PAPER • OPEN ACCESS

The analysis of partial autocorrelation function in predicting maximum wind speed

To cite this article: G M Tinungki 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **235** 012097

The analysis of partial autocorrelation function in predicting maximum wind speed

G M Tinungki

Department of Mathematics, Faculty of Mathematics and Natural Science,
Hasanuddin University, Makassar 90245, Indonesia

E-mail: *ina_matematika@yahoo.co.id*

Abstract. The stationary and non-stationary testing of the data set can be done using a plot analysis of the Partial Autocorrelation Function of the data, by viewing the maximum number of the expected value of Partial Autocorrelation. The *Autocorrelation Function* (ACF) is a function that shows the correlation between the observation of the t-time and the observation at the previous time. The autocorrelation function shows the autocorrelation coefficient, which is the correlation measurement of the observations at different times. Data taken from Statistics Indonesia -known in Indonesia as BPS (Badan Pusat Statistik)- contains the information about the maximum wind speed by month at the Paotere station in 2008 - 2017 in Makassar City. By using the data, the maximum wind speed for the next 12 months will be estimated.ie from January 2017 to December 2018. The results obtained, forecasting is done with 12 leads period ahead with 95% confidence interval.

1. Introduction

Forecasting is a process of sequentially compiling information about past events to predict future events. Forecasting aims to get a prediction to minimize the forecasting error, which can be measured using *mean Absolute Percent Error* (MAPE) [1]. Forecasting is generally used to predict what is possible to happen. Based on the logic, in common, the steps forecasting methods are collecting data, selecting data, selecting a forecasting model, using selected models to forecast, evaluating the results [2].

Stationarity means there is no drastic change in data. The data fluctuation is around a constant mean value, independent of the time and variance of the fluctuation [3]. The time series data is stationary if the mean and variance are constant, there is no trend element in the data, and no seasonal elements [4]. A thing to consider in doing the forecasting is on the error; this consideration it must be included in the forecasting method [5]. To get results, that is close to the original data; then a forecaster should put the effort in making the error as small as possible, the data is used in predicting the maximum wind speed [6].

2. Literature Review

2.1. Autocorrelation Function (ACF).

Autocorrelation Function (ACF) is a function that shows the correlation between the observation of the-t time and the observation at previous times [7]. The autocorrelation function indicates the autocorrelation coefficient, which is the measurement of the correlation between observations at different times [8]. The autocorrelation function is defined as follows [9]:

$$\rho_k = \frac{\sum_{t=1}^{N-k}(x_t-\mu)(x_{t+k}-\mu)}{\sum_{t=1}^{N}(x_t-\mu)^2} \tag{1}$$

However, considering the sample data, the autocorrelation function sample is required, which is denoted as[10],

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t-\bar{x})(x_{t+k}-\bar{x})}{\sum_{t=1}^{n}(x_t-\bar{x})^2} \tag{2}$$

With $\bar{x} = \sum_{t=1}^{n} \frac{x_t}{n}$is the mean sample.

2.2. Partial Autocorrelation Function.

The Partial Autocorrelation Function is the correlation between Z_t and Z_{t+k} after the influence of the confounding variable $Z_{t-1}, Z_{t-2}, \ldots, Z_{t-k+1}$ isremoved. Partial autocorrelation coefficients are usually denoted byϕ_{kk}.

$$\phi_{kk} = Corr\,(Z_t, Z_{t-k} \mid Z_{t-1}, Z_{t-2}, \ldots, Z_{t-k+1}) \tag{3}$$

$\phi_{kk}$is the correlation coefficient between two random variablesZ_t and$Z_{t-k}$with the provision of $Z_{t-1}, Z_{t-2}, \ldots, Z_{t-k+1}$[12].

The common method used in calculating the partial autocorrelation coefficients is the Yule-Walker equation [13]

$$\begin{aligned}
\rho_1 &= \phi_{k1} + \phi_{k2}\rho_1 + \ldots\ldots\ldots + \phi_{kk}\rho_{k-1} \\
\rho_2 &= \phi_{k1}\rho_1 + \phi_{k2} + \ldots\ldots\ldots + \phi_{kk}\rho_{k-2} \\
. \quad & \quad . \quad\quad . \quad\quad\quad\quad . \\
. \quad & \quad . \quad\quad . \quad\quad\quad\quad . \\
. \quad & \quad . \quad\quad . \quad\quad\quad\quad . \\
\rho_k &= \phi_{k1}\rho_{k-1} + \phi_{k2}\rho_1 + \ldots\ldots\ldots + \phi_{kk}
\end{aligned} \tag{4}$$

Partial autocorrelation coefficients can be estimated by using partial autocorrelation coefficients of the sample by changing the value ρon theYule-Walker equation with r, and counting for k=1,2,… to get the valueϕ_{kk} using Cramer rules [14].

