Modern Applied Science; Vol. 9, No. 11; 2015
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Imputation of Missing Values in Daily Wind Speed Data Using Hybrid AR-ANN Method

Osamah Basheer Shukur[1] & Muhammad Hisyam Lee[1]

[1] Department of Mathematical Sciences, Universiti Teknologi Malaysia, Johor, Malaysia

Correspondence: Muhammad Hisyam Lee, Department of Mathematical Sciences, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia. Tel: 60-7553-4236. Fax: 60-7556-6162. E-mail: mhl@utm.my; assomeeh@yahoo.com

Received: March 27, 2015 Accepted: April 8, 2015 Online Published: June 18, 2015

doi:10.5539/mas.v9n11p1 URL: http://dx.doi.org/10.5539/mas.v9n11p1

Abstract

Wind speed data collection process faces several problems as failure of data observing devices. Therefore, wind speed data naturally contains missing values. Imputing these missing values using an effective method is important before performing time series analysis. The classical methods as linear, nearest neighbor, and state space may not provide accurate imputations when the wind speed contains nonlinearity. In this study, the hybrid artificial neural network (ANN) and autoregressive (AR) method is proposed for imputing the missing values. ANN is a nonlinear method that is capable of imputing the missing values in wind speed data with nonlinear characteristic. AR model is used for determining the structure of the input layer for the ANN. Listwise deletion is used before AR modeling to handle the missing values. A case study is carried out using daily Iraqi and Malaysian wind speed data. The proposed imputation method is compared with linear, nearest neighbor, and state space methods. The comparison has shown that AR-ANN outperformed the classical methods. In conclusion, the missing values in wind speed data with nonlinear characteristic can be imputed more accurately using AR-ANN. Therefore, imputing the missing values using AR-ANN leads to more accurate performance of time series modeling and analysis.

Keywords: missing values, imputation, wind speed, time series, artificial neural network, Hybrid AR-ANN

1. Introduction

The collection process of wind speed data as one of meteorological time series data faces several tactical problems such as thunderstorms, failure of data observing devices, or other unforeseen errors that lead to increased complexity in data analysis. A sequential data set is required for performing analysis and modeling processes. Therefore, the missing values in wind speed data should be filled and imputed. Missing values imputation can be accomplished using simple methods such as linear, nearest neighbor or others. Although complex methods require additional expertise and specialization, they often outperform the simple methods.

In most meteorological time series data sets, nonlinearity is another problem that may hamper time series analysis using linear methods. In particular, wind speed data suffer from nonlinearity in addition to the missing values problem. In recent papers, ANN was introduced to impute missing values and to handle the nonlinearity of meteorological time series data sets in general and wind speed data set in particular. Junninen, Niska, Tuppurainen, Ruuskanen, and Kolehmainen (2004) introduced several univariate and multivariate methods for imputation of missing values in air quality data sets such as linear, nearest neighbor, self-organizing map (SOM), multi-layer perceptron (MLP), and other methods. The neural networks SOM and MLP performed better than other methods. Coulibaly and Evora (2007) introduced six different types of ANN for infilling missing daily weather records such as daily precipitation and daily extreme temperature series. Kim and Pachepsky (2010) proposed a new method to impute missing daily precipitation data set in Chesapeake Bay Watershed using two-step regression trees and artificial neural networks. Yozgatligil, Aslan, Iyigun, and Batmaz (2013) compared several imputation methods to impute the missing values of spatiotemporal meteorological time series. MLP has been used as a neural network for in filling Turkish meteorological time series data sets. He, Cao, Cao, and Wen (2013) studied the recovery of missing values for wind time series data set in particular. They developed an ensemble learning method based on multiple ANN with MLP. Using ANN to impute missing values was not limited to meteorological time series data as Kornelsen and Coulibaly (2012) introduced ANN as the most

effective method for missing values infilling in soil moisture as hydrometeorological time series data set.

In this study, a hybrid AR-ANN method was proposed based on AR model to reform the problems of this study by imputing missing values and jointly handling the nonlinearity problem. Feed-forward back propagation will be used as a neural network algorithm. AR model will be used only for determining the structure of the input layer for the ANN. AR order can be determined by observing the significant lags in partial autocorrelation function (PACF). Listwise deletion will also be used as the simplest method before AR modeling to handle missing value problems in wind speed time series data sets.

Hybrid AR-ANN model has been used in many recent papers for handling the nonlinearity in full data set of wind speed without any missing value observation. Li and Shi (2010) compared one-hour ahead forecasts for hourly wind speed data set using three different types of artificial neural networks. They used an autocorrelation function (ACF) and a PACF to determine the ANN inputs. Guo, Zhao, Haiyan, and Wang (2012) proposed many methods for wind speed forecasting. One of these methods was a feed-forward neural network whose inputs were determined based on the AR order. H. Liu, Tian, and Li (2012) proposed new hybrid AR-ANN that is similar to the proposed method in the current study. They confirmed that the performance of their method in terms of its predictions was consistently better than others.

