Research on Short-Time Wind Speed Prediction in Mountainous Areas Based on Improved ARIMA Model

Zelin Zhou 1, Yiyan Dai 2,*, Jun Xiao 3,4, Maoyi Liu 5, Jinxiang Zhang 2 and Mingjin Zhang 2

1 China 19th Metallurgical Corporation, Chengdu 610031, China
2 Department of Bridge Engineering, Southwest Jiaotong University, Chengdu 610031, China
3 CCC Second Highway Engineering Co., Ltd., Xi’an 710199, China
4 Shaanxi Union Research Center of University and Enterprise for Bridge Intelligent Construction, Xi’an 710199, China
5 Chongqing Construction Investment (Group) Co., Ltd., Chongqing 400010, China

* Correspondence: dyy@my.swjtu.edu.cn

Abstract: In rugged mountain areas, the lateral aerodynamic force and aerodynamic lift caused by strong winds are the main reasons for the lateral overturning of trains and the destruction of buildings and structures along the railroad line. Therefore, it is important to build a strong wind alarm system along the railroad line, and a reasonable and accurate short-time forecast of a strong wind is the basis of it. In this research, two methods of constructive function and time-series decomposition are proposed to pre-process the input wind speed for periodic strong winds in mountainous areas. Then, the improved Auto-Regressive Integrated Moving Average model time-series model was established through the steps of a white noise test, data stationarity test, model recognition, and order determination. Finally, the effectiveness of the improved wind speed prediction was examined. The results of the research showed that rational choice of processing functions has a large impact on wind speed prediction results. The prediction accuracy of the improved ARIMA model proposed in this paper is better than the results of the traditional Seasonal Auto-Regressive Integrated Moving Average model, and it can quickly and accurately realize the short-time wind speed prediction along the railroad line in rugged mountains. In addition, the improved ARIMA model has verified its universality in different mountainous places.

Keywords: wind speed forecast; data pre-processing; auto-regressive integrated moving average; seasonal auto-regressive integrated moving average; rugged mountain areas

1. Introduction

With economic development and social progress, more and more railways have been built in rugged mountainous areas. Many studies have therefore been carried out on wind parameters in the mountainous area [1,2]. High winds in mountainous areas can cause damage to buildings along the railway line and affect construction operations [3]. To prevent disasters in advance, it is necessary to establish a strong wind warning system [4]. However, the non-linearity and non-stationary characteristics of mountain winds pose a challenge to the prediction of wind speeds [5,6].

The main methods of predicting wind speed in mountainous areas include numerical weather prediction, neural networks, and statistical methods. Numerical weather prediction (NWP) can take into account the relationship between wind speed and topography based on Weather Research and Forecasting (WRF) [7], but it can only achieve mesoscale forecasts and needs to be combined with other models to achieve high accuracy [8]. The application of neural networks in wind speed prediction is mature, and it has a strong non-linear fitting ability, which can achieve good prediction effects for non-stationary wind in mountainous areas. Many neural network models have achieved good results in mountain wind speed prediction, such as Artificial Neural Network (ANN) [9], Back Propagation...
(BP) [10], ELMAN [11], Long Short-Term Memory (LSTM) [12], etc. Moreover, hybrid models usually behave better than single models [13–15]. However, the training time of the neural network is longer, and although the hybrid model can get better prediction results, the steps are more complicated.

The statistical method is widely applied in the prediction of wind speed in mountainous areas. The Auto-Regressive Integrated Moving Average (ARIMA) model is representative of the statistical method [16], which is simple to operate as a short-term wind speed prediction. In addition, fewer parameters of this model need to be determined, and the method for determining the parameters is well established [17,18]. However, the input data required by the model need to be stationary and do not include potential seasonal factors [19]. Wind speeds in mountainous areas are often non-stationary and the original data without processing input directly into the model cannot achieve the desired results. Therefore, Seasonal Auto-Regressive Integrated Moving Average (SARIMA) was proposed to solve the above problem and has been applied in wind speed prediction in rugged mountainous areas [20,21]. However, the SARIMA has more parameters than the ARIMA model and is more difficult to manipulate.

