

Article

On Similarity Measures for Stochastic and Statistical Modeling

Konstantinos Makris ^{1,†}, Iliá Vonta ^{1,†}  and Alex Karagrigoriou ^{2,*,†} 

¹ Department of Mathematics, School of Applied Mathematical and Physical Sciences, National Technical University of Athens, GR-15780 Athens, Greece; konstantinosmakris@yahoo.gr (K.M.); vonta@math.ntua.gr (I.V.)

² Laboratory of Statistics and Data Analysis, Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, GR-83200 Samos, Greece

* Correspondence: alex.karagrigoriou@aegean.gr

† These authors contributed equally to this work.

Abstract: In this work, our goal is to present and discuss similarity techniques for ordered observations between time series and non-time dependent data. The purpose of the study was to measure whether ordered observations of data sets are displayed at or close to, the same time points for the case of time series and with the same or similar frequencies for the case of non-time dependent data sets. A simultaneous time pairing and comparison can be achieved effectively via indices, advanced indices and the associated index matrices based on statistical functions of ordered observations. Hence, in this work we review some previously defined standard indices and propose new advanced dimensionless indices and the associated index matrices which are both easily interpreted and provide efficient comparison of the series involved. Furthermore, the proposed methodology allows the analysis of data with different units of measurement as the indices presented are dimensionless. The applicability of the proposed methodology is explored through an epidemiological data set on influenza-like-illness (ILI). We finally provide a thorough discussion on all parameters involved in the proposed indices for practical purposes along with examples.



Citation: Makris, K.; Vonta, I.; Karagrigoriou, A. On Similarity Measures for Stochastic and Statistical Modeling. *Mathematics* **2021**, *9*, 840. <https://doi.org/10.3390/math9080840>

Keywords: similarity measures; time series; dimensionless indices; index matrices; multivariate indices

Academic Editor: Christophe Chesneau

Received: 27 February 2021

Accepted: 7 April 2021

Published: 12 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 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 time series analysis, in addition to trend, seasonality, periodicity and stationarity, another concept of great significance is the similarity between two or more time series. The similarity between two or more time series focuses on the study of similarities and common changes between the series. Different techniques have been proposed over the years for measuring similarities like simple mathematical measures (see, e.g., ([1,2]), data transformations (see, e.g., [3]), algorithmic methods (see, e.g., [4]) or measures of divergence (see, e.g., [5–8]). Finally, measures of dissimilarity have been thoroughly studied (see, e.g., [9–11]).

Data series and, in particular, time series analysis involve, among others, pattern matching, anomaly identification and frequent pattern detection. All these tasks are directly associated with time series similarity techniques, some of which are mentioned above. We observe frequently that data series similarity is visualization-dependent. Indeed, it is quite common, for instance, that neuroscientists manually inspect the electroencephalogram (EEG) data of their patients, using visual analysis tools, so as to identify patterns of interest (see, e.g., [12]). Surveillance systems also rely on visual tools to monitor incidence data and compare the disease behavior in various regions for the purpose of predicting or hopefully preventing an epidemic. Finally, physiologic time-series databases often require finding similar temporal patterns of physiological responses resembling those of a prototype case (see, e.g., [13]). Detection of these complex physiological patterns not only enables demarcation of important clinical events but can also elucidate hidden dynamical

structures that may be suggestive of disease processes. In all such cases, it is important to have similarity techniques, which in conjunction with visual analysis tools will enable analysts to complete their tasks quickly and accurately.

In this work, our goal is to present and discuss similarity techniques for ordered observations between time series and non-time dependent data. The coupling analysis can be obtained via a direct method ([1]) which is a simple squared measure of distance between two series.

Since we are focusing on similarities between two series, the comparison can be achieved more effectively via indices, advanced indices and the associated (index) matrices. More specifically, as we can see below, the comparison concentrates on the extreme parts of the series and the index quantifies the degree of similarity between these specific parts of the series. In general, such indices referred in this work as *MKN* indices, compare the simultaneous time pairing of the K maximum and/or minimum values (denoted by the parameter M) between N time series based on some statistical function denoted by μ . A standard *MKN* index was formally defined in [1]. Applications of such indices include the measuring of displacements, rotations, moments and forces for two types of floating wind turbines ([14]) and epidemiological data ([15]).

For details on the origin of these ideas, one may refer to a review paper by Makris et al. ([16]).

In this work, we present the evolution of the idea of the similarity indices and propose advanced dimensionless indices and the associated index matrices which are both easily interpreted and provide more effective comparison of the series involved than the one achieved by the standard indices previously proposed. The rest of the paper has been organized as follows. In Section 2, we present some preliminary definitions and review results about standard indices illustrated by an example. A generalization to the multivariate case is also presented. In Section 3, we propose new advanced dimensionless indices and index matrices for efficient comparison of time series. A discussion on the related concept of cointegration is also included. In Section 4, a dataset on influenza-like-illness (ILI) cases in Greece is examined for illustrating the usefulness and the importance of the proposed methodology. In Section 5, we discuss the parameter μ and present an application of the indices in economics and marketing defining a novel elasticity. In Section 6, we study the effect of the parameter M and in Section 7, the indices are defined for non-time dependent data and a modified direct measure for this type of data is also discussed. Finally, in Section 8, we discuss the parameter N and its possible reduction in size. The paper concludes with some general comments and conclusions.

2. Preliminary Definitions

Most of the quantities defined in this work depend on three parameters which are denoted by M , K and N where

- M takes two values $M = Min \equiv 0$ and $M = Max \equiv 1$ depending on whether we are dealing with the maximum or the minimum values of a series or a data set,
- K represents the number of ordered observations used for the analysis, $K \in \{1, \dots, n\}$ where n is the sample size and
- N is the number of time series or data sets involved in the analysis with $N \geq 2$.

A class of dimensionless indices was recently defined by Makris and Vonta [1] that depends on a basic statistical characteristic μ like the mean (average), the variance, correlation, etc. of K ordered (largest or smallest) observations of a series. More specifically, a basic characteristic of K ordered observations is compared (through division) with a basic characteristic of all (total) observations involved in the analysis, with the latter not necessarily the same as the one used for the K ordered observations. The indices are considered to be dimensionless since we divide the same type of quantities (the same statistical characteristic). Even if two different characteristics are used, the index will remain dimensionless as long as the characteristics are in the same unit of measurement (e.g., mean for the numerator and standard deviation for the denominator or vice versa).

For the definition, consider two time series i and j and let $t^{i(k)}$ the time point at which the k th ordered observation of the series i has occurred. Having available K time points corresponding to the K largest (or smallest) ordered observations of a time series i , which plays the role of the basis for the index evaluation, we proceed and calculate the basic statistical characteristic μ of the K observations of the second series j conditionally on the time points that the baseline time series i displays its K largest (or smallest) values.