The stationarity conditions for the wind speed data were omitted in this stage, and the parameters, signs, and residual series were also omitted from the terms of AR model, because the AR model was used only for determining the structure of the input layer for the ANN such as done in Refs. (Khashei & Bijari, 2010; H. Liu et al., 2012). In many papers, AR model was also used as missing values imputation method. Alosh (2009) studied the impact of missingness on a transition model for longitudinal count data as an extension of the integer-valued autoregressive (INAR). He also considered an application of the generalized autoregressive model for a longitudinal epilepsy data set. Choong, Charbit, and Yan (2009) introduced an autoregressive-model-based missing values imputation method for microarray temporal data set. Their model was especially effective for the situation where a particular time point contains many missing values. Honaker and King (2010) suggested using a first-order autoregressive model AR(1) to deal with the time series properties of the data after performing a listwise deletion as an imputation method of missing values. Applying the listwise deletion was suggested to produce consistent and unbiased attributes of parameters especially for performing data analysis and modeling using model or software that requires a sequential data set (Cheema, 2014; Honaker & King, 2010).

Many other methods of missing values imputation will be compared with hybrid AR-ANN method proposed in this study. Linear method and nearest neighbor method will be presented as more simple methods for imputing missing values. Hybrid AR-ANN method and state space method will be presented as complex methods that need more expertise and scientific specialization. A linear method is summarized by connecting two data points with a linear equation line. It was used for comparing with more complex methods in many recent papers such as in Refs. (Junninen et al., 2004; Kornelsen & Coulibaly, 2012; Norazian, Shukri, Azam, & Al Bakri, 2008). A nearest neighbor method is the simplest imputation method of missing values that can be summarized by replacing the missing values by the nearest neighbor data point. Siripitayananon, Chen, and Jin (2003) proposed their new method by modifying the classical nearest neighbor method to fill the missing wind data every 15 minutes. A nearest neighbor method was also used as a simple method for comparing with other proposed methods such as in Refs. (Junninen et al., 2004; Liew, Law, & Yan, 2011; Waljee et al., 2013). The state space model has been used as an imputation method of missing values. Sarkka, Vehtari, and Lampinen (2004) applied the state space system of Kalman filter and Kalman smoother methods to predict the missing parts of time series data. Tsay (2005) mentioned that the Kalman filter and state smoothing recursion through applying state space model could be used to impute missing values. Root mean square error (RMSE) measurement will be computed for the error of missing values imputation for all imputation methods and all data sets as a statistical criterion to evaluate the adequacy and accuracy of these methods.

The missing values have been distributed into three different parts and with three different proportions. A total of five missing data sets were prepared for both Iraq and Malaysia. Hybrid AR-ANN has been proposed in this study and will be compared with linear, nearest neighbor, and state space methods. The proposed method outperformed other imputation methods.

This paper is organized as follows: Section 2 states the data and framework of this study and presents the proposed method and the other methods theoretically. Section 3 displays and discusses the results and the computational steps of the methods in Section 2. Section 4 provides the conclusions of this study.

2. Material and Method

2.1 Data and Framework of the Study

In this study, two data sets of daily wind speed were collected from two different meteorological stations. Iraqi wind speed is the first data set that was collected from the Mosul Dam Meteorological Station in Mosul, Iraq. It is for four hydrological years (1 October 2000 – 30 September 2004). Malaysian wind speed is the second data set that was collected from the Muar Meteorological Station in Johor, Malaysia. It is for four hydrological years (1 October 2006 – 30 September 2010). The missing values have been distributed into three different parts 1, 3, 6 respectively, and with three different proportions 10%, 20%, 30% respectively. Five sets of data is the total number of missing data sets for both Iraq and Malaysia. First three data sets include 10% of missing values that were distributed into one, three, and six equal parts respectively. Last two data sets were distributed into six equal parts that includes 20%, and 30% of missing values respectively. The framework of this study includes the following:

a) Constructing the most appropriate hybrid AR-ANN after performing the listwise deletion for missing values in the data sets.

b) Applying the linear, nearest neighbor, and state space imputation methods.

c) Comparing the proposed method with other methods to determine what model would provide the best adequacy.

Figure 1 explains the framework of this study.