To solve the above problem of the ARIMA model and improve the quality of data feed into the ARIMA model. This research proposes two methods for preprocessing the data, including the constructive function method and the time-series decomposition method. The parameter regularity of these two methods is discussed based on the historical wind speed data from different sites. Then, the processed data are input into the ARIMA model to predict the future wind speed. The results show that the improved ARIMA model based on the proposed two methods can conduct accurate forecasting and behave better than the SARIMA model. To further confirm the reliability of the improved ARIMA model, these models are used to forecast another wind speed in the mountains. The results also show the superiority of the model. Finally, some main conclusions are present.

2. Methodology
2.1. ARIMA

The Auto-Regressive Integrated Moving Average model (ARIMA) is a combination of the Auto-Regressive model (AR) and the Moving Average model (MA), which is a means of forecasting future data and trends based on historical data. The expression of the ARIMA with parameters \( p, d, \) and \( q \) is shown in Equation (1).

\[
X_t - \varphi_1X_{t-1} - \cdots - \varphi_pX_{t-p} = \varepsilon_t - \theta_1\varepsilon_{t-1} - \cdots - \theta_q\varepsilon_{t-q}
\]

where \( p \) is the parameter for the autoregressive term, and \( q \) is the parameter for the moving average term. \( d \) stands for the differential number of times, and original data become stationary data after \( d \) differentials times. \( X_t \) is a stationary time series with zero mean. \( \varepsilon_t \) is stationary white noise with zero mean and the variance is \( \sigma^2 \). The process of the ARIMA model is shown in Figure 1.

![Figure 1. The process of the ARIMA model.](image_url)

As shown in Figure 1, to confirm the quality of data that are input to the model, the original data should be conducted in the white noise test and the stationarity test. The white noise test adopts the LB statistics test [22]. The original hypothesis \( H_0 \) is that if
the \( p \)-value obtained from the test is less than 0.05, the original hypothesis is considered to be rejected (i.e., the original data are not white noise). The stationarity test uses the Augmented Dickey–Fuller test (ADF). The original hypothesis \( H_0 \) is that there is a unit root (i.e., the data are non-stationary). If the original hypothesis is accepted, differential treatment is performed to make the data stationary. In contrast, if the original hypothesis is rejected, the data are considered stationary. The Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) are used to determine the suitable type of model for the input data. Moreover, the Bayesian Information Criterion (BIC) is applied for determining the appropriate parameters of the model, which is shown in Equation (2).

\[
BIC = 2k \ln(n) - 2 \ln(L)
\] (2)

where \( k \) is the number of the parameters. \( L \) stands for the likelihood function, and \( n \) is the sample number.

2.2. SARIMA

The Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model is based on the ARIMA theory and takes into account the influence of seasonal terms. The SARIMA model has a good prediction effect for data with strong periodicity. The general form is SARIMA \((p, d, q) \times (P, D, Q)s\), where the significance of parameters \( p, d, \) and \( q \) is the same as that of the ARIMA model. The parameter \( P \) represents the order of seasonal autoregression, \( Q \) is the order of the seasonal moving average, and \( D \) is the order of the seasonal differential. In addition, the \( s \) is the cycle length of the season. The expression of SARIMA can be seen in Equations (3)–(5).

\[
\varphi(B)\varphi(B^s)(1 - B^s)^D X_t = c + \theta(B)\Theta(B^s)\epsilon_t
\] (3)

\[
\varphi(B^s) = 1 - \varphi_1B^s - \cdots - \varphi_P B^{s-P}
\] (4)

\[
\Theta(B^s) = 1 - \Theta_1B^s - \cdots - \Theta_Q B^{s-Q}
\] (5)

where \( X_t \) is a stationary time series. \( B^s \) denotes the seasonal backward shift operator. \( \epsilon_t \) is stationary white noise with zero mean and the variance is \( \sigma^2_\epsilon \). The modeling process of the SARIMA is similar to the ARIMA. The difference is that the SARIMA takes into account the potential periodicity in the order determination, which is reflected in the seasonal term of the model. Because of eliminating the influence of periodic factors, SARIMA prediction results are often better than the ordinary ARIMA model. Additionally, the SARIMA model does not need to process the input data to remove the trend term and the seasonal term. The essence of the SARIMA is the ARIMA multiplicative model considering seasonal attributes. The SARIMA model has a good predictive effect on periodic data. However, the parameters of the SARIMA are more than the ARIMA, so it takes longer time than the general ARIMA and the uncertain parameters will also affect the prediction results.

