SHORT-TERM FORECASTING OF WIND SPEED IN VIETNAM USING A STATISTICAL MODEL

DỰ BÁO NGẮN HẠN TỐC ĐỘ GIÓ Ở VIỆT NAM SỬ DỤNG MÔ HÌNH THỐNG KÊ

Trung Kien Hoang[1,*], Le Minh[1]

DOI: http://doi.org/10.57001/huih5804.2024.237

ABSTRACT

This article proposes an Autoregressive Integrated Moving Average time series model following Box-Jenkins method for a wind speed prediction. The model is based on the historical data measured at Phuoc The, Binh Thuan province from 0:00 am to 11:50 pm daily in one year. With the horizon of forecast sample of 20 predictions, it is found that early forecasts yielded good results when being verified by mean absolute error and root-mean-square error criteria.

Keywords: *ARIMA; short-term forecasting; wind energy.*

TÓM TẮT

Bài báo này trình bày nghiên cứu về dự báo tốc độ gió sử dụng mô hình chuỗi thời gian trung bình trượt kết hợp tự hồi quy dựa trên phương pháp Box-Jenkins. Mô hình dự báo sử dụng dữ liệu đo gió thực thế tại Phước Thể, Bình Thuận trong vòng một năm với dạng dữ liệu 10 phút. Phương pháp được kiểm nghiệm với mẫu dự báo 20 điểm kế tiếp, kết quả cho thấy chất lượng dự báo tốt khi so sánh với dữ liệu đo thực tế dựa trên các tiêu chuẩn về trung bình sai số tuyệt đối và căn bậc hai của trung bình bình phương sai số.

Từ khóa: *ARIMA, dự báo ngắn hạn, năng lượng gió.*

[1]Department of Applied Engineering and Technology, University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology, Vietnam

[*]Email: hoang-trung.kien@usth.edu.vn

Received: 10/4/2024

Revised: 16/5/2024

Accepted: 25/7/2024

ABBREVIATION

ACF	Autocorrelation Function
PACF	Partial Autocorrelation Function
AR	Autoregressive
MA	Moving average
ADF	Augmented Dickey-Fuller
ARIMA	Autoregressive Integrated Moving Average
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
AIC	Akaike Information Criterion

1. INTRODUCTION

Renewable energy has recently witnessed considerable development to handle the energy crisis and environmental pollution and reach the goal of emission reduction set by various countries. In Vietnam, the government has been carrying out action plans toward the net zero target set in the 2021 United Nations Climate Change Conference, those plans are reflected in the Power Development Plan 8 [1]. Due to the very high potential of wind energy, electricity generation based on this primary renewable energy source is gaining more and more attention with rapid technological advancement. Due to the uncertain nature of the wind, wind power integration is facing a big challenge to ensure a smooth and safe operation of the power grid. Increasing wind energy penetration requires an elaborate consideration of wind turbine control, reserve limit as well as dispatch of power units. It is, therefore, necessary to predict accurately wind speed for the efficient operation of the system. According to different planning strategies, the time horizon for wind speed prediction varies. They are often categorized into four types [2]: ultra-short-term (a few seconds to 30 minutes, useful for real-time grid operations and turbine control), short-term (from 30 minutes to 6 hours, useful for load dispatch planning), medium-term (from 6 hours to 24 hours, useful for operational security in electricity market), and long-

term (from 1 day to 7 days, useful for maintenance scheduling and optimize the operating cost).

As mentioned, the short-term wind speed forecast can provide significant information for grid operators to make proper decisions. The time-series analysis methods are common approaches used for wind speed forecasting. In this research, the statistical method will be used to predict the future values of wind speeds based on historical ones. The Autoregressive Integrated Moving Average (ARIMA) will be addressed due to its simple but efficient algorithm compared to other techniques. For example, numeric weather prediction is a physical-based method requiring a description of the area such as the roughness and obstacles as well as weather parameters of temperature, pressure, etc. [3]. Also, it is inexpensive to build and develop and it proves to be accurate for short-term predictions. Nevertheless, considerable errors arise when the interval of prediction increases, and any error of the recorded data in the past would propagate to future predictions since the model is heavily dependent on the earlier events [4]. In this research, the ARIMA method will be employed for a short-term forecast of the wind speed in a specific region in Vietnam, which is the Huong Linh, located in Quang Tri province. The region to choose this region is due to the higher potential of wind energy in that area with an average of wind speed about 6-8 m/s. The article is organized as follows: Section II presents the forecasting model using ARIMA method, section III shows key results and discussion for the wind speed forecast in Phuoc The, Binh Thuan province followed by the conclusion.