3. Methodology

3.1 Data.

Data taken from Statistics Indonesia -known in Indonesia as BPS (Badan Pusat Statistik) contains the information about the maximum wind speed by month at the Paotere station in 2008 – 2017, in Makassar City. By using the data, the maximum wind speed for the next 12 months will is estimated, i.e., from the month January 2018 to December 2018. The ARIMA model identification is done by viewing the existing patterns of ACF and PACF of sample data.

3.2. Stages of model identification:

1. Plot the Time series data and select the appropriate transformation
   From the time series data plot, the trend patterns, seasonality that may exist, an outlier, nonconstant variance, normality, and stationarity can be known. The usable transformation is Box-Cox's.

2. Calcúlate and test ACF and PACF
   Example:
   If ACF falls slowly and PACF is significantly different in lagone, do difference do Dickey-Fuller test. Differences are usually done at d = 0,1,2.
3. Calculate and test ACF and PACF
   Example:
   The ARIMA tentative model identification process (p, d, q) can be done by identifying ACF and PACF characteristics of an ARIMA model (Table 6.3). If the ACF and PACF characteristics of the stationary data are recognized, thenARIMA model (p, d, q) from the data can be determined. For example ACF plotis significantly different in the lag -1 and plot of PACF decreases exponentially, then data is identifiedfollowing MA model (1).
4. Test the Deterministic trend, if d > 0

4. Results and discussion

If the data is not stationary, modifications are required to produce stationary data. One of the common ways used is the *differencing method*. To determine whether the *series* is stationary or non-stationary, viewing the plot of the series or its difference form can be used. The process of differencing can be done for some period until the data become stationary by subtracting data with previous data. According to [3] a handy notation in the differencing method is the *backward shift* B operator, as follows:

$$BX_t = X_{t-1} \tag{5}$$

Notation B has the effect of shifting data 1 period backward. And then will change the data 2 periods back, as follows:

$$BBX_t = B^2X_t = X_{t-2} \tag{6}$$

If a time series is not stationary, then the data can be drawn closer to stationary by making the first differencing.

$$X_t^{'} = X_t - X_{t-1} \tag{7}$$

$$X_t^{'} = X_t - BX_t = (1-B)X_t \tag{8}$$

The first differencing is denoted by (1-B). The purpose of the differencing calculation is achieving the stationarity, and in general, if there is a difference of the-*d*ordo to reach stationarity, it is written as follows:

$$(1-B)^dX_t \tag{9}$$

Plot the data in mini-tab, to find out whether the data is stationary or not. It will be checked by viewing the plot, FAK graph and FAKP graph. To determine the mean value of maximum wind speed this following equation is used:

$$\overline{x_t} = \frac{\sum_{t=1}^{N} x_t}{N} \tag{10}$$

Based on the calculation using mini-tab 17 obtained the following results:

Time series plot in figure 1 shows that the plot is not constant. After viewing the data distribution plot, the FAK illustrates that the data is not stationary because many time lags go off from the significant limits, and the information drops close to zero slowly (exponential), this is reinforced by the calculation result in the of FAK values in mini-tab, where the FAK value is not close to zero significantly yet. FAKP diagram describes that the data is interrupted after lag 1. So, the information is not stationary mean.