Definition 1 (see [1]). For two time series i and j , the $\mu_{j|i}$ index is defined by

$$\mu_{j|i}^{MK2} = \begin{cases} \mu_{j|i}^{MaxK2} = \frac{\mu_{j|i}^{maxK}}{\mu_{j|i}^T} = \frac{\mu(j|t^{i\{n-K+1\}})}{\mu(j^{Total})} \\ \mu_{j|i}^{MinK2} = \frac{\mu_{j|i}^{minK}}{\mu_{j|i}^T} = \frac{\mu(j|t^{i\{K\}})}{\mu(j^{Total})} \end{cases} \quad (1)$$

where the notation $t^{i\{K\}}$ denotes the K time points where the series i displays its K smaller values and $t^{i\{n-K+1\}}$ the K time points where the series i displays its K larger values. In addition, j^{Total} is the notation for all the observations of the time series j .

Observe that the above index can be obtained using as a basis the series j , resulting in the index $\mu_{i|j}^{MK2}$. Naturally the index is not symmetric since the time points of the K ordered observations of one series do not necessarily occur at the exact same time points of the other series. Extending the idea of the above index one could also evaluate the index for the same series obtaining the indices $\mu_{i|i}^{MK2}$ and $\mu_{j|j}^{MK2}$. The combination of the above indices can be represented in a 2×2 matrix which, in general, is denoted by $[\mu]^{MKN}$ where μ represents the statistical function (characteristic) to be used in the analysis. The matrix created below refers to the case of the K smallest observations of the baseline series i when the basic characteristic μ is the standard deviation (*Std*):

$$[Std]^{MinK2} = \begin{pmatrix} Std_{1|1}^{MinK2} & Std_{2|1}^{MinK2} \\ Std_{1|2}^{MinK2} & Std_{2|2}^{MinK2} \end{pmatrix}$$

From the matrix above, observe that the diagonal elements refer to the baseline time series itself while the off-diagonal elements refer to one time series conditionally on the ordered time points of the other. Thus, the evaluation of the statistical characteristic (the standard deviation here) for each row k of the matrix is done conditionally on the time series k .

Remark 1. Note that in some instances the numerator will be always smaller or larger than the denominator. Indeed, if the mean (average) is used as the statistical characteristic then for the diagonal elements, the numerator will be always smaller than the denominator if the K smallest observations are used and larger if the K largest observations are used. In addition, in the case where the maximum values are studied, it is valid that the values of the indices that appear in the diagonal elements are most often larger than the off-diagonal elements in the same row and this is because the values in the same row are calculated conditionally on the time points where the diagonal elements are defined (see Example 1 below). In the case where the minimum values are examined, the elements in the diagonal most often take the minimum values in each row due to the same conditional argument.

The general form of the matrix for a general statistical characteristic μ for N time series given in [1] is of dimension $N \times N$ and is presented as follows:

$$[\mu]^{MKN} = \begin{pmatrix} \mu_{1|1}^{MKN} & \mu_{2|1}^{MKN} & \dots & \mu_{N|1}^{MKN} \\ \dots & \dots & \dots & \dots \\ \mu_{1|N}^{MKN} & \mu_{2|N}^{MKN} & \dots & \mu_{N|N}^{MKN} \end{pmatrix} \quad (2)$$

Remark 2. In the case where the maximum values are studied ($M = \text{Max}$) and there are ties, for example when a value, say $i_{(k)}$, is tied with other values within the range under investigation, e.g., $i_{(n)}, \dots, i_{(n-K+1)}$, then we select among those time points the one that maximizes the index μ^{MKN} in the remaining $N - 1$ time series because this way, we are being more conservative in terms of the similarity between the N series. Notice that the time point that provides the maximum index for each of the remaining $N - 1$ time series might be different for each series. This remark will be become clear in the following example. Analogously, for $M = \text{Min}$ we will be choosing each time the time point that minimizes the index in the remaining $N - 1$ time series.

Example 1. For a better understanding of the previous Remarks, consider three time series A , B and C consisting of 20 values (see Figure 1):

$A = (8, 9, 7, 19, 6, 7, 4, 7, 5, 6, 19, 10, 9, 21, 8, 9, 5, 9, 11, 12)$,

$B = (4, 5, 6, 11, 4, 5, 3, 4, 10, 4, 6, 5, 4, 12, 5, 4, 3, 6, 5, 7)$ and

$C = (13, 14, 15, 12, 13, 14, 8, 13, 7, 13, 25, 14, 13, 30, 14, 13, 15, 15, 14, 16)$.

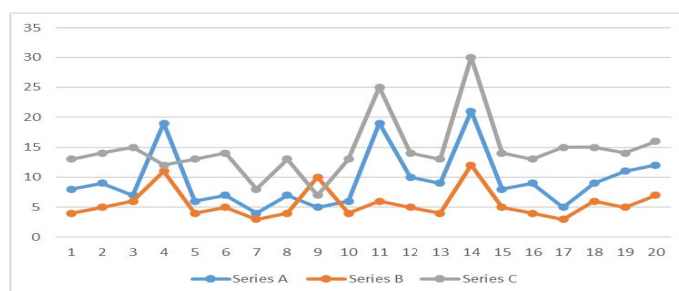


Figure 1. Time series A, B and C.

For the time series A , we observe that $A_{(20)} = 21$ and occurs at $t^{A(20)} = 14$ while $A_{(19)} = 19$ and appears twice in the series at time points 4 and 11, namely $t^{A(19)} = 4$ & 11. Obviously, for $K = 1$ and for the characteristic μ being the average (AV), $\mu_{A|A}^{Max13} = 2.20$ while for $K = 2$, $\mu_{A|A}^{Max23} = 2.09$. Observe now that for $K = 2$, $\mu_{B|A}^{Max23} = 2.04$ if the 4th observation of the series B is used and $\mu_{B|A}^{Max23} = 1.59$ if the 11th observation of the series B is used. Similarly, we get $\mu_{C|A}^{Max23} = 1.44$ if the 4th observation of the series C is used. Instead, if the 11th observation is used in the calculation then $\mu_{C|A}^{Max23} = 1.89$. The results are summarized in Table 1 for $K = 1, 2, 3$.

Table 1. The μ^{MaxK3} indices for three time series A , B and C .

K	t	$\mu_{A A}^{MaxK3}$	$\mu_{B A}^{MaxK3}$	$\mu_{C A}^{MaxK3}$
1	$t^{A(20)} = 14$	2.20	2.12	2.06
2	$t^{A(20)} = 14$ & $t^{A(19)} = 4$	2.09	2.04	1.44
2	$t^{A(20)} = 14$ & $t^{A(19)} = 11$	2.09	1.59	1.89
3	$t^{A(20)} = 14$ & $t^{A(19)} = 4$ & $t^{A(18)} = 11$	2.06	1.71	1.53

Based on Remark 2, the first row of the matrix $[AV]^{Max23}$ is given as

$$(2.09 \quad 2.04 \quad 1.89).$$

Analogously, we will deal now with the case of the minimum values of the time series. For the time series A , we observe that $A_{(1)} = 4$ and occurs at $t^{A_{(1)}} = 7$ while $A_{(2)} = 5$ and appears twice in the series at time points 9 and 17, namely $t^{A_{(2)}} = 9 \& 17$. Obviously, for $K = 1$ and for the characteristic μ being the average (AV), $\mu_{A|A}^{Min13} = 0.42$ while for $K = 2$, $\mu_{A|A}^{Min23} = 0.47$. Observe now that for $K = 2$, $\mu_{B|A}^{Min23} = 1.15$ if the 9th observation of the series B is used and $\mu_{B|A}^{Min23} = 0.53$ if the 17th observation of the series B is used. Similarly we get $\mu_{C|A}^{Min23} = 0.52$ if the 9th observation of the series C is used. Instead, if the 17th observation is used in the calculation then $\mu_{C|A}^{Min23} = 0.79$. The results are summarized in Table 2 for $K = 1, 2, 3$.

