Geochem. Geophys. Geosyst.* **2012**, *13*, 8–13. [CrossRef]
20. Brodsky, E.E.; Mori, J. Creep events slip less than ordinary earthquakes. *Geophys. Res. Lett.* **2007**, *34*, L16309. [CrossRef]
21. Dixon, T.H.; Jiang, Y.; Malservisi, R.; McCaffrey, R.; Voss, N.; Protti, M.; Gonzalez, V. Earthquake and tsunami forecasts: Relation of slow slip events to subsequent earthquake rupture. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 17039–17044. [CrossRef] [PubMed]
22. Xie, S.; Dixon, T.H.; Protti, M.; Malservisi, R.; Jiang, Y.; Muller, C. Shallow versus deep slow slip events on the Nicoya megathrust observed with GPS. In *AGU Fall Meeting Abstracts*; American Geophysical Union: Washington, DC, USA, 2020; Volume 2020, pp. 003–0018.
23. Linkimer, L.; Barquero, R.; Vargas, A.; Rojas, W.; Taylor, M.; Araya, M.C. Actividad sísmica en Costa Rica durante el 2012. *Rev. Geológica América Cent.* **2013**, *49*, 141–148. [CrossRef]
24. Takahashi, H. Static strain and stress changes in Eastern Japan due to 2011 off the Pacific coast of Tohoku Earthquake, as derived from GPS data. *Earth Planets Space* **2011**, *63*, 741–744. [CrossRef]
25. Kulkarni, M.N.; Radhakrishnan, N.; Rai, D. Global positioning system in disaster monitoring of Koyna Dam, western Maharashtra. *Surv. Rev.* **2006**, *37*, 490–497. [CrossRef]
26. Blewitt, G.; Hammond, W.C.; Kreemer, C. Harnessing the GPS data explosion for interdisciplinary science. *Eos* **2018**, *99*. [CrossRef]
27. Blewitt, G.; Kreemer, C.; Hammond, W.C.; Gazeaux, J. MIDAS robust trend estimator for accurate GPS station velocities without step detection. *J. Geophys. Res. Solid Earth* **2016**, *121*, 2054–2068. [CrossRef]
28. Barba, P.; Rosado, B.; Ramírez-Zelaya, J.; Berrocoso, M. Comparative Analysis of Statistical and Analytical Techniques for the Study of GNSS Geodetic Time Series. *Eng. Proc.* **2021**, *5*, 21.
29. Ramírez-Zelaya, J.; Peci, L.M.; Fernández-Ros, A.; Rosado, B.; Pérez-Peña, A.; Gárate, J.; Berrocoso, M. Q-Str2-Models: A software in PyQGIS to obtain Stress–Strain models from GNSS geodynamic velocities. *Comput. Geosci.* **2023**, *172*, 105308. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Investigation of FIBA World Cup 2019: Evidence Using Advanced Statistical Analysis and Quantitative Tools [†]

Christos Katris

Department of Mathematics, University of Patras, 26504 Patras, Greece; chriskatris@upatras.gr

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: The purpose of this study is the quantitative investigation of the basketball tournament of the FIBA World Cup 2019. Firstly, it identified the performance of a team by using Principal Components Analysis (PCA). Then, the contributions of shooting, rebounding, turnover, and free-throw factors are identified and compared with Offense vs. Defense in terms of their contribution to the team's performance. Moreover, other factors are identified that affected the performance, the teams which performed better than expected are detected, and finally, machine learning models which enhance the 'Power Rankings' for the prediction of the final position of the teams in the tournament are suggested.

Keywords: basketball analytics; team performance; multiple regression; k-means clustering; machine learning models

1. Introduction

It is widely known and accepted that traditional statistics cannot accurately describe some aspects of basketball, and for this reason, an advanced statistics revolution has taken place in research on basketball in order to produce statistics that are more meaningful and useful for the analysis of the game. The advanced statistics for basketball can be found in the works [1,2]. However, these analyses are valid only in a league format, where all teams play with all other teams, and seasons last for a long time. When the situation is a tournament which is a fast-track competition where teams do not face all other teams, these statistics could be misleading. Moreover, the view of this work is "macroscopic", i.e., the aim is to specify factors that can lead to overall (performance-based) success in the tournament and not to winning in a single game. The aim of this paper is to offer a quantitative method of answering questions in a tournament situation, such as the FIBA World Cup, and to be a starting point for analyzing tournaments in other sports. The focus of most previous papers regarding research on basketball has mostly been on league situations and comparisons or factors of discrimination between winning and losing teams. The focus here is on overall tournament performance and not only on single-game winning factors.

Some previously published related works include the work [3] that explored the factors that influenced the performance of the Chinese team in the 15th Men's World Basketball Championship; they found that the team's ability was imbalanced, that a flexible attack strategy was needed in order to increase attacking ability, and that players' mental regulation needed improving greatly. Furthermore, in work [4], the authors for the matches of the Chinese basketball team in the 14th Men's World Basketball Championship analyzed all kinds of causes of the losses and gains in the match and indicated that speediness, agility, precision, and antagonism are the everlasting trends in the world basketball, while in work [5], the authors used regression analysis to examine the influence of certain basketball elements (FIBA standard indicators of performance) on the final result of a basketball game (they considered games from the 13th, 14th, and 15th Men's World Basketball Championships). Additionally, in the paper [6], the authors compared the Chinese team with the

Citation: Katris, C. Investigation of FIBA World Cup 2019: Evidence Using Advanced Statistical Analysis and Quantitative Tools. *Eng. Proc.* **2023**, *39*, 85. <https://doi.org/10.3390/engproc2023039085>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 14 July 2023



Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

six other top teams of the 2006 Men's World Basketball Championship in terms of statistics. They analyzed the gaps and detected the weaknesses of the Chinese team. In work [7], the author determined which basketball performance indicators can discriminate winners from losers using a dataset of 76 matches from the world championships in Spain in 2014, of which the official statistical parameters were downloaded from FIBA. Finally, in work [8], they compared and analyzed differences between the technical styles of the Chinese and American men's basketball teams in the 15th FIBA World Cup.

The explanation of the aims of this paper follows. A crucial term is team performance. The consideration of only the final ranking of the team in the tournament is obviously misleading. Maybe a team is playing very well in all games but has a blackout in a knockout match, and then the rating is unfair for this team. On the other hand, if performance is the extent of victory, then a strong team might be lucky in the draw of the groups and easily win against their first opponents but, when facing another strong team, not be able to cope with the situation. The previous examples have led to the consideration of performance as a multivariate measure, with the target being to extract a single value for the performance of each team. To achieve this, we used principal component analysis (PCA). The next goal of this work was to determine which factors contributed to the performance of a team. The basis for this analysis is the concept of the four factors of Dean Oliver as a standard for determining the winner of a game. Another big debate is whether offense or defense is more important for success in such tournaments. Both questions are answered with the use of multivariate regression. Additionally, we studied the effects of other factors, such as (i) the height of the team, (ii) the age of the team, (iii) the coach's experience with the team, (iv) the players' percentage (pcg.) usage of the ball (or from the first five players), (v) the distance shooting in a team, (vi) the balance in team scoring, and (vii) the efficiency of small players. Multiple regression and the correlation of variables with team performance were the tools for measuring these effects. Another very popular debate is whether a team performed as well as expected in the tournament. In this manuscript, we make an attempt to determine whether team performance is compatible with pre-tournament expectations, which are specified with the help of hierarchical k-means clustering based on variables that were found to affect the performance of teams. Groups of teams were formed according to their pre-tournament characteristics, and post-tournament actual performance was compared with the expected performance of the teams. A final question that was studied is whether we can have better pre-tournament predictions than power rankings. We employed machine learning models (Random Forests and neural networks) for the prediction of the final positions (based on performance) of the teams in the tournament. Power rankings incorporate information and knowledge from experts that should not be wasted, and this is the reason why they were considered among the inputs in our models. Moreover, they were the benchmark for our models, i.e., we were interested in whether a model could enhance power rankings, and if so, then the model was considered useful. The models were compared in terms of correlation (through the pseudo-R-square measure) with performance-based final positions.

A brief overview of the problems that are studied in this work is the following: the measurement of the performance of a team in a tournament, the detection of which factors played an important role in the performance of a team, whether a team fulfilled expectations in the tournament, and, finally, the suggestion of an improvement in 'Power Rankings'.

The rest of the paper is as follows: Section 2 presents the definitions and meanings of the statistical measures and the statistical methods that were used to tackle the problems in this work. Section 3 is an overview of the questions and problems we considered for the tournament and a detailed description of the procedures we used to deal with them. Section 4 presents the data analysis, and Section 5 contains the summary and the conclusions of the paper.

2. Statistical Definitions, Measures, and Tools

In this section are briefly presented the elements which are used throughout this work. The Principal Component Analysis (PCA) method is a statistical method that was introduced by Pearson and later independently developed and named by Hotelling, and the aim is to express multivariate data with fewer dimensions. A detailed analysis of this method can be found in the book [9].

The correlation of 2 variables can be measured using a coefficient that quantifies this correlation (the value of the coefficient is between -1 and 1 , the magnitude displays the strength of the correlation while the sign displays the direction of the correlation). In this work, we use two such coefficients for completeness: Pearson correlation (r) (details about this method can be found in [10]) and the Spearman rank correlation coefficient (ρ) (details about this method can be found in [11]).

In statistics, linear regression is a linear approach of the form $y = Xb + \epsilon$, which is used to model the relationship between a (dependent) variable and one or more explanatory (independent) variables. Details about linear regression can be found in many statistics books, such as [12]. It is known that factors that affect the outcome of a game are the shooting factor, turnover factor, rebounding factor, and free throw factor, and they are introduced and described in the works [1,2]. Their formulas are mentioned briefly: The shooting factor (*Sh.F.*) formula for both offense and defense is $(FG + 0.5 \times 3P)/FGA$. The turnover factor (*To.F.*) formula for both offense and defense is $TOV/(FGA + 0.44 \times FTA + TOV)$. The rebounding factor (*Reb.F.*) formula for offense is $ORB/(ORB + Opp\ DRB)$, while the formula for defense is $DRB/(Opp\ ORB + DRB)$. The free throw factor (*FT.F.*) formula for both offense and defense is FT/FGA . Possessions of a team are computed through the formula: $FGA + 0.475 \times FTA - ORB + TO$. The possessions are calculated for both the offensive and defensive teams, and the average is considered to decide a game's overall possessions.

Random Forests (are described in [13], were introduced in [14], and each node is split using the best split among a subset of predictors randomly chosen at that node. The output is the mean of all trees for regression. This strategy performs very well against other classifiers and is robust against overfitting. Neural networks are computing systems that are inspired by biological neural networks that constitute animal brains. An overview of neural networks can be found in reference [15]. K-means clustering ([16]) is a popular method for cluster analysis in data mining. In this work, we use the method of Hierarchical k-means clustering ([17]), and the method is implemented in the R package 'factoextra' ([18]).

3. FIBA World Cup 2019: Problems and Procedures to Solve Them

To be decided the success of a team in tournament competition, they are used some metrics. Because the most important definition is the performance of a team, the Winning percentage is naturally the first used metric. However, in the case of a tournament is not a suitable metric because teams do not face all other teams (only a subset of them after a draw). Another measure of the performance of a team is the point difference (PD) between the team and its opponents (this metric displays the dominance of the team). Another metric of success of a team could be the final ranking of the team in the tournament. This metric is also inappropriate. In order to achieve a complete metric of the success of a team, we consider all the above metrics, and we derive an overall metric of success (team score) with the use of the concept of Principal Component Analysis (which explains a large portion of variance).

Furthermore, in this work, it is specified whether the four factors (*Sh.F.*, *TO.F.*, *Reb.F.*, and *FT.F.*) affect the overall performance of the team. To achieve this, we use a multiple regression model (Model 1) with these factors as independent variables and performance as dependent variable. The factors for each team are calculated based on team statistics per game (were extracted from the site of basketball reference—<https://www.basketball-reference.com/international/fiba-world-cup/2019.html> accessed on 1 March 2023).

Additionally, this work replies to another very interesting question, which is whether offence or defense played the most important role in the performance of a team in the

tournament. To answer this question, are formulated, applied, and compared two multiple regression models (Models 2 and 3).

$$Y = a + b_1(Sh.F_{offense}) + b_2(Sh.F_{defense}) + b_3(To.F_{offense}) + b_4(To.F_{defense}) + b_5(Reb.F_{offense}) + b_6(Reb.F_{defense}) + b_7(FT.F_{offense}) + b_8(FT.F_{defense}) + \epsilon \quad (\text{Model 1})$$

$$Y = a + b_1(Sh.F_{offense}) + b_2(To.F_{offense}) + b_3(Reb.F_{offense}) + b_4(FT.F_{offense}) + \epsilon \quad (\text{Model 2})$$

$$Y = a + b_1(Sh.F_{defense}) + b_2(To.F_{defense}) + b_3(Reb.F_{defense}) + b_4(FT.F_{defense}) + \epsilon \quad (\text{Model 3})$$

Furthermore, many effects are tested for their effect on the performance of a team in the tournament. Firstly, are tested the effects which are related to player usage percentage (usg%). The formula of the concept of usage percentage (usg%) is the following: $usg\% = 100 \times ((FGA + 0.44 \times FTA + TOV) \times (Tm MP/5)) / (MP \times (Tm FGA + 0.44 \times Tm FTA + Tm TOV))$. The usage percentage (usg%) is an estimate of the percentage (pcg.) of the team's offensive attempts (plays), which are used by a player while he is on the floor.

Except for the usg%, we consider the position of the player with the greatest usg% in the team (or the avg. position of the five players with the greatest usg%), the played minutes of the player with the greatest usg% in the team (or the avg. played minutes of the five players with the greatest usg%) and the percentage of plays of the player with the greatest usg% in the team (or the avg. percentage of plays of the five players with the greatest usg%). The effect of the player with the greatest usg% (or of five players with the greatest usg%) is tested with multiple regression models (Models 4 and 5, respectively).

$$Y = a + b_1(usg\% \text{ of first Player}) + b_2(\text{Position of first Player}) + b_3(\text{Minutes of first Player}) + b_4(\% \text{ of Plays of first Player}) + \epsilon \quad (\text{Model 4})$$

$$Y = a + b_1(\text{Avg. usg\% of first 5 Players}) + b_2(\text{Avg Position of first 5 Players}) + b_3(\text{Avg. Minutes of first 5 Players}) + b_4(\% \text{ of Plays of first 5 players}) \quad (\text{Model 5})$$

Next, there is tested if the players who are competing in a specific league (League Effect) can affect the performance of a team in the competition. The most important leagues (and their weights for building an overall League Effect score) is an ad-hoc decision. There are considered players who play in the NBA, the Euroleague, the Eurocup, the Basketball Champions League (BCL), and the NCAA. In this work, the scores for the leagues are respectively 1, 1, 0.5, 0.5 and 0.5. Other effects which are tested include whether they affect the performance of a team, the heights of players of the team (this is measured by the average height of the players of the team and by the number of players in a team with a height over 200 cm.), the ages of the players of the team (this effect is measured by the average age of the players of the team and by the number of players in the team with age over 30 years old), the coach experience to the bench of the team (in Years), and the importance of shooting (this is measured by the percentage of 3 point attempts over the overall attempts and by the points scored from players who plays in the positions 1, 2, and 3 (small players) versus the points scored from players who plays in the positions 4 and 5 (high players)). These effects are tested with regression models (Models 6–10).

$$Y = a + b_1(\text{League Effects} + \epsilon) \quad (\text{Model 6})$$

$$Y = a + b_1(\text{Avg. Height}) + b_2(\text{number of players with height over 200 cm}) + \epsilon \quad (\text{Model 7})$$

$$Y = a + b_1(\text{Avg. Age}) + b_2(\text{Number of players with age over 30 yearsold}) + \epsilon \quad (\text{Model 8})$$

$$Y = a + b_1(\text{Coach experience in the team (in Years)}) + \varepsilon \quad (\text{Model 9})$$

$$Y = a + b_1(\% \text{ of 3pt Attempts}) + b_2(\text{Pts of small.vs. Pts of high players}) + \varepsilon \quad (\text{Model 10})$$

Moreover, two formulas are defined:

$$(i) \quad \text{Efficiency of Small Players} = \frac{\text{Pts from Small Players (PG,SG,SF)}}{\text{Pts from Tall Players (PF,C)}} - \frac{\%usg.from Small Players}{\%usg.from Tall Players}$$

$$(ii) \quad \text{Balance} = \frac{\text{Pts from Small Players (PG,SG,SF)}}{\text{Pts from Tall Players (PF,C)}} + \frac{\%usg.from Small Players}{\%usg.from Tall Players} - 2.$$

Additionally, the intention is to check whether the team pace (tempo) affects the performance, i.e., faster or slower teams found to perform well (this is measured by the number of possessions of a team per game of the competition). These additional effects are tested with regression models (Models 11–13) and with Spearman and Pearson correlations.

$$Y = a + b_1(\text{Efficiency of Small Players}) + \varepsilon \quad (\text{Model 11})$$

$$Y = a + b_1(\text{Balance}) + \varepsilon \quad (\text{Model 12})$$

$$Y = a + b_1(\text{Possesions per Game}) + \varepsilon \quad (\text{Model 13})$$

4. Data Analysis

This section contains the data analysis of the tournament and the conclusions from this analysis. The fundamental concept here is the calculation of the total score for each team which represents its performance. The tables with the data are in the link https://docs.google.com/document/d/1QMERMfeckZNY9LZT1BCHdbl7UK6tctNc/edit?usp=share_link&ouid=118393132040933122489&rtfpof=true&sd=true, (accessed on 14 July 2023). The conclusions from the analyses which were implemented are presented here. All methods in Sections 4.1–4.3 were implemented in the statistical software R (Version 3.6.0).

4.1. Calculation of Team Performance and Descriptive Analysis

At first, we calculate the team scores using PCA. Table S1 displays both the score of each team and the ranking of the team according to this score. We consider the success of each team, the percentage of wins in its games, the point difference on average in its games, and the final position of the team in the ranking of the FIBA World Cup 2019 (we consider the value 4/Final Position). The goal is to achieve an overall score for each team. This score reflects the overall performance and takes into account all three aforementioned dimensions of the performance. The first component, after the application of the PCA method, explains almost 80% of the variance of the three initial variables, so we consider it as a measure of performance. The variables were rescaled before the analysis in order to have unit variance.

4.2. Analysis of the Four Factors in the Performance of a Team

In this subsection, at first, we calculate the offensive and the defensive factors, and then we run a multiple regression with these factors as independent variables and Team Score as the dependent variable. The regression shows that these factors explain approximately 80% of team performance (R^2 values). The signs of the factors are expected for all factors. According to the p-value, we found the following: the shooting factor (offensive and defensive) affected the score of each team, the free-throw factor (in defense) and the turnover factor (in defense) were found to affect the score of each team at the 5% level. The

other factors were found to be statistically insignificant. Moreover, we attempt to compare offense and defense. The procedure is following:

- *Offense vs. Defense:* We consider only the offensive factors and their contribution to the team's success, then only the defensive factors and their contribution to the team's success, and then we compare the contribution of the factors to the explanation of success through multiple regression models. From the above models, we conclude that the defensive functioning was found to be more influential than the offensive functioning according to R^2 values (while both were found to be statistically significant).

4.3. Exploration of Effects

In this subsection, many effects were tested through regression analysis and correlation measurement using the team performance (team score) as the dependent variable.

4.3.1. Effect of the Player (of Five Players) with the Greatest Usage

There are considered the five players (who play more than 10 min.) with the greatest usage in the game. They are taken into account and tested the next variables for these players: the average usage, the average position (i.e., PG = 1, SG = 2, SF = 3, PF = 4, C = 5), the average minutes, and the percentage of plays. There are considered the five players (who play more than 10 min.) with the greatest usage in the game. They are taken into account and tested the next variables for these players: the average usage, the average position (i.e., PG = 1, SG = 2, SF = 3, PF = 4, C = 5), the average minutes, and the percentage of plays. Table S7 displays the correlation results, and Table S8 displays the results of the regression of each factor with team performance (team score). There is no significant correlation between any of the factors with team performance, and the regression is not significant, so there is no significant effect of the usage, position, or minutes on the team score. The same applies to the effect of the leader on the team's performance. Tables S9 and S10 display the results.

4.3.2. League Effects

The goal of this subsection is to study the effect of the players who compete in more competitive leagues on the performance of the team in the tournament. We consider the number of players who compete in the NBA, the Euroleague, the Eurocup, and on BCL, and the NCAA. Table S11 displays the correlation, and Table S12 the regression of these factors with the team score. There was found to be a significant correlation between team performance and the number of NBA players and between team performance and the number of Euroleague players in a team. The regression was found to be significant, too (F-value). Specifically, the number of NBA players and the number of Euroleague players affect the performance significantly, while the number of Eurocup players, the number of BCL players, and the number of NCAA players do not seem to affect the team's performance. Furthermore, we derive a value for each team using the following formula: *Top-League Effect* = No. of NBA players + No. of Euroleague players + $0.5 \times$ No. Eurocup players + $0.5 \times$ No. of BCL players + $0.5 \times$ No. of NCAA players. The Regression is significant, and this effect explains over 45% of team performance (Table S13).

4.3.3. Height Impact, Age Impact, and Coach Experience in the Team

The first goal of this subsection is the study of the effect of the height of available players of a team. A common question is if increased height leads to increased chances of winning. We consider as variables: (i) the average (avg.) height of players and (ii) the number (no.) of players with height over 200 cm. Table S14 displays the correlation between the variables, and Tables S15 and S16 display the results of the regression of each factor with team performance (team score). It was found significant correlation and regression between average height of the team and team performance.

Furthermore, it studied the effect of the age of available players of a team. Another common belief is that increased age leads to decreased performance. This is maybe rational

for leagues, but is this true for a tournament? We consider (i) the average (avg.) age of the players of the team and (ii) the number (no.) of the players of the team with ages over 30 years old. There were not found significant correlations and regressions between the considered variables and the performance of the team (see Tables S17–S19).

Finally, it studied the effect of coach experience on the performance of a team. A common belief is that the long-term incorporation of a coach into a team leads to increased chemistry, thus, performance. We consider, as a variable, the years (yrs.) of the coach in the team. There is a significant correlation and regression between “Coach experience” and “team performance” (see Tables S20 and S21).

4.3.4. Existence of Shooters and Other Effects

In this subsection, the first goal is to study the effect of the existence of shooters inside the roster of a team on its performance. A common belief is that the existence of many shooters inside the roster of a team leads to an increased offensive threat, thus to an improved performance of the team. In order to test this fact, we consider as variables: (i) the percentage (%) of 3 pt attempts and (ii) the points scored from the players who play in the positions of PG, SG, and SF (‘small’ players) divided with the points scored from the players who play in the positions of PF and C (‘high’ players).

Furthermore, for the effects of (i) the team balance, (ii) the efficiency of small players’ and (iii) the pace of the team in its performance, it was found that there were no significant correlations and regressions.

4.4. Detect Surprises and Upsets—Clustering (with Hierarchical k-Means)

The goal of this subsection is to suggest a procedure to detect the positive and negative surprises of the tournament. At the core of this procedure is the generation of clusters of teams. From the previous analysis (in Section 4.3) there are considering the variables which were found to significantly affect the performance of a team (these are the Coach’s Experience in the team, the Average Height of the team, and the number of players who compete within Top Leagues) and the Power-Rankings before the tournament. Figure S3 presents the elbow method (a graph of the total within the sum of squares (WSS) which is explained by the increase of clusters). The decision about the number of clusters is a number for which the addition of an additional cluster is not improving much the total WSS, which is explained. In the case of the FIBA World Cup 2019 tournament, the generation of 3 clusters of teams is the decision. The first cluster represents the strongest teams (S), the second cluster represents the 2nd tier teams (M), and the third cluster represents the weakest teams (W) of the tournament. From Table S28 and Figure S4, are observed some notable facts: (i) stronger teams display higher offensive and defensive efficiencies, and (ii) we observe that Higher Pace is not a characteristic of stronger teams necessarily and rather displays the style of play of each team (see Figure S2). However, the weakest teams play, on average, on a lower tempo. Finally, we consider ranking according to clustering as expected from the team characteristics beforehand; this ranking is compared with the actual ranking of the tournament, and the procedure for the detection of surprises is implemented (see Table S29 for surprise detection).

4.5. Improve Power Rankings—Predict Team Position

The goal of this subsection is the improvement of the virtual ranking of the teams before the tournament. The accuracy of such ranking is very important for betting reasons and for coaches and fans who can adjust their expectations about the performance of the team they support. The main tool for such a ranking is the so-called ‘Power Rankings’, which are released before the tournament and take into account all the relevant information. The improvement of ‘Power Rankings’ is the aim of this subsection. The main idea is the use of ‘Power Rankings’ as input to other models. To achieve such an improvement, we use the machine learning methods of Random Forests and Feed-Forward Neural Networks (ANN). The analysis is performed with the use of the package Rattle [19] of the R statistical

software. The parameters of the models are for Random Forests 500 trees and two variables and for ANN 1 hidden layer with 10 nodes. Input variables are: (i) the 'Power Rankings' (which are released officially by FIBA) and (ii) the significant variables from Section 4.3 (the Coach Experience in the Team, the Average Height of the team, and the number of players of the team who compete in Top-Leagues), whilst the output of the models is the team ranking. The crucial part is the evaluation of the models. At first, we build the models, i.e., we consider the entire dataset in order to train the models, and the training performance of them is measured using as metric the pseudo-R-square. Secondly is evaluated the predictive accuracy of the models with the following procedure: there are considered 10 teams randomly, and then their ranking is predicted based on the model. The experiment is repeated 10,000 times. The evaluation is performed using the metric of pseudo-R-square (the mean, the standard deviation and a 90% confidence interval are calculated). We observe that the training accuracy of the models is greater than that of Power Rankings, and Random Forest is the preferred approach (see Table S31).

5. Summary and Conclusions

In this work was made a quantitative analysis of the FIBA World Cup 2019. Firstly, it was determined as a metric of the performance of a team in a tournament. Furthermore, was studied the importance of the four factors in the performance of the team and was answered the question of offense vs. defense. Moreover, the coach's experience in the team, the average height of the team, and the number of players who compete in top leagues are found to affect the performance of a team. Next, a procedure was presented, which was based on clustering in order to detect 'surprises' in the tournament. Finally, a procedure with the aim of improving Power Rankings through machine learning methods was suggested. This work can serve as a source of thought for tournament analysis in basketball and other sports.

The Data Analysis of this paper can be found in the following link: https://docs.google.com/document/d/1QMERMfeckZNY9LZT1BCHdbl7UK6tctNc/edit?usp=share_link&oid=118393132040933122489&rtopf=true&sd=true, (accessed on 14 July 2023).

Supplementary Materials: The data used for this paper and all the relevant analysis (tables and figures) are provided with the paper in a link as supplementary material of the paper (https://docs.google.com/document/d/1QMERMfeckZNY9LZT1BCHdbl7UK6tctNc/edit?usp=share_link&oid=118393132040933122489&rtopf=true&sd=true).

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used for this paper and all the relevant analysis (tables and figures) are provided with the paper in a link as an additional file as Supplementary Material of the paper.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Kubatko, J.; Oliver, D.; Pelton, K.; Rosenbaum, D.T. A starting point for analyzing basketball statistics. *J. Quant. Anal. Sport.* **2007**, *3*, 1. [CrossRef]
2. Oliver, D. *Basketball on Paper: Rules and Tools for Performance Analysis*; Potomac Books, Inc.: Sterling, VA, USA, 2004.
3. Wu, M.; Yan, J. Statistic Analysis on Chinese Men's Basketball Team in the 15th World Basketball Championship. *J. Chengdu Sport Univ.* **2007**, *3*, 63–66.
4. Zhang, H.; Pan, S. China the Development Strategies for China's Man Basketball Studied with the Case of the 14th World Man's Basketball Championship. *Sport. Sci.* **2002**, *6*, 42–44.
5. Simović, S.; Komić, J. Analysis of influence of certain elements of basketball game on final result based on differetiant at the XIII, XIV and XV World Championship. *Acta Kinesiol.* **2008**, *2*, 57–65.
6. Wang, D.; Zhou, Y. Analysis of the Gaps in Main Techniques between Chinese and European Teams in 2006 FIBA World Championship. *J. Mianyang Norm. Univ.* **2008**, *5*, 125–129.

7. Čaušević, D. Game-related statistics that discriminate winning and losing teams from the world championships in Spain in 2014. *Homo Sport*. **2015**, *17*, 16–19.
8. Zhang, Z.D.; Sun, Z.J. Analysis of different technical styles between China and American Men's Basketball in the 15th FIBA World Championship. *J. Nanyang Norm. Univ.* **2006**, *12*, 79–82.
9. Jolliffe, I. *Principal Component Analysis*; Springer: Berlin Heidelberg, Germany, 2011; pp. 1094–1096.
10. Sharma, A.K. *Text Book of Correlations and Regression*; Discovery Publishing House: Delhi, India, 2005.
11. Myers, L.; Sirois, M.J. Spearman correlation coefficients, differences between. In *Encyclopedia of Statistical Sciences*; John Wiley & Sons: Hoboken, NJ, USA, 2004; Volume 12.
12. Seber, G.A.; Lee, A.J. *Linear Regression Analysis (Vol. 329)*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
13. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
14. Breiman, L. Random forests. *Mach. Learning*. **2001**, *45*, 5–32. [CrossRef]
15. Macukow, B. Neural networks—state of art, brief history, basic models and architecture. In Proceedings of the IFIP International Conference on Computer Information Systems and Industrial Management, Vilnius, Lithuania, 14–16 September 2016; Springer: Cham, Switzerland; pp. 3–14.
16. Celebi, M.E. (Ed.) *Partitional Clustering Algorithms*; Springer: Berlin/Heidelberg, Germany, 2014.
17. Xu, T.S.; Chiang, H.D.; Liu, G.Y.; Tan, C.W. Hierarchical K-means method for clustering large-scale advanced metering infrastructure data. *IEEE Trans. Power Deliv.* **2015**, *32*, 609–616. [CrossRef]
18. Kassambara, A.; Mundt, F. Package 'factoextra'. Extract and visualize the results of multivariate data analyses. Available online: <https://www.rdocumentation.org/packages/factoextra/versions/1.0.7> (accessed on 1 March 2023).
19. Williams, G.J. Rattle: A data mining GUI for R. *R J.* **2009**, *1*, 45–55. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Impact of Migration Processes on GDP [†]

Olena Rayevnyeva ¹, Kostyantyn Stryzhychenko ^{2,*} and Silvia Matúšová ¹

¹ Department of Economics and Finance, Bratislava University of Economics and Management, Furdekova 16, 85104 Bratislava, Slovakia; olena.raev@gmail.com (O.R.); silvia.matusova@vsemba.sk (S.M.)

² Department of Economics and Finance, Simon Kuznets Kharkiv National University of Economics, av. Nauki 9-a, 61111 Kharkiv, Ukraine

* Correspondence: ukf.kendo@gmail.com

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: The globalization process and the war in Ukraine show us that migration is one of the strongest global trends in the modern economy. For this paper, we determined three types of migration, depending on the intention of the people involved, these being labor, educational, and refugee migration. Each type has a different influence on the macroeconomic process. However, in this paper, we investigate the influence of general migration on GDP. We analyze five factors that have major influences on GDP, namely, migration (I), interest rate (IR), active population (AP), export (E), and the consumer price index (CPI). For the purposes of this paper, vector autoregressive models (VAR models) were chosen to perform the analysis. We used the Granger causality test to investigate the lag structure and identified the exogenous variables in the VAR model, such as GDP, migration, and the active population. We investigated the cross-influence between these factors and found that migration has a negative effect on the active population and a positive effect on GDP, while GDP growth leads to a decrease in migration. The Akaike and Schwartz criteria showed the high quality of the VAR models. The impulse analysis of shock influences identifies the structure of the reaction seen in GDP and migration, depending on their shock factors. Using decomposition analysis, we found that migration and GDP influence each other by 10–14%, which can improve the forecasting of these factors and the study of structural migration by the use of these three types.

Keywords: migration; GDP; VAR-model; impact

Citation: Rayevnyeva, O.;

Stryzhychenko, K.; Matúšová, S.

Impact of Migration Processes on GDP. *Eng. Proc.* **2023**, *39*, 86. <https://doi.org/10.3390/engproc2023039086>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 14 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction to the Migration Process

1.1. Analysis of References to the Current Migration Processes in Europe and the Impact of Migration on GDP

The end of the 20th and the beginning of the 21st century are characterized by the significant transformation of the international market space, which can be recognized by its effects of globalization and integration. The world market has a key influence on the processes that take place in national markets. It defines the competitive space factors within which national enterprises operate. The openness of national markets, in turn, significantly changes the quality of migration processes. Under these open conditions, workers looking for better employment conditions outside national enterprises can move with few or no problems, in search of better countries in which to work. The quality of educational migration also changes significantly when students endeavor to find the best combination of cost and standard of education for their needs. These processes were quite transparent in the European countries, where the creation of the European Union resulted in new conditions for the various types of migration, in order to benefit both the population and the EU member states. However, the military conflict between Russia and Ukraine in 2022 brought a new round of forced migration to pass, which has made its own adjustments to all types of migration.

Millions of Ukrainians, mostly women, were forced to leave their country and migrate to EU countries with their families, bringing with them children who study not only in primary and elementary schools but also in universities. This assumes an a priori increase in labor migration by the educated population and educational migration. In addition, according to the estimates made by the Ptoukha Institute for Demography and Social Studies of the National Academy of Sciences of Ukraine [1], from the two largest educational centers of Ukraine alone, Kyiv and Kharkiv, 70% of women with higher education qualifications left Ukraine for the European Union.

According to Josep Borrell, the High Representative of the European Union for Foreign Affairs and Security Policy, and Vice-President of the European Commission, migration is currently being discussed in Europe, primarily as a challenge. At the end of 2022, according to Frontex, the number of migrants arriving in Europe via the Central Mediterranean or Western Balkan routes had increased again in 2022 by 51% and 136%, respectively [2].

In particular, Russia's war against Ukraine, which began in February 2022, triggered the largest displacement of people seen in Europe since World War II, while only 4 million Ukrainians received temporary protection [3]. According to the Frontex-European Border and Coast Guard Agency, in 2022, about 15 million Ukrainians came to Europe, of whom 4 million Ukrainians received temporary protection and approximately 3 million wished to remain in the European Union [3]. That is, one-fifth of the people who have completed higher education, as well as those receiving higher education in the future, will contribute to an increase in the GDP of the host country. In this regard, the task of determining the impact on the GDP of individual factors, including the migration component, is of interest.

This research is aimed at developing tools for assessing and modeling the impact of socioeconomic factors on the country's GDP.

A study assessing and identifying the impact made on the volume of GDP by various socio-economic factors showed that modern authors distinguish the following as the dominant factors [4–9]: migration (I), interest rate (IR), active population (AP), export (E), and the consumer price index (CPI). Based on the hypothesis that these factors have both a direct and a lag effect on the volume of GDP, this paper proposes to use vector autoregressive models to assess the various impacts.

1.2. Literature Review

Gross domestic product (GDP) is one of the determinants of a country's economic growth. This is why the task of studying the factors that have a diverse influence on its change is always relevant. The need for constant study of such factors is also explained by the fact that this task is permanent. This means that any change of situation in both the world and the national markets, which are associated with the evolution of the development of the economy and society, changes the degree of influence of these socioeconomic factors on the country's GDP.

One of the factors that is gaining more and more influence on each country's GDP is migration. The impact of the migration process on a country's GDP has long been the focus of international research. Thus, a report by the United Nations Development Program (UNDP), published on 21 October 2020 and named "Refugees and Migrants", analyzed the main trends in the field of migration [10]. The UNDP chief, Achim Steiner, emphasized the fact that migrants play an important role in economic recovery, especially after a crisis: although constituting only 3.5 percent of the world's population, according to data from 2015, migrants produce 9 percent of global GDP. According to studies by the International Monetary Fund and the World Bank, an increase of 3 percent of immigrants in developed countries would increase their global GDP by USD 356 billion by 2025.

The influence of migration on the main macroeconomic indicators of the development of countries has been studied by many authors. For example, Heinisch, K., and Wohlrabe, K., (2016) emphasize the power of migration's impact on macroeconomic indicators, showing that it is necessary to analyze the structure of the economically active population in terms of particular refugees with different levels of education [11].

Kudaeva, M., and Redozubov, I. (2021) prove that there is a strong relationship between GDP and migration. Thus, based on an analysis of the impulse responses of the SVAR model, it was determined that a 1% shock to the migration process increases the real GDP by 0.1% [12].

The works of many authors are devoted to the analysis and assessment of the influence of various socio-economic factors on changes in GDP.

For example, Zhuravskaya, K.G. (2016) analyzes the impact of the M2 monetary aggregate, international reserves, consumer price inflation, domestic lending to the private sector by banks, the general tax rate, the discount rate, the population, the dollar exchange rate in the national currency, and the market capitalization of companies where their shares are listed on the stock exchange on the GDP of countries with varying levels of economic development [13]. On the basis of the author's studies, the factors and the strength of their influence on the level of GDP were analyzed, and cross-country differences in the process of GDP formation were identified.

Alex Reuben Kira (2013), working against the background of a study of the dynamics of change in Tanzania's GDP and by using a Keynesian model, shows a significant impact on this macroeconomic indicator of consumption, namely, the government's final expenditure and household final expenditure, along with exports [14].

Hongbo Guo and Zewei Zhang (2022) identify the main factors influencing GDP, comprising: gross saving; the consumer price index; unemployment; population and the real interest rate [15]. Based on the results of their regression analysis, the authors prove the existence of a significant influence of exogenous factors on GDP and emphasize the need to develop an appropriate state policy that will maintain the stability of the development of these factors.

Artur Ribaj and Fitim Mexhuani (2021) also prove the existence of a strong correlation between GDP and the gross saving factor. Based on augmented Dickey–Fuller tests, Johansen cointegration tests, and the Granger causality test, the authors determined that saving stimulates investment, production, and employment, which leads to overall economic growth [16].

Based on a study of a period of economic liberalization in Ghana, Emmanuel Nketiah, Xiang Cai, Mavis Adjei, and Bekoe Bernard Boamah (2020) proved the existence of a close relationship between foreign direct investment, trade openness, and GDP [17].

1.3. Three Types of Migration

Analysis of the migration process and of the literary sources [18–20] can help us to identify three types of migration, depending on their impact on the national economy:

- (i) Labor migration;
- (ii) Educational migration;
- (iii) Refugee migration.

Each type of migration has its own impact on GDP. The first type receives GDP in the machine-building branches of the economy, while the second type affects scientific and technological areas of the economy; the third type results in pressure on the economy regarding social support for refugees.

2. Mathematical Tools for Studying the Impact of Migration on GDP

2.1. Methodological Aspects of Model-Building

For this paper, we studied the process of the impact of migration on GDP from the point of view of system analysis. In this regard, we propose the following algorithm for studying the impact of the migration process on GDP (see Figure 1).

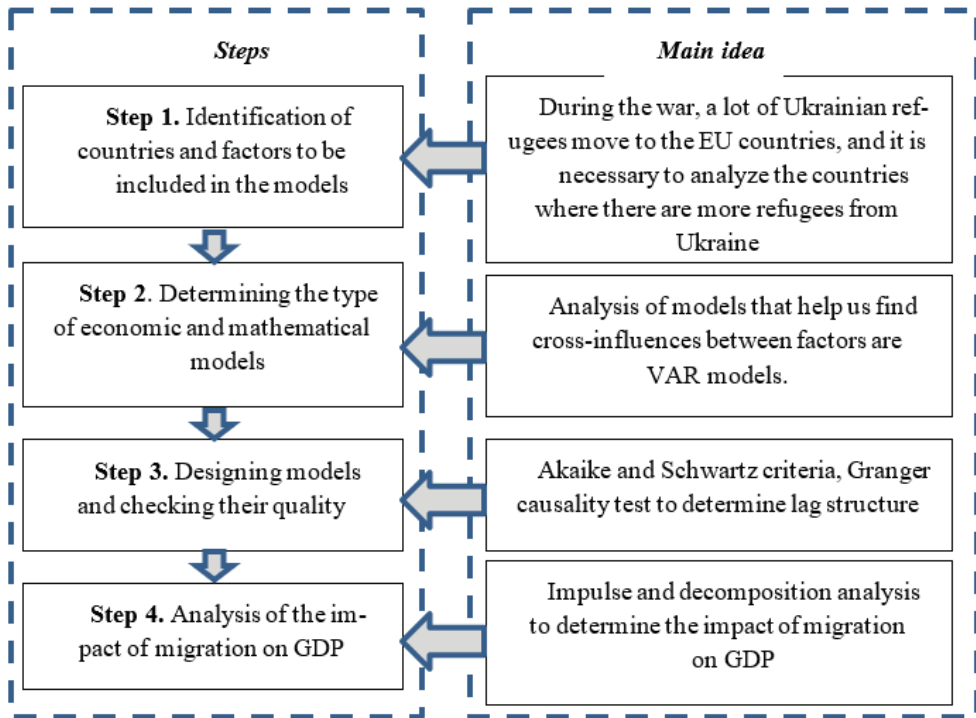


Figure 1. Algorithm for studying the impact of migration on GDP.

2.2. Mathematical Models of Migration's Influence on GDP

Migration has a lag influence on numerous economic processes that is based on the nature of migration. In this case, we used vector auto-regression models (VAR models) to perform the analysis, as follows:

- (i) General VAR model and determination of the migration impact lag structure.

The general VAR (p) model, with n variables, is represented by:

$$X_{1,t} = a_{10} + a_{11}X_{1,t-1} + \dots + a_{1p}X_{1,t-p} + \dots + b_{1p}X_{n,t-p}$$

$$X_{2,t} = a_{20} + a_{21}X_{1,t-1} + \dots + a_{2p}X_{1,t-p} + \dots + b_{2p}X_{n,t-p}$$

...

$$X_{n,t} = a_{n0} + a_{n1}X_{1,t-1} + \dots + a_{np}X_{1,t-p} + \dots + b_{np}X_{n,t-p}$$

where n is the number of variables; p is the optimal lag of the VAR model.

- (ii) Estimation of the lag's influence in the VAR model. Generally, two main criteria exist for the determination of the lag's influence; these are the Akaike criteria and the Schwarz criteria. We also used a Granger causality test for determining the lag structure.
- (iii) Impact impulse analysis. We used momentum analysis to determine the percentage impact of each factor on GDP. Decomposition analysis will help us to determine the part of the variance that depends on the changing pattern of the exogenous factors.

3. Model Design and Impact

3.1. Estimation of the Model's Parameters

The model was calculated on data from the Polish economy. This choice is due to the fact that Poland is considered the most attractive country for relocation in the European Union by Ukrainian migrants, which finding is based on a retrospective analysis of the data for 1990–2021 [21–24]. In addition, during the recent period of Russian aggression in Ukraine, Poland received about 6 million Ukrainian refugees. In our study, we assessed the following factors:

GDP per capita (current rate in USD) is the gross domestic product, divided by the midyear population—variable “GDP”;

Exports as the capacity to import (constant LCU) equals the current price value of the export of goods and services deflated by the import price index – variable “E”;

Net migration (quantity of persons) represents the net total of migrants during a particular period, calculated as the number of immigrants minus the number of emigrants, including both citizens and noncitizens—variable “I”;

Interest rate (%)—variable “IR”;

Population aged 15–64 (total)—variable “AP”;

Consumer price index (%)—variable “CPI”.

To conduct the analysis, we first investigated the stationary value of the time series using the ADF test. The results are shown in Table 1.

Table 1. ADF test of the time series.

Series	t-Stat	Prob.
DLOG_AP	−9.5662	0.0000
DLOG_CPI	−9.1768	0.0000
DLOG_E	−6.1593	0.0000
DLOG_GDP	−4.9360	0.0004
DLOG_I	−6.8486	0.0000
DLOG_IR	−3.6155	0.0532
LOG_AP	−1.9639	0.3001
LOG_CPI	−2.2329	0.1995
LOG_E	−1.0199	0.7307
LOG_GDP	−2.1610	0.2237
LOG_I	−1.7057	0.4186
LOG_IR	0.7998	0.9920

We used the first differences to calculate the ADF test. The test showed us that the first time series differences are stationary; thus, we could use this characteristic in the construction of the VAR models.

In our study, we adopted the hypothesis that five factors influence GDP. To explore this hypothesis regarding exogenous and endogenous factors, the Granger causality test is used in the model. The results are presented in Table 2.

Table 2. Exogenous and endogenous factors in the model.

DLOG_AP-Chi-sq	24.274920	0.0009
DLOG_CPI-Chi-sq	11.16410	0.3449
DLOG_E-Chi-sq	14.51292	0.1509
DLOG_GDP-Chi-sq	24.51292	0.0015
DLOG_I-Chi-sq	26.725775	0.0007
DLOG_IR-Chi-sq	5.386847	0.8639

The data in Table 2 show that three variables are likely to be exogenous in the model. These are GDP, migration, and the active population. Hence, we recalculated the VAR models. The resulting parameters of the VAR model are presented in Table 3.

Table 3. The parameters of the VAR model.

	DLOG_AP	DLOG_GDP	DLOG_I
DLOG_AP(-1)	0.454114	3.512704	-8.771671
DLOG_AP(-2)	0.537753	6.638702	27.09934
DLOG_GDP(-1)	0.008067	0.290427	-3.008223
DLOG_GDP(-2)	-0.004253	-0.233841	0.799000
DLOG_I(-1)	-0.000387	0.012125	-0.301089
DLOG_I(-2)	-0.000335	0.001815	-0.497146
C	-0.000643	-0.022083	-0.689128
DLOG_CPI	0.004353	-0.209005	1.500807
DLOG_E	-0.011102	1.080142	-4.583303
DLOG_IR	-0.000654	-0.018509	0.271095
R-squared	0.891028	0.737911	0.891095
AkaikeAIC	-8.616383	-7.571587	8.443811
SchwarzSC	-8.132499	-7.087704	8.927694
Akaike information criterion		-7.859639	
Schwarz criterion		-6.407989	

The model’s parameters were calculated using the Eviews program.

Measuring the quality of the model shows us that the VAR(2) model is of high quality. These models show the cross-influences between GDP, migration, and the active population. The influence graph is shown in Figure 2.

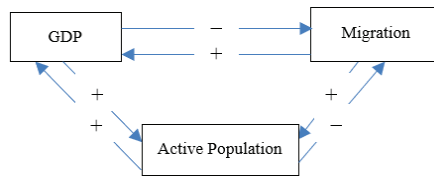


Figure 2. Cross-influences between GDP, migration, and the active population.

Figure 2 demonstrates that migration negatively affects the active population and positively affects GDP, while GDP growth leads to a decrease in migration.

3.2. Impact of the Models

For our analysis of the impact of the models, we used impulse and decomposition analysis techniques. The impulse response function is an important tool for conducting sensitivity analysis of the VAR indicators model regarding the action of external shocks. Figure 3 shows an impulse analysis of the model indicators on GDP and migration.

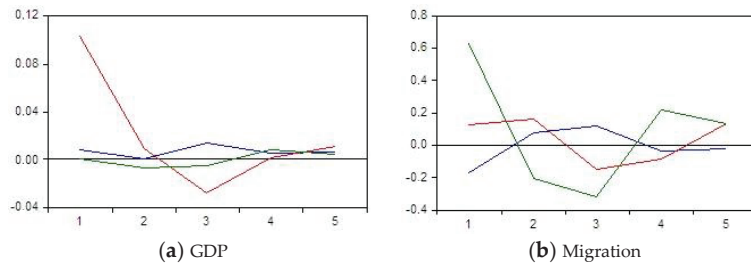


Figure 3. Impulse analysis of GDP and migration.

Figure 3 shows that after four years, migration shocks will trigger a change in GDP to the same extent as GDP shocks. Conversely, a GDP shock has a direct impact on the migration process. That is, if daily life in a particular country worsens, then migration will increase.

The second analysis tool used in this paper is decomposition analysis, as presented in Table 4.

Table 4. Variance decomposition of migration and GDP.

Period	Variance decomposition of Migration			
	S.E.	DLOG_AP	DLOG_GDP	DLOG_I
1	0.002826	6.569453	0.0000046	93.43054
2	0.003244	6.558724	12.37724	81.06403
3	0.003836	7.065201	11.46494	81.46986
4	0.004250	6.569463	14.29596	79.13457
5	0.004671	6.393239	14.94129	78.66547
Period	Variance decomposition of GDP			
	S.E.	DLOG_AP	DLOG_GDP	DLOG_I
1	0.095695	4.641632	85.35837	10.00000
2	0.099833	4.294604	85.00568	10.69971
3	0.102142	4.686875	84.60249	10.71063
4	0.102684	4.739765	84.15611	11.10412
5	0.103043	5.094311	83.78465	11.12104

An analysis of the results that were obtained allows us to conclude that migration has a 10% impact on GDP. Thus, migration has a very significant impact. During all economic periods, this impact is stable. However, GDP only begins to influence migration in the second year, which indicates that the population does not immediately respond to the deterioration of the economy in a particular country. This finding is associated with the process of adaptation by the population in the economic space of Poland and covers all the above types of migration. In general, the impact of GDP on migration is about 11–14%.

It should also be added that since the influence of migration and GDP are of almost equal degree to each other, an interesting chain reaction will be observed. Migration into the country leads to an increase in GDP, while an increase in GDP leads to a decrease in migration out of the country.

4. Conclusions and Future Prospects

The following results were obtained in the current work:

- (i) The migration process is a complex scenario consisting of three components (labor migration, educational migration, and refugee migration) that have a cumulative impact on GDP and can change the structure of the country's economically active population.
- (ii) We have built VAR models that characterize the impact of the following indicators on GDP: migration (I), interest rate (IR), active population (AP), exports (E), and the consumer price index (CPI). The Granger causality test made it possible to find such exogenous factors as GDP, migration, and the active population. This test also shows us that migration negatively affects the active population and positively affects GDP, while GDP growth leads to a decrease in migration.
- (iii) Impulse and decomposition analyses show that migration and GDP have cross-influences of about 10% and 14%. Migration has the most significant and stable impact on GDP. In addition, the population does not immediately respond to the deterioration of the economy in the country, and migration begins to increase with a decrease in GDP in the second year. Based on the above decomposition analysis, it is clear that migration into the country leads to an increase in GDP, and an increase in GDP leads to a decrease in migration out of the country.

Further research will be related to the allocation of the influence of these three components on the country's GDP, to improve the efficiency of migration process management in Europe and Ukraine.

Author Contributions: Conceptualization, O.R., K.S. and S.M.; methodology, O.R. and K.S.; software, K.S.; validation, K.S.; formal analysis, O.R.; investigation, O.R., K.S. and S.M.; resources, O.R., K.S. and S.M.; data curation, K.S.; writing—original draft preparation, O.R., K.S. and S.M.; writing—review and editing, O.R.; visualization, K.S.; supervision, O.R.; project administration, O.R. and S.M.; funding acquisition, O.R. All authors have read and agreed to the published version of the manuscript.

Funding: The reported study was funded by the EU's Next-Generation EU through Recovery and Resilience Plan for Slovakia, under project number 09I03-03-V01-00083.

Informed Consent Statement: Not applicable.

Data Availability Statement: An Official Website of the European Union. Homepage: Available online: https://www.eeas.europa.eu/eeas/migration-key-element-our-foreign-policy_en (accessed on 9 March 2023). Frontex-European Border and Coast Guard Agency Homepage: Available online: <https://frontex.europa.eu/media-centre/news/news-release/frontex-stands-with-ukraine-koMh1h> (accessed on 17 March 2023) Worldwide Immigration Trends Report: Available online: https://www.fragomen.com/trending/worldwide-immigration-trends-reports/index.html?gclid=CjwKCAjw6vyiBhB_EiwAQJRopmEjn5ZEKMBLRGxityEGUwnsobcCGZhu6Gzj1_enquuQ0euA7YnBoChlYQAvD_BwE (accessed on 27 April 2023). World Migration Report 2022. International Organization for Migration: Available online: <https://worldmigrationreport.iom.int/wmr-2022-interactive> (accessed on 27 April 2023). An Official Website of the Eurostat. Homepage: Available online: <https://ec.europa.eu/eurostat> (accessed on 27 March 2023). An Official Website of the State Statistics Service of Ukraine. Homepage: Available online: <https://www.ukrstat.gov.ua/> (accessed on 12 March 2023).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. How Many Ukrainians Are Left? Who Will Not Return? What Will Ukrainian Society and Economy Be Like? Interview with Sociologist Ella Libanova. Available online: <https://forbes.ua/war-in-ukraine/skilki-zalishilos-ukraintsiv-khto-ne-povernetsya-yakim-bude-ukrainske-suspilstvo-ta-ekonomika-intervyu-z-sotsiologineyu-elloyu-libanovoyu-17012023-11104> (accessed on 9 March 2023).
2. An Official Website of the European Union. Homepage. Available online: https://www.eeas.europa.eu/eeas/migration-key-element-our-foreign-policy_en (accessed on 9 March 2023).
3. Frontex-European Border and Coast Guard Agency Homepage. Available online: <https://frontex.europa.eu/media-centre/news/news-release/frontex-stands-with-ukraine-koMh1h> (accessed on 17 March 2023).
4. Brynjolfsson, E.; Diewert, W.E.; Eggers, F.; Fox, K.J.; Gannamaneni, A. The Digital Economy, GDP and Consumer Welfare: Theory and Evidence. 2018. Available online: https://www.oecd.org/naec/Brynjolfsson_MOCE-GDP-B_OECD_2018-07-26.pdf (accessed on 12 March 2023).
5. Martin, F. Underestimating the Real Growth of GDP, Personal Income, and Productivity. *J. Econ. Perspect.* **2017**, *31*, 145–164.
6. Cervellati, M.; Sunde, U. Life Expectancy and Economic Growth: The Role of Demographic Transition. Available online: <https://docs.iza.org/dp4160.pdf> (accessed on 12 March 2023).
7. Woo, J.; Kumar, M.S. Public Debt and Growth. Available online: <https://www.imf.org/external/pubs/ft/wp/2010/wp10174.pdf> (accessed on 12 March 2023).
8. Upreti, P. Factors Affecting Economic Growth in Developing Countries. *Major Themes Econ.* **2015**, *17*, 37–54.
9. Labra, R.; Rock, J.A.; Álvarez, I. Identifying the key factors of growth in natural resource-driven countries. A look from the knowledge-based economy. *Ens. Sobre Politics Econ.* **2016**, *34*, 74–89. [CrossRef]
10. Report of the United Nations Development Program (UNDP), published on 21 October 2020. Available online: <https://news.un.org/ru/story/2020/10/1388792> (accessed on 17 April 2023).
11. Heinisch, K.; Wohlrabe, K. The European Refugee Crisis and the Natural Rate of Output, IWH Discussion Papers, No. 30/2016, Leibniz-Institut für Wirtschaftsforschung Halle (IWH). 2016. Available online: https://www.researchgate.net/publication/260096281_Theories_and_Typologies_of_Migration_An_Overview_and_A_Primer (accessed on 7 May 2023).
12. Kudaeva, M.; Redozubov, I. The Impact of Migration Flows on Economic Activity and the Labor Market of Russia in General and in the Regional Aspect. 2021. Available online: http://www.cbr.ru/statichnol/file/131869/wp_khab_dec.pdf (accessed on 20 April 2023).
13. Zhuravskaya, K. Statistical Analysis of the Factors Shaping the GDP. *Agro-Food Economics*. 2016. Available online: <http://apej.ru/article/04-04-16> (accessed on 20 April 2023).
14. Kira, A.R. The Factors Affecting Gross Domestic Product (GDP) in Developing Countries: The Case of Tanzania. *Eur. J. Bus. Manag.* **2013**, *5*, 4. Available online: www.iiste.org (accessed on 27 April 2023).

15. Guo, H.; Zhang, Z. An Empirical Study of Factors Influencing Australia's GDP. In Proceedings of the 2022 International Conference on Mathematical Statistics and Economic Analysis (MSEA 2022), Dalian, China, 27–29 May 2023; pp. 581–586.
16. Ribaj, A.; Mexhuani, F. The impact of savings on economic growth in a developing country (the case of Kosovo). *J. Innov. Entrep.* **2021**, *10*, 1. [CrossRef]
17. Nketiah, E.; Cai, X.; Adjei, M.; Boamah, B.B. Foreign Direct Investment, Trade Openness and Economic Growth: Evidence from Ghana. *Open J. Bus. Manag.* **2020**, *8*, 39–55. [CrossRef]
18. International Migration: Drivers, Factors and Megatrends (2020). International Centre for Migration Policy Development (ICMPD). Available online: <https://www.icmpd.org/file/download/51472/file/Policy%2520Paper%2520-%2520Geopolitical%2520Outlook%2520on%2520International%2520Migration.pdf> (accessed on 1 May 2023).
19. Courtney, B.; Dustmann, C.; Preston, I. The Labor Market Integration of Refugee Migrants in High-Income Countries. *J. Econ. Perspect.* **2020**, *34*, 94–121.
20. Pew Research Center. *Around the World, More Say Immigrants Are a Strength Than a Burden*; Spring 2018 Global Attitudes Survey: Washington, DC, USA, 2019.
21. Worldwide Immigration Trends Report. Available online: https://www.fragomen.com/trending/worldwide-immigration-trends-reports/index.html?gclid=CjwKCAjw6vviBhB_EiwAQJRopmEjn5ZEKmbLRGrxitYEGUwnsobcCGZhu6Gzj1_enquuQ0euA7YnBoChlYQAvD_BwE (accessed on 27 April 2023).
22. World Migration Report 2022. International Organization for Migration. Available online: <https://worldmigrationreport.iom.int/wmr-2022-interactive> (accessed on 27 April 2023).
23. An Official Website of the Eurostat. Homepage. Available online: <https://ec.europa.eu/eurostat> (accessed on 27 March 2023).
24. An Official Website of the State Statistics Service of Ukraine. Homepage. Available online: <https://www.ukrstat.gov.ua/> (accessed on 12 March 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Defining Sports Performance by Using Automated Machine Learning System [†]

Kalle Saastamoinen ^{1,*}, Tuomas E. Alanen ², Pasi Leskinen ², Kai Pihlainen ³ and Joonas Jehkonen ⁴

¹ Department of Military Technology, National Defence University, FI-00861 Helsinki, Finland

² Naval Academy, Finnish Defence Forces, FI-00191 Helsinki, Finland; tuomas.alanen@helsinki.fi (T.E.A.); pasi.leskinen@mil.fi (P.L.)

³ Training Division, Defence Command, Finnish Defence Forces, FI-00131 Helsinki, Finland; kai.pihlainen@mil.fi

⁴ Shared Service Center, Information Management, Finnish Defence Forces, FI-04401 Helsinki, Finland; joona.jehkonen@mil.fi

* Correspondence: kalle.saastamoinen@mil.fi

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: We wanted to determine whether we could use an automated machine learning system called Azure for the selection process and placement of conscript training in such a way that AI can make decisions for the right conscript training program individually. To test this, we had four separate datasets and access to the Microsoft Azure automated machine learning environment. According to the test sets we performed, we see that, by using an automated machine learning environment, it was possible to reach the precision level of the decisions we wanted. The main obstacle was not the used automated machine learning environment itself, but the quality of the data used for learning. We also made improvement suggestions regarding how data could be collected and what kind of data we should measure to make predictive data analysis better and be more usable in the future.

Keywords: health; forecasting; automated; data-analysis

Citation: Saastamoinen, K.; Alanen, T.E.; Leskinen, P.; Pihlainen, K.; Jehkonen, J. Defining Sports Performance by Using Automated Machine Learning System. *Eng. Proc.* **2023**, *39*, 87. <https://doi.org/10.3390/engproc2023039087>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 14 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

There are many possible applications of artificial intelligence (AI) that can be used by educational, research and military institutions: (1) Making it possible to easily and at least partially automatically collect data like advertising tools do in many social media applications; (2) Analyzing data with the help of artificial intelligence applications, neural networks and learning algorithms. (3) Genuinely and measurably benefiting from the information obtained, which has an impact on the development of the implementation of systems.

In the following, we can also see risks/challenges in terms of utilizing artificial intelligence: (1) Not enough data are obtained or these do not actually support profiling; (2) The data are in non-usable format; (3) The total costs of the system do not match the achievable benefits; (4) Legal or information security-related obstacles are insurmountable; (5) The absence of a terminal device supports the collection of necessary data, such as a smart watch or other devices that track exercise.

The beneficiaries are:

1. The individual in question, so that they can monitor their development towards the target level. The system proposes a customized study path for the development of those skills, knowledge or physical characteristics that have the greatest risk of falling off the target path. From the point of view of learning analytics, the learner is one of the four areas of the cycle and the other three areas are the data, analysis, and

- action [1]. If only reports are generated from the learning data, on the basis of which no action is taken, the activity is omitted and the closed cycle of learning is not formed.
2. The trainer can monitor the development of the mass-produced group as a whole and as individuals. See where there are the most learning difficulties and, on this basis, make development ideas to make education more effective and/or to encourage and support individuals.
 3. The head of the unit can monitor the development of the group's overall performance and find opportunities for creating priorities in terms of the training content.
 4. The manager sees the mass-produced group and individuals and their suitability for the tasks in hand. If the suitability does not meet the demands, the system suggests the most suitable second option, or the manager can return to the root cause of why the suitability does not meet the demands.
 5. The training branch can examine the implementation of training at the troop division, defense branch, and general staff level. It is essential to get grounds for changing existing operational models and training practices, if the change is beneficial from the perspective of adjusting performance, economy, or time. We can utilize the results of the analysis in the use of different target groups, and these results must be relevant for each target group and produce new value. The end users of the target group must be taken into account when designing the views.
 6. The competence centers monitor the best results and practices in their own industry. Then, these analyze background variables and share the best practices for everyone to use.

Can we use AI for the selection process and placement of conscript training in such a way that AI can make decisions for the right conscript training program individually? In addition to this, the goal is to monitor and support conscripts education throughout their military service. In this article, we analyze the data collected from different sources with the help of artificial intelligence and compare the results to the physical requirements of different tasks.

2. AZURE, Automated Machine Learning and Voting Ensemble

Microsoft Azure offers data analysis services as part of the Azure Machine Learning service package [2]. The service package includes data analysis capabilities, the training and production of machine learning models, as well as version control and monitoring. The service package includes the Automated Machine Learning solution for producing models.

Ensemble Method

Azure uses ensemble methods to combine multiple machine learning algorithms together in order to create a more accurate and reliable prediction [3]. This is performed by running multiple models against the same dataset and then combining the results of each model into a single, more accurate result. Ensemble methods are used in Azure for tasks such as object detection, image classification, natural language processing (NLP), recommendation systems, anomaly detection, and time series forecasting. As an example, the Azure Ensemble includes the following models: 1. random forests; 2. gradient-boosted decision trees (GBDTs); 3. logistic regression; 4. support vector machines (SVMs); and 5. neural networks (NNs).

3. Problem and Data Description

In 2020, the Finnish Defense Forces introduced target levels for the physical performance of professional soldiers, which consist of endurance and muscle condition. The corresponding target levels for conscripts have been drawn up as part of the development of the selection process for approximately 400 positions. The purpose of this study was to utilize the durability classifications of the target levels, which are presented as the five result limits of the Cooper test. In the starting situation, there were five separate datasets that were not even related to the official task-specific aptitude test of conscripts at the user level

(user_id), and for this reason, making predictions from the aptitude test would have been impossible from the outset. For this reason, no task-specific comparison was made, but the focus was on determining the predictability of the Cooper variable based on the given data. We were given four separate datasets, which were gathered from 6000 volunteers, from these 3282 used wearable devices. The datasets were as follows:

File “exercise-data.csv” contains users’ exercise data on a weekly basis, e.g., the user’s total amount of exercise in minutes during the week. Number of rows: 283,537; and number of unique users: 6000.

File “survey-data.csv” contains information from the survey filled in by users, e.g., the result of the Cooper test. All questions are multiple-choice questions, e.g., the Cooper result is only available in categories such as “2071–2900 m” or “over 3100 m”. Number of lines: 4356; and number of unique users: 3236. Note that there could be more than one row in the data from the same user.

File “weight-data.csv” contains users’ weight data at the timestamp level. Number of rows: 109,521; and number of unique users: 2476.

File “steps-sleep-data.csv” contains users’ sleep and step counts compiled on a weekly basis, e.g., the user’s average amount of sleep during the week. Number of rows: 624,000; and number of unique users: 6000. Lots of values are missing from the columns.

The following exercise behavior variables were formed from the dataset:

- Total duration = total duration of training in minutes;
- Total count = number of training times;
- Total distance = total distance in kilometers;
- Total steps = total number of steps;
- Endurance duration = duration of endurance training;
- Endurance count = number of endurance training sessions;
- Strength duration = duration of strength training;
- Strength count = number of strength training sessions;
- Endurance target met = fulfillment of the endurance exercise recommendation (2.5 h/week) (yes/no);
- Strength target met = meeting the strength training recommendation (2 times/week) (yes/no);
- Steps avg = weekly average of daily steps;
- Steps min = the smallest number of steps per day of the week;
- Steps max = maximum number of steps per day of the week;
- Sleep avg = average amount of sleep during the week;
- Sleep min = the lowest daily amount of sleep per week;
- Sleep max = maximum daily amount of sleep per week;
- Weight = body weight (kg).

The main problem was the quality of the data, missing data points, and the absence of a target variable corresponding to the actual question. For example, the measured Cooper’s test results and the results of the muscle fitness tests were missing. From Cooper’s test, only the user’s own categorical assessment of their result was available with an accuracy of a few hundred meters. In addition, the Cooper test questionnaire was made even a couple of years earlier than the actual use of the fitness application started in 2020. Naturally, a person could have a completely wrong idea of their own Cooper condition. Partly for these reasons, but especially because the Cooper test result is given as a categorical variable, the regression model made from the data was not able to predict the assessment given by the user in the survey very well, with the explanation rates remaining poor. In order to reach higher degrees of explanation with the directly given data, it should have contained the exact measured values of the test results to be explained. In addition, it would be good to have all the measured data in the same file in csv format. For their part, these would have improved and facilitated the data analysis. The data should also have clear and objectively measurable values that aim to improve and from which it would be reasonable to make predictions, e.g., Cooper, muscle fitness tests, muscle–fat ratio, BMI, etc.

4. Results

In the classification of the target `endurance_target_met` in the exercise dataset, the accuracy was 100%, as can be seen in Figure 1. We converted all the values in the numbers and aggregated with the arithmetic mean in the regression.

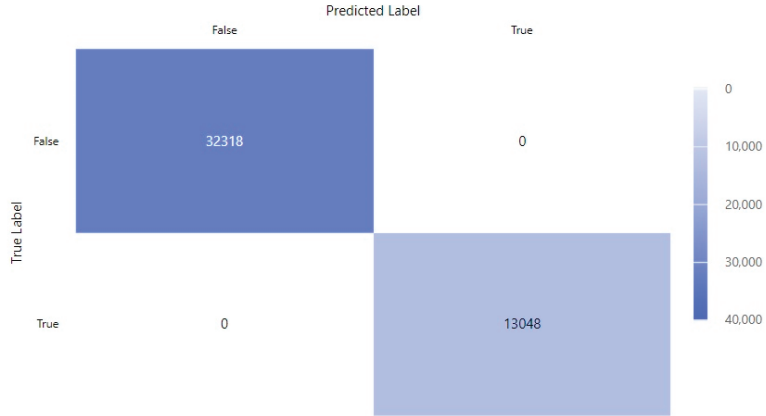


Figure 1. Confusion matrix of the classification exercise data with the target `endurance_target_met`.

There are 624,001 rows in the steps–sleep data. There are a lot of missing data in the columns, that is why the explanation rate r^2 with the regression target `sleep_avd` is only 0.024%. When all measurements containing missing data are removed from the data, 92,006 measurements remain. When this is regressed, the explanation rate r^2 is 92.2%. When these values were aggregated with the arithmetic mean, the explanation rate r^2 is 97.9%. From Figure 2, it can be seen that the minimum and maximum sleep were the main features estimating the average amount of sleep during the week.

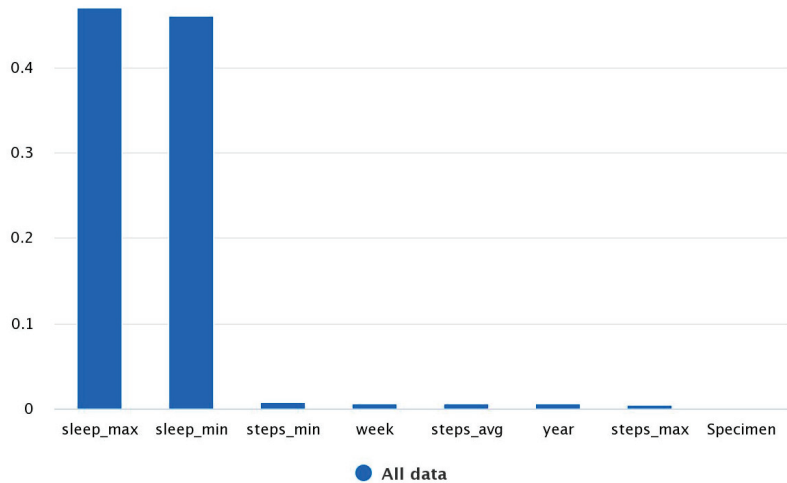


Figure 2. Feature importance of the regressing steps–sleep data with the target `sleep_avd`.

In the survey data, when classifying the object Cooper, the accuracy is only 26.3%. When the query intervals of the Cooper result are replaced by numbers 1–7, the accuracy decreases to 22.2%. When all survey data values are replaced with numerical equivalents, the accuracy was 25.5%. When the survey dataset is combined with the exercise dataset using the INNER JOIN operation with respect to users (`user_id`), at the same time, the

values of the survey data are converted into numbers, the classification result increases to 95.8%. If we include the weight dataset to this using the INNER JOIN operation, the classification result again increases to 98.2%, as can be seen in Figures 3 and 4.

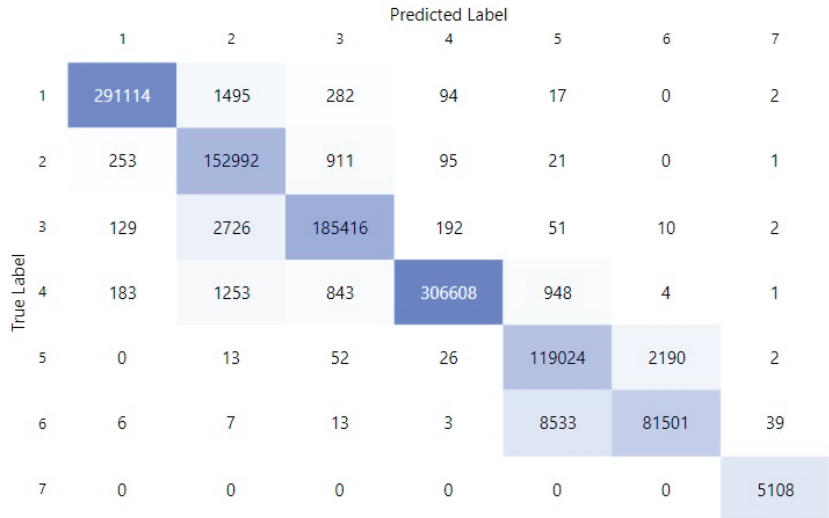


Figure 3. Confusion matrix of the classification combined datasets with the target cooper.

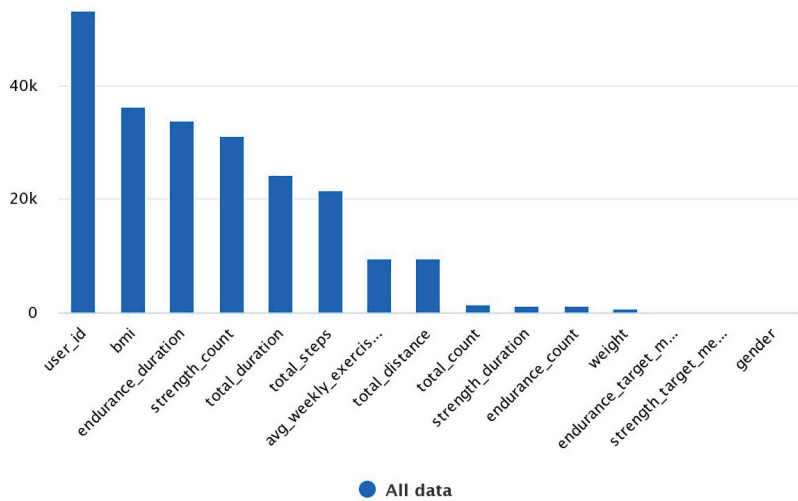


Figure 4. Feature importance of the classification combined datasets with the target cooper.

With the given data, we reach more than 90% accuracy with all tested measurements. However, in all other cases, except for exercise data, this required the deletion of the missing data or the merging of data. Combining the weight data with other data partly improved the obtained results.

5. Discussion and Future

In this study, it was found that, using directly provided data, it was reasonably difficult to reach the 90% target value given in the forecasts. In previous surveys, it has been possible to predict endurance fitness with typically 50–60% accuracy. For example, Matsuo et al. [4] reported that, in the adult population, age, sex, body mass index and level of physical

activity explained 59% of the variation in the measured maximal oxygen uptake capacity. In a recent research publication by Santtila et al. [5] on Finnish conscripts, the result of the 12 min running test was able to be predicted based on the self-reported amount of activity and sitting, the assessment of readiness to perform conscript service, educational background, smoking, and body composition with a 52% explanation rate (mean error 8.8%/207 m). These accuracy values are not sufficient in terms of their usability to make predictions from conscript training choices.

By modifying the given data by combining and manipulating it, we reached the target value, i.e., more than 90% predictability, at least in the prediction of the tested variables (Cooper, endurance goal and average amount of sleep). In the analyses of this report, the explanation rates reached a maximum of 100%. Endurance condition could not be predicted in the end, because the source material did not contain enough results of a condition test newly developed for the application. As a final result, it was found that, by developing the data to be used by adding data points, it is possible to become closer to the set goals, and in order to reach these, the necessary measures were already taken in part, even without this investigation, to obtain the missing data points.

