Menemui Matematik (Discovering Mathematics)
Vol. 43, No. 1: 21 - 26 (2021)

Forecasting and Evaluation of Time Series with Multiple Seasonal Component

Fatin Zafirah Zamri[1], Nur Haizum Abd Rahman[1*] and Hani Syahida Zulkafli[1]
[1]*Department of Mathematics and Statistics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor*
[1*]nurhaizum_ar@upm.edu.my

ABSTRACT

Seasonality is one of the components in time series analysis and this seasonal component may occur more than one time. Thus, modelling the seasonality by using one seasonal component is not enough and could produce less forecast accuracy. Autoregressive Integrated Moving Average (ARIMA) models is the fundamental method in developing the seasonal ARIMA for one seasonality or more than one seasonality. Therefore, to validate the method performance, the hourly air quality data with double seasonality were carried out as the case study. The model identification step to determine the order of ARIMA model was done by using MINITAB program and the model estimation step by using SAS program and Excel. The results showed that the double seasonal ARIMA able to model and forecast the air quality data with high frequency.

Keywords: Box-Jenkins, Forecasting, Model Identification, Seasonal

INTRODUCTION

Time series is a set of observations that have been recorded at a specific time. The main objective of time series analysis is to develop a mathematical model that can forecast future observations based on previous data. Time series data may involve four different components which are trend, seasonal, cyclical, and irregular. Trend is an upward and downward movement that shows in the time series data over a period of time. Seasonal is refer to the same pattern that repeats periodically within a calendar year meanwhile cyclical also refers to recurring up and down movements around trend levels however the period of cycle is greater than a year and not regular as seasonal variations. Irregular is refer to the erratic movements in a time series caused by unpredictable events usually define as error.

The Box-Jenkins or ARIMA is classified as a linear model that is capable in presenting different type of series components and able to model both stationary and non-stationary series. Therefore, Box-Jenkins method widely used in many fields especially in statistics, management science, marketing and business operation (Azka et al., 2020; Benvenuto et al., 2020; Khanarsa and Sinapiromsaran, 2017; Urrutia et al., 2017). Box-Jenkins methods is crucial in forecasting which it inclusive Autoregressive (AR) model, the Integrated (I) model and the Moving Average (MA) model (Hanke and Wichern, 2005). Autoregressive (AR) model provides forecast as linear function of finite number of past values, while Moving Average (MA) model forecasts based on a linear combination of a finite number of past errors. Autoregressive Moving Average (ARMA) model is a mixed between Autoregressive (AR) and Moving Average (MA) model.

ARMA, AR and MA models are suitable for stationary data. On the other hand, Autoregressive Integrated Moving Average (ARIMA) model is used when the data is non-stationary. Since there are seasonality that repeatedly occur in many types of data, researcher tend to develop seasonal ARIMA (SARIMA) model and mainly used for non-stationary data with seasonal pattern (Rahman et al., 2019; Azka et al., 2020). However, time series could contain multiple seasonal cycles of different lengths (Hassan, et al., 2012). As an example, in a time series data, the seasonality may occur in yearly, weekly, and daily variations. Due to the presence of double seasonal pattern in the data such as hourly and weekly seasonality, the double seasonal

ARIMA model is more suitable in modelling such data. This study will present the step in modelling double SARIMA and study the accuracy of the model building.

MATERIALS AND METHODS

The stationary test is used to determine either the series is stationary or non-stationary. For this study, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is used. The purpose of KPSS test is to test the null hypothesis which is stationary based on *p*-value where,

H_0: A series is stationary.
H_a: A series is nonstationary

According to Shin and Schmidt (1992), the components of KPSS are representation of y_t as the sum of a linear deterministic trend, a random walk, and a stationary error:

$$y_t = \psi t + r_t + \varepsilon_t \tag{2}$$

where r_t is a random walk:

$$r_t = r_{t-1} + u_t, u_t \sim WN(0, \sigma_u^2) \tag{3}$$

The error ε_t is stationary that makes stationarity hypothesis is $\sigma_u^2 = 0$ and y_t is trend stationary. KPSS test statistics can be expressed in the form of:

$$KPSS = \left(T^{-2} \sum_{t=1}^{T} \hat{S}_t^2 \right) / \hat{\lambda}^2 \tag{4}$$