2. METHODOLOGY

2.1. ARIMA model

The data in this article is recorded at Phuoc The, Binh Thuan in one year, from 01/01/2015 to 31/12/2015. Each data point represents the average value of the wind speed in 10-minute time intervals. ARIMA takes on the observed values in the past, models them as a linear combination, and allows one to predict the forthcoming values. Specifically, the AR model makes use of the values from the previous period while the MA model relates the value being predicted with the earlier residuals. The general form of the ARIMA method involves parameters p, d, and q. The parameter d determines the degree of first differencing required. p and q are the orders (of lags) of the AR and MA, respectively described as:

$$y_t = c + \sum_{i=1}^{p} \Phi_i y_{t-i} + e_t + \sum_{i=1}^{q} \theta_i e_{t-i} \quad (1)$$

where c is a constant, Φ_i represents the autoregressive (AR) parameter, θ_i is the moving average (MA) parameter, e_t is the error at time t.

2.2. Stationary verification and solution to non-stationary time series

Stationary stochastic processes are defined to have a constant mean, constant variance and the covariance between values y_t and y_{t-k} of lags k must be the same for any value of t [5]. AR and MA models can only be applied when the samples form a stationary time series. Non-stationary time series are those with trend or seasonality. Therefore, to address this problem, a method called differencing is implemented to enhance stationarity. The differenced series illustrates the change between successive observations of the recorded wind speed. Occasionally, the differenced data will not appear to be stationary, hence, differencing the data a second time might be necessary. In practice, it is seldom to go beyond second-order difference [6]. For time series that have seasonal patterns, the difference taken into account is the number of units of time between consecutive seasons. The presence of a unit root resulting in the non-stationary pattern of the series is inspected using the Augmented Dickey-Fuller (ADF) test. The null hypothesis of the test is that a unit root is present, that is the modulus of the parameter is equal to unity, if the p-value of the ADF statistic is smaller than the critical value which is conventionally chosen to be 5%, the null hypothesis is rejected.

2.3. Model identification and parameters estimation

Initial guesses of the orders are determined based on the autocorrelation function (ACF) and partial autocorrelation function (PACF) based on their geometric patterns with only the values exceeding the confidence interval would be considered. Moreover, ACF can also be applied to investigate the stationarity of the time series. For example, a series' ACF with an increasing trend decreases slowly while the ACF of a seasonal series also exhibits a seasonal pattern. In detail the calculation of the ACF is given by:

$$\mathrm{corr}(y_t, y_{t-k}) = \rho_k = \frac{y_k}{y_0} \quad (2)$$

where ρ_k is the value of the ACF between lags k, y_k is the covariance between lags k and y_0 is the variance of the time series.

The relation between ACF and PACF can be described by:

$$\rho_j = \theta_{k1}\rho_{j-1} + ... + \theta_{kk}\rho_{j-k} \quad (3)$$

(j = 1, 2, …, k)

The variables θ_{kj} will be solved with θ_{kk} be the value of the PACF at lag k.

The parameters of the ARIMA(p,d,q) model are determined such that they maximize the log-likelihood function. Also, the log-likelihood function can be used subsequently in Akaike information criterion (AIC) to determine the best model out of the several chosen ones. Specifically, the log-likelihood function is given as:

$$L(\Phi,\theta,\delta_a^2) = -\left[\frac{n}{2}\ln\left(\delta_a^2\right) + \frac{S(\Phi,\theta)}{2\delta_a^2}\right] \quad (4)$$

where:

$$S(\Phi,\theta) = \sum_{t=2}^{n} e_t^2 \quad (5)$$

$S(\Phi,\theta)$ is the conditional sum-of-square function needless to calculate the log-likelihood function directly, the parameters are estimated such that $S(\Phi,\theta)$ takes on the minimum value [7]. Identification of the orders determined based on ACF and PACF might sometimes be insufficient. Therefore, the AIC provides a means to measure the comparative evaluation among time series models. Mathematically AIC is defined by (6) [8].

$$AIC = 2k - 2\ln\left(\hat{L}\right) \quad (6)$$

where $\hat{L}$ is the maximum value of the likelihood function of the model, k is the number of parameters.