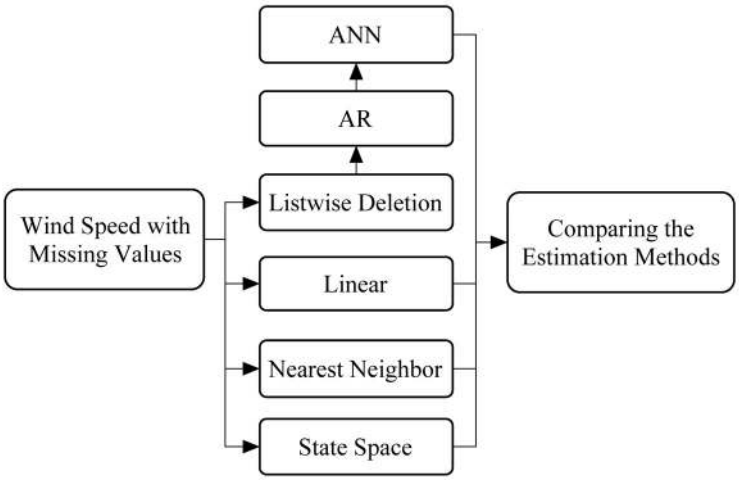

Figure 1. The framework of the study

2.2 Imputation Methods of Missing Values

2.2.1 A hybrid AR-ANN Method

A hybrid AR-ANN will be proposed for imputing the missing values. AR model was used only for determining the input layer structure of ANN. Listwise deletion will be used as the simplest imputation method before AR modeling to deal with missing values in wind speed time series data sets. hybrid AR-ANN was also proposed to handle the nonlinearity of wind speed data. Use of the multilayer feed-forward back propagation neural network for time series forecasting was supported by the ANN toolbox in MATLAB software. Determining the training functions and the transfer functions types of hidden and output layers and other requirements were necessary to create the most appropriate ANN structure.

The types of transfer functions are tan-sigmoid, which generates nonlinear outputs between -1 and +1, log-sigmoid which generates nonlinear outputs between 0 and 1, and linear which generates linear outputs between -1 and +1. Selecting a suitable transfer function is important for obtaining good results. The best training functions for back propagation algorithms are Levenberg-Marquardt and Bayesian regularization. The number of neurons in a hidden layer must be correctly calculated to create an appropriate ANN (Shukur, Fadhil, Lee, & Ahmad, 2014).

The nonlinearity of wind speed data requires the selection of a nonlinear transfer function such as tan-sigmoid and log-sigmoid for hidden layer to filter the nonlinearity. Figure 2 demonstrates the structure of feed-forward back propagation and the transfer functions types.

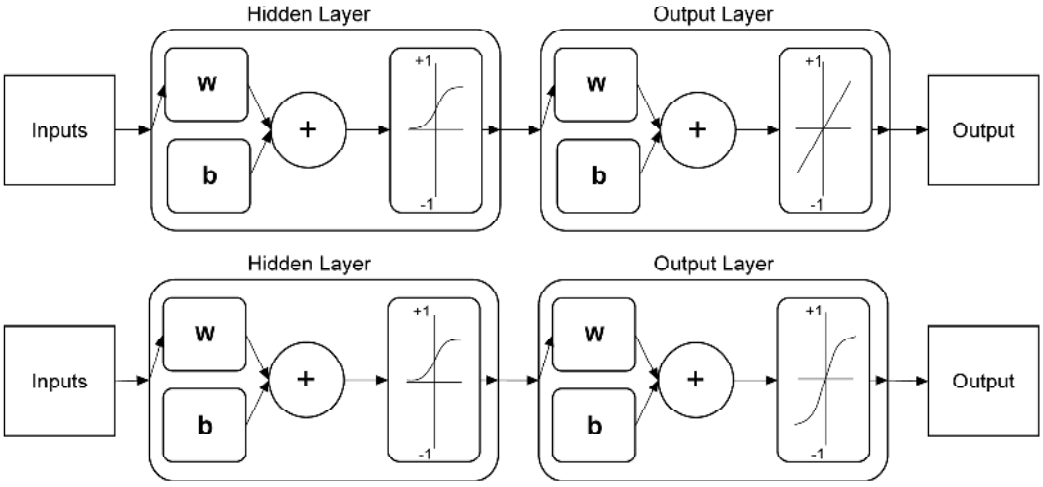

Figure 2. ANN structure and transfer function types

A general expression of the non-seasonal autoregressive AR(p) model is shown such as follows.

$$(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)Y_t = a_t \tag{1}$$

where Y_t is the original time series variable, a_t is the residual series at the current time, AR(p) is a p^{th} order of the autoregressive component, B^i is an i^{th} order of backshift operator, and ϕ is the parameter of the AR model (L.-M. Liu, 2006).

The ACF and the PACF for the original series reflect the number of significant orders for the AR model. The autocorrelation between two observations can be calculated using the variance and covariance as follows.

$$\rho_k = corr(Y_t, Y_{t-k}) = \frac{\operatorname{cov}(Y_t, Y_{t-k})}{\sqrt{\operatorname{var}(Y_t)\operatorname{var}(Y_{t-k})}} = \frac{\operatorname{cov}(Y_t, Y_{t-k})}{\operatorname{var}(Y_t)} = \frac{\gamma_k}{\gamma_0} \tag{2}$$

where $\operatorname{var}(Y_t) = \operatorname{var}(Y_{t-k})$. $\rho_0 = 1$ and $|\rho_k| \le 1$ for $k = \pm 1, \pm 2, \pm 3, \ldots$. $\rho_k = \rho_{-k}$ since the covariance is. $\gamma_k = \gamma_{-k}$. The partial autocorrelation between two observations represents the conditional correlation between them. The partial autocorrelation can be computed using Yule-Walker equation system and successive Cramer's rule. The dying out style in the ACF and the cutting off style after p lags in the PACF of any data set indicate that AR(p) is the most suitable model for this data set. If these features are not available in the ACF and PACF of the original data set, the first or second successive difference, or the first or second seasonal difference for the original data set can be taken to reach to the appropriate decision about AR model.