Table 2. The μ^{MinK3} indices for three time series A, B and C .

K	t	$\mu_{A A}^{MinK3}$	$\mu_{B A}^{MinK3}$	$\mu_{C A}^{MinK3}$
1	$t^{A_{(1)}} = 7$	0.42	0.53	0.55
2	$t^{A_{(1)}} = 7 \& t^{A_{(2)}} = 9$	0.47	1.15	0.52
2	$t^{A_{(1)}} = 7 \& t^{A_{(2)}} = 17$	0.47	0.53	0.79
3	$t^{A_{(1)}} = 7 \& t^{A_{(2)}} = 9 \& t^{A_{(3)}} = 17$	0.49	0.94	0.69

Based on Remark 2, the first row of the matrix $[AV]^{Min23}$ is given as

$$(0.47 \quad 0.53 \quad 0.52).$$

2.1. Multivariate Indices

For a better comparison between time series, a multivariate index could be used based not on a single but rather on multiple statistical characteristics. Thus, μ could be increased in dimension in order to include more than one statistical functions and be a vector of higher dimension. For example, for the control of four statistical characteristics simultaneously (i.e., average $[AV]$, standard deviation $[Std]$, coefficient of variation $[CV]$ and covariance $[Cov]$, the index μ can have dimension four creating a generalized matrix of dimension $N \times N \times 4$ defined in (3) with $\mu = [AV, Std, CV, Cov]$:

$$\begin{pmatrix} [AV, Std, CV, Cov]_{1|1}^{MKN} & [AV, Std, CV, Cov]_{2|1}^{MKN} & \dots & [AV, Std, CV, Cov]_{N|1}^{MKN} \\ \vdots & \vdots & \ddots & \vdots \\ [AV, Std, CV, Cov]_{1|N}^{MKN} & [AV, Std, CV, Cov]_{2|N}^{MKN} & \dots & [AV, Std, CV, Cov]_{N|N}^{MKN} \end{pmatrix}. \quad (3)$$

We should stress that each element of the matrix (3) is itself a vector of dimension 4, which contains the values of the four indices that correspond to the four statistical measures, for the case under consideration.

Multivariate indices could become extremely useful in stochastic ordering and majorization both of which play a key role in many areas of statistics as for example, in reliability theory and engineering. For instance, in regard to majorization, an n -dimensional vector $\mathbf{x}$ is said to be majorized by another n -dimensional vector $\mathbf{y}$, denoted by $\mathbf{x} \stackrel{m}{\leq} \mathbf{y}$ if

$$\sum_{i=1}^j x_{(i)} \leq \sum_{i=1}^j y_{(i)}, \quad j = 1, \dots, n-1 \quad \text{and} \quad \sum_{i=1}^n x_{(i)} = \sum_{i=1}^n y_{(i)}.$$

where $x_{(i)}$ and $y_{(i)}$ refer to the ordered elements of the two vectors. We use the term *strict majorization* if in the above we use strict inequality. Furthermore, if the inequality holds for all j s including the case $j = n$, we are dealing with *weak majorization*. For details about majorization please see [17].

Before closing the section and moving to the next one where some new advanced indices will be proposed, it is important to mention that the indices discussed in this work are applicable to both stationary and non-stationary processes. Although the series to be compared are expected to be of the same nature and as such the comparison is meaningful, in general, the use of stationary or non-stationary data and whether a differencing should be applied, is a challenging problem within the framework of similarity measures, that goes beyond the scope of the present work and is left as an open problem for a future project.

3. Advanced Dimensionless Indices

For better and efficient comparison between series, we propose in this section two new advanced index matrices that depend on the same three parameters M, K and N as the ones in Definition 1. The first index matrix is obtained when each element of each row in matrix (2) is divided by the diagonal element of the same row. The resulting new index matrix is denoted by $[\mu_1]^{MKN}$. As expected, this matrix has all its diagonal elements equal to one while all off-diagonal elements take non-zero values.

In a similar way, we propose the 2nd index matrix $[\mu_2]^{MKN}$ where each element of each column in matrix (2) is divided by the diagonal element of the same column. Both index matrices can be viewed as similarity measures between the series involved. The proposed index matrix measures are given below:

$$[\mu_1]^{MKN} = \begin{pmatrix} \frac{\mu_{1|1}^{MKN}}{\mu_{1|1}^{MKN}} & \frac{\mu_{2|1}^{MKN}}{\mu_{1|1}^{MKN}} & \cdots & \frac{\mu_{N|1}^{MKN}}{\mu_{1|1}^{MKN}} \\ \frac{\mu_{1|1}^{MKN}}{\mu_{1|1}^{MKN}} & \frac{\mu_{2|1}^{MKN}}{\mu_{1|1}^{MKN}} & \cdots & \frac{\mu_{N|1}^{MKN}}{\mu_{1|1}^{MKN}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\mu_{1|N}^{MKN}}{\mu_{N|N}^{MKN}} & \frac{\mu_{2|N}^{MKN}}{\mu_{N|N}^{MKN}} & \cdots & \frac{\mu_{N|N}^{MKN}}{\mu_{N|N}^{MKN}} \end{pmatrix} \quad (4)$$

and

$$[\mu_2]^{MKN} = \begin{pmatrix} \frac{\mu_{1|1}^{MKN}}{\mu_{1|1}^{MKN}} & \frac{\mu_{2|1}^{MKN}}{\mu_{2|2}^{MKN}} & \cdots & \frac{\mu_{N|1}^{MKN}}{\mu_{N|N}^{MKN}} \\ \frac{\mu_{1|1}^{MKN}}{\mu_{1|1}^{MKN}} & \frac{\mu_{2|2}^{MKN}}{\mu_{2|2}^{MKN}} & \cdots & \frac{\mu_{N|2}^{MKN}}{\mu_{2|2}^{MKN}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\mu_{1|N}^{MKN}}{\mu_{1|1}^{MKN}} & \frac{\mu_{2|N}^{MKN}}{\mu_{2|2}^{MKN}} & \cdots & \frac{\mu_{N|N}^{MKN}}{\mu_{N|N}^{MKN}} \end{pmatrix} \quad (5)$$

The above proposed advanced MKN indices (index matrices) are dimensionless while their interpretation is much more clear than that of the standard indices presented in Definition 1 because they provide efficient pairwise comparison in pairs between series and could be considered as the percentage of similarity between two time series i and j . For instance, if for the matrix μ_1 the value of the $(1,2)$ -element $\mu_{1(1,2)} = 0.75$, this means that the second time series is 75% similar as compared with the first series, or in other words, the index of the second series is 75% the value of the corresponding index of the first series, when the calculations are conditional on the time points where the first time series achieves its K maximum or minimum values. The conclusion is the same if this index turns out to be equal to 1.75.