Based on the results, it can be concluded that machine learning methods can be used to determine good enough predictions about conscripts conditions. In the future, exercise behavior data must only be objectively collected by measuring using wearable/portable devices or under controlled test conditions. Muscle and endurance test results can be used to predict the conscript's fitness class, as long as the measured results exist in one file. On a more general level, it can be stated that by utilizing existing datasets, knowledge-based management can also be developed in the education industry. Various datasets must be tested with courage, so that their capabilities in managing information can be recognized.

Author Contributions: Conceptualization, K.P., T.E.A. and J.J.; methodology, K.S.; validation, K.S. and K.P.; formal analysis, K.S.; investigation, K.S., T.E.A. and K.P.; resources, P.L.; data curation, K.S.; writing—original draft preparation, K.S.; writing—review and editing, K.S.; visualization, K.S.; supervision, P.L.; project administration, P.L.; funding acquisition, P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: No statements provided.

Informed Consent Statement: We have used only anonymous data during this research.

Data Availability Statement: Data is available through Kalle Saastamoinen, kalle.saastamoinen@mil.fi.

Acknowledgments: We are grateful to HeiaHeia Ltd. for providing us with the original data, Digia Ltd. and Oracle Ltd. for valuable and constant support during this research. We especially thank data scientist Julius Nieminen from Digia Ltd. who did more than his share analyzing data, and special thanks also go to the data scientist Bob Peulen from Oracle Ltd. for giving us the idea of combining the datasets for better results.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Clow, D. The learning analytics cycle: Closing the loop effectively. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, BC, Canada, 29 April–2 May 2012; pp. 134–138.
2. Goswami, M.; Franks, L.; Salgado, S.; Nagata, S.; Ndem, R.; Ovhal, P.; Gilley, S.; Jain, S.; Gold, B.; Wu, J.; et al. What Is Automated ML? AutoML—Azure Machine Learning | Microsoft Learn. Available online: <https://learn.microsoft.com/en-us/azure/machine-learning/concept-automated-ml> (accessed on 21 March 2023).
3. Asanka, D. Building Ensemble Classifiers in Azure Machine Learning. Available online: <https://www.sqlshack.com/building-ensemble-classifiers-in-azure-machine-learning/> (accessed on 21 March 2023).

4. Matsuo, T.; So, R.; Takahashi, M. Workers' physical activity data contribute to estimating maximal oxygen consumption: A questionnaire study to concurrently assess workers' sedentary behavior and cardiorespiratory fitness. *BMC Public Health* **2020**, *20*, 1–10.
5. Santtila, M.; Pihlainen, K.; Vaara, J.; Nindl, B.C.; Heikkinen, R.; Kyröläinen, H. Aerobic fitness predicted by demographics, anthropometrics, health behaviour, physical activity and muscle fitness in male and female recruits entering military service. *BMJ Mil. Health* **2022**, *Epub ahead of print*. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Proceeding Paper

Forecasting Transitions in Digital Society: From Social Norms to AI Applications [†]

Daniel Ullrich ¹ and Sarah Diefenbach ^{2,*}

¹ Department of Computer Science, Ludwig-Maximilians-Universität München, 80337 Munich, Germany; daniel.ullrich@ifi.lmu.de

² Department of Psychology, Ludwig-Maximilians-Universität München, 80802 Munich, Germany

* Correspondence: sarah.diefenbach@psy.lmu.de

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: The use of AI and digitalization in many areas of everyday life holds great potential but also introduces significant societal transitions. This paper takes a closer look at three exemplary areas of central social and psychological relevance that might serve as a basis for forecasting transitions in the digital society: (1) social norms in the context of digital systems; (2) surveillance and social scoring; and (3) artificial intelligence as a decision-making aid or decision-making authority. For each of these areas, we highlight current trends and developments and then present future scenarios that illustrate possible societal transitions, related questions to be answered, and how such predictions might inform responsible technology design.

Keywords: digital society; social norms; social scoring; artificial intelligence (AI); future scenarios

1. Introduction

“I am sorry, but I have to inform you that we cannot undertake the surgery.” The news was upsetting for Anna, as surgery was the only option to stop the potentially fatal disease. Sure, surgery was risky, too. But without surgery, there was no hope left other than that the disease would cure itself. Anna had only recently been diagnosed with the rare disease, and the outcome was difficult to predict. The decision of treatment, however, was not only that of her doctor. In fact, her doctor based all his decisions on “Health Guardian”, an artificial intelligence (AI) system generating treatment recommendations based on incredible amounts of data. In Anna’s case, Health Guardian recommended not to do surgery. Anna’s mother, who joined the consultation, desperately asked whether there could be a mistake and whether the doctor was of the same opinion. The doctor was in a dilemma: Personally, he was not necessarily against surgery. He would even have argued in favor of surgery, had Health Guardian voiced any uncertainty. But he knew that compared to his own, naturally limited, perspective, the AI could factor in far more data. And that was what it came down to. Although the AI results were called “recommendations”, they were actually decisions. As the responsible doctor, he would have to present extremely good reasons to oppose the AI—but no such reasons were apparent in the current case. So the doctor had no choice but to console Anna and her family. At least there was still a sliver of hope for a natural recovery.

In recent years, artificial intelligence (AI) has achieved impressive successes in various domains such as visual perception [1], pattern recognition [2], expert and decision-making systems, games such as Chess and Go [3,4], or computer strategy games [5]. At the same time, critics still question whether these performances represent “real intelligence” [6,7].

In fact, the formation of “intelligence” in such systems is hardly comprehensible to us and exceeds the horizon of human understanding [8]. This is compounded by the

Citation: Ullrich, D.; Diefenbach, S. Forecasting Transitions in Digital Society: From Social Norms to AI Applications. *Eng. Proc.* **2023**, *39*, 88. <https://doi.org/10.3390/engproc2023039088>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 14 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

fact that such systems can hardly be repaired. While a human may understand the basic mechanisms, the specific design becomes so complex that it is no longer possible for a human to discern how to fix bad parts of the system without damaging other parts. Often, the only solution is a completely new start, namely the training of a new system with modified start parameters that hopefully will not end up with the same errors.

The general lack of transparency in AI technologies [9] is one of the factors in the doctor's dilemma in the above-mentioned example of AI in the operating room: AI decisions are hardly traceable by nature. AI systems refer to patterns detected in the example material (from the past) and then try to make predictions for the future and come up with new examples. However, which exact variables are considered, how these are weighted, and which correlations between these variables do exist remain hidden from the user (and mostly also from the programmer) [10]. Hence, re-turning to the case of the decision for or against surgery, the deciding factors for the Health Guardian's decision remain obscured. Was it only about the predicted effectiveness and risk of the intervention? How were other variables taken into account, such as the cost-to-benefit ratio, budget of the healthcare system, and bed occupancy rate of hospitals? What about other waiting patients who needed surgery more urgently? It does not appear unlikely that artificial intelligence will consider the constraints of relevant stakeholders, especially in societies where resources in the healthcare system are more limited than in others.

2. Overview and Method

The AI case is exemplary for the current challenges and questions around digital transitions that our society is faced with: What does it mean if current technological trends and developments continue? What are the psychological effects and consequences of social interaction? Which moral considerations play a role, and which decisions have to be made? This paper takes a closer look at three exemplary areas of central social and psychological relevance that might serve as a basis for forecasting transitions in the digital society: (1) social norms in the context of digital systems; (2) surveillance and social scoring; and (3) artificial intelligence as a decision-making aid or decision-making authority.

For each of these areas, we highlight current trends and developments and use the method of future scenarios to illustrate possible societal transitions and related questions to be answered. Thereby, we aim to contribute to several fields of research. First, the examples and implications discussed here may inspire future research and design in the fields of human-computer interaction (HCI) and AI. Moreover, regarding forecasting and future studies in general, this article may illustrate how qualitative analyses and future-related reflections on current technological and societal trends may complement more quantitative and statistical methods.

Of course, the here-applied method of future scenarios comes with particular limitations related to some fundamental problems with predictions. Typically, when trying to forecast the future, current developments are analyzed, and one then tries to project them into the future and anticipate their interactions with other developments. However, numerous examples demonstrate how difficult this is, even for experts.

In the 1950s, when people were asked how they imagined the year 2000, they assumed that people would travel in flying cars powered by miniaturized nuclear engines, as also depicted in an artwork by Frank Rudolph Paul, an illustrator of science fiction magazines in that time [11]. Two currently successful existing technologies, the car and nuclear power, were taken as a basis and projected into the future. However, the dangers and technical limits of nuclear power could not be anticipated. Could the people of the past have made a better prediction if they had studied nuclear power more intensively? Possibly. But even if one misjudgment is taken into account and corrected, there are still many others.

In the second half of the last century, researchers at the Massachusetts Institute of Technology (MIT) published a study on the future of the world economy [12]. The key question was to predict the (assumed) necessary collapse of the current economic system based on exponential growth. Numerous parameters, such as population growth and den-

sity, aging of society, movement of goods, government budget, and debt, were considered in the prediction model. According to the model calculations, the time of collapse would be within the next 100 years, i.e., around the year 2070. However, many parameters and events that turned out to be relevant later on were naturally not considered, e.g., the disintegration of the Soviet Union, the rapid rise of China as a world power, and the significance of climate change for the planet. The significance of these developments was not foreseeable when the calculations were made and thus was not adequately taken into account in the forecast model. In the meantime, the forecast model was updated with new parameters [13]—we will see whether the predictions hold true this time.

A basic problem here is that so-called disruptive events, findings, or technologies are not taken into account. Disruptive technologies are technical innovations that replace or displace established products or services and interrupt the success of previously prevailing approaches [14]. One example would be the Internet, which has opened up many new areas of business, but at the same time brought about the collapse of many previously successful business models. A few years earlier, no one would have predicted the disruptive character of the Internet, and in turn, many predictions that disregarded the influence of the Internet were faulty.

In the end, we must remind ourselves that predictions are still a kind of thought experiment and do not allow for perfect knowledge of what will actually happen. However, this should not diminish the importance of such thought experiments. Even non-perfect thought experiments are still better than not thinking at all. Such thought experiments reveal what could happen and indicate possible alternative courses of action. Thought experiments emphasize that we are not mere passengers being overrun by the future but can actively help to shape it.

3. Social Norms

Social norms are the unwritten rules of beliefs, attitudes, and behaviors that are considered acceptable in a particular social group or culture [15]. Social norms represent shared beliefs regarding appropriate ways to feel, think, and behave [16]. In this way, social norms provide order and predictability in society [15]. For example, in German culture, if we make an appointment, we expect the other person to arrive on time. In contrast to legal norms (e.g., laws), social norms occur spontaneously rather than being planned deliberately and are enforced informally [17]. Typically, social norms only become evident when conflict arises, i.e., if someone's behavior contradicts our informal understanding of what is appropriate, such as cutting in line, entering an office without knocking, or starting to eat before everyone is seated at the table [18]. The same seems to apply to the digital space. Many conflicts in the context of social media and digital communication can be interpreted as social norm conflicts [19].

Regarding the forecast of societal transitions in the digital age, the differences or transfer of norms between the digital and non-digital spaces is an interesting aspect. In order to understand and possibly foresee such transitions, we will first take a look at some particular possibilities and characteristics of the digital space, which in turn affect the formation, change, and enforcement of certain social norms.

- Distance between interaction partners: In many channels of digital communication, interaction consists only of writing and reading text. Social cues we adhere to in face-to-face conversation (human characteristics such as appearance, voice, and physical presence) are missing. Therefore, it is terribly easy to forget that one is not interacting with texts but with humans, who have their own motives, their own value system, feelings, and emotions, and who can be hurt or offended by one's own actions. In consequence, one might not even notice having hurt the counterpart on the other end of the digital channel, and empathic mechanisms that could show the consequences of one's own actions are not activated [20];
- Avatar and control, instead of authenticity: On many social media platforms, users are represented through an avatar, which can easily be exchanged if this seems convenient.

A re-creation of another account is quickly carried out, allowing one to restart with a clean slate (assuming interactions in anonymous or pseudonymous space). Such a new start and identity change are very difficult in the non-digital space. And even on platforms where the avatar/identity cannot be easily changed, the user has much greater control over what information is revealed about him or her. In particular, involuntary aspects of communication (facial expressions, affective reactions, and voice color) are greatly reduced in digital space [21];

- Felt anonymity: The fact that other interaction partners often appear as avatars and the fact that you yourself do not know who the other person is exactly create an illusion of complete anonymity. Even though, technically speaking, users can actually be identified and are only anonymous to each other, this feeling of anonymity still has psychological consequences. This pseudo-anonymity can be sufficient to make people feel “safe” and disregard regular social norms. Like hooded demonstrators, seemingly anonymous users may no longer feel obliged to follow social rules [21,22]. Not all users make use of this “freedom,” but a significant portion do;
- Digital-exclusive mechanisms: The digital space provides various interaction mechanics that are unknown or even impossible in the non-digital space. One example is ghost banning. Ghost banning is a technique that is typically used against so-called trolls (i.e., internet forum troublemakers who derive satisfaction from provoking other users with polarizing statements). If a troll was just simply banned (deleted), this would not solve the problem for a long time since the user could easily create a new account and start again. Ghost banning, however, is a process through which a user is invisibly banned from a social network, website, or online community. The user retains the ability to browse through and use the available features without knowing that his or her actions are invisible to other users. This, in turn, prevents the user from interfering with other users [23]. Colloquially speaking, when an admin ghost bans a troll, this puts the troll in an invisible cage where they are unaware that other users cannot see their posts [24]. Initially, the ghost-banned troll has no way to determine his invisibility to others and can at best wonder about the lack of reactions to his provocations. Only if the troll would log in with another user’s account and obtain their perspective on the online world could he or she find out what is going on. Transferring the technique of ghost-banning to the non-digital space, one could imagine an invisibility cloak you can put on troublemakers without the person noticing. What is pure fiction in the real world is everyday life in the digital realm: every user receives his or her own individual view of the (digital) world, and the differences are seldom communicated.

Already nowadays, due to the ubiquitous use of digital interaction channels, corresponding digital norms are gaining more and more weight, which are in turn influenced by the peculiarities of the digital space.

A Possible Future Scenario

Social norms are implicitly learned and adhered to, and norms from the non-digital space influence those from the digital space, and vice versa [19]. We can conclude that as digitally mediated social interaction becomes more and more pervasive in everyday life, we are exposed to norms from the digital space to a greater extent. This, in turn, increases the relative influence of these norms. Ultimately, this could lead to a situation where norms from the digital space dominate over traditional norms that originated in the non-digital world.

Taking into account the characteristics of the digital space mentioned above, this could result in a greater level of rudeness and less consideration of the other’s emotional world. A side effect could also be the development of avoidance strategies against direct, non-digital interaction. In particular, people might stick to non-synchronous digital channels, such as text messaging, as a protective shield to insulate themselves from the possibly distressing interaction of the interaction partner [25], the so-called buffer effect [26]. In fact, there is already a perceptible trend among younger people to avoid direct synchronous interaction,

such as face-to-face conversations or telephone calls (e.g., [27]). Instead, they are turning to more distant, mediated communication wherever possible. Instead of dealing with one's own empathic reactions, non-digital contact is more and more evasive. As a result, empathic skills are used and trained less frequently, which, again, increases the preference for digital channels—a self-reinforcing dynamic.

Along with these predictions, we must also consider that, of course, the repertoire of traditional norms acquired over centuries in the non-digital world still continues to shape our behavior. In other words, the current observable state is still skewed in favor of conservative norms, and the future influence of norms from the digital world will become even stronger. A fictitious society starting from “zero” would presumably be even more strongly influenced by norms from the digital world. Following these thoughts, every existing society will be influenced increasingly by digital norms over time—if solely for the reason that older people, who tend to be representatives of conservative norms, die and are replaced by those who come after them and who are more strongly influenced by norms from the digital world.

4. Surveillance and Social Scoring

When the Internet emerged, the first goal was to create a failsafe communication infrastructure that would continue to function even if parts of it broke down [28]. Only later did additional (primarily economic-driven) goals emerge, such as creating specific social networks, tracking users' paths, and presenting targeted advertising. Thus, the early days of the Internet were characterized primarily by freedom: Freedom in users' actions and freedom from control. This period is also referred to as the golden age of the Internet or the Wild West period without rules [29].

However, as the popularity of the Internet increased, the economic potential of big data and large user groups became more and more recognized. First and foremost, this was the display of advertisements and the creation of numerous digital trading places [30]. In addition, the dissemination of news and information also played an increasingly important role. With more and more people obtaining their information from the Internet, the senders of information gained a steadily growing reach [31]. A natural follow-up question was how to maximize influence on users and how to establish information sovereignty: who determines which of two contradictory pieces of information is “correct”?

Accordingly, it did not take long for various stakeholders to discover the worldwide web and its users for their interests, and they began to extend their influence: Politicians, news portals, the advertising industry, providers of consumer products, activists, and individual opinion leaders as well as “influencers” [32]. As such, the Internet can be seen as the antithesis of the classic democratic society, in which information sovereignty is concentrated in the hands of the state or a small group of people. On the Internet, on the other hand, everyone is a sender and a receiver; everyone can potentially participate in opinion formation [33,34] and is, thus, a potential competitor to the major established media—a state of affairs that (traditional) media and politics losing control do not necessarily find desirable. This is accompanied by attempts at surveillance and information control, such as upload filters or sabotage of encryption technologies. Typically, these are justified with popular goals such as criminal prosecution, referring to relatively small groups of offenders (e.g., child pornography, illegal black markets). However, the negative effects and potential misuse of surveillance technologies affect all users equally.

A Possible Future Scenario

With the increasing digitalization of everyday life, the potential for surveillance increases as well. With every online action, users leave their digital footprints, becoming more and more transparent citizens. On the users' side, the awareness of monitoring leads to adapted behavior, and even the mere awareness of potentially being monitored creates distress—a symptom also known as the “chilling effect” [35]. Of course, this chilling

effect can be deliberately utilized to steer user behavior in the desired direction. Since not everyone needs to be monitored, this method is also cost-effective.

At the same time, alternative ways of surveillance, such as AI-based algorithms, will become more popular. Where once actual humans had to detect offenses in the social media world, algorithms can slip into the monitoring role. For example, such algorithms can automatically detect copyright infringements, (child) pornography, or certain keywords that are taboo on the platforms. However, the effect of such interventions has so far been negligible, since even being banned from a platform does not generally represent a serious consequence for these users.

With the introduction of social scoring, this has fundamentally changed. Social scoring takes the monitoring aspect to a new level and turns implicit, casual influence into an explicit, targeted one: with the use of social scoring—citizens receive points for desired behaviors and deductions for undesired behaviors—desirable behavior is explicitly prescribed (e.g., [36,37]). When such social scores affect real-life chances (e.g., when looking for a job or when searching for an apartment), violations against desired behaviors have specific and tangible consequences for users. Naturally, any criticism of this system will be classified as an undesirable action as well. Withdrawal from such a system will become almost impossible as soon as critical functionalities (freedom to travel, payment functions, prioritization in the search for housing, jobs, hiring criteria analogous to a police clearance certificate) are linked to the social score. In the end, the self-reinforcing spiral of social scoring systems may result in more and more extreme and comprehensive rules until all areas of human behavior are covered.

As these considerations reveal, the basic idea of social scoring already contains much negative potential. Therefore, no matter what disruptive event of the future might stop it or not, it seems important to consider now whether we want to prevent the establishment of such a concept through our actions today.

5. AI as Decision-Making Aid or Decision-Making Authority

Artificial intelligence is already being used to support complex decisions, for example in the fields of insurance [38,39], medicine (for example, diagnostics and pattern recognition in image processing mentioned by Kermany et al. [40], Esteva et al. [41]), and HR, where artificial intelligence can help identify the most suitable candidate for an advertised position [42,43]. Across all these applications, the possibilities of artificial intelligence (in particular, machine learning) are limited by three main factors:

- The specification of the method, algorithm, or network topology;
- The computing power for training the AI;
- The number of available data sets matching possible input data and output data (for example, a large collection of different animal images, each with an indication of which animal is depicted).

In many application domains, the current technical possibilities regarding all three factors are sufficient to create AIs that deliver results that are equal to or superior to those of humans. Especially for the last factor, i.e., the data sets that link input patterns with correct results, progress results as a kind of by-product of the activities of current users (e.g., of social media platforms). Every new set of stored user data generates new training data. Hence, the situation is becoming better every day—at least for those who can store and utilize the data.

A Possible Future Scenario

As soon as AI methods are able to replace human labor or skills of equal quality, there is no question of whether these methods will be applied. Not using such methods would result in a significant competitive disadvantage, maybe even being put out of the market. As methods and data collections continue to evolve, AI will find its way into more and more fields as a decision support tool, such as jurisdiction [44,45], partner choice [46,47], and many more.

With the invasion of AI into ever new domains, many questions arise, beginning with the most fundamental one: Should AI be allowed to enter all domains of human society or are there any barriers?

Moreover, what if AI delivers recommendations that are politically incorrect and therefore undesirable? How can it be ensured that the training data is “neutral” so that no bias is transferred to the trained AI?

Moreover, who is responsible for the indirect consequences of AI recommendations, and what kind of events can be traced back to an algorithm?

For example, in a recent case before the US Supreme Court, a mother whose daughter died along with 130 other people in connection with the ISIS terrorist attack in November 2015 in Paris alleged that Google’s YouTube algorithms effectively amplified Islamic State-produced materials in support of the extremists that killed her daughter [48]. As with many online media platforms, YouTube’s recommendation algorithm basically aims to suggest relevant items to users by directing them to videos that are similar to those they have previously selected and watched. YouTube’s recommendations thus mirror the user’s apparent interest. However, the family of the terror victim argues that YouTube’s recommendations expose people to (ever more) hateful content, radicalize viewers, and ultimately encourage them to make terrorist attacks of their own [49]. To date (as of February 2023), the case is still under trial. With ever more complex AI systems and algorithms in the future, such legal and moral questions will probably become more complex as well.

In connection to this, another block of questions refers to the transparency of AI: Is there a right to understand on what basis an AI makes concrete recommendations—and how could such a right ever be realized if, by nature, AI decisions remain a black box to some extent?

With the current state of technology, it is certain that AI can neither offer error-free decision-making nor transparent reasons for its decisions. At the same time, these shortcomings do not mean that AI will not be applied, especially when considering the advantages on the other side.

What will be essential, then, is how people feel about AI and its role in important decisions in society. Would it be desirable, for example, if an AI that has access to your data and will regard your interests would decide about the future and regulations in a country instead of human politicians?

When asked that question, a survey found overall high ratings in favor of AI: In the European region, the approval rate is 51% on average, with particularly strong support for AI in Spain (66%), Italy (59%), and Estonia (56%). In China, 75% are in favor of AI as a political decision-maker, whereas in the USA, only 40% want to delegate political decisions to AI [50].

Independent from the application domain, it seems likely that the use of AI will become more mainstream and that technological progress will more or less override the discussion about which applications are desirable or ethical.

6. Outlook

The use of AI and digitalization in many areas of work and private life will continue to increase in the future and hold great potential overall. Unpleasant tasks can be delegated to technology; AI can take over tasks that overwhelm or bore humans (and possibly vice versa). However, what we need to keep in focus are the major societal changes that might come with the use of AI. A system based on supply and demand for (human) work performance can hardly be maintained in its current form if artificial agents are competing with humans. New ideas for living and working together are needed. While there is probably still some time left before the big breakthrough of artificial agents, no one knows exactly how much time. When that day comes, there needs to be an action plan defining the space we want to grant AI in society. Otherwise, we will only be able to react to a factual reality instead of designing a desirable future.

Altogether, these considerations show that the innocent golden age of AI and digitization is over. Simply accepting their effects and side effects on our society is not acceptable. Conscious technology design requires us to predict how technology will continue to develop, what effects we can expect on our society, and how we can counter these influences with foresight. As in the physical world, our behavior in the digital space is influenced by design decisions [19]. In order to promote desired, prosocial behavior and reduce antisocial behavior, it needs a deliberate consideration of how certain features of technology affect social dynamics and the world we live in. Not everything that is technically feasible is morally acceptable. There is no such thing as neutral design.

Even with conscious design decisions, developing solutions that actually work flawlessly continues to be a challenge. For example, the approaches chosen to promote prosocial behavior can again have undesirable side effects. Trying to prevent antisocial behavior by making users completely transparent means trading one problem for another. The same applies to surveillance and social scoring. The negative effects of social scores must be researched in advance so as not to create a factual situation from which it will be nearly impossible to escape later on.

In sum, the development of good solutions that are morally and socially acceptable is one of the current core tasks in the context of digitalization.

Author Contributions: Conceptualization, D.U. and S.D.; methodology, D.U.; writing—original draft preparation, D.U.; writing—review and editing, S.D.; project administration, D.U. and S.D.; funding acquisition, S.D. All authors have read and agreed to the published version of the manuscript.

Funding: Part of this research was funded by the German Research Foundation (DFG), Project Perform (425412993) as part of the Priority Program SPP2199 Scalable Interaction Paradigms for Pervasive Computing Environments.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This is a revised version of “Ullrich, D. (2022). Zukunftsvisionen. In: S. Diefenbach, S. and P. von Terzi, P. (Eds). Digitale Gesellschaft neu denken. Chancen und Herausforderungen in Alltags- und Arbeitswelt aus psychologischer Perspektive. Stuttgart, Germany: Kohlhammer.” published in German by Kohlhammer publisher. Permission was granted by Kohlhammer publisher.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. DeepFace: Closing the gap to human-level performance in face verification. *2014 IEEE Conf. Comput. Vis. Pattern Recognit.* **2014**, *5*, 1701–1708.
2. Foggia, P.; Percannella, G.; Vento, M. Graph matching and learning in pattern recognition in the last 10 years. *Int. J. Pattern Recognit. Artif. Intell.* **2014**, *28*, 1450001. [CrossRef]
3. Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* **2020**, *588*, 604–609. [CrossRef] [PubMed]
4. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef]
5. Vinyals, O.; Babuschkin, I.; Czarnecki, W.M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D.H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **2019**, *575*, 350–354. [CrossRef]
6. Fjelland, R. Why general artificial intelligence will not be realized. *Humanit. Soc. Sci. Commun.* **2020**, *7*, 10. [CrossRef]
7. Crawford, K. Microsoft’s Kate Crawford: ‘AI Is Neither Artificial nor Intelligent’ (Z. Corbyn, Interviewer) [Interview]. 2021. Available online: <https://www.theguardian.com/technology/2021/jun/06/microsofts-kate-crawford-ai-is-neither-artificial-nor-intelligent> (accessed on 4 April 2023).

8. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
9. Larsson, S.; Heintz, F. Transparency in artificial intelligence. *Internet Policy Rev.* **2020**, *9*, 1–16. [CrossRef]
10. Kim, T.W.; Routledge, B.R. Why a right to an explanation of algorithmic decision-making should exist: A trust-based approach. *Bus. Ethics Q.* **2021**, *32*, 75–102. [CrossRef]
11. Novak, M. The World Will Be Wonderful in the Year 2000! Available online: <https://www.smithsonianmag.com/history/the-world-will-be-wonderful-in-the-year-2000-110060404/> (accessed on 4 April 2023).
12. Meadows, D.H.; Meadows, D.L.; Randers, J.; Behrens, W.W. *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*; Universe Books: New York, NY, USA, 1972.
13. Herrington, G. Update to limits to growth: Comparing the World3 model with empirical data. *J. Ind. Ecol.* **2021**, *25*, 614–626. [CrossRef]
14. Danneels, E. Disruptive technology reconsidered: A critique and research agenda. *J. Prod. Innov. Manag.* **2004**, *21*, 246–258. [CrossRef]
15. McLeod, S.A. Social Roles. Simply Psychology. 2008. Available online: www.simplypsychology.org/social-roles.html (accessed on 4 April 2023).
16. Turner, J.C. *Social Influence*, 16th ed.; Thomson Books: Belmont, CA, USA, 1991.
17. Hechter, M.; Opp, K.-D. *Social Norms*; Russel Sage Foundation: New York, NY, USA, 2001.
18. Diefenbach, S. Social norms in digital spaces: Experience reports on wellbeing and conflict in the teleworking context and implications for design. *Z. Für Arb.* **2022**, *77*, 56–77. [CrossRef] [PubMed]
19. Diefenbach, S.; Ullrich, D. Disrespectful technologies: Social norm conflicts in digital worlds. In *Advances in Usability, User Experience and Assistive Technology*; Ahram, T.Z., Falcão, C., Eds.; Springer International Publishing: Basel, Switzerland, 2019; pp. 44–56.
20. Carrier, L.M.; Spradlin, A.; Bunce, J.P.; Rosen, L.D. Virtual empathy: Positive and negative impacts of going online upon empathy in young adults. *Comput. Hum. Behav.* **2015**, *52*, 39–48. [CrossRef]
21. Suler, J. The online disinhibition effect. *CyberPsychology Behav.* **2004**, *7*, 321–326. [CrossRef]
22. Macdonald, C. Avatars, disconnecting agents: Exploring the nuances of the avatar effect in online discourse. *Open Sci. J.* **2020**, *5*, 1–8. [CrossRef]
23. Techopedia. Ghost Banning. 2019. Available online: <https://www.techopedia.com/definition/29190/ghost-banning> (accessed on 4 April 2023).
24. Slang.net. Ghost Banning. Censoring a Social Media User's Posts. 2022. Available online: https://slang.net/meaning/ghost_banning (accessed on 4 April 2023).
25. O'Sullivan, B. What you don't know won't hurt me: Impression management functions of communication channels in relationships. *Hum. Commun. Res.* **2000**, *26*, 403–431. [CrossRef]
26. Tretter, S.; Diefenbach, S. The buffer effect: Strategic choice of communication media and the moderating role of interpersonal closeness. *J. Media Psychol. Theor. Methods Appl.* **2022**, *34*, 265–276. [CrossRef]
27. Colbert, A.; Yee, N.; George, G. The digital workforce and the workplace of the future. *Acad. Manag. J.* **2016**, *59*, 731–739. [CrossRef]
28. Leiner, B.M.; Cerf, V.G.; Clark, D.D.; Kahn, R.E.; Kleinrock, L.; Lynch, D.C.; Postel, J.; Roberts, L.G.; Wolff, S. A brief history of the internet. *ACM SIGCOMM Comput. Commun. Rev.* **2009**, *39*, 22–31. [CrossRef]
29. Palacios, A. The Internet's "Wild West" Era: A Love Letter to the Early 00's Internet. 2019. Available online: https://medium.com/@alejandropalacios_98575/the-internets-wild-west-era-a-love-letter-to-the-early-00-s-internet-3075722f79ae (accessed on 4 April 2023).
30. Taylor, K. One Statistic Shows How Much Amazon Could Dominate the Future of Retail. Business Insider. 2021. Available online: <https://www.businessinsider.com/retail-apocalypse-amazon-accounts-for-half-of-all-retail-growth-2017-11> (accessed on 15 September 2022).
31. Beisch, N.; Schäfer, C. Ergebnisse der ARD/ZDF-Onlinestudie 2020. Internetnutzung mit großer Dynamik: Medien, Kommunikation, Social Media. *Media Perspekt.* **2020**, *9*, 462–481.
32. Moffett, S.; Santos, J. Social media as an influencer of public policy, cultural engagement, societal change and human impact. In Proceedings of the European Conference on Social Media: ECSM 2014, Brighton, UK, 10–11 July 2014; pp. 312–319.
33. Bakshy, E.; Rosenn, I.; Marlow, C.; Adamic, L. The role of social networks in information diffusion. In Proceedings of the 21st International Conference on World Wide Web—WWW'12, Lyon, France, 16–20 April 2012; Association for Computing Machinery: New York, NY, USA, 2012; pp. 519–528.
34. Burbach, L.; Halbach, P.; Zieffle, M.; Calero Valdez, A. Opinion formation on the internet: The influence of personality, network structure, and content on sharing messages online. *Front. Artif. Intell.* **2020**, *3*, 45. [CrossRef]
35. Büchi, M.; Festic, N.; Latzer, M. The Chilling Effects of Digital Dataveillance: A Theoretical Model and an Empirical Research Agenda. *Big Data Soc.* **2022**, *9*, 20539517211065368. [CrossRef]
36. Hoffrage, U.; Marewski, J.N. Social Scoring als Mensch-System-Interaktion. In *Social Credit Rating*; Everling, O., Ed.; Springer Fachmedien: Wiesbaden, Deutschland, 2020; pp. 305–329.

37. Kostka, G. China's social credit systems and public opinion: Explaining high levels of approval. *New Media Soc.* **2019**, *21*, 1565–1593. [CrossRef]
38. Eling, M.; Nuessle, D.; Staubli, J. The impact of artificial intelligence along the insurance value chain and on the insurability of risks. *Geneva Pap. Risk Insur.—Issues Pract.* **2021**, *47*, 205–241. [CrossRef]
39. Riikinen, M.; Saarijärvi, H.; Sarlin, P.; Lähteenmäki, I. Using artificial intelligence to create value in insurance. *Int. J. Bank Mark.* **2018**, *36*, 1145–1168. [CrossRef]
40. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.S.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131. [CrossRef]
41. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
42. Upadhyay, A.K.; Khandelwal, K. Applying artificial intelligence: Implications for recruitment. *Strateg. HR Rev.* **2018**, *17*, 255–258. [CrossRef]
43. Nawaz, N.; Mary, A. Artificial intelligence chatbots are new recruiters. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 1–5. [CrossRef]
44. Sourdin, T. *Judges, Technology and Artificial Intelligence: The Artificial Judge*; Edward Elgar Publishing: Cheltenham, UK, 2021.
45. Vermeys, N. The computer as the court: How will artificial intelligence affect judicial processes? In *New Pathways to Civil Justice in Europe*; Kramer, X., Biard, A., Hoevenaars, J., Themeli, E., Eds.; Springer: Basel, Switzerland, 2021.
46. Agudo, U.; Matute, H. The influence of algorithms on political and dating decisions. *PLoS ONE* **2012**, *16*, e0249454. [CrossRef]
47. Scavarelli, C.M. The Future of Dating (No. 6) [Song]. 2018. Available online: <https://soundcloud.com/user-145965453> (accessed on 4 April 2023).
48. ABC News. Family of American Terror Victim Asks Supreme Court to Curb Immunity for Social Media. Available online: <https://abcnews.go.com/Politics/family-american-terror-victim-asks-supreme-court-curb/story?id=96463560> (accessed on 4 April 2023).
49. LegalEagle. The Supreme Court Could Destroy the Internet Next Week. 2023. Available online: <https://www.youtube.com/watch?v=hzNo5IZCq5M> (accessed on 4 April 2023).
50. Jonsson, O.; de Tena, C.L. European Tech Insights 2021. Part II Embracing and Governing Technological Disruption. Center for Governance of Change. 2021. Available online: <https://docs.ie.edu/cgc/IE-CGC-European-Tech-Insights-2021-%28Part-II%29.pdf> (accessed on 4 April 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Forecasting Neonatal Mortality in Portugal [†]

Rodrigo B. Ventura ¹, Filipe M. Santos ¹, Ricardo M. Magalhães ¹, Cátia M. Salgado ¹, Vera Dantas ²,
Matilde V. Rosa ², João M. C. Sousa ^{1,*} and Susana M. Vieira ¹

¹ IDMEC, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal; rodrigo.boal.ventura@tecnico.ulisboa.pt (R.B.V.); filipempantos@tecnico.ulisboa.pt (F.M.S.); ricardo.m.a.magalhaes@tecnico.ulisboa.pt (R.M.M.); catia.salgado@tecnico.ulisboa.pt (C.M.S.); susana.vieira@tecnico.ulisboa.pt (S.M.V.)

² Social Data Lab, 1000-179 Lisbon, Portugal; veradantas@socialdatalab.pt (V.D.); matildevalenterosa@socialdatalab.pt (M.V.R.)

* Correspondence: jmsousa@tecnico.ulisboa.pt

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: In order to achieve a more efficient allocation of healthcare resources in the near future, it is crucial to understand the patterns and causes of excess mortality and hospitalizations. Neonatal mortality still poses a significant challenge, particularly in developed nations where the mortality rates are already low and healthcare resources are generally available to most of the population. Furthermore, the low mortality rates mean that the data available for modeling are often very limited, restricting the modeling methods that can be used. It is also important that the chosen methods allow for explainable, non-black-box models that can be interpreted by healthcare professionals. Considering these challenges, the work hereby presented thoroughly analyzed the time series of the neonatal mortality rates in Portugal between 2014 and 2019 in terms of trend and seasonal patterns. The applicability and performance of different data-based methods were also analyzed. Furthermore, the mortality rates were studied in terms of their relation to environmental variables, such as temperature and air pollution indicators, with the goal of establishing causal relations between such variables and excess mortality. The preliminary results show that ARMA, neural and fuzzy models are able to forecast the studied mortality rates with good performance. In particular, neural models have the best predictive performance, while fuzzy models are well suited to obtain interpretable models with acceptable performance.

Keywords: neonatal mortality; time series forecasting; ARMA; neural models; fuzzy models

Citation: Ventura, R.B.; Santos, F.M.; Magalhães, R.M.; Salgado, C.M.; Dantas, V.; Rosa, M.V.; Sousa, J.M.C.; Vieira, S.M. Forecasting Neonatal Mortality in Portugal. *Eng. Proc.* **2023**, *39*, 89. <https://doi.org/10.3390/engproc2023039089>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 14 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The survival and prosperity of children are at the heart of any thriving society. This is especially true for developed countries, where advances in medical technology and better living standards have led to a longer life expectancy, altering the population distribution among different age groups.

Children are the future labor force, responsible for keeping the economy moving and caring for the older generation while preparing the next generation for the challenges ahead. Regardless of a country's location or socioeconomic status, children are the foundation of any nation's success, and their welfare is vital to ensuring a prosperous future.

One of the most significant achievements in global health is the decline in child mortality rates [1]. The number of deaths in children under the age of five has fallen drastically from 19.6 million in 1950 to 9.6 million in 2000 and 5.0 million in 2019 [2], indicating a worldwide decrease of 74.5% over the last 70 years.

However, there is still much work to be done. The United Nations' Sustainable Development Goals aim to end preventable deaths of newborns and children under the

age of five by 2030 [3], with a target of reducing neonatal mortality to no more than 12 per 1000 live births and under-five mortality to no more than 25 per 1000 live births.

Child mortality rates are influenced by a variety of factors, which vary greatly across different regions of the world. Developing countries face significant challenges such as malnutrition, lack of access to clean water and inadequate healthcare and public safety infrastructure [4]. Addressing these issues is crucial to reducing child mortality rates. In contrast, European nations have some of the lowest child mortality rates globally [5], reflecting their strong commitment to children's health and well-being. Ultimately, a country's geographic location, socioeconomic status and cultural profile have a significant impact on child mortality rates and the factors that determine them.

Neonatal mortality refers to the death of infants within the first 28 days of life [6]. It is still one of the most significant challenges facing the global health community, particularly in low- and middle-income countries, and a critical indicator of a country's health status, healthcare system and socioeconomic development. According to the World Health Organization (WHO), an estimated 2.4 million neonatal deaths occurred globally in 2019, less than half of the 5.0 million that occurred in 1990 [7].

Neonatal mortality is preventable, and several interventions can help to reduce it. These include improving access to quality healthcare services, particularly during pregnancy, childbirth and the postnatal period [8]. Ensuring adequate nutrition for pregnant women and infants is also crucial [9], as well as promoting healthy behaviors such as breastfeeding and providing essential newborn care. Reducing neonatal mortality has far-reaching benefits beyond saving lives. It can lead to a healthier population, increased economic productivity [10] and improved social outcomes. Addressing the underlying socioeconomic determinants of neonatal mortality may also play a critical role.

Examining neonatal mortality data in a country such as Portugal, where rates are already extremely low, is still crucial for several reasons. First, even in countries with low neonatal mortality rates, there can be regional or socioeconomic disparities that need to be addressed. By analyzing the data, policymakers can identify areas where there may be higher rates of neonatal mortality and develop targeted interventions to address these disparities.

Secondly, analyzing neonatal mortality data can help to identify trends and patterns over time. Even if the overall rates are low, there may be changes in specific causes of neonatal mortality or demographic groups that require attention. For example, there may be an increase in the number of premature births, which would require additional resources and interventions to address.

Thirdly, neonatal mortality studies can help to evaluate the effectiveness of interventions and policies aimed at reducing neonatal mortality. Even in countries with low rates, there may still be room for improvement, and analyzing the data can help to identify areas where additional efforts are needed.

While countries such as Portugal may have made significant progress in reducing neonatal mortality, there is still a long way to go to achieve the Sustainable Development Goals. Therefore, it is critical to raise awareness and advocate for continued efforts to reduce neonatal mortality globally. By highlighting the importance of this issue and sharing best practices, policymakers and public health officials can continue to work towards improving the health outcomes of mothers and newborns worldwide.

The aim of this study is to forecast and explain neonatal mortality in Portugal using machine learning techniques. We analyzed a range of mortality and environmental variables to identify the most significant predictors of neonatal mortality. Furthermore, we developed three predictive models that can forecast neonatal mortality rates in Portugal over the next couple of years. The results of this study will provide valuable insights into the determinants of neonatal mortality in Portugal and may have implications for the design of public health interventions to reduce neonatal mortality rates in the country.

The remainder of this paper is structured as follows. Section 2 shows the data used in this study, as well as how the data was prepared. In Section 3, a forecast of neonatal

mortality is presented based on the past values of the time series using ARMA models. In Sections 4 and 5, models using the past values of the time series as well as exogenous values are presented, namely neural and Takagi–Sugeno fuzzy models. Finally, in Section 6, an overview of the conclusions drawn from this study is presented.

2. Data Preparation

For this study, we had access to data regarding all neonatal deaths provided by Directorate-General for Health, “Direção Geral de Saúde” (DGS), from January 2014 to December 2019. Furthermore, we used exogenous data to develop the models in Sections 4 and 5, such as temperature and various pollution indicators.

Regarding the neonatal mortality data, firstly, the data were aggregated by month of death as we had the exact date. Then, data were divided into a trend and deviation from such trend, as in [11]. Firstly, we computed the moving average by considering the current month and the 11 that preceded it, which aided in smoothing out data fluctuations and improving the representation of the underlying trend, as in [12]. In order words, the average calculated for December 2014 uses data from January 2014 to December 2014. Note that this is the first sample that was actually used, because samples from previous months would require data from 2013.

Next, we calculated the deviation from the moving average to each real value, which highlighted unusual patterns in the data. For this, data were divided into training and test sets. It was established that data from December 2014 to December 2017 is used as training data, and those from January 2018 to December 2019 are the test data.

Regarding exogenous variables, such data were provided for each municipality individually, similar to [13] but at a larger scale. However, this study focuses on forecasting mortality in the country, so national values would be preferable. To obtain a national value for each exogenous variable for each month, the average weighted by the population was calculated, in the following way:

$$X(t) = \frac{\sum_{i=1}^N x_i(t)p_i(t)}{\sum_{i=1}^N p_i(t)} \tag{1}$$

where t represents time; i a municipality; x_i a raw (by municipality) exogenous variable, such as temperature; X the respective transformed (nationwide) variable; and p_i the population of a municipality. By calculating the weighted sum in this way instead of a simple average, we guarantee that the values of the more heavily populated areas have more impact on the final value since more people are exposed to those conditions. Table 1 presents the exogenous variables used for the modeling procedure.

Table 1. Exogenous variables.

Variable	Unit
Mean Temperature	°C
NO ₂ Concentration	µg/m ³
PM10 Concentration	µg/m ³
PM2.5 Concentration	µg/m ³

3. ARMA Modeling

As an initial approach, an AutoRegressive Moving Average (ARMA) model was utilized to predict neonatal mortality in 2018 and 2019 due to its widespread usage and established effectiveness in modeling time series data, see [14–16]. ARMA models have proven to be particularly useful in scenarios where a clear dependence exists between the current value of the time series and its past values.

We employed an ARMA model to predict the moving average segment of the time series. Subsequently, we created another ARMA model to predict the deviation segment of the time series. For both ARMA models, data regarding the last 12 months were utilized. Eventually, summing up the two predictions derived the ultimate prediction value, accounting for both the trend and variations in the data, which enhanced the prediction performance.

Overall, the procedure facilitated the accurate analysis of the neonatal mortality time series and generated precise predictions for future values. Incorporating the moving average and deviations from it, along with utilizing ARMA models for prediction, allowed the trend and variations in the data to be accounted for, resulting in more accurate predictions.

Although there are differences between the predicted and actual values, this can be attributed to noise and variability in the data, as well as the assumptions and limitations of the ARMA models utilized. The obtained results (Figure 1) indicated that ARMA models are effective in capturing the underlying patterns and dependencies in the data, with an RMSE of 3.042 and a VAF of 0.36 indicating an average difference of 3 units from the actual values, which is reasonable given the time series data's range of values. However, it is crucial to note that the RMSE is just one measure of prediction accuracy, and there may be other metrics or measures that can provide different insights [17].

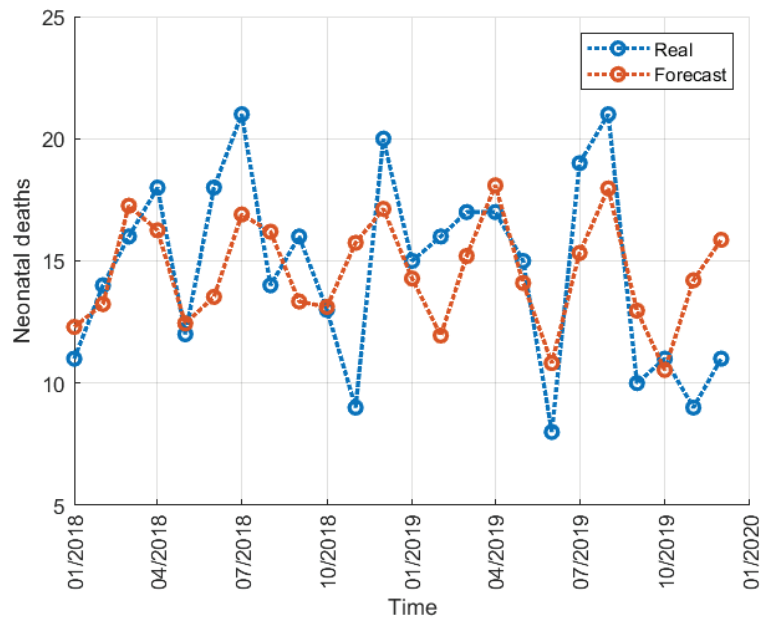


Figure 1. Neonatal deaths forecast using yearly and monthly ARMA model.

4. Neural Network Modeling

ARMA models used only the previous values of the time series to predict its future values. Considering such results, the next step is to use also exogenous variables with the goal of achieving better predictive performance. Exogenous variables are variables that are measured alongside the forecast time series and can help determine its future values. In the context of the work hereby presented, the exogenous variables are the environmental variables presented in Table 1.

In this section, we propose to use an artificial neural network (NN) model, as in [18,19], to incorporate the exogenous variables into the modeling process. The proposed NN model takes as inputs the exogenous variables, as well as the previous values of the time series. Following the procedure for the ARMA model, the 12 previous values of the 4 exogenous

variables as well as the 12 previous values of the time series moving average and deviations are given as input to the NN model, which then forecasts the mortality for the next time period. Therefore, the proposed network has a total of 72 inputs and one output. Considering such characteristics, the proposed NN is a shallow network with 73 neurons in the hidden layer. Table 2 shows the other relevant properties of the NN model.

Table 2. Neural network model parameters.

Parameter	Value
Structure	1 layer, 73 neurons
Activation	Relu
Solver	lbfgs
Max. Iterations	500

The results (Figure 2) show that the NN model achieved better predictions than the ARMA models with an RMSE of 2.394 and a VAF of 0.61. The improvement in RMSE may be attributed to the inclusion of exogenous variables or to the non-linear nature of the NN. These results suggest that incorporating external factors and utilizing non-linear models can enhance the accuracy and predictive power of time series models.

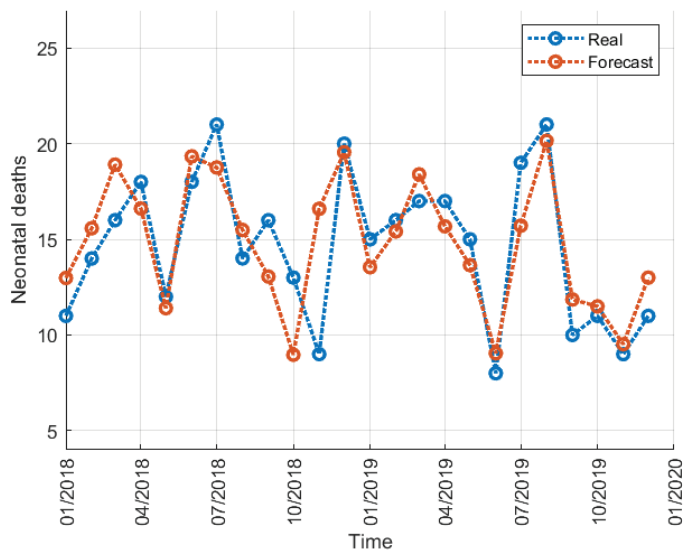


Figure 2. Neonatal deaths forecast using the NN model.

5. Fuzzy Modeling

NN models are known as black-box models [20], meaning that, while their predictive performance may be outstanding, they are also very hard to interpret, and the reasoning between the inputs and the output is often not clear. This problem, known as interpretability, is one of the most important yet still open problems in the field of machine learning [21,22].

While much research has been carried out in the recent years on this topic, the chosen approach in the work hereby presented is the use of fuzzy logic to model the forecasting problem. Fuzzy inference systems are a class of data-based learning algorithms that are particularly well suited for modeling problems where interpretability is a priority [23–25]. Furthermore, Takagi–Sugeno fuzzy systems are universal learners [26], meaning that, similarly to NN models, they are able to approximate any function provided that sufficient degrees of freedom are given to the model.

The modeling approach for the fuzzy models is exactly the same as the one described for the NN in Section 4, with the same inputs and output. The parameters of the obtained fuzzy model are as shown in Table 3.

Table 3. TSK model parameters.

Parameter	Value
Clustering Method	Fuzzy C-means
Cluster Validation	Silhouette Coefficient
Number of Rules	31
Consequent Type	Affine

Regarding the predictive performance of the obtained model, Figure 3 shows the forecasting results which correspond to an RMSE of 3.569 and VAF of 0.15, meaning that the model is not so good as the previous models. However, it is important to mention that the main advantage of fuzzy models is not necessarily their predictive performance, but their interpretability. So, the predictive performance is still acceptable and comparable to the other models. Still, regarding the topic of interpretability, the obtained fuzzy model has a large number of fuzzy rules that makes its interpretability hard without a further refinement of its rules by merging the fuzzy sets. However, the fuzzy model is still by far the most interpretable model obtained in this work.

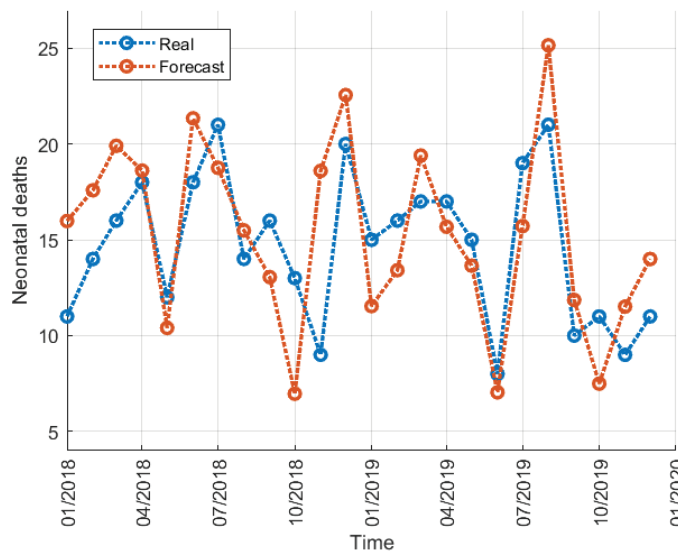


Figure 3. Neonatal deaths forecast using TSK fuzzy system.

6. Discussion and Conclusions

In this study, three different methodologies to forecasting mortality were utilized. The first one was a junction of two ARMA models, one to predict the trend and another one to predict the deviation from the trend. The results obtained show that the majority of the variations that can be observed in the time series can be predicted by just looking at the mortality rates of the last 12 months. With the aim of improving the prediction made by the ARMA models, an NN model was utilized due to its nonlinear nature and its capability to include exogenous variables to make predictions. The results improved substantially as expected. Finally, a Takagi–Sugeno fuzzy model was utilized to create IF–THEN rules to predict mortality, as these are more interpretable to a human. This model

achieved reasonable predictions when paired with a high number of rules, which hurts the interpretability of the task at hand.

Overall, according to the metrics (RMSE and VAF), the NN model is clearly the best performer. However, upon comparing Figures 1 and 3, it is observable that the reason why ARMA models perform so well is that they always make more conservative predictions. In other words, the model underestimates periods of higher mortality and overestimates periods of lower mortality, always keeping its prediction close to the average value of the time series. On the other hand, the fuzzy model more effectively captures the magnitude of the highs and the lows of the time series. The model only performs poorly on both metrics because sometimes the prediction of peaks of mortality is offset by a month, dooming the score on both metrics.

When the prediction is offset from the real curve, it can be observed that prediction comes before the actual peak. These might be useful in certain scenarios because if a peak does not occur in the predicted month, it will most likely occur in the following month. In future work, other interpretable modeling techniques should be used to try to identify which variables and in what instances are critical to predicting neonatal mortality.

Author Contributions: Conceptualization, V.D., M.V.R., S.M.V. and J.M.C.S.; methodology, R.B.V., F.M.S., R.M.M., C.M.S., V.D., M.V.R., S.M.V. and J.M.C.S.; software, R.B.V., F.M.S. and R.M.M.; validation, C.M.S., V.D., M.V.R., S.M.V. and J.M.C.S.; formal analysis, C.M.S., S.M.V. and J.M.C.S.; resources, S.M.V., and J.M.C.S.; data curation, R.B.V., F.M.S. and R.M.M.; writing—original draft preparation, all authors; writing—review and editing, C.M.S., V.D., M.V.R., S.M.V. and J.M.C.S.; supervision, S.M.V. and J.M.C.S.; project administration, S.M.V.; funding acquisition, C.M.S., V.D., M.V.R. and S.M.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Portuguese Foundation for Science Technology, FCT, through IDMEC, under LAETA, by the FCT project AI4Life DSAIPA/DS/0054/2019, and by the FCT PhD scholarships 2022.14216.BDANA, 2022.12077.BDANA and MPP2030-FCT ID 22405888735. No potential competing interests are reported by the authors.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to anonymisation of data.

Informed Consent Statement: Patient consent was waived due to anonymisation.

Data Availability Statement: Data is available for researchers under the conditions defined by the Directorate-General for Health, “Direção Geral de Saúde” (DGS), Portugal.

Conflicts of Interest: No potential competing interests are reported by the authors.

References

- Centers for Disease Control and Prevention: Ten Great Public Health Achievements—Worldwide, 2001–2010. May 2011. Available online: <https://pubmed.ncbi.nlm.nih.gov/21697806/> (accessed on 5 July 2022).
- Wang, H.; Abbas, K.M.; Abbasifard, M.; Abbasi-Kangevari, M.; Abbastabar, H.; Abd-Allah, F.; Abdelalim, A.; Abolhassani, H.; Abreu, L.G.; Abrigo, M.R.; et al. Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950–2019: A comprehensive demographic analysis for the Global Burden of Disease Study 2019. *Lancet* **2020**, *396*, 1160–1203. [CrossRef] [PubMed]
- Kancherla, V.; Roos, N.; Walani, S.R. Relationship between achieving Sustainable Development Goals and promoting optimal care and prevention of birth defects globally. *Birth Defects Res.* **2022**, *114*, 773–784. [CrossRef] [PubMed]
- Saheed, R.; Hina, H.; Shahid, M. Water, Sanitation and Malnutrition in Pakistan: Challenge for Sustainable Development. *Glob. Econ. Rev.* **2021**, *6*, 1–14. [CrossRef]
- Raghupathi, V.; Raghupathi, W. The influence of education on health: An empirical assessment of OECD countries for the period 1995–2015. *Arch. Public Health* **2020**, *78*, 20. [CrossRef] [PubMed]
- Lee, H.Y.; Leslie, H.H.; Oh, J.; Kim, R.; Kumar, A.; Subramanian, S.V.; Kruk, M.E. The association between institutional delivery and neonatal mortality based on the quality of maternal and newborn health system in India. *Sci. Rep.* **2022**, *12*, 6220. [CrossRef] [PubMed]
- Mitiku, H.D. Neonatal mortality and associated factors in Ethiopia: A cross-sectional population-based study. *BMC Women's Health* **2021**, *21*, 156. [CrossRef] [PubMed]
- Tekelab, T.; Choienta, C.; Smith, R.; Loxton, D. The impact of antenatal care on neonatal mortality in sub-Saharan Africa: A systematic review and meta-analysis. *PLoS ONE* **2019**, *14*, e0222566. [CrossRef]

9. Christian, P.; Mullany, L.C.; Hurley, K.M.; Katz, J.; Black, R.E. Nutrition and maternal, neonatal, and child health. In *Seminars in Perinatology*; WB Saunders: Philadelphia, PA, USA, 2015; pp. 361–372.
10. Setyadi, S.; Syaifudin, R.; Desmawan, D. Human Capital and Productivity: A Case Study of East Java. *Econ. Dev. Anal. J.* **2020**, *9*, 202–207. [CrossRef]
11. Gupta, R.; Pal, S.K. Trend Analysis and Forecasting of COVID-19 outbreak in India. *MedRxiv* **2020**. [CrossRef]
12. Büyüksahin, Ü.Ç.; Ertekin, Ş. Improving forecasting accuracy of time series data using a new ARIMA-NARX hybrid method and empirical mode decomposition. *Neurocomputing* **2019**, *361*, 151–163. [CrossRef]
13. Padilla, C.; Lalloué, B.; Pies, C.; Lucas, E.; Zmirou-Navier, D.; Séverine, D. An ecological study to identify census blocks supporting a higher burden of disease: Infant mortality in the lille metropolitan area, France. *Matern. Child Health J.* **2014**, *18*, 171–179. [PubMed]
14. Malladi, R.K.; Dheeriyaa, P.L. Time series analysis of cryptocurrency returns and volatilities. *J. Econ. Financ.* **2021**, *45*, 75–94. [CrossRef]
15. Kumar, J.; Singh, A.K. Performance assessment of time series forecasting models for cloud datacenter networks' workload prediction. *Wirel. Pers. Commun.* **2021**, *116*, 1949–1969. [CrossRef]
16. Gu, Q.; Dai, Q. A novel active multi-source transfer learning algorithm for time series forecasting. *Appl. Intell.* **2021**, *51*, 1326–1350. [CrossRef]
17. Hewamalage, H.; Ackermann, K.; Bergmeir, C. Forecast evaluation for data scientists: Common pitfalls and best practices. *Data Min. Knowl. Discov.* **2023**, *37*, 788–832. [PubMed]
18. Salis, V.E.; Kumari, A.; Singh, A. Prediction of gold stock market using hybrid approach. In *Emerging Research in Electronics, Computer Science and Technology*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 803–812.
19. ul, Sami, I.; Junejo, K.N. Predicting future gold rates using machine learning approach. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 92–99.
20. Letzgs, S. Towards transparent NARX wind turbine power curve models. *arXiv* **2022**, arXiv:2210.12104.
21. Molnar, C.; Casalicchio, G.; Bischl, B. Interpretable machine learning—a brief history, state-of-the-art and challenges. In Proceedings of the ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, 14–18 September 2020; Springer International Publishing: Cham, Switzerland, 2021; pp. 417–431.
22. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.* **2020**, *32*, 18069–18083. [CrossRef]
23. Mencar, C.; Alonso, J.M. Paving the way to explainable artificial intelligence with fuzzy modeling: Tutorial. In Proceedings of the Fuzzy Logic and Applications: 12th International Workshop, WILF 2018, Genoa, Italy, 6–7 September 2018; Revised Selected Papers; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 215–227.
24. Magdalena, L. Fuzzy Systems Interpretability: What, Why and How. In *Fuzzy Approaches for Soft Computing and Approximate Reasoning: Theories and Applications: Dedicated to Bernadette Bouchon-Meunier*. 2021. pp. 111–122. Available online: https://link.springer.com/chapter/10.1007/978-3-030-54341-9_10 (accessed on 1 June 2022).
25. Shabelnikov, A.N.; Kovalev, S.M.; Sukhanov, A.V. Interpretability of fuzzy temporal models. In *Proceedings of the Third International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'18)*; Springer International Publishing: New York, NY, USA, 2019; Volume 13, pp. 223–234.
26. Xie, R.; Wang, S. A wide interpretable Gaussian Takagi–Sugeno–Kang fuzzy classifier and its incremental learning. *Knowl.-Based Syst.* **2022**, *241*, 108203. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

A Comparison between Successive Estimate of TVAR(1) and TVAR(2) and the Estimate of a TVAR(3) Process [†]

Johannes Korte *, Jan Martin Brockmann and Wolf-Dieter Schuh

Institute of Geodesy and Geoinformation, University of Bonn, Nussallee 17, 53115 Bonn, Germany; brockmann@geod.uni-bonn.de (J.M.B.); schuh@uni-bonn.de (W.-D.S.)

* Correspondence: korte@geod.uni-bonn.de; Tel.: +49-228-73-3576

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: In time series analyses, the auto-regressive (AR) modelling of zero mean data is widely used for system identification, signal decorrelation, detection of outliers and forecasting. An AR process of order p is uniquely defined by p coefficients and the variance in the noise. The roots of the characteristic polynomial can be used as an alternative parametrization of the coefficients, which can be used to construct a continuous covariance function of the AR process or to verify that the AR process is stationary. In a previous study, we introduced an AR process of time variable coefficients (TVAR process) in which the movement of the roots was specified as a polynomial of order one. Until now, this method was analytically derived only for TVAR processes of orders one and two. Thus, higher-level processes had to be assembled by the successive estimation of these process orders. In this contribution, the analytical solution for a TVAR(3) process is derived and compared to the successive estimation using a TVAR(1) and TVAR(2) process. We will apply the proposed approach to a GNSS time series and compare the best-fit TVAR(3) process with the best-fit composition of TVAR(2) and TVAR(1) process.

Keywords: AR process of order 3; non-stationarity; time-varying AR coefficients; time-variable roots from polynomials

Citation: Korte, J.; Brockmann, J.M.; Schuh, W.-D. A Comparison between Successive Estimates of TVAR(1) and TVAR(2) and the Estimate of a TVAR(3) Process. *Eng. Proc.* **2023**, *39*, 90. <https://doi.org/10.3390/engproc2023039090>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 18 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The auto-regressive process is a means in time series analysis and, among other things, this method is used to estimate discrete covariance functions (see [1] [p. 32, eq. (182)]) or to decorrelate observations by filtering [2–4]. Under the assumption of a non-stationary process, there are two possible states: first, that the roots of the characteristic polynomial are no longer in the unit circle, and therefore no covariances can be calculated, and secondly, the case we are looking at here, where the root changes over time, but always stays in the unit circle. This case has often appeared in the literature, but the TVAR coefficients, when used, were given a form of motion like polynomial motions (see [5,6]), trigonometric function [7], modified Legendre polynomials [8] and spherical sequences [9]. What all these methods often overlook is the resulting movement of the roots, as can be seen, for example, in [10]. In [11], it was derived how a TVAR process has to be calculated successively from TVAR(1) and TVAR(2) process estimates for a linear root movement. This method is now to be extended so that it can be applied directly to the TVAR(3) process without testing different series of TVAR(1) and TVAR(2) processes. Divided into five chapters, this extension is presented here. In Section 2, the necessary condition for the linear roots of a TVAR process of arbitrary order is derived again. In Section 3, additional restrictions are derived as sufficient conditions. In Section 4, the theory is tested on a GNSS time series and contrasted with successive estimations. The discussion of the results, as well as an overlook of further research directions, can be found in Section 5.

2. Successive Estimation Using TVAR(1) and TVAR(2) Processes

Ref. [10] defines the time variable auto-regressive process of the order p (TVAR (p) process) using the recursion formula

$$S_t = \alpha_1(t)S_{t-1} + \alpha_2(t)S_{t-2} + \dots + \alpha_p(t)S_{t-p} + \mathcal{E}_t. \tag{1}$$

Here $\alpha_1(t), \alpha_2(t), \dots, \alpha_p(t)$ are the time-varying coefficients of the TVAR(p) process, and \mathcal{E}_t is an i.i.d. sequence with variance $\sigma_{\mathcal{E}}^2$. As long as a single time $t = \tau$ is considered separately, the representation in (1) corresponds to a time-stable AR process of order p (TSAR(p) process):

$$S_t = \alpha_1(\tau)S_{t-1} + \alpha_2(\tau)S_{t-2} + \dots + \alpha_p(\tau)S_{t-p} + \mathcal{E}_t.$$

According to [12] [p. 167], the TSAR Process is stationary if the roots (P_1, P_2, \dots, P_p) of the characteristic polynomial

$$\begin{aligned} \chi(x) &= x^p - \alpha_1(\tau)x^{p-1} - \alpha_2(\tau)x^{p-2} - \dots - \alpha_p(\tau) \\ &= (x - P_1(\tau))(x - P_2(\tau))\dots(x - P_p(\tau)) \end{aligned}$$

are in the unit circle ($\|P_k(\tau)\| < 1 \forall k = 1, 2, \dots, p$). Ref. [11] has shown how TVAR processes with linear root movements are successively estimated using TVAR(1) and TVAR(2) processes, and the advantages of this method compared to the estimation of time constant AR processes. The general TVAR estimate consists of two steps: First, the coefficients of the TVAR process are replaced by polynomials, where the order of the polynomial is equal to the order of the coefficient

$$\alpha_k(t) = \sum_{j=0}^k \beta_j^{(k)} t^j. \tag{2}$$

The general solution of $\beta_j^{(k)}$ can be calculated by a least squares adjustment using the system of equations

$$\begin{aligned} \begin{bmatrix} S_p \\ S_{p+1} \\ S_{p+2} \\ \dots \\ S_{p+n} \end{bmatrix} &= TM\beta, \text{ for } n \geq \frac{p^2 + 3p}{2}, \text{ with} \\ T &= \begin{bmatrix} S_{p-1} & S_{p-2} & S_{p-3} & \dots & S_0 \\ S_p & S_{p-1} & S_{p-2} & \dots & S_1 \\ \dots & \dots & \dots & \dots & \dots \\ S_{n-1} & S_{n-2} & S_{n-3} & \dots & S_{n-p} \end{bmatrix}, \\ M &= \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & t & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & t & 0 & \dots & 0 & t^2 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & 0 & t & \dots & 0 & 0 & t^2 & \dots & 0 & t^3 & \dots & 0 \\ & & & \ddots & & & & & \ddots & & & & \ddots & & & & \\ 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & t & 0 & 0 & \dots & t^2 & 0 & \dots & t^3 \end{bmatrix} \text{ and} \\ \beta &= [\beta_0^{(1)} \beta_0^{(2)} \dots \beta_0^{(p)} \beta_1^{(1)} \beta_1^{(2)} \dots \beta_1^{(p)} \beta_2^{(2)} \dots \beta_2^{(p)} \dots \beta_{p-1}^{(p-1)} \beta_{p-1}^{(p)} \beta_p^{(p)}]^T. \end{aligned}$$

The order of the TVAR process is irrelevant for this calculation. Second, additional restrictions are put in place to ensure linear root movements. For orders 1 and 2, these restrictions were set out in the paper [11]. Now, this procedure is to be extended to the TVAR(3) process.

3. Restrictions of TVAR(3) Process with Linear Root Motion

The condition for the linear root movement arises from the transformation from the parameters $\alpha_k(t)$ ($\beta_j^{(k)}$) to the roots $P_k(t)$, respectively. Therefore, the representation of the roots from the coefficients for the time-constant case is considered first.

3.1. Calculation of the Roots from the Time-Stable Coefficients

The roots of a third-order polynomial

$$b(x) = x^3 - \alpha_1(\tau)x^2 - \alpha_2(\tau)x - \alpha_3(\tau) \tag{3}$$

can be calculated after [13] [p. 22f] in a two-step procedure. The first step eliminates the monomial x^2 in (3) by substituting x with $y = x - \frac{\alpha_1(\tau)}{3}$. For the resulting polynomial over y

$$b(y) = y^3 + \underbrace{\left[-\frac{\alpha_1^2(\tau)}{3} - \alpha_2(\tau) \right]}_{c_1(\tau)} y - \underbrace{\left[2\frac{\alpha_1^3(\tau)}{27} - \frac{\alpha_1(\tau)\alpha_2(\tau)}{3} - \alpha_3(\tau) \right]}_{c_2(\tau)}, \tag{4}$$

the first roots $\bar{P}_1(\tau)$ of the Equation (4) can be determined using the Cardano solution:

$$\bar{P}_1(\tau) = S_1 + S_2,$$

where

$$S_1 = \sqrt[3]{-\frac{c_2(\tau)}{2} + \sqrt{\left(\frac{c_2(\tau)}{2}\right)^2 + \left(\frac{c_1(\tau)}{3}\right)^3}}$$

and $S_2 = \sqrt[3]{-\frac{c_2(\tau)}{2} - \sqrt{\left(\frac{c_2(\tau)}{2}\right)^2 + \left(\frac{c_1(\tau)}{3}\right)^3}}.$

This root is back-substituted to find the roots of the characteristic polynomial of the third order in (3):

$$P_1(\tau) = \bar{P}_1(\tau) + \frac{\alpha_1(\tau)}{3} = S_1 + S_2 + \frac{\alpha_1(\tau)}{3}.$$

The other two roots then result from

$$P_2(\tau) = -\frac{S_1 + S_2}{2} + \frac{\alpha_1(\tau)}{3} + i\sqrt{3}\frac{S_1 - S_2}{2}$$

and $P_3(\tau) = -\frac{S_1 - S_2}{2} + \frac{\alpha_1(\tau)}{3} - i\sqrt{3}\frac{S_1 - S_2}{2}.$

3.2. Restrictions for Linear Root Motion

These findings for a discrete point in time can be transferred to functions over time to derive the restrictions for linear root movements. However, this chapter does not derive the general condition for linear root motion; instead, the three sumands of $P_1(t)$

$$S_1(t) = \sqrt[3]{-\frac{c_2(t)}{2} + \sqrt{\left(\frac{c_2(t)}{2}\right)^2 + \left(\frac{c_1(t)}{3}\right)^3}} \stackrel{!}{=} f_1 + g_1t, \quad (5)$$

$$S_2(t) = \sqrt[3]{-\frac{c_2(t)}{2} - \sqrt{\left(\frac{c_2(t)}{2}\right)^2 + \left(\frac{c_1(t)}{3}\right)^3}} \stackrel{!}{=} f_2 + g_2t \quad (6)$$

$$\text{and} \quad -\frac{\alpha_1(t)}{3} \stackrel{!}{=} f_3 + g_3t \quad (7)$$

are individually converted into linear functions by restrictions. Because of the three conditions, all linear combinations of these functions (S_1 , S_2 and α_1) are automatically linear. This particularly applies to $P_1(t)$, $P_2(t)$ and $P_3(t)$.

The conditions in (5), (6) and (7) have to be rewritten according to the conditions at the parameters $\beta_j^{(k)}$. After the derivation of the coefficients in (2), $\alpha_1(t) = \beta_0^{(1)} + \beta_1^{(1)}t$ is already a polynomial of first degree and the condition (7) is always met. To simplify the conditions in (5) and (6), both conditions are potentiated by 3 and then linked to each other via addition or multiplication:

$$(f_1 + g_1t)^3 + (f_2 + g_2t)^3 \stackrel{!}{=} -c_2(t) \quad \text{and} \quad (8)$$

$$(f_1 + g_1t)(f_2 + g_2t) \stackrel{!}{=} -\frac{c_1(t)}{3}. \quad (9)$$

This creates two new conditions: the first contains a third-order polynomial and the second a second-order polynomial, for which the equations are exactly satisfied if both sides have the same polynomial coefficients. Thus, the polynomial of order 3 in (8) includes four restrictions

$$f_1^3 + f_2^3 = 2\left(\frac{\beta_0^{(1)}}{3}\right)^3 + \frac{\beta_0^{(1)}\beta_0^{(2)}}{3} + \beta_0^{(3)} \quad (10)$$

$$3(f_1^2g_1 + f_2^2g_2) = 2\left(\frac{\beta_0^{(1)}}{3}\right)^2\beta_1^{(1)} + \frac{\beta_0^{(1)}\beta_1^{(2)} + \beta_1^{(1)}\beta_0^{(2)}}{3} + \beta_1^{(3)} \quad (11)$$

$$3(f_1g_1^2 + f_2g_2^2) = 2\beta_0^{(1)}\left(\frac{\beta_1^{(1)}}{3}\right)^2 + \frac{\beta_0^{(1)}\beta_2^{(2)} + \beta_1^{(1)}\beta_1^{(2)}}{3} + \beta_2^{(3)} \quad (12)$$

$$g_1^3 + g_2^3 = 2\left(\frac{\beta_1^{(1)}}{3}\right)^3 + \frac{\beta_1^{(1)}\beta_2^{(2)}}{3} + \beta_3^{(3)}, \quad (13)$$

and the polynomial of order 2 in (9) adds three further restrictions

$$f_1f_2 = \left(\frac{\beta_0^{(1)}}{3}\right)^2 + \frac{\beta_0^{(2)}}{3} \quad (14)$$

$$3(f_1g_2 + f_2g_1) = 2\frac{\beta_0^{(1)}\beta_1^{(1)}}{3} + \beta_1^{(2)} \quad (15)$$

$$g_1g_2 = \left(\frac{\beta_1^{(1)}}{3}\right)^2 + \frac{\beta_2^{(2)}}{3}. \quad (16)$$

With the help of the formulas (10), (13), (14) and (16), the variables

$$(f_{1,2})^3 = \left(\frac{\beta_0^{(1)}}{3}\right)^3 + \frac{\beta_0^{(1)}\beta_0^{(2)}}{6} + \frac{\beta_0^{(3)}}{3} \pm w_1$$

$$(g_{1,2})^3 = \left(\frac{\beta_1^{(1)}}{3}\right)^3 + \frac{\beta_1^{(1)}\beta_2^{(2)}}{6} + \frac{\beta_3^{(3)}}{3} \pm w_2$$

where

$$w_1 = \sqrt{\left(\frac{\beta_0^{(3)}}{2}\right)^2 - \left(\frac{\beta_0^{(2)}}{3}\right)^3 - \frac{(\beta_0^{(1)}\beta_0^{(2)})^2}{108} + \frac{(\beta_0^{(1)})^3\beta_0^{(3)}}{9} + \frac{\beta_0^{(1)}\beta_0^{(2)}\beta_0^{(3)}}{6}}$$

and

$$w_2 = \sqrt{\left(\frac{\beta_3^{(3)}}{2}\right)^2 - \left(\frac{\beta_2^{(2)}}{3}\right)^3 - \frac{(\beta_1^{(1)}\beta_2^{(2)})^2}{108} + \frac{(\beta_1^{(1)})^3\beta_3^{(3)}}{9} + \frac{\beta_1^{(1)}\beta_2^{(2)}\beta_3^{(3)}}{6}}$$

can be determined. These results are used in the remaining restrictions (11), (12), and (15).