The stationarity conditions for the wind speed data were omitted in this stage, and the parameters, signs, and residual series were also omitted from the terms of AR model, because the AR model was used only for determining the structure of the input layer for the ANN. Therefore, the total number of ANN inputs will be equal to p. This approach can be called hybrid AR-ANN model (H. Liu et al., 2012) or ANN (Khashei & Bijari, 2010, 2011; Zhang, 2003).

The Akaike information criterion (AIC) will be plotted to measure the adequacy of the best AR model and to confirm the results of ACF and PACF.

Listwise deletion method simply discards observations with missing values. It may produce consistent and unbiased parameters imputations. It was applied for performing data analysis and modeling using AR that

requires a sequential data set.

2.2.2 The Classical Methods

In linear imputation method, two data points are connected with a line of linear model. The linear model will be modified from the simple linear regression as follows:

$$f(x) = b_0 + b(x - x_0) \tag{3}$$

where x is an independent variable with n observations. x_0 is the last known observation before the missing value and x is the first known observation after the missing values. The first and last data point must be valid to apply the linear method and impute missing values. The coefficients b_0 and b are functions for variables x_0 and x respectively, where $b_0 = f(x_0)$ and $b = [f(x) - f(x_0)]/(x - x_0)$.

A nearest neighbor method is the simplest imputation method for imputing the missing values. It can be summarized by replacing the missing values by the last nearest known neighbor data point before the missing value or by the first nearest known neighbor data point after the missing value. K nearest neighbor method has the same strategy but by replacing the missing values by the last K nearest known neighbor data points before the missing value or by the first K nearest known neighbor data points after the missing value (Liew et al., 2011). Most researchers used the normal nearest neighbor method such as Refs. (Junninen et al., 2004; Waljee et al., 2013).

The state space model can be used as a method of missing values imputation. The state space system includes state equation (SE) and observation equation (OE) as follows

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{C}^T a_t \tag{4}$$

and

$$\mathbf{Y}_t = \mathbf{C}\mathbf{X}_t \tag{5}$$

where X_t is m-dimensional state vector $\mathbf{X}_t = \begin{bmatrix} X_{1,t} & X_{2,t} & \ldots & X_{m,t} \end{bmatrix}^T$. $\mathbf{A}$ is $m \times m$ state transition matrix $\mathbf{A} = \begin{bmatrix} K_1 & K_2 & K_3 & \cdots & K_m \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}_{m \times m}$, where $(K_1, K_2, \ldots, K_m)$ are the parameter values, and m is the number of these parameters (Madsen, 2007). $\mathbf{C}$ is $1 \times m$ observation transition matrix $\mathbf{C} = [1 \quad 0 \quad 0 \quad \cdots \quad 0]_{1 \times m}$.

The output of OE represented the fitted series after imputing the missing values. The difference $\mathbf{Y}_t - \mathbf{C}\mathbf{X}_t$ will produce the error series of missing values imputation using state space method. Linear, nearest neighbor, state space methods have been computed using MATLAB software programming.

3. Results and Discussion

The Iraqi wind speed data for the period spanning (1 October 2000 – 30 September 2004) and Malaysian wind speed data for the period spanning (1 October 2006 – 30 September 2010) are plotted in Figure 3.