The comparison through the matrix measure $[\mu_1]$ is indirect and becomes more apparent when the values of one time series are multiples of the values of the other, because the element $\mu_{1(i,j)}^{MKN} = \frac{\mu_{j|i}^{MKN}}{\mu_{i|i}^{MKN}}$ and therefore the statistical function μ based on the values of the time series j in the numerator at the time points of maximum (or minimum) values of the time series i is compared against the statistical function μ based on the values of the time

series i in the denominator. That is, the closer the indices in value to the proportionality parameter (or its inverse depending on the case) between two time series, the more confident we are about their similarity in terms of the occurrence of their maximum (or equivalently minimum) values.

In the case of the matrix index $[\mu_2]$, observe that the comparison is direct as the index $\mu_{2(i,j)}^{MKN} = \frac{\mu_{ji}^{MKN}}{\mu_{ij}^{MKN}}$ and therefore the statistical function μ of the time series j in the numerator is based on the time points of maximum (or minimum) values of the time series i and compared against the statistical function μ based on itself in the denominator.

Thus, if for two time series 1 and 2 the series 2 has an index value $\mu_{2(1,2)}^{MaxKN} = 1$, then this means that the second time series presents its K maximum values at the exact same time points as the first time series, and vice versa, if time series 1 has index $\mu_{2(2,1)}^{MaxKN} = 1$, then the first series presents its maximum values at the same time points as the second series. In fact, in general, we have:

$$\mu_{2(i,j)}^{MaxKN} = 1 \Leftrightarrow \mu_{2(j,i)}^{MaxKN} = 1.$$

Although this work is devoted to similarity, we cannot overlook the fact that it is, at the same time, interrelated to the concepts of causality and cointegration which are briefly mentioned below for the sake of completeness. The notion of causality is rather common in economic time series although causality issues could also be found in reliability or engineering. The difficulties of establishing a causal relationship between economic variables led Granger ([18]) to develop the economic concept of causality known as *Granger Causality*. On the other hand, the concept of cointegration was established later and discussed thoroughly in [19] where the associated statistical inference including tests used to identify the long-term relationships between two or more series, was explored. The phenomenon is also quite common since, in general, economic theory forces certain pairs of series staying close to each other and moving alongside. For testing cointegration, one could use the Engle–Granger Augmented Dickey–Fuller test for cointegration (EG-ADF test, [19]) based on the classical Dickey–Fuller test, if two series are involved or the Johansen test if more than two series are involved ([20]). For further reading, the interested reader is referred to [21,22] and to the interesting review article by Hubrich et al. [23].

In Microeconomics, another concept closely related to the above is the concept of elasticity which is a measure of the sensitivity of a variable to a change in another variable. For instance, although the prices of some goods are inelastic, this is not always the case. This issue, which is of great importance in marketing, is explored through an example in the following section where a special price elasticity index is presented and discussed.

In the sections that follow, the quantities that play a key role in the analysis of the proposed index matrices of this section, namely the function (or characteristic) μ and the parameters M , K and N , will be thoroughly explored.

4. An Epidemiological Application

In this section, we examine time series on influenza-like-illness (ILI) rates for the purpose of identifying differences between (geographical) regions as well as differences between each region-rate and the country-rate. Usually, the purpose of such analyses is the identification of areas for which further monitoring and actions for reducing the spread of a disease are needed.

The data have been drawn from the Sentinel system of the Hellenic National Public Health Organization (EODY) for the period 2004–2014. From the data entered into the system by the physicians, the ILI rate is calculated weekly as the number of ILI cases per 1000 visits ($1000 \cdot \text{cases} / \text{visits}$), a rate that displays the spread of the disease to the community.

In this work, we will present and compare six time series which report the weekly ILI-rate for the time period of the 28th week of 2004 up to and including the 39th week of

2014 in four geographical regions in which Greece is divided. The time series which report the ILI-rate for the four regions will be denoted by ILI-1 to ILI-4. For the entire country, two series are available, the overall ILI-rate as a weighted average of ILI-1 through ILI-4 (based on appropriate weights depending on the population size per region), denoted by ILI and the ILI-total that reports the total number of cases per 1000 visits ($1000 \times \text{cases}/\text{visits}$). Through the matrix $[\mu]^{MKN}$ given in (2), we compare the similarity of the six time series based the μ^{MKN} indices where the parameter μ is the average, $N = 6$, $M = \text{Max}$ and the parameter K takes two values, $K = 10$ (Table 3) and $K = 20$ (Table 4).

The $\mu_{(1|1)}^{MKN}$ element of the matrix in Table 3 is equal to 8.6051 which means that the average of the 10 largest observations of the series ILI is about 8.6 times bigger than the average of all ILI observations. Observe that the average of the 10 observations of each of the series ILI-1, ILI-2 and ILI-3 (8.2869, 8.4281, 8.4036) evaluated conditionally on the time points where the country-rate displays its 10 largest observations, are very similar but the same is not true for the 4th country region where a much lower similarity value is observed (7.6249). This implies that the rate of the disease in that specific region (Aegean Sea islands and Crete) is not as high as in all other regions. The time series ILI-total displays an even smaller similarity (7.36) with the time series ILI, a fact that is probably attributed to the better definition of the ILI-rate as a weighted average of ILI1 – ILI4.

From the results of Table 4 (for $K = 20$), we observe that all diagonal elements as well as almost all off-diagonal elements are reduced in size as compared to the elements of Table 3. Observe that as K increases, the diagonal elements approach 1.

Table 3. Matrix $[AV]^{Max,10,6}$.

	ILI	ILI-Total	ILI-1	ILI-2	ILI-3	ILI-4
ILI	8.6051	7.3692	8.2869	8.4281	8.4036	7.6249
ILI-total	4.2769	4.8284	4.2289	4.4310	4.5389	3.9419
ILI-1	9.7851	7.2102	10.1454	9.1647	9.1323	7.6818
ILI-2	8.0030	7.3993	7.5104	8.5804	7.7295	6.7081
ILI-3	8.9237	7.2820	8.3950	8.5294	9.4920	7.2900
ILI-4	7.0564	6.3136	6.5287	6.3817	7.0027	8.3180

Table 4. Matrix $[AV]^{Max,20,6}$.

	ILI	ILI-Total	ILI-1	ILI-2	ILI-3	ILI-4
ILI	7.5738	7.2367	7.2424	7.1852	7.1960	6.6832
ILI-total	4.3343	4.4772	4.0483	4.2104	4.1190	3.7095
ILI-1	7.8491	7.4434	8.3022	7.2914	7.2111	6.6973
ILI-2	7.4669	7.1833	6.9574	7.7247	6.8522	6.1845
ILI-3	7.4722	7.4135	6.6209	7.0901	7.8949	5.9697
ILI-4	6.6155	6.0219	6.0356	5.9091	6.4685	7.3293

The matrix $[\mu_1]^{MKN}$ given in (4) for μ_1 being the average and $K = 10$ is presented in Table 5. According to the definition of $[\mu_1]^{MKN}$, the diagonal elements are all equal to 1. Observe that the first three geographical regions display very similar behavior as compared with the country-rate (ILI) with values at least equal to 96%, while the fourth region is much less similar (88% similarity). In Table 6 for $K = 20$, we observe that the differences between all time series are alleviated and the indices approach 1. The fourth geographical region constitutes an exception since it reaches about 88% similarity as compared with the whole country and even less similarity as compared with the other three regions.