4. Application: Two GNSS Time Series

To test the theory, the time series of an altitude component of a GNSS station (shown in Figure 1) is used. The data were provided by the Federal Institute of Hydrology (BFG), which operates a GNSS monitoring network for georeferencing and monitoring selected measuring stations.

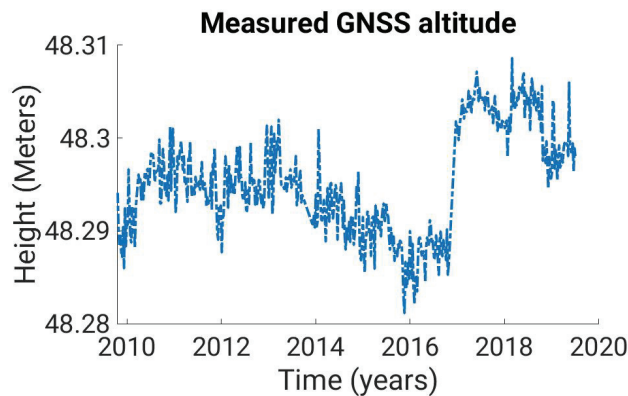


Figure 1. A GNSS time series S_t at the GNSS station in Pogum, Germany (TGPO) from end of 2009 to the begin of 2019.

Before the TVAR estimation can be performed, data jumps, data holes and trend must be removed from the observations. To eliminate the jump in the time series, all observations on the left side of the data jump are reduced by their mean value and the same is carried out for the right-side observations. The data gaps are interpolated by third-order splines. In order to establish the stationarity of the time series, each observation is reduced by its predecessor:

$$\hat{S}_t = S_t - S_{t-1}.$$

The offset of the resulting time series is eliminated by subtracting with the median of the time series. The result is shown in Figure 2.

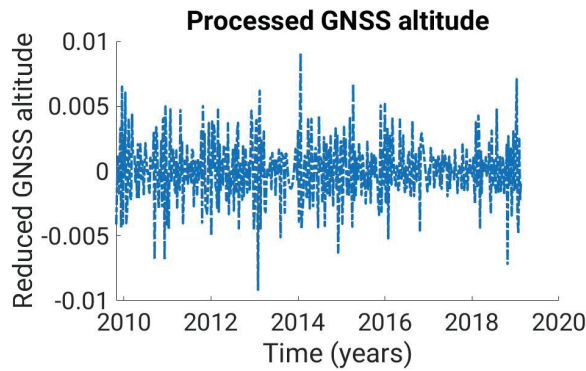


Figure 2. Reduced time series \hat{S}_t . This is created from the time series S_t by removing the data jump, extrapolating the data gaps and eliminating the trend.

In order to validate and compare the TVAR estimates later, we estimate time-stable AR(3) processes for a sliding window with the width of 100 observations. The roots of the AR(3) processes are shown in Figure 3. The time course goes from dark blue to light red, thus illustrating the temporal variability of the roots.

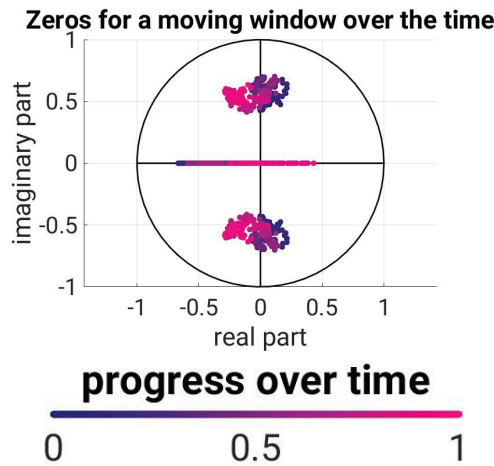


Figure 3. The roots of stationary AR(3) processes from each 100 consecutive observations. Here, the same-colored points correspond to the evaluation of a window, and the color gradient represents the temporal assignment.

Now, the TVAR(3) process is estimated: On the one hand, this can be established via the direct method presented here for order 3 processes. The result is shown in Figure 4. Here, the linear movement of the roots is shown in the same color gradient as in Figure 3; the progress of the window had been displayed. For comparison, the roots from Figure 3 are shown as grey sequence in Figures 4 and 5.

On the other hand, the successive estimation method from [11] was used to describe this time series. There are three ways to appreciate the TVAR(3) process:

1. Via three TVAR(1) processes.
2. Via a TVAR(1) process followed by a TVAR(2) process.
3. Via a TVAR(2) process followed by a TVAR(1) estimate.

All three possibilities were realized and evaluated using the sum of squared residuals:

$$SSR = \frac{e^T e}{n}.$$

Here, $e_t = \hat{S}_t - \alpha_1(t)\hat{S}_{t-1} - \alpha_2(t)\hat{S}_{t-2} - \alpha_3(t)\hat{S}_{t-3}$ is the error that occurs in the TVAR(3) estimation, and n is the number of errors per estimate. Method 3 had the smallest SSR, and was therefore used for the comparison. The resulting root movement is again contrasted with the result from the windowing and is shown in Figure 5.

Linear root motion for TVAR(3) estimation

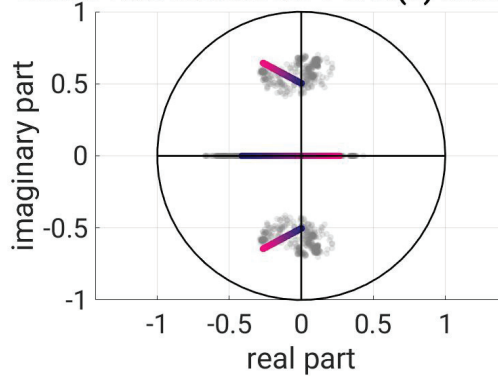


Figure 4. Time course of the roots according to the direct TVAR(3) estimate.

Successive TVAR estimation

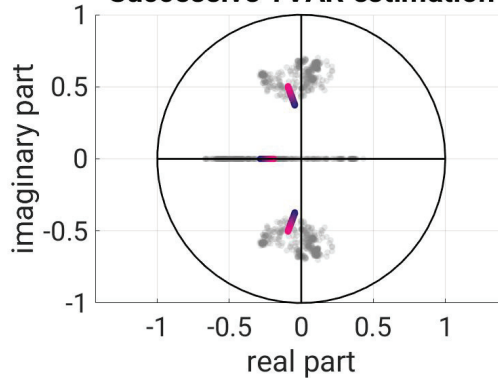


Figure 5. Time course of the roots according to the best-fitted successive TVAR(3) estimate.

In Figure 4, it can immediately be noted that the root motion of the direct calculation moves longitudinally through the roots of the windowing, while the roots of the successive estimation in Figure 5 show little movement and do not lie in the middle of the roots of the windowed version. Over time, the beginning and end points of the successive estimation are quite accurate in the first and last window, but the course of time is almost completely overlooked. The time course of the direct estimation has been approximated well by its counterpart.

5. Conclusions and Outlook

In this elaboration, a method has been developed which allows for TVAR(3) processes with linear root movements to be turned into estimates without sequential procedures.

For this purpose, the TVAR(3) estimation was extended by three nonlinear conditions to obtain linear root motions. In an application, it was shown that the solution obtained by the direct estimation of a TVAR(3) process with linear root movements better fits the data than the successive estimation of TVAR(1) and TVAR(2) processes. Therefore, the method set out here shows a useful extension of the TVAR process estimation with linear root motions. In further studies, the conditions for the TVAR(4) process can be determined to expand the evaluation possibilities. Polynomial degrees higher than 4 are not possible. This follows from the realization that there is no way to analytically determine the roots for these polynomials (see [14]).

Author Contributions: Conceptualization, J.K.; methodology, J.K.; software, J.K.; validation, J.K., J.M.B. and W.-D.S.; formal analysis, J.K.; investigation, J.K.; resources, J.K.; data curation, J.K.; writing—original draft preparation, J.K.; writing—review and editing, J.K.; visualization, J.K.; supervision, W.-D.S.; project administration, W.-D.S.; funding acquisition, W.-D.S. All authors have read and agreed to the published version of the manuscript.

Funding: His research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Grant No. 435703911 (SCHU 2305/7-1 ‘Nonstationary stochastic processes in least squares collocation—NonStopLSC’).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schuh, W.D. Signalverarbeitung in der Physikalischen Geodäsie. In *Erdmessung und Satellitengeodäsie: Handbuch der Geodäsie, herausgegeben von Willi Freeden und Reiner Rummel*; Rummel, R., Ed.; Springer Reference Naturwissenschaften; Springer: Berlin/Heidelberg, Germany, 2017; pp. 73–121. [CrossRef]
2. Schubert, T.; Brockmann, J.M.; Schuh, W.D. Identification of Suspicious Data for Robust Estimation of Stochastic Processes. In Proceedings of the IX Hotine-Marussi Symposium on Mathematical Geodesy, Rome, Italy, 18–22 June 2018; Novák, P., Crespi, M., Sneeuw, N., Sansò, F., Eds.; Springer International Publishing: Cham, Switzerland, 2021; International Association of Geodesy Symposia; pp. 199–207. [CrossRef]
3. Schuh, W.D.; Krasbutter, I.; Kargoll, B. Korrelierte Messung—Was Nun? In *Zeitabhängige Messgrößen—Ihre Daten Haben (Mehr-)Wert*; Neuner, H., Ed.; Wißner: Augsburg, Germany, 2014; Volume 74, pp. 85–101.
4. Schuh, W.D.; Brockmann, J.M. Numerical Treatment of Covariance Stationary Processes in Least Squares Collocation. In *Handbuch der Geodäsie: 6 Bände*; Freeden, W., Rummel, R., Eds.; Springer Reference Naturwissenschaften; Springer: Berlin/Heidelberg, Germany, 2019; pp. 1–37. [CrossRef]
5. Charbonnier, R.; Barlaud, M.; Alengrin, G.; Menez, J. Results on AR-modelling of Nonstationary Signals. *Signal Process.* **1987**, *12*, 143–151. [CrossRef]
6. Kargoll, B.; Omidalizandani, M.; Alkhatib, H.; Schuh, W.D. Further Results on a Modified EM Algorithm for Parameter Estimation in Linear Models with Time-Dependent Autoregressive and t-Distributed Errors. In *Proceedings of the Time Series Analysis and Forecasting*; Rojas, I., Pomares, H., Valenzuela, O., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 323–337.
7. Grenier, Y. Time-Dependent ARMA Modeling of Nonstationary Signals. *IEEE Trans. Acoust. Speech Signal Process.* **1983**, *31*, 899–911. [CrossRef]
8. Hall, M.; Oppenheim, A.V.; Willsky, A. Time-Varying Parametric Modeling of Speech. In Proceedings of the 1977 IEEE Conference on Decision and Control Including the 16th Symposium on Adaptive Processes and A Special Symposium on Fuzzy Set Theory and Applications, Tokyo, Japan, 7–9 December 1977, pp. 1085–1091. [CrossRef]
9. Slepian, D. Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty—V: The Discrete Case. *Bell Syst. Tech. J.* **1978**, *57*, 1371–1430. [CrossRef]
10. Kamen, E.W. The poles and zeros of a linear time-varying system. *Linear Algebra Its Appl.* **1988**, *98*, 263–289. [CrossRef]
11. Korte, J.; Schubert, T.; Brockmann, J.M.; Schuh, W.D. *On the Estimation of Time Varying AR Processes*; International Association of Geodesy Symposia; Springer: Berlin/Heidelberg, Germany, 2023; pp. 1–6. [CrossRef]

12. Dahlhaus, R. Fitting Time Series Models to Nonstationary Processes. *Ann. Stat.* **1997**, *25*, 1–37. [CrossRef]
13. Ludyk, G. *CAE von Dynamischen Systemen*; Springer: Berlin/Heidelberg, Germany, 2013. [CrossRef]
14. Ayoub, R.G. Paolo Ruffini's Contributions to the Quintic. *Arch. Hist. Exact Sci.* **1980**, *23*, 253–277. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Proceeding Paper

Approximation of Weymouth Equation Using Mathematical Programs with Complementarity Constraints for Natural Gas Transportation [†]

Cristian Alejandro Blanco-Martínez ^{1,*}, David Augusto Cardenas-Peña ¹, Mauricio Holguín-Londoño ¹, Andrés Marino Álvarez-Meza ² and Álvaro Angel Orozco-Gutiérrez ¹

¹ Automatic Research Group, Universidad Tecnológica de Pereira, Pereira 660003, Colombia; dcardenas@utp.edu.co (D.A.C.-P.); mau.hol@utp.edu.co (M.H.-L.); aaog@utp.edu.co (Á.A.O.-G.)

² Signal Processing and Recognition Group, Universidad Nacional de Colombia, Manizales 170003, Colombia; amalvarezme@unal.edu.co

* Correspondence: cristian.blanco@utp.edu.co

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Environmental demands around the world have led to an increasing interest in natural gas due to its advantages over other hydrocarbons used in power generation, which has led to the search for the best way to solve the transportation problem associated with this resource. In this paper, we propose a methodology that allows us to address the non-convexity related to the Weymouth equation that makes the optimization problem so difficult. The mentioned equation, in charge of relating the flows through the pipelines and the pressures at the nodes, is characterized by having a discontinuity in the form of a sign function. The proposal of this work is based on the use of Mathematical Programs with Complementarity Constraints (MPCC) to achieve a good approximation since it allows make certain continuous variables to behave as discrete variables in such a way that it is possible to avoid having to pose a mixed integer programming problem and this one. This approach showed a smaller approximation error (or at least equal) with other approximations used in the state of the art when tested in three different networks: one of 8 nodes, one of 48 nodes tested in other related works, and one of 63 nodes representing the Colombian natural gas transportation system.

Keywords: natural gas; Weymouth equation; discontinuous functions; MPCC

Citation: Blanco-Martínez, C.A.; Cardenas-Peña, D.A.; Holguín-Londoño, M.; Álvarez-Meza, A.M.; Orozco-Gutiérrez, Á.A. Approximation of Weymouth Equation Using Mathematical Programs with Complementarity Constraints for Natural Gas Transportation. *Eng. Proc.* **2023**, *39*, 91. <https://doi.org/10.3390/engproc2023039091>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 18 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Natural gas, as an energy source, has achieved great relevance worldwide in recent years due to two fundamental causes: Firstly, it allows reliable supply and continuous development supporting economic growth, which is highly related to energy consumption [1]. Secondly, the low greenhouse gas emission of natural gas makes it attractive for environmental care and sustainable development.

According to a study from 2020, the global demand for natural gas was 1788 billion cubic meters (bcm) in the same year, and the 2040 projection reached 2142 bcm, despite the new regulations for consumption decrease in Europe and the Middle East [2]. Particularly for Latin American countries, constituting the largest consumers of domestic natural gas, the statistics rise from 96 bcm in 2020 to 148 bcm in 2040. For most above countries, natural gas must also counteract the reduction of hydroelectric generation during dry seasons while supplying residential, commercial, and industrial demands [3]. Hence, there is a need for natural gas systems that fully supply all kinds of demand at a minimal fuel.

In general, natural gas systems are composed of four fundamental elements: injection fields (or re-gasification plants), providing the fuel at regulated pressure; gas pipelines, transporting the gas from sender to receiver nodes; compressors, raising the input-to-output

pressure; and end users, establishing the fuel demand. Several authors have proposed different ways to model the above elements and their interconnections, thus supplying the demand through optimization techniques [4]. Though the extensive work on each element, the gas pipelines remain a rather complex modeling problem since the physical relationship between the pressures at its ends and the flow through it, known as the Weymouth equation, holds a sign function determining the flow direction. As a nonconvex and discontinuous equality constraint, such a function poses a strong challenge in optimization [5].

The challenge imposed by the Weymouth equation promoted the development of optimization approaches with mathematical complexities without compromising the computational cost [6]. The first family of approaches turned the signum function into a linear combination with binary auxiliary variables yielding a mixed integer optimization problem [7]. Despite integrating discontinuities, mixed integer optimization problems present a significant source of nonconvexities reducing the probability of reaching a global optimum [8]. Further, the computational complexity of mixed integer programming is considerably larger than other optimization approaches [9]. As another solution, optimization through heuristic algorithms straightforwardly deals with nonlinear constraints [10]. However, their high sensitivity to initial conditions leads to suboptimal solutions [9]. Linearization and convexification strategies relax the Weymouth constraint reducing the computational complexity [11]. For instance, the binary auxiliary variables weigh piecewise linear functions that approximate the nonlinearities [12]. Another linear approximation relies on the Taylor series to replace nonlinear equations with a series of linear inequalities [13]. As an example of relaxation through convexification, the Second-order cone (SOC) programming introduces continuous and binary auxiliary variables and guarantees a global optimum on the approximation [14]. More recently, a polynomial regression holding odd coefficients approximates the Weymouth equation, its first and second derivative at the ends of a predefined operating interval [15]. Despite the reduced complexity and compatibility with conventional solvers, previous strategies result in Weymouth approximations that infringe on physical pipeline behavior, some of them to a great extent.

For reducing approximation errors, this work formulates the Weymouth equation in terms of mathematical programming with complementarity constraints (MPCC) that, instead of imposing an equality constraint, solves an optimization problem. MPCC expresses the signum function as an optimization problem with linearly constrained continuous variables behaving as binary, with two advantages: Firstly, the gas transport optimization avoids solving a complex mixed integer problem [16]. Secondly, MPCC not only constrains the original problem, but also minimizes the Weymouth approximation error.

The paper agenda is as follows: Section 2 describes both the objective function and constraints in the optimization problem. Section 3 proposes the problem solution using MPCC for modeling the Weymouth equation. Section 4 compares proposed MPCC against three Weymouth approximation approaches for 8-node, 48-node, and 63-node networks, the latter a case study of the Colombian gas transportation system. Finally, Section 5 concludes with the main findings and proposes future work.

2. Problem Formulation for Natural Gas Transport

The natural gas transmission network can be represented as a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where the set of vertices \mathcal{N} corresponds to the system nodes. The nodes with gas injection are known as wells, denoted by $\mathcal{W} \subset \mathcal{N}$, with an associated gas flow f_j^w . The nodes demanding gas are known as users, $\mathcal{U} \subset \mathcal{N}$, holding an individual cost for diversified loads. Due to the operational rationing scenarios, this study considers a virtual flow f_R^u at each user accounting for the unsupplied gas demand. For simplicity, nodes cannot be wells and users simultaneously, i.e., $\mathcal{W} \cap \mathcal{U} = \emptyset$. The set of directed edges $\mathcal{E} = \{(n, m) \mid n, m \in \mathcal{N}\}$ connect node pairs through two kinds of transmission elements, pipelines $\mathcal{T} \subset \mathcal{E}$, and compressing stations $\mathcal{C} \subset \mathcal{E}$. Note that edges must necessarily correspond to either a pipeline or a compressor, that is, $\mathcal{T} \cup \mathcal{C} = \mathcal{E}$, $\mathcal{T} \cap \mathcal{C} = \emptyset$. Since gas pipelines admit bidirectional flow $f_t^l : \forall t \in \mathcal{T}$, the edge direction is arbitrarily chosen, being f_t^l positive when flowing in the

chosen direction and negative otherwise. For gas flow through compressors $f_T^c : \forall c \in \mathcal{C}$, the values must always be positive, as bi-directional stations are not considered.

Since the objective of natural gas transport is supplying the user’s demand at the lowest cost, the optimization problem in Equation (1) takes place by including the following operation costs: C_I^w for injecting gas into the system by production wells, C_T for pipeline transportation, C_T^c for pressure boosting by compressing stations, and C_R^u for unsupplied demand penalties. The set $\mathcal{F} = \{f_I^w, f_T^t, f_T^c, f_R^u\}$ gathers all flows as decision variables for the optimization problem. Hence, each term in the summation becomes a function of the number of natural gas units used in its respective element.

$$\min_{\mathcal{F}} \sum_{w \in \mathcal{W}} C_I^w f_I^w + \sum_{t \in \mathcal{T}} C_T^t f_T^t + \sum_{c \in \mathcal{C}} C_T^c f_T^c + \sum_{u \in \mathcal{U}} C_R^u f_R^u \tag{1}$$

To mathematically model the technical limits and the physical behavior, the optimization problem objective function considers the following constraints: Equation (2) limits the flow injected by the production wells to its technical maximum/minimum injection capacity $\underline{F}_I^w / \overline{F}_I^w$. Equation (3) truncates the pipeline capacity at the structural maximum F_T^t and allows bidirectional flows. Equation (4) considers the maximum compression through the output-to-input pressure ratio β_c , resulting in a linear inequality constraint. \underline{P}_n and \overline{P}_n technically bound node pressures p_n in Equation (5). Regarding the rationing, Equation (6) limits the unsupplied demand between the desired zero and the respective user demand F_D^u . Finally, two equalities guarantee the physical behavior of the gas in the system. Equation (7), termed nodal gas balance, linearly matches each node’s injected with ejected gas. The Weymouth equality in Equation (8) ties the pressure at two nodes with the flow through the pipeline connecting them using a structural constant K_{ij} . Particularly, the Weymouth constraint is nonlinear, nonconvex, and disjunctive due to the sign function.

$$\underline{F}_I^w \leq f_I^w \leq \overline{F}_I^w, \quad \forall w \in \mathcal{W}, \tag{2}$$

$$-F_T^t \leq f_T^t \leq F_T^t, \quad \forall t \in \mathcal{T}, \tag{3}$$

$$p_m \leq \beta_c p_n, \quad \forall c = (n, m) \in \mathcal{C}, \tag{4}$$

$$\underline{P}_n \leq p_n \leq \overline{P}_n, \quad \forall n \in \mathcal{N}, \tag{5}$$

$$0 \leq f_R^u \leq F_D^u, \quad \forall u \in \mathcal{U}, \tag{6}$$

$$\sum_{m:(m,n) \in \mathcal{E}} f^m = \sum_{m':(n,m') \in \mathcal{E}} f^{m'}, \quad \forall n, m, m' \in \mathcal{N}, \tag{7}$$

$$\text{sgn}(f_T^t)(f_T^t)^2 = K_{ij}(p_i^2 - p_j^2), \quad \forall t = (n, m) \in \mathcal{T} \tag{8}$$

3. Problem Solution Using Complementarity Constraints for Nonconvex Functions

The sign function in Section 2 poses a challenge for conventional optimization approaches due to its non-derivability, non-linearity, and nonconvexity. This work proposes to deal with such a challenge using the mathematical technique of Mathematical Programs with Complementarity Constraints (MPCC). Complementarity refers to a relationship between variables where one or both must be at their bound, modeling mutually exclusive situations without the need for binary variables. Here, MPCC turns the discontinuous sign function into bounded continuous variables resulting from the optimization problem with the complementarity constraints in Equations (9) to (14). The Equations (9) to (14) indicate that when f_T^t is positive, f_+ will be positive and equal in magnitude to f_T^t , while f_- would necessarily be zero. Otherwise, when f_T^t is negative, it will be f_- that takes the value in magnitude of the transport flow of interest and f_+ that adjusts its value to zero. The Equation (14), in either of the two cases above, guarantees that the variable y takes the value of 1 if the sign

of f_T^t is positive or -1 if the sign of f_T^t is negative. Note that the proposed solution avoids the formulation of conventional mixed integer optimization approaches [17].

$$\min_y -yf_T^t \quad (9)$$

$$\text{s.t. } f_T^t = f_+ - f_- \quad (10)$$

$$f_+ \geq 0 \quad (11)$$

$$f_- \geq 0 \quad (12)$$

$$f_+f_- = 0 \quad (13)$$

$$f_+(1-y) + f_-(1+y) = 0 \quad (14)$$

To solve the resulting optimization problem, we resorted to the IPOPT solver, which is characterized by its use of a Primal-Dual Barrier Approach. This method works by converting the model into an unconstrained optimization problem, using a barrier function to penalize solutions that do not satisfy the constraints of the original problem. This algorithm allows convergence starting from poor initial points and incorporates a line-search filter, an important feature that helps to ensure progress towards a solution at each step by using the Armijo condition as a criterion. This condition requires that the objective function decreases by a sufficient amount relative to the step length. If the step is not sufficiently successful, the line-search filter reduces the step length and the algorithm takes a smaller step [18]. The main feature of this algorithm is that it finds the solution to the problem by moving through the feasible solution region using a central path [19]. Additionally this solver incorporates a variation of the original algorithm that solves both the primal and dual problems, which has shown superior performance compared to the standard version of the algorithm in practice [20].

4. Case Study

This work tests the efficiency of the proposed approach in three cases: A small system with one closed trajectory, a 48-node system with several closed trajectories, and a 63-node system representing the Colombian natural gas system. All cases were tested in Google Colab notebooks using the Gekko optimization tool [21] to implement the internal complementarity constraints for the sign function.

To validate the performance of the solution obtained by the proposed approach, it was compared with three approaches used in the state of the art to solve this same problem: replacing the equality constraint with a series of linear inequalities using the Taylor Series method [13], convexifying the problem using second-order cone programming (SOC) [14] and approximating the equation in a defined interval using a polynomial of degree five with odd coefficients only [15].

To compare the proposal presented in this study with other approximations, it was decided to take the solution obtained with each solver and evaluate it in the Weymouth equation equated to zero, so quantifying the approximation error.

4.1. 8-Node Natural Gas System

In the first instance, an 8-node database (Figure 1) was used for the study for two main reasons. The first is that being a small network, it presented an additional facility when corroborating the results. Despite the above, this network had a closed trajectory, an additional difficulty since it required the use of bidirectional pipelines, making it a very good starting point for the different approaches used.

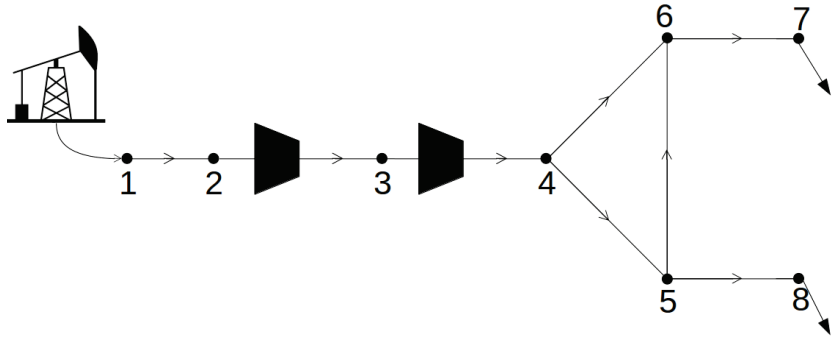


Figure 1. 8-node natural gas system.

The optimization problem was solved using each of the four approaches. None of the results obtained had to reach the point of rationing the hydrocarbon and, as shown in the Table 1, in three of them the result of the objective function was the same. Apart from the value of the objective function, which represents the operating cost of the system, in this study, it is of interest to know how good the solutions achieved with each of the approaches are. To understand it better, it can be seen that each of the pipelines has an associated equation of form Equation (8), so if the respective pressures and flows of these elements obtained when the problem was solved are taken, each Weymouth equation can be evaluated in order to quantify the amount of error in the approximation.

Table 1. Value of the objective function using each approach in each system.

System	Taylor	SOC	Polynomial	MPCC
8-node	194,133	194,133	229,169	194,168
48-node	11,095,000	11,095,000	11,099,349	13,121,240
63-node	4,517,783	4,517,783	-	4,704,223

The boxplots in Figure 2 contain the resulting values when the solution obtained by each of the respective approximations was evaluated in the equation equal to zero, i.e., the value to which it should ideally tend. Here a paired test yielded that the MPCC approximation gave the smallest error.

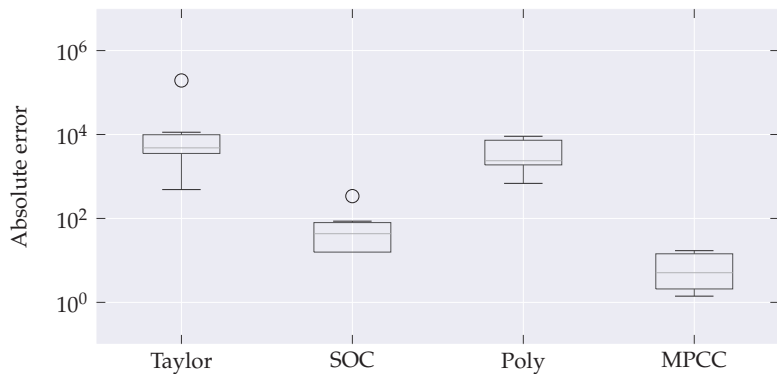


Figure 2. Absolute approximation error in Weymouth equation for the 8 nodes network. The boxplots illustrate the error dispersion for each considered approach. The whiskers bound the error first and third quartiles, and the circles denote outlying approximation errors.

4.2. 48-Node Natural Gas System

The second tested case is the 48-node database used by [22], among other authors in state of the art. This network, which can be seen in Figure 3, is composed of 9 injection fields, 8 compressor stations, and 22 gas demand nodes. This system was selected since its structure has several loops, which represent an additional difficulty in the solution of the problem and therefore it is a good way to test the robustness of the tested models.

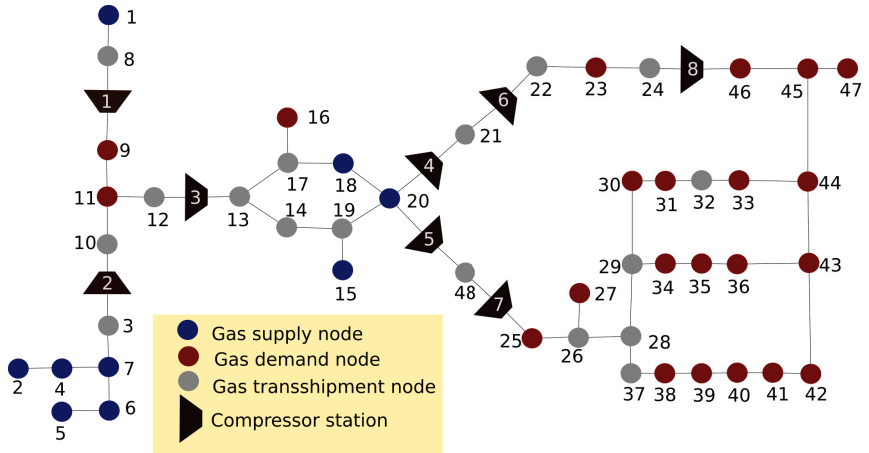


Figure 3. 48-node natural gas system [23].

Unlike the previous system, in this case, all approaches performed quite similarly in terms of the value of the objective function. Table 1 shows how the difference between approaches was practically negligible. Despite the above, this behavior was not maintained in the results obtained when the errors of the approaches were compared. As seen in Figure 4, the approximation using the Taylor series presented the highest error, followed by the polynomial approximation. In this case, the error presented by the Taylor series and MPCC was quite similar, being a statistical test the one that determined that the latter approach had a significantly lower error.

In this case, it is necessary to highlight the increase in error values with respect to those obtained in the 8-node network due to the difference between the systems. The fact of not only having many more elements but also connecting them in more complex configurations is a sample of how the difficulty of this type of optimization problem has quite a high scalability, forcing it to have sufficiently robust models for its solution.

4.3. 63-Node Natural Gas System

The third network corresponds to the Colombian natural gas injection and transportation system illustrated in Figure 5. This system is composed of 13 injection fields, 14 compressor stations, and 26 consumer nodes. Despite being radial, this system considers gas pipelines with bidirectional flows since the change of gas demanded throughout the year is related to the country's meteorology. For this case, the system introduced in [15] was updated in the Atlantic coast region by grouping elements and fixing the new operational constraints. These changes resulted in a new system with a total of 63 nodes.

Figure 6 presents the absolute errors in approximating the Weymouth equation for the 63-node system. The first observation from the results is that the polynomial approach fails to converge, probably because the optimal flows and pressures fall outside the Weymouth approximation interval. Secondly, the errors for this system are lower than those for the previous one, mainly due to lacking closed trajectories that alleviate the complexity. In terms of performance, despite reaching a cost function value about 4% more expensive, the proposed MPCC-based solution significantly decreases the error on the Weymouth

approximation compared to Taylor and SOC. Lastly, it is worth noting that MPCC yields two outlier errors, almost five orders larger than the average. After a manual exploration, we found the same outlying errors in Taylor and SOC at two pipelines between compressors. We hypothesize that the pressures at the ends and the flow through the pipe are over-constrained, so the solvers only accomplish the linear relationships of compressing ratios and gas balance instead of the complex Weymouth equation. Therefore, the MPCC-based solution provides more realistic gas transportation solutions due to a systematical reduction of Weymouth approximation errors.

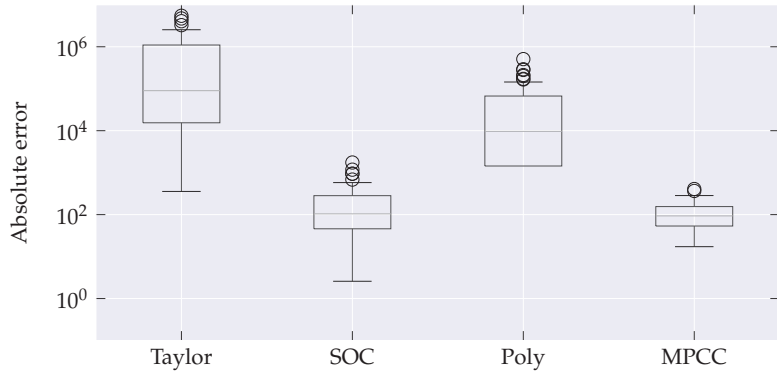


Figure 4. Absolute approximation error in Weymouth equation for the 48 nodes network. The boxplots illustrate the error dispersion for each considered approach. The whiskers bound the error first and third quartiles, and the circles denote outlying approximation errors.

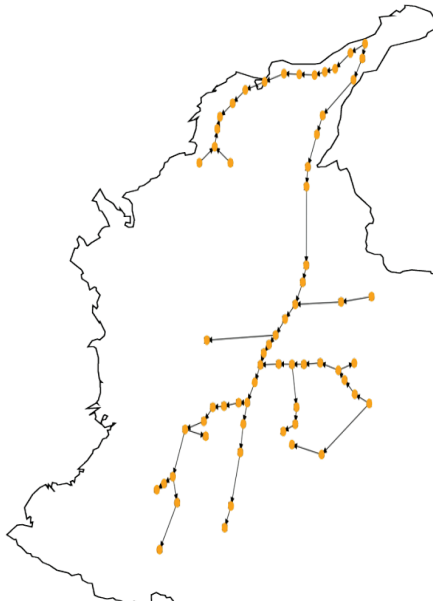


Figure 5. 63-node natural gas system (Colombian system).

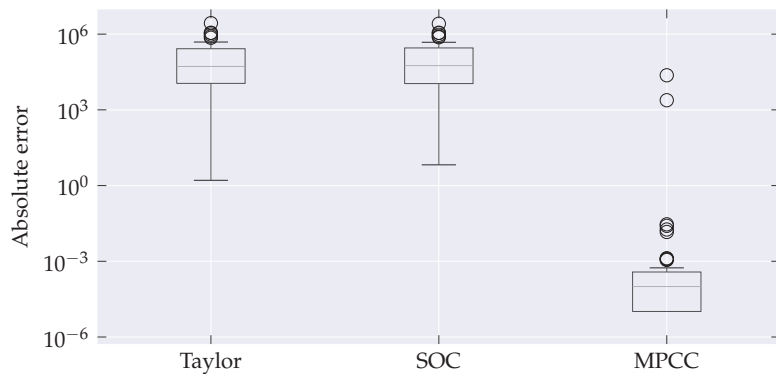


Figure 6. Absolute approximation error in Weymouth equation for the Colombian network. The boxplots illustrate the error dispersion for each considered approach. The whiskers bound the error first and third quartiles, and the circles denote outlying approximation errors.

5. Concluding Remarks and Future Work

This work proposed a solution for dealing with the Weymouth equation within the framework of natural gas transportation through complementarity constraints. MPCC formulated the nonconvex signum function in the Weymouth equation in terms of continuous bounded variables instead of binary ones, avoiding mixed integer programming. The proposed solution was contrasted against Taylor series, SOC, and polynomial approaches regarding the absolute approximation error at three study cases: An 8-node network with one closed trajectory, a 48-node network with multiple closed trajectories, and a 63-node radial network representing the Colombian gas transportation system. For experimental integrity, the IPOPT library solved all programming problems for all contrasted approaches. Experimental results evidenced that the proposed MPCC solution attained an approximation error smaller than contrasted approaches. Therefore, approximating the Weymouth equation using MPCC yielded pressures and flows satisfying the technical limits and physical behavior demanded by natural gas transportation problems.

For future work, we devise the following research directions. Firstly, the problem formulation will be extended to stochastic programming for natural gas transportation scenarios without a deterministic demand. Such scenarios are common for power generation that relies heavily on hydroelectric inputs and thermal power plants to fulfill the remaining demand. Secondly, a study of potential expansion plans must be conducted by evaluating various investment options in a robust and reliable solution for gas optimization problems.

Author Contributions: Data Extraction, Á.A.O.-G. and M.H.-L.; Validation, D.A.C.-P., and A.M.Á.-M.; Original Draft Preparation, C.A.B.-M., Á.A.O.-G., and D.A.C.-P.; Review and Editing, C.A.B.-M., and D.A.C.-P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Minciencias project: “Desarrollo de una herramienta para la planeación a largo plazo de la operación del sistema de transporte de gas natural de Colombia”—código de registro 69982—CONVOCATORIA DE PROYECTOS CONECTANDO CONOCIMIENTO 2019 852-2019.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable

Acknowledgments: Thanks to the Maestría en Ingeniería Eléctrica, graduate program of the Universidad Tecnológica de Pereira.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alam, M.S.; Paramati, S.R.; Shahbaz, M.; Bhattacharya, M. Natural gas, trade and sustainable growth: Empirical evidence from the top gas consumers of the developing world. *Appl. Econ.* **2017**, *49*, 635–649. [CrossRef]
2. Ahmad, T.; Zhang, D. A critical review of comparative global historical energy consumption and future demand: The story told so far. *Energy Rep.* **2020**, *6*, 1973–1991. [CrossRef]
3. Restrepo-Trujillo, J.; Moreno-Chuquen, R.; Jiménez-García, F.N. Strategies of expansion for Electric Power Systems based on hydroelectric plants in the context of climate change: Case of analysis of Colombia. *Int. J. Energy Econ. Policy* **2020**, *10*, 66–74. [CrossRef]
4. He, C.; Wu, L.; Liu, T.; Shahidehpour, M. Robust Co-Optimization Scheduling of Electricity and Natural Gas Systems via ADMM. *IEEE Trans. Sustain. Energy* **2017**, *8*, 658–670. [CrossRef]
5. Liu, F.; Bie, Z.; Wang, X. Day-ahead dispatch of integrated electricity and natural gas system considering reserve scheduling and renewable uncertainties. *IEEE Trans. Sustain. Energy* **2019**, *10*, 646–658. [CrossRef]
6. Yao, L.; Wang, X.; Ding, T.; Wang, Y.; Wu, X.; Liu, J. Stochastic day-ahead scheduling of integrated energy distribution network with identifying redundant gas network constraints. *IEEE Trans. Smart Grid* **2019**, *10*, 4309–4322. [CrossRef]
7. Üster, H.; Dilaveroglu, S. Optimization for design and operation of Natural Gas Transmission Networks. *Appl. Energy* **2014**, *133*, 56–69. [CrossRef]
8. Klatzer, T.; Bachhiesl, U.; Wogrin, S. State-of-the-art expansion planning of integrated power, natural gas, and Hydrogen Systems. *Int. J. Hydrogen Energy* **2022**, *47*, 20585–20603. [CrossRef]
9. Raheli, E.; Wu, Q.; Zhang, M.; Wen, C. Optimal coordinated operation of Integrated Natural Gas and Electric Power Systems: A review of modeling and solution methods. *Renew. Sustain. Energy Rev.* **2021**, *145*, 111134. [CrossRef]
10. He, C.; Zhang, X.; Liu, T.; Wu, L.; Shahidehpour, M. Coordination of interdependent electricity grid and natural gas network—A review. *Curr. Sustain. Energy Rep.* **2018**, *5*, 23–36. [CrossRef]
11. Singh, M.K.; Kekatos, V. Natural gas flow solvers using convex relaxation. *IEEE Trans. Control Netw. Syst.* **2020**, *7*, 1283–1295. [CrossRef]
12. Correa-Posada, C.M.; Sánchez-Martín, P. Gas Network Optimization: A comparison of Piecewise Linear Models. Optimization Online. 2014, pp. 1–24. Available online: <https://optimization-online.org/2014/10/4580/> (accessed on 5 May 2022).
13. Fodstad, M.; Midthun, K.T.; Tomasgard, A. Adding flexibility in a natural gas transportation network using interruptible transportation services. *Eur. J. Oper. Res.* **2015**, *243*, 647–657. [CrossRef]
14. Schwele, A.; Ordoudis, C.; Kazempour, J.; Pinson, P. Coordination of Power and Natural Gas Systems: Convexification Approaches for Linepack Modeling. In Proceedings of the 2019 IEEE Milan PowerTech, Milan, Italy, 23–27 June 2019; pp. 1–6. [CrossRef]
15. García-Marín, S.; González-Vanegas, W.; Murillo-Sánchez, C. MPNG: A MATPOWER-Based Tool for Optimal Power and Natural Gas Flow Analyses. *IEEE Trans. Power Syst.* **2022**, 1–9. [CrossRef]
16. Baumrucker, B.; Biegler, L. MPEC strategies for cost optimization of Pipeline Operations. *Comput. Chem. Eng.* **2010**, *34*, 900–913. [CrossRef]
17. Baumrucker, B.; Renfro, J.; Biegler, L. MPEC problem formulations and solution strategies with chemical engineering applications. *Comput. Chem. Eng.* **2008**, *32*, 2903–2913. [CrossRef]
18. Wächter, A.; Biegler, L.T. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program.* **2005**, *106*, 25–57. [CrossRef]
19. Nemirovski, A.S.; Todd, M.J. Interior-point methods for optimization. *Acta Numer.* **2008**, *17*, 191–234. [CrossRef]
20. Dahl, J.; Andersen, E.D. A primal-dual interior-point algorithm for nonsymmetric exponential-cone optimization. *Math. Program.* **2021**, *194*, 341–370. [CrossRef]
21. Beal, L.D.R.; Hill, D.C.; Martin, R.A.; Hedengren, J.D. GEKKO Optimization Suite. *Processes* **2018**, *6*, 106. [CrossRef]
22. Chen, S.; Conejo, A.J.; Sioshansi, R.; Wei, Z. Unit Commitment With an Enhanced Natural Gas-Flow Model. *IEEE Trans. Power Syst.* **2019**, *34*, 3729–3738. [CrossRef]
23. García-Marín, S.; González-Vanegas, W.; Murillo-Sánchez, C.E. MATPOWER/mpng: MPNG. 2019. Available online: <https://github.com/MATPOWER/mpng> (accessed on 17 July 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Analyzing Mobility Patterns of Complex Chronic Patients Using Wearable Activity Trackers: A Machine Learning Approach [†]

Alejandro Polo-Molina ^{1,*}, Eugenio F. Sánchez-Úbeda ¹, José Portela ^{1,2}, Rafael Palacios ¹, Carlos Rodríguez-Morcillo ¹, Antonio Muñoz ¹, Celia Alvarez-Romero ³ and Carlos Hernández-Quiles ⁴

¹ Instituto de Investigación Tecnológica (IIT), Universidad Pontificia Comillas, Escuela Técnica Superior de Ingeniería ICAI, 28015 Madrid, Spain; euge@comillas.edu (E.F.S.-Ú.); jportela@comillas.edu (J.P.); rafael.palacios@iit.comillas.edu (R.P.); carlos.rodriguez@iit.comillas.edu (C.R.-M.); antonio.munoz@iit.comillas.edu (A.M.)

² Facultad de Ciencias Económicas y Empresariales, ICADE, Universidad Pontificia Comillas, 28015 Madrid, Spain

³ Computational Health Informatics Group, Institute of Biomedicine of Seville, Virgen del Rocío University Hospital, CSIC, University of Seville, 41013 Seville, Spain; celia.alvarez@juntadeandalucia.es

⁴ Internal Medicine Department, Virgen del Rocío University Hospital, 41013 Seville, Spain; quiles_es@yahoo.es

* Correspondence: apolo@comillas.edu

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: This study suggests using wearable activity trackers to identify mobility patterns in chronic complex patients (CCPs) and investigate their relation with the Barthel index (BI) to assess functional decline. CCPs are individuals who suffer from multiple, chronic health conditions that often lead to a progressive decline in their functional capacity. As a result, CCPs frequently require the use of healthcare and social resources, placing a significant burden on the healthcare system. Evaluating mobility patterns is critical for determining a CCP's functional capacity and prognosis. To monitor the overall activity levels of CCPs, wearable activity trackers have been proposed. Utilizing the data gathered by the wearables, time series clustering with dynamic time warping (DTW) is employed to generate synchronized mobility patterns of the mean activity and coefficient of variation profiles. The research has revealed distinct patterns in individuals' walking habits, including the time of day they walk, whether they walk continuously or intermittently, and their relation to BI. These findings could significantly enhance CCPs' quality of care by providing a valuable tool for personalizing treatment and care plans.

Keywords: Barthel index; chronic complex patients; dynamic time warping; functional decline; mobility patterns; time series clustering

Citation: Polo-Molina, A.; Sánchez-Úbeda, E.F.; Portela, J.; Palacios, R.; Rodríguez-Morcillo, C.; Muñoz, A.; Alvarez-Romero, C.; Hernández-Quiles, C. Analyzing Mobility Patterns of Complex Chronic Patients Using Wearable Activity Trackers: A Machine Learning Approach. *Eng. Proc.* **2023**, *39*, 92. <https://doi.org/10.3390/engproc2023039092>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 18 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Chronic complex patients (CCPs) are characterized by a set of comorbidities that often lead to progressive functional decline, as well as increased use of healthcare and social resources. In addition, CCPs tend to be older adults with a high degree of polypharmacy, which can exacerbate underlying health conditions. The assessment of functional decline in CCPs can be an important tool for healthcare professionals to tailor treatment and care for these patients.

The mobility patterns observed in CCPs are linked to their functional capacity, determined by the Barthel index (BI) and, consequently, to their prognosis [1,2]. Given that modifications in CCPs' mobility patterns can indicate changes in their functional status, they can serve as valuable prognostic factors. Therefore, the BI has been identified as a reliable measure of CCPs' functional capacity and prognosis, highlighting the crucial role of mobility patterns in assessing and predicting CCPs' clinical outcomes.

Taking this into account, the aim of this study is to examine the relationship between declining mobility and alterations in a patient's clinical status. Given the importance of measuring activity levels, wearable activity trackers are proposed to assess the mobility of CCPs. In particular, the research sought to evaluate the various mobility patterns, derived from the data gathered, of CCPs and their connection to the BI. To accomplish this goal, the study was designed as a descriptive study. Ethical approval was obtained from the regional health organization before the study began.

The combination of synchronized mobility patterns with the BI is a unique and original approach that allows healthcare professionals to identify temporal variations in patient movement and underlying factors affecting patient mobility.

The study presented in this paper is part of the chronic-IoT project, which is a coordinated effort funded by the Ministry of Science, Innovation and Universities, through the 2019 Research Challenges call of the State Research Agency (ref. PID2019-110747RB-C21). With a duration of 36 months, from June 2020 to the end of May 2023, the work conducted in chronic-IoT is based on the development and validation of behavioral models, based on machine learning (ML) and the IoT environment, to predict changes in the functional capacity of CCPs through the analysis of mobility patterns measured by activity wristband devices.

This collaborative project involved the participation of two institutions in Spain: the Virgen del Rocío University Hospital (HUVR) in Seville and the Institute for Research in Technology (IIT) at ICAI School of Engineering (ICAI) of Universidad Pontificia Comillas in Madrid. As a coordinator, HUVR played a crucial role in the project, leveraging its expertise in healthcare to contribute to the research objectives. Meanwhile, IIT at ICAI brought its technological knowledge and research capabilities to the table, complementing HUVR's strengths.

The present paper is organized as follows: Section 2 provides a literature review of prior research studies that investigated the utilization of wearable activity trackers and DTW time series clustering in medical contexts. Section 3 outlines the methodology employed in the research concerning data collection, data pre-processing and time series clustering. Section 4 presents the findings of the research, which are then followed by the conclusions in Section 5.

2. Literature Review

Over the years, the use of wearable activity trackers has garnered significant attention in the healthcare industry, particularly in the assessment of the physical conditions of patients. There have been many approaches to try to relate physical condition and wearables. Specially, as mentioned in [3], patient monitoring and behavioral changes are two main topics regarding the use of wearables in medical research. Wearables have the potential to provide continuous, objective, and non-invasive monitoring of a patient's physiological parameters. Moreover, the validity of the measures taken by the wearables have been studied regarding steps taken [4], heart rate [5] or sleep quality [6]. Combined, this type of technology can aid in the early detection and management of chronic diseases, improving the overall health outcomes of patients.

As previously noted, two of the most extensively studied areas are patient monitoring and behavioral changes. Numerous studies have investigated the utility of wearable activity trackers as a tool for monitoring patient information. Within this domain, some studies have examined the feasibility of using wearables to monitor patients during rehabilitation [7,8]. Additionally, other studies have explored how data collected by wearable activity trackers can be leveraged to provide feedback that facilitates faster planning and intervention [9].

In the context of research on behavioral changes in relation to wearable technology, the most commonly explored approach involves investigating whether wearable activity trackers have a positive effect on physical activity [10]. However, there are relatively few articles that focus on the relationship between the data collected from wearables and the deterioration of patient health [11].

Regarding the wearable device used in this study, a commercially available device was utilized instead of a medical one. Similarly, ref. [12] examined the acceptance and usage of commercially available wearable activity trackers among adults over 50 with chronic illnesses. The study found that while participants generally perceived the devices as easy to use, they identified challenges in maintaining sustained use.

On the other hand, ML is considered one of the most prominent fields in light of the development of data-driven solutions aimed at gaining a better understanding of a variety of problems. In recent years, the utilization of ML has experienced exponential growth across diverse domains, including healthcare [13], finance [14], and marketing [15]. Among the different ML techniques available, time series analysis is the most relevant area for this paper's objective, characterizing time variable mobility patterns. DTW is a widely adopted metric for measuring the distance between time series data, even if there are differences in length or phase. Originally introduced in the field of data mining [16], DTW has found numerous applications in various domains, including speech recognition and medicine.

In the field of biomedical signal processing, it has been used to analyze electrocardiograms (ECGs) to classify ECG frames [17,18]. In addition, DTW has been employed to cluster EEG waveforms, and has been demonstrated to be effective in discriminating between waves with minor disparities in frequency, amplitude, peak location, or initial phase [19]. In comparison to other methods that rely on waveform features or peak-aligned difference computation, DTW resulted in more homogeneous clusters, as demonstrated in experimental studies involving stimulated and actual EEG data.

DTW has also shown promise in applications related to human movement analysis, such as gait analysis. For example, ref. [20] used DTW to compare the gait patterns of patients with Parkinson's disease and healthy controls, identifying significant differences between the two groups.

In addition, to the best of our knowledge, the methodology that integrates pattern mobilities with the BI represents a distinctive and original approach. Furthermore, a novel methodology based on cross-correlation is proposed for the synchronization of DTW mobility patterns, which are treated as circular data. This method allows for the identification of temporal variations in the movement of patients and enables the creation of a synchronized representation of these patterns, providing insights into the underlying factors that contribute to patient mobility. By utilizing this approach, healthcare professionals can better understand the progression of patient mobility and develop effective interventions to enhance patient outcomes.

3. Methodology

This section describes the methodology used to obtain the study results. It covers data collection and data pre-processing including aggregating and smoothing to generate the mean and coefficient of variation (CV) mobility profiles. These profiles are clustered using a K-means clustering algorithm based on DTW distances and a decision tree analysis is used to understand obtained patterns.

3.1. Data Collection

During the first phase of the study, patients were recruited based on their BI scores, measuring a patient's ability to perform daily activities whereby higher scores indicate greater independence. The patients were divided into three groups based on their BI scores: those with total dependence (A) ($BI \leq 20$), severe dependence (B) (BI in $(20,60]$), and moderate/mild dependence or independence (C) ($BI > 60$). The study included a total of 36 patients from the Internal Medicine Department of the Virgen del Rocio University Hospital of Seville, all of whom met the criteria of chronic patients with complex health needs defined according to the Integrated Patient Care Process of the Andalusian Ministry of Health. Patients in a situation of agony, those with limited vital prognosis, and psychiatric or neurodegenerative diseases were excluded from the study. Moreover, some patients were excluded from the analysis due to a lack of data.

Out of the considered participants, 64% were male and 36% were female. It was found that 16/36 ($\approx 44\%$) of the patients had BI of type B, while the other 20/36 ($\approx 56\%$) had an BI of type C. No patients of index A were considered due to the limited range of movement. The mean age of male participants was 75.78 years (SD = 7.15), while the mean age of female participants was 74.69 years (SD = 9.82).

The second phase of the study involved the implementation of an Internet of Things (IoT) infrastructure to collect patient mobility measures. After careful consideration, the most appropriate technology for their needs was selected. The IoT-based infrastructure consisted of wearables to measure the mobility activities of patients, with a focus on minimizing disruption to their daily routines. Wearables allowed the researchers to measure physical activity through the number of steps taken, cardiac activity and the sleep time of the 36 patients in the study.

3.2. Data Pre-Processing

As previously indicated, the wearable activity tracker is capable of collecting information on a patient's number of steps taken, heart rate, and sleep duration. The device automatically captures the number of steps and heart rate at irregular intervals, which are subsequently aggregated into fixed intervals to maintain consistent data granularity.

3.2.1. Activity Profiles Based on the Mean

To generate the mean activity profiles, the number of steps taken are added in one-hour intervals, and the median heart rate is computed for the same interval. Additionally, in an attempt to ensure data quality, in cases where the median heart rate is missing, the number of steps is also marked as empty. This procedure is implemented because a null heart rate value might indicate that the wearable device is not properly positioned, thereby possibly resulting in inaccurate measurements. In terms of sleep data, the activity tracker provides daily information on the total duration and quality of sleep, which is further categorized into multiple variables.

Based on data collected every hour over multiple days per each patient, the mean step profile is constructed. The mean steps profiles are generated after smoothing the time series. In order to smooth the time series, a centered rolling window with a size of three is computed. After this, data is grouped by the hour such that each patient has a mean representation of their activity throughout a 24 h period. The resulting data allows us to gain insights into a CCP's physical activity levels and obtain a more accurate picture of their daily activity patterns as shown in Figure 1.

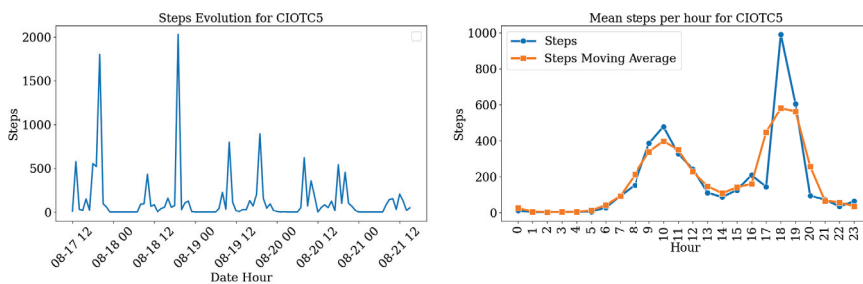


Figure 1. Generation of the 24 h mean profiles for CIOTC5.

Furthermore, it is worth noting that, given that the objective of the study is to understand mobility patterns and not just the raw number of steps, it is necessary to normalize the mean step count curves in order to compare them across the patients. This difference is even more noticeable between patients with a BI of type B and type C, as the latter group tends to have a much higher mean step count volume (Figure 2). To address this, a normalization process was performed by subtracting the minimum value and dividing by

the range (max–min). This transformation ensures that all values fall between 0 and 1 for each patient.

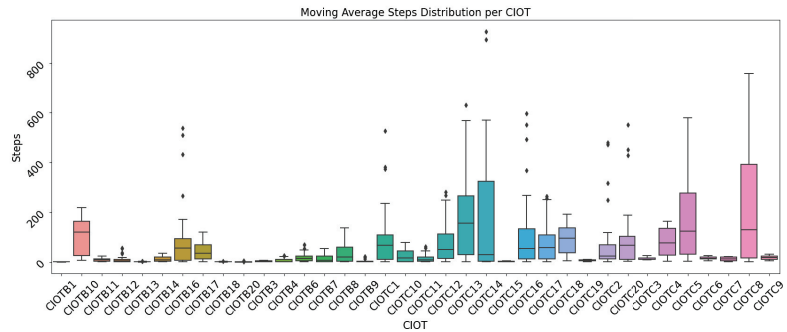


Figure 2. The distribution of steps per day and CIOT after the application of a moving average.

Given that sleep data are only available on a daily basis, the median values were computed for each patient to obtain a general representation of their sleep patterns and quality. Among all the available sleep-related data, the focus was on the median amount of sleep, which was further divided into deep sleep, shallow sleep, REM and wake time, as well as the median bedtime and wake-up time. These variables provide insight into the overall sleep patterns of the patients, including the duration and quality of their sleep, as well as their sleep–wake cycle.

3.2.2. Activity Profiles Based on the Coefficient of Variation

In addition to the 24 h mean profiles presented earlier, the CV profiles were also incorporated. Specifically, for each day in the dataset from 00:00 to 23:59, the CV was calculated for each day and hour-based on the aggregated 5 min data, and then the median was computed for all hours for a given patient.

To obtain smoother profiles, a centered moving average with a window size of three was applied to the 24 h CV profiles. As a result, each curve represents the smoothed median CV for each hour and patient. Missing values are assigned for cases where there is no movement during a specific hour or when the mean is zero.

3.3. Time Series Clustering Using DTW

Time series clustering is a powerful analytical technique used to identify patterns and relationships among time series data. By grouping similar time series together, this method can help extract meaningful insights and reveal underlying patterns that may not be visible when examining individual series in isolation.

In this case, since the goal is to generate mobility patterns based on the normalized average hourly profiles for each patient, a temporal clustering algorithm was used to identify existing mobility patterns. For this study, the time series K-means algorithm from the tslearn library was applied, as it is widely regarded as a standard in the literature [21,22]. However, it is important to note that other methods could have been considered as well.

It is essential to highlight that time-shifts are insignificant within a certain range of maximum hours, as the goal is to create mobility patterns independent of the specific hour and primarily based on shape. Therefore, DTW is the preferred distance function to measure the similarity between time series.

DTW evaluates the similarity between two time series by finding the best alignment between them, which involves time-axis stretching or compressing. This method is well-suited for the current study, as a group of patients may demonstrate similar mobility patterns with only slight variations in time. Given this scenario, a Sakoe–Chiba [23] radius

of 3 h was considered as mobility patterns may be slightly out-of-sync by a few hours, yet still exhibit significant differences between the morning and afternoon.

To synchronize time series data following the creation of clusters, cross-correlation was employed to determine the optimal time lags for comparing both series. The objective was to move the mobility patterns through the time-axis in order to find the best alignment in terms of correlation. Cross-correlation is a mathematical function that measures the similarity between two signals as a function of the time lag applied to one of them [24]. Once the optimal time lag has been determined, the time series can be synchronized by shifting one of the series by the optimal time lag.

When examining the mobility patterns of individuals, it is useful to consider the average mobility patterns over a 24 h cycle. However, due to differences in individual sleep and work schedules, these patterns may not necessarily align perfectly with each other. This can result in a phase shift, where the 24 h cycle of one individual is shifted forward or backward relative to another individual.

To account for these phase shifts, it is useful to treat the 24 h cycle as circular data. In circular data, the end of the cycle precedes the beginning, forming a continuous circle rather than a linear sequence. This allows for the accurate representation of phase shifts and the analysis of cyclic patterns as observed in Figure 3.

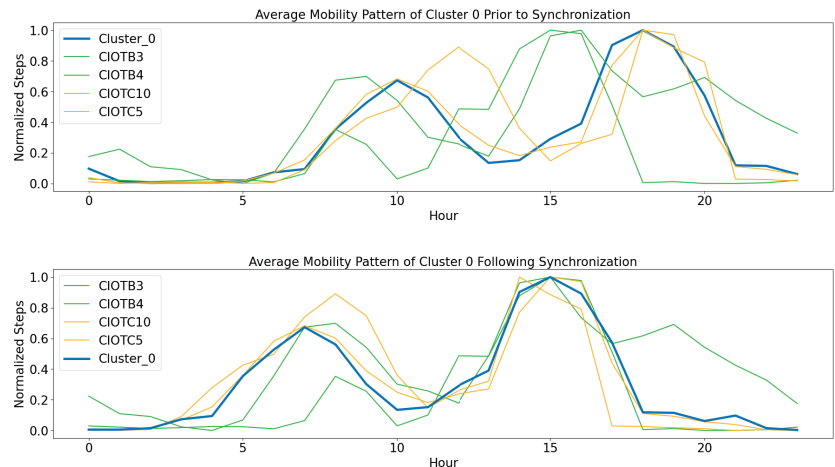


Figure 3. Cross-correlation synchronization of DTW clustering.

4. Results

This section presents the mean mobility patterns resulting from the application of DTW time series K-means. Additionally, the clusters derived from the CV profiles are presented. Finally, a comparison of cluster members is made to better understand the relationship between the distribution of steps and walking behavior.

4.1. Clustering Mean Activity Profiles

After running the time series K-means algorithm considering the DTW clustering distance and a Sakoe–Chiba radius of three, six clusters were selected using the elbow methodology. The mobility patterns acquired are shown in Figure 4. According to Table 1, cluster 3, which represents patients with a more stable daily mobility pattern, appears to be associated with a BI of type C.

Table 1. Distribution of sample percentages across Barthel types within each of the most populated clusters.

Cluster	N° Samples	CIOTB	CIOTC
Cluster 0	4	50%	50%
Cluster 1	8	50%	50%
Cluster 2	13	46%	54%
Cluster 3	7	14%	86%

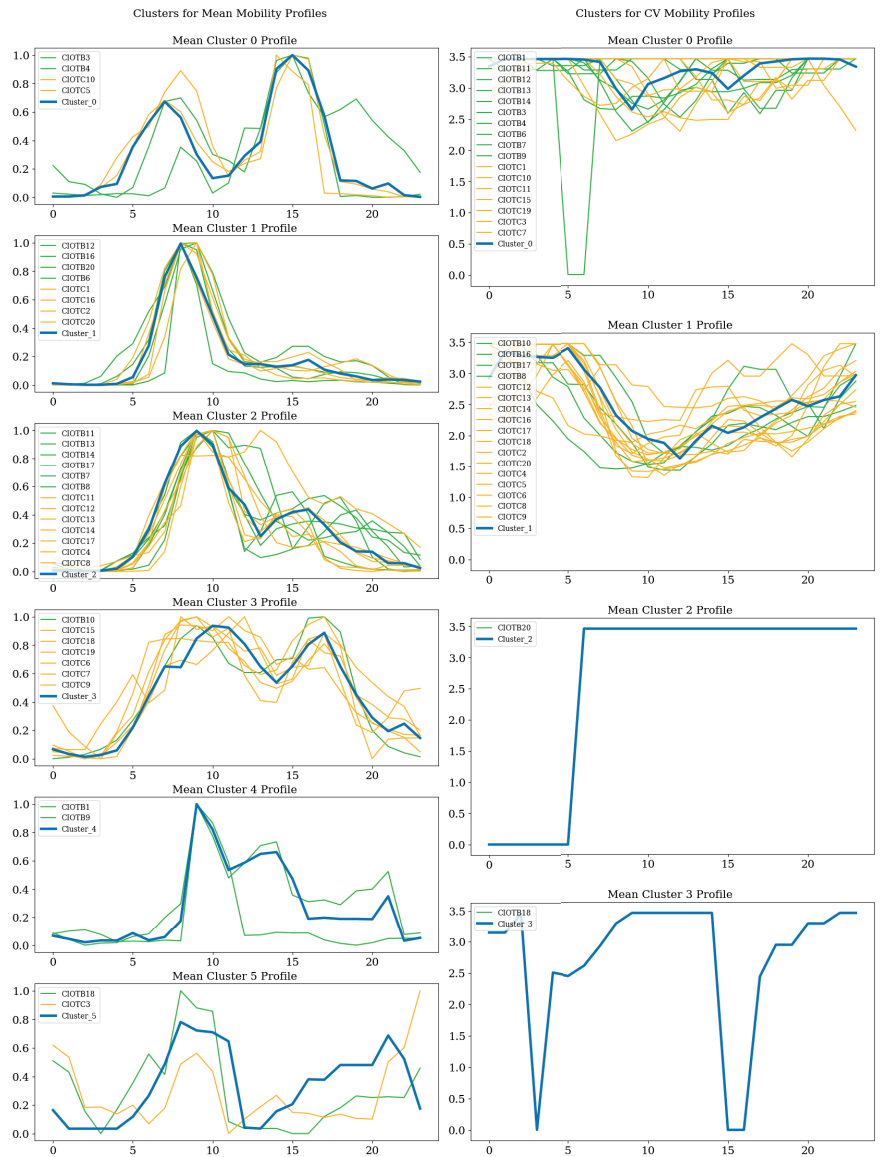


Figure 4. Comparison of the clusters of mean and coefficient of variation (CV) activity profiles.

To investigate this relationship further, a decision tree was trained to predict the cluster to which a given patient belongs based on their sleep and mobility patterns (see Figure 5). By examining the splits made by the decision tree, it is possible to gain an insight into the different mobility patterns present in the dataset. To optimize the hyperparameters of the decision tree, a stratified K-fold approach was used with $k = 5$, accounting for both the criterion (gini or entropy) and the minimum impurity decrease. Our analysis revealed that the optimal hyperparameters for the decision tree were gini as the criterion and a minimum impurity decrease of 0.06.

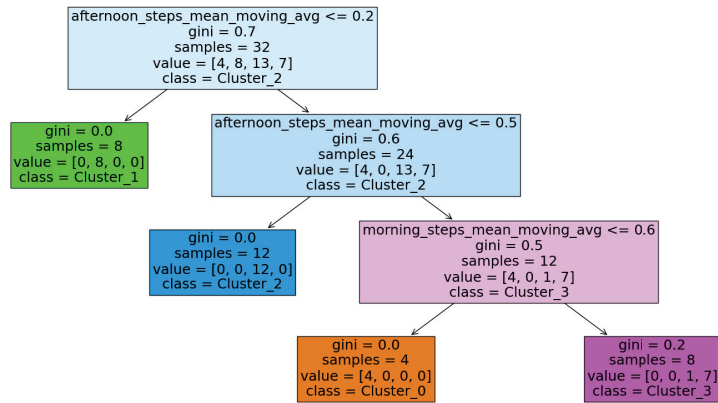


Figure 5. Decision tree for patient classification into the four main clusters.

The features considered in the tree were the median sleep time, (divided into deep sleep, shallow sleep, REM, and wake time), the median hour of sleeping, the median hour of waking up, and the mean normalized steps divided into four sections: Dawn [0,6], Morning [7, 13], Afternoon [14, 19], and Night [20,24]. Note that normalized steps were used here; therefore, how much each patient was walking during the afternoon relative to their overall walking patterns was evaluated.

It was discovered that the primary disparity between the clusters was the average normalized steps taken during the afternoon. When fewer steps were taken during the afternoon, the patient was categorized into cluster 1. Following this division, patients were classified based on the number of steps taken during the afternoon. If the normalized steps taken during the afternoon exceeded 0.5, indicating considerable movement during that time, patients were classified into either cluster 3 or 0, depending on their activity levels during the morning. Conversely, patients who were not as active during the afternoon were classified as cluster 2.

4.2. Clustering Coefficient of Variation Activity Profiles

The objective of the clustering analysis performed above was to characterize the distinct patterns of mobility in terms of the average amount of walking by each patient during each hour. However, these clusters lack information on the variability of mobility within each patient. It is important to recognize that patients do not all walk in the same way, and thus it is necessary to consider both the mean and variance for each hour. This approach allows for discrimination not only by the amount of movement within each hour but also by the type of movement performed, such as whether it is continuous or interrupted. In cases where movement is interrupted during a one-hour span, the standard deviation within the data collected every 5 min will be higher than in cases where movement is constant, resulting in a standard deviation closer to zero.

To address these differences, a DTW time series K-means clustering was performed based on the hourly median of the CV for every day. The analysis in Figure 4 and Table 2

indicates the presence of two predominant patterns of physical activity with respect to the CV. The primary distinguishing feature between cluster 0, predominantly composed of patients with a BI of type B, and cluster 1, with patients having a BI of type C, is the reduced variability in the duration of morning and afternoon walks. This finding suggests that patients in these clusters may exhibit more consistent patterns of locomotion, as opposed to a start–stop movement pattern.

Table 2. Distribution of sample percentages across Barthel types within each of the most populated clusters.

Cluster	N° Samples	CIOTB	CIOTC
Cluster 0	17	59%	41%
Cluster 1	18	24%	76%

4.3. Relation between Mean and CV Profiles

To gain a better understanding of the correlation between clusters based on the CV and 24 h mobility patterns, reference can be made to the Sankey diagram presented in Figure 6. It is noteworthy that patients in cluster 3 of the mean profiles appear to correspond to those in cluster 1 of the CV profiles. Upon analyzing the decision tree depicted in Figure 5 and the clusters shown in Figure 4, it becomes evident that patients with more stable movement patterns, without significant differences between morning and afternoon, tend to walk in a more continuous manner. Furthermore, there appears to be a relationship between the BI and CV cluster. In an attempt to establish this relationship statistically, a two-sample z-test for proportions was performed with a 90% confidence interval, resulting in statistical significance [25].



Figure 6. Correlation between clusters predominantly populated by 24 h mobility patterns and the cluster generated regarding the CV.

5. Conclusions

In this study, we analyzed the different patterns of mobility and their relationship with the patient’s clinical status. Specifically, we intended to build a better understanding of how CCPs move through the day and how it can be related to their BI.

To do so, a time series clustering algorithm was used using 24 h mean and CV profile data using DTW as the similarity measure. It was found that there are four main patterns of mobility, considering the mean profiles, depending on their levels of movement during the morning and afternoon. Moreover, those clusters can be related to those obtained using the CV patterns and it was concluded that patients with greater mobility during the afternoon seem to have a more continuous way of walking rather than a start–stop pattern. Specifically, those who tend to walk in a more continuous way were mostly related to a BI of type C.

Overall, this study highlights the potential for using wearables to gather data on patient mobility and clinical condition, which could be used to improve the care provided to chronic patients with complex health needs. The study’s findings could contribute to the

growing body of research on how technology can be used to monitor and improve patient health outcomes.

Lastly, future research may further analyze the relationship between the information provided by the activity tracker and the detection of patient degradation based on the mobility patterns described in this study.