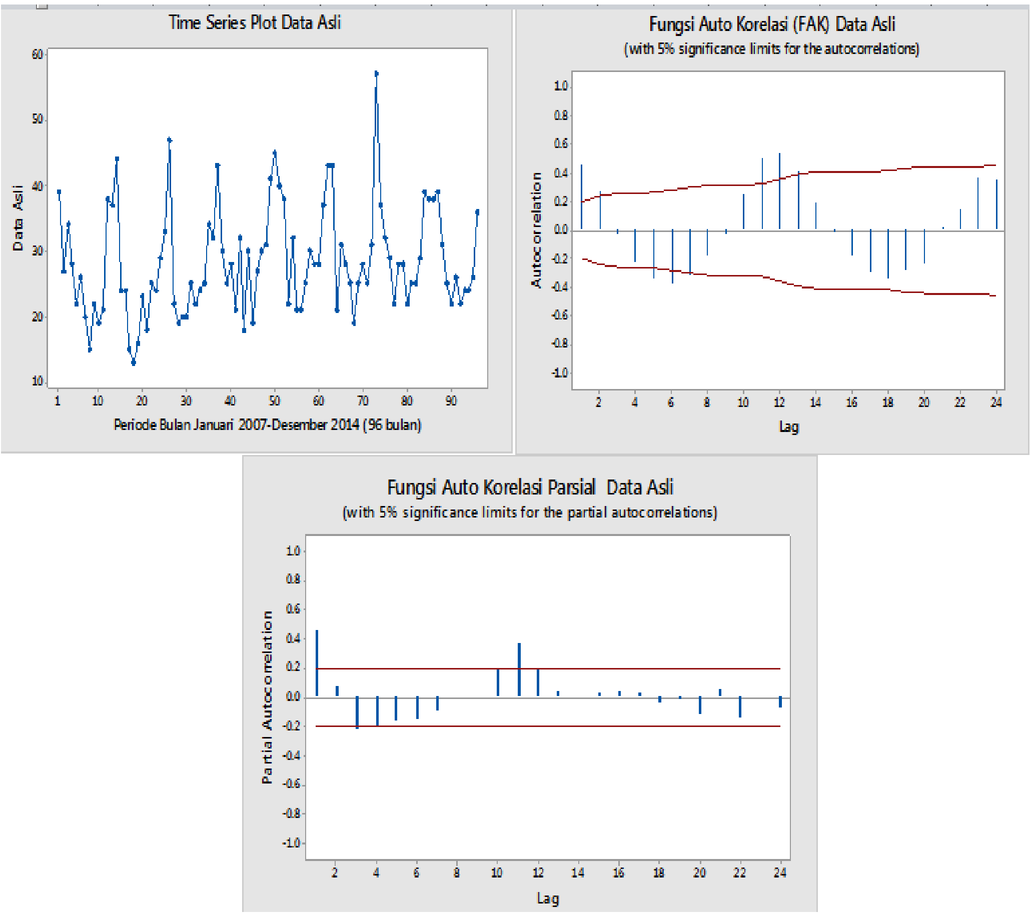


Figure 1. Plot time series, FAK and FAKP function.

The stationary and non-stationary testing can be done by analyzing the plot of FAK (Auto-Correlation Function) and FAKP (Partial Auto-Correlation Function) from data. The maximum number of estimated value of FAK and FAKP is 96/4 = 24. Using Minitab, the value of FAK and FAKP estimates up to lag 24 can be obtained. Based on the calculation by using mini-tab 17, the following result is obtained:

4.1 Autocorrelation Function: Real Data.

Partial Autocorrelation Function (PACF) used to measure the level of closeness between $x_t$andx_{t+k}, if the effect of the time lag$1,2,3,\dots,k-1$is removed. The partial autocorrelation formula or ϕ_{kk} is:

$$\phi_{kk} = \frac{r_k - \sum_{j=1}^{k-1} \phi_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j} r_j} \tag{11}$$

With $\phi_{kk}$is partial autocorrelation between$x_t$andx_{t+k}.

4.2 Data Stationary Testing.

There are two types of stationary: stationary invariance and stationary in the mean. If it is not stationary to the variant, the Box-Cox transformation is carried out, whereas if it is not stationary to the mean, the differencing is carried out.

a. Testing stationery invariance, a time series data, which is stationary from time to time, have a constant and unchanged data fluctuation. If the data is not stationary invariance, the transformation box-cox can be used on the mini-tab. The result is as follows:

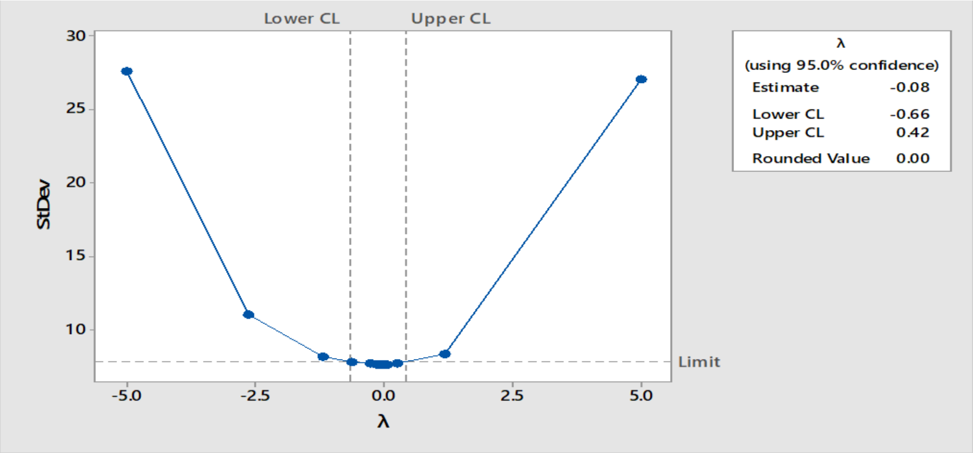


Figure 2. Box cox of real data.

Based on the results of box-cox testing shown in figure 2, it can be seen that the is not stationary to the variance yet since the *rounded Value* $\neq 1$. Thus, it needs to be transformed according to the Rounded Value stationer. Since$\lambda = 0$ then $\ln x_t$

b. The stationary test in the mean is the fluctuation of the data around a constant mean value, a way of doing stationary testing when the data is not stationary in mean is using the differencing method to produce stationary data. Differencing is carried out by reducing the value in a period with the amount in the previous period.