Table 5. Matrix $[\mu_1]^{Max,10,6}$.

	ILI	ILI-Total	ILI-1	ILI-2	ILI-3	ILI-4
ILI	1	0.8564	0.9630	0.9794	0.9766	0.8861
ILI-total	0.8858	1	0.8758	0.9177	0.9401	0.8164
ILI-1	0.9645	0.7107	1	0.9033	0.9001	0.7572
ILI-2	0.9327	0.8624	0.8753	1	0.9008	0.7818
ILI-3	0.9401	0.7672	0.8844	0.8986	1	0.7680
ILI-4	0.8483	0.7590	0.7849	0.7672	0.8419	1

Table 6. Matrix $[\mu_1]^{Max,20,6}$.

	ILI	ILI-Total	ILI-1	ILI-2	ILI-3	ILI-4
ILI	1	0.9555	0.9562	0.9487	0.9501	0.8824
ILI-total	0.9681	1	0.9042	0.9404	0.9199	0.8285
ILI-1	0.9454	0.8966	1	0.8782	0.8686	0.8067
ILI-2	0.966	0.9299	0.9007	1	0.8871	0.8006
ILI-3	0.9465	0.9390	0.8386	0.8980	1	0.7561
ILI-4	0.9026	0.8216	0.8235	0.8063	0.8825	1

Epidemiological differences accurately identified among regional series are useful to health officials since they provide a useful tool for identifying, as early as possible, disease outbreaks in certain regions and is beneficial to the society in general, for early detection, prevention and spread of extreme, possibly harmful, events.

5. The Function μ

The measure μ which refers to the ratio of the statistical functions considered is an important factor in the analysis of the data through the indices.

In previous sections μ was chosen to be a basic statistical function like the average, the standard deviation, the coefficient of variation, etc. In this section, we deal with the case where μ is a differential (backshift) operator denoted by d referring to the first differences of a time series. We denote by d^K the differential based on first differences between the K maximum (or minimum) values of a time series. Thus, d^K is a $K - 1$ dimensional vector defined by $(i_{(n)} - i_{(n-1)}, i_{(n-1)} - i_{(n-2)}, \dots, i_{(n-K+2)} - i_{(n-K+1)})$ while d^T denotes the operator applied to all (total) values of the time series and is of dimension $n - 1$. The corresponding index is defined below.

$$d^{MKN} = \frac{d^K}{d^T} \quad (6)$$

with an alternative version being defined as

$$\mu * d^{MKN} = \frac{d^K}{\mu^T} \quad (7)$$

where the denominator relies on any statistical function including any ordered observation. Indeed, in (7) with $M = Max$ and $K = 2$, the operator d can be the classical first difference between the two largest observations for the series i , namely $i_{(n)} - i_{(n-1)}$ divided or weighted by either the mean of all observations (see (8)) or the maximum of the two largest values (see (9)) or the minimum among the two largest observations (see (10)). Observe

that the last one is nothing but the well-known *relative or percentage change*. The relevant expressions are provided below:

$$\mu * d^{a,Max21} = \frac{d^2}{\mu^T} = \frac{i_{(n)} - i_{(n-1)}}{AV(i_1, \dots, i_n)} \quad (8)$$

$$\mu * d^{b,Max21} = \frac{d^2}{\max} = \frac{i_{(n)} - i_{(n-1)}}{i_{(n)}} \quad (9)$$

and

$$\mu * d^{c,Max21} = \frac{d^2}{\min} = \frac{i_{(n)} - i_{(n-1)}}{i_{(n-1)}} \quad (10)$$

In contrast to the above definitions, the index (11) calculates the ratio of two indices. More specifically it calculates the ratio of the percentage change of the time series i to the percentage change of the time series j conditionally on the time points the time series i displays its K maximum (or minimum) values. A comment would be that the function d^T could be equal to 1 or any other constant. Its existence is to achieve weighting and to ensure the dimensionless property of the indices.

We define the new index as

$$\mu * d_{ij}^{p,Max22} = \frac{\mu_{i|i}^{c,[Max22]}}{\mu_{j|i}^{c,[Max22]}} = \frac{\frac{i_{(n)} - i_{(n-1)}}{i_{(n-1)}}}{\frac{j(i_{(n)}) - j(i_{(n-1)})}{j(i_{(n-1)})}} \quad (11)$$

Example 2. For the case of the index defined in (11), and for the case of three maximum values of the time series i , (i.e., $K = 3$) a vector of dimension 2 can be derived in the numerator of the index and therefore the index is actually the 2-dimensional vector

$$\mu * d^{c,Max31} = \frac{d^3}{\min} = \left(\frac{i_{(n)} - i_{(n-1)}}{i_{(n-1)}}, \frac{i_{(n-1)} - i_{(n-2)}}{i_{(n-2)}} \right) \quad (12)$$

while for $M = \text{Min}$ and $K = 4$ we have

$$\mu * d^{c,Min41} = \frac{d^4}{\min} = \left(\frac{i_{(2)} - i_{(1)}}{i_{(1)}}, \frac{i_{(3)} - i_{(2)}}{i_{(2)}}, \frac{i_{(4)} - i_{(3)}}{i_{(3)}} \right). \quad (13)$$

Example 3. (Price elasticity of demand). The index defined in (11) has many applications, mainly in economics and marketing. For example, if the time series i in (11) is the demand for a good, say A , measured in Q units and if the time series j reports the corresponding price values, say P , of the good, then the resulting index is a measure of the response of the maximum (or minimum) quantity demanded of a good, relative to the change in its price, with all other factors considered constant (see index (15) below). In other words, the index expresses the percentage change of the maximum (or minimum) quantity demanded of the good to the percentage change of its price, known as the price elasticity of the maximum (or minimum) demand denoted by $E_{(D,Q,Max)}$, i.e., a novel form of the known price elasticity of demand E_D ([24]). For the classical case with $K = 2$ for two series, we have

$$\mu * d_{ij}^{p,Max22} = \frac{\frac{Q_{(n)} - Q_{(n-1)}}{Q_{(n-1)}}}{\frac{P(i^{Q_{(n)}}) - P(i^{Q_{(n-1)}})}{P(i^{Q_{(n-1)}})}} \quad (14)$$

which can also be denoted by

$$\mu * d_{ij}^{p,Max22} = \frac{\Delta Q_{(n)}}{\Delta P|Q_{(n)}} \equiv E_{(P|Q)_{(n)}} \quad (15)$$

where $\Delta Q_{(n)}$ stands for the ratio $\frac{Q_{(n)} - Q_{(n-1)}}{Q_{(n-1)}}$. Note that in the general K case we have

$$\mu * d_{PQ}^{p, MaxKN} = \left(\frac{\Delta Q_{(n)}}{\Delta P|Q_{(n)}}, \dots, \frac{\Delta Q_{(n-K+2)}}{\Delta P|Q_{(n-K+2)}} \right) \equiv (E_{(P|Q)_{(n)}}, \dots, E_{(P|Q)_{(n-K+2)}}) \quad (16)$$

It should be noted that if the price is replaced by income then the resulting index will be a novel elasticity of demand for income while if the price is substituted by demand for a complementary good (relative to good A) the resulting index will be a novel cross-elasticity of demand.