Author Contributions: Conceptualization, A.P.-M., E.F.S.-Ú., J.P., C.A.-R. and C.H.-Q.; methodology, A.P.-M., E.F.S.-Ú. and J.P.; software, A.P.-M.; validation, E.F.S.-Ú.; formal analysis, A.P.-M., E.F.S.-Ú. and J.P.; investigation, A.P.-M.; resources, C.A.-R. and C.H.-Q.; data curation, A.P.-M., E.F.S.-Ú., J.P., C.A.-R. and C.H.-Q.; writing—original draft preparation, A.P.-M.; writing—review and editing, A.P.-M., E.F.S.-Ú., J.P., R.P., C.R.-M., A.M., C.A.-R. and C.H.-Q.; visualization, A.P.-M.; supervision, E.F.S.-Ú. and J.P.; project administration E.F.S.-Ú., C.A.-R. and C.H.-Q.; funding acquisition, E.F.S.-Ú., J.P., R.P., C.R.-M., A.M., C.A.-R. and C.H.-Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the chronic-IoT project (Agencia Estatal de Investigación, PID2019-110747RB-C21/AEI/10.13039/501100011033), which has received funding from the Ministry of Science, Innovation and Universities of the Government of Spain and the State Research Agency. Also, this research has been co-supported by the Carlos III National Institute of Health, through the IMPaCT-Data program (code IMP/00019), and through the Platform for Dynamization and Innovation of the Spanish National Health System industrial capacities and their effective transfer to the productive sector (code PT20/00088), both co-funded by European Regional Development Fund (FEDER) ‘A way of making Europe’.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Restrictions apply to the availability of these data. The data were obtained from households participating in the chronic-IoT project (Agencia Estatal de Investigación, PID2019-110747RB-C21/AEI/10.13039/501100011033). The dataset is neither public nor available in the way that it has been used as a source in this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Silguero, S.A.A.; Martínez-Reig, M.; Arnedo, L.G.; Martínez, G.J.; Rizo, L.R.; Soler, P.A. Enfermedad crónica, mortalidad, discapacidad y pérdida de movilidad en ancianos españoles: Estudio FRADEA. *Revista Española de Geriatría y Gerontología* **2014**, *49*, 51–58. [CrossRef] [PubMed]
2. Cech, D.J.; Martin, S. *Functional Movement Development Across the Life Span*, 3rd ed.; W.B. Saunders: Saint Louis, MO, USA, 2012; p. iv. [CrossRef]
3. Shin, G.; Jarrahi, M.H.; Fei, Y.; Karami, A.; Gafinowitz, N.; Byun, A.; Lu, X. Wearable activity trackers, accuracy, adoption, acceptance and health impact: A systematic literature review. *J. Biomed. Inform.* **2019**, *93*, 103153. [CrossRef] [PubMed]
4. Alinia, P.; Cain, C.; Fallahzadeh, R.; Shahrokni, A.; Cook, D.; Ghasemzadeh, H. How Accurate Is Your Activity Tracker? A Comparative Study of Step Counts in Low-Intensity Physical Activities. *JMIR Mhealth Uhealth* **2017**, *5*, e106. [CrossRef] [PubMed]
5. Martín-Escudero, P.; Cabanas, A.M.; Dotor-Castilla, M.L.; Galindo-Canales, M.; Miguel-Tobal, F.; Fernández-Pérez, C.; Fuentes-Ferrer, M.; Giannetti, R. Are Activity Wrist-Worn Devices Accurate for Determining Heart Rate during Intense Exercise? *Bioengineering* **2023**, *10*, 254. [CrossRef] [PubMed]
6. Siyanbade, J.; Abdulrazak, B.; Sadek, I. Unobtrusive Monitoring of Sleep Cycles: A Technical Review. *BioMedInformatics* **2022**, *2*, 204–216. [CrossRef]
7. Cook, D.J.; Thompson, J.E.; Prinsen, S.K.; Dearani, J.A.; Deschamps, C. Functional Recovery in the Elderly After Major Surgery: Assessment of Mobility Recovery Using Wireless Technology. *Ann. Thorac. Surg.* **2013**, *96*, 1057–1061. [CrossRef] [PubMed]
8. Roe, J.; Salmon, L.; Twigg, J. Objective measure of activity level after total knee arthroplasty with the use of the ‘Fitbit’ device. *Orthop. J. Sport. Med.* **2016**, *4*, 2325967116S00012. [CrossRef]
9. Shinde, A.M.; Gresham, G.K.; Hendifar, A.E.; Li, Q.; Spiegel, B.; Rimel, B.; Walsh, C.S.; Tuli, R.; Piantadosi, S.; Figlin, R.A. Correlating wearable activity monitor data with PROMIS detected distress and physical functioning in advanced cancer patients. *J. Clin. Oncol.* **2017**, *35*, e21689–e21689. [CrossRef]
10. Washington, W.D.; Banna, K.M.; Gibson, A.L. Preliminary efficacy of prize-based contingency management to increase activity levels in healthy adults. *J. Appl. Behav. Anal.* **2014**, *47*, 231–245. [CrossRef] [PubMed]

11. Lunney, M.; Wiebe, N.; Kusi-Appiah, E.; Tonelli, A.; Lewis, R.; Ferber, R.; Tonelli, M. Wearable Fitness Trackers to Predict Clinical Deterioration in Maintenance Hemodialysis: A Prospective Cohort Feasibility Study. *Kidney Med.* **2021**, *3*, 768–775.e1. [CrossRef] [PubMed]
12. Mercer, K.; Giangregorio, L.; Schneider, E.; Chilana, P.; Li, M.; Grindrod, K. Acceptance of Commercially Available Wearable Activity Trackers Among Adults Aged Over 50 and With Chronic Illness: A Mixed-Methods Evaluation. *JMIR mHealth uHealth* **2016**, *4*, e7. [CrossRef] [PubMed]
13. Shehab, M.; Abualigah, L.; Shambour, Q.; Abu-Hashem, M.A.; Shambour, M.K.Y.; Alsalibi, A.I.; Gandomi, A.H. Machine learning in medical applications: A review of state-of-the-art methods. *Comput. Biol. Med.* **2022**, *145*, 105458. [CrossRef] [PubMed]
14. Ahmed, S.; Alshater, M.M.; Ammari, A.E.; Hammami, H. Artificial intelligence and machine learning in finance: A bibliometric review. *Res. Int. Bus. Financ.* **2022**, *61*, 101646. [CrossRef]
15. Ngai, E.W.; Wu, Y. Machine learning in marketing: A literature review, conceptual framework, and research agenda. *J. Bus. Res.* **2022**, *145*, 35–48. [CrossRef]
16. Berndt, D.J.; Clifford, J. Using Dynamic Time Warping to Find Patterns in Time Series. In Proceedings of the KDD Workshop, Seattle, WA, USA, 31 July–1 August 1994.
17. Huang, B.; Kinsner, W. ECG frame classification using dynamic time warping. In Proceedings of the IEEE CCECE2002. Canadian Conference on Electrical and Computer Engineering. Conference Proceedings (Cat. No.02CH37373), Winnipeg, MB, Canada, 12–15 May 2002; Volume 2, pp. 1105–1110. [CrossRef]
18. Yao, X.; Wei, H.L. A Modified Dynamic Time Warping (MDTW) and Innovative Average Non-self Match Distance (ANSMD) Method for Anomaly Detection in ECG Recordings. In *Proceedings of the Recent Advances in AI-enabled Automated Medical Diagnosis*; CRC Press: Boca Raton, FL, USA, 2022; pp. 281–303.
19. Huang, H.C.; Jansen, B. EEG waveform analysis by means of dynamic time-warping. *Int. J. -Bio-Med. Comput.* **1985**, *17*, 135–144. [CrossRef] [PubMed]
20. Steinmetzer, T.; Bonninger, I.; Priwitzter, B.; Reinhardt, F.; Reckhardt, M.C.; Erk, D.; Travieso, C.M. Clustering of Human Gait with Parkinson's Disease by Using Dynamic Time Warping. In Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOB), San Carlos, Costa Rica, 18–20 July 2018; pp. 1–6. [CrossRef]
21. Tavenard, R.; Faouzi, J.; Vandewiele, G.; Divo, F.; Androz, G.; Holtz, C.; Payne, M.; Yurchak, R.; Rußwurm, M.; Kolar, K.; et al. Tslern, A Machine Learning Toolkit for Time Series Data. *J. Mach. Learn. Res.* **2020**, *21*, 1–6.
22. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, USA, 2001.
23. Sakoe, H.; Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech, Signal Process.* **1978**, *26*, 43–49. [CrossRef]
24. Weisstein, E.W. Cross-Correlation. MathWorld—A Wolfram Web Resource. 2003. Available online: <http://mathworld.wolfram.com/Cross-Correlation.html> (accessed on 23 February 2023).
25. Stine, R.; Foster, D. *Statistics for Business: Decision Making and Analysis*; Addison-Wesley: Boston, MA, USA, 2011.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Reduced Order Modeling with Skew-Radial Basis Functions for Time Series Prediction [†]

Manuchehr Aminian ^{1,*} and Michael Kirby ^{2,*}¹ Department of Mathematics and Statistics, California State Polytechnic University, Pomona, CA 91768, USA² Department of Mathematics, Colorado State University, Fort Collins, CO 80523, USA

* Correspondence: maminian@cpp.edu (M.A.); michael.kirby@colostate.edu (M.K.)

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: We present a sparsity-promoting RBF algorithm for time-series prediction. We use a time-delayed embedding framework and model the function from the embedding space to predict the next point in the time series. We explore the standard benchmark data set generated by the Mackey–Glass chaotic dynamical system. We also consider a model of temperature telemetry associated with the mouse model immune response to infection. We see that significantly reduced models can be obtained by solving the penalized RBF fitting problem.

Keywords: skew-radial basis functions; reduced order models; sparse optimization; time-series prediction

1. Introduction

Radial basis functions (RBFs) provide an attractive option for the fitting of data in high-dimensional domains. They are appealing for their simplicity and compact nature while enjoying universal approximation theorems suggesting, in principle, that they are potentially as powerful as other methodologies. Traditional RBF expansions are of the form

$$f(x) = \sum_i w_i \phi(\|x - c_i\|) \quad (1)$$

where $f: \mathbb{R} \rightarrow \mathbb{R}$, and the function $\phi(\cdot)$ is selected from a variety of options including functions such as $\{r, r^3, r^2 \ln r, \exp(-r), \exp(-r^2), \dots\}$ and placed at points $\{c_i\}$. Unlike the fully nonlinear optimization problems encountered with multilayer perceptrons, recurrent neural networks, and deep convolutional networks, the weights in RBF expansions can be determined by solving a least-squares problem. The RBF centers $\{c_i\}$ can be found using a variety of approaches including random selection, clustering, or nonlinear optimization. The added property of skewness, introduced in [1], makes an RBF-based approach more powerful at fitting asymmetric features in the data.

Skew-radial basis functions (sRBF) introduce the symmetry-breaking function $s_i: \mathbb{R} \rightarrow \mathbb{R}$ for each center, i.e.,

$$f(x) = \sum_i w_i s_i(x; D_i) \phi(\|x - c_i\|_{W_i}) \quad (2)$$

The matrix D_i consists of the skew shape parameters, and W_i is a diagonal weighting for the Euclidean inner product.

Radial basis functions were introduced as an alternative to artificial neural networks for function approximation based on the flexibility of the optimization problem [2]. They have proven particularly useful for both reduced-order and adaptive, or online modeling of data-streams. Given that the contribution of each individual basis function to the complexity of the model is easily interpreted, they may be sequentially placed where

Citation: Aminian, M.; Kirby, M. Reduced Order Modeling with Skew-Radial Basis Functions for Time Series Prediction. *Eng. Proc.* **2023**, *39*, 93. <https://doi.org/10.3390/engproc2023039093>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 21 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

they are most needed until the data are fitted [3]. Alternatively, one can employ linear programming to adapt a model as changes are detected in the distributions of the data [4]. A comprehensive review of the RBF literature may be found in [5].

2. Sparse Skew RBFs

Building off of traditional RBF (1) and skew RBF formulation (2), we introduce sparsity via L_1 penalization for the number of centers via the solution of the optimization problem

$$\min_{\{w_i, W_i, D_i, c_i\}} \left\| y - \sum_{i=1}^{n_c} w_i s_i(x; D_i) \phi(d_i(x, c_i; W_i)) \right\| + \alpha \sum_{i=1}^{n_c} |w_i| \tag{3}$$

We follow the choices in [1] (Section 4.3) for choice of s_i and ϕ . The per-center inner product and induced norm for a single center c_i with $u, v \in \mathbb{R}^n$ are

$$\langle u, v \rangle_{W_i} = u^T W_i v; \quad \|u\|_{W_i} = \sqrt{\langle u, u \rangle_{W_i}} \tag{4}$$

with a symmetric positive definite W_i . The metric is $d_i(x, c_i; W_i) = \|x - c_i\|_{W_i}$, and choice of base RBF $\phi(r) = \exp(-r^2)$ unless stated otherwise. The skew functions s_i have shape parameters that are optimized for each basis function. We specialize to skew matrices which are diagonal; $D_i = \text{diag}(\lambda_i)$, $\lambda_i \in \mathbb{R}^n$; so

$$s_i(x; \lambda_i) = \frac{1}{\pi} \arctan(\lambda_i^T (x - c_i)) + \frac{1}{2} \tag{5}$$

so $s_i : \mathbb{R}^n \rightarrow (0,1)$. Using diagonal D_i parameter matrix produces a skew which is interpreted as having a direction and magnitude of effect; see Figure 1 for an illustration of this idea with a single sRBF in \mathbb{R}^2 .

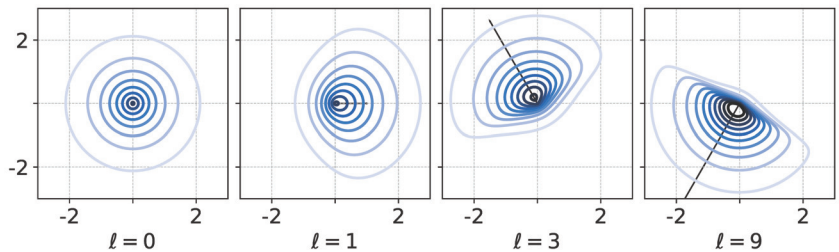


Figure 1. Visualization of level sets a single skew RBF as in Equation (2) in \mathbb{R}^2 using $W = I$ and diagonal $D = \text{diag}(\ell \cos(\theta), \ell \sin(\theta))$. Setting $\ell = 0$ ($D = 0$) restores symmetry. Various θ and ℓ (vectors shown in panels) directly correspond to a direction and magnitude of asymmetry in the RBF. Further shape control comes with choice of Euclidean W -norm (not shown).

We have solved the optimization problem in Equation (3) to compute a parsimonious expansion given by Equation (2). For scalar time series, this is performed using a time-delay embedding of the series $\{x_1, x_2, x_3, \dots\}$, where the fitting problem becomes a map from an l -dimensional embedding space to the next time-point of interest, e.g.,

$$x_{n+1} = f(x_n, x_{n-T}, x_{n-2T}) \tag{6}$$

where typically we choose $l = 3$, and T is a delay time for sampling the time series at uncorrelated points. In this approach we are solving this optimization problem in batch mode, i.e., we assume that the complete data set is available for training. This is in contrast to [6], where RBFs are sequentially placed based on a statistical test, or in [4], where the RBFs are placed incrementally for streaming data. Note that [4] also employs the one-norm in the optimization problem, but it is solved using the dual simplex algorithm rather than gradient-based methods we use here. The optimization problem in Equation (3)

is readily solved in Python, via PyTorch, which provides automatic differentiation and optimization tools.

3. Numerical Results

3.1. Details of Implementation

Each scalar time series is mean-centered using values from the training data prior to time-delay embedding.

We use PyTorch as a framework to implement the mixed-objective loss (3). PyTorch provides a large collection of tools for a user to define arbitrary loss functions and optimization techniques. We subclassed `torch.nn.Module`, allowing us to pass a collection of parameters and save the history during the learning phase. Updates were found using `torch.optim.SGD` with learning rate 0.1 and sometimes enabling Nesterov momentum, though we expect the details of these choices have no noticeable impact on our downstream conclusions.

The entirety of input/output training data (X, y) with $X \in \mathbb{R}^{n \times N}$, $y \in \mathbb{R}^N$ (N samples) are given to the module in the training phase (in contrast to online learning). For larger time series, we trained using simple uniform random batches of 5% of the training data on each iteration. Unless stated otherwise, the objective is the one-step prediction for $y_n := x_{n+1}$ based on input $X_n := (x_n, x_{n-T}, x_{n-2T})$.

3.2. Mackey–Glass Revisited

We begin our exploration of the sparse skew RBFs by solving the optimization problem given by Equation (3) for various parameter configurations. For this first numerical experiment, we use the Mackey–Glass chaotic time series, a standard for benchmarking RBF algorithms [3]. We employ 1197 points time-delayed and embedded into three dimensions to predict the next point. We used 800 points for training, 200 points for validation, and reserved an additional 195 for testing. In all our experiments, the prediction accuracies on training, validation, and test data were very similar, indicating no overfitting was taking place. Note that our current goal is to explore the behavior of the proposed algorithm for different parameters, rather than perform a head-to-head comparison with other algorithms. This frees us to select the modeling set-up from scratch and not be bound to choices made in other papers. Here we take $T = 1$.

It is useful to explore the sparsity behavior of the model as a function of the parameter α ; we select the values shown in Table 1. All the results in this paper used 50,000 epochs for training. In Table 1, we see that, as expected, the number of skew RBFs with weight $|W_i| > 1e-3$ decreases with increasing α . Interestingly, for $\alpha = 0.002$, we see a sharp drop in the absolute value of the weights. The resulting model requires only two RBFs and has relatively low error. Note that we also saw similar behavior with model size $n_c = 20,100$.

Table 1. This table shows the variability of the model order (number of numerically nonzero sRBF weights) as a function of the sparsity parameter α .

Number of Centers = 50				
Sparsity α	$\ w\ _1$	Train Acc	Val Acc	RBFs
0	26.84	0.0013	0.0012	50
0.0001	19.47	0.0009	0.0007	50
0.002	2.14	0.0024	0.0021	2
0.01	1.66	0.0036	0.0033	1
0.1	1.28	0.0177	0.0169	1

In Figure 2, we plot the absolute values of the skew RBF expansion weights w_i sorted in decreasing order. We see how the distribution of weights is impacted by the sparsity parameter with $\alpha = 0.002$ versus $\alpha = 0$. Here, with the sparsity parameter $\alpha = 0.002$, the model selects two optimal RBFs that produce an error of 0.0021 versus the 50 skew RBF

model (with $\alpha = 0$) that results in an error of 0.0012. Interestingly, the 1-norm penalized model with $\alpha = 0.0001$ still uses 50 skew RBFs but has the smallest error of all the models. In Figure 3 (top), we plot the model prediction that uses 50 RBFs and $\alpha = 0$. In contrast, in Figure 3 (bottom) we see the results of an RBF with two basis functions approximating the Mackey–Glass time series. The smaller model fails to capture the positive extreme values, but only 2 skew RBFs are used in contrast to 50 skew RBFs.

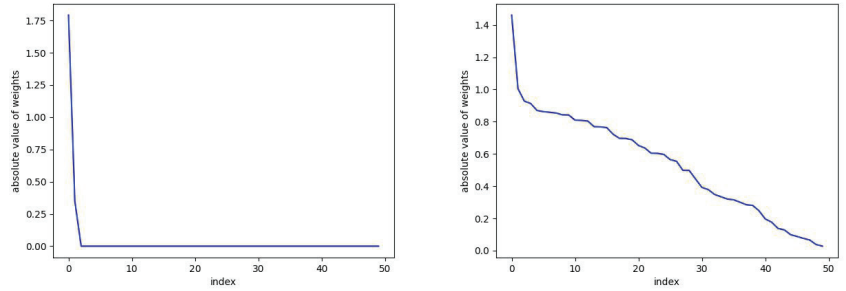


Figure 2. The parameter α can be seen here to promote sparsity. On the left, $\alpha = 0.002$ and the number of required skew RBFs is 2. On the right, we take $\alpha = 0$ and there is no sparsity and 50 skew RBFs are used in the model. See Figure 3 for the corresponding predictions.

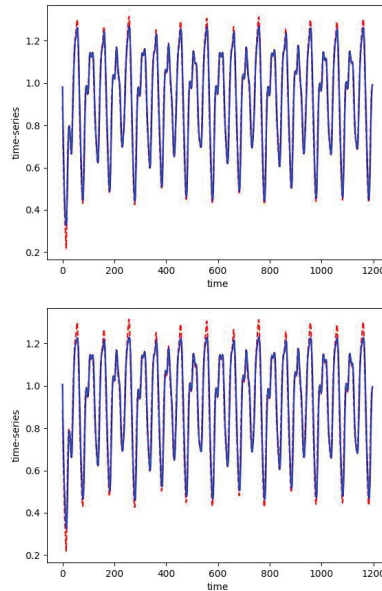


Figure 3. The one-step prediction (blue) of the Mackey–Glass time series (red). The first 800 points are training data. For both simulations we take $n_c = 50$. Top: as expected, with $\alpha = 0$, the solution of the optimization problem requires 50 skew basis functions. Bottom: with $\alpha = 0.002$ the sparse solution only uses 2 skew RBFs.

We remark that this approach differs from the benchmarking in [6], where noise was added to the data and the mapping used a four-dimensional domain. The approach used in [6] requires the presence of noise since the approach employs a noise test on the residuals. Here we are using an optimization that uses all the available data. This is in contrast to methods that build the model in a streaming fashion, i.e., *online*, adding one point at a time as they become available [4,6].

3.3. Mouse Telemetry Data

The collection of data here are the result of experiments on laboratory mice conducted by the Andrews-Polymenis and Threadgill labs [7–9]. The broad question is to identify mechanisms of tolerance and understand the variety of immune response to a few specific diseases in mice. The time series are the result of surgically embedding a device which tracks internal body temperature (degrees Celsius) and activity (physical movements) sampled at 1/60 Hz (once per minute). Mice are left alone for approximately 7 days, then infected, and then further observed for an additional 7–14 days before being euthanized. We focus on the temperature time series here.

The process of numerically finding a zero-autocorrelation time to find a delay for TDE leads us to a delay $T = 4 \sim 6$ h. Alternatively, a theoretical argument for studying autocorrelation of the monochromatic signal $\sin(2\pi t/\tau)$ leads to guidance of choosing $T = \tau/4$, i.e., one-fourth of the period. Assuming mice exhibit circadian behavior (a period of 24 h), this agrees with numerical studies and so we use $T = 24/4 = 6$ h. The training process and results of this process are visualized in Figure 4.

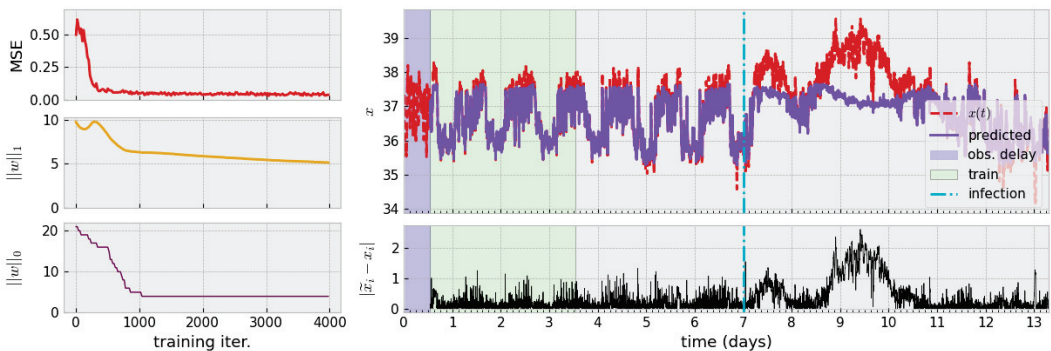


Figure 4. (Left): Decay of loss during training and diagram of associated time series. The mean-square error quickly saturates while one-norm of weight vector w slowly decays. Only a handful of non-zero model centers $\|w\|_0$ persist past the first 1000 iterations. (Right): Approximate mean-normalization is applied, followed by a delay (blue region) for TDE. The model is trained on three days of data (green region). The original time series (red, dashed) and model prediction of the learned model (solid purple) are shown. Pointwise prediction error of $x(t)$ is shown beneath. Parameters: $\alpha = 0.01, n_c = 21, T = 360, l = 3$.

3.4. Other Applications

3.4.1. Iterated Prediction

We study how our sRBF implementation can be directly interpreted and how it can be applied to anomaly detection. One method of evaluating the success of the trained model is to study iterated prediction rather than repeated one-step prediction. Given an initial training set, the task is to repeatedly feed the output of the model as new input; e.g.,

$$f(x_t, x_{t-T}, x_{t-2T}) \rightarrow \hat{x}_{t+1} \tag{7}$$

$$\Rightarrow f(\hat{x}_{t+1}, x_{t-T+1}, x_{t-2T+1}) \rightarrow \hat{x}_{t+2} \tag{8}$$

$$\Rightarrow \dots \tag{9}$$

The broad question is to ask, “Did the model in fact learn the shape of my data?” The answer may be *yes quantitatively* if the time series continues to have small prediction error without further training or online learning applied. More holistically, we are interested in whether the time series learns a (quasi)periodic shape by approaching a limit cycle. Figure 5 illustrates this idea with a noiseless sine wave. One-step prediction (red) is successful; iterated prediction (purple) receives no further ground-truth, with some evidence towards the existence of a limit cycle. Similar application with a mouse time series in Figure 5

(bottom) succeeds in one-step prediction, but the iterated prediction exhibits behavior akin to an exponential decay to a fixed point.

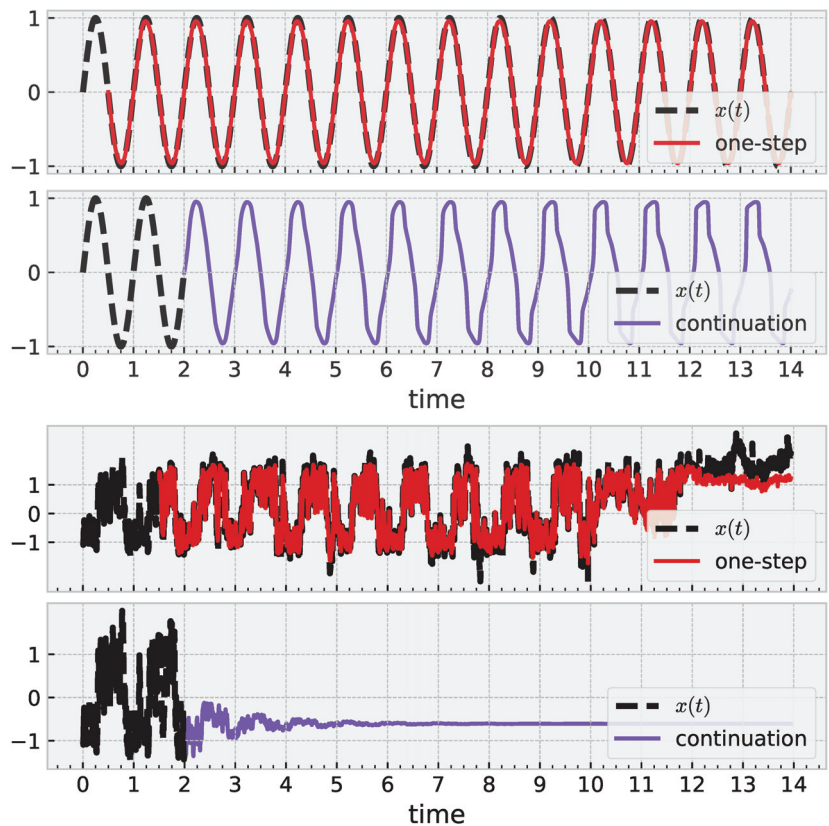


Figure 5. Results illustrating how a trained sRBF model learns geometric structure. **(Top):** With $\sin(t)$, a learned model leads to a limit cycle. **(Bottom):** Similar approach with noisy mouse data fails. Parameters $\alpha = 0.01, n_c = 21, l = 3$. After training, both models decrease to 3 active centers.

3.4.2. Visualization of Trained sRBF Models

When the embedding dimension of a scalar time series is $l = 2$, we can reason about and explain the model with the full set of learned parameters in Equation (2). Figure 6 illustrates this idea with one such mouse time-series model. Following the expectation for L_1 penalization in the training objective, we consider weights $|w_i| < 1e - 2$ to be numerically zero and then build the full function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ in analogy to the single sRBF illustrations in Figure 1. We emphasize careful interpretation with this figure. Red crosses represent locations of sRBF centers, which may not directly align with local extrema due to the asymmetry. Next, contour colors represents the value associated with the prediction for x_{t+1} , which cannot be directly visualized in the plane here. Ideally, we would like to understand how the learned model positions centers and shapes parameters to match the embedded time series in \mathbb{R}^2 but requires subsequent visualization of time-series values. We leave this for future work.

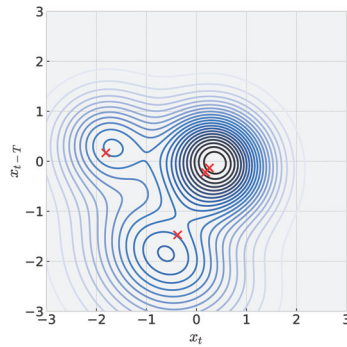


Figure 6. Visualization of sRBF model trained with mouse time series embedded in dimension $l = 2$. Red crosses mark sRBF centers, which may not be located with local maxima due to asymmetry. Parameters: $\alpha = 0.01, n_c = 21, l = 2$.

3.4.3. Anomaly Detection

An important application area for time series modeling, especially with health data such as these, is anomaly detection. An anomaly can signal need for treatment, adjustment, intervention, etc. Additionally, it is important to carefully consider how to train such a model and how to evaluate success (or failure) depending on the specific application. Figure 7 illustrates these ideas with one such temperature time series. Here, the shaded green region marks training data, which is assumed “nominal” or “healthy” data (and in general requires knowledge of the application area). Square marks in darker shades of green on the time series denote detected anomalies. The definition of an anomaly can depend on the specific modeling approach; but often involves producing a scoring system (some function which maps input data to $[0, \infty)$ or $[0, 1]$) which can be thresholded to produce a binary decision of “nominal” or “anomaly”. Ultimately, the threshold is a free parameter, and one mediates between false positives and negatives based on the choice. This figure illustrates two automated methods for choosing a threshold based on quantiles. The more sensitive of the two shown is a decision based on a new pointwise prediction error $|\hat{x}_{t+1} - x_{t+1}|$ being greater than 99% of such errors in the training data. Increasing this threshold reduces sensitivity but hopefully also reduces false positives (the meaning of which is also dependent on the application). Anomaly detection remains challenging without applying a full time-series analysis pipeline which denoises or directly models noise, and we leave a more thorough investigation of the application of skew RBFs to this task to future work.

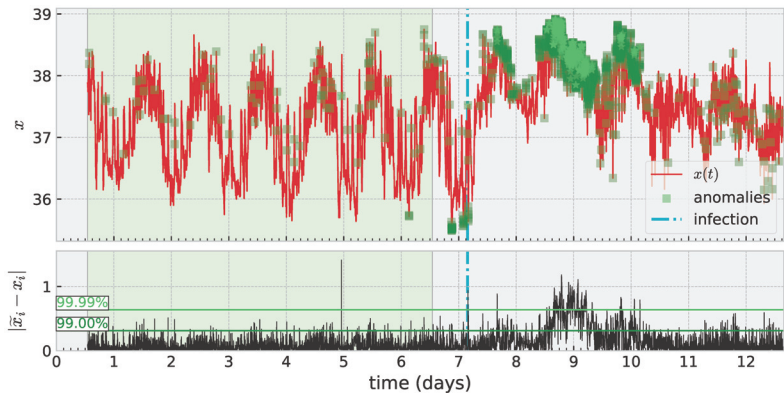


Figure 7. Anomaly detection based on thresholding error in one-step prediction $|\hat{y} - x_{t+1}|$ exceeding a threshold of error built from training data.

4. Conclusions

In summary, we have proposed an approach for computing reduced order skew RBF approximations. This approach is complementary to others in the literature that seek to construct a parsimonious fit with RBFs. This batch method approach is appealing for its algorithmic simplicity. The fact that the number of required RBFs can be determined by the steep drop in expansion weights makes it appealing for automatic model order determination. In the future, it would be interesting to explore the annealing of the sparsity coefficient during training.

We note that there has been a renewed interest in RBFs in view of the observation that kernel methods, under certain circumstances, can be viewed as equivalent to deep neural networks [10]. The use of a weighted Euclidean inner product may be viewed as a *featurization* of the data, while the kernel expansion completes the data fitting. Thus, this renewed attention to the optimization problems associated with radial basis functions may be of broad interest.

Author Contributions: M.K. and M.A. contributed equally to the conceptualization, methodology, data curation, software development, visualization, and writing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jamshidi, A.; Kirby, M. Skew-Radial Basis Function Expansions for Empirical Modeling. *SIAM J. Sci. Comput.* **2010**, *31*, 4715–4743. [CrossRef]
2. Broomhead, D.; Lowe, D. Multivariable Functional Interpolation and Adaptive Networks. *Complex Syst.* **1988**, *2*, 321–355.
3. Jamshidi, A.A.; Kirby, M.J. A Radial Basis Function Algorithm with Automatic Model Order Determination. *SIAM J. Sci. Comput.* **2015**, *37*, A1319–A1341. [CrossRef]
4. Ma, X.; Aminian, M.; Kirby, M. Incremental Error-Adaptive Modeling of Time-Series Data Using Radial Basis Functions. *J. Comput. Appl. Math.* **2019**, *362*, 295–308. [CrossRef]
5. Jamshidi, A. Modeling Spatio-Temporal Systems with Skew Radial Basis Functions: Theory, Algorithms and Applications. Ph.D. Dissertation, Colorado State University, Department of Mathematics, Fort Collins, CO, USA, 2008.
6. Jamshidi, A.; Kirby, M. Modeling Multivariate Time Series on Manifolds with Skew Radial Basis Functions. *Neural Comput.* **2011**, *23*, 97–123. [CrossRef] [PubMed]
7. Aminian, M.; Andrews-Polymenis, H.; Gupta, J.; Kirby, M.; Kvinge, H.; Ma, X.; Rosse, P.; Scoggin, K.; Threadgill, D. Mathematical methods for visualization and anomaly detection in telemetry datasets. *Interface Focus* **2020**, *10*, 20190086. [CrossRef]
8. Scoggin, K.; Gupta, J.; Lynch, R.; Nagarajan, A.; Aminian, M.; Peterson, A.; Adams, L.G.; Kirby, M.; Threadgill, D.W.; Andrews-Polymenis, H.L. Elucidating mechanisms of tolerance to *Salmonella typhimurium* across long-term infections using the collaborative cross. *MBio* **2022**, *13*, e01120-22. [CrossRef] [PubMed]
9. Scoggin, K.; Lynch, R.; Gupta, J.; Nagarajan, A.; Sheffield, M.; Elsaadi, A.; Bowden, C.; Aminian, M.; Peterson, A.; Adams, L.G.; et al. Genetic background influences survival of infections with *Salmonella enterica* serovar Typhimurium in the Collaborative Cross. *PLoS Genet.* **2022**, *18*, e1010075. [CrossRef]
10. Radhakrishnan, A.; Beaglehole, D.; Pandit, P.; Belkin, M. Feature learning in neural networks and kernel machines that recursively learn features. *arXiv* **2022**, arXiv:2212.13881.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

A Deep Learning Model for Generalized Surface Water Flooding across Multiple Return Periods [†]

Syed Kabir *, David Wood and Simon Waller

JBA Risk Management Limited, Skipton BD23 3FD, UK; david.wood@jbarisk.com (D.W.); simon.waller@jbarisk.com (S.W.)

* Correspondence: syed.kabir@jbarisk.com

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Flood modelling is essential for addressing a range of scientific and engineering challenges. In recent years, the high computational demands of solving shallow water equations numerically have led researchers to explore machine-learning-based emulators for predicting floods and flood risk. Specifically, the proliferation of convolutional neural networks in solving different scientific problems has encouraged researchers to investigate their applicability in flood modelling. Most of these studies, however, have focused on specific locations or hydrological conditions, meaning that their findings may not be directly applicable to other situations without additional data and further training. We present here a U-Net model, a popular deep learning algorithm, which has the capacity to approximate maximum flood depths across multiple return periods while maintaining catchment generalizability. The model was trained using the outputs from a 2D hydraulic model (JFlow) to predict maximum water depths for a set of rainfall return periods (20, 100 and 1000 years). The pre-trained model was then applied to estimate depths in three unseen catchment areas. Our results demonstrate that U-Net can be used to approximate water depths in previously unseen catchments with significantly less computational time compared to the 2D model.

Keywords: rapid flood modelling; machine learning; deep learning; catchment generalization; flood inundation

Citation: Kabir, S.; Wood, D.; Waller, S. A Deep Learning Model for Generalized Surface Water Flooding across Multiple Return Periods. *Eng. Proc.* **2023**, *39*, 94. <https://doi.org/10.3390/engproc2023039094>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 24 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The shallow water equations, derived from the Navier–Stokes equations, are commonly used to model hydrological processes and flood dynamics. Numerically solving these governing equations offers a reliable method for describing the physical process of water flow. However, applying such methods in large-scale applications can prove challenging and time-consuming [1–3]. This often leads to a conflict between the necessity for precise results and the practical feasibility of obtaining them [4]. The issue becomes particularly important for larger domains with high spatial resolutions (i.e., small raster grid sizes) [2].

Considerable research effort has been dedicated to enhancing the performance of conventional numerical models. Strategies include simplifying the equations by disregarding the inertial and advection terms of the momentum equation [5], leveraging high-performance computing facilities [6], and utilizing graphics processing units (GPUs) [7,8]. Alternatively, non-physically based models, such as the transition rules of the cellular automata method [9], have been used to predict water depths over large areas. While this non-physically based approach accelerates the hydrological calculations, its primary drawback lies in its sensitivity to time steps and spatial resolutions [2], and an increase in spatial resolution can potentially result in a tenfold increase in simulation time [9].

There is a need for innovative modelling approaches that can address the technical challenges of generating actionable information while alleviating the computational load.

One promising solution is to use machine learning (ML) models that can emulate the outputs of the computationally expensive 2D hydrodynamic models. While ML techniques for rainfall–runoff forecasting have been in use for a few decades, studies applying ML to flood inundation modelling are more limited [3].

Recently, deep learning (DL), and more specifically convolutional neural networks (CNNs), have been increasingly applied in data-driven flood modelling [3]. However, most research has focused on creating models for specific drainage systems, which restricts the applicability of such modelling approaches. For example, Kabir et al. [3] developed a CNN-based fluvial flood inundation model tested in the downstream of the Eden catchment (UK). Similarly, a Gaussian process-based neural network model was tested in the same catchment area [10]. Guo et al. [1] applied an autoencoder-type model (a type of DL method often used for image reconstruction/image-to-image translation) to predict the maximum flood depths of an urban catchment.

In [11], do Lago et al. constructed a conditional generative adversarial network (cGAN) and used both topographical features and rainfall data for flood predictions in an urban catchment. They trained the cGAN model using data from multiple sub-catchments and tested it on sub-catchments outside the training areas. The U-Net, a neural network architecture widely used for image segmentation, has also been developed for predicting maximum flood depths [12]. Recently, Guo et al. [2] used the U-Net model to estimate flood depths for a 100-year storm event. In this study, the authors considered catchment generalizability, meaning the model was tested in areas beyond the training datasets with different boundary conditions. In [2], the authors demonstrated that only topographical features can be used to predict maximum water depths. These studies indicate the potential for further research in utilizing data-driven models that can be generalized to different topographical inputs.

In this study, we describe the development of a new U-Net model that emphasizes both spatial and temporal generalizability. In other words, our model can predict maximum flood depths for design storms (synthetic storm events created based on historic data) of multiple return periods while maintaining spatial transportability.

2. Method and Materials

2.1. Problem Statement

DL-based flood models need a substantial volume of flood data and high-quality terrain features for training, as well as substantial efforts to create inputs of uniform dimensionality, necessitating a systematic representation of river catchments of varying sizes [2]. Yet it is often the case that there is insufficient historical flood data on a national scale and high-resolution digital elevation models (DEMs) are not universally available, all likely contributing to the paucity of DL studies in this area.

This study aims to address these challenges by developing a new DL-based model capable of streamlining the prediction process at the catchment scale. We make use of high-resolution DEM data to extract terrain features and introduce a systematic data discretization method designed to accommodate drainage systems of varying sizes, thereby effectively training the model to predict maximum flood depths for 3 design storms (i.e., 20-, 100- and 1000-year return periods). As this is a supervised learning task, the model is trained using input–output instances where the inputs consist of various terrain features and the outputs are the maximum depths estimated from simulations using a detailed 2D hydraulic model [13].

2.2. Study Area and Data

For this study, we collected terrain data—the primary inputs to the DL model—corresponding to 28 catchments from across England, UK. These selected catchments cover most of the country (Figure 1), and these datasets exhibit an overlap at the boundaries with adjacent catchments. Of the 28 catchments, 25 serve as the training and validation

sets, and the remaining 3 are for testing. These 3 test catchments are located in 3 different regions (south, north and centre) of the country.

The target data for the DL model are maximum surface water flood maps generated by a 2D hydraulic model in response to 3 different design storms. We use the proprietary hydraulic model JFlow, developed by JBA Consulting, which solves the 2D shallow water equations and leverages graphics processing units to facilitate large-scale simulations in a swift and efficient manner. A model description and example applications appear elsewhere in the literature [13].

The UK surface water flood map utilizes precipitation depth at a 5 km grid resolution, using the rainfall intensity duration frequency (IDF) model described by [14], (often referred to as the FEH13 model). These IDF curves are translated into event hyetographs for each 5 m × 5 m DEM grid cell used by JFlow using the ReFH2 method that produces a storm profile augmented by losses due to soil storage and urban/rural land classification. The method is described in [15,16].

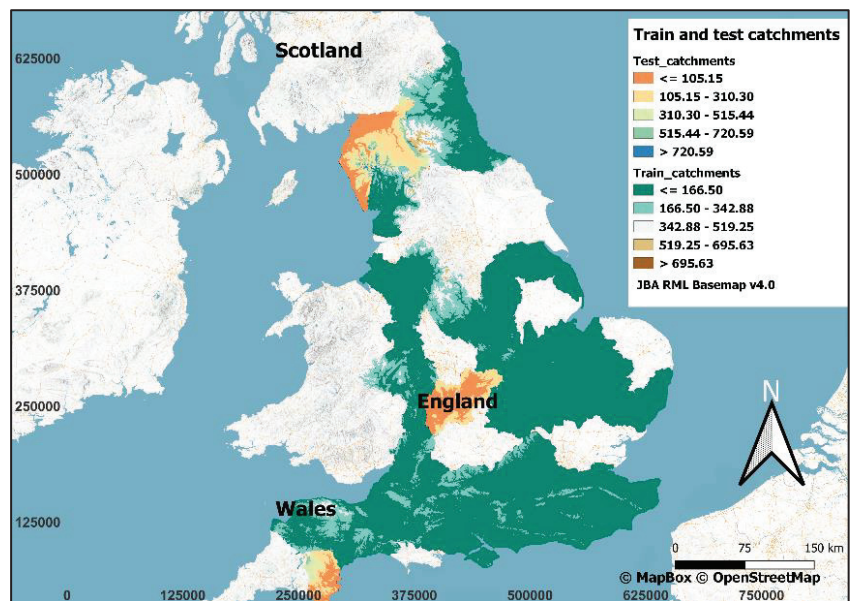


Figure 1. All training (includes validation catchments) and 3 test catchments.

2.3. U-Net

The U-Net architecture, proposed by Ronneberger and colleagues in 2015 [17], is an autoencoder-like structure equipped with skip connections and is predominantly utilized in image segmentation tasks. The U-Net consists of a contracting pathway designed to encapsulate the context, and a symmetrically expanding pathway that facilitates accurate localization. Skip connections, bridging the contracting and expanding pathways, permit the model to utilize low-level features for high-precision segmentation. U-Net has attained significant popularity within the realm of medical imaging, where it has demonstrated unparalleled performance across a spectrum of segmentation tasks.

2.4. Error Statistics

To assess the performance of the proposed U-Net model in emulating the results of JFlow, the model predictions in terms of maximum water depths are directly compared with the outputs from the hydraulic model. The root-mean-square error (RMSE) [18] and the modified index of agreement (D1) [19] are used to evaluate the overall model performance in capturing the maximum flood depths. In addition, the critical success index

(CSI), as used by [11], was used to assess the spatial performance of the predicted maps. The expressions of these indices are described in Table 1.

Table 1. Evaluation metrics used in this study.

Indicator	Formula ¹	Range and Optimal Score
RMSE	$\sqrt{\frac{\sum_i^N (O_i - P_i)^2}{N}}$	[(0, ∞), 0]
D1	$1 - \frac{\sum_i^N O_i - P_i }{\sum_i^N (O_i - \bar{O} + P_i - \bar{O})}$	[(0, 1), 1]
CSI	$\frac{\text{Hits}}{\text{Hits} + \text{False pos.} + \text{Misses}}$	[(0, 100), 100%]

¹ N is the sample size; O_i, P_i and \bar{O} are the ‘observed’, ‘predicted’ and ‘observed mean’ values. Hits are the flooded cells in both U-Net and JFlow, false positives are the flooded cells predicted by only the U-Net model and misses refer to the cells only predicted by the JFlow model.

3. Experimental Details

This section provides the key details related to the data pre-processing, the U-Net construction and the model training procedure.

3.1. Data Pre-Processing

In the domain of data-driven modelling, the efficacy of a model is significantly influenced by the quality and relevance of the input data. For flood depth modelling specifically, Löwe et al. [12] identified 11 potential terrain datasets, each encapsulating a wide range of topographical features. However, we could not repeat this given the substantial computational resources and more extensive network architecture that using 11 datasets would necessitate. Our primary focus was on exploring the feasibility of constructing a transferable data-driven model capable of estimating maximum flood depths across various return periods. As such, identifying the optimal set of model inputs was beyond the scope of this investigation and, consequently, we acknowledged that the terrain features used in our model have not been optimized.

The data pre-processing consisted of a four-step process. For the first step, terrain features such as surface elevation, flow accumulation and slope were computed from the DEM. In addition, a drainage mask (binary raster where cells within channels were encoded as ones and the remaining cells were zeros) was used as the fourth input. The resolution of the input data was downgraded from 5 m to 10 m to expedite training times.

The second step involved dividing terrain features by their respective maximum values to rescale the input datasets within a range of 0–1. Additionally, invalid cells were replaced with zero and the target datasets were filtered by assigning a value of zero to depths less than 0.1 m.

For the third step, a systematic patch generation method was used to develop training data patches. This process involved padding the zeros along the catchment boundaries to equalize their sizes, followed by selecting a patch size of 1024 × 1024 using a moving window technique. During this phase, data augmentation techniques, such as vertical and horizontal flipping, were used to increase the size of the training data samples.

For the fourth and final step, the patches from step 3 were stacked to form raster maps composed of multiple image channels. The dimensions of an input patch were set to 2 × 1024 × 1024 × 4, where 2 refers to the batch size (comprising the actual patch and an augmented patch, either vertically or horizontally), 1024 refers to the size of the patch and 4 refers to the number of channels. The dimensions of an output patch were set to 2 × 1024 × 1024 × 3, where the 3 refers to the flood depths corresponding to the 3 design storms (with return periods of 20, 100 and 1000 years).

3.2. U-Net Architecture and Training

We used a U-Net with the aim of maintaining detailed spatial patterns in the outputs while also ensuring a large ‘receptive field’, which refers to the ‘visible pixels’ of the input layer for each output pixel [20]. From a hydrological perspective, a larger receptive field

facilitates the capture of water flow from upstream to downstream, which results in water pooling in smaller regions, as the model can learn from global terrain information as opposed to merely local terrain patterns [21]. Consequently, the model's latent layer (the last layer of the encoder section) possesses a receptive field larger than the input size. We selected an input size of 1024×1024 for the U-Net model to retain local and global information, while also ensuring the model's size was compatible with the computing device and did not exceed memory capacity.

Figure 2 shows the architecture of our U-Net model. Overall, the model comprises four layers, with each layer consisting of two convolutional layers followed by a 'maxpooling' layer. We use the 'Leaky Relu' activation function in all layers other than the output layer, which uses the 'rectified linear unit (Relu)' activation function. The 'Leaky Relu' activation function offers two advantages: it circumvents the vanishing gradient problem [22] and mitigates the 'dead neuron' issue associated with the 'Relu' activation function [23]. The 'Relu' function was used in the output layer to ensure that predicted values are always above zero. The 'kernel size' of the encoder section was set to 5×5 , the upsampling (decoder) was 3×3 and the 'maxpooling' size was 2×2 .

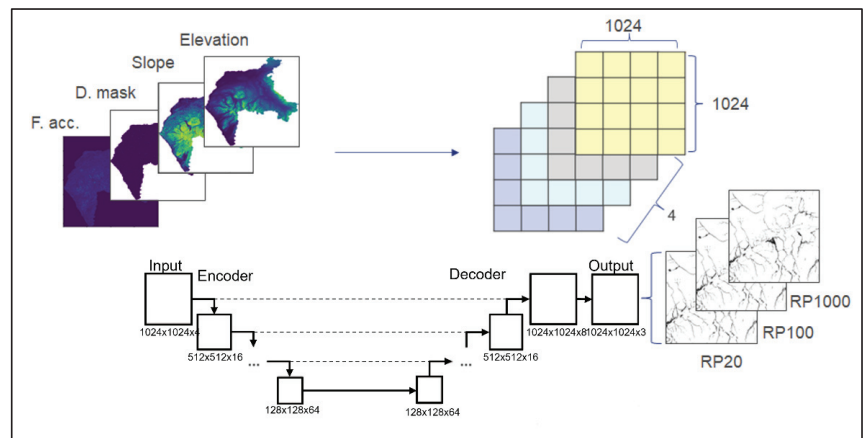


Figure 2. The U-Net model architecture. Systematically generated patches from the terrain features are stacked as image channels and fed as model inputs; water depths corresponding to 3 return periods are the network outputs. For clarity, not all network layers are shown.

Our U-Net model was constructed using the 'Keras' application programming interface (API) in conjunction with Tensorflow 2.5.0. The model's training was facilitated by the Adam Optimizer [24], which operated with a learning rate of. Due to memory constraints associated with the available graphics card (Nvidia Quadro RTX 5000), a batch size of two was implemented. The model was trained over a span of 2000 epochs, with the mean square error (MSE) serving as the loss function.

Finally, the training process was repeated 3 times using the same network architecture for 3 different training and validation datasets (each time, 20 different catchments were used for training and 5 for validation from a set of 25 catchments). This was done to observe any significant differences in the predicted flood maps when different training and validation data were used. Training the U-Net three times means that we have three models with three different model parameters (weights and biases). These three models can be used independently to predict maximum water depths or can be treated as a three-member ensemble model.

4. Results and Discussion

4.1. Training and Validation Loss

The associated training and validation loss for 2000 epochs is shown in Figure 3. The models continue to exhibit convergence tendencies beyond the 2000 epoch mark, but we ceased training at this point to prevent model overfitting, thereby maintaining their generalizability and performance on unseen data. An additional consideration was the substantial computational time required for the training of a single model (~20 days on average).

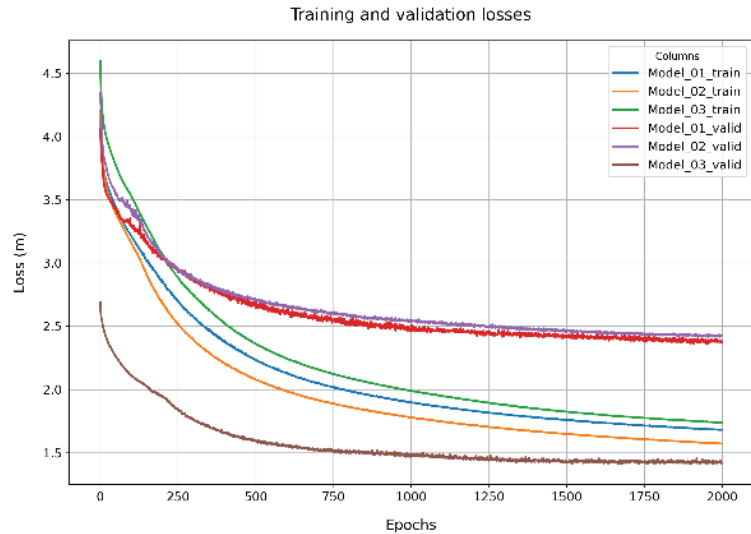


Figure 3. Training and validation losses for all three U-Net models.

4.2. Assessing the Outputs—Larger Depth for a Larger Event

After the model training, we evaluated the depths corresponding to various return periods. The models need to avoid producing counterintuitive outcomes, such as predicting elevated water levels for storms of lower return periods compared to those with higher return periods. As previously noted, the catchments exhibit overlapping boundaries. Consequently, a comparison of results would not be accurate if a portion of the test catchments were employed for training purposes. To address this issue, we chose three subdomains from each of the catchments for the purpose of comparing depth maps. This approach ensures that the comparisons were based on distinct geographic areas, eliminating the potential bias that could result from overlapping catchment boundaries in the training and test datasets.

Aside from a few minimal discrepancies, we found that the models were indeed capable of predicting increased depths for higher return periods. For instance, following the aggregation process, where maximum depths from all three models were combined into one maximum depth map, one centrally situated test catchment (Area 5404) had a single pixel where the depth associated with a return period of 100 exceeded that of a return period of 1000. Overall, such inconsistencies were noted in 17 pixels spread across the 3 test catchments. Comparatively, the total count of pixels in the case of JFlow maps amounted to 13,790. These observations indicate that the U-Net model's learning was guided more by global terrain features than local ones, demonstrating its capability for generalization.

4.3. Comparison of U-Net and JFlow Map Outputs

To compare the flood maps, we converted the predicted and the reference (JFlow outputs) water depths into categorical maps. Depth values less than 0.1 m were set to 0 (dry)

and otherwise to 1 (wet). The CSI, a commonly used metric for categorical forecasting, was utilized to ascertain the model's ability to accurately discern wet and dry cells. The CSI encompasses both false alarms and misses, thereby providing a more balanced assessment of actual model performance. The CSI scores indicate that the U-Net model was less accurate than JFlow in detecting wet cells (Table 2). However, a closer look at the flood maps revealed that the U-Net model accurately detected wet cells in regions characterized by channels, tributaries, valleys and sinks, but less accurately in urban environments. This can be attributed to the distinct terrain characteristics of urban areas and the bias in our training data, which predominantly represent rural or semi-urban areas. We also found that the model struggled to accurately simulate flooding along transport lines (roads, railways) and impervious urban areas. At the same time, a clear success was that the model did not predict the presence of water in areas where that would be implausible. Moreover, the model was successful in identifying hotspots, which are areas where water predominantly accumulates.

Table 2. Error statistics for the test catchments.

Storm	Area	CSI (%)	RMSE (m)	D1
20-year	Area 7400 (North)	20	0.04	0.69
100-year		28	0.05	0.64
1000-year		40	0.09	0.60
20-year	Area 5404 (Centre)	31	0.04	0.74
100-year		37	0.05	0.76
1000-year		46	0.07	0.78
20-year	Area 4600 (South)	37	0.03	0.72
100-year		43	0.04	0.74
1000-year		45	0.07	0.74

4.4. Comparison of U-Net Model and JFlow Depth Outputs

As with the map outputs, we conducted a comparative analysis of the U-Net model's predicted depths against those from JFlow. For the purposes of comparison of depth, all cells with a water depth less than 0.1 m were assigned a value of 0. This adjustment was implemented consistently across both models. The discrepancies in water depth predictions were systematically quantified using RMSE and D1. The RMSE metric assigns relatively high weights to large errors and was particularly useful when such errors are deemed undesirable. The modified index of agreement, represented by D1, has the advantage of appropriately weighting errors and differences, without inflation due to squared values.

Our analysis revealed a stronger concurrence in the depth maps, though the model consistently underestimated the depths by a smaller margin. Higher D1 values indicate that the U-Net model demonstrates good performance in estimating water depths. However, a worse performance was found for 'Area 7400' compared to the other two areas (Table 2). This can be attributed to the unique nature of the terrain in that catchment, where the surface elevation was higher compared to those in the training data.

Figure 4 compares a U-Net-model-predicted flood against an equivalent JFlow-simulated map, also showing the difference between the two. The comparison is consistent with the quantitative error measurements in Table 2, and it is evident that the U-Net model consistently underestimates both the extent and depths of flooding across the test catchments. However, despite the overall discrepancies in the predicted depths, most of the errors tend to cluster within the lowest error band (Figure 4C). While there are instances of large errors, these errors do not occur in areas that should remain void of water accumulation.

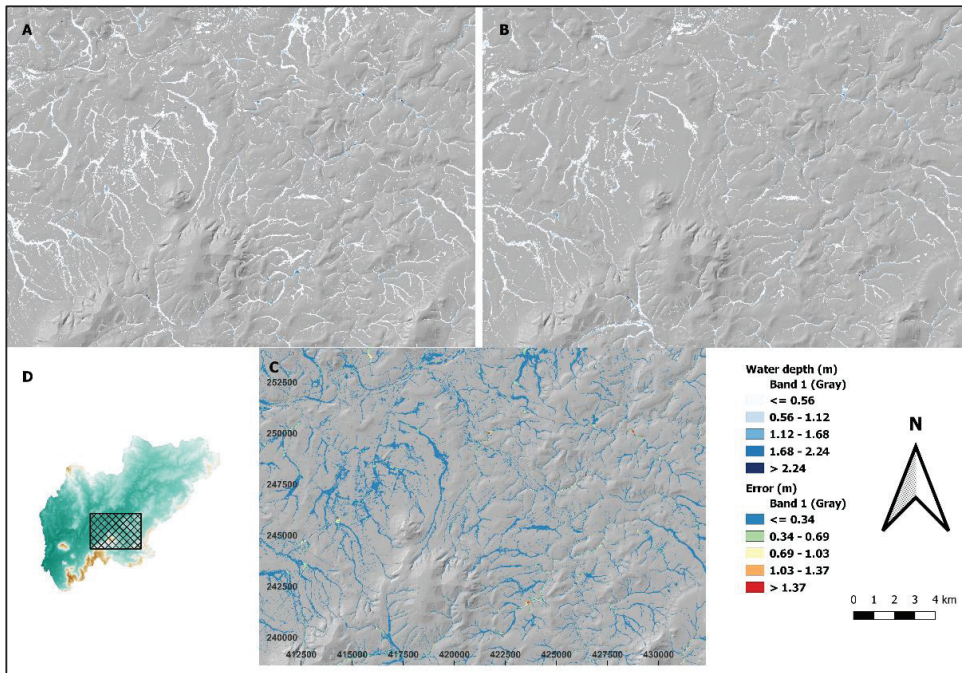


Figure 4. A visual comparison of water depths between JFlow and U-Net for a 1000-year return period storm event. (A) JFlow-simulated flood map; (B) U-Net-predicted flood map; (C) error map; and (D) map location.

5. Conclusions

We have presented a generalised data-driven model for predicting maximum flood water depths, which demonstrates that advanced DL algorithms can detect flood zones efficiently using terrain attributes.

A key objective of formulating a data-driven model was the rapid production of flood maps. The three U-Net models trained within the scope of this study demonstrate this potential. In this study, we formulated and evaluated three U-Net models for swift flood prediction utilizing topographical features (we aggregated outputs from these models to produce a single flood map for each return period). The findings suggest that the models have the potential to predict flood depths in uncharted catchments across multiple return periods. Nevertheless, the models also frequently underestimate both the depth and extent of water. This underscores the need for further refinement of the model to enhance its accuracy while maintaining its speed, highlighting an avenue for future research and development.

Temporally speaking, each model was capable of estimating depths for three return periods across a domain of approximately 1248 square kilometres within an impressive timeframe of roughly 13 s. Such efficiency may prove valuable in rapid response and planning scenarios, despite the trade-offs inherent in the data-driven approach.

However, our objective was not exclusively to construct a model for either fully urban or rural areas, but rather a hybrid model that encompasses both. The performance of the model could potentially be enhanced through the optimization of network architecture, fine-tuning of hyperparameters and a systematic search for suitable input data. In [12], the authors proposed a forward selection methodology that could potentially be utilized to identify the most suitable set of inputs. Furthermore, it is essential to recognize that

strategic selection of terrain features can significantly streamline the process of exploratory data analysis, substantially reducing the time invested in this phase.

Additionally, our observation of significant variability in the water depths predicted by the three models underscores the necessity for a comprehensive assessment of uncertainty. These strategies could collectively contribute to a more robust and accurate model, reinforcing its predictive capacity while also providing a more nuanced understanding of the inherent uncertainty in such predictions.

A trade-off exists between the water depth estimations produced by a hydraulic model and those derived from a data-driven model. Hydraulic models, underpinned by physical laws and centuries' worth of scientific theory and formulae, are generally deemed more reliable. Conversely, data-driven models do not inherently account for physical constraints, such as mass balance. Given this, one good scenario would be to have a data-driven model capable of generating flood maps expeditiously while maintaining an acceptable margin of error.

Author Contributions: Conceptualization, S.K. and D.W.; methodology, S.K.; software, S.K.; validation, S.K.; formal analysis, S.K.; investigation, S.K.; resources, D.W.; data curation, S.K.; writing—original draft preparation, S.K.; writing—review and editing, S.K., D.W. and S.W.; visualization, S.K.; supervision, D.W.; project administration, S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was internally funded by JBA Risk Management.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Training data are proprietary.

Acknowledgments: The DTMs were supplied by Airbus Defence and Space (AD&S), Environment Agency (EA), Natural Resources Wales (NRW) and Bluesky. © Environment Agency copyright and/or database right 2015. All rights reserved. Contains Natural Resources Wales information © Natural Resources Wales and database right. All rights reserved.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Guo, Z.; Leitão, J.P.; Simões, N.E.; Moosavi, V. Data-driven flood emulation: Speeding up urban flood predictions by deep convolutional neural networks. *J. Flood Risk Manag.* **2021**, *14*, e12684. [CrossRef]
- Guo, Z.; Moosavi, V.; Leitão, J.P. Data-driven rapid flood prediction mapping with catchment generalizability. *J. Hydrol.* **2022**, *609*, 127726. [CrossRef]
- Kabir, S.; Patidar, S.; Xia, X.; Liang, Q.; Neal, J.; Pender, G. A deep convolutional neural network model for rapid prediction of fluvial flood inundation. *J. Hydrol.* **2020**, *590*, 125481. [CrossRef]
- Kochkov, D.; Smith, J.A.; Alieva, A.; Wang, Q.; Brenner, M.P.; Hoyer, S. Machine learning-accelerated computational fluid dynamics. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2101784118. [CrossRef] [PubMed]
- Bates, P.D.; De Roo, A.P.J. A simple raster-based model for flood inundation simulation. *J. Hydrol.* **2000**, *236*, 54–77. [CrossRef]
- Neal, J.; Dunne, T.; Sampson, C.; Smith, A.; Bates, P. Optimisation of the two-dimensional hydraulic model LISFOOD-FP for CPU architecture. *Environ. Model. Softw.* **2018**, *107*, 148–157. [CrossRef]
- Crossley, A.; Lamb, R.; Waller, S. Fast solution of the shallow water equations using GPU technology. In Proceedings of the BHS Third International Conference—Managing Consequences of a Changing Global Environment, Newcastle upon Tyne, UK, 19–23 July 2010.
- Xia, X.; Liang, Q.; Ming, X. A full-scale fluvial flood modelling framework based on a high-performance integrated hydrodynamic modelling system (HiPIMS). *Adv. Water Resour.* **2019**, *132*, 103392. [CrossRef]
- Guidolin, M.; Chen, A.S.; Ghimire, B.; Keedwell, E.C.; Djordjević, S.; Savić, D.A. A weighted cellular automata 2D inundation model for rapid flood analysis. *Environ. Model. Softw.* **2016**, *84*, 378–394. [CrossRef]
- Donnelly, J.; Abolfathi, S.; Pearson, J.; Chatraborty, O.; Daneshkhah, A. Gaussian process emulation of spatio-temporal outputs of a 2D inland flood model. *Water Res.* **2022**, *225*, 119100. [CrossRef] [PubMed]
- do Lago, C.A.F.; Giacomoni, M.H.; Bentivoglio, R.; Taormina, R.; Gomes, M.N.; Mendiondo, E.M. Generalizing rapid flood predictions to unseen urban catchments with conditional generative adversarial networks. *J. Hydrol.* **2023**, *618*, 129276. [CrossRef]

12. Löwe, R.; Böhm, J.; Jensen, D.G.; Leandro, J.; Rasmussen, S.H. U-FLOOD—Topographic deep learning for predicting urban pluvial flood water depth. *J. Hydrol.* **2021**, *603*, 126898. [CrossRef]
13. Lamb, R.; Crossley, M.; Waller, S. A fast two-dimensional floodplain inundation model. *Water Manag.* **2009**, *162*, 363–370. [CrossRef]
14. Stewart, E.J.; Jones, D.A.; Svensson, C.; Morris, D.G.; Dempsey, P.; Dent, J.E.; Collier, C.G.; Anderson, C.A. Reservoir Safety—Long Return Period Rainfall. Available online: https://assets.publishing.service.gov.uk/media/602e43e2e90e0709e3127489/_long_return_report_1.pdf (accessed on 21 July 2023).
15. Kjeldsen, T.R. The revitalised FSR/FEH rainfall-runoff method. In *FEH Supplementary Report No. 1*; Centre for Ecology & Hydrology: Wallingford, UK, 2007.
16. Kjeldsen, T.R.; Miller, J.D.; Packman, J.C. Modelling design flood hydrographs in catchments with mixed urban and rural land cover. *Hydrol. Res.* **2013**, *44*, 1040–1057. [CrossRef]
17. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
18. Barnston, A.G. Correspondence among the Correlation, RMSE, and Heidke Forecast Verification Measures; Refinement of the Heidke Score. *Weather. Forecast.* **1992**, *7*, 699–709. [CrossRef]
19. Willmott, C.J. On the Evaluation of Model Performance in Physical Geography. In *Spatial Statistics and Models*; Gaile, G.L., Willmott, C.J., Eds.; Springer: Dordrecht, The Netherlands, 1984; pp. 443–460. [CrossRef]
20. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. *arXiv* **2017**, arXiv:1701.04128.
21. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-Trained CNNs Are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness. 2018. Available online: <http://arxiv.org/abs/1811.12231> (accessed on 16 May 2023).
22. Hochreiter, S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **1998**, *6*, 107–116. [CrossRef]
23. Nair, V.; Hinton, G.E. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on International Conference on Machine Learning, Madison, WI, USA, 21–24 June 2010; pp. 807–814.
24. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Macroeconomic Adverse Selection in Machine Learning Models of Credit Risk [†]

Joseph L. Breeden ^{*‡} and Yevgeniya Leonova [‡]

Deep Future Analytics LLC, Santa Fe, NM 87505, USA; leonova@deepfutureanalytics.com

* Correspondence: breeden@deepfutureanalytics.com

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

‡ These authors contributed equally to this work.

Abstract: Macroeconomic adverse selection is computed as a time series of forecast residuals via the vintage origination model for an industry dataset of auto loans. The adverse selection time series are computed separately as model residuals using logistic regression, neural networks, and stochastic gradient boosted trees to predict defaults in the first 24 months of a loan. Panel data versions of these models with lifecycle and environment inputs from a segmented Age-Period-Cohort analysis were also estimated. The estimates show that panel data methods make better use of available data to provide faster estimates of adverse selection risk in recent vintages and incorporate defaults at any age of the loan. The nonlinear modeling advantages of neural networks and stochastic gradient boosted trees did not significantly alter the estimates of adverse selection. Overall, all methods confirmed that macroeconomic adverse selection was dramatically higher in 2021 and 2022 for US auto loan originations.

Keywords: adverse selection; credit scoring; survival models; neural networks; stochastic gradient boosted trees

1. Introduction

The COVID-19 pandemic brought rapid, dramatic swings in economic conditions and consumer behavior. Monitoring of credit quality has shown deterioration in many loan categories. In auto lending, credit quality deterioration appears to have begun in the second quarter of 2021 and extended at least through the end of 2022. When normalized for changes in the credit quality of borrowers using logistic regression models, the residual credit risk appears to correlate to the rapid rise in new and used car prices and the rise in auto loan interest rates. This suggests a period of macroeconomic adverse selection similar to what was observed between 2006 and 2009.

Since the 2009 mortgage crisis, the lending industry has widely adopted new methods from machine learning and artificial intelligence in lending. Adoption for credit risk assessment and underwriting has been slower than other industries because of regulatory demands, but research and experimentation are extensive [1] and deployment will continue to grow. Given that machine learning has greater flexibility for finding nonlinear patterns, some proponents have suggested that such methods may be able to incorporate the structure that is showing up as macroeconomic adverse selection in regression-based models.

The current research analyzes auto loan data for originations from 2002 through 2022. Origination scores predicting the likelihood of being 60+ days past due (DPD) are estimated using logistic regression, discrete time survival models, stochastic gradient boosted trees (SGBT), SGBT with Age-Period-Cohort (APC) inputs, neural networks, and neural networks with APC inputs. For each model, time series of the residual errors are estimated by origination (vintage) month and compared across models. This research is the first to compare the estimation of adverse selection time series by vintage across regression

Citation: Breeden, J.; Leonova, Y. Macroeconomic Adverse Selection in Machine Learning Models of Credit Risk. *Eng. Proc.* **2023**, *39*, 95.

<https://doi.org/10.3390/engproc2023039095>

<https://doi.org/10.3390/engproc2023039095>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 24 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

and machine learning methods. Although some lenders quantify adverse selection across product types on internal data, our study is the first to publish a history of macroeconomic adverse selection for auto loans on a broad industry dataset.

This study finds clear advantages to panel data methods with APC inputs over traditional scoring methods for rapidly estimating adverse selection within a portfolio. This advantage is both in better use of the available data and having a stronger baseline versus age of the loan and calendar date against which to compare.

Section 2 provides an overview of the literature. Section 3.1 describes the available data. Section 3.2 provides brief descriptions of the modeling techniques used. Section 4 provides the results.

2. Literature Review

Adverse selection broadly means that the credit quality of borrowers is not what was expected when the loans were written according to the loan-origination model employed. The earliest work on adverse selection [2] described how this can happen when lenders compete on loan pricing and terms. Given a choice, borrowers will apply for the loan with better terms first. The borrowers rejected by the bank with better pricing then apply to the lender with less attractive pricing. The bank with higher pricing may have expected higher yields, but by attracting only the riskier borrowers, the losses could be higher and yields actually fall.

Adverse selection through competitive pressure can be described as microeconomic adverse selection. Sometimes trends of adverse selection are apparent across the industry. Dubbed macroeconomic adverse selection [3], this has been found to correlate to changes in the cost of the goods being purchased and changes in the cost of borrowing [4]. The theory is that depending upon macroeconomic conditions, the pool of borrowers may shift in ways not observable from the data typically available on the borrowers.

The work by Breeden in 2011 [3] observed this through three credit cycles between 1990 and 2006 for mortgages. Breeden and Canals-Cerda in 2016 [4] performed a detailed analysis of credit quality before and through the 2009 mortgage crisis to adjust for all loan-application information and found that half of the credit quality deterioration could not be explained by poor underwriting. Instead, it appeared to correlate to the cost of homes and mortgage interest rates. This effect can be observed in all product categories. Calem, Canon, and Nakamura (2011) [5] related adverse selection in home equity lines of credit to county-level unemployment and consumer confidence.

3. Materials and Methods

3.1. Data

Auto loan performance data from 26 lenders was modeled in order to assess adverse selection trends. This dataset included 1,244,651 loans originated between January 2005 and December 2022. Typical origination variables were used, including Bureau Score, Loan-To-Value (LTV) ratio, Debt-To-Income (DTI) ratio, channel (direct or indirect), collateral type (new or used), term, state, and anonymized lender ID.

Behavioral variables such as delinquency that are included in a typical behavior score are a post-origination attempt to adapt to the difference between origination expectation and post-origination reality. For that reason, including delinquency in our models would dilute the measure of adverse selection. However, some information is available immediately after the loans are originated that would not be included in a traditional origination score. Primary among these are the offered interest rate on the loan (APR), and the balance of the loan. The annual percentage rate (APR) offered by the lender may incorporate information not made available for creation of the origination scores, such as adjustments for specific dealers, form of employment, broader relationship with the lender, etc. Therefore, we created a measure $\Delta APR(i, v, L)$ that computed the difference between the APR on a specific loan within a specific vintage $APR(i, v, L)$ and the average APR for that vintage by

lender $\overline{APR}(v, L)$ where v is the vintage and L is the lender in order to incorporate some of this missing knowledge.

$$\Delta APR(i, v, L) = APR(i, v, L) - \overline{APR}(v, L) \tag{1}$$

To make measures of adverse selection useful to portfolio managers, the estimation needs to occur as early as possible in the life of a vintage. Therefore, default D has been defined as the date on which an account first becomes ≥ 60 days past due (DPD). When logistic regression (LR), neural networks (NN), or stochastic gradient boosted trees (SGBT) were used to create a traditional origination credit score, the outcome period for default was the first 24 months of the life of the loan. A total of 2428 such defaults exist within the dataset, a 24-month default rate of 0.2%.

When using models that are normalized to a lifecycle or hazard function, defaults at any loan age are relevant. The total number of defaults through the entire life of the loans was 55,435, a 4.45% lifetime default rate. The panel data approach thus incorporates defaults from vintages that have not yet reached 24 months old and considers additional defaults later in the lifetime of the loans, which will be important to interpreting the results obtained later.

3.2. Algorithms

Several modeling techniques were compared in order to determine the robustness of estimations of macroeconomic adverse selection as driven by external forces rather than simply estimation noise.

3.2.1. Logistic Regression

Logistic regression is the traditional method for creating origination scores. For a thorough introduction to credit scoring, see Thomas, Crook, and Edelman [6] or Anderson [7]. The cumulative probability of default for the first 24 months of the loan is predicted as

$$\text{logit}(D_i) \sim \sum_{j=1}^n c_j s_{ij} + c_0 \tag{2}$$

where the c_j are the n estimated coefficients for the scoring factors s_{ij} for account i . The set of scoring factors s are chosen to optimize the Akaike Information Criterion (AIC). Some of the variables are binned to capture nonlinearities.

After the model has been created, the residual error by vintage is estimated by creating a second regression with the original model $M(c, s_i)$ as a fixed input.

$$\text{logit}(D_i) \sim M(c, s_i) + \sum_v g_v \delta_v \tag{3}$$

where δ_v is a delta function for vintage date v and g_v is the corresponding coefficient. Fixed effects by vintage (dummy variables) could have been included in the original regression, Equation (2). In some situations, this can shift some of the explanatory power of the scoring factors to the vintage fixed effect. Since the goal here is to obtain maximum explanatory power from the scoring factors and use the vintage effects only to measure the residuals, a two-step process was employed.

3.2.2. Age-Period-Cohort Models

Age-Period-Cohort (APC) models [8–10] for vintage analysis explain the risk of default at each observation period as a combination of functions of the age a of the loan, the calendar date t , and the vintage date v . These functions can be spline approximations, non-parametric, or other forms, but are generally not tied to specific scoring factors.