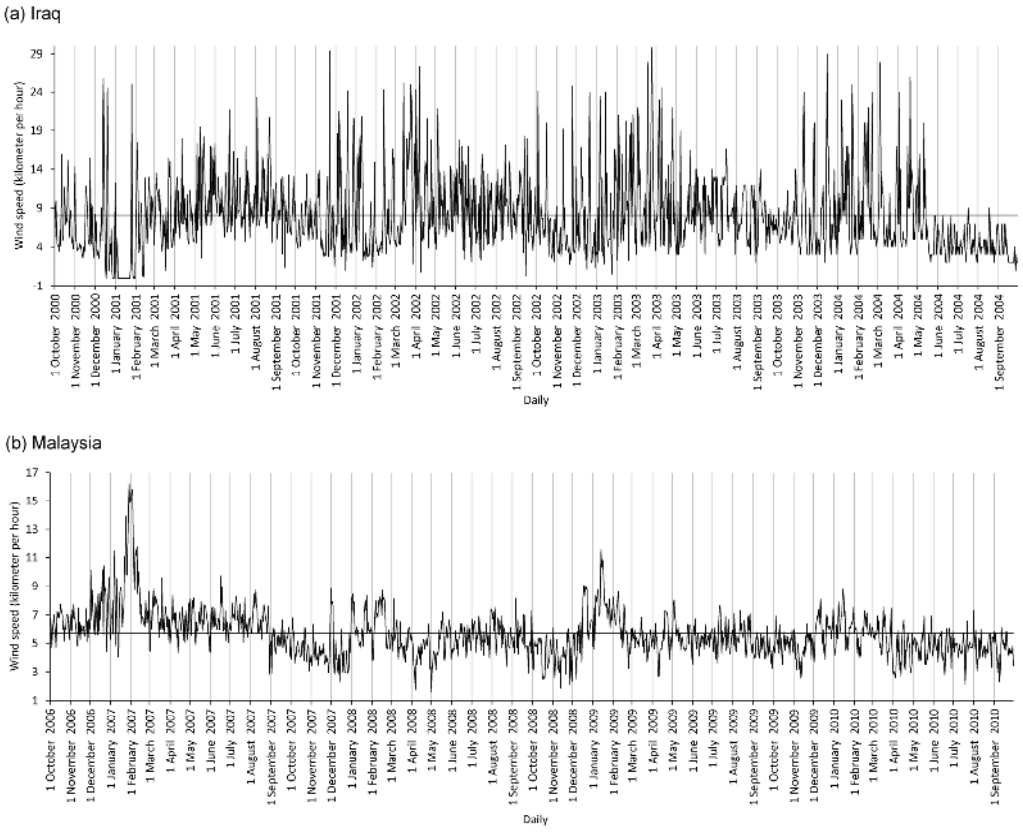

Figure 3. Time series plot of wind speed for Iraq and Malaysia

Full data sets in Figure 3 represent the target series. The output series of each method was compared with the target series in order to get the error series of method imputation.

From Figure 3, it is clear that the pattern and the magnitude of Iraqi wind speed data set is different from those of Malaysian data set. The differences between the Iraqi and Malaysian wind speed data sets can be attributed to the differences in the meteorological environments of Iraq and Malaysia. Iraq faces four monsoon winds seasons yearly, whereas Malaysia faces only two, making the Iraqi data set more complex.

Neural network toolboxes in MATLAB include many types of training algorithms and training functions. Feed forward and back-propagation algorithm with the Levenberg Marquardt and the Bayesian regularization training algorithms were used in this study to produce better results. Bayesian regularization training algorithm with log sigmoid transfer function for hidden layer and tan sigmoid transfer function for output layer gave the best results compared with others.

After performing the listwise deletion method by discarding the missing values, the PACF for the original series will reflect the number of significant order for the AR model. The stationarity conditions were omitted in this stage, because the AR model was used only for determining the structure of the input layer for the ANN. ACF and PACF for all the five data sets had similar styles for both Iraq and Malaysia. Figure 4 illustrates the ACF and PACF of the original Iraqi wind speed data set, and it also demonstrates the ACF and PACF of the original Malaysian wind speed data set.

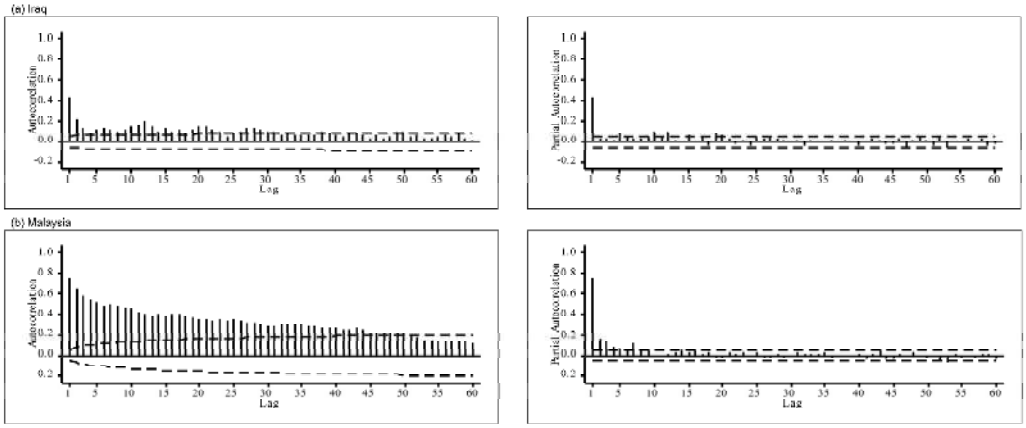

Figure 4. ACF and PACF for Iraqi and Malaysian wind speeds

From Figure 4, ACF exhibited the slow dying out style of the Iraqi and Malaysian non-stationary data sets. The PACF in Figure 4 illustrates the cutting off style after 12 for the Iraqi data set and after 7 for the Malaysian data set. Therefore, AR(12) and AR(7) can be proposed based on the ACF and PACF results for the Iraqi and Malaysian data sets, respectively.

The AIC can be used to measure the adequacy of the best AR model. AIC was plotted to confirm AR(P) results (Figure 5) for the original Iraqi and Malaysian data set series for $AR(1), AR(2), \ldots, AR(20)$.