2.3. Improved ARIMA

The pre-processing of data can eliminate the non-stationarity of original data, which solves the problem of low prediction accuracy when the order parameter of the ARIMA model is too low, and the model is complex and time-consuming when the order parameter is too high. Moreover, the potential regularity of input data cannot be identified by the model. To solve the problem, this research proposes two methods to preprocess the input data to achieve the purpose of optimizing the ARIMA model, namely the constructive function method and the time-series decomposition method. The process of the improved ARIMA model is shown in Figure 2.
To describe conveniently, the training set data are recorded as \( \{X_{t1}\} \), in which the data used to input the ARIMA model are recorded as \( \{X_{t2}\} \). \( \{X_{t3}\} \) is a term that stands for the constructive function or the time-series decomposition. There is a relationship between \( \{X_{t1}\} \), \( \{X_{t2}\} \), and \( \{X_{t3}\} \) as in Equation (6). The \( \{X_{t4}\} \) stands for the testing set of data.

\[
\{X_{t2}\} = \{X_{t1}\} - \{X_{t3}\}
\] (6)

### 2.3.1. Constructive Function Method

Due to wind in rugged mountainous areas being complex and non-stationary, a periodic function \( \{X_{t3}\} \) is constructed to eliminate the non-stationarity of the data for making the input model data stable. The specific step is to reduce the training set \( \{X_{t1}\} \) by a periodic function \( \{X_{t3}\} \) to get \( \{X_{t2}\} \), and predict \( \{X_{t1}\} \) after removing the trend term. The expression of \( \{X_{t3}\} \) is shown in Equation (7).

\[
\{X_{t3}\} = K \cdot \cos\left(\frac{i}{\omega}\right) + C
\] (7)

where \( K \) is the magnification of periodic functions, and \( i \) is the number of the data. \( \omega \) represents the period of the periodic function, and \( C \) is a constant.

### 2.3.2. Time-Series Decomposition Method

To remove the influence of the period term in the training set data on the prediction, this research adopts the statsmoedls to decompose the historical measured data. The statsmoedls can extract the period components from a one-dimensional time series. The data apply a convolution filter to estimate trends to obtain results. Then, the trend is deleted from the sequence, and the average value of the detrended sequence for each period is the returned periodic component \([23,24]\).

The statsmoedls model decomposes the training set data into a trend term, a period term \( \{X_{t3}\} \), and a residual term. The decomposition principle is to estimate the trend by applying a convolution filter to the data. Then, the trend is removed from the time series, and the average of the trend series at different times is a periodic component returned. Moreover, to increase the effectiveness of the period term on input model data, the decomposition term is multiplied by a magnification coefficient as a treatment function, the expression of which is shown in Equation (8).

\[
\{X_{t3}\} = K \cdot S(\varphi, t)
\] (8)

where \( K \) is a magnification coefficient, \( S \) is the decomposition function of the seasonal term of the statsmoedls. \( \varphi \) represents the sampling period for decomposition.

### 3. Cases Studies

#### 3.1. Original Data

This research selected the measured 1-min average wind speed data in two mountainous areas as objects of study, with high wind speed at site A and low
wind speed at site B. In addition, wind speed varies greatly at both locations, with high amplitude of maximum and minimum wind speeds. The number of data points is 1500, of which 1000 are divided as the training set \( \{X_{t1}\} \) and the remaining 500 as the test set \( \{X_{t2}\} \). The data visualization of A and B is plotted in Figure 3.