Because $a = t - v$, a model-specification error exists if no constraints are imposed. In applications to credit risk analysis, the following representation is common.

$$D \sim b_0 + b_1a + F'(a) + b_2v + G'(v) + H'(t) \tag{4}$$

where b_0 is the intercept, b_1 and b_2 are the linear coefficients for a and v , and $F'(a)$, $G'(v)$, and $H'(t)$ are the nonlinear functions that have zero mean and no linear component. For explanation, these are usually combined as $F(a) = b_0 + b_1a + F'(a)$, $G(v) = b_2v + G'(v)$, $H(t) = H'(t)$ where $F(a)$ is called the lifecycle measuring the timing of losses through the life of the loan, $G(v)$ is the vintage function measuring credit risk by vintage, and $H(t)$ is the environment function measuring the net impact from the environment (primarily economic conditions). The primary advantage of APC models is the ability to separate these effects, so the credit risk function captures the full amount of credit quality variation, but cleaned of impacts from the macroeconomic environment and normalized for differences in the age of the loans. The credit risk function does not adjust for loan-level changes in underwriting, so it is not a perfect measure of adverse selection. However, if the analysis is segmented by key measures such as bureau score and term, a net residual credit risk function can be extracted that can serve as an adverse selection measure.

3.2.3. Discrete Time Survival Models

Discrete time survival models [11,12] are a form of panel regression where each account is observed each month to predict default/no default. The regression equation can include nonparametric lifecycle and environment functions as in APC models and scoring factors as in logistic regression. Cox proportional hazards [13] models are the original continuous time formulation of this, where the APC-style lifecycle is a discrete time version of the Cox PH hazard function.

Previous work has shown that survival models that are estimated via a partial likelihood estimation as with Cox PH or a logistic regression estimation of the full panel model have instabilities in the context of credit risk modeling [14]. The instability occurs, in part, because the model-specification error of the APC model appears as colinearity between the scoring factors, lifecycle, and environmental factors of the survival model.

Breeden [15] proposed a solution to this where an initial APC decomposition is performed as described above and the lifecycle and environment functions are taken as fixed inputs to a second panel logistic regression estimation.

$$D \sim F(a) + H(t) + \sum_{j=1}^n c_j s_{ij} + c_0 \tag{5}$$

This two-step process resolves any colinearities between scoring factors, lifecycle, and environment while retaining maximum explanatory power for the scoring factors. In practice, this has proven to create scores that are more stable through changes in the environment while retaining account-level predictive accuracy.

Adverse selection is measured in a final step as described in Equation (3) except as a panel logistic regression.

3.2.4. Artificial Neural Network

Using artificial neural networks (NN) for credit risk forecasting has been the subject of numerous publications [16–18]. The problem design is similar to creating a logistic regression credit score, but with the network allowing for nonlinearity and interaction effects that would need to be discovered manually and encoded into the inputs of a regression model.

The available training data for auto loan defaults is not particularly complex compared to many applications of neural networks, and thus is not a showcase for the nonlinear wonders of machine learning. However, it is sufficient to address the question of whether

adverse selection is a model-specific error or due to a hidden variable that is not discoverable by any algorithm using only traditional data.

The neural network architecture was correspondingly simple. The network had an input layer, five fully connected layers with softplus activation functions, and a sigmoid output node. Softplus is less efficient and some argue less interpretable than ReLU activation functions, but it had better convergence performance in this context. The target was the same binary indicator of default within 24 months as used in the logistic regression model with a binary cross-entropy loss function.

Neural networks such as this do not function well when defaults comprise only 0.2% of the training data. Previous research has shown that at least a 4:1 or 3:1 ratio is needed for proper network estimation [19,20]. In this case, all default accounts were included and four times as many non-default accounts were randomly sampled from the dataset. The resulting network predictions need to be balanced back in order to match the overall default probability of the original training dataset.

3.2.5. NN + APC

Within the domain of credit risk modeling, having data from 2005 through 2022 is considered a significant amount of history. Compared to economic cycles, it is not. One problem with neural networks or any scoring technique with a wide (24 month) outcome period is that fragments of an economic cycle get confused with scoring attributes. The primary theoretical advantage of discrete time survival models over logistic regression is creating a distinction between environmental trends and credit risk trends that are explainable from scoring factors.

Analogous to the discrete time survival models, the lifecycle and environment from the APC models can be provided as inputs to the neural network with the data arranged as a panel of repeated observations for the accounts until default or payoff [21]. The network architecture is arranged such that the APC inputs $O(a, t) = F(a) + H(t)$ in units of log-odds of default are passed to the final node as an offset without modification. The neural network is used only as a replacement for the credit risk component, effectively modeling the account-level residuals around the long-term trends of lifecycle and environment.

For proper estimation, the dataset still requires balancing. The input offset needs to be adjusted with an additive constant for any change in default probabilities due to rebalancing. The revised offset, O' , is

$$O' = O(a, t) + \left(\log \left(\frac{\bar{p}}{1 - \bar{p}} \right) - \bar{O} \right) \quad (6)$$

When the network produces forecasts, the original offset $O(a, t)$ is used without the rebalancing adjustment factor. As with the plain NN, the final dataset for model estimation under-sampled the loans that never default in order to achieve a 4:1 ratio with loans that eventually will default. Model training was performed on 80% of this balanced dataset and cross-validation on 20% to determine the stopping point.

The part of the network dedicated to processing the origination factors can have the same architecture as that used without the APC inputs. However, providing APC inputs often allows for a simpler network architecture. The target variable for the network was default that occurs at any point in the life of the loan, as done in the DTSM, allowing the larger panel dataset to be modeled.

3.2.6. Stochastic Gradient Boosted Trees

Decision trees are as old as credit risk modeling [22,23]. The multidimensional space described by the scoring attributes is split with hyperplanes to separate good from bad accounts. A slightly more sophisticated version fits a regression model within each terminal node of the tree, as in CART [24]. Stochastic gradient boosted trees [25,26] are essentially an ensemble modeling approach where each new regression tree is weighted to explain the

data points that were less explainable by the preceding set of regression trees. Trees are added until no significant improvement is obtained on a test set.

Tree-based methods do not suffer from multicollinearity problems as regression does, so additional inputs can be provided without destabilizing the model. Therefore, the SGB Tree was provided with all of the inputs given to the logistic regression and neural network models as well as factor variables for state and lender. These additional inputs might allow the algorithm to better handle outliers. The target variable is again whether an account defaults within the first 24 months, as used in the logistic regression and neural network origination scoring models. Unlike the neural network approach, no balancing of default and non-default data is required for model convergence.

In most credit scoring competitions, SGBT has been a winning approach. Recent research by Grinsztajn, Oyallon, and Varoquaux [27] suggests that tree-based models will perform better than neural networks for tabular data structures where neighboring input factors may have no ordering or continuity. Neural networks have been found to excel in sound and image processing applications where the inputs are neighboring pixels in an image or sequential points in the time sampling.

For the current work, the goal is not to declare a winner, but rather simply to compare the residuals of these methods versus vintage origination date. For consistency of comparison, vintage date is again excluded from the inputs and adverse selection is quantified via a final logistic regression as in Equation (3) where $M(s_i)$ is the full ensemble of trees applied to forecasting account i and held as a fixed input when measuring the vintage residuals.

3.2.7. SGBT + APC

Some implementations of stochastic gradient boosted trees allow for the same kind of fixed inputs as logistic regression and the NN+APC algorithm above. Again using $O(a, t) = F(a) + H(t)$ as a fixed input allows us to create an SGBT credit risk panel model that is centered around the long-term trends of lifecycle and environment. As observed with NN+APC, the resulting hybrid model can be both simpler and more robust out-of-sample as compared to the stand-alone SGBT model.

The inputs to the credit scoring SGBT model were the same as for the DTSM using the full panel dataset, where defaults occurring at any age are included. This is the same dataset used for NN+APC models. Because of the volume of data, the model was estimated on a 5% random sample of loans, including the full history for each loan. During training, 80% of the 5% sample was used for training and 20% for cross-validation to determine the stopping point. Model residuals by vintage were estimated by applying the models to the full dataset.

4. Results

Nine separate models were estimated using seven different techniques, including the APC decomposition. For regression models, measures of LTV and term were binned to allow for nonlinearities in their relationship to default. collateral type and channel categorical variables are measured relative to their reference levels, which are indicated with a 0 estimate. Not all lenders reported DTI, so a separate flag for DTI missing was included and DTI missing was interacted with DTI to capture the correlation to default.

All of these independent variables are available at loan origination, which is the traditional design of an origination score. For purposes of measuring adverse selection, we are concerned with the loans that are actually booked. Therefore, we can additionally incorporate information available just after origination. We call this a “post-origination score”. The most useful factor was found to be $\Delta APR(i, v)$ as defined in Equation (1). Adding $\Delta APR(i, v)$ to the model improved the in-sample fit, lowering AIC from 23,800 to 23,314, and actually improved the significance of the channel and collateral type coefficients. The rest of the models were estimated post-origination so that the adverse selection measure removes as much structure as possible from the independent variables.

To confirm that the models were estimated properly, receiver operating characteristic (ROC) curves were estimated for each model. Because of the different datasets for the models with 24-month outcome periods and models with APC inputs using panel data, the Gini coefficients are unlikely to be directly comparable. Further, the NN and SGBT models were estimated on samples and applied to the full dataset, so most of those test results are out-of-sample. Regardless of the many test differences, the results in Table 1 show that all models are working as expected. Logistic regression performs normally given that it uses only the first 24 months of performance data and therefore is missing a majority of the defaults that occur later. APC has the lowest Gini coefficient of the panel methods, because it has no loan-level information and makes no attempt to be a scoring model. The DTSM, NN, and APC models all perform comparably given the uncertainty in estimating the test statistics and different handling of the data. For example, SGBT + APC Post-Orig Score has the most sampling disadvantages of the models, yet performed comparably.

Table 1. Gini coefficients for the models tested.

Method	Traditional	+ APC
LR/DTSM Orig	0.664	0.853
LR/DTSM Post-Orig	0.703	0.853
NN	0.839	0.822
SGBT	0.868	0.836
APC		0.773

With confirmation that the models are performing properly, the following analysis compares the model residuals by vintage. Figure 1 computes the change in adverse selection by vintage comparing origination to post-origination logistic regression scores. The figure shows that including APR information in the score does refine our understanding of adverse selection in 2016–2017, where the origination score residuals are overestimated and 2018–2019 where the origination score residuals were underestimated. The scale of ± 0.1 in units of change in log-odds is roughly equivalent to a $\pm 10\%$ change in credit risk. This is not large compared to the underlying measures of adverse selection shown in subsequent graphs, but not immaterial. In general, we conclude that post-origination models provide some advantage when measuring adverse selection as a way to incorporate underwriting policy changes that might not otherwise be captured in the models.

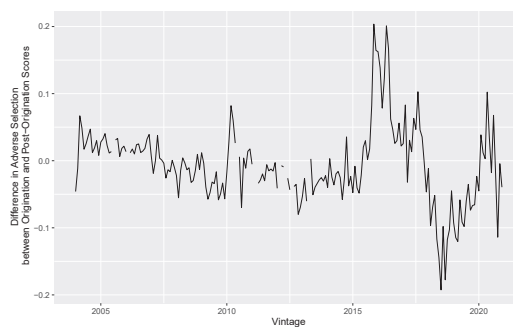


Figure 1. The difference between macroeconomic adverse selection measures for a logistic regression score using information available at origination versus that which is available post-origination but excluding behavioral data.

The next step was to compare adverse selection as measured for logistic regression, neural networks, and stochastic gradient boosted trees, Figure 2. These models are post-origination credit scores using default in the first 24 months as the target variable. Tests were run using other outcome periods, but they were less effective. Extending the outcome

period to 36 months captures significantly more loan defaults, but it delays the measurement of adverse selection by that same three years. In order to have business value, the waiting time for estimating adverse selection must be as short as possible. At the other extreme, we could have looked at the first 12 months in the life of a loan to assess residual credit risk. Some lenders even focus on first payment default as an early warning indicator. Although potentially useful, we would run out of data with which to construct models. The trade-offs are challenging.

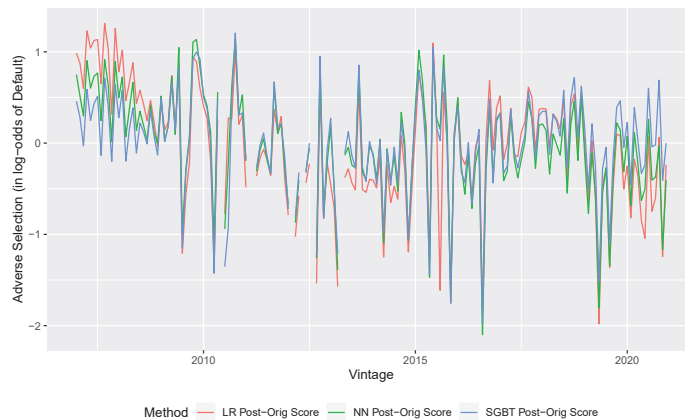


Figure 2. A comparison of macroeconomic adverse selection measures from logistic regression, neural networks, and stochastic gradient boosted trees using defaults within the first 24 months of a loan as the target variable.

Most notable when comparing these measures of adverse selection is the overall similarity. This suggests that adverse selection is more of an attribute of the loans than of the models. Although the NN and SGBT models can capture much more nonlinearity and interaction between variables, the modeling technique does not fully explain the structure within the data. The model residuals (adverse selection) for NN and SGBT are closer to the through-the-cycle average during 2020, 2014, and 2007, but they are not flattened entirely. Notably, the periods of high risk from 2008 through 2009 are still present and are consistent with prior mortgage studies of heightened adverse selection during that period. Better loan quality from 2011 through 2014 is also consistent with the prior mortgage results, so although the estimates are volatile by monthly vintage, they are broadly consistent with expectations.

One solution to the trade-off between quicker response and more defaults is the use of a survival modeling approach where defaults at any age are compared to a baseline expectation from a hazard function or lifecycle. This kind of analysis could be implemented in many ways. Beginning with an Age-Period-Cohort decomposition provides a complete measure of credit risk by vintage but without explanation, Figure 3. That APC decomposition uses lifecycles segmented by bureau score and term, so it is adjusted for dominant scoring factors, but not LTV, DTI, channel, or collateral type.

Taking lifecycle and environment as fixed inputs to a panel logistic regression (DTSM) allows us to further adjust for lender shifts in origination volume by LTV, DTI, channel, or collateral type. The adverse selection measured for the DTSM is overlaid in Figure 3. The comparison of APC vintage function to DTSM adverse selection shows that they are very similar. A small divergence occurs between 2012 and 2016, but recent measures are very well aligned. This suggests that the segmented APC analysis is a quick, computationally efficient way to capture most of the adverse selection problem, although there can be situations where an account-level score brings further refinement. Those advantages might

be more acute when measured for a single lender where the volume by loan attributes can swing more rapidly.



Figure 3. A comparison of the credit risk estimate by vintage from an Age-Period-Cohort analysis and the residuals by vintage from a discrete time survival model with the same APC lifecycles and environment as fixed inputs.

Figure 4 compares the three methods of panel estimation with APC inputs. From a data perspective, the comparison is still not fair. Even though large servers were used for the analysis, the full panel dataset (performance data for every month of every loan) is 32,028,587 rows of data. That far exceeds what could be processed using standard libraries for stochastic gradient boosting (gbm) and neural networks (keras) in R in reasonable time. Therefore, the NN and SGBT models downsampled the non-default loans so that the model training sets were only 342,895 rows of observations. Conversely, the APC vintage decomposition uses all of the data in vintage aggregate form and the DTSM used all observations within the panel data. Regardless of sampling and algorithm used, the vintage-aggregate residuals for each model are remarkably similar.

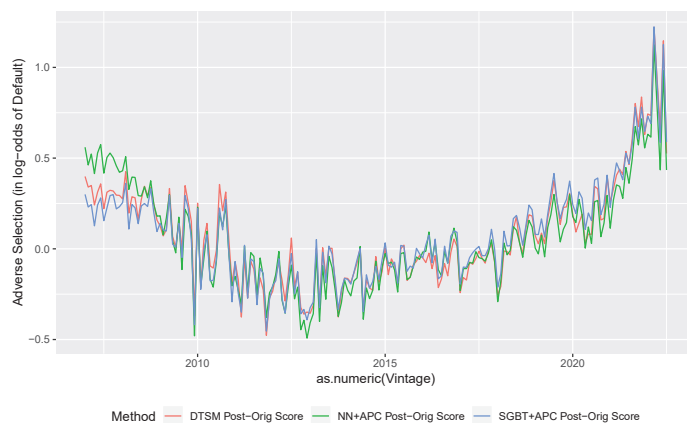


Figure 4. A comparison of the macroeconomic adverse selection estimates from the best DTSM, NN, and SGBT models.

The biggest difference in estimating adverse selection by vintage is seen to be the difference between traditional scoring data and panel data. Comparing Figure 2 to Figure 4

makes clear that scoring methods are significantly more volatile in their adverse selection estimates, simply due to the smaller number of defaults available for modeling, only 4.4% of the total number of lifetime defaults observed. Moreover, because of the 24-month lag in estimation, the scoring approach provides no indication of recent trends.

The recent trends in adverse selection are quite important. All of the panel approaches with APC inputs show that adverse selection has been dramatically higher since February 2021. Within the auto lending industry, this is assumed to be caused by the jump first in the cost of new and used vehicles and later by the increase in the cost of borrowing. Those pressures are hypothesized to have pushed the “value shoppers” out of the market, leaving the less flexible or financially savvy buyers. This is the same dynamic observed leading up to the 2009 recession.

In the second half of 2022, those selection pressures began to ease with the cost of vehicles coming down and auto loan APRs decreasing by the end of the year. The dramatic drops seen at the end of these trends for the most recent months are based upon only a few months of observations and have correspondingly large uncertainties.

These results provide compelling evidence that a panel approach with APC inputs is superior to measuring adverse selection from a traditional scoring approach, but it leaves open which modeling technique is best.

5. Conclusions

The concept of macroeconomic adverse selection became clear during the period 2006 through 2009 when poor quality loans were originated beyond what lenders could expect from their usual scoring inputs. The conditions of rapidly rising home prices and rising interest rates appear to have created an unappealing environment for financially cautious borrowers. The macroeconomic conditions in 2021–2022 resemble this prior period, but with even more extreme rates of change. This led us to suspect that adverse selection would again occur.

This study was undertaken in part to confirm this intuition about the presence of macroeconomic adverse selection in recent auto originations, which was shown here. In addition, the analysis demonstrated that models which create scores estimated relative to lifecycle and environment measures from APC or survival analysis can more rapidly and accurately identify emerging periods of adverse selection. This is valuable from a business perspective so that measuring adverse selection becomes actionable intelligence rather than a retrospective curiosity.

Contrary to some suggestions, machine learning models cannot explain adverse selection as missing nonlinear structure. Rather, the adverse selection measured from neural network and stochastic gradient boosted tree models, even with APC inputs, had residual credit risk by vintage that was statistically unchanged relative to discrete time survival models. This confirms that residual credit risk by vintage should not be viewed as model error but rather as a real indicator of macroeconomic adverse selection. Ideally, sociodemographic data not currently available for model development might quantify the presence or absence of value shoppers, but they would have to be data that are not restricted due to discrimination risks or privacy concerns.

Author Contributions: Conceptualization, J.L.B. and Y.L.; methodology, J.L.B. and Y.L.; software, J.L.B. and Y.L., validation, J.L.B. and Y.L.; writing, J.L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data used in this study is proprietary to the contributing institutions and not publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Breeden, J. A survey of machine learning in credit risk. *J. Credit. Risk* **2021**, *17*, 3. [CrossRef]
2. Wilson, C. Adverse selection. In *Allocation, Information and Markets*; Palgrave Macmillan: London, UK, 1989; pp. 31–34.
3. Breeden, J.L. Macroeconomic adverse selection: How consumer demand drives credit quality. In Proceedings of the Credit Scoring and Credit Control XII Conference, Edinburgh, UK, 30 August–1 September 2011.
4. Breeden, J.L.; Canals-Cerdá, J.J. Consumer risk appetite, the credit cycle, and the housing bubble. *J. Credit. Risk* **2018**, *14*, 1–30. [CrossRef]
5. Calem, P.S.; Cannon, M.; Nakamura, L.I. *Credit Cycle and Adverse Selection Effects in Consumer Credit Markets-Evidence from the Heloc Market*; FRB of Philadelphia: Philadelphia, PA, USA, 2011; Working Paper No. 11–13.
6. Thomas, L.; Crook, J.; Edelman, D. *Credit Scoring and Its Applications*; SIAM: Singapore, 2017.
7. Anderson, R. *Credit Intelligence & Modelling: Many Paths through the Forest*; Oxford University Press: Oxford, UK, 2019.
8. Fu, W. *A Practical Guide to Age-Period-Cohort Analysis: The Identification Problem and Beyond*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018.
9. Holford, T.R. The estimation of age, period and cohort effects for vital rates. *Biometrics* **1983**, *39*, 311–324. [CrossRef] [PubMed]
10. Mason, W.M.; Fienberg, S. *Cohort Analysis in Social Research: Beyond the Identification Problem*; Springer: Berlin/Heidelberg, Germany, 1985.
11. De Leonardis, D.; Rocci, R. Assessing the default risk by means of a discrete-time survival analysis approach. *Appl. Stoch. Model. Bus. Ind.* **2008**, *24*, 291–306. [CrossRef]
12. Stepanova, M.; Thomas, L. Survival analysis methods for personal loan data. *Oper. Res.* **2002**, *50*, 277–289. [CrossRef]
13. Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. Ser.* **1972**, *34*, 187–220. [CrossRef]
14. Breeden, J.L.; Leonova, E.; Bellotti, A. *Instabilities Using Cox ph for Forecasting or Stress Testing Loan Portfolios*; Researchgate: Berlin, Germany, 2019.
15. Breeden, J.L. Incorporating lifecycle and environment in loan-level forecasts and stress tests. *Eur. J. Oper. Res.* **2016**, *255*, 649–658. [CrossRef]
16. Angelini, E.; Di Tollo, G.; Roli, A. A neural network approach for credit risk evaluation. *Q. Rev. Econ. Financ.* **2008**, *48*, 733–755. [CrossRef]
17. Desai, V.S.; Crook, J.N.; Overstreet, G.A., Jr. A comparison of neural networks and linear scoring models in the credit union environment. *Eur. J. Oper. Res.* **1996**, *95*, 24–37. [CrossRef]
18. Khashman, A. Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Syst. Appl.* **2010**, *37*, 6233–6239. [CrossRef]
19. Laurikkala, J. Improving identification of difficult small classes by balancing class distribution. In Proceedings of the Artificial Intelligence in Medicine: 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001, Cascais, Portugal, 1–4 July 2001; Springer: Berlin/Heidelberg, Germany, 2001; pp. 63–66.
20. Sundarkumar, G.G.; Ravi, V. A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. *Eng. Appl. Artif. Intell.* **2015**, *37*, 368–377. [CrossRef]
21. Breeden, J.L.; Leonova, E. When big data isn't enough: Solving the long-range forecasting problem in supervised learning. In Proceedings of the 2019 International Conference on Modeling, Simulation, Optimization and Numerical Techniques (SMONT 2019), Shenzhen, China, 27–28 February 2019; Atlantis Press: Amsterdam, The Netherlands, 2019; pp. 229–232.
22. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
23. Ali, K.; Pazzani, M. Error reduction through learning multiple descriptions. *Mach. Learn.* **1996**, *24*, 172–202. [CrossRef]
24. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 1984.
25. Bastos, J. *Credit Scoring with Boosted Decision Trees*; Technical Report MPRA Paper No. 8034; CEMAPRE, School of Economics and Management (ISEG), Technical University of Lisbon: Lisbon, Portugal, 2007.
26. Chang, Y.-C.; Chang, K.-H.; Wu, G.-J. Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Appl. Soft Comput.* **2018**, *73*, 914–920. [CrossRef]
27. Grinsztajn, L.; Edouard Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data? *arXiv* **2022**, arXiv:2207.08815.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Proceeding Paper

Sensor Virtualization for Anomaly Detection of Turbo-Machinery Sensors—An Industrial Application †

Sachin Shetty ^{1,*}, Valentina Gori ^{2,*}, Gianni Bagni ^{2,*} and Giacomo Veneri ^{2,*}

¹ Baker Hughes, Doddanakundi Industrial Area 2, Bengaluru 560037, India

² Baker Hughes (Nuovo Pignone Tecnologie), Via Felice Matteucci 2, 50127 Firenze, Italy

* Correspondence: sachin.shetty@bakerhughes.com (S.S.); valentina.gori@bakerhughes.com (V.G.); gianni.bagni@bakerhughes.com (G.B.); giacomo.veneri@bakerhughes.com (G.V.)

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: We apply a Granger causality and auto-correlation analysis to train a recurrent neural network (RNN) that acts as a virtual sensor model. These models can be used to check the status of several hundreds of sensors during turbo-machinery units' operation. Checking the health of each sensor is a time-consuming activity. Training a supervised algorithm is not feasible because we do not know all the failure modes that the sensors can undergo. We use a semi-supervised approach and train an RNN (LSTM) on non-anomalous data to build a virtual sensor using other sensors as regressors. We use the Granger causality test to identify the set of input sensors for a given target sensor. Moreover, we look at the auto-correlation function (ACF) to understand the temporal dependency in data. We then compare the predicted signal vs. the real one to raise (in case) an anomaly in real time. Results report 96% precision and 100% recall.

Keywords: virtual sensor; anomaly detection; time series multi-regression; Granger causality; turbo-machinery

1. Introduction

Turbo-machinery units are equipped with hundreds of sensors to monitor their health during functioning [1,2]. Some of these sensors measure primary physical quantities, which can affect the overall health of the machine. Thus, detecting the improper behavior of sensors or mechanical equipment is a critical task in energy [3,4] and the mechanical industry or, generally speaking, in every IOT-related industry [5]. Detecting unexpected behavior is also a challenging task [2,6]; indeed, in many real-world problems, samples from the unexpected classes are of insufficient sizes to be effectively modeled using supervised algorithms [7]. Anomaly detection identifies novelty cases by training only on samples considered normal and then identifying the unusual cases [8–10].

1.1. Problem Statement

In this domain, monitoring some sensors is important because they can trigger alerts; in that case, a machine shutdown and manual inspections are required, with an associated cost. Sometimes the triggers are false since they are caused by a sensor failure, not by a machine issue. Hence, early detection is required to avoid undesired shutdowns. Indeed, if a sensor is about to break, service operations can exclude this sensor from the control strategy.

We want to detect possible faults (anomalies) in the sensors installed on our turbo machines (Figure 1) to prevent unnecessary inspection/shutdown efforts by site engineering while making sure that correct triggers, instead, are not ignored.

Citation: Shetty, S.; Gori, V.; Bagni, G.; Veneri, G. Sensor Virtualization for Anomaly Detection of Turbo-Machinery Sensors—An Industrial Application. *Eng. Proc.* **2023**, *39*, 96. <https://doi.org/10.3390/engproc2023039096>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 27 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

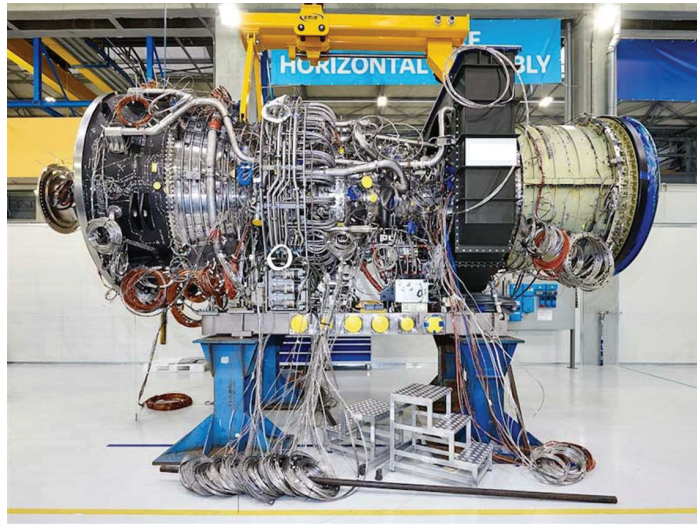


Figure 1. Turbine: a turbo machine is a system that transfers energy between a rotor and a fluid, including both turbines and compressors. While a turbine transfers energy from a fluid to a rotor, a compressor transfers energy from a rotor to a fluid.

The challenge consists in dealing with these aspects:

- **Early** detection is required: only a prompt action allows to avoid the high potential costs of unnecessary shutdowns.
- **Up to few thousand sensors** need to be checked **daily**.
- **Recall is key**: anomalies detected by the tool will be checked by operators and vice versa, where if no alert is given, the anomaly may remain undetected.
- **Precision should be kept under control**: too many false positives would increase the set of signals to be checked and may invalidate the benefits.

1.2. Related Works

Many other authors have tried to solve similar problems with different techniques: Malhotra et al. [11] apply recurrent neural networks (RNNs) for anomaly detection on aircraft. Park et al. [12] and Pereira [13] uses variational recurrent autoencoder and clustering to detect anomalous time series in healthcare. Geiger et al. [14] applies generative adversarial networks (GANs) and LSTM to identify the temporal correlations of time-series distributions (see also [15,16]). Zheng et al. [17] apply long short-term memory for residual useful life estimation. In a similar research, Strazzer et al. [1] confirm that LSTM outperforms the not recurrent neural network also in the domain adaptation. Zhang et al. [18] extend reinforcement learning (RL) and the Markov decision process [19] to build a general framework for fault prediction and residual useful life estimation. Several other authors (Yang [20], Pawełczyk and Sepe [21]) use machine-learning-based prediction models for gas turbine operating parameters estimation (see also [22] for a small review). They find that machine learning techniques are applicable to any of the gas turbine parameters when reference physics-based models and large sensor measurements datasets are available to validate the accuracy of the data-driven algorithms developed. Escobedo [23] uses the Bayesian technique and feature extraction to scale up to a broad large mechanical equipment fleet.

2. The Dataset

Our data are output from all sensors installed on a turbo machine [1,2] and are acquired at a frequency of one sample per second. Different kinds of sensors like temperature,

pressure, speed sensors have been acquired and comprise our database. Among these sensors, the ones which are critical for machine control are considered “output sensors” in our work. In fact, those are the sensors whose health needs to be monitored to be sure that an eventual alarm triggered by them is actually due to a machine failure, not to a probe failure. The remaining sensors can be used as input features for building virtual sensor models (digital twins) of the first set of sensors. In this work, we focus on one target sensor only to explain the process more easily.

The dataset was collected during 14 months of machine operation (1 s sampling interval). It was split into training (10 months data), validation (1 month data) and test (3 months data) sets. The training data have no reported anomaly, while the validation and test sets have some anomalies reported.

3. The Model

3.1. Selection of Input Sensors

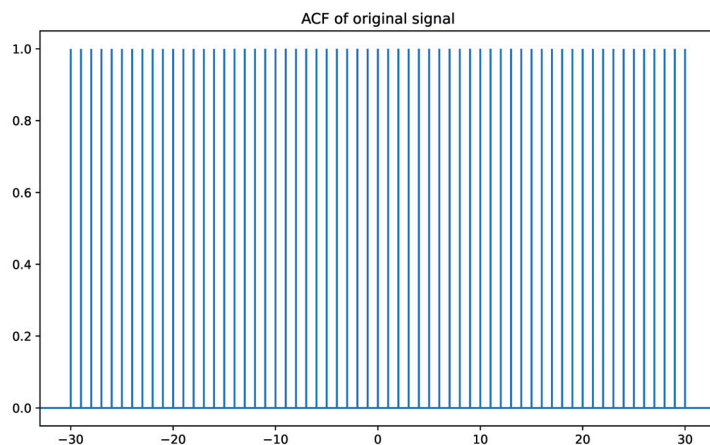
There are more than 200 sensors that can be used to build a virtual sensor for each output sensor. We used the Granger Causality [24,25] test to determine the subset of input sensors that have a causal effect on the target. For the target sensor shown here, we identified around 15 input sensors to be used to reconstruct the same.

3.2. Selection of Lookback Window

We used the auto-correlation to find the temporal relations in both input sensors and target sensor to obtain the best “window size” to train the LSTM model.

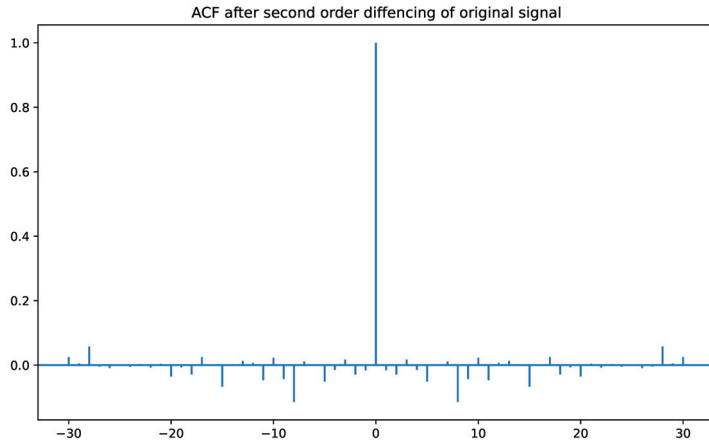
Figure 2a shows the auto-correlation function (ACF) graph for one of the sensors: we can see high correlation values among all the lags, which represent the presence of strong trend (non-stationary series). Hence, we need to make this series stationary by differentiating to view if there is any seasonality present in the data.

Figure 2b shows the auto-correlation function (ACF) graph after second-order differentiation: there is no significant seasonality and a small degree of trend. Thus, we can conclude the absence of seasonality but the presence of strong trend (strong correlation among first few lag values). Moreover, we know from subject matter experts that in turbo-machinery applications the thermocouple thermal inertia is less than 5 s [26]. Hence, we chose a sliding window of five samples for this temperature-measuring sensor selected as output. Indeed, after five samples, ACF shows highly sparse values.



(a)

Figure 2. Cont.



(b)

Figure 2. (a) Auto-correlation function plot for one of target sensor. (b) Auto-correlation function plot after second-order derivative to make series stationary.

3.3. Model Training

We used a deep learning model with two long short-term memory (LSTM) layers of 32 nodes each, with *tanh* activation, followed by four fully connected layers with *ReLU* activation. We used Adam optimizer to train the model and a callback on the validation set to stop the training. We used a semi-supervised [2] approach and trained the model on non-anomalous data only to build a virtual sensor acting like a digital twin of the sensor itself [11,27]. In Figure 3, we can see that the model is able to correctly reproduce the actual signal.

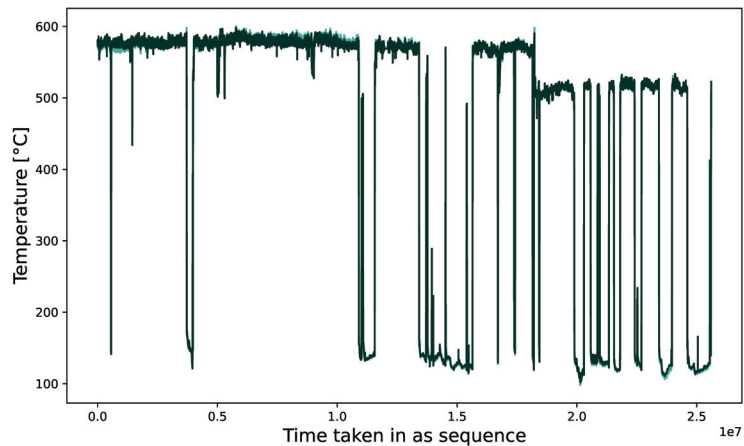


Figure 3. The picture shows the good fitting between the virtual sensor (dark green) and the actual sensor (light green) for the training set. Values were arbitrarily scaled to maintain data confidentiality.

3.4. Inference Logic

Once the model has been trained, we use it to reconstruct the signal in a time region when sensor anomalies may have occurred. To distinguish between anomalous and non-anomalous samples, we identified a criterion based on the level of agreement of the actual sensor with respect to the virtual one.

Given the actual signal y_i , with $i = 0, \dots, T$, where T is the signal length, and the related virtual signal \hat{y}_i , with $i = 0, \dots, T$, the discrepancy $\Delta y_i = \text{abs}(\hat{y}_i - y_i)$, $i = 0, \dots, T$ can be calculated. We declare y_i to be anomalous if its related Δy_i is higher than expected. This expected value was derived by looking at the values of Δy_i of non-anomalous samples in the validation set. Furthermore, given that the validation set contains both anomalous and non-anomalous samples, we leveraged the different Δy_i distribution between non-anomalous and anomalous samples to determine the threshold value. Figure 4 shows the distribution of the discrepancy Δy_i , $i = 0, \dots, T_v$, where T_v is the validation set length in the case of anomalous (orange) and non-anomalous (blue) points.

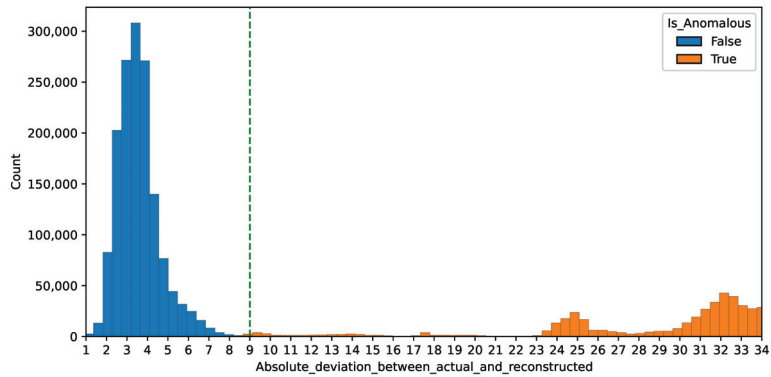


Figure 4. Deviation between the actual and the reconstructed signal in the validation set for anomalous (orange) and non-anomalous (blue) samples.

In this example, we can see two non-overlapping distributions: here, we decided to use a threshold of 9 to best discriminate between anomalous and non-anomalous samples.

Another possibility is to leverage the ROC curve to identify the optimal value for the threshold that optimally balances the true positive and false positive rate.

Figure 5 shows the model performance at test time. We can see a good agreement between the actual and the virtual signal in the region where no sensor anomalies occurred (rightmost part of the plot) and, instead, a discrepancy between them in a region where sensor anomalies are present (leftmost part of the plot).

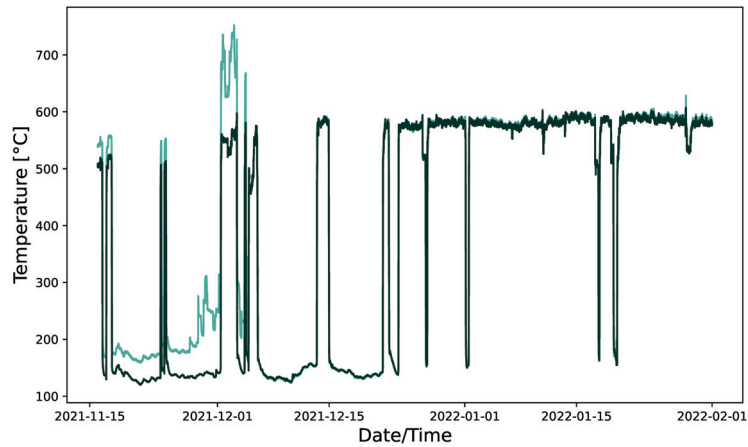


Figure 5. The picture shows the superposition of the actual signal (light green) and the reconstructed one (dark green). The region ranging from mid November to early December shows a discrepancy between the two: here, a sensor anomaly is highlighted by the model and confirmed by subject matter experts (SMEs). The remaining part of the test set shown here has no anomalies highlighted by the model nor by SMEs. *Values were arbitrarily scaled to maintain data confidentiality.*

4. Results

Table 1 shows the reconstruction performance of the model on training, validation and test sets. For validation and test sets, the metrics are evaluated only using subsets where no anomalies occurred. Please note that the error Δ is defined here as the deviation between the actual signal y and the reconstructed signal \hat{y} : $\Delta = y - \hat{y}$, then ME is the mean error, MAPE is the mean absolute percentage error, and P90 is the 90th percentile of the absolute value of the error Δ .

Table 1. Model performance on training, validation and test sets.

	ME	MAPE	P90
Training set	0.12	0.61	5.06
Validation set (non-anomalous samples only)	1.45	1.61	5.97
Test set (non-anomalous samples only)	1.89	0.65	6.52

For what concerns the anomaly detection performance, when applying the model to the test set, we are able to detect anomalous signals with 96% precision and 100% recall as summarized in Table 2.

Table 2. Anomaly detection performance on the test set.

	Precision	Recall
Full test set	96%	100%

5. Conclusions

In this work, we presented a real industrial application of sensor anomaly detection in the domain of energy and turbomachinery. We applied a semi-supervised deep learning technique, which can be used to perform anomaly detection in an industrial context. In particular, we applied anomaly detection to turbo-machinery units by training a virtual sensor model for a given sensor. We first selected input features through Granger causality and leveraged auto-correlation and subject matter expertise to identify the best window size for the recurrent neural network chosen (LSTM).

This method can be scaled and extended to almost all the sensors installed on the unit, for a complete sensor anomaly detection system.

Furthermore, once the model has been trained for a single sensor, we can later retrain the model using data collected over time, with a continual learning approach [28] so that the algorithm is able to also take into account data-shift phenomena.

Our next plans focus on the deployment of the inference algorithm on edge devices, i.e., on the MarkVIe system. For this purpose, some model distillation may be required (for a review, see [29]). In particular, we need to detect potential sensor faults as early as possible so that we can exclude the sensor from the control system, thus avoiding undesired shutdowns.

Author Contributions: S.S. and V.G.: developed the model; G.B. defined the problem and data analysis; G.V. reviewed the model and contributed to model definition and performance analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research received internal funding.

Institutional Review Board Statement: Baker Hughes internal review process: # 1440.

Informed Consent Statement: Authors authorize the use and disclosure of the following information for this research.

Data Availability Statement: Proprietary Data.

Acknowledgments: Thanks to Luca Strazzer, Andrea Panizza and Laure Barrière for the valuable discussion.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Strazzer, L.; Gori, V.; Veneri, G. DANNTe: A Case Study of a Turbo-Machinery Sensor Virtualization under Domain Shift. *arXiv* **2021**, arXiv:2201.03850.
2. Gori, V.; Veneri, G.; Ballarini, V. Continual Learning for anomaly detection on turbomachinery prototypes—A real application. In Proceedings of the 2022 IEEE Congress on Evolutionary Computation (CEC), Padua, Italy, 18–23 July 2022; pp. 1–7. [CrossRef]
3. Michelassi, V.; Allegorico, C.; Cioncolini, S.; Graziano, A.; Tognarelli, L.; Sepe, M. Machine Learning in Gas Turbines. *Mech. Eng.* **2018**, *140*, S54–S55. [CrossRef]
4. Wu, Y.; Yuan, M.; Dong, S.; Lin, L.; Liu, Y. Remaining useful life estimation of engineered systems using vanilla LSTM neural networks. *Neurocomputing* **2018**, *275*, 167–179. [CrossRef]
5. Capasso, A. *Hands-On Industrial Internet of Things: Create a Powerful Industrial IoT Infrastructure Using Industry 4.0*; Packt Publishing: Birmingham, UK, 2018.
6. Hodge, V.J.; Austin, J. A Survey of Outlier Detection Methodologies. *Artif. Intell. Rev.* **2004**, *22*, 85–126. [CrossRef]
7. Zimek, A.; Schubert, E.; Kriegel, H.P. A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat. Anal. Data Mining Asa Data Sci. J.* **2012**, *5*, 363–387.
8. Akçay, S.; Atapour-Abarghouei, A.; Breckon, T.P. GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training. *arXiv* **2018**, arXiv:cs.CV/1805.06725.
9. Akçay, S.; Atapour-Abarghouei, A.; Breckon, T.P. Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection. *arXiv* **2019**, arXiv:cs.CV/1901.08954.
10. Nanduri, A.; Sherry, L. Anomaly detection in aircraft data using Recurrent Neural Networks (RNN). In Proceedings of the 2016 Integrated Communications Navigation and Surveillance (ICNS), Herndon, VA, USA, 19–21 April 2016; pp. 5C2–1–5C2–8.
11. Malhotra, P.; Vig, L.; Shroff, G.M.; Agarwal, P. Long Short Term Memory Networks for Anomaly Detection in Time Series. In Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, (ESANN 2015), Bruges, Belgium, 22–24 April 2015; pp. 89–94.
12. Park, D.; Hoshi, Y.; Kemp, C.C. A Multimodal Anomaly Detector for Robot-Assisted Feeding Using an LSTM-Based Variational Autoencoder. *IEEE Robot. Autom. Lett.* **2018**, *3*, 1544–1551. [CrossRef]
13. Pereira, J.; Silveira, M. Learning Representations from Healthcare Time Series Data for Unsupervised Anomaly Detection. In Proceedings of the 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), Kyoto, Japan, 27 February–2 March 2019; pp. 1–7.
14. Geiger, A.; Liu, D.; Alnegheimish, S.; Cuesta-Infante, A.; Veeramachaneni, K. TadGAN: Time Series Anomaly Detection Using Generative Adversarial Networks. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022; pp. 33–43. [CrossRef]

15. Li, Y.; Peng, X.; Zhang, J.; Li, Z.; Wen, M. DCT-GAN: Dilated Convolutional Transformer-based GAN for Time Series Anomaly Detection. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 3632–3644. [CrossRef]
16. Sabuhi, M.; Zhou, M.; Bezemer, C.P.; Musilek, P. Applications of Generative Adversarial Networks in Anomaly Detection: A Systematic Literature Review. *IEEE Access* **2021**, *9*, 161003–161029. [CrossRef]
17. Zheng, S.; Ristovski, K.; Farahat, A.; Gupta, C. Long Short-Term Memory Network for Remaining Useful Life estimation. In Proceedings of the 2017 IEEE International Conference on Prognostics and Health Management (ICPHM), Dallas, TX, USA, 19–21 June 2017; pp. 88–95.
18. Zhang, C.; Gupta, C.; Farahat, A.; Ristovski, K.; Ghosh, D. Equipment Health Indicator Learning Using Deep Reinforcement Learning. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases*; Brefeld, U., Curry, E., Daly, E., MacNamee, B., Marascu, A., Pinelli, F., Berlingerio, M., Hurley, N., Eds.; Springer: New York, NY, USA, 2019; pp. 488–504.
19. Jacobs, W.R.; Edwards, H.; Li, P.; Kadiramanathan, V.; Mills, A.R. Gas turbine engine condition monitoring using Gaussian mixture and hidden Markov models. *Int. J. Progn. Health Manag.* **2018**, *9*, 1–15. [CrossRef]
20. Yang, C.; Gunay, B.; Shi, Z.; Shen, W. Machine Learning-Based Prognostics for Central Heating and Cooling Plant Equipment Health Monitoring. *IEEE Trans. Autom. Sci. Eng.* **2020**, *18*, 346–355. [CrossRef]
21. Pawełczyk, V.; Fulara, S.; Sepe, M.; De Luca, A.; Badora, M. Industrial gas turbine operating parameters monitoring and data-driven prediction. *Eksploat. I Niezawodn.* **2020**, *22*, 391–399. [CrossRef]
22. Yan, Z.; Sun, J.; Yi, Y.; Yang, C.; Sun, J. Data-Driven Anomaly Detection Framework for Complex Degradation Monitoring of Aero-Engine. *Int. J. Turbomach. Propuls. Power* **2023**, *8*, 3. [CrossRef]
23. Ernesto Escobedo, E.; Arguello, L.; Sepe, M.; Parrella, I.; Cioncolini, S.; Allegorico, C. Enhanced early warning diagnostic rules for gas turbines leveraging on bayesian networks. In Proceedings of the ASME Turbo Expo 2020: Turbomachinery Technical Conference and Exposition, Virtual, 21–25 September 2020.
24. Granger, C. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
25. Rosoł, M.; Młyńczak, M.; Cybulski, G. Granger causality test with nonlinear neural-network-based methods: Python package and simulation study. *Comput. Methods Programs Biomed.* **2022**, *216*, 106669. [CrossRef]
26. Straubinger, D.; Illés, B.; Busek, D.; Codreanu, N.; Géczy, A. Modelling of thermocouple geometry variations for improved heat transfer monitoring in smart electronic manufacturing environment. *Case Stud. Therm. Eng.* **2022**, *33*, 102001. [CrossRef]
27. Tran, K.P.; Nguyen, H.D.; Thomassey, S. Anomaly detection using Long Short Term Memory Networks and its applications in Supply Chain Management. *IFAC-PapersOnLine* **2019**, *52*, 2408–2412. [CrossRef]
28. Rebuffi, S.A.; Kolesnikov, A.; Sperl, G.; Lampert, C.H. iCaRL: Incremental Classifier and Representation Learning. *arXiv* **2017**, arXiv:cs.CV/1611.07725.
29. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge Distillation: A Survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Uncertainty and Business Cycle: An Empirical Analysis for Uruguay [†]

Bibiana Lanzilotta, Gabriela Mordecki, Pablo Tapie and Joaquín Torres *

Instituto de Economía, Facultad de Ciencias Económicas y de Administración, Universidad de la República, Montevideo 11200, Uruguay; bibiana.lanzilotta@fcea.edu.uy (B.L.); gabriela.mordecki@fcea.edu.uy (G.M.); pablo.tapie@fcea.edu.uy (P.T.)

* Correspondence: joaquin.torres@fcea.edu.uy

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: As a small and open economy, Uruguay is highly exposed to international and regional shocks that affect domestic uncertainty. To account for this uncertainty, we construct two geometric uncertainty indices (based on the survey of industrial expectations about the economy and the export market) and explore their association with the Uruguayan GDP cycle between 1998 and 2022. Based on the estimated linear ARDL models that showed negative but weak relationships between the uncertainty indices and the GDP cycle, we test for the existence of structural breaks in these relationships. Although we find a significant break in 2003 for both indices and another in 2019 for one of them, Wald tests performed on the non-linear models only confirm the structural break in the early 2000s in the model with the index based on export market expectations. In this non-linear model, we find that the negative influence of uncertainty fades after 2003. The evidence of a differential influence before and after this date remains, even when controlling for the variability in non-tradable domestic prices. Two implications arise from these results. First, the evidence of relevant changes that made the Uruguayan economy less vulnerable from 2003 onward. Second, the importance of the expectation about the future of the export market in the macroeconomic cycle of a small and open economy like Uruguay.

Keywords: uncertainty; macroeconomic cycle; expectations; Uruguay; structural breaks; ARDL models; non-linear models

JEL Classification: C53; E32; E37; E71

Citation: Lanzilotta, B.; Mordecki, G.; Tapie, P.; Torres, J. Uncertainty and Business Cycle: An Empirical Analysis for Uruguay. *Eng. Proc.* **2023**, *39*, 97. <https://doi.org/10.3390/engproc2023039097>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 28 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

A recent and growing trend in the literature that seeks to understand the fundamentals that explain the movements of macroeconomic variables has focused on uncertainty as a relevant factor. Intuitively, economic agents making decisions may not have complete information or the capacity to correctly process the information they possess. This can lead to decisions being made under uncertainty. Moreover, the 2008 global financial crisis made the importance of quantifying uncertainty and having indicators that measure its impact in real-time to detect early signs of the economic situation and contribute to timely decision-making even more evident [1].

However, uncertainty is not a directly measurable phenomenon. Therefore, the economic literature has developed different strategies to capture agents' uncertainty. Many of the strategies are based on the assumption that prediction errors increase when the uncertainty rises, stock markets become more volatile and the expectations of different economic agents are significantly divergent. Several studies measure uncertainty through the magnitude of forecast errors or by developing dispersion-based indicators of expectations [2–5]. More recent techniques, based on machine learning, developed new indicators

constructed through text analysis ([6–8]; among others), including some indicators based on news through processes that could involve some subjectivity [2].

A recent branch of the empirical literature considers survey-based measures of uncertainty. This approach relies on the fact that, in periods of higher uncertainty, there are more discrepancies between the forecasts experts or managers [9–11]. This underpins the construction of uncertainty indicators that exploit the dispersion or divergence between agents' expectations or forecasts. The underlying hypothesis is that lack of predictability and large divergence between forecasters and managers are signs of increased economic uncertainty, and this type of measure captures the uncertainty of decision-makers, who play an important role in investment and innovation decisions. Some empirical applications of this kind of uncertainty indices for macroeconomic forecast are, among others [12,13].

Empirical research on this topic refers mostly to developed economies. However, previous studies, such as [14], found substantial heterogeneity in reactions to uncertainty shocks across countries using an open-economy VAR approach. Compared to developed countries, emerging economies took a longer time to recover, and they relate this effect to the depth of financial markets. Evidence from emerging economies is still scarce (see [15]).

With the aim of contributing to the empirical literature on developing economies, we computed a survey-based uncertainty index for Uruguay following the methodology proposed by [2]. The authors, noting that most survey-based uncertainty indicators do not take into account the responses of agents who do not expect changes in the future [16], propose a time-varying disagreement metric that incorporates information from the three categories of responses. Thus, they construct a positional indicator of disagreement that can be interpreted as the percentage of disagreement between responses. A recent application of this index [17] examines the uncertainty impact on unemployment in European countries one year after the emergence of COVID-19, using two indicators that exploit the European Commission's survey of business expectations.

The Uruguayan economy has certain characteristics. First, it is a small and open economy located in South America between two large, highly volatile economies, Argentina and Brazil, with which it forms Mercosur. For this economy, the dynamic and influence of uncertainty on the economy are studied in [18,19], using different methods and measures. Ref. [18] proposed a composite index measure of macroeconomic uncertainty that, following the methodology of [20], combines external uncertainty captured by Brazil's Economic Political Uncertainty (EPU) index (Fundação Getúlio Vargas) and the Global index (Baker, Bloom, and Davis), with domestic uncertainty measured as the standard deviation of 12-month exchange rate forecasts collected by the Central Bank of Uruguay (BCU). On the other hand, Ref. [19] analyze the dynamics of manufacturing firms' expectations from a network approach, finding that higher uncertainty affects the coordination of groups of firms.

In contrast to the previous studies, this paper considers an alternative uncertainty index for Uruguay, based on economic trend surveys. Following the proposal of [2], we use the industrial monthly survey since 1998, obtained by the Uruguayan Chamber of Industry. We focus on agents' expectations about the country's situation in relation to the economy as a whole, both at present and at the end of the next six months. This survey covers 170 companies and asks about expectations for the next 6 months in relation to sales in the domestic market, sales in the foreign market (if applicable), and their expectations for the sector, the company, and the economy. The answer options are: worse, same, better, and don't know. We explore the relationship between the Uruguayan GDP cycle and uncertainty indices by applying linear and nonlinear models and structural breaks tests. This empirical strategy is in line with that proposed by recent research analyzing how economic uncertainty affects the economy in the short run ([13,21,22], among others).

The rest of the paper is organized as follows. Section 2 presents the data and methodology for the empirical analysis, introducing the uncertainty index for Uruguay based on this methodology. Section 3 presents the empirical results, and the final section, the main conclusions, and policy implications.

2. Data and Methodology

2.1. Data

For the proposed analysis, this article relies mainly on three different sources of information. First, the data used to construct uncertainty indices come from a monthly survey conducted by the CIU. This survey was created in 1997 and one of its objectives is to monitor firms in the manufacturing industry regarding the evolution of different variables. The potential responses were “better, worse, the same (or doesn’t know)”. These responses are later recorded as 1, -1 , and 0, respectively. The questions refer to industrialists’ perceptions of the following: (1) The evolution of the national economy in the next six months; (2) If the respondent firm exports, it is asked whether the physical units exported will increase, decrease, or remain the same. Companies are also asked about (3) the evolution of their own sector and (4) the evolution of the company’s domestic sales.

The sample used in this survey contained approximately 200 firms and was first constructed using as a benchmark a different sample designed by the INE. The sample is dynamic [23] in the sense that firms may enter or leave for different reasons, such as the closure of a firm or the entry into the market of a new relevant firm. By means of an analysis carried out by CIU, it is possible to state that the sample is representative of the manufacturing industry. This database contains monthly observations from October 1998 to August 2022, totaling 287 observations.

The change in the percentage of responses to the question regarding the evolution of the economy (question 1) and the evolution of the volume exported in the next six months (question 2) can be seen in Figures A1 and A2, respectively, in the Appendix A.

It can be seen that, for both questions, most companies tend to answer “the same” and that the “worse” answers are, on average, higher than the “better” answers, revealing a certain lack of optimism among companies. There is also a similar evolution of the different response options between the two questions, although differences in level are evident.

Second, the cycle of the Uruguayan economy is extracted from GDP data from Uruguayan Central Bank (BCU). To obtain the cycle, we used the Structural Time Series Analyser, Modeller and Predictor (STAMP) econometric software [24], based on quarterly data from the second quarter of 1980 to the third quarter of 2022. The estimation was performed using the logarithm of the Uruguayan GDP with the following selected options in STAMP: level selected as fixed, the slope as stochastic, the seasonal as stochastic, and an irregular component. The estimated cycle is a short cycle; an intervention analysis was introduced and the estimation method used was maximum likelihood via the BFGS numerical score algorithm.

The estimation for the cycle and the other unobserved components can be seen in Figure A3. The resulting reduction in variance in both the seasonal and the cycle components towards the observations of the latter is noteworthy. Further statistics and results from this estimation can be seen in Table A1.

Finally, in order to consider a factor linked to the domestic market, we also used the year-on-year rate of non-tradable price index (NTP), constructed updating the methodology proposed by [25]. The evolution of the non-tradable inflation can be seen in Figure A4.

2.2. Methodology

This subsection presents the methodology used to compute the uncertainty indexes, based on the aforementioned industrial survey, and to test the existence of a relationship between the economic cycle and the uncertainty index.

2.2.1. Uncertainty Indexes

To construct our geometric uncertainty index we follow [2]. This index is based on the discrepancy of responses to surveys where the answers are (or can be coded as) qualitative.

The index also includes respondents that think the variables will remain stable or will not change. The indicator is calculated as follows:

$$uncertainty_t^k = 1 - \frac{\sqrt{(R_t - \frac{1}{3})^2 + (C_t - \frac{1}{3})^2 + (F_t - \frac{1}{3})^2}}{\sqrt{\frac{2}{3}}} \quad (1)$$

where R_t denotes the share of respondents that answer that the variable will rise in the next period, while C_t is the share that answer will remain constant, F_t the share that considers that will fall and k refers to questions 1 and 2 (the Index constructed with question 1 will now be referred to as Economic Uncertainty, while the one constructed with question 2 will be referred to as Export Uncertainty), which were used to construct the geometric discrepancy index. (an analysis was also conducted for questions 3 and 4, but no significant results were found.) The highest level achievable of uncertainty is reached when the share of responses is equal, i.e., when one-third of the respondents think the variable will rise, one-third think it will remain constant and the last third think it will fall. The evolution of both uncertainty indexes is shown in Figure 1, annotated with the main international and national events in the period.

In the case of economic uncertainty, the mean of the index for the entire period is 0.45, showing two major spikes in 2008 and 2020, which are likely to be related to the financial crisis in 2008 and the appearance of the COVID-19 pandemic. The index related to export uncertainty shows a mean of 0.58, indicating higher levels of expectations misalignment on average, in comparison with the economic uncertainty index. Furthermore, the export uncertainty index does not peak in 2008 and 2020, but rather in 2002, from when a drop in misalignment is evidenced. On the one hand, it is interesting that the indexes present different evolutions because this can have a direct impact on how they relate to the GDP cycle. On the other hand, changes in their own evolution during the period may also suggest that the association with the economic cycle is not constant.

Both of the indexes are $I(0)$, and details of the test [26,27] can be seen in Table A2 in the Appendix A.

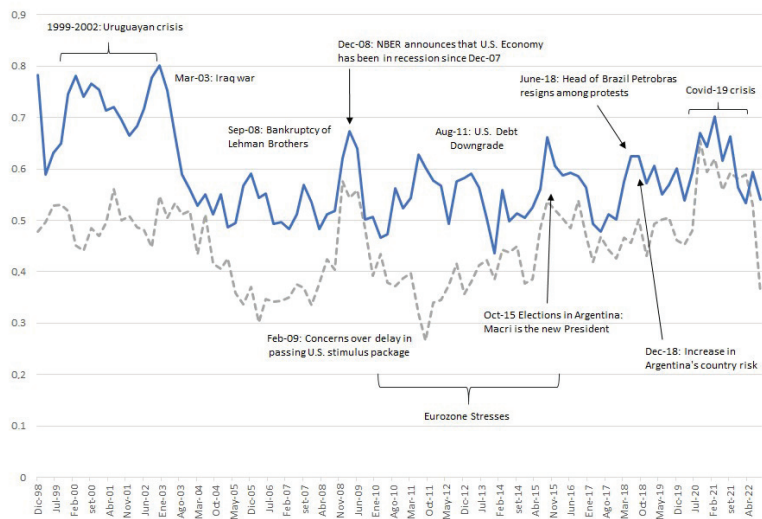


Figure 1. Geometric Uncertainty Indexes (own calculations based on CIU data). Note: Export Uncertainty Index in blue line, the Economic Uncertainty Index in dotted line.

2.2.2. Modeling the Link between the Macroeconomic Cycle and Uncertainty

Our first strategy is to estimate an autoregressive distributed lag (ARDL) model, where the economic cycle enters as the dependent variable and the uncertainty index is used as a regressor. This model can be represented as:

$$y_t^k = \sum_{i=1}^p \gamma_i y_{t-i} + \sum_{j=0}^r \beta_j x_{t-j} + \epsilon_t \quad (2)$$

where y_t is the value of the economic cycle at time t , k refers to the Index 1 or 2, x_t is the uncertainty index at time t and ϵ_t is supposed to be white noise.

No constant was introduced as the mean of the economic cycle is null by definition. The amount of lags introduced in the model was selected using the Akaike information criterion (AIC). The coefficient covariance matrix used was HAC with the Bartlett Kernel and Newey–West Fixed Bandwidth methods.

Then, we tested the existence of structural breaks using the methodology proposed by [28]. Basically, the method consists of considering all the possible partitions given m pre-established breaks and h minimum observations per sub-sample. Then, the sum squared of residuals (SSR) is computed for each partition, looking for the partition where it is minimized. In this article, we first employed a contrast to test the hypothesis of the non-existence of breaks versus the existence of a fixed number of breaks. Secondly, we performed and presented the test of l breaks versus $l + 1$ breaks to test whether the breaks are significantly different from each other. It is worth noting that, as the model utilized in Equation (2) is multivariate, the tests described previously admit structural breaks at all parameters.

Finally, if we find significant breaks in the relationship between uncertainty and the GDP cycle, we will proceed to estimate a new specification that takes this into account. Specifically, the objective is to evidence the existence of possible changes in the dynamics of the association between both series.

3. Results

This section presents the results obtained from the different estimations that were performed. The section is divided into three parts: (i) linear bivariate estimation and break search; (ii) bivariate estimation with breaks; and (iii) estimation incorporating the price index of non-tradable goods and services as a control.

3.1. Linear Bivariate Association

Tables 1 and 2 present the results of the ARDL estimation of the relationship between the GDP cycle and the economic and export uncertainty indices, respectively. As can be seen, the estimation reveals a relationship with an autoregressive component; at the same time, the contemporaneous coefficient and the first lag of the uncertainty index are also significant. However, the coefficients themselves, beyond the best modeling found, are only significant in the case of economic uncertainty. In fact, a negative relationship is found, i.e., higher levels of misalignment between economic expectations are associated with lower levels in the GDP cycle. Although not statistically significant, this association was also found in the model with foreign market expectations.

However, Uruguay, during the period of analysis, experienced important economic events (crisis in 2002, institutional and government changes, and price shocks, among others); therefore, it may make sense that the association between the series in question has not remained constant. In fact, if we look at the mean and standard deviation in the correlation between the series partitioning the sample in five-year periods, an indication of this may be evident (Table A3).

Table 1. ARDL model with Index 1 (own estimations).

Variable	Coefficient	t-Statistic	Prob
Economic Cycle (−1)	1.297712	13.59380	0.0000 ***
Economic Cycle (−2)	−0.450312	−5.399764	0.0000 ***
Uncertainty Eco	−0.025935	−2.247897	0.0270 **
Uncertainty Eco (−1)	0.024972	2.168351	0.0327 **
Adjusted R-squared: 0.862085			
Durbin-Watson Stat: 1.871796			
Jarque Bera Prob for residuals: 0.634181			
Note: Significance levels: 1% *** 5% **.			

Table 2. ARDL model with Index 2 (own estimations).

Variable	Coefficient	t-Statistic	Prob
Economic Cycle (−1)	1.302897	14.21119	0.0000 ***
Economic Cycle (−2)	−0.452242	−5.067577	0.0000 ***
Uncertainty Expo	−0.015864	−1.074760	0.2853
Uncertainty Expo (−1)	0.014817	1.008693	0.3158
Adjusted R-squared: 0.859729			
Durbin-Watson Stat 1.876346			
Jarque Bera Prob for residuals 0.552310			
Note: Significance levels: 1% ***.			

As mentioned, given the possibility that the association is not constant, we tested for the possible existence of breaks. The results for the Bai and Perron sequential L vs. L+1 test about the presence of significantly different structural breaks in our models are presented in Table 3.

Table 3. Bai and Perron sequential tests for the existence of structural breaks (own estimations).

Model	Trim	Max Breaks	Break Dates
ARDL 1	0.15	5	2003Q2, 2019Q2
	0.10	5	2003Q2, 2019Q2
ARDL 2	0.15	5	2003Q1
	0.10	5	2003Q1

Note: Break dates significant at 5%.

Indeed, structural breaks are found in the relationship between the variables. Table 3 shows that for the first model, there are two breaks—in 2003Q2 and 2019Q2—while in the second there is only one break, in 2003Q1. The first break is directly linked to the economic crisis experienced by the country during 2002–2003. In the case of 2019, the break may be linked to a slowdown in the GDP growth rate, accompanied by the subsequent fall caused by the COVID pandemic. The next subsection presents the estimation of the model, incorporating the nonlinearity given by the breaks that were found.

3.2. Non-Linear Bivariate Association

Given the presence of significative break dates and using the ARDL models estimated previously, we incorporate the resulting partitions as interactions. Then, a new specification for the economic uncertainty index can be written as:

$$\begin{aligned}
 y_i^1 = & (1 - \mathbb{1}_{eco1} - \mathbb{1}_{eco2}) * \gamma_1 y_{t-1} + (1 - \mathbb{1}_{eco1} - \mathbb{1}_{eco2}) * \gamma_2 y_{t-2} + (1 - \mathbb{1}_{eco1} - \mathbb{1}_{eco2}) * \beta_1 x_t + \\
 & (1 - \mathbb{1}_{eco1} - \mathbb{1}_{eco2}) * \beta_2 x_{t-1} + \mathbb{1}_{eco1} * \gamma_1 y_{t-1} + \mathbb{1}_{eco1} * \gamma_2 y_{t-2} + \mathbb{1}_{eco1} * \beta_1 x_t + \\
 & \mathbb{1}_{eco1} * \beta_2 x_{t-1} + \mathbb{1}_{eco2} * \gamma_1 y_{t-1} + \mathbb{1}_{eco2} * \gamma_2 y_{t-2} + \mathbb{1}_{eco2} * \beta_1 x_t + \mathbb{1}_{eco2} * \beta_2 x_{t-1} + \epsilon_t
 \end{aligned}
 \tag{3}$$

where y and x_t are the same variables as in Equation (2), $\mathbb{1}_{eco1}$ is a dummy variable equal to 1 when the date is between [2003Q3, 2019Q2] and $\mathbb{1}_{eco2}$ is another dummy equal to 1 when the observation is in [2019Q3, 2022Q3]. Similarly, the specification for the export uncertainty index can be written as:

$$y_i^2 = (1 - \mathbb{1}_{exp}) * \gamma_1 y_{t-1} + (1 - \mathbb{1}_{exp}) * \gamma_2 y_{t-2} + (1 - \mathbb{1}_{exp}) * \beta_1 x_t + (1 - \mathbb{1}_{exp}) * \beta_2 x_{t-1} + \mathbb{1}_{exp} * \gamma_1 y_{t-1} + \mathbb{1}_{exp} * \gamma_2 y_{t-2} + \mathbb{1}_{exp} * \beta_1 x_t + \mathbb{1}_{exp} * \beta_2 x_{t-1} + \epsilon_t \tag{4}$$

where $\mathbb{1}_{exp}$ is a dummy variable equal to 1 when the dates are between [2003Q2,2022Q3]. In both new specifications it is interesting to see if, in addition to finding a significant relationship, the association between uncertainty and the GDP cycle is significantly different between the periods established by the breaks.

Table 4 shows the results of the model estimation (3). In this model, unlike what was found in the ARDL estimation, only the autoregressive component is significant. The GDP cycle presents a strong inertial factor in its dynamics. Uncertainty regarding the future state of the economy does not seem to be related to the economic cycle, as the coefficients are not significant. Although they are significant in the second period, the Wald test of the joint significance of the coefficients does not allow for us to reject the hypothesis that the sum is 0. The results of this Wald tests can be seen in Table 5.

Table 4. Results for model with economic uncertainty incorporating structural breaks.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Economic Cycle (-1) * (1 - $\mathbb{1}_{eco1}$ - $\mathbb{1}_{eco2}$)	1.304264	0.273616	4.766768	0.0000 ***
Economic Cycle (-2) * (1 - $\mathbb{1}_{eco1}$ - $\mathbb{1}_{eco2}$)	-0.407803	0.308015	-1.323972	0.1891
Uncertainty Eco * (1 - $\mathbb{1}_{eco1}$ - $\mathbb{1}_{eco2}$)	-0.035868	0.028754	-1.247423	0.2158
Uncertainty Eco (-1) * (1 - $\mathbb{1}_{eco1}$ - $\mathbb{1}_{eco2}$)	0.030622	0.029146	1.050632	0.2965
Economic Cycle (-1) * $\mathbb{1}_{eco1}$	1.012357	0.106402	9.514434	0.0000 ***
Economic Cycle (-2) * $\mathbb{1}_{eco1}$	-0.349082	0.103437	-3.374817	0.0011 ***
Uncertainty Eco * $\mathbb{1}_{eco1}$	-0.027985	0.014033	-1.994257	0.0494 **
Uncertainty Eco (-1) * $\mathbb{1}_{eco1}$	0.029711	0.013889	2.139243	0.0354 **
Economic Cycle (-1) * $\mathbb{1}_{eco2}$	1.623731	0.097551	16.64488	0.0000 ***
Economic Cycle (-2) * $\mathbb{1}_{eco2}$	-0.901717	0.167721	-5.376283	0.0000 ***
Uncertainty Eco * $\mathbb{1}_{eco2}$	-0.024665	0.020418	-1.207973	0.2305
Uncertainty Eco (-1) * $\mathbb{1}_{eco2}$	0.017909	0.016007	1.118846	0.2664
Adjusted R-squared	0.868747	Durbin-Watson Stat		2.002169
Jarque Bera Prob for residuals	0.856963			

Note: Significance levels: 1% *** 5% **. Own estimations.

Table 5. Wald test coefficient restrictions for model with economic uncertainty.

Null Hypothesis	t-Statistic	Probability
Uncertainty Eco (1) = 0	-0.840298	0.4032
Uncertainty Eco (1) = 0	0.912886	0.3639
Uncertainty Eco (2) = 0	-1.199455	0.2338

In turn, Table 6, which shows the results regarding uncertainty in international trade, shows interesting results. There is a negative and significant association between this uncertainty index and the GDP cycle; this is only found for the first period. This seems to evidence a possible reduction in the negative effect of uncoordinated expectations on the business cycle. After the economic crisis of 2002–2003, uncertainty regarding international trade does not seem to be significantly associated with the GDP cycle. As in the previous model, the Wald Tests of joint significance are presented in Table 7.

If we look at the Figures 1 and A3, corresponding to the uncertainty index and the GDP cycle, it is possible to observe two marked trends after the fall: (i) an important reduction in the level of uncertainty, shown by an increase in the share of firms that respond that the expectation regarding exports is that it will remain “the same”; and (ii) a reduction in the variability of the economic cycle.

Table 6. Results for model with Index 2 incorporating structural breaks (own estimations).

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Economic Cycle (−1) * (1− $\mathbb{1}_{exp}$)	1.119265	0.142532	7.852730	0.0000 ***
Economic Cycle (−2) * (1− $\mathbb{1}_{exp}$)	−0.143645	0.122415	−1.173425	0.2438
Uncertainty Expo * (1− $\mathbb{1}_{exp}$)	0.046752	0.025535	1.830874	0.0705 *
Uncertainty Expo (−1) * (1− $\mathbb{1}_{exp}$)	−0.054458	0.024681	−2.206475	0.0300 **
Economic Cycle (−1) * $\mathbb{1}_{exp}$	1.192547	0.109440	10.89678	0.0000 ***
Economic Cycle (−2) * $\mathbb{1}_{exp}$	−0.450539	0.087777	−5.132737	0.0000 ***
Uncertainty Expo * $\mathbb{1}_{exp}$	−0.024633	0.015735	−1.565455	0.1211
Uncertainty Expo (−1) * $\mathbb{1}_{exp}$	0.024789	0.015084	1.643424	0.1039
Adjusted R-squared	0.881744	Durbin–Watson Stat		1.877005
Jarque Bera Prob for residuals	0.516868			

Note: Significance levels: 1% *** 5% ** 10% *. Own estimations.

Table 7. Wald test coefficient restrictions for model with export uncertainty.

Null Hypothesis	t-Statistic	Probability
Uncertainty Expo (1) = 0	−3.468117	0.0008 ***
Uncertainty Expo (2) = 0	0.108110	0.9142
Uncertainty Expo (1) = Uncertainty Expo (2)	−2.980876	0.0037 ***

Significance levels: 1% ***.

Among the factors that may be behind these results, it is possible to consider that in a small and open country like Uruguay, the short-term dynamics of GDP (which is what we are observing more clearly when using the cycle) are more relevant to foreign market conditions than to the domestic economy itself. This is mainly because uncertainty regarding possible changes in international markets can be transferred more quickly to GDP.

This significant relationship is relevant for understanding how uncoordinated expectations (in this case, from the industrial sector) affect GDP dynamics. However, it is possible that other factors are also relevant both to the dynamics of the economy and as a channel through which uncertainty is transferred to GDP.