To solve the non-stationary data, the first differencing process on the data is done, as follows:

$$x'_t = x_{t-}x_{t-1} \tag{12}$$

for t=2,3, 4….96 the following result is obtained:

$$x'_2 = x_{2-}x_{2-1}$$
$$x'_2 = x_{2-}x_1$$
$$x'_3 = x_{3-}x_{3-1}$$
$$x'_3 = x_{3-}x_2$$
$$\vdots$$
$$x'_{96} = x_{96-}x_{95}$$

From the above data, the mean value can be calculated by using the following formula:

$$\overline{x_t} = \frac{\sum_{t=1}^{N} x_t}{N} \tag{13}$$

After finishing the first differencing process, then made plot and FAK and FAKP diagram on mini-tab from data result of the differencing process. The result as follows:

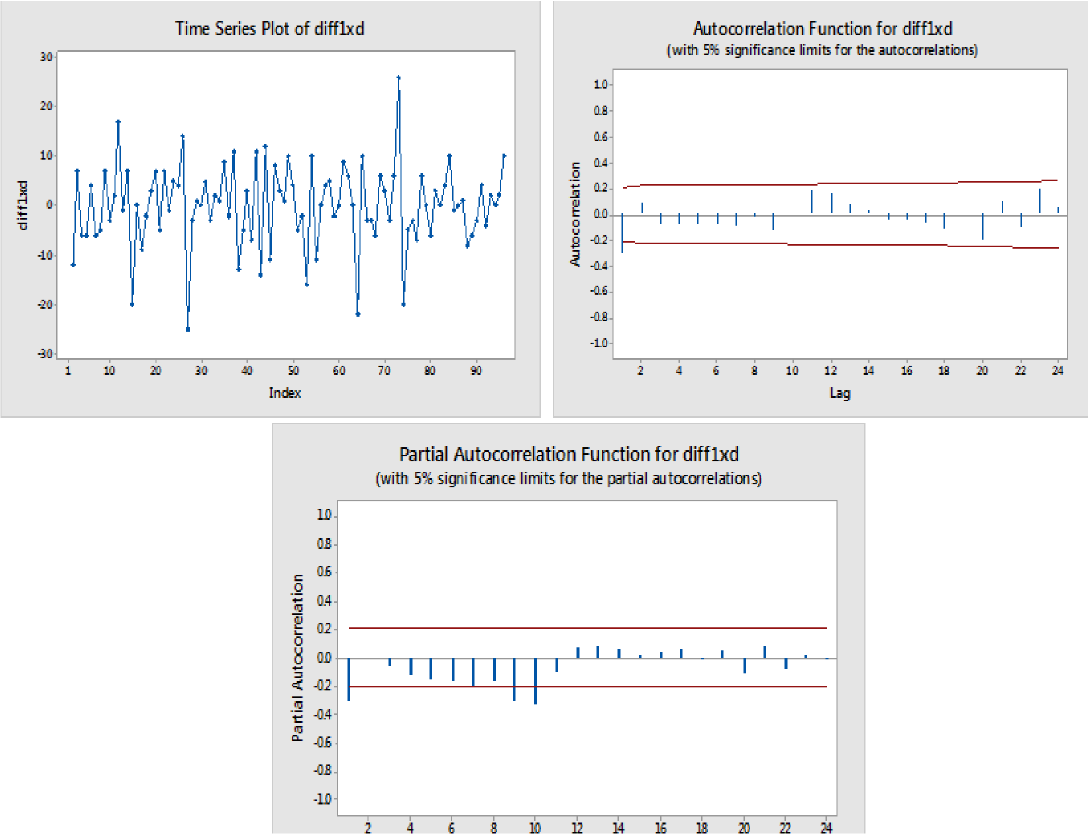


Figure 3. PlotFAK and FAKP data *differencing 1* non-seasonal (d=1).

Based on figure 3, the plot of the result of the differencing process of the data shows the stationary on the mean, and in the FAK and FAKP diagrams. The maximum wind speed data after the one-time differencing process, the FAK graph shows that the autocorrelation value falls slowly on a seasonal basis. This indicates the maximum wind speed data is not stationary yet in the seasonal mean. Therefore, to make stationary, one seasonal differencing 12 (D = 1) is carried out. The result as follows:

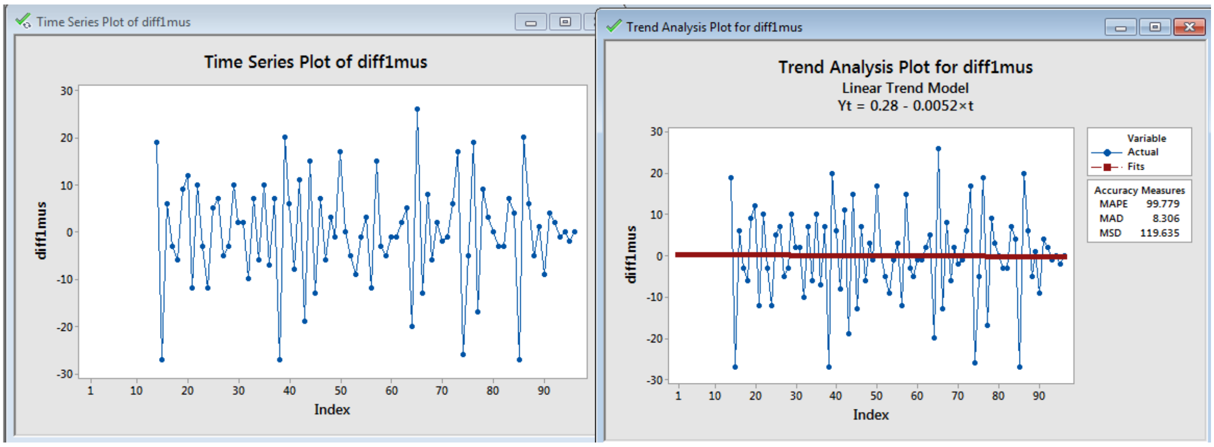


Figure 4. Plot time series & plot trend.

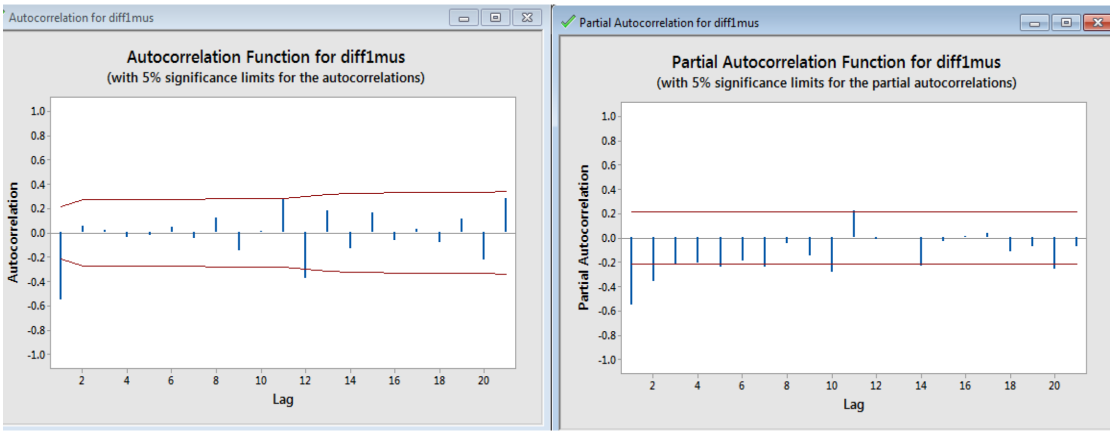

Autocorrelation for diff1mus
Autocorrelation Function for diff1mus
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
Lag
Partial Autocorrelation for diff1mus
Partial Autocorrelation Function for diff1mus
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
Lag

Figure 5. Diagram FAK and FAKP *differencing* 1 lag 12 seasonal

Based on figure 4 and figure 5, it is seen that the data is stationary. Next is to specify the model of the initial prediction of ARIMA seasonal model. The FAKP and FAKP chart patterns are cut off and each is cut on *lag* 1non seasonal and *lag* 1 seasonal (lag 12). So, the initial prediction of seasonal ARIMA model isARIMA (1,1,0) $(1,1,0)^{12}$. To obtain the most appropriate model, which is overfitting the temporary model, it can be done by changing the AR and MA ordo of the ARIMA model (1,1,0) $(1,1,0)^{12}$. thus, ARIMA (0,1,1) $(0,1,1)^{12}$ is obtained.

4.3.Model Identification.
Based on the initial estimated model, ARIMA $(1,1,0)(1,1,0)^{12}$ and ARIMA $(0,1,1)(0,1,1)^{12}$, with d=D=1 and s=12 are obtained.
Model classification
Box-Jenkins(ARIMA) model divided into 3 groups; autoregressive model (AR), moving average (MA), mixed ARIMA (autoregressive moving average) model that has characteristics of the first two models.

4.3.1.Diagnostic testing diagnostics. Testing can is divided into two parts; parameter significance test and fit model test (white noise assumption test and normal distribution).

4.3.2. Parameter Significance Test. In general, the ACF for the model AR (1) is

$$r_k = \begin{cases} 1 & ; k = 0 \\ \emptyset_1^k & ; k > 0 \end{cases} \tag{14}$$

The ACF for AR model 1 indicates that the autocorrelation value is smaller or close to zero along with the increasing lag (k), so it can be concluded that the ACF form of the AR model 1 is decreasing exponentially.