6. Parameter M and Cross-Correlation Indices

The values of the parameter M , one involving the minimum ($M = 0$) and one the maximum ($M = 1$) values of a data set, can be used in combination, connecting maximum and minimum values between time series simultaneously. This proposal, that is, to connect the occurrence of maximum and minimum values between time series, is inspired by variables that are complementary, such as the prices of two substitute goods (such as the matches and the lighter), where, as it is well-known when the price of a good rises the price of the substitute good goes down, with the result that at the time points one good receives its maximum prices, the other (namely the substitute good) receives its minimum prices.

For two time series i and j , when we are interested in comparing the information from the maximum and the minimum values simultaneously, we propose the definition below with the notation $\mu^{M[Cross]KN}$ for the new index. More specifically this definition entails two parts, one for the case $i = j$ and one for the case $i \neq j$. In the first case, the K values of the time series j used to calculate the function μ are the time points where the K maximum values of the time series j are presented, whereas for the case $i \neq j$ the K values of the time series j used in order to calculate the function μ are the time points where the K minimum values of the time series i are presented. In this way, a cross-coupling of the time points of occurrence of the maximum and minimum values of the time series i and j is established. Obviously, another definition arises when we replace in the above the maximum values with the minimum values and vice versa.

$$\mu_{j|i}^{M[Cross]KN} = \begin{cases} \mu_{j|i}^{M[Cross]KN} = \frac{\mu_{j|j}^{maxKN}}{\mu_{j|j}^T} = \frac{\mu(j|t^{\{n-K+1\}})}{\mu(j^{Total})} & \text{for } i = j \\ \mu_{j|i}^{M[Cross]KN} = \frac{\mu_{j|i}^{minKN}}{\mu_{j|i}^T} = \frac{\mu(j|t^{\{i\{K\}\})}}{\mu(j^{Total})} & \text{for } i \neq j \end{cases} \quad (17)$$

Based on the definition (17), a matrix defined in (18) can be created and denoted by $[\mu]^{M[Cross]KN}$. Through this matrix and based on the values of each of N time series in general, it can be seen whether a cross relationship of maximum and minimum values between them exists.

$$[\mu]^{M[Cross]KN} = \begin{pmatrix} \mu_{1|1}^{M[Cross]KN} & \mu_{2|1}^{M[Cross]KN} & \dots & \mu_{N|1}^{M[Cross]KN} \\ \dots & \dots & \dots & \dots \\ \mu_{1|N}^{M[Cross]KN} & \mu_{2|N}^{M[Cross]KN} & \dots & \mu_{N|N}^{M[Cross]KN} \end{pmatrix} \quad (18)$$

As an example, consider two time series 1 and 2. If the cell value $\mu_{1|1}^{M[Cross]KN}$ (which is calculated based on the K maximum values of the time series 1) is the same or close to the value of the cell $\mu_{2|1}^{M[Cross]KN}$ (which is calculated on the values of the time series 2 conditionally on the time points where the K minimum values of the time series 1 are presented), then this means that wherever the time series 1 presents its K maximum values, the time series 2 will present its K minimum values. Alternatively, if the value of the cell $\mu_{1|2}^{M[Cross]KN}$ is very close to the value of the cell $\mu_{2|2}^{M[Cross]KN}$, then this means that wherever the time series 2 presents its maximum values, the time series 1 presents its minimum

values. One of the aforementioned results does not imply the second and vice versa, but when both are valid then there will be a time cross correlation between the two time series in terms of their K maximum and minimum values.

Caution is required for the possibility of spurious correlation. Since the cross-correlation mentioned earlier may be due to the presence of a hidden, confounding factor, certain measures should be taken for investigating such a possibility. It should be noted that spurious correlation is not uncommon and it is surfaced not only in economics and finance but also in behavioral sciences.

7. Non-Time Dependent Data

7.1. Parameter K

The parameter K is the most important parameter for the indices μ^{MKN} , as it is the one that affects the condition under which the calculations of the numerator of the indices is being done. In this section, a change in the calculation condition will be presented and various cases will be discussed through examples.

In the previous sections, the indices were defined for time-dependent data (time series). In this section, the indices will be defined for data sets which are independent of time. The parameters μ , M and N remain unchanged as in the previous sections in terms of definitions, while the parameter K in the setting of this section corresponds to the frequency of occurrence of K distinct maximum or minimum values (as opposed to correspondence with the time points of occurrence of K maximum or minimum values).

Consider two random variables X and Y . A random sample $(X_1, \dots, X_{n'})$ is drawn from the same distribution as X and similarly a random sample $(Y_1, \dots, Y_{n'})$ is drawn from the distribution of Y . The ordered realization samples are defined as $(X_{(1)}, \dots, X_{(n')})$ and $(Y_{(1)}, \dots, Y_{(n')})$, respectively.

Furthermore, consider the ordered realizations of distinct values in each data set, that is,

$$(x_{(1)} < x_{(2)} < \dots < x_{(n_1)}) \text{ and } (y_{(1)} < y_{(2)} < \dots < y_{(n_2)})$$

and let the corresponding frequencies of occurrence for each ordered distinct value of X and Y be $(f_{(1)}^X, \dots, f_{(n_1)}^X)$ and $(f_{(1)}^Y, \dots, f_{(n_2)}^Y)$, respectively.

We define indices between the two variables $X \equiv i$ and $Y \equiv j$ as follows

$$\mu_{ji}^{MK[f]2} = \begin{cases} \mu_{ji}^{MaxK[f]2} = \frac{\mu_{ji}^{MaxK[f]2}}{\mu_{ji}^T} = \frac{\mu(j|f^{i\{n_i-K+1\}})}{\mu(j^{Total})} \\ \mu_{ji}^{MinK[f]2} = \frac{\mu_{ji}^{MinK[f]2}}{\mu_{ji}^T} = \frac{\mu(j|f^{i\{K\}})}{\mu(j^{Total})} \end{cases} \quad (19)$$

which for N series can be summarized in a $N \times N$ matrix as follows

$$[\mu]^{MK[f]N} = \begin{pmatrix} \mu_{1|1}^{MK[f]N} & \mu_{2|1}^{MK[f]N} & \dots & \mu_{N|1}^{MK[f]N} \\ \dots & \dots & \dots & \dots \\ \mu_{1|N}^{MK[f]N} & \mu_{2|N}^{MK[f]N} & \dots & \mu_{N|N}^{MK[f]N} \end{pmatrix} \quad (20)$$

We should point out here based on the above, that

- The realizations of the random variables $X = i$ and $Y = j$ are of the same length but not necessarily with the same number of distinct ordered values.
- $f^{i\{n_i-K+1\}} = \{f_{(n_i)}^i, \dots, f_{(n_i-K+1)}^i\}$ is the notation for the collection of frequencies with which the K distinct maximum observations of the variable i occur and similarly $f^{i\{K\}} = \{f_{(1)}^i, \dots, f_{(K)}^i\}$ is the notation for the collection of frequencies with which the K distinct minimum observations of the variable i occur.