![Data visualization](image)

**Figure 3.** Modeling data visualization: (a) Dataset division of A site; (b) dataset division of B site.

The first-order twelve-step difference removes the trend and period terms from the original data, due to the fact that the original data cannot pass the stationarity test and the white noise test. After conducting first-order twelfth-order differencing, the \( p \)-values of the ADF test are \( 1.8 \times 10^{-18} \) in site A and \( 2.0 \times 10^{-14} \) in site B, which is significantly less than the hypothetical value. Furthermore, both the SARIMA model and the two methods proposed in this research can eliminate trend and period terms from the data, so the three methods mentioned above were used to conduct the wind prediction for sites A and B to illustrate the superiority of the proposed methods.

3.2. Modeling
3.2.1. Data Pre-Processing

This research takes site B as an example and preprocesses the training set data of site B. The processing functions are the constructed periodic function and the decomposition function based on the statsmodels decomposition. For data pre-processing, the original training set data are subtracted from the processing function. The processing functions are shown in Figures 4 and 5, respectively. The data for the input model are plotted in Figure 6.

![Periodic function](image)

**Figure 4.** Constructed periodic function.
3.2.2. Selection of Parameters

After preprocessing the data, an improved ARIMA model can be established as shown in Figure 2. To further improve the accuracy of wind speed prediction, the parameters of the processing function are optimized and analyzed. To evaluate the quality of the prediction results, the corresponding indicators for its accuracy evaluation are introduced as shown in Equations (9)–(12).

\[
R^2 = 1 - \frac{\sum (\hat{X}(i) - X(i))^2}{\sum (X - X(i))^2}
\]  
(9)

\[
MAE = \frac{1}{N}\sum_{i=1}^{N}|X(i) - \hat{X}(i)|
\]  
(10)

\[
MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{X(i) - \hat{X}(i)}{X(i)}\right|
\]  
(11)
\[ \text{MSE} = \frac{1}{N} \sum_{i=1}^{N} [X(i) - \hat{X}(i)]^2 \]  

(12)

where \(X(i)\) is the measured wind speed, \(\hat{X}(i)\) is the predicted wind speed, and \(\bar{X}\) stands for the mean value of the measured wind speed value. Furthermore, the \(R^2\) and \(\text{MAPE}\) are selected as the main evaluation indicators. The parameters that affect the results of predicted wind speed are the magnification coefficient \(K\), the period of the constructed function \(\omega\) of the constructive function method, the magnification coefficient \(K\), and the sampling period \(\phi\) of the time-series decomposition method.

The variation of parameters will affect the prediction results and observing changes in evaluation indicators can determine the optimal parameters. The results of the parameter discussion are shown in Figure 7, where the parameters are taken at intervals and densely selected from the parameters that behave well. In Figure 7, the results of fitting the sample points using the function fitting method are shown and given the practical time-consuming issues. A simple polynomial function can be implemented to capture the trend in the variations of parameters.

It can be seen from Figure 7 that changing the parameters of the processing function will affect the accuracy of the prediction results. Moreover, the prediction results of \(A\) and \(B\) can be obtained at the same regularity due to parameter changes. To visualize the regularity, the results are fitted by polynomial functions as shown in Figure 7. In the case of parameter changes, there are the following regularity:

1. For the time-series decomposition method, it can be seen from the periodic term in Figure 5 that the length of the repeated period of the decomposed subsequence is \(\phi\). When the value of \(\phi\) is 0, the length of the decomposition sequence under each period is 0. When \(\phi > 500\), the number of cycles in the training set is less than 2. So, the parameter \(\phi\) is selected as \(0 < \phi \leq 500\), which is reasonable. From Figure 7a,b, with the increase of \(\phi\) value, the overall trend of \(\text{MAPE}\) decreases, and the overall trend of \(R^2\) increases. The results show that the higher the value of \(\phi\), the more the series obtained from the statsmodels decomposition is able to extract the periodicity of the wind speed to enhance the prediction results.