Therefore, in the next section, we re-analyze the relationship between the series, incorporating a non-tradable price index in the estimations. By including this series in the analysis, we seek to control for a possible price effect of goods and services not associated with the foreign market.

3.3. Controlling for Variability in Domestic Prices

Following the steps of the previous subsections, first, an ARDL model is estimated to establish the relevant components in the linear relationship (the non-tradable price index was included in its seasonal difference, i.e., the year-on-year rate of non-tradable inflation).

As can be seen in Tables 8 and 9 the relevance of the autoregressive component is again evident; all three lags are significant for both models. Uncertainty maintains its significance, even with the inclusion of the non-tradable price index. Moreover, unlike the bivariate ARDL of the first subsection, the export uncertainty index now has a negative and significant association at 10%. Although, in both cases, the relationship between expectations misalignment and the business cycle seems to be negative, the difference found between the models is that, in the case of the economic uncertainty index, the first lag in the variable also appears to be relevant (something not found for the other model). In contrast, the NTP index seems to be positively associated with the cycle.

Table 8. ARDL model with economic uncertainty and non-tradable price index as regressors (own estimations).

Variable	Coefficient	t-Statistic	Prob
Economic Cycle (−1)	1.328141	14.11823	0.0000 ***
Economic Cycle (−2)	−0.625139	−5.160872	0.0000 ***
Economic Cycle (−3)	0.164179	1.745806	0.0843 *
Uncertainty Eco	−0.030652	−2.528639	0.0132 **
Uncertainty Eco (−1)	0.022047	1.778425	0.0788 *
NTP Index	−0.072540	−1.129097	0.2619
NTP Index (−1)	0.116385	1.785335	0.0777 *

Adjusted R-squared: 0.865329
 Durbin–Watson Stat: 2.023571
 Jarque Bera Prob for residuals: 0.499441

Note: Significance levels: 1% *** 5% ** 10% *.

Table 9. ARDL model with export uncertainty and non-tradable price index as regressors (own estimations).

Variable	Coefficient	t-Statistic	Prob
Economic Cycle (−1)	1.318875	16.29477	0.0000 ***
Economic Cycle (−2)	−0.628313	−5.455212	0.0000 ***
Economic Cycle (−3)	0.205955	2.028425	0.0455 **
Uncertainty Expo	−0.012640	−1.723270	0.0883 *
NTP Index	−0.024012	−0.351361	0.7261
NTP Index (−1)	0.109869	1.955052	0.0537 *

Adjusted R-squared: 0.877273
 Durbin–Watson Stat 2.032154
 Jarque Bera Prob for residuals: 0.535540

Note: Significance levels: 1% *** 5% ** 10% *.

Subsequently, we analyzed the possible existence of structural breaks. Given the number of observations available (also reduced by the inclusion of the seasonal difference of the NTP Index), we established in the break tests that only a maximum of two breaks can be found. If this restriction is not established, a third structural break is found in 2012Q3, and in 2007Q2 for the economic uncertainty model and export uncertainty model, respectively. As can be seen in Table 10, the breaks that were found are extremely similar to those found in Section 3.1: one linked to the crisis and the other prior to the pandemic, which was also linked to a period of slowdown in the Uruguayan economy. The main difference is that, for the new model specification with export uncertainty, a second structural break is found at 2019Q2.

Table 10. Bai and Perron sequential tests for the existence of structural breaks (own estimations).

Model	Trim	Max Breaks	Break Dates
ARDL 1	0.15	2	2003Q2, 2018Q2
ARDL 2	0.15	2	2003Q2, 2019Q2

Note: break dates significant at 5%.

Finally, both models are estimated considering the breaks found. First, several points emerge from the specification with the economic uncertainty index. For simplicity, only the uncertainty results are shown. The extended estimations of the models can be found in the Appendix (Tables A4 and A5). As Table 11 shows, uncertainty is negatively associated with the GDP cycle in the first period, a result that was not found in the estimation of the previous subsection. On the other hand, the relationship is not significant for the second

period and is positive and significant at 10% for the third period. Wald test for significant differences between the periods results in the rejection of the hypothesis that there are no different effects. In other words, the nonlinearity of the association is confirmed.

With respect to the positive effect found in the third period, some considerations may be made. By constructing the indicator where uncertainty refers to the uncoordinated expectations of the firms, it is possible to increase uncertainty by improving expectations. As an example, if we start from a scenario where the responses are distributed as follows: 70% “the same”, 20% “worse” and 10% “better”, and we move to a new scenario where 50% “the same”, 20% “worse” and 30% “better”, the expectations improve while the misalignment among firms increases.

Second, in line with the bivariate model, the specification with the export uncertainty index shows a negative and significant association in the first period, which dissipates in the following periods (Table 12).

In summary, the effects of both models are consistent with the idea that the Uruguayan economy managed to reduce the effects of uncertainty after the economic crisis of 2002–2003. As mentioned, this is directly related to the combination of two marked trends, an improvement in the coordination of firms’ expectations and a reduction in the volatility of the cycle.

Table 11. Results for model with economic uncertainty.

	Coefficient	Std. Error	t-Statistic	Prob.
<i>First period (1999Q1-2003Q2)</i>				
Uncertainty Eco	−0.078016	0.027857	−2.800543	0.0065 ***
Uncertainty Eco (−1)	0.020252	0.032630	0.620650	0.5367
Wald test: Uncertainty Eco (1) = 0	−0.057764	0.011436	−5.051039	0.0000 ***
<i>Second period (2003Q3-2018Q2)</i>				
Uncertainty Eco	−0.031572	0.014348	−2.200449	0.0309 **
Uncertainty Eco (−1)	0.027546	0.021599	1.275347	0.2062
Wald test: Uncertainty Eco (2) = 0	−0.004026	0.014667	−0.274513	0.7845
<i>Third period (2018Q3-2022Q3)</i>				
Uncertainty Eco	0.037021	0.027618	1.340480	0.1842
Uncertainty Eco (−1)	0.062965	0.035732	1.762144	0.0822 *
Wald test: Uncertainty Eco (3) = 0	0.099986	0.058531	1.708271	0.0918 *
Wald tests				
Null Hypothesis			t-Statistic	Prob.
Uncertainty Eco (1) = Uncertainty Eco (2)			−2.859274	0.0055 ***
Uncertainty Eco (1) = Uncertainty Eco (3)			−2.645159	0.0100 ***
Uncertainty Eco (2) = Uncertainty Eco (3)			−1.729928	0.0878 *

Note: Significance levels: 1% *** 5% ** 10% *. Own estimations.

Table 12. Results for Model with Export Uncertainty.

	Coefficient	Std. Error	t-Statistic	Prob.
<i>First period (1999Q1-2003Q2)</i>				
Uncertainty Expo	−0.029535	0.009953	−2.967433	0.0040 ***
<i>Second period (2003Q3-2019Q2)</i>				
Uncertainty Expo	−0.008266	0.011044	−0.748455	0.4564
<i>Third period (2019Q3-2022Q3)</i>				
Uncertainty Expo	0.088481	0.063845	1.385867	0.1697
Wald tests				
Null Hypothesis			t-Statistic	Prob.
Uncertainty Expo (1) = Uncertainty Expo (2)			−1.397758	0.1661
Uncertainty Expo (1) = Uncertainty Expo (3)			−1.826405	0.0716 *
Uncertainty Expo (2) = Uncertainty Expo (3)			−1.492682	0.1396

Note: Significance levels: 1% *** 10% *. Own estimations.

4. Main Remarks

The relationship between uncertainty and business cycles has been extensively studied in the economic literature. The main idea is that, during periods of high uncertainty, businesses and consumers become more cautious in their spending, which can lead to a decrease in economic activity and a recession. On the other hand, periods of low uncertainty can lead to increased spending and economic growth.

Overall, the literature suggests that the impact of uncertainty on business cycles can vary across economies and may depend on factors such as the level of financial development and the structure of the economy. However, there is a general consensus that higher uncertainty can lead to a decrease in economic activity and lower productivity.

This paper considers an uncertainty index for Uruguay (a small South American country that is highly exposed to international and regional shocks), based on economic trend surveys. We follow the proposal of [2], using the industrial monthly survey that has been led since 1998 by the Uruguayan Chamber of Industry. Similar to recent research that analyzed the way economic uncertainty affects the economy in the short term ([13,21,22], among others), we explore the relationship between the Uruguayan GDP cycle and uncertainty indices by applying linear and nonlinear models.

The estimated ARDL linear models showed negative but weak relationships between the uncertainty indices and the GDP cycle. The tests for the existence of structural breaks in these relationships show a significant break in the year 2003 for both indices, and another in 2019 for one of them. The Wald tests performed on the nonlinear models only confirm the structural break in the early 2000s in the model with the index based on export market expectations. Before 2003, the effect of uncertainty over the Uruguayan economy is significant and negative, as [15] found for the Brazilian economy. After 2003, this negative effect decreases. This result holds even when controlling for the variability of non-tradable domestic prices. This result is probably associated with the improvement in institutional factors and in the soundness of the Uruguayan financial system. The authors of [14] pointed out that the greatest effect of uncertainty in emerging economies compared to developed economies is the depth of financial markets in the latter.

Two implications can be derived from these results. First, there is evidence of relevant changes that made the Uruguayan economy less vulnerable as of 2003. Second, the monitoring of the evolution of agents' expectations about the future of the export market over the macroeconomic cycle is important in a small and open economy such as the Uruguayan economy.

Author Contributions: All authors contributed equally to the different stages of this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by: Comisión Sectorial de Investigación Científica (CSIC)—Proyectos de I+D 2020, Universidad de la República, Uruguay.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found at <https://www.ciu.com.uy/monitoreo-industrial/expectativas-empresariales-industriales/> and <https://www.bcu.gub.uy/Estadisticas-e-Indicadores/Paginas/Series-Estadisticas-del-PIB-por-industrias.aspx>. Data access date: 15 January 2023.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

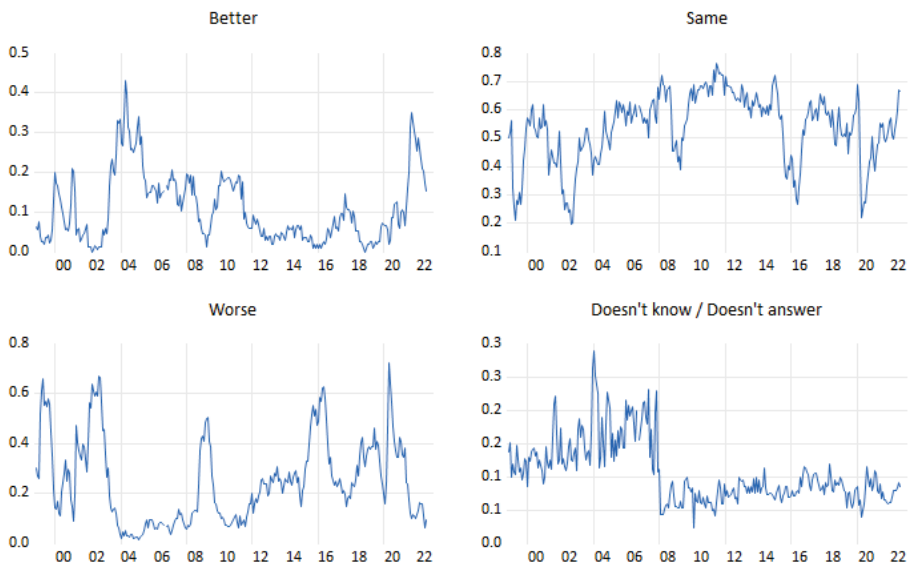


Figure A1. Change in share of responses to the question: What are your firm expectations about the evolution of the country's economy in the next 6 months? (Own construction based on CIU data).

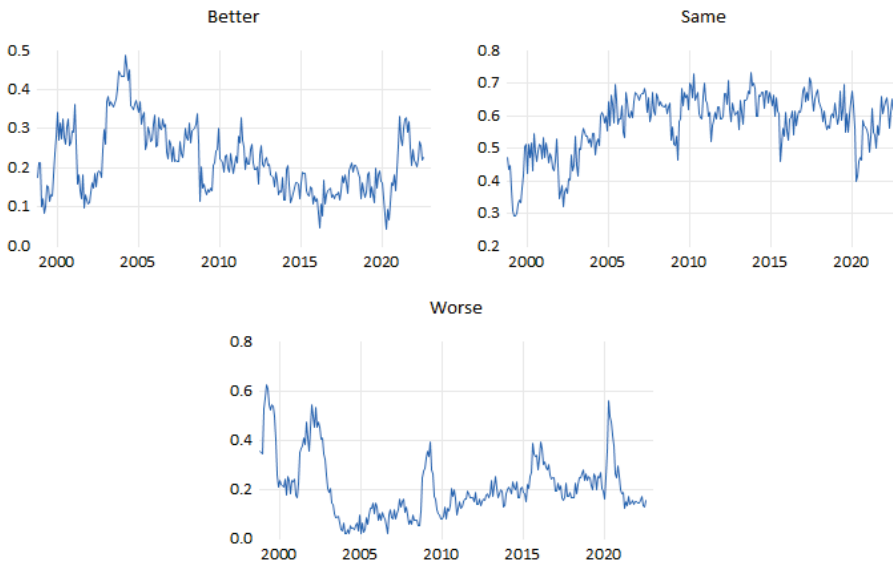


Figure A2. Change in share of responses to the question: If your firm exports, what are your expectations about your external sales in units in the next 6 months compared to last year? (Own construction based on CIU data).

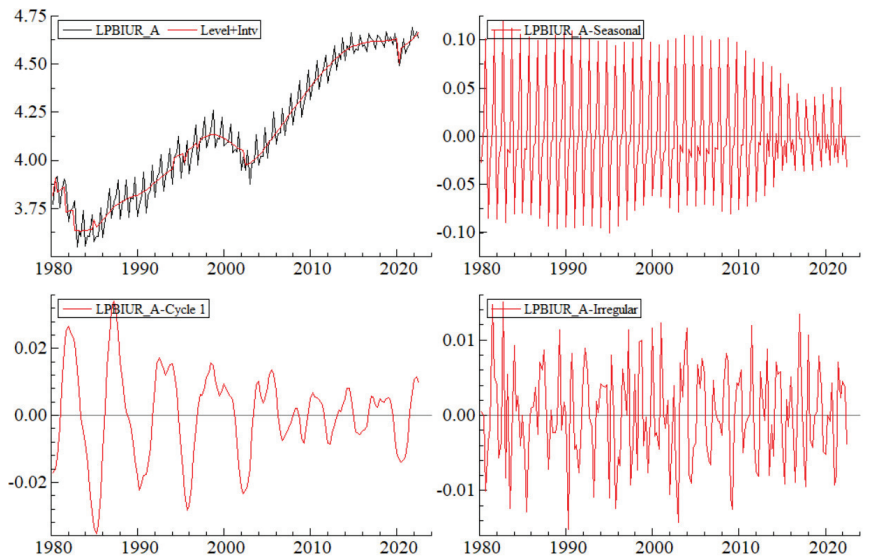


Figure A3. Estimation of unobserved components of the Uruguayan GDP (Own construction).

Table A1. Results of unobserved components' estimation.

Model Estimated	
Y = Level + Seasonal + Irregular + Cycle + Interventions	
Standard deviations of component residuals	
Level	0
Seasonal	0.00191589665
Irregular	0.00539346827
Cycle	0.01207414593
Model Diagnostic Statistics	
Normality (Bowman – Shenton)	2.0214
T	170
Rs ²	0.72238
Cycle other parameters	
Standard Deviation	0.0331662479
Period in Years	7.13498
Damping Factor	0.93116
Frequency	0.22015

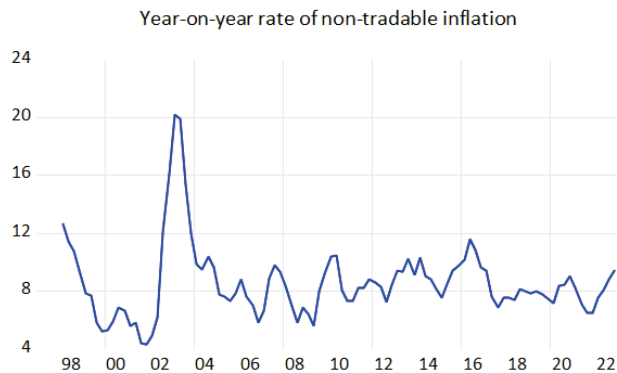


Figure A4. Year-on-year rate of non-tradable price index (NTP).

Table A2. Augmented Dickey–Fuller test for Index 1 and 2. MacKinnon one-sided *p*-values.

Null Hypothesis	Included in Test Equation	T-Statistic	Prob
Index 1 has a unit root (0)	Constant	−3.382373	0.0140 **
Index 2 has a unit root (0)	Constant	−3.745660	0.0048 ***

Significance levels: 1% *** 5% **. Lag length selected automatically based on SIC criterion, number of lags between brackets.

Table A3. Evolution of mean and S.D. for the economic cycle and uncertainty indexes (own construction).

Five-Year Period Mean	Cycle	Index 1	Index 2
[1998Q4–2003Q3]	−0.007172	0.498835	0.711100
[2003Q4–2008Q3]	0.005583	0.387073	0.527276
[2008Q4–2013Q3]	−0.000898	0.407762	0.559700
[2013Q4–2018Q3]	0.005037	0.459868	0.550142
[2018Q4–2022Q3]	−0.008590	0.525412	0.597186

Table A3. Cont.

Five-Year Period Standard Deviation	Cycle	Index 1	Index 2
[1998Q4-2003Q3]	0.038733	0.032445	0.063504
[2003Q4-2008Q3]	0.011634	0.056597	0.032582
[2008Q4-2013Q3]	0.009007	0.079864	0.057885
[2013Q4-2018Q3]	0.008959	0.047416	0.058000
[2018Q4-2022Q3]	0.020782	0.078845	0.050813

Table A4. Results for model with economic uncertainty (extended).

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Economic Cycle (-1) * $(1 - \mathbb{1}_{eco1} - \mathbb{1}_{eco2})$	0.971293	0.143547	6.766363	0.0000
Economic Cycle (-2) * $(1 - \mathbb{1}_{eco1} - \mathbb{1}_{eco2})$	-0.460914	0.227425	-2.026664	0.0463
Economic Cycle (-3) * $(1 - \mathbb{1}_{eco1} - \mathbb{1}_{eco2})$	0.556891	0.166095	3.352847	0.0013
Uncertainty Eco * $(1 - \mathbb{1}_{eco1} - \mathbb{1}_{eco2})$	-0.078016	0.027857	-2.800543	0.0065
Uncertainty Eco (-1) * $(1 - \mathbb{1}_{eco1} - \mathbb{1}_{eco2})$	0.020252	0.032630	0.620650	0.5367
NTP Index * $(1 - \mathbb{1}_{eco1} - \mathbb{1}_{eco2})$	0.015281	0.110945	0.137738	0.8908
NTP Index (-1) * $(1 - \mathbb{1}_{eco1} - \mathbb{1}_{eco2})$	0.260635	0.116805	2.231379	0.0287
Economic Cycle (-1) * $\mathbb{1}_{eco1}$	1.086471	0.131641	8.253275	0.0000
Economic Cycle (-2) * $\mathbb{1}_{eco1}$	-0.558807	0.183351	-3.047751	0.0032
Economic Cycle (-3) * $\mathbb{1}_{eco1}$	0.177890	0.125142	1.421504	0.1594
Uncertainty Eco * $\mathbb{1}_{eco1}$	-0.031572	0.014348	-2.200449	0.0309
Uncertainty Eco (-1) * $\mathbb{1}_{eco1}$	0.027546	0.021599	1.275347	0.2062
NTP Index * $\mathbb{1}_{eco1}$	0.023900	0.102736	0.232640	0.8167
NTP Index (-1) * $\mathbb{1}_{eco1}$	0.006676	0.082092	0.081323	0.9354
Economic Cycle (-1) * $\mathbb{1}_{eco2}$	1.238076	0.260182	4.758494	0.0000
Economic Cycle (-2) * $\mathbb{1}_{eco2}$	-0.554057	0.257644	-2.150478	0.0348
Economic Cycle (-3) * $\mathbb{1}_{eco2}$	0.402342	0.265752	1.513978	0.1343
Uncertainty Eco * $\mathbb{1}_{eco2}$	0.037021	0.027618	1.340480	0.1842
Uncertainty Eco (-1) * $\mathbb{1}_{eco2}$	0.062965	0.035732	1.762144	0.0822
NTP Index * $\mathbb{1}_{eco2}$	-0.150417	0.331084	-0.454317	0.6509
NTP Index (-1) * $\mathbb{1}_{eco2}$	-0.517170	0.174352	-2.966238	0.0041

Table A5. Results for model with export uncertainty (extended).

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Economic Cycle (-1) * $(1 - \mathbb{1}_{exp1} - \mathbb{1}_{exp2})$	1.132934	0.113630	9.970396	0.0000
Economic Cycle (-2) * $(1 - \mathbb{1}_{exp1} - \mathbb{1}_{exp2})$	-0.564487	0.200161	-2.820169	0.0061
Economic Cycle (-3) * $(1 - \mathbb{1}_{exp1} - \mathbb{1}_{exp2})$	0.460023	0.176247	2.610102	0.0108
Uncertainty Expo * $(1 - \mathbb{1}_{exp1} - \mathbb{1}_{exp2})$	-0.029535	0.009953	-2.967433	0.0040
NTP Index * $(1 - \mathbb{1}_{exp1} - \mathbb{1}_{exp2})$	0.056315	0.101043	0.557341	0.5789
NTP Index (-1) * $(1 - \mathbb{1}_{exp1} - \mathbb{1}_{exp2})$	0.137541	0.115078	1.195204	0.2356
Economic Cycle (-1) * $\mathbb{1}_{exp1}$	1.083669	0.119190	9.091904	0.0000
Economic Cycle (-2) * $\mathbb{1}_{exp1}$	-0.554825	0.158812	-3.493603	0.0008
Economic Cycle (-3) * $\mathbb{1}_{exp1}$	0.181646	0.118082	1.538306	0.1280
Uncertainty Expo * $\mathbb{1}_{exp1}$	-0.008266	0.011044	-0.748455	0.4564
NTP Index * $\mathbb{1}_{exp1}$	0.064436	0.093712	0.687595	0.4937
NTP Index (-1) * $\mathbb{1}_{exp1}$	9.79×10^{-5}	0.079311	0.001234	0.9990
Economic Cycle (-1) * $\mathbb{1}_{exp2}$	1.212852	0.316458	3.832585	0.0003
Economic Cycle (-2) * $\mathbb{1}_{exp2}$	-0.084595	0.552603	-0.153085	0.8787
Economic Cycle (-3) * $\mathbb{1}_{exp2}$	-0.302975	0.207107	-1.462895	0.1475
Uncertainty Expo * $\mathbb{1}_{exp2}$	0.088481	0.063845	1.385867	0.1697
NTP Index * $\mathbb{1}_{exp2}$	0.018109	0.258704	0.069998	0.9444
NTP Index (-1) * $\mathbb{1}_{exp2}$	-0.759109	0.302274	-2.511328	0.0141

References

1. Bloom, N. The impact of uncertainty shocks. *Econometrica* **2009**, *77*, 623–685.
2. Claveria, O.; Monte, E.; Torra, S. Economic uncertainty: A geometric indicator of discrepancy among experts' expectations. *Soc. Indic. Res.* **2019**, *143*, 95–114. [CrossRef]
3. Claveria, O. Uncertainty indicators based on expectations of business and consumer surveys. *Empirica* **2021**, *48*, 483–505. [CrossRef]
4. Jackson, L.E.; Kliesen, K.L.; Owyang, M.T. The nonlinear effects of uncertainty shocks. *Stud. Nonlinear Dyn. Econom.* **2019**, *24*, 20190024. [CrossRef]
5. Christou, C.; Naraidoo, R.; Gupta, R. Conventional and unconventional monetary policy reaction to uncertainty in advanced economies: Evidence from quantile regressions. *Stud. Nonlinear Dyn. Econom.* **2019**, *24*, 20180056. [CrossRef]
6. Baker, S.R.; Bloom, N.; Davis, S.J. Measuring economic policy uncertainty. *Q. J. Econ.* **2016**, *131*, 1593–1636. [CrossRef]
7. Aromi, J.D. Linking words in economic discourse: Implications for macroeconomic forecasts. *Int. J. Forecast.* **2020**, *36*, 1517–1530. [CrossRef]
8. Algaba, A.; Ardia, D.; Bluteau, K.; Borms, S.; Boudt, K. Econometrics meets sentiment: An overview of methodology and applications. *J. Econ. Surv.* **2020**, *34*, 512–547. [CrossRef]
9. Ferderer, J.P. The impact of uncertainty on aggregate investment spending: An empirical analysis. *J. Money Credit. Bank.* **1993**, *25*, 30–48. [CrossRef]
10. Bachmann, R.; Elstner, S.; Sims, E.R. Uncertainty and economic activity: Evidence from business survey data. *Am. Econ. J. Macroecon.* **2013**, *5*, 217–249. [CrossRef]
11. van Aarle, B.; Moons, C. Sentiment and uncertainty fluctuations and their effects on the Euro Area business cycle. *J. Bus. Cycle Res.* **2017**, *13*, 225–251. [CrossRef]
12. Claveria, O.; Pons, E.; Ramos, R. Business and consumer expectations and macroeconomic forecasts. *Int. J. Forecast.* **2007**, *23*, 47–69. [CrossRef]
13. Sorić, P. Consumer confidence as a GDP determinant in New EU Member States: A view from a time-varying perspective. *Empirica* **2018**, *45*, 261–282. [CrossRef]
14. Carrière-Swallow, Y.; Céspedes, L.F. The impact of uncertainty shocks in emerging economies. *J. Int. Econ.* **2013**, *90*, 316–325. [CrossRef]
15. Ferreira, P.C.; Vieira, R.M.B.; da Silva, F.B.; de Oliveira, I.C. Measuring Brazilian economic uncertainty. *J. Bus. Cycle Res.* **2019**, *15*, 25–40. [CrossRef]
16. Claveria, O. Qualitative survey data on expectations. Is there an alternative to the balance statistic? In *Economic Forecasting*; Nova Science Publishers, Inc.: Barcelona, Spain, 2010; pp. 181–190. [CrossRef]
17. Claveria, O.; Sorić, P. Labour market uncertainty after the irruption of COVID-19. *Empir. Econ.* **2022**, *64*, 1897–1945. [CrossRef]
18. Lanzilotta, B.; Merlo, G.; Mordecki, G.; Umpierrez, V. Understanding Uncertainty Shocks in Uruguay Through VAR Modeling. *J. Bus. Cycle Res.* **2023**, 1–21. [CrossRef]
19. Brida, J.G.; Lanzilotta, B.; Rosich, L.I. On the dynamics of expectations, uncertainty and economic growth: An empirical analysis for the case of Uruguay. *J. Emerg. Mark.* **2022**. [CrossRef]
20. Azqueta-Gavaldón, A. Developing news-based economic policy uncertainty index with unsupervised machine learning. *Econ. Lett.* **2017**, *158*, 47–50. [CrossRef]
21. Sorić, P.; Claveria, Ó. Employment uncertainty a year after the irruption of the COVID-19 pandemic. In *AQR–Working Papers 2021 AQR21/04*; University of Barcelona: Barcelona, Spain, 2021.
22. Apaitan, T.; Luangaram, P.; Manopimoke, P. Uncertainty in an emerging market economy: Evidence from Thailand. *Empir. Econ.* **2022**, *62*, 933–989. [CrossRef]
23. CIU. *Encuesta Mensual Industrial. Revisión Metodológica*; Technical Report; Cámara de Industrias del Uruguay: Montevideo, Uruguay, 2017.
24. Koopman, S.; Harvey, A.; Doornik, J.; Shepard, N. *Structural Time Series Analyser, Modeller and Predictor: STAMP 8.2*; International Series of Monographs on Physics; Timberlake Consultants Ltd.: London, UK, 2009.
25. Bergara, M.; Dominioni, D.; Licandro, J.A. Un modelo para comprender la “enfermedad uruguaya”. *Rev. Econ.* **1995**, *2*, 39–76.
26. Dickey, D.; Fuller, W. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *JASA J. Am. Stat. Assoc.* **1979**, *74*, 427–431. [CrossRef]
27. MacKinnon, J.G. Numerical distribution functions for unit root and cointegration tests. *J. Appl. Econom.* **1996**, *11*, 601–618. [CrossRef]
28. Bai, J.; Perron, P. Computation and analysis of multiple structural change models. *J. Appl. Econom.* **2003**, *18*, 1–22. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Data-Driven Spatio-Temporal Modelling and Optimal Sensor Placement for a Digital Twin Set-Up †

Mandar Tabib ^{1,*}, Kristoffer Skare ², Endre Bruaset ² and Adil Rasheed ^{1,2,*}

¹ Mathematics and Cybernetics, SINTEF Digital, 7037 Trondheim, Norway

² Department of Engineering Cybernetics, Norwegian University of Science and Technology, 7034 Trondheim, Norway

* Correspondence: mandar.tabib@sintef.no (M.T.); adil.rasheed@ntnu.no (A.R.)

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: A computationally efficient predictive digital twin (DT) of a small-scale greenhouse needs an accurate and faster modelling of key variables such as the temperature field and flow field within the greenhouse. This involves : (a) optimally placing sensors in the experimental set-up and (b) developing fast predictive models. In this work, for a greenhouse set-up, the former requirement fulfilled first by identifying the optimal sensor locations for temperature measurements using the QR column pivoting on a tailored basis. Here, the tailored basis is the low-dimensional representation of hi-fidelity computational fluid dynamics (CFD) flow data, and these tailored basis are obtained using proper orthogonal decomposition (POD). To validate the method, the full temperature field inside the greenhouse is then reconstructed for an unseen parameter (inflow condition) using the temperature values from a few synthetic sensor locations in the CFD model. To reconstruct the flow-fields using a faster predictive model than the hi-fidelity CFD model, a long-short term memory (LSTM) method based on a reduced-order model (ROM) is used. The LSTM learns the temporal dynamics of coefficients associated with the POD-generated velocity basis modes. The LSTM-POD ROM model is used to predict the temporal evolution of velocity fields for our DT case, and the predictions are qualitatively similar to those obtained from hi-fidelity numerical models. Thus, the two data-driven tools have shown potential in enabling the forecasting and monitoring of key variables in a digital twin of a greenhouse. In future work, there is scope for improvements in the reconstruction accuracy by involving deep-learning-based corrective source term approaches.

Keywords: dimensionality reduction; forecasting; LSTM; POD; QR pivoting; digital twin

Citation: Tabib, M.; Skare, K.; Bruaset, E.; Rasheed, A. Data-Driven Spatio-Temporal Modelling and Optimal Sensor Placement for a Digital Twin Set-Up. *Eng. Proc.* **2023**, *39*, 98. <https://doi.org/10.3390/engproc2023039098>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 16 August 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

A digital twin (DT) [1] is defined as a virtual representation of a physical asset enabled through data and simulators for real-time predictions, optimization, monitoring, controlling, and improved decision-making. For efficient real-time predictive digital twins, the use of computationally intensive hi-fidelity numerical solvers and a large number of sensors (for control) needs to be avoided, as well as time-series prediction techniques and sparse sensor placement locations. In this work, an experimental greenhouse set-up is constructed (as seen in Figure 1) as the physical asset of DT. The physical asset (Figure 2) has sensors to measure the varying internal conditions (temperature, flow rate, humidity) inside a fully controllable environment, but requires optimal sensor placements to enable the reconstruction of a full field using only the measured sensor values.

Determining sensor placements for large data-sets involves methods such as a compressed sensing algorithm [2], which assumes that the original signal is sparse on a universal basis. It then uses the L1-norm-based optimization approach to find the sparsest solution. Once the sparse signal is recovered using compressed sensing (CS), the sensor locations can

be identified by examining the non-zero entries in the solution vector. This method does not require training data to find the basis functions as it uses a universal basis. However, this approach is not suitable for high-dimensional physical systems with a known structure, and for such systems, a method based on the data-driven QR pivoting of tailored basis [3] is more suitable. This data-driven approach uses training data to find the basis specific to the known system and this results in a lower number of optimum sensor placements for a high-dimensional system than that obtained using the CS-based method. Hence, the QR-pivoting-based sensor placement method is demonstrated to have applications in a greenhouse digital twin. For DT, there is also a need to develop faster data-driven reduced-order models [4–6] for predicting the temporal flow state of the physical asset. In this work, we employ data-driven techniques involving long-short term memory, proper orthogonal-decomposition-based decomposition and QR-pivoting to enable modelling of the temporal dynamics of key variables for a digital twin of a small-scale greenhouse. The methodology and results are discussed in the next sections, followed by the conclusion.

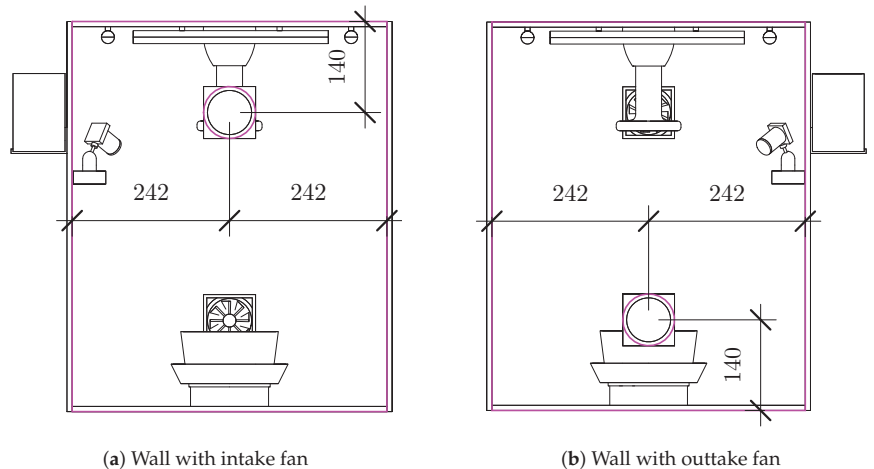


Figure 1. Schematic showing the layout of the greenhouse side walls with dimensions in millimeters.

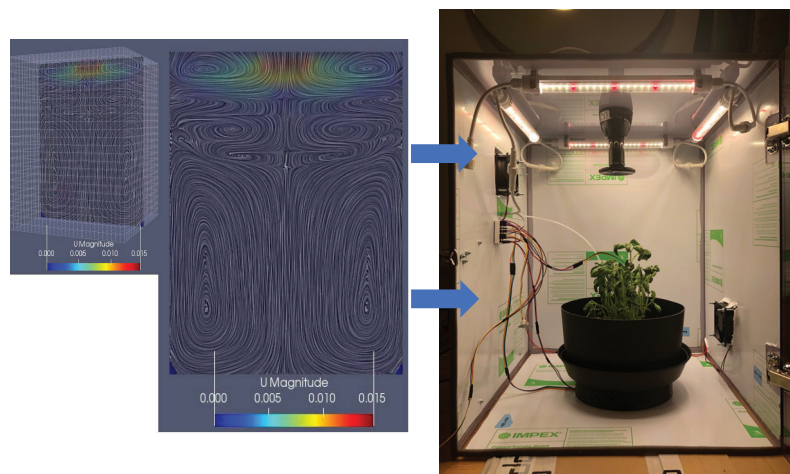


Figure 2. Bidirectionally coupled DT and Asset.

2. Methodology

The greenhouse set-up (physical set-up) and its hi-fidelity computational fluid dynamics (CFD) set-up are constructed first (Figures 1 and 2). The training data-set of temperature and velocity fields obtained from CFD simulations of the greenhouse set-up (as described in Section 2.1) are then subjected to proper orthogonal decomposition (POD). The POD decomposes the data into the dominant basis functions (which serve as a low-dimensional representation) and accompanying time-dependent coefficients (as discussed in Section 2.2). These dominant basis modes are then used to obtain optimal sparse locations for temperature measurements and to develop a reduced-order model for the flow field. The details of the data-set and methodology are provided below.

2.1. Training Data-Set and Greenhouse CFD Simulations

The training data (i.e., the full temperature field and velocity field across the greenhouse) needed by the data-driven techniques are generated using a hi-fidelity computational fluid dynamics (CFD) simulation of the greenhouse set-up. Figure 2 shows the experimental set-up and CFD simulation of the greenhouse to enable a digital twin (DT). In this work, a hi-fidelity OpenFOAM CFD solver is used for the simulation, and this solves the Navier stokes equation (i.e., the continuity and momentum conservation equations) along with the thermal equation, while the turbulence is modelled using an RANS *k-epsilon* model. The greenhouse has a fan to control inlet airflow speed, and a heater on the top of roof for temperature control. These are considered boundary conditions for the CFD solver. For CFD simulations, the parameters that are changed are: (a) Inlet air flow speed due to the fan operation in the greenhouse. For each simulation case, this is varied as follows: 0 m/s (i.e., no in-flow and the natural convective flow occurs due to the heat flux from the heater at the top), 1 m/s, 2 m/s and 3 m/s; (b) The heat flux at the top is varied to match the expected heater output from the top. The generated simulation data-set is divided into training and validation data-sets to develop and test the data-driven techniques, while the training data-set comprises simulation cases with 0 m/s, 1 m/s and 3 m/s inlet flow conditions. The simulation case of 2 m/s is used to validate/test data-driven reconstructions. The fan speed and heat flux are varied in the real-experimental set-up to enable an optimal temperature for plant growth in the greenhouse. The grid size used for numerical CFD simulation was selected after a proper grid-independence test and the total number of grids was $n = n_x n_y n_z = 14,165$. The discretization schemes employed are as follows: linearUpwind for convection term and implicit second-order backward scheme for temporal discretization. The time-step for simulation is selected to ensure the Courant number is less than 1 and total time of simulations in each simulation could develop the flow. The snapshots of temperature and velocity from each of the simulation cases are saved at a time-step of about 0.5 s. These snapshots are then subjected to proper orthogonal decomposition (POD) to obtain the dominant basis functions (which serve as a low-dimensional representation) and accompanying time-dependent coefficients (as discussed in Section 2.2). The dominant basis functions are then used to obtain optimal sparse locations for temperature measurements and develop a reduced-order model. The methodology of the LSTM-POD reduced-order model is covered briefly in Section 3.2, and in more detail in [6]. The next section describes the POD methodology and sensor placement methodology.

2.2. Proper Orthogonal Decomposition

The chosen snapshots of the temperature simulations from the training database (i.e., simulations with an inflow of 0 m/s, 1 m/s and 3 m/s, respectively) were flattened in their spatial dimensions to form a matrix $\mathbf{T} \in \mathbb{R}^{n \times m}$, where $n = n_x n_y n_z$. The matrix was then shifted by the mean of its columns (μ_T) and scaled by the standard deviation of its columns (σ_T) to obtain matrix \mathbf{X} .

$$\mathbf{X} = \frac{1}{\sigma_T} (\mathbf{T} - \mu_T) \quad (1)$$

The matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ was decomposed as shown in Equation (2) by using singular value decomposition to obtain the dominant basis functions (i.e the POD modes),

$$\mathbf{X} = \mathbf{\Psi} \mathbf{\Sigma} \mathbf{V}^T \tag{2}$$

where $\mathbf{\Psi} \in \mathbb{R}^{n \times n}$ and $\mathbf{V} \in \mathbb{R}^{m \times m}$ are orthonormal matrices, while $\mathbf{\Sigma} \in \mathbb{R}^{n \times m}$ is only non-zero on its diagonal. The values in the diagonal of $\mathbf{\Sigma}$ are called the singular values of \mathbf{X} and are ordered in descending order. The columns of $\mathbf{\Psi}$ are called the POD modes of \mathbf{X} . By truncating the matrices $\mathbf{\Psi}$, $\mathbf{\Sigma}$ and \mathbf{V} to only use the first r singular values, it can be used for dimensionality reductions. The truncated POD modes serves as low-dimensional representation of hi-dimensional data. The columns of $\mathbf{\Psi}$ and singular values are ordered by how important they are for the reconstruction of \mathbf{X} . The truncated singular value decomposition is shown in Equation (3). $\mathbf{\Psi}_r \in \mathbb{R}^{n \times r}$ and $\mathbf{V}_r \in \mathbb{R}^{m \times r}$ are the first r columns of $\mathbf{\Psi}$ and \mathbf{V} , respectively, and $\mathbf{\Sigma}_r \in \mathbb{R}^{r \times r}$ is a diagonal matrix with the first r singular values on the diagonal. Using the first r POD-modes will lead to some error, represented by the matrix \mathbf{E}_r . For many matrices consisting of structured data, the error will be small for a relatively limited choice of r if the first r modes captures most of the variance in the data.

$$\mathbf{X} = \mathbf{\Psi}_r \mathbf{\Sigma}_r \mathbf{V}_r^T + \mathbf{E}_r \approx \mathbf{\Psi}_r \mathbf{A}_r + \mathbf{E}_r \approx \mathbf{\Psi}_r \mathbf{A}_r \tag{3}$$

Similarly, any column vector of \mathbf{X} , or any vector similar enough to the columns of \mathbf{X} , can be approximated as in Equation (4). The vector \mathbf{a} can be seen as a low-dimensional representation of \mathbf{x} .

$$\mathbf{x} = \mathbf{\Psi}_r \mathbf{a} + \mathbf{e} \approx \mathbf{\Psi}_r \mathbf{a} \tag{4}$$

In this study, the number of POD modes and sparse sensors were both chosen to be $p = r = 10$. The choice of number of POD modes, r , and the number of sensors, p , is important to the performance of the method. Too small an r or p will lead to a poor performance, while too large an r or p will result in a too large model that might be slow or infeasible to use in practice. A too large p will also defeat the purpose of the method using sparse measurements. More details on the choice of $p = r = 10$ are given in Section 2.4.

2.3. Reconstruction from Sparse Measurement

This section explains how to estimate full-field (\mathbf{x}) from a sparse measurement (\mathbf{y}) given a tailored basis $\mathbf{\Psi}_r$ and the sensor locations in \mathbf{C} .

Here, \mathbf{y} is a sparse measurement of \mathbf{x} and is calculated as in Equation (5) where $\mathbf{C} \in \mathbb{R}^{p \times n}$ is the measurement matrix and \mathbf{v} is measurement noise. The measurement \mathbf{y} is sparse in the sense that it only contains information about relatively few entries in \mathbf{x} . The number of elements in \mathbf{y} is denoted as p and is the number of measurements taken of \mathbf{x} .

$$\mathbf{y} = \mathbf{C} \mathbf{x} + \mathbf{v} \tag{5}$$

$$\mathbf{y} = \mathbf{C}(\mathbf{\Psi}_r \mathbf{a} + \mathbf{e}) + \mathbf{v} \approx \mathbf{C} \mathbf{\Psi}_r \mathbf{a} \tag{6}$$

The matrix \mathbf{C} represents the part of \mathbf{x} that is measured in \mathbf{y} . It can be structured in different ways. One way is for each row of \mathbf{C} to consist of a single 1, with all other entries being 0. Then, each element in \mathbf{y} will be a direct measurement of a single element in \mathbf{x} . This case is called sparse sensor placement.

The relationship between \mathbf{y} and \mathbf{a} in Equations (6) is found by combining (5) and (4). Furthermore, \mathbf{a} and \mathbf{x} can be approximated from \mathbf{y} , as shown in Equation (7). These approximations are good as long as \mathbf{e} and \mathbf{v} are sufficiently small.

$$\mathbf{a} \approx (\mathbf{C} \mathbf{\Psi}_r)^\dagger \mathbf{y} \approx \mathbf{\Theta}^\dagger \mathbf{y} \tag{7}$$

$$\mathbf{x} \approx \mathbf{\Psi}_r \mathbf{a} \approx \mathbf{\Psi}_r (\mathbf{C} \mathbf{\Psi}_r)^\dagger \mathbf{y} \approx \mathbf{\Psi}_r \mathbf{\Theta}^\dagger \mathbf{y} \tag{8}$$

Thus, to reconstruct full-field \mathbf{x} , the condition number of $\mathbf{\Theta}$ (the row-selected basis matrix—a product of the measurement and basis matrices) should be small so that the input errors

are not amplified during the inversion in Equation (8). The condition number of Θ can be controlled by the choice of measurement matrix \mathbf{C} . A suitable sensor placement algorithm is the one that helps to find rows of Ψ_r corresponding to the point sensor location in state space that provides the optimal conditions for the inversion of matrix $\Theta = \mathbf{C}\Psi_r$. This is obtained by QR pivoting (the chosen sensor location method), as detailed in Section 2.4.

2.4. QR Pivoting for Sparse Sensor Placement

The choice of sensor locations in \mathbf{C} is important to enable optimal conditioning and inversion operation in Equation (8). The QR-decomposition with column pivoting is proposed as a computationally efficient alternative to finding optimal sensor locations [3]. This is carried out using the first p choices of column pivots when calculating the QR decomposition of $\Psi_r\Psi_r^T$ in the algorithm as the p sensor locations. Using $p = r$, one can calculate the QR decomposition of Ψ_r^T instead. Each row of \mathbf{C} is a row of zeros besides a single element that is set to 1. The resulting p sensor locations (selected pivots) are then used to find the elements in \mathbf{C} that should be set to 1. The value i -th chosen pivot is the index of the element on the i -th row of \mathbf{C} that is set to 1. The first pivot location is chosen by finding the row of the matrix $(\Psi\Psi_r^T$ or $\Psi)$ with the largest ℓ_2 -norm. The index of this row is the first chosen sensor location. The matrix $(\Psi\Psi_r^T$ or $\Psi)$ is then modified before the next iteration. Each row is subtracted by the projection of the row on the row corresponding to the chosen sensor location. This is repeated until the desired number of sensor locations is chosen. In this form, the algorithm can easily be modified to include constraints on the possible choices of sensor placement. This can be very useful; for example, if some sensor locations are not practical to use. Another case when this is useful is if one wants a maximum or minimum number of sensors inside an area.

Regarding the choice of number of modes r and number of sensors p in this work, explained variance is often used to choose the number of POD modes. This can be a useful tool to find a lower bound for good choices of number of modes r . Here, first two modes captures most of the variance in both the temperature and velocity fields. However, it is wiser to use a larger r because some time steps might not be modeled well by the POD-modes even though the explained variance is high. This can happen if many time steps are similar and a few time steps are different from the others. The explained variance might be high because the POD modes model the time steps that are similar to each other. The few time steps that are modeled poorly will not necessarily effect the explained variance. This could be the case in this application, where the time steps toward the end of the simulation are very similar because the system moves toward a steady state. However, the early time steps are very different from the steady state and could therefore be modeled poorly, even though the explained variance is high. Therefore, it is wiser to use a higher r .

However, there are some restrictions to the choice of r and number of sensors p : (a) The first restriction on the choice of r and p comes from the number of elements in the POD-modes. If $p > r$, the matrix $\Psi_r\Psi_r^T \in \mathbb{R}^{n \times n}$ is constructed in the sensor placement algorithm. However, if n is large enough, then constructing and using this matrix is infeasible. In this application, $n = n_x n_y n_z = 141,659$, which makes it infeasible to choose $p > r$. Instead, if $p = r$, then the matrix $\Psi_r \in \mathbb{R}^{n \times r}$ is used instead, which makes it much easier to work with for a small choice of r . (b) Another constraint is that one actually has to be able to measure at the chosen sensor locations when the method is used in practice. Therefore, one can not choose a p that is larger than the number of sensors one can access. In this application, a maximum of 10 temperature sensors can be used to estimate temperature from sparse measurements. Since $r = p$, the maximum constraint for r is also 10. We ended up using all 10 available physical sensors because the use of more sensors should not hurt the performance, as 10 is still a relatively small number compared to n . If using more sensors than necessary, one could use a subset of these sensors in future to estimate the temperature field and use the remaining sensors to validate how good the resulting estimate is. Therefore, $p = r = 10$ was chosen.

3. Results and Discussion

3.1. Flow Reconstruction from Sensor Placement for Test Data

The POD modes of the temperature data from training set were used as the tailored basis. The first two of those dominant temperature POD modes are shown in Figure 3. The POD modes (ψ) are then used to find good sensor locations (pivots in C) in the temperature grid, as shown in Figure 4.

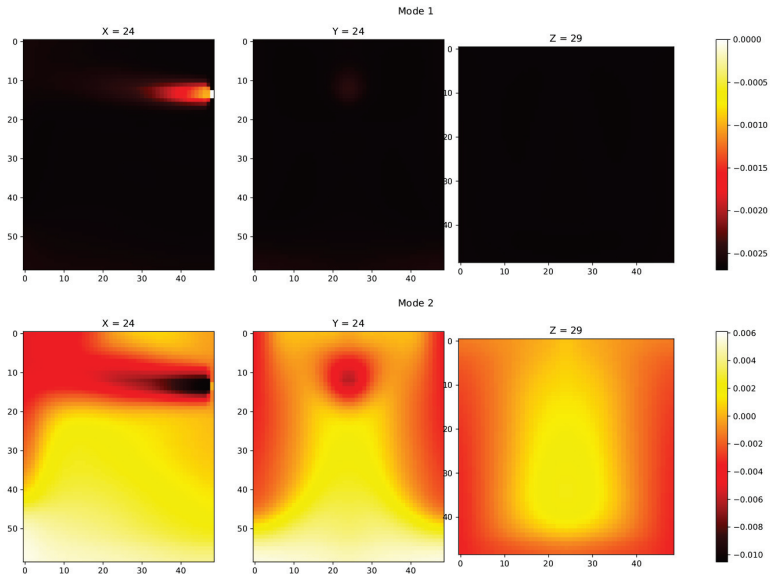


Figure 3. Mode 1 and Mode 2 from POD decomposition.

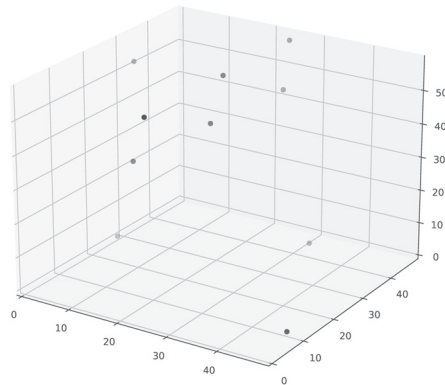


Figure 4. The chosen sensor locations.

These are the physical locations that should be measured for the reconstruction. Thus, one can estimate the whole temperature grid from a given sparse temperature measurement y . The temperatures measured in y correspond to the location in the measurement matrix C . The sparse sensor locations and basis functions obtained using the training data-set are then used to reconstruct the full temperature field for the unseen test case with an inflow of 2 m/s, and this involves using sparse measurements at test case (2 m/s case) at specified locations for the full temperature field reconstruction. The reconstructed flow field is then compared with the full CFD temperature field for the test case. Figure 5 shows examples of temperature estimations obtained from sparse measurements.

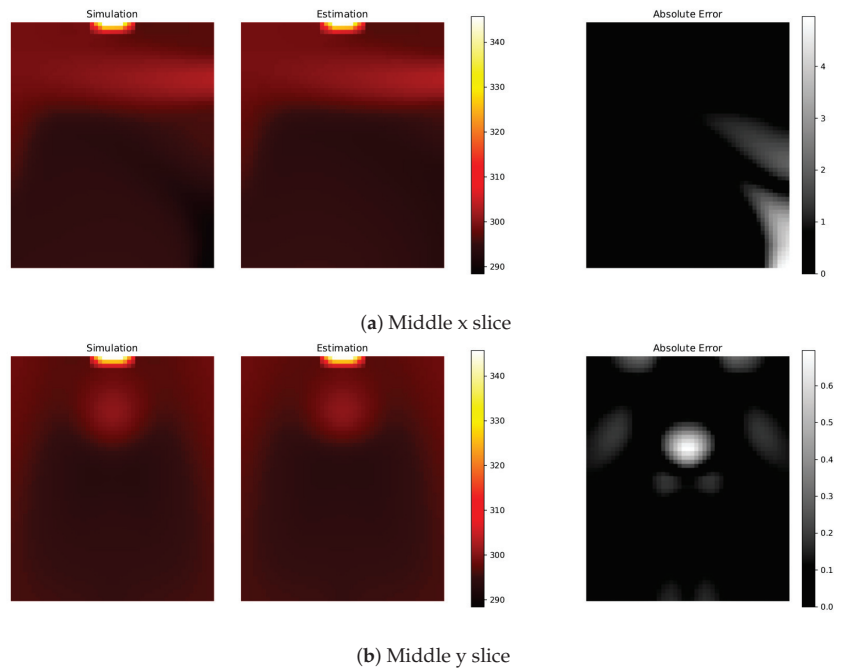


Figure 5. Comparing simulation data and their estimates from sparse measurements. This is for the time step with the largest MAE. Everything is measured in Kelvin.

Since the data are distributed in a 3D space, it is hard to show the entire estimate; these figures show the x and y slices in the middle of the 3D space. Figure 5a shows an example of one of the worst estimates with errors up to 4 Kelvin in the bottom right corner in Figure 5a. For a time-step with an error close to the average error (figure not shown), there are no errors above 1 K.

As seen in Figure 6 the error is largest in the first time steps of the time series estimated from sparse measurements. In the beginning of the simulation, the temperature is harder to model than the more homogeneous temperature that dominates the later time steps. The sparse measurements are used to find the linear combination of the POD modes that best fits the current state. If the current temperature field does not fit well with the POD modes, the estimate of the sparse measurement will not provide a result that is close to the ground truth. This is independent of the number of sensor locations that are used, as long as the number of POD modes does not increase.

A histogram of all the absolute errors at every spatial location and time step is shown in Figure 7. In this histogram, the vast majority of the errors in temperature estimation are below 0.5 Kelvin. Therefore, there seems to be only a few time steps and spatial positions with large errors, which provide the large errors shown in Figures 5 and 6. The errors are low in the rest of the region. Thus, the flow reconstruction from the sensor can be considered good, with scope for improvements in the future. In next section, we will see results for the reconstruction of a flow field using a data-driven reduced-order model for efficient predictions in a DT set-up.

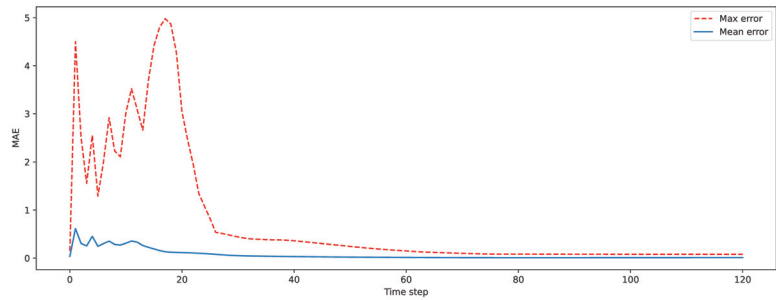


Figure 6. Maximum and average absolute error for each time step.

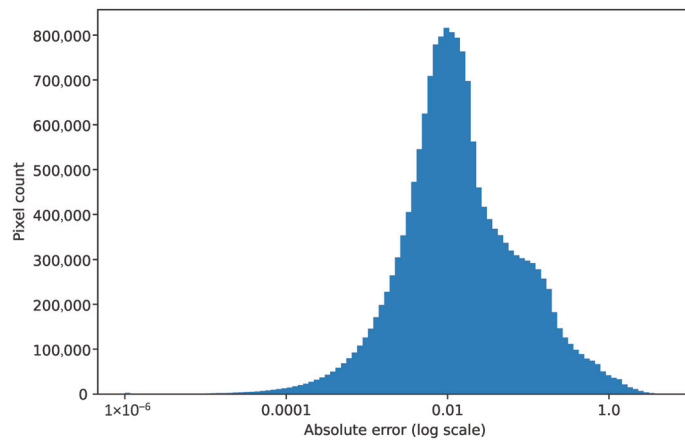


Figure 7. Histogram of the absolute error for each pixel for each time step.

3.2. Flow Reconstruction of Velocity Using LSTM-POD ROM

The training data-set for LSTM comprises an input 3D matrix containing the POD-evaluated time-coefficients and has dimensions of $N \times \sigma \times R$, corresponding to the number of samples (N), the look-back time steps ($\sigma = 3$ for this work) and the number of features (R). The R number of features correspond to the temporal coefficient values associated with R spatial basis function (modes). If needed, the flow-rate can be added as an additional feature (parameter). For LSTM output (target), a database of a 2D array of the temporal coefficients for time t is provided with dimensions $N \times R$ to train the LSTM. The LSTM is trained to map the inputs (σ previous time-steps of time coefficients) to time-coefficient values at time t for all R modes. Here, the LSTM parameters are found using the hyper-parameter optimization software optuna. The LSTM uses two layers with 30 units (neurons) and tanh activation function. The details of the ROM methodology involving LSTM can be found at [6]. Figure 8 shows the temporal dynamics of coefficients, as predicted by the trained LSTM model on test data. The trained LSTM model could qualitatively predict the temporal coefficient evolution trend when compared to the actual true coefficient. There is scope for improvements in its accuracy when using a larger training database to provide a result that is qualitatively similar to the true coefficient. These temporal coefficients are now used, along with the velocity basis functions (obtained from POD), to reconstruct the full velocity field for the unseen test case (as seen in Figure 9).

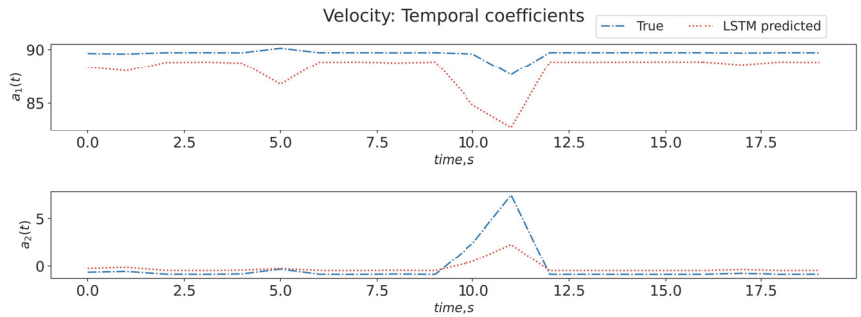


Figure 8. Temporal dynamics of coefficients captured by LSTM.

Figure 9 shows the reconstruction of velocity field at time t before the flow is fully developed using the (a) LSTM-POD methodology compared to that obtained by (b) high-fidelity CFD data.

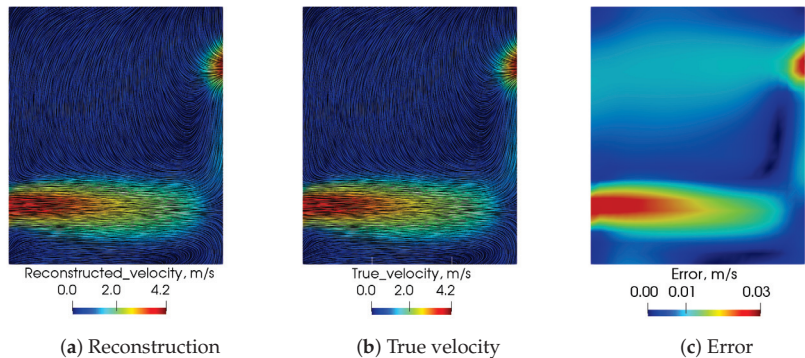


Figure 9. Comparison of prediction of velocity field between LSTM-POD reconstruction and high-fidelity CFD.

Figure 9 shows that the LSTM-POD methodology can predict qualitatively similar results (flow field) as actual CFD data in an online set-up using only a few minutes of computational time, as compared to the few hours of computational time needed by the hi-fidelity CFD with minimal errors. Thus, the methodologies developed to date have potential for use in a synthetic CFD data-set. The flow reconstruction from sensors needs to be further tested on actual sensor measurements in the experimental set-up, and there is scope for corrections to the reconstruction errors using techniques such as hybrid analytics and modelling.

4. Conclusions

The application of data-driven optimal temperature sensor placements and reduced-order models enabled us to reconstruct and predict the full-scale temperature and flow field for a synthetic dataset for a greenhouse digital twin set-up. The methodology shows promise in monitoring the spatio-temporal dynamics of key variables for a digital twin greenhouse setup. Future work involves increasing further the accuracy using an HAM-corrective source term approach to correct the reconstruction error while using sparse sensors and reduced-order models, and testing these with actual measurements from the experimental set-up.

Author Contributions: Conceptualization, A.R. and M.T.; methodology and experiments, K.S., E.B., M.T. and A.R.; CFD data curation, M.T.; greenhouse experiments and data curation, E.B.; software, K.S. and M.T.; validation, K.S. and E.B.; formal analysis, K.S. and M.T.; investigation, K.S. and M.T.; resources, A.R.; writing—original draft preparation, K.S., E.B. and M.T.; writing—review and editing, A.R. and M.T.; visualization, M.T., K.S. and E.B.; supervision, A.R. and M.T.; project administration, A.R.; funding acquisition, A.R. and M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the SEP funding from SINTEF for the project named: Highly Capable Digital twin, SEP project number: 102024647-54.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can be made available on request.

Acknowledgments: We would like to thank Disruptive Technologies for contributing with their advanced temperature and humidity sensors for the project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rasheed, A.; San, O.; Kvamsdal, T. Digital Twin: Values, Challenges and Enablers from a modeling perspective. *IEEE Access* **2020**, *8*, 21980–22012. [CrossRef]
2. Donoho, D. Compressed sensing. *IEEE Trans. Inf. Theory* **2006**, *52*, 1289–1306. [CrossRef]
3. Manohar, K.; Brunton, B.W.; Kutz, J.N.; Brunton, S.L. Data-Driven Sparse Sensor Placement for Reconstruction: Demonstrating the Benefits of Exploiting Known Patterns. *IEEE Control Syst. Mag.* **2018**, *38*, 63–86. [CrossRef]
4. Pawar, S.; Ahmed, S.E.; San, O.; Rasheed, A. Data-driven recovery of hidden physics in reduced order modeling of fluid flows. *Phys. Fluids* **2020**, *32*, 036602. [CrossRef]
5. Ahmed, S.E.; Pawar, S.; San, O.; Rasheed, A.; Tabib, M. A nudged hybrid analysis and modeling approach for realtime wake-vortex transport and decay prediction. *Comput. Fluids* **2021**, *221*, 104895. [CrossRef]
6. Tabib, M.V.; Pawar, S.; Ahmed, S.E.; Rasheed, A.; San, O. A non-intrusive parametric reduced order model for urban wind flow using deep learning and Grassmann manifold. *J. Phys. Conf. Ser.* **2021**, *2018*, 012038. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Forecasted Self: AI-Based Careerbot-Service Helping Students with Job Market Dynamics [†]

Asko Mononen ^{1,*}, Ari Alamäki ², Janne Kauttonen ², Aarne Klemetti ³, Anu Passi-Rauste ⁴ and Harri Ketamo ⁴

¹ Digital Living Lab, Laurea University of Applied Sciences, 02650 Espoo, Finland

² Digital Services, Haaga-Helia University of Applied Sciences, 00520 Helsinki, Finland

³ School of ICT, Metropolia University of Applied Sciences, 02610 Espoo, Finland

⁴ HeadAI Ltd., 28130 Pori, Finland; anu.passi-rauste@headai.com (A.P.-R.); harri.ketamo@headai.com (H.K.)

* Correspondence: asko.mononen@laurea.fi; Tel.: +358-400-679-768

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: In this article, we introduce an AI-enhanced study planning solution named Careerbot, which is a service designed to help students with their “forecasted self”. We define a new term “forecasted self” to mean a future-oriented digital twin, where a student can explore several future selves equipped with new, acquired skills for projected future jobs. The future jobs domain here includes knowledge-work related jobs related to digitalization, emerging technologies, and Industry 4.0/Society 5.0.

Keywords: artificial intelligence; big data; career coaching; data visualization; forecasted self; Industry 4.0; job market intelligence; mydata; skills data; Society 5.0

1. Introduction

In this article, we introduce an AI-enhanced study planning solution named Careerbot, which is a service designed to help students with their “forecasted self”. We define a new term “forecasted self” to mean a future-oriented digital twin, where a student can explore several future selves equipped with new, acquired skills for projected future jobs. We believe the use of this new term and approach will provide the benefits of understanding the following: (1) the essence of future orientation; (2) a holistic approach of soft skills and hard skills that are appreciated by employers; (3) the skill gap between current skills and the direction on which to focus skill acquisition, and (4) the ability to verbalize one’s skills and competences in the concrete language used in job ads by employers (in contrast to academic jargon, for example). We also use the term skills data as the unifying factor among different actors and operations: “skills data describes people’s skills, the competence needs of organisations, and the competence offerings of educational institutions. In practice, skills data can be found, for example, on employees’ CVs, companies’ job adverts, and course guides” [1].

We examine the adoption of artificial intelligence (AI) in three applied universities (3AMK) in Finland. More specifically, we analyse and discuss experiences regarding the educational AI-solution that assists higher education students by providing course suggestions, thesis topic trends, and job market data for their career and study planning. 3AMK is a strategic alliance among the three largest universities of applied sciences in Finland: Haaga-Helia, Laurea, and Metropolia (3amk.fi). 3AMK has approximately 34,000 students, 2000 staff, and 15 campuses in Helsinki, the capital region. In this paper, we conceptualize “forecasted self” based on the analysed experiences in designing and adopting the Careerbot AI-enhanced study planning service.

The adoption of AI is rapidly growing as a means to enhance students’ personal or collaborative learning and study planning in higher education. The adoption of AI

Citation: Mononen, A.; Alamäki, A.; Kauttonen, J.; Klemetti, A.; Passi-Rauste, A.; Ketamo, H. Forecasted Self: AI-Based Careerbot-Service Helping Students with Job Market Dynamics. *Eng. Proc.* **2023**, *39*, 99. <https://doi.org/10.3390/engproc2023039099>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera and Fernando Rojas Olga

Published: 4 September 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

provides new opportunities to develop study planning where AI can model and suggest competence profiles, needs, and requirements from real-time job market data. The prior literature on AI-enhanced learning and teaching shows that AI can create value for students and teachers, e.g., [2–4]. AI enables personalization, e.g., [5,6] which is an important requirement in improving and customizing learning for the special needs of each student. AI is also widely adopted for students' performance assessment, competence profiling and assessment, finding learning gaps, and predicting students' progress in the courses [7–9].

2. Background of Careerbot-Service of 3AMK in Finland

3AMK developed Careerbot-service to help their students to pursue their dream careers with the help of AI. The Careerbot-service can help 3AMK students to do the following:

- (1) Verbalize their skills with the help of AI (skills profile; current or “forecasted self” in the future);
- (2) Find jobs with their skills profiles (job market intelligence);
- (3) Find courses for skill development (upskilling, re-skilling);
- (4) Find theses/research topics, trends, and content (research intelligence).

Careerbot-service uses a language model based on AI, which has been trained with millions of news articles and with ESCO classification, for example. ESCO is the multilingual classification of European Skills, Competences, and Occupations. ESCO is part of the Europe 2020 strategy.

The data sources for Careerbot-service currently include the following:

- (a) Job market data in Finland (Työmarkkinatori, MOL, and Duunitori/employment services) with more than 400,000 job ads on a yearly basis, since January 2018;
- (b) 3AMK course data for all 15,000 courses;
- (c) Theseus—A thesis database with more than 120,000 theses available from Finland, since 2010;
- (d) Global article database, a directory of open access journals, DOAJ, with more than 8.6 million articles.

The language model, foresight data products (curriculum data, labor market data, investment data, and research papers), and the AI behind the service, Graphmind, are powered by the Finnish tech company HeadAI Ltd, from Pori, Finland. Graphmind is a graph machine-learning-based semantic computing framework accessible via REST-API. The usage of API allows Careerbot-service the flexibility to use any AI model to expand its functionality.

The basic operations behind the framework are the following:

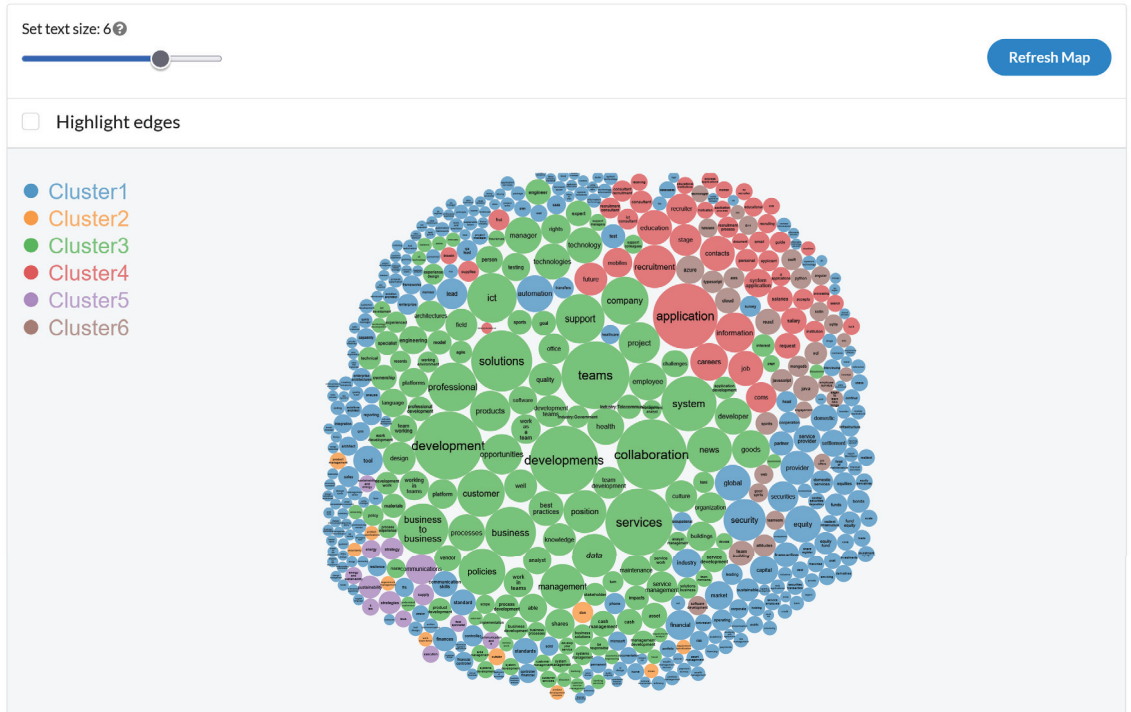
- (i) Building a digital twin (personal, curriculum, and scenarios);
- (ii) Comparing two digital twins against each other to show similarities and gaps;
- (iii) Recommending interventions from the third digital twin to bridge the gap.

In addition to the 3AMK students, the 3AMK staff have access to service. Lecturers and content creators can ensure their content is up to date. RDI staff can search for research ideas or prior research articles for supporting the new externally funded RDI projects. In addition, career coaching can use Careerbot-service in their career counseling for students, backing up visual CVs with data and vocabulary known in the work sector.

Figure 1 below demonstrates one example map, a zoomable snapshot of the most important hard and soft skills in ICT in the Helsinki region, with data from the previous cut-off date. The clusters below in the bubble chart (a) represent the same data in the top lists that are shown in (b), namely the 15 largest hits in order of relevance. There are currently 19 ready-made example maps in Finnish and English: 13 job market maps and 6 curriculum maps. The maps are updated every 1–2 months, which seems to be frequent enough to see the current changes. The same functionality can visualize the students' skills profile data, so they can attach the image to their CVs, for example. These maps leverage the semantic language model and its graphical representation of terms. This graph

is visualized using a multi-body particle simulation model to represent the graph as a collection of 2D non-overlapping disks. The algorithm pushes the most connected terms towards the center, while the less connected terms stay near the boundary. The clustering is computed using a weighted community detection algorithm [10].