2.4. Forecasting function

To predict the values of the time series, the minimum mean square error forecast is implemented. The time series function for a general ARIMA(p,d,q) at lead time l with respect to the origin time t is represented in terms of weight ψ and errors as follows:

$$\begin{aligned} y_{t+l} &= e_t(l) + \widehat{Z}_t(l) \\ &= \left(e_{t+l} + \psi_1 e_{t+l-1} + ... + \psi_{l-1} e_{t+1}\right) + \left(\psi_{l+1} e_{t-1} + ...\right) \end{aligned} \quad (7)$$

where

$$\varphi(B)\psi(B) = \theta(B) \quad (8)$$

$$\varphi(B) = \Phi(B)(1-B)^d \quad (9)$$

$e_t(l)$ is the error of the forecast and $\widehat{Z}_t(l)$ is the forecast at lead time l and the weight ψ can be calculated using (10) by equating the coefficients of the B operator.

$$\begin{aligned} &\left(1-\varphi_1 B - ... - \varphi_{p+d} B^{p+d}\right)\left(1+\psi_1 B + \psi_2 B^2 + ...\right) \\ &= \left(1-\theta_1 B - \theta_2 B^2 ... - \theta_q B^q\right) \end{aligned} \quad (10)$$

Another representation of the forecasting function can be formulated as:

$$\widehat{y}_t(l) = \sum_{j=1}^{p+d} \varphi_j \widehat{y}_t(l-j) - \sum_{j=l}^{q} \theta_j e_{t+l-j} \quad (11)$$

where $\widehat{y}_t(-j)$ implies the actual observation y_{t-j}.

A prediction $\widehat{Z}_t(l+1)$ at Z_{t+l+1} can be improved as soon as the observation at time t + 1 is obtained. The forecast at Z_{t+l+1} is then measured with respect to the new time origin t + 1.

$$\widehat{Z_{t+1}}(l) = \widehat{Z}_t(l+1) + \psi_1 e_t \quad (12)$$

3. RESULTS AND DISCUSSION

The time series is separated into two sets, the training set containing 104 data points and the testing set sample of 20 data points. Firstly, the ACF is examined in Fig. 1 for 104 data points to determine the order of the model as well as the visual assessment to make the initial guess of the order.

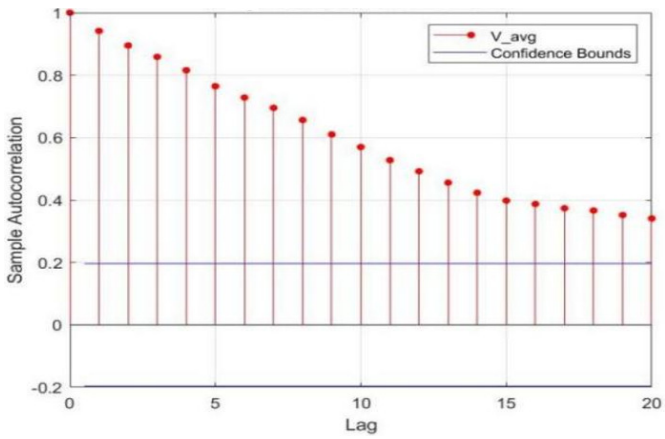


Fig. 1. ACF plot of the observations with 104 data points

As can be seen, the high value of ACF at different lags indicates a correlation between different lagged observations. The function decreases exponentially in a relatively regular manner from lag 1, therefore, AR(1) should be investigated. The AR(1) model can be represented as:

$$y_t = \theta y_{t-1} + \varepsilon_t \quad (13)$$

After multiplying both sides of (13) with y_{t-k} taking the expectation and dividing by the variance at lag 0, i.e., k = 0, for a general AR(p) model it can be shown that:

$$\Phi(B)\rho_k = 0 \quad (14)$$

Therefore, an AR(1) model ACF can be given by:

$$\rho_k = \Phi_{1\rho_{k-1}} \quad (k > 0) \tag{15}$$

With the requirement for a stationary process that requires that $\Phi(B) = 0$ has roots lying outside the unit circle, the autocorrelation function decreases exponentially from lag 1, in agreement with the plot. To check the validity of the reasoning, it should be true that $\rho_1 = \Phi_1$ at lag 1. The AR parameter Φ_1 was 0.8858 while ρ_1 was 0.88 with the error of 0.65% and the model is verified.

The investigation of the PACF from Fig. 2 also suggests an AR(1) model. The partial autocorrelations of an AR(p) can be calculated from the Yule-Walker equation (Appendix). Specifically, the PACF of AR(1) function is $\Phi_{11} = \rho_1$ at lag 1 and 0 beyond. In other words, the PACF function for this model cuts off after lag 1 which agrees with the PACF plot.

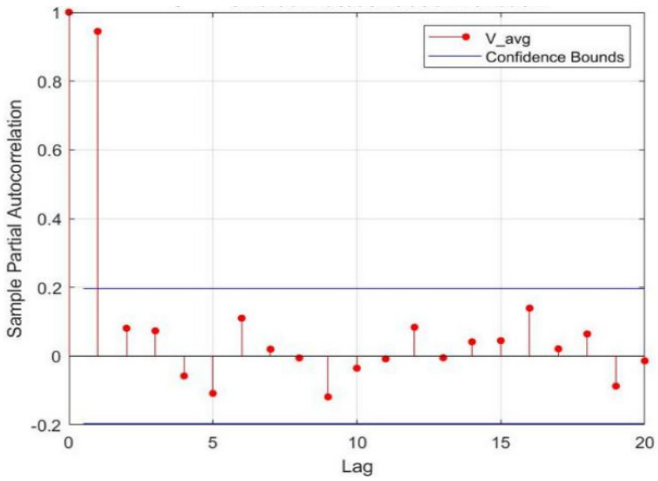
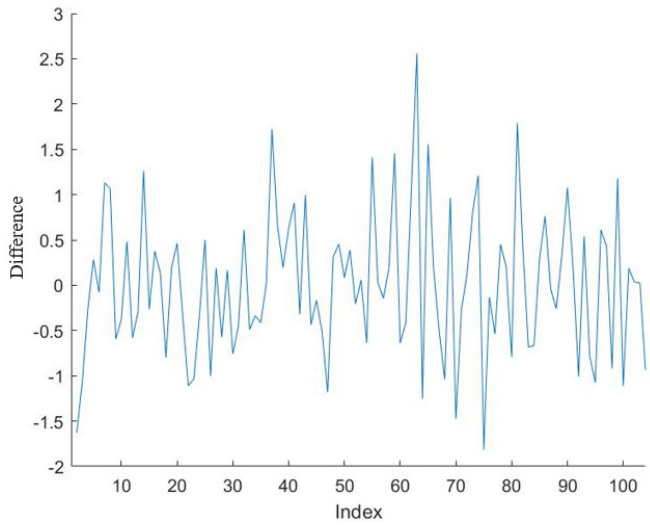
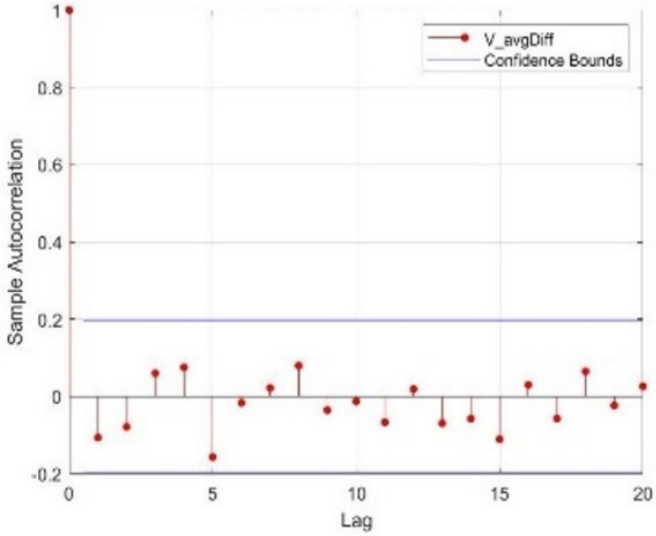