Double seasonal ARIMA (SARIMA) model was developed due to the occurrence of double seasonal pattern in the data set. This model of the general multiplicative double seasonal ARIMA (SARIMA) can be written as:

$$\begin{aligned} \phi_p(B)\Phi_{P1}(B^{S1})\Pi_{P2}(B^{S2})(1-B)^d(1-B^{S1})^{D1}(1-B^{S2})^{D2}Y_t \\ = \theta_q(B)\Theta_{Q1}(B^{S1})\Psi_{Q2}(B^{S2})\varepsilon_t \end{aligned} \tag{1}$$

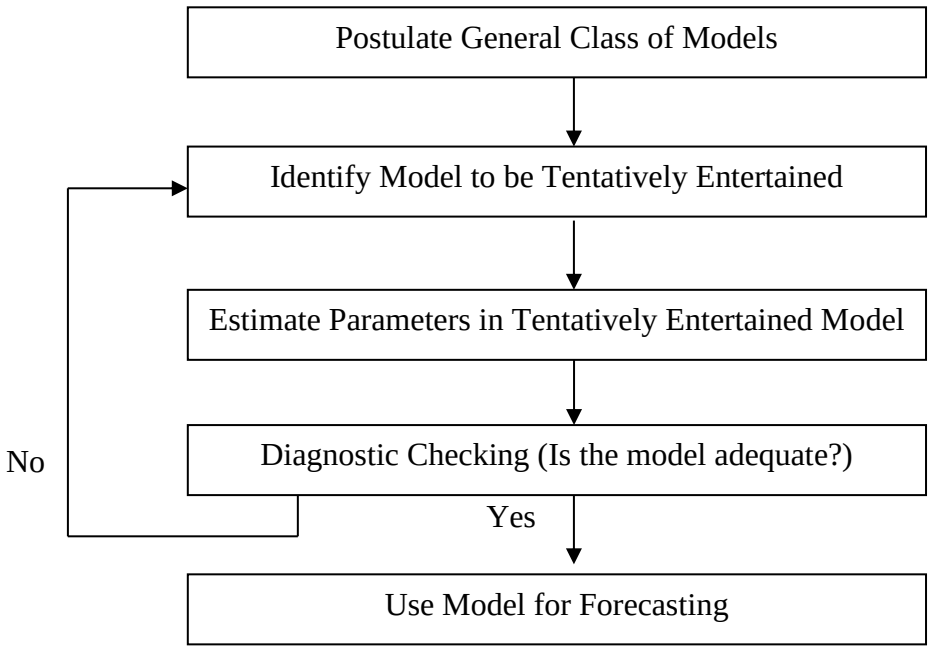
where

$$\begin{aligned} \phi_p(B) &= 1 - \phi_1 B^1 - \phi_2 B^2 - \cdots - \phi_p B^p \\ \Phi_{P1}(B^{S1}) &= 1 - \Phi_1 B^{S1} - \Phi_2 B^{2S1} - \cdots - \Phi_{P1} B^{P1S1} \\ \Pi_{P2}(B^{S2}) &= 1 - \Pi_1 B^{S2} - \Pi_2 B^{2S2} - \cdots - \Pi_{P2} B^{P2S2} \\ \theta_q(B) &= 1 - \theta_1 B^1 - \theta_2 B^2 - \cdots - \theta_q B^q \\ \Theta_{Q1}(B^{S1}) &= 1 - \Theta_1 B^{S1} - \Theta_2 B^{2S1} - \cdots - \Theta_{Q1} B^{Q1S1} \\ \Psi_{Q2}(B^{S2}) &= 1 - \Psi_1 B^{S2} - \Psi_2 B^{2S2} - \cdots - \Psi_{Q2} B^{Q2S1} \end{aligned}$$

where B denotes the backward shift operator; d, $D1$ and $D2$ denote the non-seasonal, first seasonal and second seasonal order of differences, respectively. The model can be abbreviated as SARIMA $(p,d,q)(P1,D1,Q1)^{S1}(P2,D2,Q2)^{S2}$.