Significant autocorrelation testing can be done with the hypothesis:
H_0: $r_k = 0$(Non-significantautocorrelation coefficient)
H_1: $r_k \neq 0$(Significant autocorrelation coefficient)

The statistics, which is used is:

$$t = \frac{r_k}{SE(r_k)} \tag{15}$$

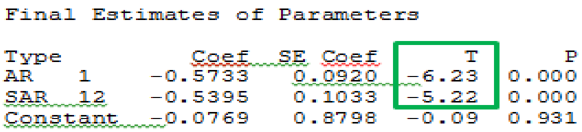
Where:

$$SE(r_k) = \sqrt{\frac{1+2\sum_{i=1}^{k-1} r_i^2}{n}} \tag{16}$$

Decision criteria: H_0 is decline if$|t_{hit}| > t_{\frac{\alpha}{2},n-1}$

Based on the delineation on mini-tab obtained value T_{hit} as follows:

- MODEL (1 1 0)(1 1 0)^{12}

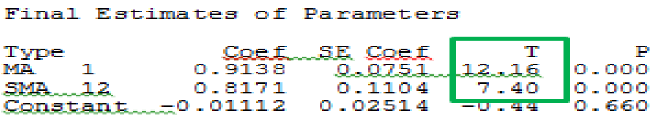


```
Final Estimates of Parameters

Type          Coef  SE Coef       T      P
AR    1    -0.5733   0.0920   -6.23  0.000
SAR  12    -0.5395   0.1033   -5.22  0.000
Constant   -0.0769   0.8798   -0.09  0.931
```

(a)

- MODEL (0 1 1)(0 1 1)^{12}



```
Final Estimates of Parameters

Type            Coef   SE Coef       T      P
MA    1       0.9138    0.0751   12.16  0.000
SMA  12       0.8171    0.1104    7.40  0.000
Constant    -0.01112   0.02514   -0.44  0.660
```

(b)

Figure 6. (a) and (b) Final Estimates of Parameters

Based on Figure 6 (a), H0 is accepted so that it is not significant, and in Figure 6 (b). H$_0$isrefusedsince$|t_{hit}| > t_{\frac{\alpha}{2},n-1}$so, the model is significant.

5. White Noise Process

The model is useful if the error value is random, meaning it does not have a specific pattern anymore. In other words, the obtained model is able to capture the existing data pattern perfectly. To see the error value randomness, the test autocorrelation coefficient value of the error is done by using one of the following statistics:

Q Box and Pierce Testing:

$$Q = n'\sum_{k=1}^{m} r_k^2 \tag{17}$$

Ljung-box testing:

$$Q = n'(n' + 2)\sum_{k=1}^{m} \frac{r_k^2}{(n'-k)} \tag{18}$$

Spreading in by Khi Square with degrees of freedom (db)=(k-p-q-P-Q)
description:

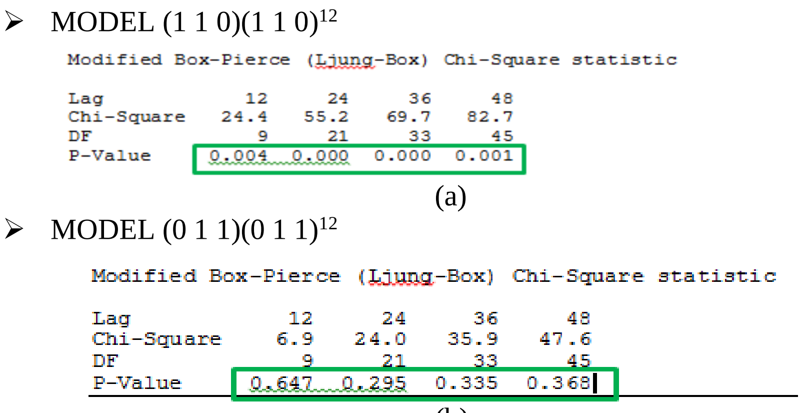
n’ = n-(d+SD)
d = non-seasonal factor differentiation ordo
D = seasonal factor differentiation ordo
S = number of periods per season
m = maximum time lag
rk = autocorrelation for the time lag 1, 2, 3, 4,..., k

Hypothesis:

H$_0$: Data is white noise
H$_1$: Data is not white noise

Testing criteria:

$Q \leq \chi^2$or p-value > α then the error value is random (data is white noise)
$Q > \chi^2$or p-value < α then the error value is not random (data is not white noise)
based on the calculation on minitab17, the results as follows:



- MODEL (1 1 0)(1 1 0)^{12}

```
Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag              12     24     36     48
Chi-Square     24.4   55.2   69.7   82.7
DF                9     21     33     45
P-Value       0.004  0.000  0.000  0.001
```

(a)

- MODEL (0 1 1)(0 1 1)^{12}

```
Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag              12     24     36     48
Chi-Square      6.9   24.0   35.9   47.6
DF                9     21     33     45
P-Value       0.647  0.295  0.335  0.368
```

(b)

Figure 7. (a) and (b) Modified Box-Pierce (Ljung-Box).

Based the figure7 (a) p-value <αthen the error value is not random so the model does not meet the white noise requirements. Based in the Image 7 (b) $Q \leq x^2$ or p-value > α then the error value is random. So, the data meets the white noise requirements.