Most important hard skills and soft skills in Helsinki, Espoo, Vantaa region in 2023-2 limited to theme ICT (EN)



(a)

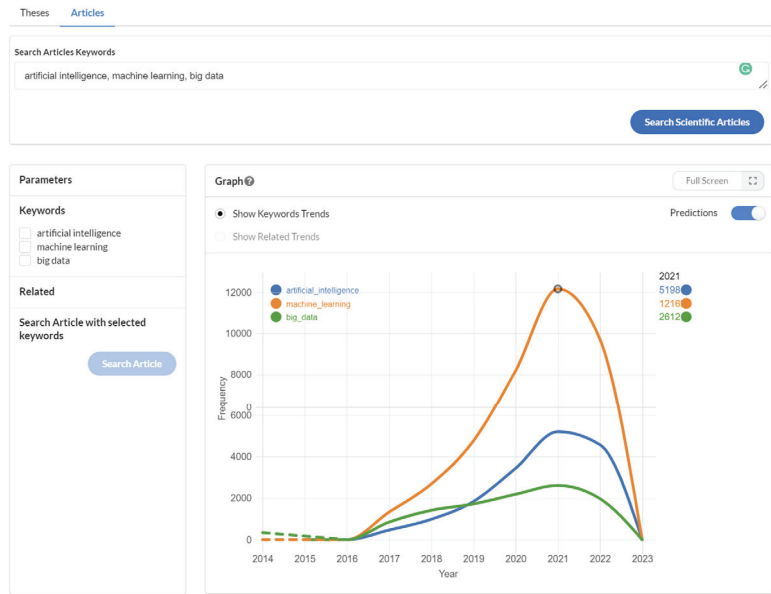
Figure 1. Cont.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
security 20	don 9.9	collaboration 33.5	application 30.7	communications 14.6	azure 13.4
equity 17.5	outside 8	development 32.4	recruitment 20.4	sustainability 10.7	team_building 12.2
automation 17.5	product_management 8	developments 32.4	information 18.8	strategies 10.7	cloud 12.2
global 17	mean 6.8	teams 32.1	careers 17.9	strategy 10.7	react 12.2
provider 16.1	uncertainty 6.8	services 31.9	job 16.6	supply 9.9	java 12.2
securities 14.6	product_development_process 6.8	solutions 29	contacts 16.6	energy 9	typescript 10.7
lead 14.6	requirement_specification 6.8	system 26.2	education 16.6	energy_and_sustainability 8	spirits 10.7
industry 14	requirement_management 6.8	business_to_business 24.1	future 15.6	execution 8	attitudes 10.7
financial 14	product_specifications 6.8	business 24.1	stage 15.1	it_law 8	sql 10.7
tool 14	work_experience 6.8	ict 23.5	coms 14.6	laws 8	javascript 10.7
capital 13.4		professional 23.2	recruiter 14.6	communication_and_it 8	mongodb 9.9
market 13.4		support 23.2	mobiles 14	field_specialist 8	python 9.9
service_provider 12.8		company 22.5	system_application 13.4	sustainability_and_energy 8	swift 9.9
settlement 12.8		customer 22.5	request 12.2		aws 9.9
domestic 12.8		policies 22.2	salary 12.2		web 9

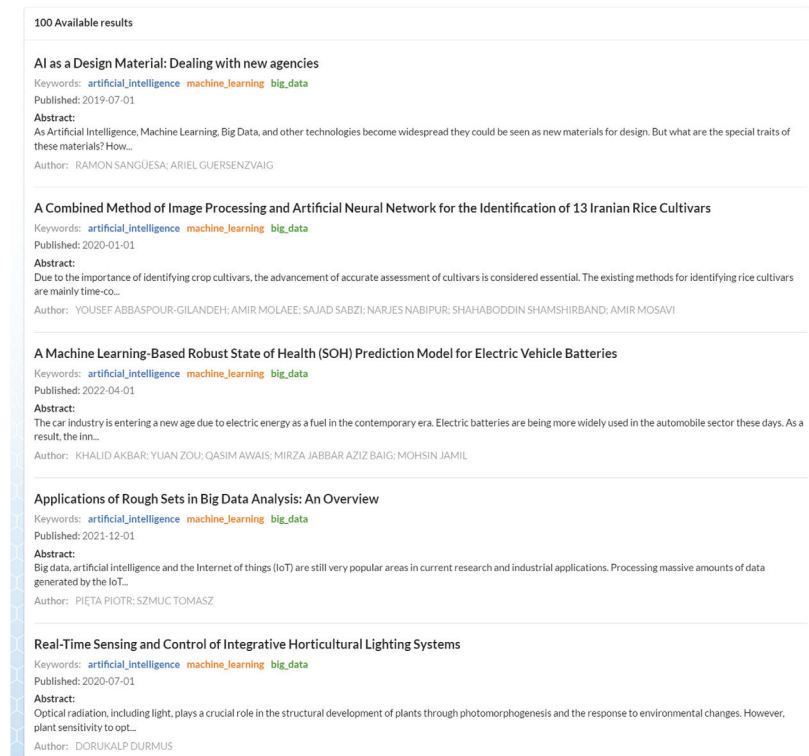
(b)

Figure 1. Zoomable example map in Careerbot-service: (a) represented as clusters and (b) same data in top 15 lists, with color-coded clusters.

Figure 2a depicts trends from the global DOAJ article database (doaj.org), with the search words artificial intelligence, machine learning, and big data. The data are updated currently until December 2021, so the year 2023 is denoted as zero. From the graph, we can conclude that of the search results, “machine learning” had been trending clearly above “artificial intelligence” and “big data” in 2021. The prediction tab is used for testing; it calculates the following years based on the historical data and fits a B-spline approximation for the data [11]. Figure 2b lists the search results page below the graph (Figure 2a), where the individual papers can be opened with a mouse click.



(a)



(b)

Figure 2. (a) Searching global DOAJ article database, with trends shown. (b) Searching global DOAJ article database, search results page. Source: Careerbot-Service.

3. Conclusions

We contribute to the discussion of AI-enhanced learning and teaching by conceptualizing “forecasted self”, a novel concept for the digital twin approach in the context of education.

“Forecasted self” is defined as a future-oriented digital twin, which allows students to explore several future selves equipped with new, acquired skills for projected future jobs of Society 5.0.

We also use the term skills data as the unifying factor among different actors and operations. The idea is to combine all the relevant distributed data sources, including internal data, mydata, and public/open data. This approach supports the EU skills data space initiative by mapping, matching, and forecasting skill-based data.

3AMK developed and adopted an AI-enhanced study planning service named Careerbot for helping higher education students with their “forecasted self”. Via API, the service can leverage any AI model running on the server or in the cloud.

The adoption of AI is rapidly growing in higher education and provides new opportunities to develop study planning, learning and teaching, including personalization and customizing learning for the individual needs of each student. Students need to be able to create their own digital competence profile (digital twin), for example, with the help of a Careerbot AI solution that simulates the competence requirements of the up-to-date and current job market data.

Author Contributions: Conceptualization & methodology: A.M. and J.K.; validation J.K., H.K. and A.P.-R.; visualization A.M.; writing—original draft preparation A.M., J.K. and H.K.; review and editing A.M., J.K., A.K., A.A. and A.P.-R.; project administration A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data from Careerbot-service is currently available only for 3AMK students and staff.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Technology Industries of Finland. Skills Data Playbook. 2022. Available online: [https://teknologiateollisuus.fi/sites/default/files/inline-files/Osaamisdatan-Playbook\[-\]-\[ENG-03-aukeamittain.pdf](https://teknologiateollisuus.fi/sites/default/files/inline-files/Osaamisdatan-Playbook[-]-[ENG-03-aukeamittain.pdf) (accessed on 23 March 2023).
2. Mononen, A.; Alamäki, A.; Kauttonen, J.; Klemetti, A.; Räsänen, E. Adopting AI-enhanced chat for personalising student services in higher education. In Proceedings of the AINL 2020 Artificial Intelligence and Natural Language Conference, Online, 5–7 July 2020.
3. Popenici, S.A.; Kerr, S. Exploring the impact of artificial intelligence on teaching and learning in higher education. *Res. Pract. Technol. Enhanc. Learn.* **2017**, *12*, 10–11. [CrossRef] [PubMed]
4. Renz, A.; Krishnaraja, S.; Gronau, E. Demystification of artificial intelligence in education—How much AI is really in the educational technology? *Int. J. Learn. Anal. Artif. Intell. Educ. (Ijai)* **2020**, *2*, 4–30. [CrossRef]
5. Chassignol, M.; Khoroshavin, A.; Klimova, A.; Bilyatdinova, A. Artificial intelligence trends in education: A narrative overview. *Procedia Comput. Sci.* **2018**, *136*, 16–24. [CrossRef]
6. Tiihonen, J.; Felfernig, A. An introduction to personalization and mass customization. *J. Intell. Inf. Syst.* **2017**, *49*, 1–7. [CrossRef]
7. Costa, E.B.; Fonseca, B.; Santana, M.A.; de Araújo, F.F.; Rego, J. Evaluating the effectiveness of educational data mining techniques for early prediction of students’ academic failure in introductory programming courses. *Comput. Hum. Behav.* **2017**, *73*, 247–256. [CrossRef]
8. Ketamo, H.; Moisio, A.; Passi-Rauste, A.; Alamäki, A. Mapping the future curriculum: Adopting artificial intelligence and analytics in forecasting competence needs. In Proceedings of the 10th European Conference on Intangibles and Intellectual Capital ECIIIC 2019, Chieti-Pescara, Italy, 23–24 May 2019; Sargiacom, M., Ed.; pp. 144–153.
9. Yang, F.; Li, F.W. Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Comput. Educ.* **2018**, *123*, 97–108. [CrossRef]

10. Clauset, A.; Newman, M.E.; Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **2004**, *70*, 066111. [CrossRef] [PubMed]
11. Unser, M.; Aldroubi, A.; Eden, M. On the asymptotic convergence of B-spline wavelets to Gabor functions. *IEEE Trans. Inf. Theory* **1992**, *38*, 864–872. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Proceeding Paper

Predictive Accuracy of Logit Regression for Data-Scarce Developing Markets: A Nigeria and South Africa Study [†]

Jonathan D. Oladeji ^{*}, Benita G. Zulch and Joseph A. Yacim

Department of Construction Economics, Faculty of Engineering, Built Environment and Information Technology, University of Pretoria, Pretoria 0002, South Africa; benita.zulch@up.ac.za (B.G.Z.); josephyacim@gmail.com (J.A.Y.)

^{*} Correspondence: jonathan.oladeji@tuks.co.za

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: This research examines how much forecasting accuracy can be achieved by modelling the relationships between listed real estate and macroeconomic time series variables using the logit regression model. The example data for this analysis included 10-year (2008–2018) transactions. The Statistical Package for Social Sciences (SPSS, version 25) and Microsoft Excel 2016 were used for descriptive and inferential analysis. The data collected on the listed real estate transactions for South Africa and Nigeria represent the largest listed real estate markets in the continent. The study found that 22.2% variance in the Nigerian real estate market was explained by the lending rate, treasure bill rate, and Consumer Price Index, while 9.4% variance in the South African real estate market was explained by changes in the exchange rate and coincident indicators. The strength and similarity of the model capacity in both countries showed that each market signal has a predictive accuracy of 75% (Nigeria) and 80% (South Africa).

Keywords: economic leading indicators; real estate; forecasting; investment; market modelling

1. Introduction

The ability to predict and model market trends is an important part of the investment decision-making process for local and foreign real estate investors, especially where quantitative data are scarce. Investors view real estate as an asset class competing against other investment opportunities in stocks and shares. For this reason, understanding real estate behaviour and the future trajectory of emerging markets provides a strong basis for investment.

Real estate is one of the sectors that contribute to gross domestic products (GDP) of countries worldwide. Hongkong and Shanghai Banking Corporation [1] reported that real estate assets were valued at USD 228 trillion in 2016 alone, while Gordon [2] noted that global real estate accounted for 60% of mainstream global assets and about three times the size of global GDP in 2015. These figures show that real estate attracts a lot of investment and has even more potential to grow.

In Figure 1, the number of listed Real Estate Investment Trusts (REITs) demonstrate growing interest in emerging African markets which must be supported by improved accuracy in forecasting and reporting to stimulate investments. As mentioned by Bello and Yacim [3], comparable data as the bedrock of most rental value assessments are sparsely available in most emerging markets [3]. This makes it difficult, if not impossible, to rely on historical data for rent forecasts. Boshoff [4] and Keng [5] suggest that listed data and economic time series are useful for evaluating the future trajectory of real estate markets, which could help investors make better capital divestment decisions. This informs the choice of two locations in sub-Saharan Africa for better decision marking. Additionally, the sizes of the two markets play a key role in the choice. Nigeria is Africa's most populated

Citation: Oladeji, J.D.; Zulch, B.G.; Yacim, J.A. Predictive Accuracy of Logit Regression for Data-Scarce Developing Markets: A Nigeria and South Africa Study. *Eng. Proc.* **2023**, *39*, 100. <https://doi.org/10.3390/engproc2023039100>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 4 September 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

country, while the population of South Africa has been estimated at approximately one-third of Nigeria's population. In recent years, Nigeria and South Africa have both been rated as Africa's largest economies [6,7].

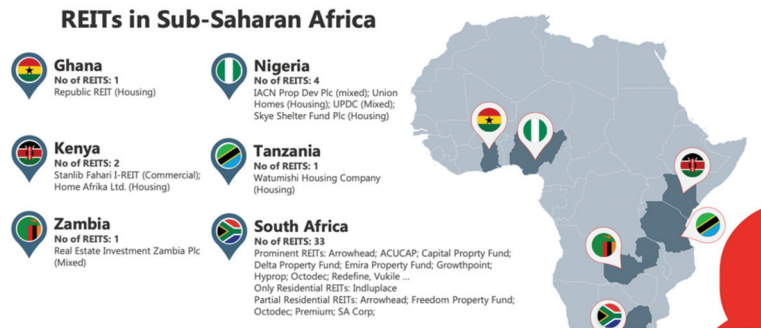


Figure 1. Growing REITS across sub-Saharan Africa.

The two economies have experienced similar enough growth patterns to compare their performance [6]. The two economies have experienced major shocks and economic prospects and are still fraught with uncertainties. Investors are likely to perceive these two economies in a similar light despite their population differences. This paper argues that confidence in the real estate sector in sub-Saharan Africa can be improved when valuable insight is achieved using leading economic indicators [8] to forecast turning points in real estate markets. This paper investigates how much forecasting accuracy can be achieved by modelling the relationships between listed real estate and macroeconomic time series using the logit regression model.

2. Literature Review

The fact that real estate is viewed as an asset class like stocks and equities informs the use of historical data and analysis for macroeconomic and financial investment analysis. Tonelli and Cowley [9] showed that an understanding of the past behaviour of the rent component was valuable for evaluating future behaviour [10,11]. However, as noted by Aron and Muellbauer [12], there is a need for caution in using merely market comparison and imputed rents. Olanrele et al. [13], in a study of the Nigerian market, focused on evaluating the causal relationship between N-REITs' dividend yield and five Money Market Indicators (MMI). They concluded that there was indeed a relationship between REIT returns and the MMI variables both in the short run (through Trace) and long run (using Max-Eigen values). Their study limited its research to data from a single REIT, the Skye REIT, and as such could not provide much insight into the future trajectory of the underlying real estate market.

Similarly, Boshoff [4] investigated listed real estate assets in South Africa and brought together two different asset classes, the stocks and bond market and the real estate market, as similar entities. The study asserted that price detection occurred in the listed real estate market, which could also be a signal of market movement in the direct real estate market. A lot of research has been carried out to promote macroeconomic modelling as a means of improving forecasting accuracy. Some popular works in these subject areas include studies such as those of Munusamy, Muthuveerappan and Baba [14]; Jadevicius, Sloan and Brown [15]; Tsolacos [16]; Jadevicius, Sloan and Brown [17]; and Boshoff [4]. Buehler and Almeida [18] discussed the investment market in the United States relative to predicting commercial market bubbles as a part of decision making for global investors. They posited that in creating predictive models, identifying the "right" set of variables that combined to trigger changes in the market was a first step in predictive modelling.

Munusamy, Muthuveerappan and Baba [14] considered a variety of literature regarding modelling types, accuracy and adoption of statistical modelling techniques. They

reported that Multiple Regression Analysis (MRA) and Artificial Neural Networks (ANNs) were most widely used. They concluded that ANNs showed an average error rate between 5 to 10% inaccuracy, while multiple regression analysis showed a higher average, which was 10 to 15%. Chrostek and Kopczewska [19] and other similar studies [20] compared the quality of prediction for several models: a classical linear model estimated with ordinary least squares (OLS), a linear OLS model including geographical coordinates, a spatial expansion model, spatial lag and spatial error models as well as geographically weighted regression. They concluded that models comprising the spatial components rendered better estimates than a-spatial models. They posit that there is evidence of the capacity of complex models such as these to predict rent behaviour.

However, other researchers like Moolman and Jordaan [21]; Tsolacos and Brooks [22]; Boshoff [4]; and Udoekanem, Ighalo and Sanusi [23] have preferred simpler models such as simple regression, vector auto regression and binomial logit regression. These studies point to the use of logit regression as a form of directional forecast, but while this method has not been tested on African data, this study uses unexplored time series data to test the predictive modelling approach. Considering the various positions on appropriate modelling techniques, binary logistic regression proves to be popular and devoid of complex parameters. This makes it desirable for further testing with African data sources. In the next section, this study's methods and results are discussed. The performance of the outcome model determines the extent to which macroeconomic data can serve to improve the accuracy of predictive rent models [8,24].

3. Methodology

South African economic data were collected from the Iress expert and Statistics South Africa (Stat SA) database. The FTSE/JSE SA Listed Property (J253) consists of the twenty largest liquid companies by market capitalisation in the Real Estate Investment and Services Sector and Real Estate Investment Trust Sector with a primary listing on the JSE. Nigerian listed real estate data were collected from the Central Bank of Nigeria, Sky Shelter REIT (SKY REIT) and the UACN property development company data. All monthly data were converted into quarterly data prior to analyses to ensure data uniformity with the exogenous data. The available data for the Nigerian listed real estate market were collected for the period 2008 Q1 to 2018 Q4. The South African macroeconomic data series were collected between the first quarter of 2003 and the fourth quarter of 2018.

To handle seasonality in the time series data, the Nigerian REIT and JSE time series data were used to create dummy binary outcomes for the purpose of logistic regression. The time series data difference of $Y_t - Y_{t-1}$ was classified based on a rise or fall. Growth in the time series represented a 0, while a fall represented a 1. This provided the data for the binary variable in both data sets. The South African dummy variable is denoted as South Africa Listed Real Estate (SALRE), while the Nigerian dummy variable is denoted as Nigeria Listed Real Estate (NLRE).

A turning point analysis was performed by using a regression probability model that outputs a dichotomous binary result that can be either 1 or 0. The threshold for turning point detection is typically set at 50% or 0.5 thresholds.

4. Results and Discussion

For SALRE indicators, the Hosmer and Lemeshow test shows a high value of 0.757 in Table 1, which proves the goodness of fit of the model. For the NLRE model, the Hosmer and Lemeshow test shows a value of 0.825, which proves the goodness of fit of the model.

Table 1. Hosmer and Lemeshow goodness-of-fit test of Logit Regression Model ($r \leq 0.05$).

	Chi-Square	Df	Sig.
South Africa (SALRE)	5.010	8	0.757
Nigeria (NLRE)	3.599	7	0.825

In Table 2, the SALRE model, the Cox and Snell and Nagelkerke R-squared were 0.256 and 0.364, respectively, which implies that the model explains about 25.6% or 36.4% variation in the dependent variables. The Cox and Snell and Nagelkerke R-Square values in the NLRE model were 0.440 and 0.587, respectively. This translates to 44% and 58.7% estimates of how much of the variation in listed real estate is explained by the model. The Cox and Snell and Nagelkerke R-squared were 0.440 and 0.587, respectively, which implies that the model explains about 44.0% or 58.7% variation in the dependent variables.

Table 2. Pseudo-R values of logit regression model ($r \leq 0.05$).

	-2 Log Likelihood	Cox and Snell R Square	Nagelkerke R Square
South Africa (SALRE)	58.920 ^a	0.256	0.364
Nigeria (NLRE)	28.920	0.440	0.587

^a denotes that the South Africa model tested for goodness of fit excludes variables with a high correlation.

T, being the state of the independent variable, is estimated to be 1 or 0, based on the logit regression rule:

T = 1 for the period that capital values decline;

T = 0 otherwise.

Therefore, the objective of using a logit approach is to estimate a response probability:

$$\Pr(T = 1 | x) = \Pr(T = 1 | x_1, x_2, \dots, x_k) \tag{1}$$

$$\Pr(T = 1 | x) = \log(p/1 - p) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \tag{2}$$

Table 3 value are included in Equation (1), the coincident indicator $\beta = 0.479$ and the exchange rate $\beta = 0.083$. The constant or intercept value was -74.738 . This model is expressed thus:

$$Y = \log(p/1 - p) = \beta_0 + \beta_1 CI + \beta_2 ER \tag{3}$$

$$Y = \Pr(T = 1 | x) = \log(p/1 - p) \text{ or } \ln(\text{ODDS}) = +0.479 (CI) + 0.083 (ER) \tag{4}$$

where:

Table 3. Indicators included in the accepted model for South Africa.

	B	S.E.	Wald	df	Sig.
GDP at market prices (R million)	0.000	0.000	0.215	1	0.643
Coincident indicator (2015 = 100)	0.479	0.156	9.467	1	0.002
M0	0.000	0.000	1.509	1	0.219
M1A	0.000	0.000	0.403	1	0.526
M1	0.000	0.000	1.135	1	0.287
M2	0.000	0.000	0.523	1	0.469
Total monetary (M3) deposits	0.000	0.000	0.169	1	0.681
Price of gold per ounce (Rand)	0.000	0.000	3.409	1	0.065
Exchange rates	0.083	0.032	6.598	1	0.010
Constant	-74.738	28.311	6.969	1	0.008

Y = SALRE

Pr = Probability

T = The indicator of a fall of capital values

p = Probability of a decline in capital values

β_0 = Model intercept
 β_x = Regression coefficient
 CI = Coincident indicator
 ER = Exchange rate

For the Nigerian data sets in Table 4, the β (beta coefficient) values showed the lending rate, treasury bill rate and Consumer Price Index/inflation, with a significance score on 0.01 level.

$$\log(p/1 - p) \text{ or } \ln(\text{ODDS}) = -21.938 + 0.143(\text{IR}) - 0.037(\text{TBR}) - 0.034(\text{CPI}) \quad (5)$$

where:

Table 4. Indicators included in the accepted model for Nigeria.

	B	S.E.	Wald	Df	Sig.
Total GDP	0.000	0.000	4.419	1	0.036
Composite Consumer Price Index (%)	-0.034	0.086	0.153	1	0.695
Prime lending/interest rate (%)	0.143	0.331	0.187	1	0.666
T-bill %	-0.037	0.064	0.339	1	0.560
Total money asset	0.000	0.000	4.087	1	0.043
Money supply (M1)	0.000	0.000	0.198	1	0.656
Currency in circulation	0.000	0.000	2.307	1	0.129
Money supply (M2)	0.000	0.000	6.249	1	0.012
Constant	-21.938	21.429	1.048	1	0.306

Y = NLRE
 Pr = Probability
 T = The indicator of a fall of capital values
 p = Probability of a decline in capital values
 IR = Lending/interest rate
 TBR = Treasury bill rate
 CPI = Consumer Price Index

Comparison of MODEL Performance in Identifying Leading Economic Indicators in Nigeria and South Africa

In Table 5, the Nigerian logit model outperforms the South African logit model by a 22.2% improvement on the null model as against the 9.4% improvement observed in the South African model. However, the misclassification rate for the Nigerian logistic model is 5% higher.

Table 5. The model misclassification rate.

Country	Classification Accuracy (%)	Improvement on Null Model (%)	Misclassification Rate
South Africa	79.7	9.4%	20%
Nigeria	75.0	22.2%	25%

In Figure 2, the area under the curve for Nigeria is 0.837, with a 95% confidence interval (0.714, 0.959). The area under the curve is also significantly different from 0.5 since the *p*-value is 0.000. Similarly, for South Africa, the area under the curve is 0.815, with a 95% confidence interval (0.704, 0.927). The classification similarities between the two models from South African and Nigerian data are visualised in Figures 3 and 4.

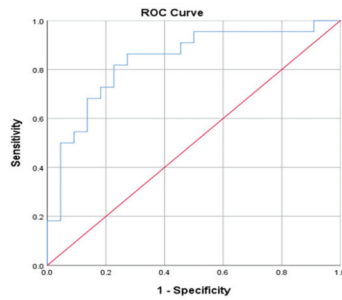


Figure 2. ROC curve for Nigerian predicted probabilities.

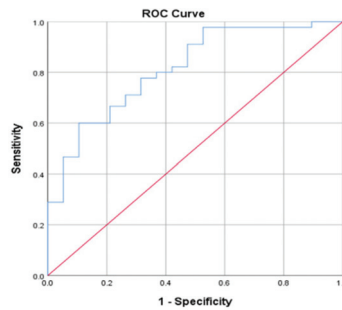


Figure 3. ROC curve for South African predicted probabilities.

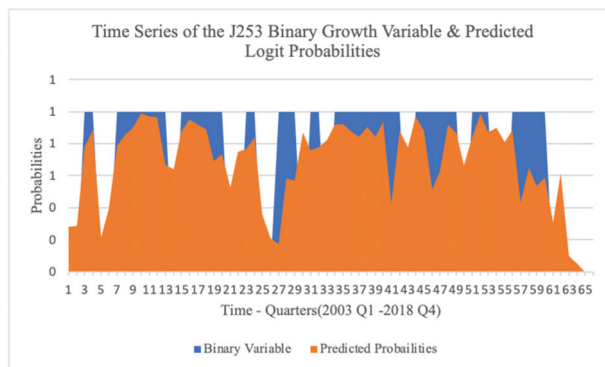


Figure 4. Time series of the J253 binary growth variable and predicted logit probabilities.

In Figures 4 and 5, the time series of the dependent variable is plotted against the probabilities predicted by the logit model in each model. This forecasting visualisation reports the in-sample forecasting results. The predicted probabilities for the South African model, as seen in Figure 4, reach a peak, while frequently coinciding with the J253 growth. The declines in the data coincide only twice in Q4 of 2003 and 2008, however. The South African model may predict growth probabilities more accurately than it predicts falls. Conversely, the probabilities of the Nigerian model coincide almost as accurately in the declines as they do in the peaks. This can be seen in Figure 5. The Nigerian market demonstrates a more responsive listed real estate market to economic indicators.

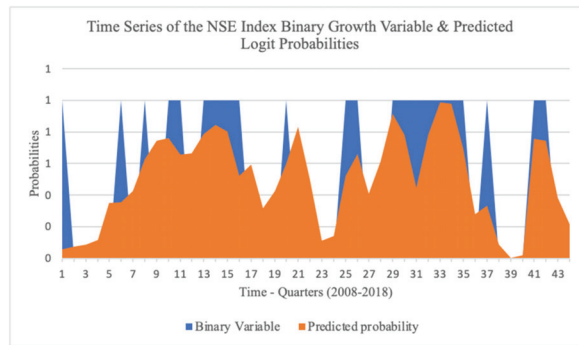


Figure 5. Time series of the NSE binary growth variable and predicted logit probabilities.

5. Conclusions

This study explored the existing literature on the selection and sources of leading indicator data for predicting the movement or changes in South African and Nigerian listed real estate. The paper investigates how much forecasting accuracy can be achieved by modelling the relationships between listed real estate and macroeconomic time series using the logit regression model. Data from South Africa and Nigeria were input into the logit regression model to evaluate how much predictive power the model has.

For predictive rent modelling to perform adequately, the use of simple probabilistic models proved valuable aligning with the approach by previous authors [24] in addressing countries and central bank policy changes relative to changes in macroeconomic indicators. The logistic regression model evaluated the performance of the best fitting model to classify the state of a dummy variable $T = 0$ or 1 , representing growth or decline in the listed real estate indicators. The test for predictive accuracy showed that 22.2% variance in the Nigerian real estate market was explained by the logit regression model, while 9.4% variance in the South African real estate was explained by changes in the exchange rate and coincident indicators. These findings agree with the results by Olanrele et al. [13] and Boshoff [4]. The strength and similarity of the model capacity in both countries showed that each market signal correctly predicts turning points in the economy for as much as 75% (Nigeria) and 80% (South Africa) of the time. The misclassification rate for the Nigerian logistic model is, however, 5% higher which is similar to the average model error margins observed by Munusamy, Muthuveerappan and Baba [14]. Meanwhile, the classification accuracy of the South African logit model is higher than that of the Nigerian logit model.

Author Contributions: Conceptualization, J.D.O., B.G.Z. and J.A.Y.; methodology, J.D.O.; software, J.D.O.; validation, J.D.O., B.G.Z. and J.A.Y.; formal analysis, J.D.O.; investigation, J.D.O., B.G.Z. and J.A.Y.; resources, J.D.O., B.G.Z. and J.A.Y.; data curation, J.D.O.; writing—original draft preparation, J.D.O.; writing—review and editing, J.D.O., B.G.Z. and J.A.Y.; visualization, J.D.O.; supervision, B.G.Z. and J.A.Y.; project administration, B.G.Z.; funding acquisition, J.D.O. All authors have read and agreed to the published version of the manuscript.

Funding: The IREBS Foundation for African Real Estate Research and the University of Pretoria Postgraduate Bursary has supported this research. The funding was provided for completing a master's dissertation including data gathering, research design, and reporting.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We acknowledge the support of the IREBS Foundation for African Real Estate Research and the Department of Construction Economics, University of Pretoria.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hongkong Shangai Banking Corporation (HSBC). Rising Potential of The World's Largest Asset Class. 2017. Available online: <https://sp.hsbc.com.my/liquid/6642.html> (accessed on 24 August 2023).
2. Gordon, S. Value of All Global Real Estate Totals \$217 Trillion. Pam Golding Properties. 2016. Available online: <https://www.pamgolding.co.za/property-research/2016/2/2/value-of-all-global-real-estate-totals-217-trillion> (accessed on 24 August 2023).
3. Bello, O.M.; Yacim, A.J. An Assessment of the Impact of Tree Shade on Rental Value of Residential Property in Maiduguri, North—Eastern, Nigeria. In Proceedings of the XXV FIG Congress, Kuala Lumpur, Malaysia, 16–21 June 2014; pp. 1–18.
4. Boshoff, D. Empirical Analysis of Space and Capital Markets in South Africa: A Review of the REEFM- and FDW Models. *S. Afr. J. Econ. Manag. Sci.* **2013**, *16*, 383–394. [CrossRef]
5. Keng, T.Y. Australian property securities funds: A survey of strategic investment issues. *Pac. Rim Prop. Res. J.* **2004**, *10*, 263–282. [CrossRef]
6. Naidoo, P. As African Leaders Meet on Growth, SA and Nigeria Are a Drag. News24. 2019. Available online: <https://www.news24.com/fin24/economy/as-african-leaders-meet-on-growth-sa-and-nigeria-are-a-drag-20190902> (accessed on 24 August 2023).
7. Rossouw, J. South Africa Is Africa's Largest Economy (Again). But What Does It Mean? The Conversation. 2016. Available online: <https://theconversation.com/south-africa-is-africas-largest-economy-again-but-what-does-it-mean-63860> (accessed on 24 August 2023).
8. Frankel, J.; Saravelos, G. Can Leading Indicators Assess Country Vulnerability? Evidence from The 2008–2009 Global Financial Crisis. *J. Int. Econ.* **2012**, *87*, 216–231. [CrossRef]
9. Tonelli, M.; Cowley, M. Forecasting Office Building Rental Growth Using a Dynamic Approach. *Pac. Rim Prop. Res. J.* **2004**, *10*, 283–304. [CrossRef]
10. Crosby, N.; Henneberry, J. Changing Investment Valuation Practices in the UK. In Proceedings of the Annual Conference of the European Real Estate Society, Milan, Italy, 23–26 June 2010.
11. Irohama, C.O.; Oluwunmi, A.O.; Simon, R.F.; Akerele, B.A. Assessing the Trend in Rental Values of Commercial Properties Along Oyemekun. *Covenant J. Res. Built Environ. (CJRBE)* **2014**, *1*, 10–29.
12. Aron, J.; Muellbauer, J.N.J. *Some Issues in Modeling and Forecasting Inflation in South Africa*; Social Research; SSRN: Rochester, NY, USA, 2009; pp. 29–31. Available online: <https://ssrn.com/abstract=1356392> (accessed on 24 August 2023).
13. Olanrele, O.O.; Adegunle, T.O.; Fateye, O.B.; Ajayi, C.A. Causal Relationship between N-REIT's Dividend Yield and Money Market Indicators. *J. Afr. Real Estate Res.* **2019**, *4*, 71–91.
14. Munusamy, M.; Muthuveerappan, C.; Baba, M. An Overview of the Forecasting Methods Used in Real Estate Housing Price. *J. Teknol.* **2015**, *5*, 189–193. [CrossRef]
15. Jadevicius, A.; Sloan, B.; Brown, A. *Examination of Property Forecasting Models—Accuracy and Its Improvement Through Combination Forecasting*; School of Engineering and the Built Environment, Edinburgh Napier University: Edinburgh, UK, 2012; p. 20.
16. Tsolacos, S. The role of sentiment indicators for real estate market forecasting. *J. Eur. Real Estate Res.* **2012**, *5*, 109–120. [CrossRef]
17. Jadevicius, A.; Sloan, B.; Brown, A. Property Market Modelling and Forecasting: A Case for Simplicity. In Proceedings of the 20th Annual European Real Estate Society Conference, Vienna, Austria, 3–6 June 2013; pp. 3–6.
18. Buehler, M.M.; de Almeida, P.R. *Understanding the Commercial Real Estate Investment Ecosystem. An Early Warning System Prototype*; World Economic Forum: Geneva, Switzerland, 2016.
19. Chrostek, K.; Kopczewska, K. Spatial Prediction Models for Real Estate Market Analysis. *Ekonomia* **2013**, *35*, 25–43.
20. Füß, R.; Stein, M.; Zietz, J. A Regime-Switching Approach to Modeling Rental Prices of U.K. Real Estate Sectors. *Real Estate Econ.* **2012**, *40*, 317–350. [CrossRef]
21. Moolman, E.; Jordaan, J. Can Leading Business Cycle Indicators Predict the Direction of the South African Commercial Share Price Index? *S. Afr. J. Econ.* **2005**, *73*, 68–78. [CrossRef]
22. Tsolacos, S.; Brooks, C. *Real Estate Modelling & Forecasting*; Cambridge University Press: New York, NY, USA, 2010.
23. Udoekanem, N.; Ighalo, J.; Sanusi, Y. Predictive Modeling of Office Rent in Selected Districts of Abuja, Nigeria. *Real Estate Manag. Valuat.* **2015**, *23*, 95–104. [CrossRef]
24. Anas, J.; Laurent, F. Detecting Cyclical Turning Points. *J. Bus. Cycle Meas. Anal.* **2004**, *2004*, 193–225. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Proceeding Paper

Extreme Characteristics of a Stochastic Non-Stationary Duffing Oscillator [†]

Samuel J. Edwards ^{1,*}, Matthew D. Collette ² and Armin W. Troesch ²

¹ Carderock Division, Naval Surface Warfare Center, West Bethesda, MD 20817, USA

² Naval Architecture and Marine Engineering, University of Michigan, 2600 Draper Dr, Ann Arbor, MI 48109, USA; mdcoll@umich.edu (M.D.C.); troesch@umich.edu (A.W.T.)

* Correspondence: samuel.j.edwards33.civ@us.navy.mil

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: Unexpected responses in dynamic systems can lead to catastrophic failures. Without full knowledge of the system, it is impossible to know whether all of the dynamics have been captured or considered. Furthermore, a large number of Monte Carlo simulations may be time-prohibitive when looking at extreme behavior. In this paper, the Matched Upcrossing Equivalent Linear System (MUELS) linearization method is applied to a series of Duffing oscillators of varying stationarities, characterized by brief excursions into domains of much larger oscillation, to test the non-linear limits of the MUELS method and the ability of the MUELS method to uncover rare dynamics. The MUELS method is a linearization scheme that searches for linear systems that have the same zero-upcrossing rate as the non-linear system of interest. These systems are then input into the Design Loads Generator (DLG) to produce an ensemble of input time series that lead to extreme linear realizations, which are then used as input into the non-linear system of interest. The MUELS method results were compared to Monte Carlo simulations in various ways including probability density functions, time series, and computational expense. It was found that the MUELS method recovers extreme behavior with relative success, seeing more accurate results for more stationary systems. The current work suggests that improvements to return period estimation and equivalent linear system parameter fidelity could produce even more accurate results.

Keywords: extreme events; non-stationary; stochastic processes; Duffing oscillator

Citation: Edwards, S.J.; Collette, M.D.; Troesch, A.W. Extreme Characteristics of a Stochastic Non-Stationary Duffing Oscillator. *Eng. Proc.* **2023**, *39*, 102. <https://doi.org/10.3390/engproc2023039102>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 21 September 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Often times in an ocean environment, the extreme responses of ships and other structures can be different than expected. Running simulations and tests does not always reveal the behavior that appears in these scenarios. Engineers designing systems that contain unknown dynamical properties, such as a domain of attraction, orders of magnitude larger than the ordinary motion would benefit from a method that could identify the presence of these very dynamics. Specifically, this paper will focus on predicting rare behavior of stochastically forced non-stationary systems containing multiple attractors.

The current landscape of extreme value prediction techniques is vast but not particularly suited to this problem. Generally, extreme value theory [1] is a solid foundation to start rare event analysis. Strictly speaking, the extreme characteristics of ocean processes cannot be viewed as time series using extreme value theory due to dependence between peaks. As such, [2] discusses extreme value theory as related to stochastic processes taking into account dependence between peaks and changes in parameters over time. The aforementioned paper focuses mainly on stationary processes, so while it provides a good starting point, the derivations made and theories stated are not directly applicable to non-stationary processes.

Extreme value theory is applied to both Gaussian and non-Gaussian dynamic systems in [3] to calculate reliability. Both generalized Pareto via peaks-over-threshold and generalized extreme value distributions were fit to small sets of data to estimate the probability of failure. In general, the two distribution types seemed to extrapolate Monte Carlo results with some levels of pre-processing involved within reason. That being said, the method relies heavily on the samples used and could break down if there are unknown dynamics or the process tends to non-stationarity. The shortcomings of the generalized Pareto via peaks-over-threshold are discussed in more detail in [4] using a marine dynamics viewpoint.

A further investigation of rare events of non-linear systems was performed in [5]. The extreme characteristics of a piecewise linear oscillator was studied by investigating the tail of the response under various circumstances. The behavior of the tail was found to be dependent on various factors but was more or less defined under the set of circumstances examined in this paper. While [5] provides an excellent derivation and study, it is limited in that the solution is specific to the model and the results are not necessarily usable outside of a piecewise linear oscillator.

Another direction that can be taken is through linearization. The basic idea of linearization is to find a linear surrogate for a non-linear process, generally with the same root mean square, so that linear analysis can be used. In [6], multiple non-linear systems were linearized using a novel approach involving harmonic averaging and statistical linearization for a system that is both deterministically and stochastically forced. The authors were able to recover the magnitude of the response spectra and the average mean square value quite well. However, insights into the transfer function phase relationships and time series comparisons would be helpful for any extreme value analysis. This paper provides a solid resource for linearization, but it would be of academic and design interest to compare time series of the responses as well as extreme characteristics.

Another well-used method for non-linear extreme characteristic study is the First Order Reliability Method (FORM). The basic idea of FORM is to find the most probable realization of an input that results in the response level of interest. In [7], FORM was used to predict statistical features of parametric roll (parametric roll is a phenomenon that occurs when a ship is (generally) perpendicular to a wave train and the relationship between wavelength and the length of a vessel reaches a certain point, resulting in extremely large rolling motions). FORM was able to capture the rarity levels of extreme roll motions as compared to Monte Carlo simulations rather well, especially when taking into account the multiple sets of most probable input realizations that lead to the response level of interest and after implementing different optimization algorithms. With FORM and in [7], the response level needs to be indicated. In situations where the response levels are unknown, FORM would not be able to efficiently flesh out the dynamics of the system.

One of the major building blocks for the method that will be used in this paper is the Design Loads Generator (DLG). The DLG is a tool that provides extreme realizations of linear systems using modified phase distribution and the asymptotic nature of extreme value theory [8]. To produce these extreme realizations, an input spectrum, transfer function, and return period of interest are input into the DLG. The DLG uses a metric for the return period called the Target Extreme Value (TEV) [9], which can be described by Equation (1).

$$TEV = \sqrt{2 \ln(n)} \frac{\bar{x}}{\sigma} \quad (1)$$

where n is the number of cycles in the return period, \bar{x} is the most probable maximum response for the return period, and σ is the standard deviation of the response. Note that the equivalence between the two terms only applies if the process is Gaussian. While the DLG is generally applied to cases of Gaussian forcing and responses, it can also be used to produce realizations of extrema in a surrogate process. Not only does the DLG provide extreme realizations of the surrogate process but also the input that leads to those extremes. These inputs are valid realizations of the input spectrum and can be used to evaluate the response of a non-linear system that is related to the surrogate process used. As such,

the inputs can also be run through other degrees of freedom or responses to investigate the behavior of a system as a whole while a single degree of freedom is experiencing an extreme. The surrogate process strategy with the DLG was used in [10] to investigate the probability of failure for a stiffened ship panel under both slamming pressures and bending stresses. Using different panel configurations, the estimation of the probability of failure using the DLG compared to Monte Carlo simulations was in the same order of magnitude for each panel while taking less than 0.4% of the time. While this implementation of the DLG has been shown to produce encouraging results, it still requires knowledge of the physics behind a system. Systems with unknown dynamics, like some non-stationary systems, could not be investigated with this method as presented without knowledge of a surrogate that could represent the system of interest.

Investigating non-stationary extremes is very important to ensure the safety and proper design of any structure. That being said, without the knowledge that the system can exhibit this type of behavior due to limited data or modeling simplifications, the design problem becomes immensely difficult. Furthermore, any time series analysis regarding the response of interest or other degrees of freedom during an extreme event remains a challenge for most of the methods mentioned above. In this paper, the Matched Upcrossing Equivalent Linear System (MUELS) [11] method was used to identify rare, unknown behaviors of non-stationary systems and to produce an ensemble of extreme realizations. The MUELS method was further developed and tested in this paper by comparing extreme probability density functions and time domain results with Monte Carlo simulations. An experiment gauging the applicability of the TEV was also performed to improve the accuracy of the MUELS method results.

2. Methodology

In this section, the problem is set up and the Duffing oscillator is described. Then, a relative stationarity test is defined for the sake of comparison between each of the three systems used in this paper. An overview of the MUELS method follows along with the Monte Carlo simulation setup.

2.1. Problem Statement

To demonstrate the capability of the MUELS method to identify extreme characteristics in non-stationary systems, Duffing oscillators with fixed system parameters excited by a sea spectrum and variable forcing factor were used. The Duffing oscillator can be representative of the roll motion in ships due to the cubic stiffness term representing the non-linear restoring force. Identifying extreme characteristics of roll motions is of utmost importance due to potential capsize or damage to crew, machinery, and cargo. The equation of motion for the Duffing oscillator is as follows:

$$\ddot{x} + d\dot{x} + ax + \beta x^3 = F_s \eta(t) \quad (2)$$

where x is displacement, \dot{x} is the velocity, \ddot{x} is the acceleration, d is the linear damping, a is the linear stiffness, β is the cubic stiffness, F_s is the forcing factor, and $\eta(t)$ is a stochastic time series drawn from an ocean-wave spectrum. For a given system, the forcing factor is the primary driver in setting the level of stationarity. In this paper, a Bretschneider spectrum [12] was used with a significant wave height of 3.0 and a modal period of 2.1 s.

Thus, the Duffing oscillator is a practical and relevant model to investigate stochastic bifurcations [13]. These bifurcations generate statistics that change with time, resulting in non-stationary processes. In this paper, these bifurcations are used as a measure of stationarity and a characteristic that may or may not be known about the system.

2.2. Stationarity Tests

In this application, the weak-sense definition of stationarity is the primary focus. A weak-sense stationary process essentially has a mean that is constant in time, i.e., no trends, and a variance that does not change with time. The non-stationary systems investigated

in this paper bifurcated into two distinct domains of attraction with differing root mean square (RMS) values. As such, the stationarity tests were performed by calculating a moving RMS of each time series. By calculating the moving RMS, any excursions into the other domain of attraction were detected by counting the number of threshold upcrossings of the moving RMS. The moving RMS is a system function in *MATLAB* that calculates the RMS of overlapping, variable-length windows centered around a given point. Since all of the processes in this paper are zero-mean, the RMS is a measure of the moving standard deviation and, therefore, variance. The key parameter in the moving RMS metric is the window size, or the number of points that are included in each calculation of the RMS. For this paper, a window size of 10,000 points was selected such that extremes from a given basin did not influence the moving RMS enough to provide any misidentified excursions into the large attractor while ensuring that individual excursions could be separated from each other. Of course, there are uncertainties or expected fluctuations with estimating the moving mean and variance. To account for these uncertainties, probability distributions of the moving RMS were estimated using a Kernel Density Estimator (KDE) and the x-value at the largest magnitude peak of said distribution was considered a principal value. Using the x-value of the largest magnitude peak as the principal value is essentially taking the most probable RMS of the most represented attractor as the basis for potential stationarity. Given the fact that the moving RMS is essentially a filter and it “smooths” out excursions with window size selection, the rarity of threshold exceedances is increased even more. Therefore, a measure of Gaussian rareness was applied to set the threshold and account for any natural variations. The rareness of an event in a Gaussian process is typically normalized by the standard deviation of the process, as is mentioned in Section 1. In this paper, the threshold was set at 10 standard deviations of the moving RMS above the mean RMS for the entire time series. The moving RMS pdfs were not necessarily Gaussian, but, by using a larger number of standard deviations, the probability of non-exceedance does increase and is sufficient for this application. To determine the standard deviation of the moving RMS, the variance of a truncated pdf of the moving RMS was calculated. The truncation point of the moving RMS pdf was determined by cutting the pdf off at a point that the principal attractor was no longer represented. An example pdf of the moving RMS along with the truncation point is shown in Figure 1.

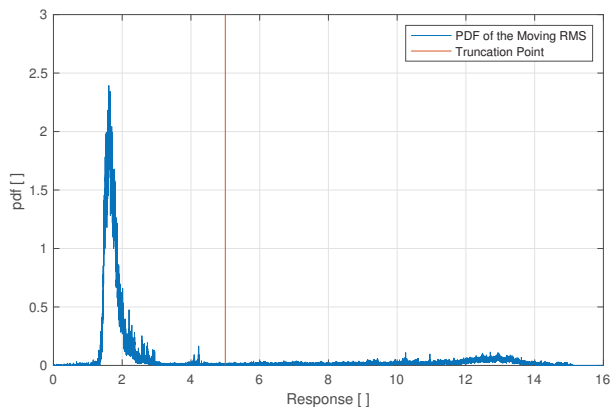


Figure 1. An example pdf showing where the truncation point was placed for estimated statistics for the dominant attractor.

It can be said with reasonable confidence that excursions above this threshold are likely the result of the RMS, and, therefore, the variance, changing with time rather than statistical uncertainty. Excursions are defined in this paper to be the amount of upcrossings of the moving RMS above the threshold. An example graph of one of these tests can be

seen in Figure 2 where a moving RMS window of 10,000 points was used and there were four excursions above the threshold.

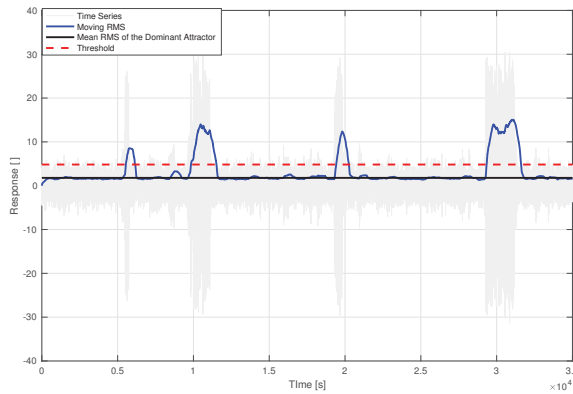


Figure 2. The moving RMS of an example Duffing oscillator compared with the threshold and the average RMS of the dominant attractor.

2.3. System Parameters

System parameter selection was performed such that there were interesting dynamics, defined here as transitions between domains of attraction, and three systems of varying non-stationarity. Table 1 lists the fixed system parameters, including the modal period, T_m , and the significant wave height, H_s , of the ITTC spectrum.

Table 1. Values for the system parameters.

Parameter	Value
d	0.02
a	1.00
β	0.04
T_m	2.10
H_s	3.00

The forcing factors were selected such that there was a system that was stationary, i.e., zero excursions in the stationarity test, a system with some non-stationarity, i.e., one or two excursions per time series, and a system with major non-stationarity, i.e., several excursions per time series. It follows that the systems with the non-stationarity feature “jump” to a larger domain of attraction. These dynamics are a result of the system parameter selection, namely, F_s and T_m . The tests discussed in Section 2.2 were used to modulate the degrees of non-stationarity. Each test was run for 10 time series of 2^{22} time steps and a time step of 0.05 s, and the number of excursions for each time series and the forcing factor were recorded and averaged. The forcing factors, threshold information, and average number of excursions are shown in Table 2. Note that fewer excursions indicate more stationary processes. Stationary processes have a very high probability of having zero excursions.

Table 2. Forcing factors selected for analysis, the standard deviation of the dominant attractor, σ_{DA} , the threshold for counting excursions, and the number of threshold exceedances.

F_s	σ_{DA}	Threshold	N_{exc}
10.0	0.85	1.06	0.0
14.7	1.36	2.58	0.8
17.0	1.78	6.41	18.2

Figures 3–5 show characteristic graphs of the stationarity tests.

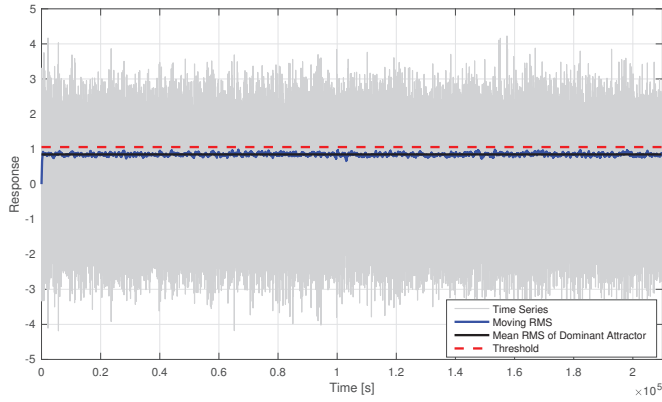


Figure 3. An example stationarity test for $F_s = 10.0$. Note that there are no excursions in this example.

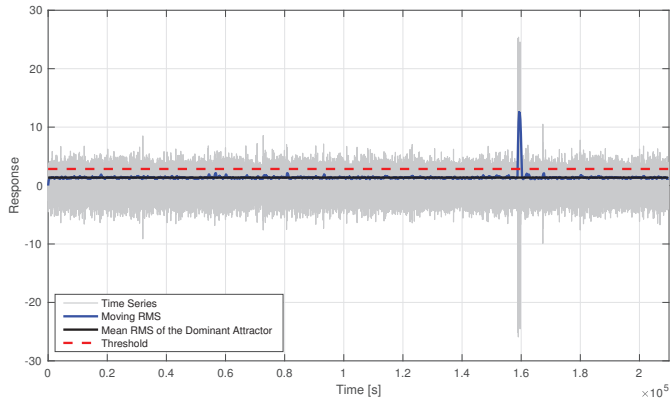


Figure 4. An example stationarity test for $F_s = 14.7$. Note that there is a single excursion in this example.

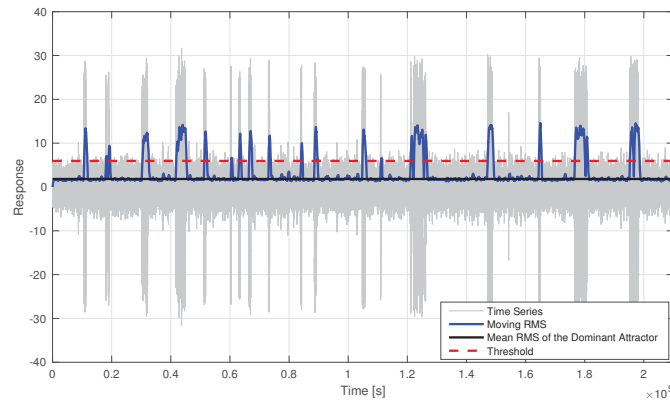


Figure 5. An example stationarity test for $F_s = 17.0$. Note that there are 19 discrete excursions in this example.

In Figures 3–5, the excursions above the given threshold increase as the forcing factor increases. The number of excursions for the $F_s = 14.7$ case ranged from zero to two

excursions in a given time series. In the $F_s = 14.7$ and $F_s = 17.0$ cases, it is clear that the variance changes with time and the processes are not stationary.

To provide a more intuitive measure of the non-stationarity, magnification curves for each system are shown in Figures 6–8 and extreme pdfs for 58-h exposure periods are in Figure 9.

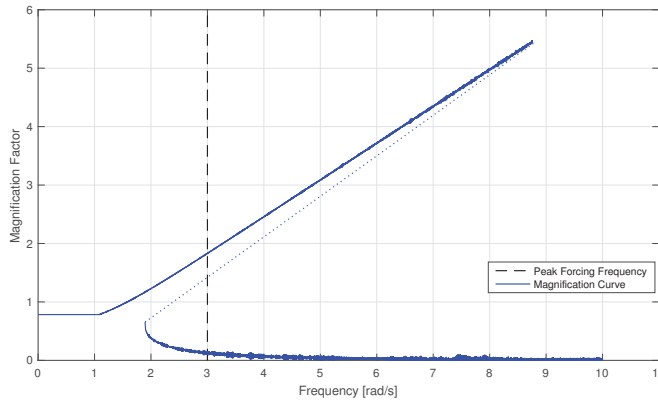


Figure 6. Magnification curve for $F_s = 10.0$ along with the peak forcing frequency. Note that the dotted line is an unstable branch.

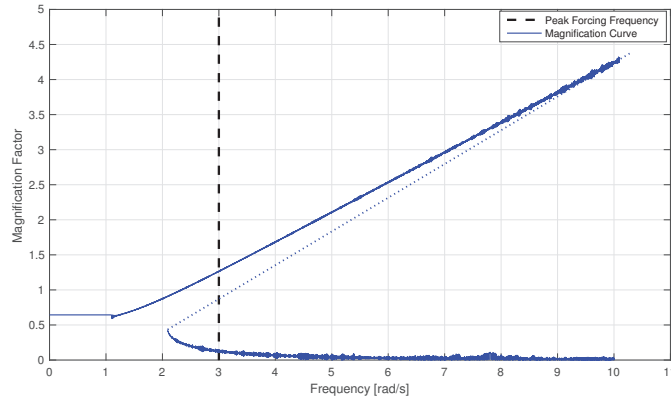


Figure 7. Magnification curve for $F_s = 14.7$ along with the peak forcing frequency. Note that the dotted line is an unstable branch.

The peak forcing frequency of 3.0 rad/s corresponds to the modal period of 2.1 s, where there are two stable responses for each forcing factor. These stable responses act as domains of attraction for the oscillator. The magnitude of the larger stable response decreases with an increasing forcing factor, which explains the increase in the frequency of excursions into the larger domain. The upper branch is generally not sustained for extended periods of time, but larger forcing factors can result in a longer duration of upper branch oscillations. Simply put, weak sense stationarity dictates that both the mean and variance remain constant in time. While the mean of each time series remains constant, it is clear that the variance would change due to the excursions into the larger domain.

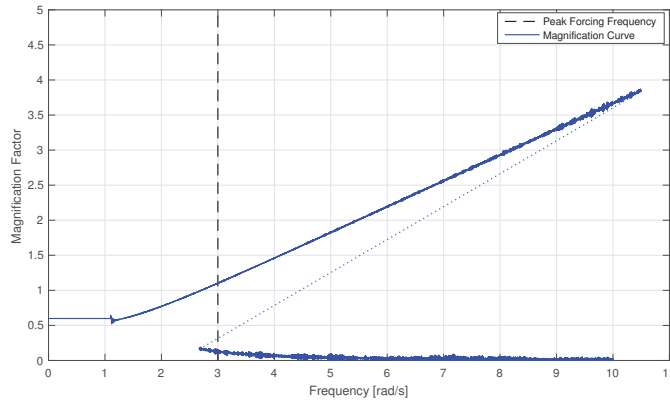


Figure 8. Magnification curve for $F_s = 17.0$ along with the peak forcing frequency. Note that the dotted line is an unstable branch.

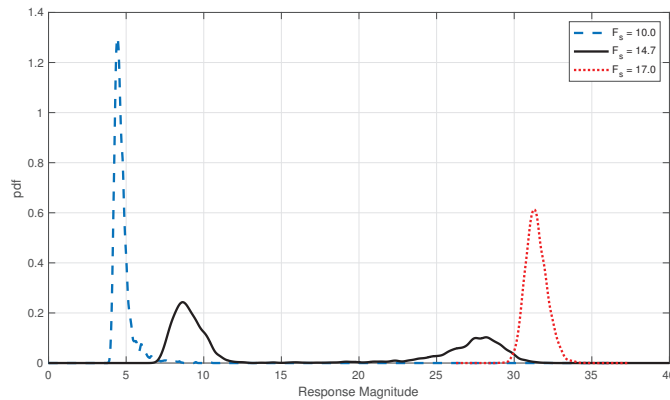


Figure 9. Kernel density estimated probability density functions for the largest value in a 58-h long time series for each forcing factor.

The three extreme PDFs for the different forcing factors give an idea of how often excursions occur. Note that these are drawn from the maximum value in each of 4000 Monte Carlo simulations of length $N = 2^{22}$ points. In the $F_s = 10.0$ case, the extreme PDF is almost entirely limited to the lower domain of attraction, while the $F_s = 14.7$ case is split between the two domains of attraction. The $F_s = 10.0$ case had five total excursions in the entire set of 4000 Monte Carlo simulations of length $N = 2^{22}$. Each time series in the $F_s = 17.0$ case had at least one excursion, and the extreme PDF reflects that.

2.4. Matched Upcrossing Equivalent Linear System (MUELS) Method

To generate extreme realizations of a non-linear system such as the Duffing oscillator, the MUELS method, developed in [11], was used. In [11], the authors used the MUELS method to estimate extreme characteristics for a set of stationary Duffing oscillators. The MUELS method uses linear systems with the same upcrossing frequency of the non-linear system of interest as surrogate processes to be input into the Design Loads Generator (DLG). A linearization scheme typically matches variance or RMS between the non-linear system of interest and the linearized system, as in [6]. Here, the goal is to find a linear system with, on average, the same number of peaks (note that a peak here implies a maximum between zero-upcrossings) as the non-linear system of interest. The linear systems used in

this paper consist of two parameters: the damping ratio, ζ , and linear natural frequency, ω_n , and are set up as in Equation (3).

$$\ddot{x}(t) + 2\omega_n\zeta\dot{x}(t) + \omega_n^2x(t) = F(t) \quad (3)$$

where $x(t)$ represents the response, $\dot{x}(t)$ is the velocity, $\ddot{x}(t)$ is the acceleration, and $F(t)$ is the forcing. A contour of constant zero-upcrossing period (for a given input spectrum) can be generated over a field of damping ratios and linear natural frequencies from which candidate linear systems can be drawn and input as transfer functions into the DLG. The DLG provides realizations of extreme linear responses for the return period of interest and the input that led to those extreme realizations. Those input time series are a valid input into the Duffing oscillator and result in conditional extremes for the system of interest. The idea driving the MUELS method is that, for each non-linear system, there likely exists at least one linear system that shares extreme characteristics with it, namely, an input that leads to extremes. The MUELS method scans equivalent linear systems with the same average upcrossing frequency and, therefore, the same number of upcrossings in a return period in an attempt to find a linear system that can be used as a surrogate for the non-linear system of interest. The current method for selecting the surrogate is to choose the set of inputs that lead to the largest most probable maximum response in the non-linear system of interest.

The MUELS method uses the Target Extreme Value (TEV), as discussed in Section 1, as a metric for the return period. The TEV measures the rareness of Gaussian processes and does not necessary share a correlation with the rareness of non-Gaussian processes. A flowchart detailing the MUELS method is shown in Figure 10.

In this paper, the DLG was set up to produce 1000 realizations of 100 s for each MUELS run. Furthermore, 2048 frequency components were used to ensure fine enough discretization for the various linear natural frequencies and resulting transfer functions. The current method to select parameters was to choose the set that results in the extreme PDF whose peak has the largest x -value. This method was used due to the lower bound property inherent to the DLG [8].

2.5. Monte Carlo Simulations

To evaluate the MUELS method, Monte Carlo simulations (MCS) were also performed. For each system, 4000 runs of 2^{22} points with a time step of 0.05 s, or 58.3 h, were generated. The time frame of 58.3 h corresponds to a TEV of about 4.80 in each forcing factor case, with slight variations following the change in upcrossing period. The MUELS method was trained with time series of length 2^{18} , or 3.6 h, and the DLG return period was selected to match the length of the Monte Carlo simulations. For the $F_s = 14.7$ case, the excursion into the more extreme domain, around 14,000 s in Figure 4, does not always appear in the 58 h time series. In fact, in the 4000 simulations, an excursion into the larger domain occurred in 57% of the simulations. This irregularity was intentional to be representative of systems for which there is a limited amount of data and that may have unknown dynamics.

The comparison of the MCS and the MUELS method was performed using a practical approach. The computational expense for the MCS and MUELS method was compared. The desired exposure period of 58.3 h plays a role in the computational expense and the comparison would differ with a different exposure period. The extreme PDF of a non-linear process for a given exposure is useful in design but is not always easy to generate. Therefore, the extreme PDFs generated from the MCS results were compared to extreme PDFs generated from the MUELS method results using selected characteristics. While the actual magnitude of the extreme values is useful to have, the time series are also vital so that the response of other degrees of freedom during an extreme event can be observed. As such, the time series structure of the MCS and MUELS method results near extremes was also compared.

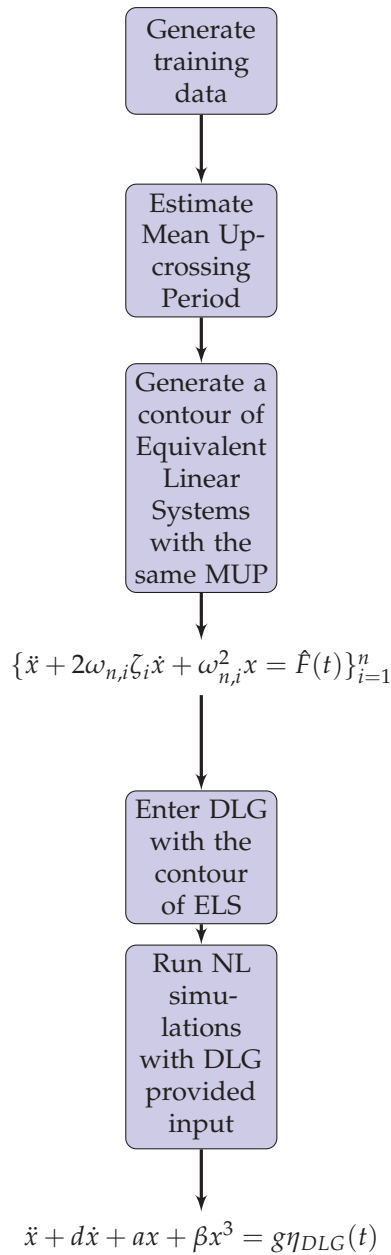


Figure 10. The Matched Upcrossing Equivalent Linear System (MUELS) method flowchart.

3. Results and Discussion

In this section, the results of the different studies are presented and discussed.

3.1. MUELS Method Performance at a Fixed TEV

For each forcing factor value, around 20 sets of parameters were input as equivalent linear systems into the DLG. While the return period for each forcing factor was the same,

the zero-upcrossing period, and therefore the TEV, changed. Figures 11–13 show the contours for each forcing factor.

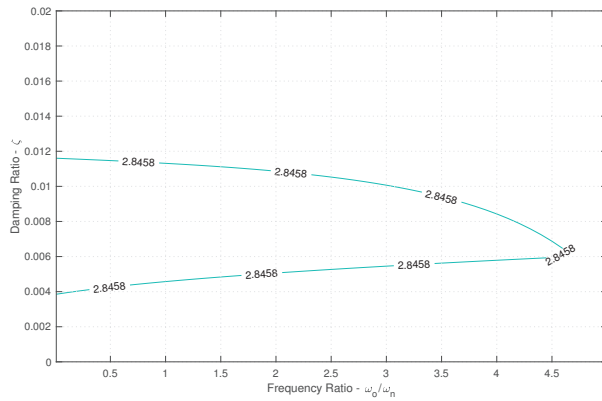


Figure 11. The equivalent linear system contour for $F_s = 10.0$ along with the zero-upcrossing frequency of 2.8458 rad/s. Note that ω_0 is the peak frequency of the input spectrum and ω_n is the linear natural frequency.

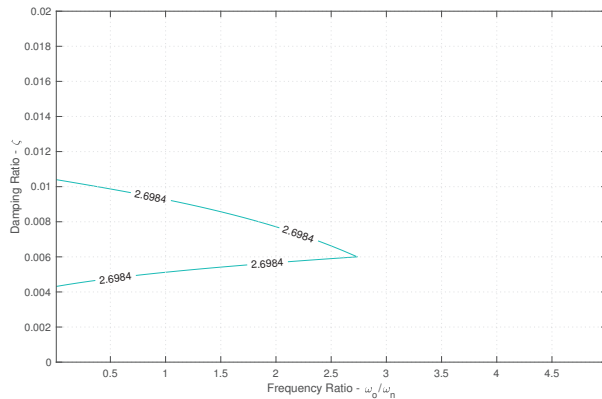


Figure 12. The equivalent linear system contour for $F_s = 14.7$ along with the zero-upcrossing frequency of 2.6984 rad/s. Note that ω_0 is the peak frequency of the input spectrum and ω_n is the linear natural frequency.

As seen in Figures 11–13, increasing the forcing factor shifts the contour to the left. As the Duffing oscillators become more and more non-linear and non-stationary due to the increased forcing factor, there are fewer equivalent linear systems available to represent the Duffing oscillators. As such, the probability that there exists a linear system that shares inputs that lead to extremes with the non-linear system of interest decreases. The parameters from these contours are sampled such that about 20 sets of parameters were selected for input into the DLG for the purpose of simplicity and speed. Furthermore, the bulk of these sets of parameters fall near the bend in the contours, at frequency ratio values above 1.0. The majority of resulting natural frequencies fall below 1.0 rad/s, which may have an effect on the performance of the MUELS method due to the distance between the ELS natural frequencies and the peak forcing frequency. While it is possible that increasing this discretization, i.e., using more parameter sets from around the contour, would increase

accuracy and performance, only around 20 parameter sets from each contour were used for this paper.

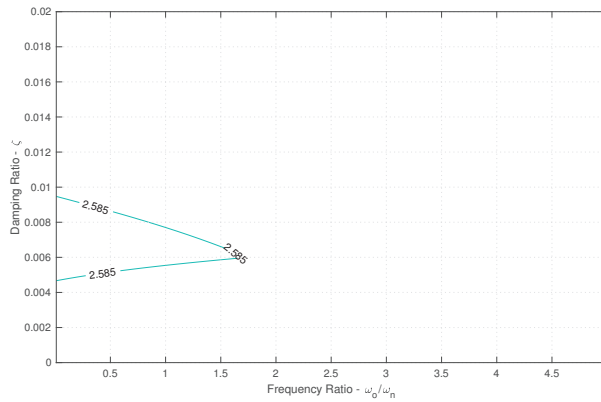


Figure 13. The equivalent linear system contour for $F_s = 17.0$ along with the zero-upcrossing frequency of 2.5850 rad/s. Note that ω_0 is the peak frequency of the input spectrum and ω_n is the linear natural frequency.

Table 3 outlines the TEV and selected parameters for each forcing factor. The parameter selection process is detailed in Section 2.4.

Table 3. The TEV for the given return period and the selected linear natural frequencies, ω_n , and damping ratios, ζ , for each forcing factor.

F_s	TEV	$\omega_{n,sel}$	ζ_{sel}
10.0	4.793	0.059	0.006
14.7	4.774	0.196	0.009
17.0	4.761	0.148	0.006

The linear natural frequencies and resulting transfer functions selected have little overlap with the energy from the input spectrum. Further investigations into the importance of prioritizing systems whose transfer functions overlap more with the input spectrum will be considered in future work.

One of the major benefits of the MUELS method is the increase in computational efficiency compared to Monte Carlo simulations. In this application, a single MUELS running for each forcing factor, including gathering training data and producing 1000 realizations, took 14,705 s on a quad-core processor. To produce 4000 Monte Carlo simulations for the same return period of 58 h took 144,840 s on eight cores. While there were more MCS produced, generating an equivalent number of MUELS realizations would add around 900 s per parameter set, or about 18,000 s for an entire MUELS run.

The current configuration of MUELS, which takes about 10–15% of the time of Monte Carlo simulations, allows for some increase in fidelity at the cost of computational effort. One area that could improve the accuracy of the MUELS method would be, as mentioned earlier, a finer discretization of the contour to examine more parameter sets.

Figures 14–16 show the selected MUELS extreme PDF and the extreme Monte Carlo PDF for each forcing factor. Note that each PDF was generated using a kernel density estimator.

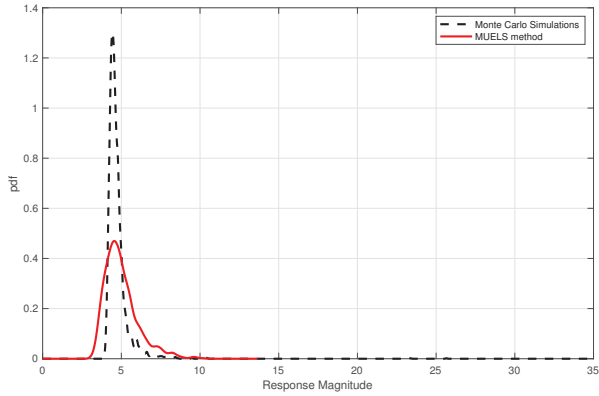


Figure 14. The extreme value PDF for the Monte Carlo simulations and the selected extreme value distribution for the MUELS method for $F_s = 10.0$.

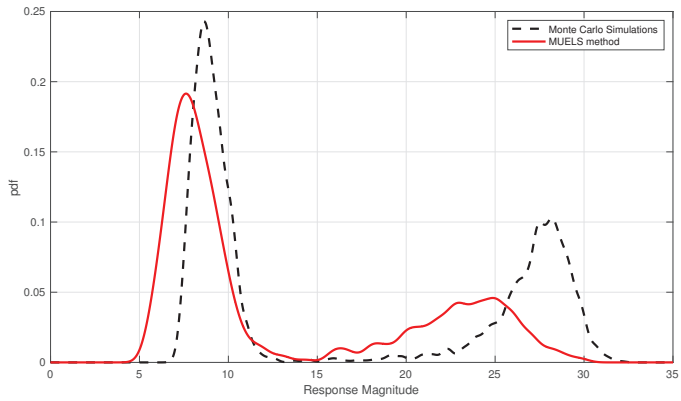


Figure 15. The extreme value PDF for the Monte Carlo simulations and the selected extreme value distribution for the MUELS method for $F_s = 14.7$.

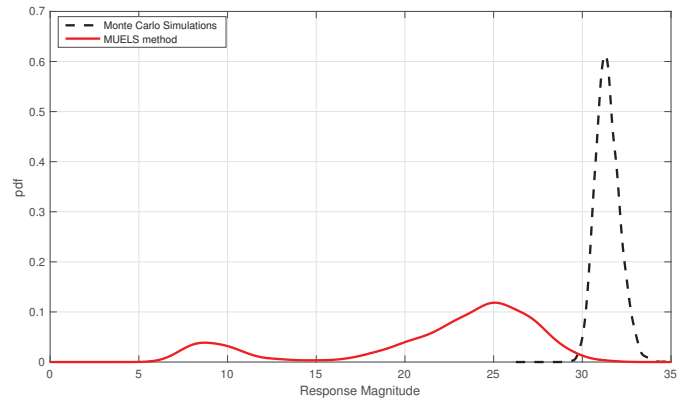


Figure 16. The extreme value PDF for the Monte Carlo simulations and the selected extreme value distribution for the MUELS method for $F_s = 17.0$.

In the $F_s = 10.0$ case in Figure 14, the MUELS method extreme PDF predicted the most probable maximum of the Monte Carlo simulations well. The MUELS method PDF has a larger standard deviation than the MCS PDF but has a large amount of overlap and, therefore, valid extreme realizations.

The MUELS method was able to recover the two attractors in the $F_s = 14.7$ case successfully. The under-prediction here could be the result of the TEV, given the levels of non-linearity that were introduced or since there are now essentially two return periods to examine: that of the small attractor and that of the large attractor. While the MUELS method does under-predict the MCS in the most probable maxima of both attractors, there is still a good amount of overlap that can provide valid extreme realizations.

In the $F_s = 17.0$ case, the MUELS method retained some realizations that did not contain excursions. Furthermore, the amount of overlap between the MUELS method PDF and the Monte Carlo PDF is reduced even more.

The immediately evident and important characteristic of the $F_s = 14.7$ PDF is the bi-modality, while the $F_s = 10.0$ and $F_s = 17.0$ cases exhibit uni-modality in the smaller domain of attraction and larger domain of attraction, respectively. The most obvious comparison we can make between the MCS and MUELS method is the x-value location of the peaks and the area of each of the peaks. It should be reiterated that each peak is representative of a different domain of attraction, as indicated in Section 2.2. As such, the area and the x-value of the maximum of each peak were used to compare the MUELS method with the Monte Carlo simulations. Table 4 shows the specified characteristics of the extreme MCS and MUELS PDFs and the mean absolute percentage error between the two.

Table 4. Comparison of pertinent PDF characteristics between the MUELS method and Monte Carlo simulations. The mean absolute percentage error (MAPE) between the MUELS method and MCS is also shown. Note that, for $F_s = 10.0$ and $F_s = 17.0$, there was only one attractor in the Monte Carlo simulations and, therefore, only one peak to compare.

Characteristic	$F_s = 10.0$			$F_s = 14.7$			$F_s = 17.0$		
	MUELS	MCS	MAPE	MUELS	MCS	MAPE	MUELS	MCS	MAPE
Peak 1 X-Value	4.55	4.44	0.03	7.62	8.64	0.12	8.71	N/A	N/A
Attractor 1 Area	1.00	1.00	0.00	0.66	0.57	0.16	0.16	N/A	N/A
Peak 2 X-Value	N/A	N/A	N/A	25.02	28.19	0.11	25.52	31.29	0.18
Attractor 2 Area	N/A	N/A	N/A	0.34	0.43	0.21	0.84	1.00	0.16

There were a limited number of excursions in the $F_s = 10.0$ Monte Carlo simulations, which is not reflected in the significant figures shown. That being said, the performance of the MUELS method for $F_s = 10.0$ produced results nearly identical to MCS. This was expected, as the $F_s = 10.0$ case is nearly linear, which resulted in a closer match between the ELS and the actual oscillator. While the MUELS PDF had more variance, as seen in Figure 14, this provides a solid foundation to produce an infinite number of extreme realizations at any return period of interest.

For $F_s = 14.7$, the MUELS method under-predicts the MCS in both peak x-value and number of simulations with excursions. The under-prediction could be due to the MUELS method reaching the non-linearity limits or it could be due to the TEV selection. For this section, the TEV was determined simply by using the return period of 58.3 h and the zero-crossing period for each forcing factor. It is important to reiterate that the TEV becomes less meaningful as more non-linearity is introduced. The TEV is still a good starting point but cannot be expected to produce accurate results without any changes made to account for non-linearity.