Three main steps that must be considered in building the ARIMA model for forecasting include; (a) tentative identification, (b) parameter estimation, and (c) diagnostic checking (Hyndman and Athanasopoulos, 2018). The tentative identification step is used to identify an

appropriate Box-Jenkins model. The identification is based on autocorrelation function (ACF) and partial autocorrelation function (PACF). When the tentative model is specified, the historical data are used to estimate the parameters of the tentatively identified model. In this study, least square method is used to estimate the parameters. Finally, the adequacy of the model is check in diagnostic checking step. These three steps could be repeated for several times until a satisfactory of model is finally selected. This process can be simplified as Figure 1 below:



Source: Hanke, J. E., & Wichern, D. W. (2005)

Figure 1: Flow Diagram for the Box-Jenkins Model-Building Strategy

Finally, the accepted model will be use in forecasting and the performance will be validate by using mean absolute percentage error (MAPE). MAPE is often used in practice because of its very intuitive interpretation in terms of relative error (de Myttenaere et al., 2016). Then, the Lewis' judgement scale (Table 1) is referred to determine the forecast accuracy from the computed MAPE values (Lewis, 1982).

Table 1: Lewis' judgement scale

MAPE	Accuracy
≤ 10%	High
10% to 20%	Good
21% to 50%	Reasonable
≥ 51%	Inaccurate

RESULTS AND DISCUSSION

The following were the results in building an ARIMA model based on the procedure from Box-Jenkins model. Three years (2014-2016) hourly air quality data were used as case study. The data were divided into two data sets, namely: (1) a training data set from 1st January 2014 until 30th November 2016 (25563 observations) to identify the model, and (2) a test data set in December 2016 with a total of 744 observations to check the model performance. The data plot was shown in Figure 1.