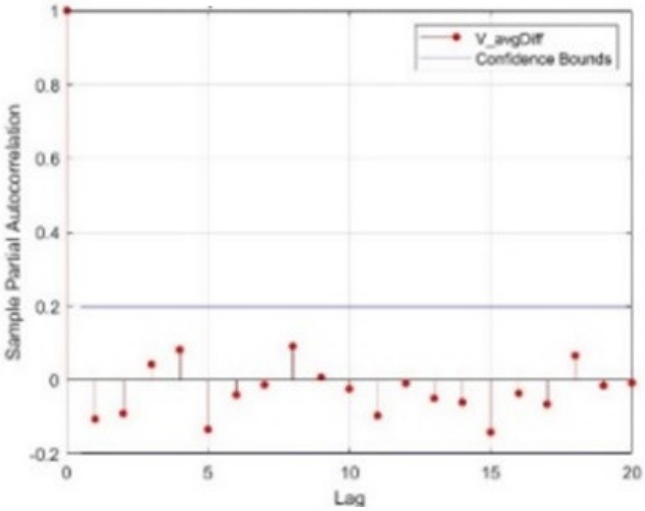
Fig. 2. PACF plot of the observations with 104 data points



(a)

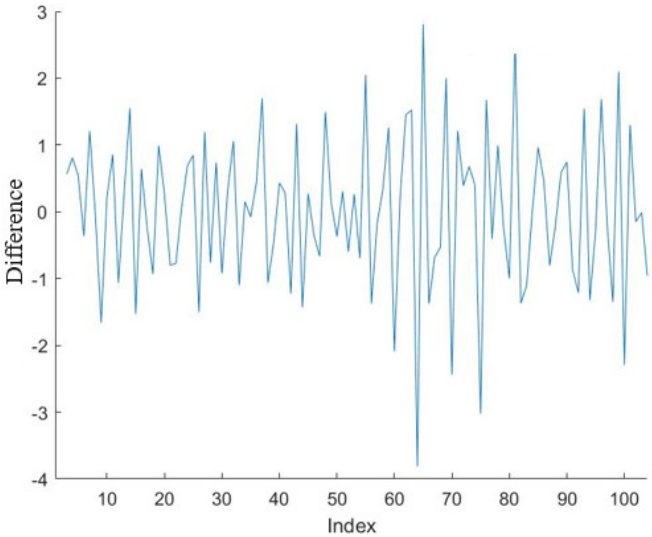


(b)

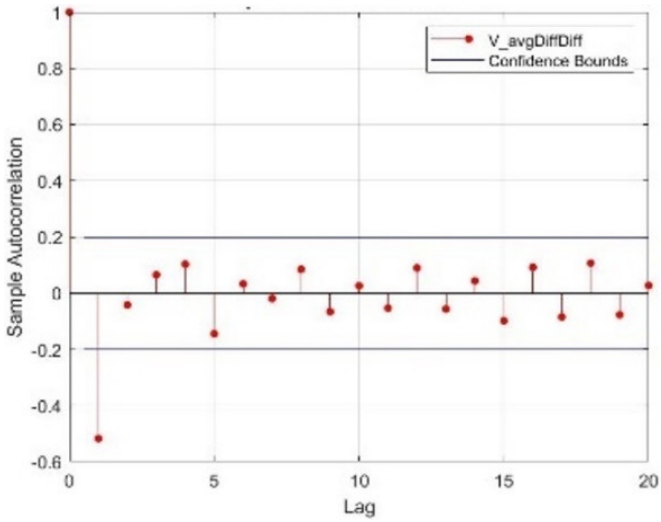


(c)