6. Residual Normal

The residual normality test is performed to find out whether the residue meets the normality assumption or not. The normality assumption test used in the research is Kolmogorov Smirnov test, with the following hypothesis.

Hypothesis:

H_0: Residual has a normal distribution

H_1: Residuals has a non-normal distribution

Testing Statistics:

$$D = KS = maksimum|F_0(x) - S_n(x)| \qquad (19)$$

Description:

$F_0(x)$ = the cumulative distribution function that occurs under the normal distribution

$S_n(x)$ = the frequency distribution function being observed

Decision criteria:

H_0 is refused if*p-value*< α

Calculations on mini-tab can be seen in the figure below:

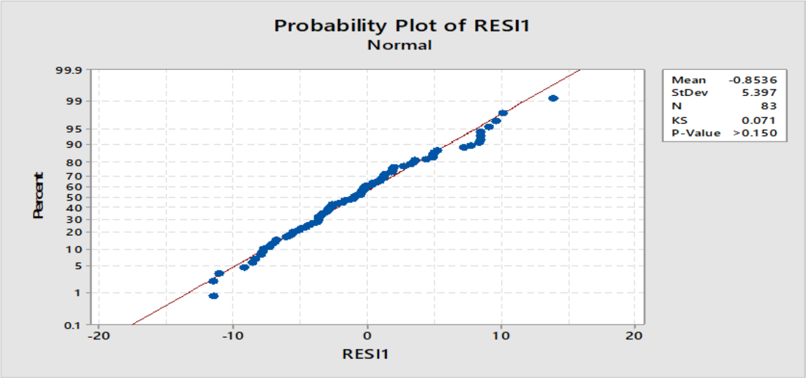


Figure 8. Plot residual.

Figure 8 shows that residuals are normally distributed since the plots of data follow the normal line. By calculating using Kolmogorov Smirnov test in mini-tab, it is obtained P-value => 0,150, then p-value value> 0,05 so that H0 is accepted and can be concluded that maximum wind speed data in Makassar for January 2007-December 2014 fulfill the normal distribution assumption.

a. Selection of the best Model

Determining the best model can be done by using the following standard error estimate:

$$S = \left[\frac{SSE}{n-n_p}\right]^{1/2} = \left[\frac{\sum_{t=1}^{n}(Z_t - \hat{Z}_t)^2}{n-n_p}\right]^{1/2} \quad (20)$$

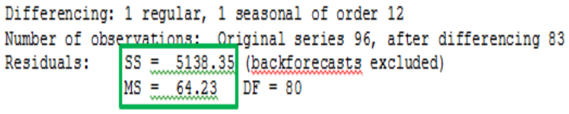
Description:

Zt = the actual value at time -t

$\hat{Z}_t$= estimated value at time-t

The best model is the model that has the least standard error estimate (S) value. Based on the calculation on mini-tab 17, the following results were obtained:

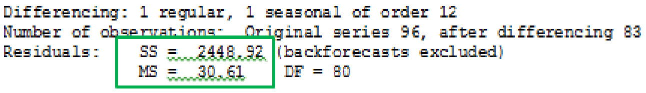
- MODEL (1 1 0)(1 1 0)^{12}



```
Differencing: 1 regular, 1 seasonal of order 12
Number of observations: Original series 96, after differencing 83
Residuals:    SS =  5138.35  (backforecasts excluded)
              MS =  64.23    DF = 80
```

(a)

- MODEL (0 1 1)(0 1 1)^{12}



```
Differencing: 1 regular, 1 seasonal of order 12
Number of observations:  Original series 96, after differencing 83
Residuals:    SS =  2448.92  (backforecasts excluded)
              MS =  30.61    DF = 80
```

(b)

Figure 9. (a) and (b) Residuals**.**

Based on figure 9 (b) it has the least estimate error standard compared with Figure 9 (a). Since all parameters in the obtained model are significant, the residue has fulfilled the white noise process, normally distributed, and has the least estimation error standard, then **ARIMA Model: (0 1 1)(0 1 1)^{12}** is appropriate.

b. Forecasting Stage

After selecting the best model, which is ARIMA (0 1 1)(0 1 1)^{12}with this following equation:

$$(1\text{-}B)(1\text{-}\emptyset_1 B)Y_t = \mu' + (1\text{-}\theta_1 B)e_t$$

$$(1\ \text{-}\emptyset_1 B - B + \emptyset_1 B^2)Y_t = \mu'+e_t\text{-}\ \theta_1 B e_t$$

$$Y_t - Y_t\emptyset_1 B - Y_t B + Y_t\ \ \emptyset_1 B^2 = \ \ \mu'+e_t\text{-}\ \theta_1 B e_t$$

$$Y_t = \mu'+(1+\emptyset_1\ \ )Y_{t-1}\text{-}\emptyset_1\ \ Y_{t-2}\ \text{-}\ e_t\text{-}\ \theta_1\ \ e_{t-1} + e_t$$

$$Y_t\text{= -0.01112+ 0.9138}Y_{t-1}\ \text{+0.8171}Y_{t-2} + e_t$$

Then, the last stage of the ARIMA process is using the best model for forecasting. Forecasting can be done for several periods, in this study, the prediction is done with 12 leads period ahead with a 95% confidence interval.