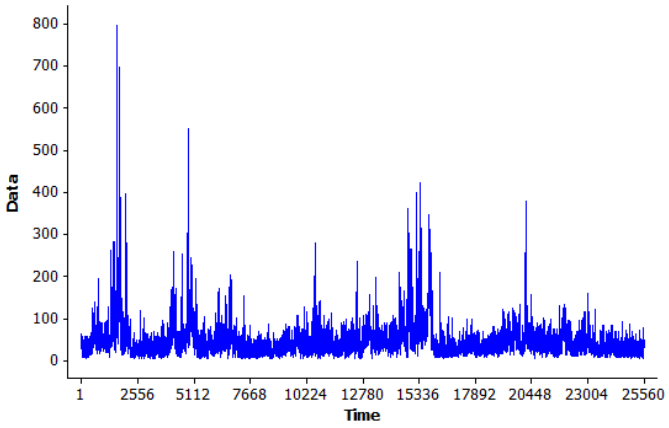


Figure 1: Time series plot

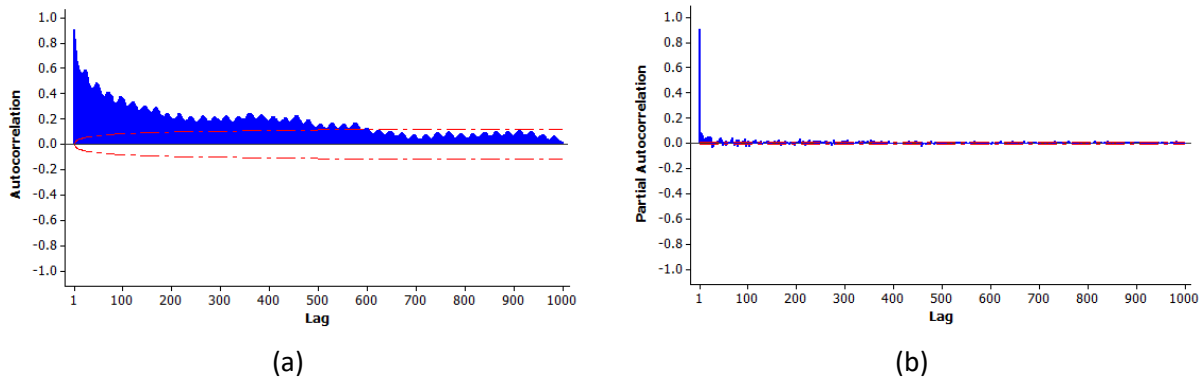
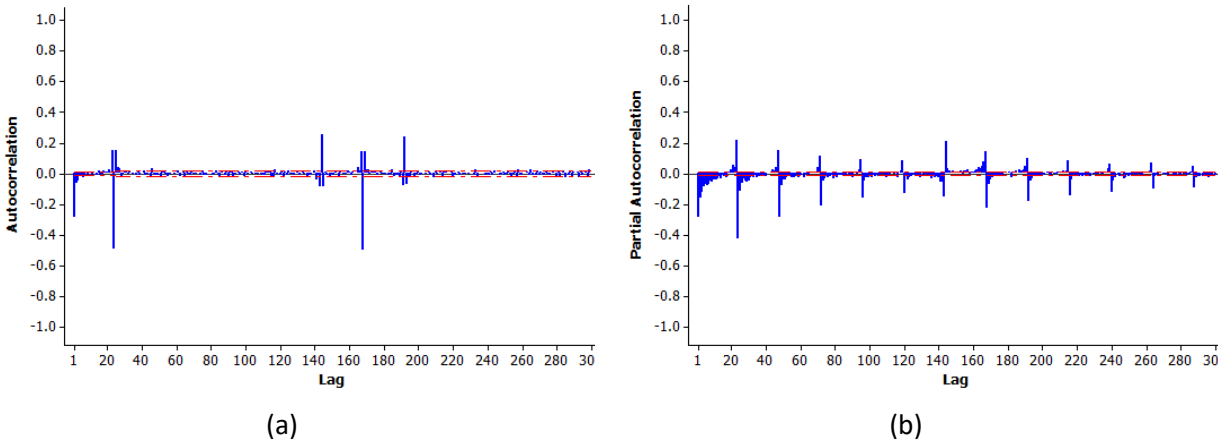


Figure 2: (a) Autocorrelation Function; (b) Partial Autocorrelation Function from original data

The plot based on Figure 1 shows the data was nonstationary since exist seasonality and increasing variations. Meanwhile, the correlation dies down slowly in ACF indicated nonstationary characteristic in Figure 2. The argument also proven by using stationary test, KPSS test. The *p*-value was 0.01, smaller than α=0.05. Transformation and three times differencing had been performed which are non-seasonal differencing (d = 1), hourly seasonal differencing (D1 = 1; S1 = 24) and weekly seasonal differencing (D2 = 1; S2 = 168) to fulfil stationary condition. By using KPSS test to the new data, the *p*-value become 0.10 which was greater than α=0.05. Therefore, the new data was stationary.

The stationary ACF and PACF were shown in Figure 3. Four double seasonal models had been identified from Figure 3. SARIMA $(0,1,1)(0,1,1)^{24}(0,1,1)^{168}$, SARIMA $(0,1,3)(0,1,1)^{24}(0,1,1)^{168}$, SARIMA $(1,1,0)(0,1,1)^{24}(0,1,1)^{168}$ and SARIMA $(1,1,1)(0,1,1)^{24}(0,1,1)^{168}$ denoted as Model 1, Model 2, Model 3, and Model 4, respectively. The models then were validated using in-sample data and four forecast horizons (out-sample) for one week, two weeks, three weeks and four weeks ahead. The results were shown in Table 2.