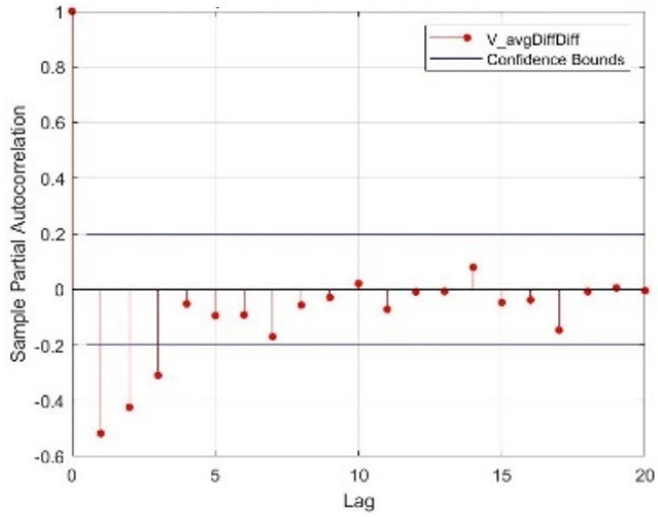
Fig. 3. First difference: a) Time series data. b) ACF plot. (c) PACF plot



(a)



(b)



(c)

Fig. 4. Second difference: a) Time series data. b) ACF plot. (c) PACF plot

However, it should be noted that the ACF decays slowly with all values outside the confidence interval under the null hypothesis that the series is completely random or white noise. Particularly, the ACF implies a high correlation between the lags. These high correlation coefficients can be attributed to the strong propagation from lag 1 to the successive lags. This statement can be verified by observing the PACF plot in Fig. 2, the only significant partial autocorrelation is between consecutive observations indicated by a spike at lag 1.

Moreover, it should be noted that the AR parameter Φ_1 of 0.9445 is relatively close to unity. With that being said, the possibility of a unit root should be considered carefully. Suppose that the AR parameter Φ_1 was actually unity, the AR(1) model is therefore expressed as:

$$y_t = c + y_{t-1} \tag{16}$$

The representation for a random walk model with growth should have been avoided since this is the indication of non-stationarity. From a different perspective, this model is equivalent to a first difference from which it can be deduced that differencing is clearly needed. For further investigation on the presence of a unit root of the time series, Augmented Dickey-Fuller (ADF) tests of the original compared with the differenced time series are carried out. With the null hypothesis that the series contains a unit root, the results of the tests are summarized in Table 1. The p-value of the test for the original series is larger than the chosen significance level. Also, the t-statistic is higher than the critical value. Therefore, it failed to reject the null hypothesis. Contrarily, in the test for the differenced series, a very low p-value is obtained meaning that such a sample can hardly be obtained if the null hypothesis was true. Consequently, the null hypothesis can be rejected.

With ACF and PACF of the differenced series being plotted in Fig. 3, the assertion that the series is stationary has again been verified, it can be observed that all of the values of the ACF and the PACF of the differenced time series are within the confidence interval. Also, the two functions are relatively similar in pattern, this is the result of the elimination of the propagating effect between lags of the ACF. Hence, the differenced series has no correlation between the lags, i.e., the ARIMA(0,1,0) random walk model can be considered, this model can be represented as follows:

$$(1-B)y_t = e_t \tag{17}$$

Table 1. Summary of the ADF test of the original series and the differenced series with the significance level be 0.05 and the null hypothesis that the time series contains a unit root

	p-value	t-statistics	Critical value	Rejected null
Original series	0.4728	-0.4831	-1.9443	False
Differenced series	1.000e-3	-11.3440	-1.9443	True

Nonetheless, over-differencing should be avoided once stationarity has been achieved since it induces unnecessary complexity in the model. The series is differenced to the second order (d = 2) for illustration. Visually, Fig. 4 seems to be stationary, however, the correlation coefficient of the ACF becomes negative and it exceeds the confidence interval, investigation of the PACF plot also shows a similar result. Over differencing might cause non-invertibility which eventually leads to biased estimates during the parameter estimation step.