For $F_s = 17.0$, the MUELS method under-predicted the MCS again. In fact, there were a number of DLG inputs that did not result in an excursion in the 100-second realization. The under-prediction here is most likely the result of both TEV selection and reaching the non-linear limits of the MUELS method. Despite this, the large attractor x-value of the peak fell within 20% of the MCS most probable maximum and there are a number of

realizations that overlap with the Monte Carlo extreme PDF. In practice, the amount of overlap would not be known, but schemes are being developed to form an acceptance–rejection method based on extreme value theory and knowledge of the system, which will enable one to estimate the amount of overlap between the true extreme value distribution and the extreme PDF from the MUELS method.

3.2. Time Series Comparison

One of the major benefits of the MUELS method is the ability to produce any number of time series realizations that lead to an extreme response. It should be reiterated that the difference between just running Monte Carlo simulations and the MUELS method is that the MUELS method uses the DLG to produce multiple sets of input realizations from different equivalent linear systems of relatively short length. After the equivalent linear system parameters are selected, the DLG is capable of producing many realizations for that set of linear parameters that potentially lead to extremes in the non-linear system of interest. That being said, it is important to compare the MUELS method time series with Monte Carlo simulations to ensure that the time series have the similar characteristics near extremes. The phase sampling procedure in the DLG results in input time series that lead to linear extremes at $t = 0$. Using the time series as input into the non-linear system will not necessarily result in an extreme or potential extreme at $t = 0$ and that is reflected in the ensemble average time series. The lag is more noticeable when compared to the Monte Carlo simulation ensemble average near extremes, which was set to have the extreme at $t = 0$, so the magnitudes were scaled and normalized to match the relationship between the peak value of the largest attractor for the Monte Carlo simulations and the MUELS method. Figures 17–19 show these normalized ensemble averages near extremes for the Monte Carlo simulations and the MUELS method for each forcing factor.

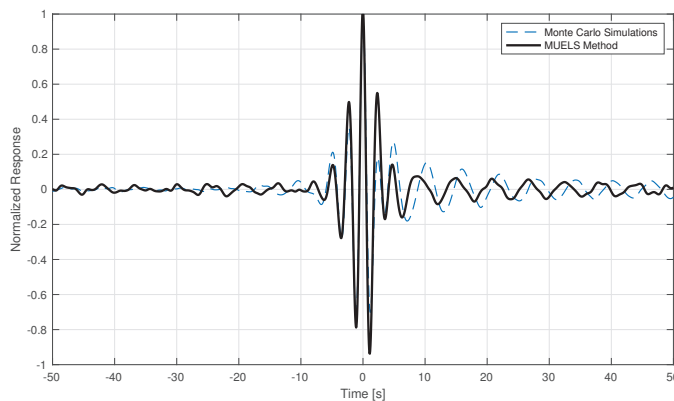


Figure 17. Ensemble average of the time series near extremes for Monte Carlo simulations and the MUELS method for $F_s = 10.0$. Note that the MUELS method results are not centered.

In the $F_s = 10.0$ case, the MUELS method and Monte Carlo simulations have very similar mean frequencies near $t = 0$ and the magnitudes of the peaks leading up to the extreme value. Since the $F_s = 10.0$ case is the most linear and, therefore, more immediately compatible with the DLG, it follows that it would produce time series that are closer to Monte Carlo simulations. It also seems to capture the dynamics shown in the Monte Carlo simulations further away from the extreme.

In the $F_s = 14.7$ case, the MUELS method ensemble average seems to have a lower characteristic frequency than the Monte Carlo simulations. This may be a result of the lag mentioned earlier as the zero-upcrossing period should remain constant due to the fact that the input time series are valid realizations of the input spectrum. It is also interesting to note that the minimum value of the MUELS method after the positive peak follows the

behavior of the Monte Carlo simulations while having a larger magnitude than the positive maximum of the MUELS method.

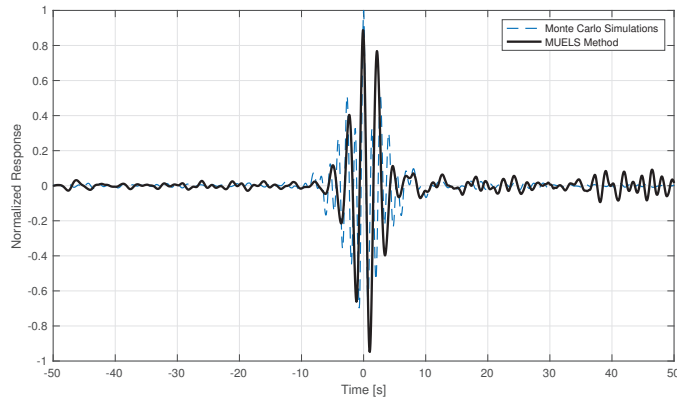


Figure 18. Ensemble average of the time series near extremes for Monte Carlo simulations and the MUELS method for $F_s = 14.7$. Note that the MUELS method results are not centered.

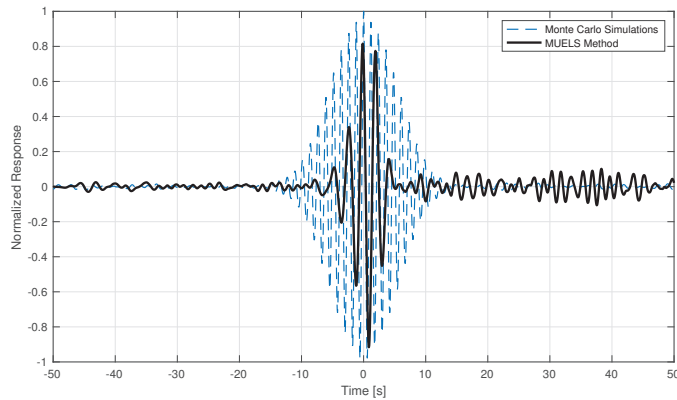


Figure 19. Ensemble average of the time series near extremes for Monte Carlo simulations and the MUELS method for $F_s = 17.0$. Note that the MUELS method results are not centered.

In the $F_s = 17.0$ case, the MUELS method again has a lower characteristic frequency than the MUELS method. The buildup to the maximum is not as gradual or symmetric, as shown in the Monte Carlo ensemble average, but again re-centering the MUELS time series would reduce some of these deviations.

A future comparison between the MUELS method and Monte Carlo simulations would center the MUELS ensemble average to have a clearer comparison between the magnitudes of the ensemble average between the MCS and MUELS method. While the re-centering would improve the MUELS method performance relative to the Monte Carlo simulations, there may be another point of improvement in the TEV selection.

4. Conclusions

In this paper, the abilities and the limits of the MUELS method were tested. Three systems of varying non-linearity and non-stationarity were used to compare the MUELS method with the conventional method of Monte Carlo simulations. The key characteristic in each of the systems was the number of excursions into a domain of attraction with

peak magnitudes two to three times larger than the base domain of attraction's peaks. In general, the MUELS method under-predicted extreme characteristics found using Monte Carlo simulations but remained within about 20%. That being said, the computational expense of the MUELS method was only 10–15% of the Monte Carlo simulations on a less computationally powerful setup. The reduced load could allow for a larger number of potential surrogate linear systems for the MUELS method to test.

One of the major benefits of the MUELS method is the ability to produce time series realizations of conditional extremes. In comparing the ensemble average of the MUELS method and Monte Carlo simulations near extremes, it was found that there was a degradation in accuracy as non-linearity increased. One main cause of this is likely the fact that, while the DLG produces extreme linear time series with a maximum at $t = 0$, there is no basis for those inputs to provide a non-linear realization with a maximum at exactly $t = 0$. Additionally, a centering of the maximum values before taking the ensemble average would certainly improve both the ensemble average magnitude and average period when compared to Monte Carlo simulations.

Future studies into using alternative TEVs to minimize the distance between the MUELS method extreme PDF, the Monte Carlo simulations, and finer discretized parameter contours could potentially improve the MUELS method performance.

Author Contributions: All sections aside from supervision: S.J.E.; Supervision and writing—review and editing: M.D.C. and A.W.T. All authors have read and agreed to the published version of the manuscript.

Funding: The work in this paper was supported by the Office of Naval Research (ONR) through the “Probabilistic Assessment of Design Events for Complex Systems Subject to Stochastic Input” project (program manager Kelly Cooper) and with Government support under and awarded by the Science, Mathematics and Research for Transformation (SMART) SEED grant established by the Department of Defense (DoD).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gumbel, E. *Statistics of Extremes*; Echo Point Books and Media: Brattleboro, VT, USA, 1958.
2. Leadbetter, M.; Rootzen, H. External Theory for Stochastic Processes. *Ann. Probab.* **1988**, *16*, 431–478. [CrossRef]
3. Grigoriu, M.; Samorodnitsky, G. Reliability of Dynamic Systems in Random Environment by Extreme Value Theory. *Prob. Eng. Mech.* **2014**, *38*, 54–69. [CrossRef]
4. Pipiras, V. Pitfalls of data-driven peaks-over-threshold analysis: Perspectives from extreme ship motions. *Prob. Eng. Mech.* **2020**, *60*, 103053. [CrossRef]
5. Belenky, V.; Glotzer, D.; Pipiras, V.; Sapsis, T. Distribution Tail Structure and Extreme Value Analysis of Constrained Piecewise Linear Oscillators. *Prob. Eng. Mech.* **2019**, *57*, 1–13. [CrossRef]
6. Zhang, Y.; Spanos, P. A Linearization Scheme for Vibrations due to Combined Deterministic and Stochastic Loads. *Prob. Eng. Mech.* **2020**, *60*, 103028. [CrossRef]
7. Jensen, J.; Choi, J.; Nielsen, U. Statistical Prediction of Parametric Roll using FORM. *Ocean Eng.* **2017**, *144*, 235–242. [CrossRef]
8. Kim, D. *Design Loads Generator: Estimation of Extreme Environmental Loadings for Ship and Offshore Applications*; University of Michigan: Ann Arbor, MI, USA, 2012.
9. Ochi, M. *Applied Probability and Stochastic Processes*; Wiley-Interscience: Hoboken, NJ, USA, 1990.
10. Seyffert, H.; Troesch, A.; Collette, M. Combined Stochastic Lateral and In-Plane Loading of a Stiffened Ship Panel Leading to Collapse. *Marine Struct.* **2019**, *67*, 102620. [CrossRef]
11. Edwards, S.; Troesch, A.; Collette, M. Estimating Extreme Characteristics of Stochastic Non-Linear Systems. *Ocean Eng.* **2021**, *225*, 109042. [CrossRef]

12. Det Norske Veritas. DNV-RP-C205 Environmental Conditions and Environmental Loads—Recommended Practice. 2010. Available online: <https://www.dnv.com/oilgas/download/dnv-rp-c205-environmental-conditions-and-environmental-loads.html> (accessed on 20 September 2023).
13. Namachchivaya, N. Stochastic Bifurcation. *App. Math. Comp.* **1990**, *39*, 101–159. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Proceeding Paper

Treatment and Analysis of Multiparametric Time Series from a Seismogeodetic System for Tectonic Monitoring of the Gulf of Cadiz, Spain [†]

Javier Ramírez-Zelaya ^{1,*}, Vanessa Jiménez ², Paola Barba ¹, Belén Rosado ¹, Jorge Gárate ¹ and Manuel Berrocoso ¹

¹ Laboratory of Astronomy, Geodesy and Cartography, Department of Mathematics, Faculty of Sciences, University of Cadiz, 11510 Puerto Real, Spain; paola.barba@uca.es (P.B.); belen.rosado@uca.es (B.R.); jorge.garate@uca.es (J.G.); manuel.berrocoso@uca.es (M.B.)

² Department of Theoretical Physics and the Cosmos, University of Granada, 18010 Granada, Spain; vanessa.jimenezmorales@hotmail.com

* Correspondence: javierantonio.ramirez@uca.es

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: The tectonic activity produced by the interaction between the Eurasian and African plates continually generates high seismic activity and the possibility of tsunamis occurring in the Gulf of Cadiz, Spain. The occurrence of these phenomena and the associated threat implies the need to implement a seismogeodesic system made up of a GNSS receiver, a seismograph–accelerograph, and an inclinometer that allows for us to study the behavior of tectonic activity in the Gulf and adjacent areas. This system is installed in the Doñana biological station, Huelva, Spain, and sends continuous records to the control center located in the University of Cadiz, generating GNSS, seismic, accelerometric, and inclinometric time series, which, together with the implementation of geodetic and geophysical techniques, is capable of providing information on tectonic activity immediately. In this manuscript, the time series generated by the system have been analyzed, in addition to a specific seismic event that occurred in the study area.

Keywords: Tectonic Monitoring; seismogeodetic systems; GNSS-GPS time series analysis; seismic hazard; geodynamic monitoring; tectonic surveillance

Citation: Ramírez-Zelaya, J.; Jiménez, V.; Barba, P.; Rosado, B.; Gárate, J.; Berrocoso, M. Treatment and Analysis of Multiparametric Time Series from a Seismogeodetic System for Tectonic Monitoring of the Gulf of Cadiz, Spain. *Eng. Proc.* **2023**, *39*, 46. <https://doi.org/10.3390/engproc2023039046>

Academic Editor: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 4 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The south of the Iberian Peninsula and North Africa region is conditioned to the great Eurasian and African plates; this region corresponds to the transition between the oceanic edge and the continental edge where the Iberian Peninsula and Africa meet in the direction of Tunisia. It includes the Betic mountain ranges, the Gulf of Cadiz, the Alboran Sea, and the northern part of Morocco, characterized by a large complex of faults giving rise to a complex tectonic evolution and moderate seismic activity as a consequence of the convergence process between the Eurasian and African plates. Additionally, opposing movements are produced due to the difference in oceanic opening velocities in the Atlantic and the structural complexity of the Alboran domain.

In the Gulf of Cadiz, seismic activity is distributed in the east–west direction along a 100 km wide band located north of the Gulf; this tectonic activity (according to the magnitude, intensity, location, depth and other characteristics of the event) leads to the possibility of tsunami occurrence in the area. The tsunami that produced the greatest natural catastrophe in Spain was recorded on 1 November 1755, as a result of an earthquake of magnitude 8.5 Mw, located about 200 km from the cape of San Vicente in the S–W direction [1,2], (Figure 1).

The consequent dangerousness of a high-magnitude earthquake and the possible associated tsunami in the Gulf of Cadiz implies the need and motivation to develop and

implement a seismogeodetic system that allows for the monitoring and surveillance of the tectonic activity in the area. This system comprises geodetic and geophysical techniques capable of providing immediate information on tectonic activity to understand, assess and minimize potential associated hazards.

The Seismogeodetic System is composed of a seismograph combined with a MEMS type-three-component (E, N, Z) accelerometer, a low-cost Tilt Data Logger, and a multi-frequency GNSS-GPS receiver (DONA). It is located in the Doñana Biological Station (EBD), in the Doñana National Park, Huelva, Spain. The main objective is to obtain a set of multiparametric time series (geodetic, seismic, acelerographic, and inclinometry) in real-time and/or deferred that, together with geodetic and geophysical techniques, can generate immediate information to monitor the tectonic activity of the Gulf of Cadiz and adjacent areas to minimize the possible associated risks. Other objectives will be the correlation between the different disciplines, time series, and results, in addition to their integration capacity in a regional EWS.

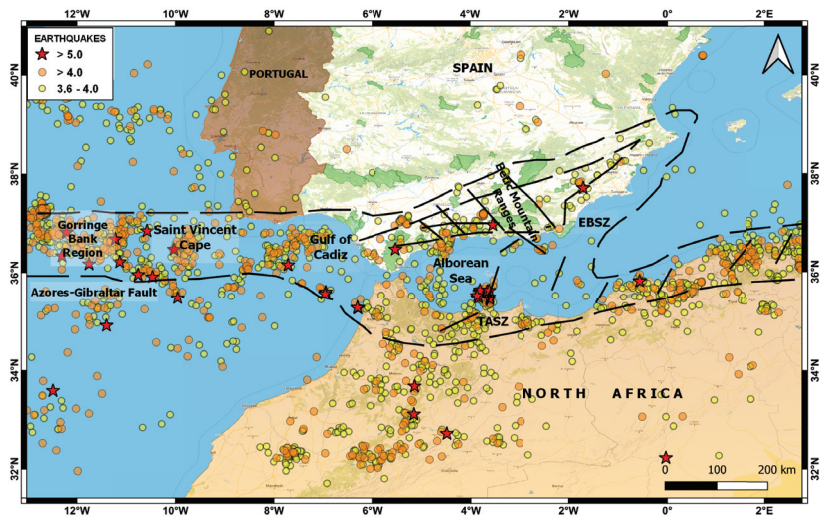


Figure 1. Map showing the geodynamic context, seismic activity (2015–2022) and main faults of the southern region of the Iberian Peninsula and North Africa. The most important faults are: Gorrige Bank Region, Gulf of Cadiz, Azores–Gibraltar Fault, Saint Vincent Cape, Alboran Sea, Betic Mountain Ranges, Eastern Betic Shear Zone (EBSZ), and Trans–Alboran Shear Zone (TASZ).

This system is complemented by a network of cGNSS stations (AYAM (“Ayamonte” Town Hall) VEJE (Public Library of “Vejer de la Frontera”), VALV (Valverde del Camino Town Hall), PGUZ (Town Hall of “Puebla de Guzmán”) and UCA1 (University of Cadiz) distributed homogeneously along the first coastline of the Gulf of Cadiz and which, like the seismogeodetic system, transmits records in real-time to the control center located in the LAGC–UCA.

This manuscript presents a description of the seismogeodesic system installed in “EBD”, the techniques used for the treatment and processing of the records, as well as an analysis of the time series generated by the system, emphasizing the GNSS–GPS records, to learn the tectonic behavior of the study area. In addition, to illustrate the scope of the system, we show the results obtained from the 4.4 Mw earthquake that occurred on 1 January 2022 in the Gulf of Cadiz.

2. Methodology

Seismogeodetic System Description (Hardware, Software and Processing Techniques)

The Seismogeodetic System is made up of the instruments: A Leica “GR30” GNSS Receiver [3], a Biaxial Digital Tilt Logger “DTL202B” [4], a Raspberry Shake “RS4D” Seismometer–accelerometer [5], a Vaisala Weather Transmitter “WXT520” [6], electrical supply equipment, and two devices (*router and switch*) for data transmission. The sensors “DTL202B” and “RS4D” are installed in a concrete chamber at 1 m depth, and the sensors “GR30” and “WXT520” were installed in a metal structure or tripod near the concrete chamber. The control center located in the LAGC–UCA is made up of three servers: a server for the virtual infrastructure “Citrix”, a main storage “NAS”, and a mirror data backup.

For the transmission and reception of the data produced by the system, a communication network was established using the following protocols: VPN, which establishes an encrypted connection over the Internet from a primary host to a destination host, and provides connection security and remote control [7]. SFTP also runs on “SSH” service and offers reliable and secure data transfer [8]. RSync service is used for automatic data storage, synchronization, and replication, and runs recursively and incrementally between two hosts [9].

For the time synchronization of the inclinometer (*DTL202B*) we used the “NTP” service, which is designed to synchronize the clocks of devices over a network connected to a time server, on a common “UTC” time base [10]. For the time synchronization of the seismic records (*seismometer “RS4D”*), we used the USB GNSS receiver “UBX-M8030” [11], which connects to different satellites to learn their position and navigation time. The data are sent from the Doñana Biological Station to the control center via the “CSIC” VPN connection, which offers greater security regarding the transmission, reception, and availability of the data. The records produced by the prototype are automatically stored on a main NAS server [12], and then distributed to the data processing, and filtering modules (*these modules are part of the virtual infrastructure of the control center*), (Figure 2).

The software used in the development and implementation of the prototype is divided into three modules (*acquisition, processing, and filtering modules*). The acquisition module manages, stores, and visualizes the data produced from the different sensors of the prototype. Seismic data generated by the “RS4D” seismograph are managed and visualized using the SWARM application [13], an open-source Java application created to visualize and analyze seismic waveforms in real-time; this can connect to different sources of static data, dynamic data, and common waveforms server: Earthworm, Winston, SeisComp, and SeedLink. GNSS observations are managed through a local data repository that facilitates data management, sharing, and data searching.

The processing module is dedicated to the treatment, quality control, and processing of the multiparametric data produced by the prototype. The software used for the seismic records is “SEISAN”: a free, multiplatform software, useful for processing the waveforms generated by the earthquakes that occurred in the Gulf of Cadiz and recorded by seismograph “RS4D” [14]. The data generated by the “DTL202B” inclinometer are processed with the proprietary “DT Logger Host” software, which also allows us to visualize the results simply and quickly.

For GNSS data processing, we use the scientific software “BERNESE”, developed by “AIUB” [15,16], and “GIPSY”, developed by “JPL” [17], both of which require a license for use and are under permanent development. The GNSS processing techniques that were used are: PPP, Relative, and Kinematic.

In the filtering module, different mathematical and statistical techniques are applied for signal processing, the correction of abnormal values, and reductions in the noise level of the time series. For this purpose, data analysis and filtering techniques are used, which are methodologically grouped into initial filters (*1- σ , 2- σ , and outliers*), analytical filters (*Kalman & Wavelets*) and statistical filters (*ARMA & ARIMA*). This filtering software was developed using multiplatform and free-to-use statistical language “R” [18].

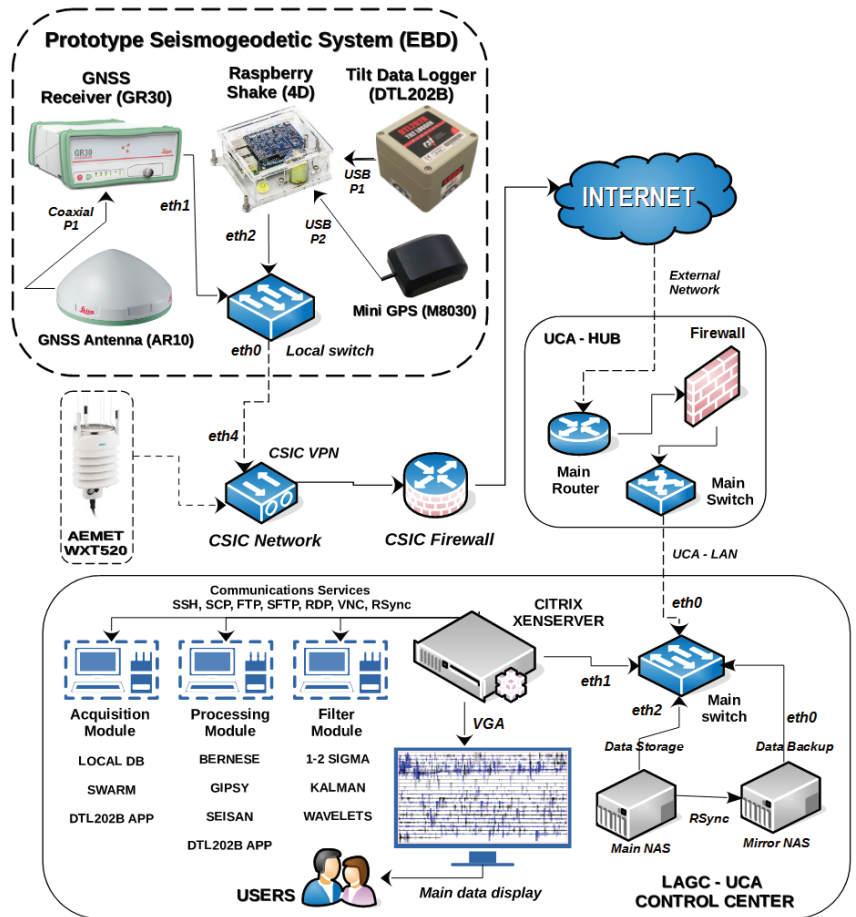


Figure 2. Figure showing the network diagram and hardware components of the prototype seismogeodetic system (communications, sensors, servers, virtual machines, NAS, mirror backup, etc.). It is divided into three parts: Prototype Seismogeodetic (Doñana Station), UCA-HUB, and Control Center (LAGC). Initially, the prototype, and the UCA-HUB are interconnected by the VPN service provided by the “CSIC”, facilitating data transmission over the Internet to the management and control center, which has a “Citrix XenServer” [19] virtual infrastructure with virtual machines that have services and applications dedicated to the automatic acquisition, processing, visualization, and filtering of data.

3. Results

Case Study: 4.4 Mw Earthquake that Occurred on 1 January 2022, in the Gulf of Cadiz

In the last two years, several earthquakes have been recorded in the Gulf of Cadiz; however, they have not been of high magnitude nor have they occurred very close to the seismogeodetic system installed in the EBD. However, a case study was included to illustrate the purpose and scope of the prototype. The earthquake analyzed in this work occurred at 21:03:49 (UTC) on 1 January 2022, of magnitude 4.4 Mw, whose epicenter was located about 130 km southwest of Doñana, Huelva, Spain (Figure 3).

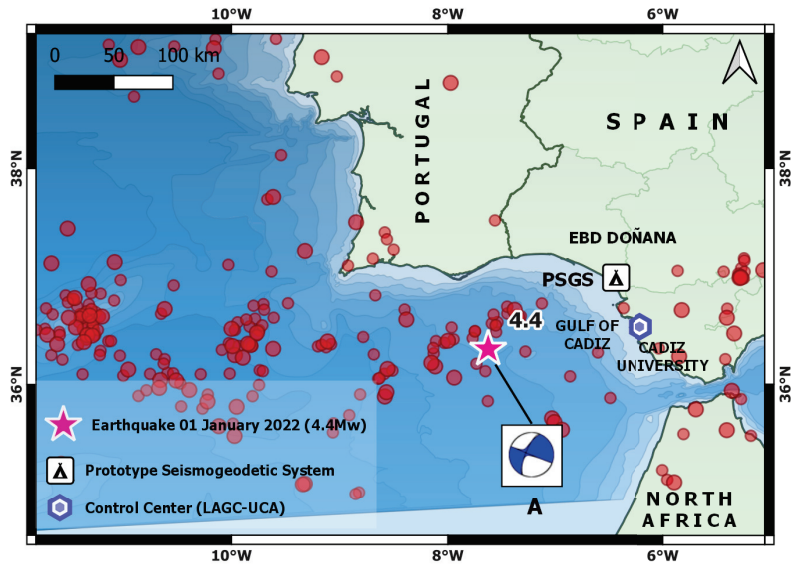


Figure 3. Map showing the location of the 4.4 Mw earthquake that occurred on 1 January 2022 in the Gulf of Cadiz ($LAT = 36.3276$; $LON = 7.6271$; $Depth 6 Km$) recorded by the prototype, the seismic events greater than 3.5 Mw occurred in the Gulf of Cadiz and surroundings between 2015 and 2022 (events taken from the public seismic catalog of IGN, Spain), and the generated focal mechanism (A).

Regarding the kinematic GNSS processing, we observe that the 3D evolution of the GNSS receiver antenna position occurred 45s after the seismic event; there is also a small but significant displacement in the “N” and “U” components. However, the component “E” shows a smaller displacement than the previous ones (Figure 4). This earthquake lacks features that allow for the production of highly significant GNSS kinematic records to be correlated with seismic, accelerometric, or inclinometry records.

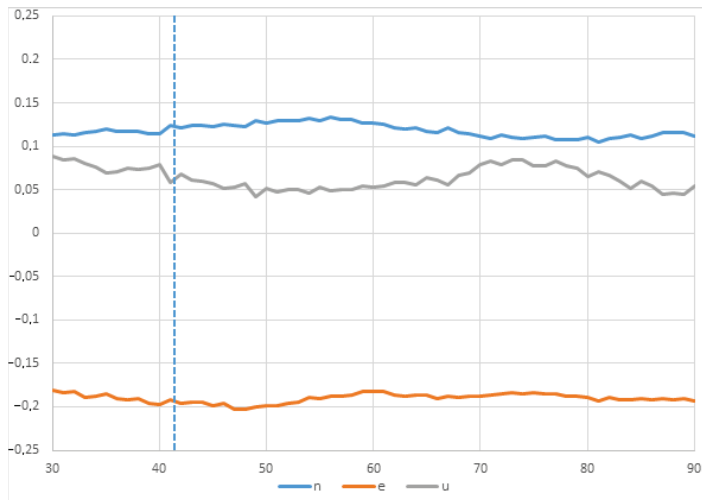


Figure 4. East, North, Up components of the GNSS time series (1Hz sample rate) for the position of the “GR30” receiver seconds after the magnitude 4.4 Mw earthquake of 1 January 2022, with epicenter about 130 km southwest of Doñana, Huelva, Spain. A small change in the trend is shown 45s (approximately) after the event occurred; this corresponds to the arrival of the seismic wave.

In the seismic signal of this earthquake, a low signal-to-noise ratio was found at certain periods of time, which allowed for the use of a first filter of 0.5 to 10 Hz and a later one of 2 to 8 Hz, (Figure 5). The study of the focal mechanism shows the following parameters: “Double pair, plane A”; average azimuth of 112° , average dip of 89° , and a slip angle of 156° ; “Double pair, plane B”; average azimuth of 203° , average dip of 86° , and slip angle of 1° . This solution presents a strike-slip faulting with NW–SE trending “P” axis, according to the NW–SE to WNW–ESE direction of Eurasian plate convergence. This mechanism is similar to previous moment tensor solutions in the Gulf of Cadiz, [20–22], (Figure 5A). In this case study, we also included the inclinometry records at the time of the seismic event on 1 January 2022 (Figure 6).

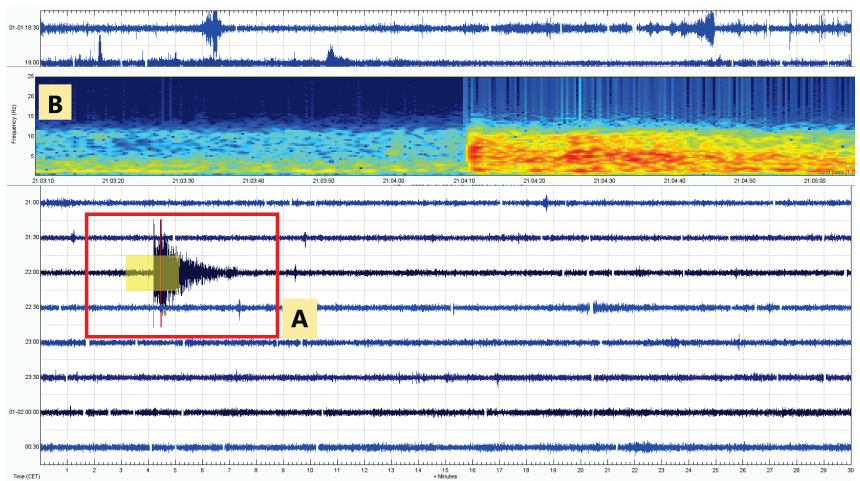


Figure 5. Figure showing the seismogram (A) and spectrogram (B) of the earthquake that occurred on 1 January 2022 at 22:03:49 (local time), using the free softwares SEISAN and SWARM, registered by the RS4D seismometer integrated in the prototype.

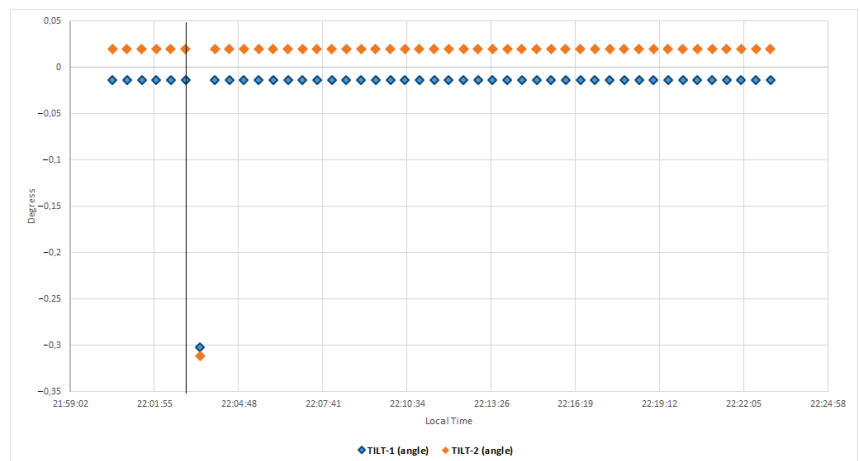


Figure 6. Figure showing the inclinometry records (30s sample rate) where the displacement produced in both sensors (Tilt 1, Tilt2) is observed, corresponding to the arrival of the seismic wave of the 4.4 Mw earthquake that occurred on January 1, 2022 in the gulf of Cadiz.

4. Conclusions

This seismogeodetic system generates multiparameter time series (seismic, accelerographic, geodetic and inclinometry) in real and deferred time, which allows for us to learn the evolution of tectonic activity in the Gulf of Cadiz and adjacent areas, as well as possible associated tsunamis. A priority of the system is the ability to provide immediate information on the tectonic activity of the Gulf of Cadiz, based on the deformation parameter and its variability (*velocity and acceleration*), in order to minimize possible risks. Another objective of the implementation of this system is the correlation of the different time series produced and their results, in addition to its integration capacity in a regional EWS.

The high-frequency GPS observations show that system GNSS is an excellent tool for measuring large displacements in areas near earthquakes, where the seismographs due to the limits in their dynamic range are saturated, impeding the correct calculation of location and magnitude, when in fact, this information is basic for the detection and rapid evaluation of the seismic event. Therefore, the seismogeodetic systems based on the integration of GNSS–GPS receptors and accelerometers complement seismic networks in moderate-magnitude earthquakes, but will be essential to the occurrence of high-magnitude earthquakes [23–25].

We analyzed the time series of the cGNSS station DONA (Figure 7), which is a fundamental part of the described seismogeodetic system. The relative processing technique was used with the cGNSS reference stations VILL and YEBE, located in the province of Madrid, Spain. Both stations belong to the international network IGS. GNSS processing was performed with the BERNESE scientific software, using the ITRF14 reference frame. The years analyzed were from 2016 to 2022; the results show the following deformation values per component (E, N, U):

$$\text{East} = 22.8 \text{ mm/y, North} = 15 \text{ mm/y, Up} = -9.1 \text{ mm/y}$$

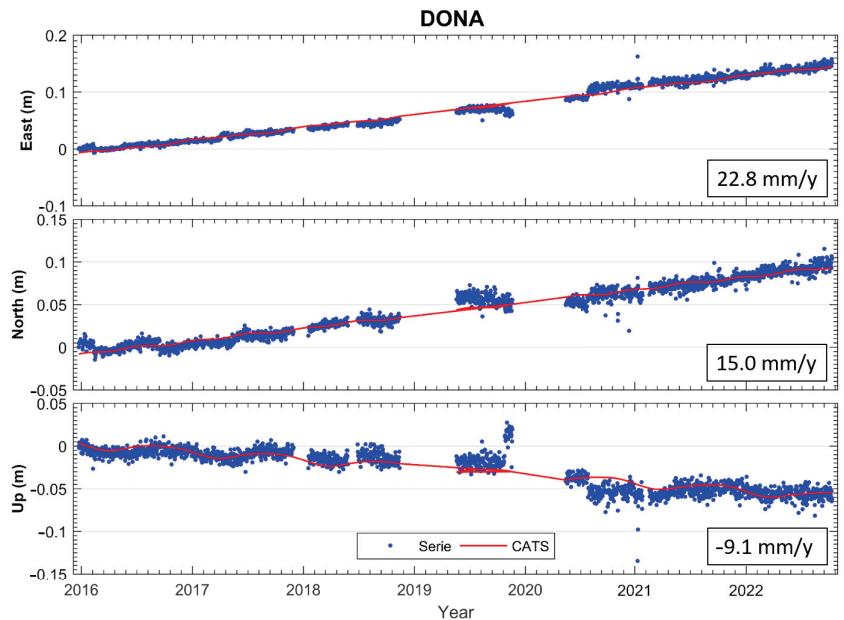


Figure 7. Figure showing the results (E, N, U) of the time series of the cGNSS station “DONA”, the GNSS processing was performed with the BERNESE scientific software using ITRF14 reference frame.

Author Contributions: Conceptualization, J.R.-Z., M.B. and J.G.; methodology, J.R.-Z., B.R. and M.B.; software, J.R.-Z., V.J. and P.B.; validation, M.B. and J.G.; formal analysis, J.R.-Z., B.R., V.J., J.G. and M.B.; investigation, J.R.-Z., V.J., J.G. and M.B.; resources, M.B.; data curation, J.R.-Z., J.G., P.B. and V.J.; writing—original draft preparation, J.R.-Z. and M.B.; writing—review and editing, J.R.-Z., M.B. and J.G.; visualization J.R.-Z. and B.R.; supervision, M.B. and J.G. All authors have read and agreed to the published version of the manuscript.

Funding: The design and implementation of the seismogeodetic system were possible thanks to the resources and funds of the astronomy, geodesy, and cartography laboratory of the University of Cadiz, directed by the principal researcher Dr. D. Manuel Berrocoso Domínguez.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The seismic data shown in this manuscript belong to the public seismic catalog of the National Geographic Institute of Spain (IGN), <https://www.ign.es/web/ign/portal/sis-catalogo-terremotos>. The data generated by the seismogeodetic system belongs to the laboratory of astronomy, geodesy and cartography of the University of Cádiz (LAGC-UCA), <https://lagc.uca.es/>.

Acknowledgments: The development and implementation of the prototype seismogeodetic system described in this manuscript has been carried out with the support of: University of Cadiz, the Spanish National Research Council (CSIC) and the working group of the Doñana Biological Station (EBD) located in the Doñana National Park, Huelva, Spain.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIUB	Astronomical Institute of the University of Bern
CATS	Create and Analyze Time Series
cGPS	Continuous Global Positioning System
CSIC	Consejo Superior de Investigaciones Científicas
EBD	Estación Biológica Doñana
EWS	Early Warning System
GPS	Global Positioning System
GNSS	Global Navigation Satellite System
IGN	Instituto Geográfico Nacional
IGS	International GNSS Services
ITRF	Internacional Terrestrial Reference Frame
JPL	Jet Propulsion Laboratory
LAGC	Laboratorio de Astronomía Geodesia y Cartografía
MEMS	Micro Electro Mechanical Systems
NAS	Network Attached Storage
NTP	Network Time Protocol
PPP	Precise Point Positioning
RSync	Remote Synchronization
UCA	Universidad de Cádiz
UTC	Universal Time Coordinated
VPN	Virtual Network Protocol

References

1. Baptista, M.A.; Miranda, J.M.; Chierici, F.; Zitellini, N. New study of the 1755 earthquake source based on multi-channel seismic survey data and tsunami modeling. *Nat. Hazards Earth Syst. Sci.* **2003**, *3*, 333–340. [CrossRef]
2. Baptista, M.A.; Miranda, J.M. Revision of the Portuguese catalog of tsunamis. *Nat. Hazards Earth Syst. Sci.* **2009**, *9*, 25–42. [CrossRef]
3. Leica Geosystems Official Website. Leica GR30 GNSS Receiver Datasheet. Available online: <https://leica-geosystems.com/en-gb/products/gnss-reference-networks/receivers/leica-gr50-and-gr30> (accessed on 11 September 2020).
4. RST Instruments Official Website. Digital Tilt Loggers Specifications. Available online: <https://rstinstruments.com/product/dtl2-01b-dtl202b-uniaxial-biaxial-digital-tilt-loggers/> (accessed on 10 November 2021).

5. Raspberry Official Website. Raspberry Shake RS4D Specifications Manual. Available online: https://manual.raspberrypi.org/_downloads/SpecificationsforRaspberryShake4DMEMSV4.pdf (accessed on 11 November 2021).
6. Vaisala Instruments Official Website. Vaisala “WKT520” User Guide. Available online: <https://www.vaisala.com/sites/default/files/documents/M210906EN-C.pdf> (accessed on 21 September 2020)
7. Cisco System Official Website. What Is a VPN?. Available online: <https://www.cisco.com/c/en/us/products/security/vpn-endpoint-security-clients/what-is-vpn.html> (accessed on 21 September 2021).
8. SSH Home Page. SFTP File Transfer Protocol-Get SFTP Client & Server. Available online: <https://www.ssh.com/academy/ssh/sftp> (accessed on 10 September 2021).
9. Rsync Official Website. Rsync Home Page. Available online: <https://rsync.samba.org/> (accessed on 11 September 2021).
10. NTP Official Website. Network Time Foundation’s NTP Support Wiki. Available online: <https://support.ntp.org/bin/view/Main/WebHome> (accessed on 1 October 2021).
11. Ublox Official Website. UBlox USB GNSS Receiver Manual. Available online: <http://bit.ly/QGPgnss> (accessed on 14 February 2021).
12. Synology Official Website. Synology Hardware Specifications. Available online: <https://www.synology.com/es-es/products/RS1221+> (accessed on 10 September 2019).
13. USGS Official Website. SWARM Home Page. Available online: <https://volcanoes.usgs.gov/software/swarm/index.shtml> (accessed on 10 January 2021).
14. SEISAN Official Website. SEISAN Home Page. Available online: <http://www.seisan.info/> (accessed on 19 January 2021).
15. Bernese GNSS Software Official Website. Bernese Software Home Page. Available online: <http://www.bernese.unibe.ch/> (accessed on 5 January 2021).
16. Dach, R.; Lutz, S.; Walser, P.; Fridez, P. *Bernese GNSS Software Version 5.2. User Manual*, Astronomical Institute, University of Bern, Bern Open Publishing: Bern, Switzerland, 2015; ISBN 978-3-906813-05-9. Available online: <https://boris.unibe.ch/id/eprint/72297> (accessed on 10 January 2021).
17. UNAVCO Official Website. Gipsy–Oasis Description. Available online: <https://www.unavco.org/software/data-processing/postprocessing/gipsy/gipsy.html> (accessed on 15 January 2021).
18. Barba, P.; Rosado, B.; Ramírez–Zelaya, J.; Berrocoso, M. Comparative Analysis of Statistical and Analytical Techniques for the Study of GNSS Geodetic Time Series. *Eng. Proc.* **2004**, *5*, 21. [CrossRef]
19. Citrix Xen Server Official Website. XenServer 7.0 Standard Edition Web Page. Available online: <https://www.citrix.com/es-es/downloads/citrix-hypervisor/product-software/xenserver-70-standard-edition.html> (accessed on 15 February 2019).
20. Stich, D.; Mancilla, F.; Morales, J. Crust–mantle coupling in the Gulf of Cadiz (SW–Iberia). *Geophys. Res. Lett.* **2005**, *32*, L13306. <https://doi.org/10.1029/2004GL020400>. [CrossRef]
21. Stich, D.; Martin, R.; Morales, J. Moment tensor inversion for Iberia–Maghreb earthquakes (2005–2008). *Tectonophysics* **2010**, *483*, 390–398. [CrossRef]
22. Martín, R.; Stich, D.; Morales, J.; Mancilla, F. Moment tensor solutions for the Iberian–Maghreb region during the IberArray deployment (2009–2013). *Tectonophysics* **2015**, *663*, 261–274. [CrossRef]
23. Larson, K. M.; Bodin, P.; Gomberg, J. Using 1–Hz GPS Data to Measure Deformations Caused by the Denali Fault Earthquake. *Science* **2003**, *300*, 1421–1424. [CrossRef]
24. Shu, Y.; Fang, R.; Geng, J.; Zhao, Q.; Liu, J. Broadband velocities and displacements from integrated GPS and accelerometer data for high–rate seismogeodesy. *Geophys. Res. Lett.* **2018**, *45*, 8939–8948. [CrossRef]
25. Bilich, A.; Cassidy, J.F.; Larson, K.M. GPS Seismology: Application to the 2002 Mw 7.9 Denali Fault Earthquake. *Bull. Seismol. Soc. Am.* **2008**, *98*, 593–606. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Proceeding Paper

Tropospheric and Ionospheric Modeling Using GNSS Time Series in Volcanic Eruptions (La Palma, 2021) [†]

Paola Barba ^{1,*}, Javier Ramírez-Zelaya ¹, Vanessa Jiménez ², Belén Rosado ¹, Elena Jaramillo ¹, Mario Moreno ¹ and Manuel Berrocoso ¹

¹ Laboratorio de Astronomía, Geodesia y Cartografía, Departamento de Matemáticas, Facultad de Ciencias, Campus de Puerto Real, Universidad de Cádiz, 11510 Puerto Real, Spain; javierantonio.ramirez@uca.es (J.R.-Z.); belen.rosado@uca.es (B.R.); elena.jaramillorosado@alum.uca.es (E.J.); mario.morenocanca@alum.uca.es (M.M.); manuel.berrocoso@uca.es (M.B.)

² Departamento de Física Teórica y del Cosmos, Facultad de Ciencias (Edificio Mecenasa), Campus de Fuentenuova, Universidad de Granada, 18010 Granada, Spain; vanessa.jimenezmorales@hotmail.com

* Correspondence: paola.barba@uca.es

[†] Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

Abstract: The signal coming from the artificial satellites of the GNSS system suffers various effects that considerably decrease the precision in solving the positioning problem. To mathematically model these effects, the atmosphere is divided into two main parts, the troposphere and the ionosphere. The troposphere can only be modelled, while the ionospheric effect can be modeled or eliminated depending on the geodetic sophistication of the receivers used. In this way, information is obtained about both layers of the atmosphere. For tropospheric modeling, the parameters of total zenithal delay (ZTD) or precipitable water vapor (PWV) will be taken, and for the ionosphere the total electron content (TEC) will be taken. In this work, statistical and analytical techniques will be applied with the R software; for example, ARMA, ARIMA models, least squares methods, wavelet functions, Kalman techniques, and CATS analysis. With this, the anomalies that occurred in the values of the ZTD and TEC in the case of the 2021 eruption of the Cumbre Vieja volcano on the island of La Palma.

Keywords: troposphere; ionosphere; ZTD; TEC; GNSS system; volcanic eruption; La Palma Island

Citation: Barba, P.; Ramírez-Zelaya, J.; Jiménez, V.; Rosado, B.; Jaramillo, E.; Moreno, M.; Berrocoso, M.

Tropospheric and Ionospheric Modeling Using GNSS Time Series in Volcanic Eruptions (La Palma, 2021). *Eng. Proc.* **2023**, *39*, 47. <https://doi.org/10.3390/engproc2023039047>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 4 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The GNSS systems composed by GPS (USA), Glonass (Russia), Galileo (European Union), and Beidou (China), in addition to solving the problem of precise positioning, provide information on the delay in the propagation of the signal as it passes through the troposphere and the ionosphere.

Most meteorological events occur in the troposphere, so the variability of the tropospheric delay depends directly on the meteorological conditions at the time of study, considering the temperature, humidity and atmospheric pressure, the PWV (Precipitable Water Vapor) value is obtained from the ZTD value. This dependence between the two parameters means that the study of tropospheric delay is related to the calculation of precipitable water vapor.

The ionosphere is characterized by its total electron content (TEC), which will affect the propagation of the GNSS-GPS signal. Ionization is caused, mainly, by ultraviolet radiation from the Sun; and with this there are maximum values of TEC during the day and minimum values at night. Solar activity is characterized through the number of sunspots produced by the Sun, solar cycles have been detected every 11 years and supercycles between 80 and 100 years; however, some works relate ionospheric disturbances with seismic and volcanic phenomena [1].

Tropospheric delay can be modeled using different tropospheric models. For the study of the ionospheric delay, the combination of the frequencies L_1 and L_2 has been used. Thus, obtaining the frequency L_4 and, by its definition, the value of the TEC. In this work, the data modeling has been carried out using the Bernese 5.2 software [2].

This work will focus on the study of the ZTD and TEC values in the island of La Palma and surroundings during the pre-eruption, eruption, and post-eruption of Cumbre Vieja volcano using different statistical and analytical techniques, such as ARMA, ARIMA, and Kalman, and STL decomposition using R 4.1.2 software.

2. Experimental Background

2.1. Geodynamic Background

La Palma is part of the volcanic archipelago of the Canary Islands that is one of the southern archipelagos of the African Atlantic border, together with Madeira, the Savage Islands, and Cape Verde [3].

The Canary Islands archipelago is located in the interior of the African Plate, presenting volcanic and tectonic activity. All these islands have been formed from volcanic eruptions; it could be said that it is a volcanically active area.

The following table shows the latest eruptions in the Canary archipelago [4]:

Year	Island	Denomination
1712	La Palma	Eruption of El Charco (Montaña Lajiones)
1730–1736	Lanzarote	Eruption of Timanfaya
1798	Tenerife	Volcano Pico Viejo (Narices del Teide)
1824	Lanzarote	Volcano de Tao, Volcano Nuevo del Fuego y Volcano nuevo
1909	Tenerife	Volcano Chinyero
1949	La Palma	Volcano Hoyo Negro, Durazanero, Llano del Banco
1971	La Palma	Volcano Teneguía
2011–2012	El Hierro	Volcano del mar de las Calmas
2021	La Palma	Volcano Cumbre Vieja

Before the eruption of the Cumbre Vieja volcano, the island of La Palma was considered an area with low seismicity; however, in 2017, seismic swarms of low intensity and at great depth began to occur. It was not until June 2021 when more and more seismic events began to be experienced on the island of La Palma, but a week before the eruption is when there were a large number of seismic swarms whose depth was decreasing, this being a powerful indication of the imminent eruption (See Figure 1).

2.2. Description of Selected Series

To carry out the study on how the eruption of the La Palma volcano has been influenced, the time series provided by the Bernese Software 5.2, corresponding to the values of ZTD and TEC, will be taken, which refer to the delay produced in the propagation of the signal from satellites to permanent GNSS stations as it passes through the atmosphere. In addition, the data will be expanded, in the case of ZTD values, making use of the GNSS stations provided by the Nevada Geodetic Laboratory (NGL).

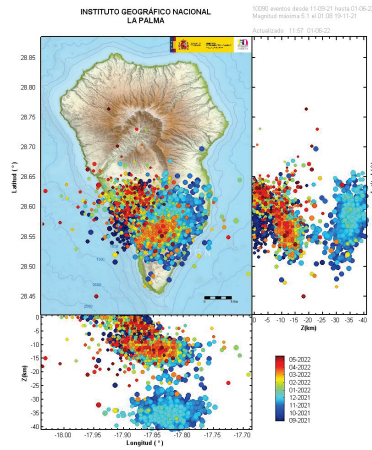


Figure 1. Earthquakes from 1 September 2021 to 1 June 2022. Image extracted from IGN.

For this purpose, GPS stations located on the different islands that make up the Canary archipelago and that belong to the MAGNET network provided by NGL will be taken [5] (See Figure 2):

- La Palma: Garafia (LPAL), Villa de Mazo (MAZO).
- La Gomera: San Sebastián de La Gomera (GOME, GOM1), Alarej6 (ALAJ).
- El Hierro: La Restinga (LRES), Frontera (FRON), Valverde (EH01).
- Tenerife: Gúímar (IZAN), Santiago del Teide (STEI), San Miguel de Abona (SNMG), Santa Cruz de Tenerife (GRAF), La Laguna (LLAG), Santa cruz de Tenerife (TN01), Puerto de la Cruz (TN02).
- Gran Canarias: La Aldea de San Nicolás (ALDE), Teror (TERR), Agüimes (AGUI), Agüineguín (ARGU), Tafira Baja (ULP2), Maspalomas (MAS1).
- Lanzarote: Haría (HRIA), Yaiza (YAIZ), Tías (TIAS), Órzo1a (LZ01), Arrecife (LZ02).
- Fuerteventura: Morro Jable (MORJ), Tarajalejo (TARA), Antigua (ANTI), La Oliva (OLIV), Puerto del Rosario (FUER), La Lajita (LAL1).

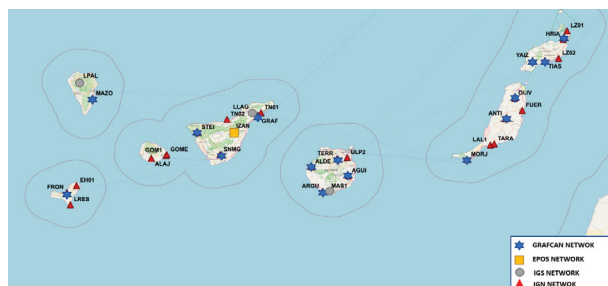


Figure 2. Permanent GPS stations of the MAGNET network.

3. Analytical and Statistical Techniques

GPS satellites have been used in this work. The GPS time series formed by the data from the ZTD and TEC values are a priori corrected when processing the data sent by the satellites to the GPS permanent stations. These values provide information on the atmospheric delay produced in the path of the GPS signal traveling from the satellites to a given GPS station.

These data usually contain outliers, so initial filtering must be used on the series obtained to eliminate them. Once the filtered series is obtained, analytical and statistical techniques can be applied to proceed to a descriptive analysis of the series.

For this study, ZTD and TEC values obtained through the Bernese 5.2 software will be taken. In addition, the number of permanent GNSS-GPS stations from which ZTD values are obtained will be increased, using the data provided by the NGL laboratory [6].

3.1. Initial Filters of the Series

The objective of any initial filtering that is applied to the GPS series consists of the elimination of data with very different values, outliers, from the rest of the series. In the case of non-linear series, this process is carried out by linear sections within the series. On the other hand, R contains a package, *forecast*, to filter time series data, that is based on the Box–Cox transform [7,8] and is completed by the *tsoutliers()* function. It is used to achieve greater linearity, homoscedasticity and a tendency towards a normal distribution of the values of the series.

3.2. Kalman

For this filtering, it is necessary to know what the dynamic linear models are like; assuming they are known, we proceed to define the Kalman filtering [9]. The Kalman filter is of a predictive–corrective type, as the parameter θ_t that determines the state of the model at time t is calculated, the estimation of the observations of the series is calculated [10]. Assuming $\theta_0 \sim N(m_0, C_0)$:

$$\theta_t = G_t\theta_{t-1} + c_t + R_tW_t$$

Furthermore, to calculate the estimate of the data of the series we will use:

$$y_t = F_t\theta_t + d_t + v_t$$

3.3. ARIMA Model

ARIMA models (integrated moving average autoregressive) are given by $ARIMA(p, d, q)$, deal with stationary time series and are made up of three models, the autoregressive (AR), the integrated (I), and the mean mobile (MA), which are defined, respectively, by p , d , and q ; uniting these three models we have the ARIMA model, which is given by

$$\phi_p(B)(1 - B)^d Y_t = \phi_0 + \theta_q(B)e_t$$

where e_t represents the errors produced at time t and Y_t the data of the series. What is more

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$$

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$$

where B is the delay operator.

3.4. ARMA Model

ARMA models, defined by $ARMA(p, q)$, deal with non-stationary series and are given by the union of autoregressive models (AR) and moving average models (MA). Therefore, joining the expressions of both models we obtain the expression of the ARMA model

$$\phi_p(B)Y_t = \phi_0 + \theta_q(B)e_t$$

where $\phi_p(B)$ and $\theta_q(B)$ are defined in the same way as in the ARIMA model.

3.5. STL Decomposition

STL decomposition (Seasonal and Trend decomposition procedure based on LOESS) additively decomposes a time series into its three components, trend, seasonality, and

irregularities [11]. The time series can contain gaps due to various factors. These do not have a negative influence on the decomposition of the time series. Local regression (LOESS) is used to estimate the three components of the series because the STL decomposition fills in the gaps in its three components. STL decomposition consists of two processes, internal and external. In the internal process, in each position the values of the trend and seasonality components are estimated and updated with the LOESS regression. In the external process, the irregularities component of the series is obtained. The trend and seasonality components are smoothed [11].

4. Zenital Total Delay and Precipitable Water Vapor Parameters

Tropospheric refraction is the delay in the signal path caused by the neutral part of the atmosphere. An electromagnetic wave always propagates in a vacuum at the speed of light, but in this case the presence of the atmosphere affects the transmission causing the waves travel slower than they would do in a vacuum. This effect, and the fact that the curvature of the trajectory should be rectilinear, are the main causes of the tropospheric delay.

The total zenith delay (ZTD) is an estimate of the delay if the signal passes through the atmosphere in the zenith direction. Multiplying this value by the appropriate mapping function provides the tropospheric delay.

For radio waves, the tropospheric delay does not depend on frequency, so its effect cannot be completely eliminated and can only be modeled. One method to model the phenomenon is based on decomposing the (ZTD) into two factors.

A first factor would be the dry component of the atmosphere (ZHD) which is responsible for 90% of the signal delay and its value is very stable. The small variations that occur are proportional to pressure changes, which makes it possible to estimate ZHD values with high accuracy.

The second factor is the wet component of the atmosphere (ZWD). Although this factor contributes less than 10% to the signal delay, the variability and instability of the atmospheric water vapor distribution is mainly responsible for the variations of the (ZTD) Thus,

$$ZTD = ZHD + ZWD. \tag{1}$$

This factor must be calculated for each measurement due to its great variability depending on the altitude, pressure and meteorological situation of the place. Following the Saastamoinen model, the ZHD value can be calculated as follows [12,13]

$$ZHD = 0.002277 \cdot \frac{P}{1 - 0.0026 \cdot \cos(2\phi) - 0.00028 \cdot h_0}, \tag{2}$$

where P is the station's atmospheric pressure (hPa), ϕ denotes the station's latitude and h_0 its respective altitude.

Using (1) the (PWV) values are calculated. Finally, the relation between the wet component of the atmosphere (ZWD) and the precipitable water vapor (PWV) is given by the following formula

$$PWV = \Pi \cdot ZWD, \tag{3}$$

$$\Pi = \frac{10^6}{\rho_w \cdot \frac{R}{m_w} \cdot \left[\frac{k_3}{T_m} + k_2 - \frac{m_w}{m_d} \cdot k_1 \right]}, \tag{4}$$

where Π is the conversion factors between the ZWD and PWV; T_m represents the average temperature in degrees Kelvin, ρ_w is the density of the liquid water, R is the universal gas constant ($R = 8314 \text{ Pa} \cdot \text{m}^3 \cdot \text{K}^{-1} \cdot \text{kmol}^{-1}$), m_w represents the molar mass of water vapor ($m_w = 18.02 \text{ kg} \cdot \text{kmol}^{-1}$), m_d represents the molar mass of the dry atmosphere ($m_d = 18.96 \text{ kg} \cdot \text{kmol}^{-1}$), y k_1 , k_2 y k_3 are the following constants, $k_1 = 77.604 \pm 0.014 \text{ K/hPa}$, $k_2 = 70.4 \text{ K/hPa}$ y $k_3 = 3.776 \pm 0.014 \times 10^5 \text{ K}^2/\text{hPa}$ [14].

5. Total Electron Content

The state of the ionosphere can be described by the electron density, n_e , which is in units of electrons per cubic meter. The impact of the ionosphere on the propagation of the signals is given by the TEC, which is called E :

$$E = \int_i^k n_e(s) ds$$

E defines the signal (s) path emitted by the satellite (i) to the receiver (k). To estimate the values of the TEC, three types of mathematical models will be defined, which are [2]:

1. Station-Specific TEC models.
2. Global TEC model.
3. Local TEC model.

In this work, the TEC values obtained from the station-specific TEC models have been used.

5.1. Station-Specific TEC Models

Station-specific TEC models are treated in exactly the same way as global models. A complete one is carried out with the set of parameters necessary to estimate the ionospheric values with respect to each station involved.

5.2. Global TEC Model

The global model for the estimation of the TEC values can also be used for regions, it is defined by

$$E(\beta, s) = \sum_{n_{max}}^{n=0} \sum_n^{m=0} \tilde{P}_{nm}(\sin\beta)(a_{nm}\cos(ms) + b_{nm}\sin(ms)),$$

where:

- n_{max} is the maximum degree of the spherical harmonic expansion.
- \tilde{P}_{nm} are the normalized associated Legendre functions of degree n and order m .
- a_{nm}, b_{nm} are the (unknown) TEC coefficients of the spherical harmonics, i.e., the global ionosphere model parameters to be estimated.
- β is the geographic latitude of the intersection point of the line receiver–satellite with the ionospheric layer.
- s is the sun–fixed longitude of the ionospheric pierce point. s is related to the local solar time (LT) according to

$$s = LT - \pi \approx UT + \lambda - \pi,$$

where UT is universal time and λ is the geographic longitude of the intersection point.

6. Application of Methodology Developed and/or Adapted R

Application of Methods

The time series obtained contain the values of the amounts of ZTD and TEC that can be seen in the atmosphere. A series of statistical and analytical techniques will be applied to these time series at different times to see what the evolution of these values has been.

The following methodology will be applied. Firstly, we will apply an initial filtering to eliminate possible outliers, then we will use the ARIMA, ARMA, and Kalman methods and the decomposition of the time series in relation to its seasonality, trend, and noise components.

To give an idea of the behavior of the series, the corresponding results of applying the methodology, previously described, for the IZAN station in the time interval 2010–2020 are visualized, thus eliminating the possible influences that have occurred in the troposphere

due to the volcano of La Palma, and, due to its location, the influence of the submarine volcano of El Hierro is also eliminated [15]. In addition, the graphs of the ZTD and PWV values corresponding to the EH01 station from September 2017 to 2018 will be shown.

The ARIMA, ARMA, and Kalman methods will be applied to the permanent stations mentioned above, thus comparing the results obtained from each of the applied techniques, and the comparison between the ZTD and TEC values during the period from August to December in the years 2018, 2019, 2020, 2021, and 2022, marking the eruptive period in each of the years, the eruption occurring in the year 2021. A translation has been applied to these data to achieve an optimal visual comparison for the different years.

7. Conclusions

This paper seeks to know the influence that the eruption of the Cumbre Vieja volcano had on the ZTD and TEC values, for which the results obtained by applying the ARMA, ARIMA, and Kalman methods at different times have been analyzed. Observing the Figures 3 and 4 shows the behavior without the volcanic influence of the IZAN station during the period 2010–2020. In both images, it can be seen that the data present a periodicity, after half the year it is seen how the time series grows, thus producing a seasonality that provides maximum points in the graph, which can be seen in the seasonal component that returns the STL decomposition. Figure 5 shows the comparison between the values of ZTD and PWV for station EH01 (El Hierro). Figure 6 shows the application of the ARMA, ARIMA and Kalman models for different stations. In the Figures 7 and 8 obtained by comparing the methods during the months of August to December in the years 2018, 2019, 2020, 2021, and 2022, the data obtained by applying the various methods have been shown. A translation has been applied to these data to achieve a better visual comparison for the different years. It can be seen, as the data corresponding to the year 2021 are slightly more linear than the rest.

You can see the STL decomposition in Figure 9 for the TEC data, in the trend component the solar cycle that occurs every 11 years [2] is observed. In addition, Figure 9 shows the STL decomposition of MAZO (La Palma), in the seasonal component it can be seen that before 2022 the TEC values are higher. In Figure 10, the application of the ARIMA, ARMA models and the Kalman technique on the filtered data is shown. In Figure 11, there is a comparison of the ARIMA model for different stations with the dates on which the active volcano was marked; a rise can be seen during this period in the TEC values.

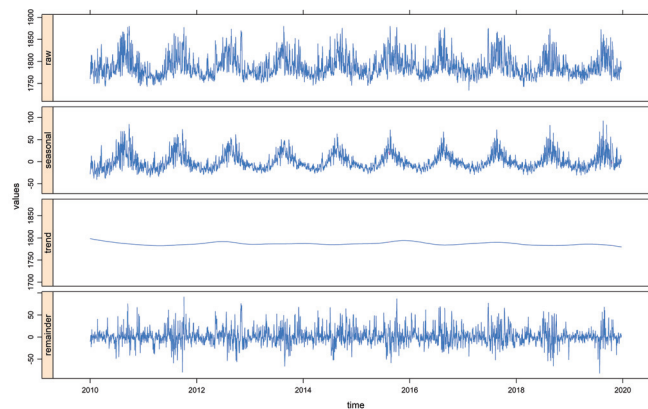


Figure 3. STL decomposition to IZAN (Tenerife) during the period 2010–2020.

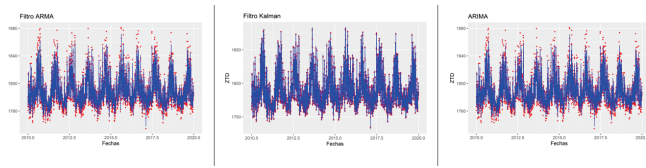


Figure 4. Application of ARMA, Kalman, and ARIMA (from left to right) on the IZAN station in the period 2010–2020. The blue graph represents the one obtained by applying the methods and the red elements are the data of the filtered series.

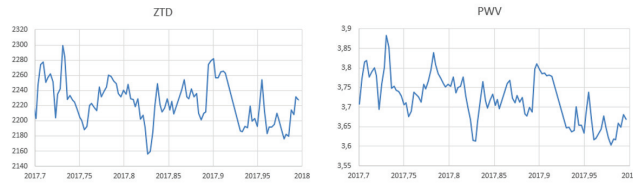


Figure 5. Comparative of ZTD and PWV values for station EH01 (El Hierro).

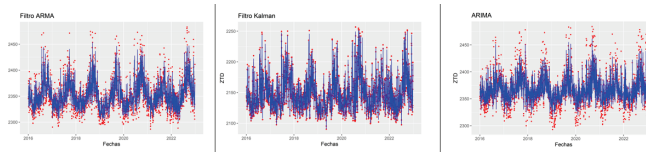


Figure 6. Application of ARMA, Kalman and ARIMA methods on ANTI (Fuerteventura), STEI (Tenerife), and FRON (El Hierro) stations, respectively.

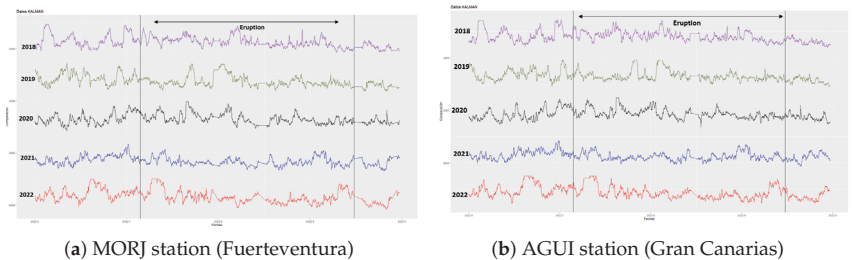


Figure 7. Kalman comparative techniques during the years 2018 to 2022. Marking as the eruption period throughout each year.

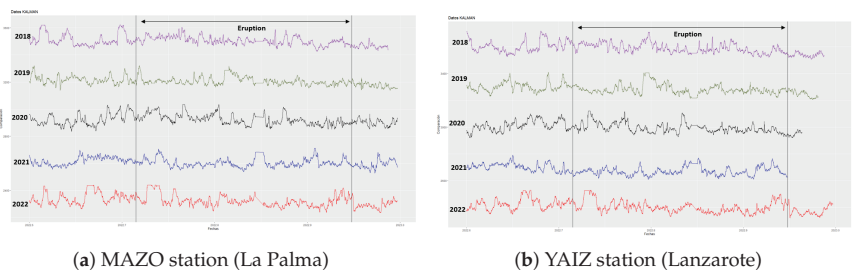


Figure 8. Kalman comparative techniques during the years 2018 to 2022. Marking as the eruption period throughout each year.

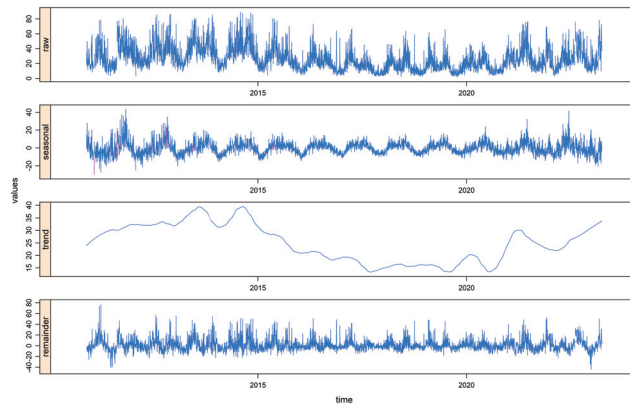


Figure 9. STL decomposition to MAZO (La Palma) during the period from 2010 to 2022.

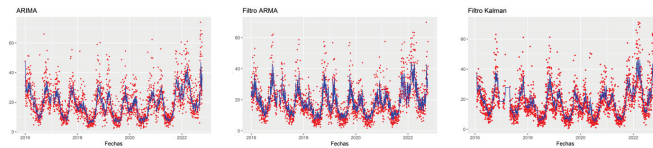


Figure 10. Application of ARIMA, ARMA and Kalman methods on ARGU, GRAF and ALAJ stations, respectively.

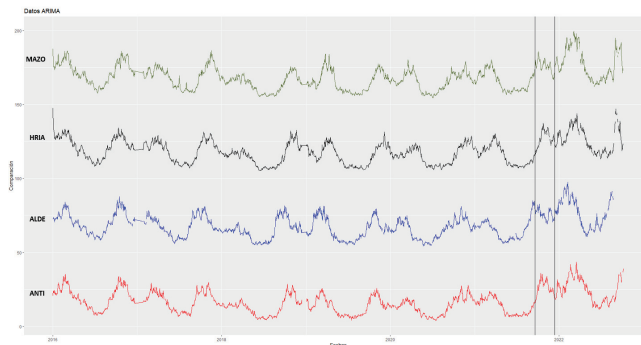


Figure 11. Comparison of the ARIMA model for the MAZO (La Palma), HRIA (Lanzarote), ALDE (Gran Canarias) and ANTI (Lanzarote) stations. Marked eruptive period.

Author Contributions: Conceptualization, P.B., V.J. and M.B.; methodology, P.B., B.R. and M.B.; software, P.B. and J.R.-Z.; validation, P.B., B.R. and M.B.; formal analysis, P.B. and M.B.; investigation, P.B. and M.B.; resources, P.B. and M.B.; data curation, P.B., E.J. and M.M.; writing—original draft preparation, P.B.; writing—review and editing, P.B. and M.B.; visualization, P.B. and M.B.; supervision, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: Funded by the “INICIA-INV” grant from the “Own Plan 2021–2022” from the University of Cádiz.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data corresponding to the tropospheric time series used are available at <http://geodesy.unr.edu/NGLStationPages/gpsnetmap/GPSNetMap.html> (accessed on 14 April 2023).

Acknowledgments: Thank the GRAFCAN Canarian network for disseminating the data and the University of Cádiz (UCA) for the financial aid “INICIA-INV” of the “Plan Propio 2021–2022”.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jin, S.; Jin, R.; Liu, X. *GNSS Atmospheric Seismology*; Springer: Berlin/Heidelberg, Germany, 2019.
2. Dach, R.; Lutz, S.; Walser, P.; Fridez, P. Bernese GNSS Software Version 5.2. User Manual, Astronomical Institute, University of Bern, Bern Open Publishing, 2015; ISBN 978-3-906813-05-9. Available online: <https://boris.unibe.ch/id/eprint/72297> (accessed on 14 April 2023).
3. González, Cárdenas, M.E.; Gosálvez, Rey, R.U.; Becerra, Ramírez, R.; Escobar, Lahoz, E. *La Erupción de Cumbre Vieja de 2021*; University of Castilla-La Mancha: Isla de La Palma, España, 2022.
4. Sevilla de Lerma, M.J. *Análisis de Series Temporales en Estaciones Permanentes GPS*. Doctoral’s Dissertation, Universidad Complutense de Madrid, Madrid, Spain, 2015.
5. Blewitt, G.; Hammond, W.C.; Kreemer, C. Harnessing the GPS data explosion for interdisciplinary science. *Eos* **2018**, *99*. [CrossRef]
6. Barba, P.; Rosado, B.; Ramírez-Zelaya, J.; Berrocoso, M. Comparative Analysis of Statistical and Analytical Techniques for the Study of GNSS Geodetic Time Series. *Eng. Proc.* **2021**, *5*, 21. [CrossRef]
7. Peña, D.; Peña, J. A normality test based on the Box-Cox transformation. *Span. Stat.* **1986**, 33–46.
8. Box, G.E.; Cox, D.R. An analysis of transformations. *J. R. Stat. Soc. Ser. (Methodol.)* **1964**, *26*, 211–243. [CrossRef]
9. Martín Rodríguez, G. Representación en el Espacio de los Estados y Filtro de Kalman en el Contexto de las Series Temporales Económicas. Documentos de Trabajo Conjuntos: Facultades de Ciencias Económicas y Empresariales, Universidad de Las Palmas y Universidad de La Laguna, DT 2002/05. **2003**, *5*, 200246. Available online: <https://acceda.riis.ulpgc.es/bitstream/10553/324/1/626.pdf> (accessed on 14 April 2023).
10. Prates, G.; García, A.; Fernández-Ros, A.; Marrero, J.M.; Ortiz, R.; Berrocoso, M. Enhancement of sub-daily positioning solutions for surface deformation surveillance at El Hierro volcano (Canary Islands, Spain). *Bull. Volcanol.* **2013**, *75*, 1–9. [CrossRef]
11. Cleaveland, R.B.; Cleaveland, W.S.; McRae, J.E.; Terpenning, I. STL: A seasonal-trend decomposition. *J. Off. Stat.* **1990**, *6*, 3–73.
12. Wu, M.; Jin, S.; Li, Z.; Cao, Y.; Ping, F.; Tang, X. High-precision GNSS PWV and its variation characteristics in China based on individual station meteorological data. *Remote Sens.* **2021**, *13*, 1296. [CrossRef]
13. Aragón Paz, J.M. Estimación de parámetros Troposféricos en Tiempo Casi Real para SUDAMérica Mediante técnicas GNSS. Doctoral Dissertation, Universidad Nacional de La Plata, Buenos Aires, Argentina, 2020.
14. Bevis, M.; Businger, S.; Herring, T.A.; Rocken, C.; Anthes, R.A.; Ware, R.H. GPS meteorology: Remote sensing of atmospheric water vapor using the Global Positioning System. *J. Geophys. Res. Atmos.* **1992**, *97*, 15787–15801. [CrossRef]
15. Rosado, B.; Ramírez-Zelaya, J.; Barba, P.; de Gil, A.; Berrocoso, M. Comparative Analysis of Non-Linear GNSS Geodetic Time Series Filtering Techniques: El Hierro Volcanic Process (2010–2014). *Eng. Proc.* **2021**, *5*, 23.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
www.mdpi.com

Engineering Proceedings Editorial Office
E-mail: engproc@mdpi.com
www.mdpi.com/journal/engproc



Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Academic Open
Access Publishing

[mdpi.com](https://www.mdpi.com)

ISBN 978-3-0365-9731-7