2. Parameter \(K\) in the time-series decomposition method amplifies the effect of a decomposed subsequence on the original data. From Figure 7c,d, it can be seen that different \(K\) will get different predictions. When \(K\) is less than the optimal value, with the increase of the \(K\) value, \(R^2\) increases, and \(\text{MAPE}\) decreases. When \(K\) is greater than the optimal value, the change regularity of \(R^2\) and \(\text{MAPE}\) is the opposite. The dashed line in Figure 7c,d indicates the location of the optimal value of \(K\). Based on the results of the parameter discussion, the optimal value of \(K\) is 0.5 times the value of the maximum value of the original wind speed divided by the maximum value of the processing function.

3. Figure 7e,f describe the changes in the indicators that evaluate the prediction results as the parameter \(\omega\) of the constructive function method changes. The change regularity of the prediction indicators can be fitted by a quadratic polynomial. From the fitting results, it can be easily seen that the extreme value of the function corresponds to the optimal prediction result. When \(\omega\) reaches the optimal value, \(R^2\) is the largest and \(\text{MAPE}\) is the smallest. The corresponding optimal \(\omega\) values of \(A\) and \(B\) are both 26, which can provide a reference for wind speed prediction in similar rugged mountainous areas.

4. Similar to \(K\) in the time-series decomposition method, \(K\) in the constructive function method is also related to the wind speed of the original data. Through Figure 7g,h when \(K = 5\) in site \(A\) and \(K = 35\) in site \(B\), \(R^2\) is the largest, and \(\text{MAPE}\) is the smallest, the model can achieve the best-predicted results. The optimal parameter \(K\) corresponding to the two sites is different because of the different wind speeds. The small wind speed in site \(A\) corresponds to the small \(K\) value, and site \(B\) is the opposite. The maximum wind speed can provide a reference for the value of \(K\).
Figure 7. Parameter discussion results: (a) $\varphi$, $A$, time-series decomposition method; (b) $\varphi$, $B$, time-series decomposition method; (c) $K$, $A$, time-series decomposition method; (d) $K$, $B$, time-series decomposition method; (e) $\omega$, $A$, constructive function method; (f) $\omega$, $B$, constructive function method; (g) $K$, $A$, constructive function method; (h) $K$, $B$, constructive function method.
3.3. Results and Analysis

According to Section 3.2.1, the prediction results are affected by parameter change, and the methods of selecting the optimal parameters have been discussed above. To describe the results that used the improved ARIMA model with the optimal parameters, the predicted results are plotted in Figures 8 and 9. Figures 8 and 9 show the prediction results of the three methods, the time-series decomposition method, the construction function method, and SARIMA, for sites A and B, respectively. Moreover, Tables 1 and 2 correspond to the evaluation indicators for the predicted results in A and B, respectively.

![Figure 8. Predicted results in site A.](image)

![Figure 9. Predicted results in site B.](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>MAPE</th>
<th>MSE</th>
<th>MAE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMA</td>
<td>23.10%</td>
<td>5.91</td>
<td>1.72</td>
<td>0.68</td>
</tr>
<tr>
<td>Improved ARIMA by constructive function method</td>
<td>9.28%</td>
<td>0.84</td>
<td>0.72</td>
<td>0.95</td>
</tr>
<tr>
<td>Improved ARIMA by time-series decomposition method</td>
<td>9.22%</td>
<td>0.99</td>
<td>0.76</td>
<td>0.94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>MAPE</th>
<th>MSE</th>
<th>MAE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMA</td>
<td>17.02%</td>
<td>0.09</td>
<td>0.21</td>
<td>0.94</td>
</tr>
<tr>
<td>Improved ARIMA by constructive function method</td>
<td>7.75%</td>
<td>0.01</td>
<td>0.09</td>
<td>0.99</td>
</tr>
<tr>
<td>Improved ARIMA by time-series decomposition method</td>
<td>6.92%</td>
<td>0.02</td>
<td>0.09</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Based on the prediction results, it can be seen that the method proposed in this research is better able to obtain future one-step predictions in short-time wind speed prediction in rugged mountainous areas. However, as seen in Figure 7, if the parameters are not chosen appropriately, the proposed method’s prediction results are rather inferior to the SARIMA model, which shows the importance of parameter selection. If the appropriate parameters are selected, the predicted results for sites A and B are much better than the SARIMA model. Moreover, the proposed method is easy to operate, stable, and time-consuming, and has no requirements for the configuration of the operating equipment.