Other models have been tested also in order to have an overview of all possible models for the time series. There are two models other than ARIMA(0,1,0) that stand out and their parameters are summarized in Table 2 by adding an MA(1) model, ARIMA(0,1,1) is selected to prevent over differencing usually indicated by a highly negative correlation in the first lag. These models are selected based on the Akaike Information Criterion (AIC) and their statistics. It can be reaffirmed that ARIMA(0,1,0) model gives the best fit due to its lowest AIC value. Although the ARIMA(1,0,0) is included in the table, compared to the other two integrated models, it has the highest AIC value as expected from an undifferenced model. Therefore, among the three models, it is the least favored. Table 3 interprets the statistics for the variance of the residuals from fitting the three models discussed above. Although all tabulated residuals p-value for the models are statistically significant, two stationary models, i.e., ARIMA(0,1,0) and ARIMA (0,1,1) reveal higher values of t-statistics and correspondingly smaller p-value compared to AR(1) model confirming the goodness of fit of differenced models for the time series being investigated. The forecast for 20 data points which is equivalent to 2 hours and 10 minutes are plotted in Fig. 5 for the two differenced models.

Table 2. Three possible models with their parameters and AIC summarized

Model	Parameters			AIC
	Constant	AR	MA	
(0,1,0)	0	0	0	249.156
(0,1,1)	0	0	$\Phi_1 = -0.1311$	249.713
(1,0,0)	c = 0.4454	$\Phi_1 = 0.9445$	0	251.6713

The forecast is implemented such that the time origin t in the algorithm is updated constantly as presented by (13) in section 2. In other words, for each origin t, the value at the lead time l = 1 is predicted. It can be observed that the predictions based on ARIMA(0,1,0) and ARIMA(0,1,1) are relatively similar. This can be attributed to the fact that the extra MA(1) parameter is unnecessary and barely improves over-differencing of the time series being studied. In Fig. 3, lag-1 autocorrelation although negative, it does not cross the determined confidence interval as usually seen in over-differenced series. To evaluate the performance, an error analysis is carried out by utilizing RMSE and MAE expressed by (18) and (19), respectively with the prediction of 20 data points to investigate the capability of the ARIMA model in further predictions.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}\left(y_t - \widehat{y_t}\right)^2}{T}} \quad (18)$$

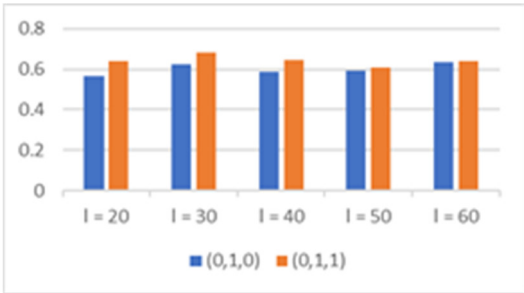
$$\text{MAE} = \frac{\sum_{t=1}^{T}\left|y_t - \widehat{y_t}\right|}{T} \quad (19)$$

where T is the number of points in the time series data.

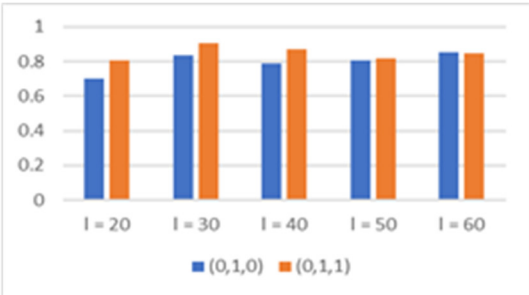
Table 3. Descriptive statistics for the residuals of the models

Model	Value	Standard error	t-statistics	p-value
(0,1,0)	0.6389	0.0852	7.4954	6.6081e-14
(1,0,0)	0.6215	0.0880	7.0616	1.6458e-12
(0,1,1)	0.6304	0.0830	7.5925	3.1383e-14

Fig. 5. Forecasts for 20 data points versus actual data



(a)



(b)

Fig. 6. Performance comparison between ARIMA(0,1,0) and ARIMA(0,1,1): a) MAE. b) RMSE

Generally, the error increases with the number of predictions as expected. Also, the model ARIMA(0,1,0) has a better performance compared to that of the model ARIMA(0,1,1) at all values of time horizons examined. The results can again be explained by the inclusion of the redundant MA(1) term, adding complexity during the parameter estimation process. Moreover, the most optimal MAE and RMSE are presented by model ARIMA(0,1,0) when l equals 20 which are 0.499 and 0.701, respectively, compared to 0.575 and 0.808 of the ARIMA(0,1,1) model.