- If, for example, the function μ is the average, the index $AV_{(j/i)}^{MaxK^{(f)}N}$ is calculated as a ratio where in the numerator we have the average of so many maximum values of the data set j as the sum of the frequencies in the set $f^{i_{\{n_1-K+1\}}}$ dictate. That is, we take the average of $f_{(n_1)}^i + \dots + f_{(n_1-K+1)}^i$ maximum values of the data set j , and we divide by the average of all the values of the data set j .
- Through the indices $AV_{(j/i)}^{MaxK^{(f)}N}$, we evaluate whether two data sets have their K maximum distinct values occurring with the same frequencies. A similar result holds for $M = Min$.
- Let the notation $X_{\{n_1-K+1\}}$ stand for the collection of distinct ordered times of the data set X of the form $\{x_{(n_1)}, \dots, x_{(n_1-K+1)}\}$. We consider a similar notation for the data set Y .

An example will be presented below for a better understanding of the concepts.

Example 4. Let $K = 2$, $M = Max$ and $N = 2$ and focus on the frequencies of occurrence of the first two distinct maximum values of a realization of the random variable X . Let also a realization of a random variable Y . Suppose that $x_{(n_1)}$ appears $f_{(n_1)}^X$ times and $x_{(n_1-1)}$ appears $f_{(n_1-1)}^X$ times, so that the calculation of an index of X based on itself, with the parameter μ defined as the average (AV), is as follows:

$$\mu_{X|X}^{MaxK^{[f]}=2,2} = \frac{AV_{X|X}^{K[f]=2}}{AV_{X|X}^T} = \frac{AV(X_{\{n_1-1\}} | f^{i_{\{n_1-1\}}})}{AV(X^{Total})} = \frac{\frac{f_{(n_1)}^X \cdot x_{(n_1)} + f_{(n_1-1)}^X \cdot x_{(n_1-1)}}{f_{(n_1)}^X + f_{(n_1-1)}^X}}{\bar{X}}.$$

For the evaluation of the index for Y calculated based on the frequencies of the K maximum distinct observations of X , we have

$$\mu_{Y|X}^{MaxK^{[f]}=2,2} = \frac{AV_{Y|X}^{K[f]=2}}{AV_{Y|X}^T} = \frac{AV(Y | f^{i_{\{n_1-1\}}})}{AV(Y^{Total})} = \frac{\frac{\sum_{k=n_1'-(f_{(n_1)}^X+f_{(n_1-1)}^X)+1}^{n_1'} Y_{(k)}}{f_{(n_1)}^X + f_{(n_1-1)}^X}}{\bar{Y}}.$$

Observe that for $K > 2$ the above expressions take the form

$$\mu_{X|X}^{MaxK^{[f]},2} = \frac{AV_{X|X}^{K[f]}}{AV_{X|X}^T} = \frac{AV(X_{\{n_1-K+1\}} | f^{i_{\{n_1-K+1\}}})}{AV(X^{Total})} = \frac{\frac{\sum_{k=n_1-K+1}^{n_1} f_{(k)}^X \cdot x_{(k)}}{\sum_{k=n_1-K+1}^{n_1} f_{(k)}^X}}{\bar{X}}.$$

and

$$\mu_{Y|X}^{MaxK^{[f]},2} = \frac{AV_{Y|X}^{K[f]}}{AV_{Y|X}^T} = \frac{AV(Y | f^{i_{\{n_1-K+1\}}})}{AV(Y^{Total})} = \frac{\frac{\sum_{k=n_1-K+1}^{n_1} \left(\sum_{l=n_1-K+1}^{n_1} f_{(l)}^X \right) + 1 Y_{(k)}}{\sum_{k=n_1-K+1}^{n_1} f_{(k)}^X}}{\bar{Y}}.$$

The above quantities are naturally rewritten in case the roles of X and Y are reversed so that the indices are calculated conditionally on the frequencies of K distinct maximum values of the random variable Y .

7.2. Direct Measure Based on Frequencies

In this subsection, we propose and briefly discuss a special measure V defined for the purpose of comparing directly the frequencies of K distinct maximum (or minimum) values of two data sets that do not depend on time. The definition of the V measure applied to frequencies is based on the definition of the direct measure V of Makris and Vonta ([1])

defined for time points. For two non-dependent data sets which are realizations of size n of two random variables X and Y , we have

$$V_{X,Y}^{MK[f]N} = \begin{cases} V_{X,Y}^{MinK[f]N} = \frac{1}{K} \sum_{r=1}^K (f_{(r)}^X - f_{(r)}^Y)^2 \\ V_{X,Y}^{MaxK[f]N} = \frac{1}{K} \sum_{r=n-K+1}^n (f_{(r)}^X - f_{(r)}^Y)^2 \end{cases} \quad (21)$$

where $f_{(k)}^Q$ is the frequency of the k th distinct ordered observation of the random variable Q . Observe that

$$V_{X,Y}^{MK[f]N} \geq 0,$$

$$V_{X,Y}^{MK[f]N} = 0, \text{ if } X = Y$$

and

$$V_{X,Y}^{MK[f]N} = V_{Y,X}^{MK[f]N}.$$

The above properties imply that the measure $V_{X,Y}^{MK[f]N}$ is a proper statistical measure.

8. Parameter N

The parameter N stands for the number of data sets (which may be time dependent or not) that are analyzed, thus creating in each case an index matrix of dimension $N \times N$.

In case many data sets are to be analyzed, namely the value of N is large and therefore the Matrix $[\mu]$ is difficult to deal with, there is a need to discard some data sets that are not important to be present in the analysis (e.g., when the values of the μ indicators are very small). We introduce therefore an additional parameter ν and we have a new notation $N^{[\nu]}$ which displays the dependence of N on ν . Our purpose for doing that is to discard from the final analysis those data sets that are not significant based on a criterion.

The criterion for keeping a data set is its closeness to another data set which can be measured by a typical distance, e.g., the Euclidean distance or the absolute value based on the indices $\mu_{(i,j)}^{MKN}$. The selection of the data sets that remain in the matrix is performed in two steps. In the first step, the ν smallest absolute differences for $j = 1, \dots, N$,

$$d_{(i,j)}^{MKN} = |\mu_{(i,i)}^{MKN} - \mu_{(i,j)}^{MKN}|$$

are kept in each row with the result being that the initial $N \times N$ matrix becomes an $N \times \nu$ matrix. Let us, for simplicity, denote these differences by $\{d_{i1}, d_{i2}, \dots, d_{i\nu}\}$ for each row $i = 1, \dots, N$. In the second stage, the absolute differences on each row are summed up as $s_i = \sum_{\lambda=1}^{\nu} d_{i\lambda}, i = 1, \dots, N$.