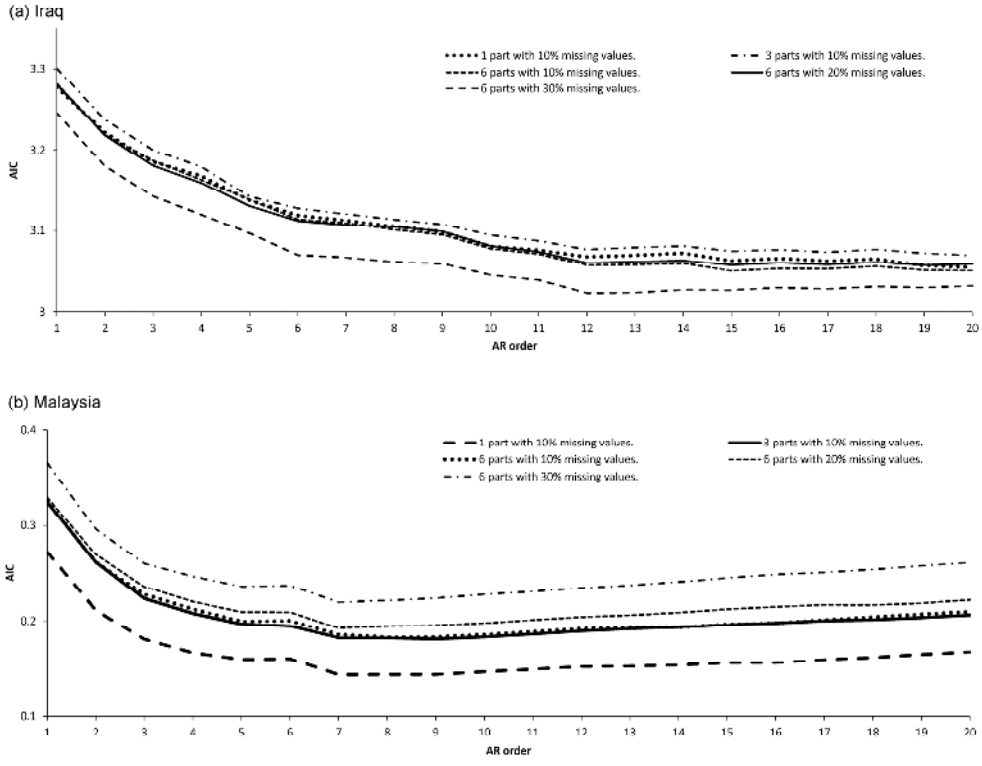

Figure 5. AIC of the ARIMA for Iraqi and Malaysian data sets

From Figure 5, AIC had similar styles for all five Iraqi and Malaysian data sets. Figure 5 shows the quick dying

out style of the AIC values that stabilized after the 12th and 7th values for the Iraqi and Malaysian data sets, respectively. The AIC plots confirmed that the selection of AR(12) and AR(7) was performed correctly for all missing data sets. The parameter imputations of all AR models were significant and the p-values were less than the significant level 5%. The AR(p) model for the wind speed data of Iraq and Malaysia can be expressed respectively as follows:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_{12} Y_{t-12} + a_t \quad (6)$$

and

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_7 Y_{t-7} + a_t \quad (7)$$

The inputs structure of the ANN for both data sets can be considered based on the order of the AR(12) and AR(7) models. In other words, the ANN inputs are equal to 12 and 7 for the Iraqi and Malaysian data sets, respectively, for all five missing data sets. In MATLAB, the input variables must be inserted as rows into one variable. The targets of ANN were the original time series data for Iraqi and Malaysian wind speed. The target series must be inserted as a row into one variable.

MATLAB has improved some of its toolboxes for neural networks. Neural network toolboxes in MATLAB use many types of training algorithms, training functions and transfer functions. The structure of ANN has been constructed by determining the requirements such as follows.

a) Feed-forward back propagation must be determined as a network type.

b) The inputs of ANN for all missing data sets of Iraq were ($Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots, Y_{t-12}$) and for all missing data sets of Malaysia were ($Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots, Y_{t-7}$). Input and target data variables must be imported to construct ANN structure using ANN toolboxes in MATLAB.

c) Levenberg Marquardt and Bayesian regularization are used as training algorithms. Bayesian regularization training algorithm was found to be the best training function in this study. It provided the best results besides using log-sigmoid and tan-sigmoid transfer functions for hidden and output layers respectively.

d) The number of neurons in the hidden layer was determined to be (no. of inputs$\times 2+1$) (Sheela & Deepa, 2013). The number of neurons in the hidden layer was 25 for all missing data sets of Iraq and 15 for all missing data sets of Malaysia, respectively.

e) The weights and biased part were determined randomly depending on ANN toolboxes strategies and all previously determined requirements.

After completing ANN construction, training processes will be the next step to obtain the output series which is also known as fitted series. The difference between fitted series and target series generated the error series or the residual series using ANN method based on AR model. Table 1 explains the RMSE values of the error of missing values imputation for all data sets using hybrid AR-ANN method.

Table 1. The parameter estimators and the testing values of ARIMA

Method	MAPE				
	Set 1	Set 2	Set 3	Set 4	Set 5
Linear					
Iraq	1.88	1.62	2.17	2.56	3.15
Malaysia	0.37	0.47	0.50	0.74	1.02
Nearest Neighbor					
Iraq	1.87	1.61	1.89	2.49	3.48
Malaysia	0.36	0.45	0.40	0.72	0.79
State Space					
Iraq	2.95	2.95	2.95	2.95	2.95
Malaysia	1.74	1.74	1.74	1.74	1.74
Hybrid AR-ANN					
Iraq	1.36	1.51	1.63	2.28	2.84
Malaysia	0.35	0.39	0.361	0.54	0.68

From Table 1, RMSE values for the missing values imputations using hybrid AR-ANN method were 1.88, 1.62, 2.17, 2.56, and 3.15 for the five missing data sets of Iraq, respectively, and 0.37, 0.47, 0.50, 0.74, and 1.02 for

the five missing data sets of Malaysia, respectively.