4. CONCLUSION

The ARIMA model has been discussed for a short-term forecast of the wind speed in Quang Tri, Vietnam. The key process in this method is to determine the number of lags (p and q) and check if differencing might be required to make the data stationary. In this research, over-differencing or adding a moving average increased the complexity of the model leading to lower performance. The ARIMA(0,1,0) model results in the very close forecast to the time series with the training data set. The results suggest that, for the purpose of wind speed forecast, unnecessary parameters such as the moving average could be ignored to avoid possible convergence failure of the algorithm. In addition, to improve the forecast quality the training data should be updated right after the observation is available. It is noted that the seasonal patterns have yet to be taken into account, therefore, only when this factor is ignored, the ARIMA(0,1,0) model provides the best fit. Therefore, the seasonal character of the time series shall be studied in the future which is suggested by the Seasonal ARIMA model to yield a better fit to the observations and improved forecast consequently.

APPENDIX

Appendix 1. Backshift operator

B operator is defined to cause any observation that it multiplies with to shift backward by one unit time. In general, for an n integer:

$$B^n y_t = y_{t-n}$$

Appendix 2. Yule-Walker equation

Partial Autocorrelation Function (PACF) considers the correlation between two lags directly without accounting for the effects of the lags in between like the Autocorrelation Function (ACF). The PACF is usually calculated using Yule-Walker equation:

$$\rho_j = \Phi_{k1}\rho_{j-1} + ... + \Phi_{k(k-1)}\rho_{j-(k-1)} + \Phi_{kk}\rho_{j-k} \quad (j = 1, 2, \ldots, k)$$

where ρ_j is the j-th autocorrelation ($\rho_0 = 1$), Φ_{jk} is the j-th autoregressive parameters of an AR(k) model.

The equation can also be represented as:

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \Phi_{k1} \\ \Phi_{k1} \\ \vdots \\ \Phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix}$$

or succinctly:

$$P_k \Phi_k = \rho_k$$

REFERENCES

[1]. Government news, *Decision approving national power development plan 8*. 2014. Accessed 22 April 2024 <URL: https://en.baochinhphu.vn/decision-approving-national-power-development-plan-8-111230614195813455.htm>

[2]. Jung J., Broadwater R.P., "Current status and future advances for wind speed and power forecasting," *Renew. Sustain. Energy Rev.*, 31, 762-777, 2014.

[3]. Hafini S., Liu Xiaolei, Lin Z., Lotfian S., "A critical review of wind power forecasting methods - Past, present and future," *Energies*, 15, 1-24, 2020.

[4]. Olaofe Z. O., *Wind energy generation and forecasts: a case study of Darling and Vredenburg sites*. Msc Thesis Cape Town: University of Cape Town, 2013.

[5]. Box G. E. P., Jenkins G. M., Reinsel G. C., Ljung G. M., *Time series analysis: forecasting and control* (5th Edition). Wiley, USA, 2015.

[6]. Hyndman R. J., Athanasoupoulos G., *Forecasting: Principle and practice* (2nd edition. OTexts, Australia, 2018.

[7]. Cryer J. D., Chan K. S., *Time series analysis with applications in R* (2nd edition). Springer, Germany, 2008.

[8]. Hastie T., Tibshirani R., Friedman J., *The elements of statistical learning* (2nd edition). Springer, Germany, 2001.

THÔNG TIN TÁC GIẢ

Hoàng Trung Kiên, Lê Minh

Khoa Công nghệ và Kỹ thuật ứng dụng, Trường Đại học Khoa học và Công nghệ Hà Nội, Viện Hàn lâm Khoa học và Công nghệ Việt Nam