Finally, the ν smallest of those sums are selected and the corresponding rows are kept into the matrix (so the $N \times \nu$ matrix reduces to the final $\nu \times \nu$ matrix) and gives rise to a matrix denoted by $[\mu]^{MKN^{[\nu]}}$. To be more explicit, the resulting matrix contains the rows with row sum

$$\{s_{(1)}, s_{(2)}, \dots, s_{(\nu)}\} \quad (22)$$

9. Conclusions

In this work, a method of data analysis was presented based on the indices μ^{MKN} , which are calculated as a ratio of two statistical functions. We have studied the parameters involved in the indices and how these parameters affect the indices. More specifically, we examined the parameter (function) μ with various examples and an application to economics and marketing on price elasticity of demands. We have also studied the parameter M and how we can use the maximum and minimum values simultaneously to perform a cross-correlation of times. Furthermore, we examined the reduction of the size of the parameter N by discussing how some data sets that are not significant (important) to be

present in the analysis, can be discarded (e.g., when the values of their indices are very big), namely, we reduce the size of the parameter N and thus the number of the data sets kept in the analysis. Finally, the indices for data independent of time (namely random variables) are discussed through explanatory examples.

The proposed indices can be used as powerful statistical tools in similarity matching problems involving pattern matching, anomaly identification and/or frequent pattern detection. The applicability of the proposed methodology goes beyond neuroscience or physiology and epidemiology all of which have been mentioned in this work. Indeed, such techniques can play a vital role in various scientific fields like financial mathematics, economics, management, geosciences, stylometry or music retrieval. Such examples include among many others, the identification of companies with similar growth patterns, products with similar selling patterns and seismic waves not similar in spotting geological irregularities. Finally, music retrieval and plagiarism in literature and music will greatly benefit from the implementation of the proposed methodology.

Author Contributions: Conceptualization, K.M.; Data curation, K.M.; Formal analysis, K.M.; Methodology, K.M., I.V. and A.K.; Project administration, I.V. and A.K.; Supervision, I.V. 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 the Hellenic National Public Health Organization and are available from the corresponding author with the permission of the Hellenic National Public Health Organization.

Acknowledgments: The authors wish to express their appreciation to the anonymous reviewers and the Academic Editor for their valuable and constructive comments that greatly improved the quality of the manuscript. The authors would like to thank the Department of Epidemiological Surveillance and Intervention of the Hellenic National Public Health Organization for providing the influenza-like illness (ILI) rate data, collected weekly through the sentinel surveillance system. This work is part of the Doctoral Thesis of the first author. The first author wishes to acknowledge the financial support from the Papakyriakopoulos scholarship, of the Department of Mathematics of the National Technical University of Athens. The last author wishes to acknowledge the Laboratory of Statistics and Data Analysis of the University of the Aegean.

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

References

1. Makris, K.; Vonta, I. Presentation of coupling analysis techniques of maximum and minimum values between N sets of data using Matrix $[\mu]^{[MKN]}$. *Int. J. Math. Eng. Manag. S* **2021**, under review. Available online: [http://www.math.ntua.gr/\\$\sim\\$svonta/Makris_IJMEMS_preprint.pdf](http://www.math.ntua.gr/$\sim$svonta/Makris_IJMEMS_preprint.pdf) (accessed on 25 March 2021).
2. Iglesias F.; Kastner W. Analysis of Similarity Measures in Time Series Clustering for the Discovery of Building Energy Patterns. *Energies* **2013**, *6*, 579–597.
3. Lin J.; Keogh E.; Leonardi S.; Chiu B. *A Symbolic Representation of Time Series with Implication for Streaming Algorithm*; University of California: La Jolla, CA, USA, 2003.
4. Serra J.; Arcos J.L. *A Competitive Measure to Assess the Similarity Between Two Time Series*; Spanish National Research Council: Barcelona, Spain, 2012.
5. Kullback, S.; Leibler, R. On Information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
6. Csiszar, I. Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat on Markhoschen Ketten. *Magyar Tudományos Akadémia Közleményei* **1963**, *8*, 84–108.
7. Cressie, N.; Read, T.R.C. Multinomial goodness-of-fit tests. *J. R. Stat. Soc. B* **1984**, *5*, 440–454.
8. Mattheou, K.; Lee, S.; Karagrigoriou, A. A model selection criterion based on the BHHJ measure of divergence. *J. Statist. Plann. Infer.* **2009**, *139*, 128–135.
9. Toma, A. Optimal robust M-estimators using divergences. *Stat. Prob. Lett.* **2009**, *79*, 1–5.
10. Huber-Carol, C.; Balakrishnan, N.; Nikulin, M. S.; Mesbah, M. *Goodness of Fit Tests and Model Validity*; Birkhauser: Boston, FL, USA, 2002.

11. Meselidis, C.; Karagrigoriou, A. Statistical inference for multinomial populations based on a double index family of test statistics. *J. Statist. Comput. Simul.* **2020**, *90*, 1773–1792.
12. Jing, J.; Dauwels, J.; Rakthanmanon, T.; Keogh, E.; Cash, S.S.; Westover, M.B. Rapid Annotation of Interictal Epileptiform Discharges via Template Matching under Dynamic Time Warping. *J. Neurosci. Methods* **2016**, *274*, 179–190.
13. Saeed M.; Lieu C. ; Raber G.; Mark RG. MIMIC II: A massive temporal ICU patient database to support research in intelligent patient monitoring. *Comput. Cardiol.* **2002**, *29*, 641–644.
14. Makris, K. Statistical Analysis of Random Waves in SPAR-Type and TLP-Type Wind Turbines. Master's Thesis, National Technical University of Athens, Athens, Greece, 2017. doi:10.26240/heal.ntua.6628. (In Greek)
15. Makris, K. Statistical Analysis of Epidemiological Time Series Data. Master's Thesis, National Technical University of Athens, Athens, Greece, 2018. doi:10.26240/heal.ntua.8270. (In Greek)
16. Makris, K.; Karagrigoriou, A.; Vonta, I. On divergence and dissimilarity measures for multiple time series. In *Applied Modelling Techniques and Data Analysis*; Dimotikalis, I., Ed.; iSTE WILEY: London, UK, 2021; pp. 249–261.
17. Marshall, A.W.; Olkin, I.; Arnold B.C. *Inequalities: Theory of Majorization and Its Applications*; Springer: New York, NY, USA, 2011.
18. Granger, C.W.J. Investigating causal relation by econometric and cross-sectional method. *Econometrica* **1969**, *37*, 424–438.
19. Engle, R.F.; Granger, C.W.J. Co-integration and error correction: representation, estimation, and testing. *Econometrica* **1987**, *55*, 251–276.
20. Johansen, S. Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica* **1991**, *59*, 1551–1580.
21. Bierens, H. Nonparametric cointegration analysis. *J. Econ.* **1997**, *77*, 379–404.
22. Quintos, C. Fully modified vector autoregressive inference in partially nonstationary models. *J. Am. Stat. Assoc.* **1998**, *93*, 783–795.
23. Hubrich K.; Luetkepohl H.; Saikkonen P. A review of system cointegration tests. *Econ. Rev.* **2001**, *20*, 247–318.
24. Marshall, A. *Principles of Economics*; Macmillan: New York, NY, USA, 1890.