MATLAB software programming was used in this study to impute the missing values using linear method, nearest neighbor method, and state space method. By comparing the fitted series with the target series, the error series of missing values imputation was obtained and RMSE values were computed.

From Table 1, RMSE for the missing values imputation using linear method were 1.88, 1.62, 2.17, 2.56, and 3.15 for the five missing data sets of Iraq, respectively. As for the five missing data sets of Malaysia, the results were 0.37, 0.47, 0.50, 0.74, and 1.02, respectively. RMSE for the five missing data sets of Iraq using nearest neighbor method were 1.87, 1.61, 1.89, 2.49, and 3.48, respectively, while the results of the five missing data sets of Malaysia were 0.36, 0.45, 0.40, 0.72, and 0.79, respectively. RMSE for the missing values imputation using state space method were 2.95 for all missing data sets of Iraq, and 1.74 for all missing data sets of Malaysia.

The previous results in Table 1 indicated that a hybrid AR-ANN method outperformed the classical methods of missing values imputation for all missing data sets of Iraq and Malaysia. The capability of a hybrid AR-ANN method for handling the nonlinearity and imputing the missing values jointly is the main reason for the high accuracy of results compared with the results of the classical methods.

The similarity of RMSE values of state space method can be attributed to the similar behaviors of the different data sets for both Iraq and Malaysia. Similar behaviors were clear from the ACF and PACF figures for those data sets of Iraq and Malaysia and also from the similarity of their AIC figures and AR models as in Figure 4 and Figure 5. The impact of the place and the ratio of missing values were also very clear from the results in Table 1. Once the proportions and the number of missing parts were changed, the adequacies of results were also changed.

From Table 1, Malaysian results were more adequate than Iraqi result for all methods. This may be due to the differences in the meteorological environments of Iraq and Malaysia and indicate that the meteorological environment of Iraq is more complex than Malaysia.

For more obvious comparison between the accuracy of a hybrid AR-ANN imputation results and the accuracy of the results of other classical imputation methods, the improvement percentage that can be written as in Equation (8) was computed using RMSE values in Table 1 to reflect the percentage of improvement after imputing the missing values.

$$\text{Improvement} = \frac{\text{RMSE}_{\text{method1}} - \text{RMSE}_{\text{method2}}}{\text{RMSE}_{\text{method1}}} \times 100\% \tag{8}$$

Method 2 in Equation (8) refers to the hybrid AR-ANN imputation method, while Method 1 represents the other classical imputation methods sequentially. Figure 6 displays the improvement percentage of missing values imputation of all estimated Iraqi wind speed data sets using a hybrid AR-ANN method compared to the classical methods, while Figure 7 displays the improvement percentage of missing values imputation of all estimated Malaysian wind speed data sets using a hybrid AR-ANN method compared to the classical methods.

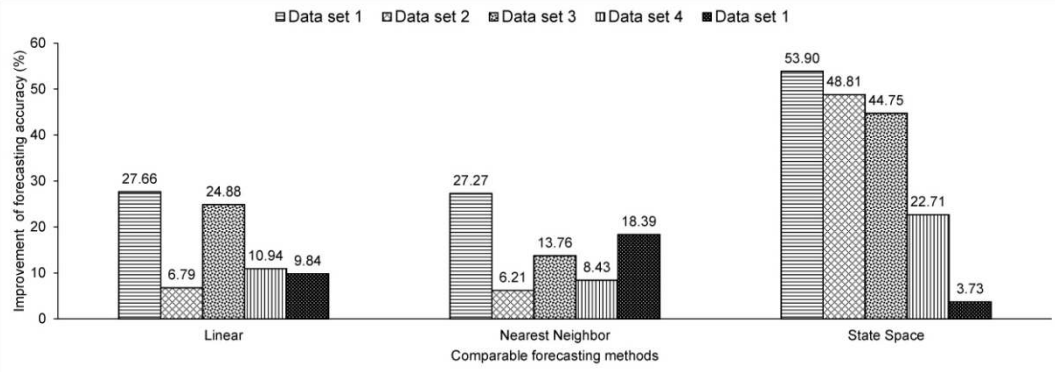

Figure 6. The improvement percentage of all imputed Iraqi missing datasets using AR-ANN compared to other classical methods

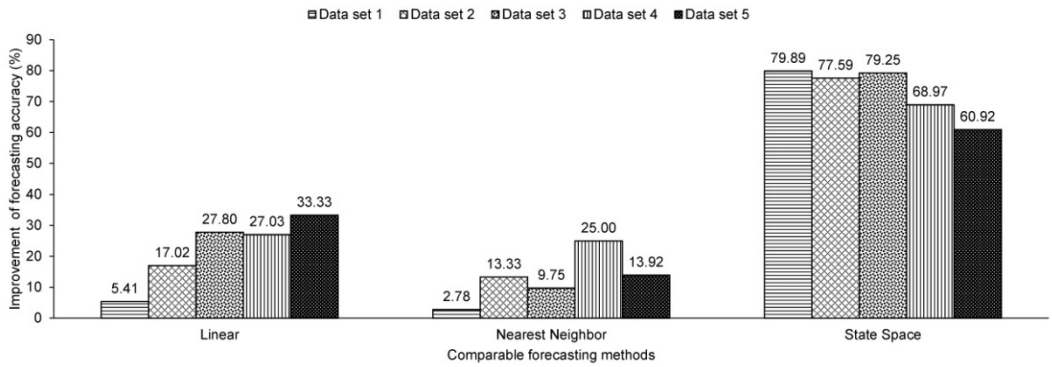

Figure 7. The improvement percentage of all imputed Malaysian missing datasets using AR-ANN compared to other classical methods