However, the prediction results in Figures 8 and 9 are based on the optimal parameters after discussion. It is therefore necessary to validate the methods for determining the parameters obtained in Section 3.2.1 to verify the applicability of the improved ARIMA model. A set of wind speeds in rugged mountainous areas was arbitrarily selected to build the model and the selected dataset was divided into a training set and a testing set. Then, the optimal parameters for the appropriate training set were selected according to the regularity identified in Section 3.2.1. The predicted results of the improved ARIMA with the optimal parameters are plotted in Figure 10. The evaluation indicators are shown in Table 3.

![Figure 10. Predicted results of validation data.](image)

Table 3. Forecast evaluation indicators of validation data.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAPE</th>
<th>MSE</th>
<th>MAE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMA</td>
<td>17.02%</td>
<td>0.09</td>
<td>0.21</td>
<td>0.94</td>
</tr>
<tr>
<td>Improved ARIMA by constructive function method</td>
<td>7.75%</td>
<td>0.01</td>
<td>0.09</td>
<td>0.99</td>
</tr>
<tr>
<td>Improved ARIMA by time-series decomposition method</td>
<td>6.92%</td>
<td>0.02</td>
<td>0.09</td>
<td>0.99</td>
</tr>
</tbody>
</table>

The results of the validation show that the improved ARIMA with the optimal parameters determined by the proposed method still behaves better than the SARIMA model. This demonstrates the applicability of the proposed method to wind speed prediction in mountainous areas.

4. Conclusions

This research proposed two methods to improve the quality of the input data for the ARIMA model, including the constructive function method and the time-series decomposition method. Based on the history of wind speed of two different sites, A and B, the parameters of the proposed methods are discussed. The preprocessed data were input into the ARIMA models for predicting the future wind speed. Furthermore, to verify the reliability of the proposed improved ARIMA, another dataset of wind speed in the mountain area was conducted for prediction. Some conclusions were obtained, as below.

1. This research proposes two different methods (i.e., constructive function method and time-series decomposition method) for pre-processing the data input to the ARIMA model, which improves the quality of the wind speed input to the ARIMA model.
2. To solve the problem of parameter uncertainty in the proposed methods, based on the wind speed of two typical mountain wind fields, this research carries out a discussion on parameter optimization. The discussion gets the regularity for determining the relevant parameters of the proposed methods, and the regularity of identifying parameters is validated using the other dataset of mountain wind fields.

3. Combining the ARIMA and proposed methods (i.e., constructive function method and time-series decomposition method), the improved ARIMA model is conducted for prediction in rugged mountainous areas. The results show the proposed model behaves well and operates easily, and the evaluation indicates that the improved ARIMA model has superior performance to the SARIMA model. The improved ARIMA model can provide a new method for prediction in mountainous areas.

Author Contributions: Conceptualization, Y.D. and M.Z.; methodology, Z.Z.; software, J.X.; validation, Z.Z., J.X. and M.L.; formal analysis, J.Z.; investigation, M.Z.; resources, M.Z.; data curation, M.L.; writing—original draft preparation, Y.D.; writing—review and editing, M.Z.; visualization, Z.Z.; supervision, Z.Z.; project administration, J.X.; funding acquisition, Z.Z. 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.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to Unit-related requirements.

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

References


20. Zhang, W.; Lin, Z.; Liu, X. Short-Term Offshore Wind Power Forecasting—A Hybrid Model Based on Discrete Wavelet Transform (DWT), Seasonal Autoregressive Integrated Moving Average (SARIMA), and Deep-Learning-Based Long Short-Term Memory (LSTM). *Renew. Energy* 2022, 185, 611–628. [CrossRef]

21. Liu, X.; Lin, Z.; Feng, Z. Short-Term Offshore Wind Speed Forecast by Seasonal ARIMA—A Comparison against GRU and LSTM. *Energy* 2021, 227, 120492. [CrossRef]