7. Calculating Forecasting Errors with Using MAPE (Mean Square Percent Error)

Calculations using MAD and MSE will yield large values even up to thousands, to avoid this, MAPE or Mean Absolute Percent Error can be used, where errors are calculated based on the percent absolute error. MAPE calculates the deviation between the actual data and the forecast value, then calculate the mean percentage. The MAPE formula as follows:

$$MAPE = \frac{\sum(Kesalahan\ Persen\ Absolut)}{n} \qquad (21)$$

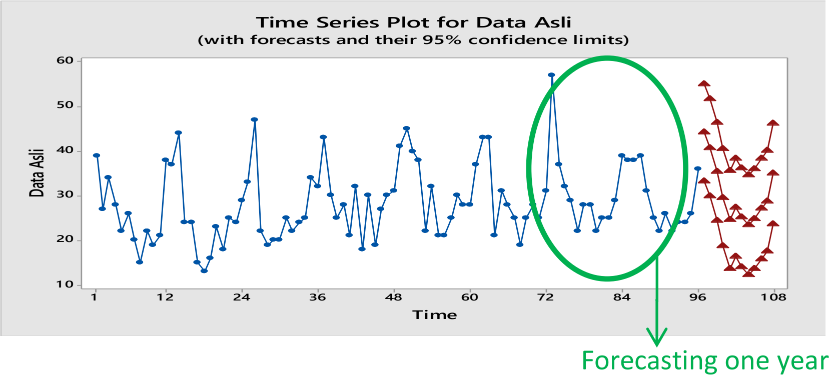


Figure 10. Plot *forecast.*

Figure 10 shows the plot of forecasting for 1 year ahead.

8. Conclusion

Based on the result and discussion, it can be concluded that:
1.The effect of maximum wind speed forecasting for the next 1 year in 2017 and 2018 as follows:

Table 1. Maximum Wind Speed Forecasting at Makassar's Maritime Station Paotere (Knots) 2017

MONTH	2017
01. January	44,1366
02. February	40,7731
03. March	35,3959
04. April	29,6019
05. May	24,6398
06. June	27,3244
07. July	25,1354
08. August	23,4268
09. September	24,8646
10. October	27,0529
11. November	28,7767
12. December	34,939

Seen from the table 1above, the highest wind speed was in February 2018 and the lowest wind speed was in August 2018.

2. Function PartialAutocorrelation can be used to check the stationary and non-stationary data sets which can also indicate the correlation between the observed-time and the observation at previous times, to predict the future.

References

[1] Bernhardt C 2007*Modeling of Electricity Prices by Linear Time Series Models, and Value-at-Risk Estimation Using Methods from Extreme Value Theory* (Munchen: Technische Universität München).

[2] Dettling M 2008 *Applied Time Series Analysis*(Zürich: Hochschule für Angewandte Wissenschaften).

[3] Makridakis S, Wheelwright S C and Mcgee V E1995 *Metode dan Aplikasi Peramalan* (Jakarta: Erlangga).

[4] Gottman J M 2009 *Design and Analysis of Time-Series Experiments* (USA: Colorado Associated University Press).

[5] Wilks D S 2006 *Statistical Methods in the Atmospheric Sciences 2nd Edition*(*International Geophysics Series*vol 91), ed R Dmowska *et al* (Oxford: Elsevier) .

[6] Islam M A 2014 A study on the performance of symmetric and asymmetric garch models in estimating stock returns volatility *Int. J. Empir. Financ.***2** 182-192.

[7] Teusch A 2006 Introduction to the Spectral and Time Series Analysis with Examples from Geodesy (München: Verlag der Bayerischen Akademie der Wissenschaften).

[8] Madsen H 2008*Time Series Analysis* (Boca Raton: Chapman and Hall/CRC).

[9] Dubrova T A and Arhipova M Y 2004 *Statistical Methods for Forecasting The Economy* (Moscow: Mir).

[10] Toutenburg H and Heumann C 2008 *Descriptive Statistics: An Introduction to Methods and Applications with R and SPSS*(Berlin: Springer-Verlag Berlin Heidelberg).

[11] Vaidogas E R and Juocevicius V 2012 A critical estimation of data on extreme winds in Lithuania *J. Environ. Eng. Landsc. Manag.***19** 178-188.

[12] Bucur R D and Harja M 2012Homogeneous areas delimitation by considering the energy demand for plants growing in covered spaces *Environ. Eng. Manag. J.* **11**253-257.

[13] Nikitin A Y and Sosunova I A 2003 *Analysis and Prediction of Time Series Ecological Observations and Experiments* (Irkutsk: Irkutsk State University) p 81.