Figure 6, Figure 7, and Table 1 confirmed that the missing values imputation results of hybrid AR-ANN method outperformed all other studied classical methods. This happens because of the hybrid AR-ANN method is capable to impute the missing values and handle the nonlinearity in wind speed data sets.

4. Conclusion

The imputation of missing values is important before the modeling and analyzing of time series. The comparison between the proposed method and the classical imputation methods had shown that hybrid AR-ANN significantly outperformed the others. In conclusion, the missing values in wind speed data with nonlinear characteristic can be imputed more accurately using the proposed method. Therefore, imputing the missing values using the proposed method leads to more accurate performance of time series modeling and analysis.

References

Alosh, M. (2009). The impact of missing data in a generalized integer-valued autoregression model for count data. *Journal of biopharmaceutical statistics, 19*(6), 1039-1054.

Cheema, J. R. (2014). A Review of Missing Data Handling Methods in Education Research. *Review of Educational Research*, 1-22. http://dx.doi.org/10.3102/0034654314532697

Choong, M. K., Charbit, M., & Yan, H. (2009). Autoregressive-model-based missing value estimation for DNA microarray time series data. *Information Technology in Biomedicine, IEEE Transactions on, 13*(1), 131-137.

Coulibaly, P., & Evora, N. (2007). Comparison of neural network methods for infilling missing daily weather records. *Journal of hydrology, 341*(1), 27-41.

Guo, Z., Zhao, W., Haiyan, L., & Wang, J. (2012). Multi-step forecasting for wind speed using a modified EMD-based artificial neural network model. *Renewable Energy, 37*, 241-249.

He, H., Cao, Y., Cao, Y., & Wen, J. (2013). Ensemble learning for wind profile prediction with missing values. *Neural Computing and Applications, 22*(2), 287-294.

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science, 54*(2), 561-581.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment, 38*(18), 2895-2907.

Khashei, M., & Bijari, M. (2010). An artificial neural network (p,d,q) model for time series forecasting. *Expert Systems with Applications, 37*(1), 479-489.

Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Appl. Soft. Comput., 11*, 2664–2675.

Kim, J.-W., & Pachepsky, Y. A. (2010). Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *Journal of hydrology, 394*(3), 305-314.

Kornelsen, K., & Coulibaly, P. (2012). Comparison of Interpolation, Statistical, and Data-Driven Methods for Imputation of Missing Values in a Distributed Soil Moisture Dataset. *Journal of Hydrologic Engineering, 19*(1), 26-43.

Li, G., & Shi, J. (2010). On comparing three artificial neural networks for wind speed forecasting. *Appl. Energy.,*

87, 2313–2320.

Liew, A. W.-C., Law, N.-F., & Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in bioinformatics, 12*(5), 498-513.

Liu, H., Tian, H., & Li, Y. (2012). Comparison of two new ARIMA-ANN and ARIMA-Kalman hybrid methods for wind speed prediction. *Appl. Energy., 98*, 415–424.

Liu, L. M. (2006). *Time Series Analysis and Forecasting* (2nd ed.). Illinois, USA: Scientific Computing Associates Corp.

Madsen, H. (2007). *Time Series Analysis*: CRC Press.

Norazian, M. N., Shukri, Y. A., Azam, R. N., & Al Bakri, A. M. M. (2008). Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia, 34*, 341-345.

Sarkka, S., Vehtari, A., & Lampinen, J. (2004). *Time series prediction by Kalman smoother with cross-validated noise density.* Paper presented at the Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on.

Sheela, K. G., & Deepa, S. (2013). Review on methods to fix number of hidden neurons in neural networks. *Mathematical Problems in Engineering, 2013*.

Shukur, O. B., Fadhil, N. S., Lee, M. H., & Ahmad, M. H. (2014). Electricity Load Forecasting using Hybrid of Multiplicative Double Seasonal Exponential Smoothing Model with Artificial Neural Network. *Jurnal Teknologi, 69*(2).

Siripitayananon, P., Chen, H. C., & Jin, K. R. (2003). A modified nearest neighbors approach for estimating missing wind data. *International Journal of Smart Engineering System Design, 5*(3), 149-160.

Tsay, R. S. (2005). *Analysis of financial time series* (Vol. 543): John Wiley & Sons.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., ... Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open, 3*(8).

Yozgatligil, C., Aslan, S., Iyigun, C., & Batmaz, I. (2013). Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theoretical and applied climatology, 112*(1-2), 143-167.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing., 50*, 159-175.