(a) (b)

Figure 3: (a) Autocorrelation Function; (b) Partial Autocorrelation Function from stationary data

Table 2: MAPE for in-sample and out-sample forecast

Forecast	Model 1	Model 2	Model 3	Model 4
In-Sample	2.6901	2.6820	0.6189	**0.5077**
Out-Sample				
One week	5.1084	5.1068	1.0578	**0.9448**
Two weeks	4.2608	4.2634	1.001	**0.9509**
Three weeks	4.2595	4.2559	1.1574	**1.0108**
Four weeks	4.1797	4.1805	1.4430	**1.3077**

From Table 2, all the models gave high accuracy since the forecast errors were below 10%. However, Model 4 provide the best forecasting model compared to other three models. As shown, the errors were 0.5077%, 0.9448%, 0.9509%, 1.0107% and 1.3077% for in-sample, one week, two weeks, three weeks, and four weeks ahead, respectively. Thus, Model 4 can be written as:

$$(1-\phi B)(1-B)(1-B^{24})(1-B^{168})Y_t = (1-\theta B)(1-\Theta_1 B^{24})(1-\Psi_1 B^{S2})\varepsilon_t \qquad (2)$$

CONCLUSION

This paper has discussed double seasonal ARIMA in forecasting data with two seasonality, hourly and weekly. The forecast accuracy for in-sample and out-sample real data were tested using mean absolute percentage error (MAPE). All the identified model produced higher accuracy since below 10% of error with Model 4, SARIMA $(1,1,1)(0,1,1)^{24}(0,1,1)^{168}$ gave the best forecast result. This study gives valuable contribution into forecasting the data with multiple seasonal components. In addition, this model can be set as the benchmark method for modelling and forecasting time series data with great accuracies. Future work can include other model identification, subset and additive into Box-Jenkins method.

ACKNOWLEDGEMENTS

This project is supported by the Universiti Putra Malaysia under Putra Grant Vot No: 9587700

REFERENCES

Azka, M., Wiradinata, S. A., Faisal, M., & Suhartono. (2020). Double Seasonal ARIMA for Forecasting Electricity Demand of Kuaro Main Gate in East Kalimantan. *IOP Conference Series: Materials Science and Engineering*.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief*.

de Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean Absolute Percentage Error for regression models. *Neurocomputing*, **192**, 38–48.

Hanke, J. E., & Wichern, D. W. (2005). *Business Forecasting* (8th ed.). New Jersey: Pearson/Prentice Hall.

Hassan, S. N., Ahmad, M. H., Suhartono, & Mohamed, N. (2012). A comparison of the forecast performance of double seasonal ARIMA and double seasonal ARFIMA models of electricity load demand. *Applied Mathematical Sciences*, **6(133–136)**, 6705–6712.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting : principles and practice. In *OTexts* (2nd ed.). Retrieved from OTexts.com/fpp2.

Khanarsa, P., & Sinapiromsaran, K. (2017). Multiple ARIMA subsequences aggregate time series model to forecast cash in ATM. *2017 9th International Conference on Knowledge and Smart Technology: Crunching Information of Everything, KST 2017*, 83–88.

Khashei, M., Bijari, M., & Hejazi, S. R. (2012). Combining seasonal ARIMA models with computational intelligence techniques for time series forecasting. *Soft Computing*, **16(6)**.

Lewis, C. D. (1982). *Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting*. In *Butterworth Scientific*.

Rahman, N. H. A., Lee, M. H., Suhartono, & Latif, M. T. (2019). Hybrid seasonal ARIMA and artificial neural network in forecasting southeast Asia City Air Pollutant Index. *ASM Science Journal*, **12**(Special Issue 1).

Shin, Y., & Schmidt, P. (1992). The KPSS stationarity test as a unit root test. *Economics Letters*, **38(4)**, 387–392.

Urrutia, J., Lean, D. J., & Mingo, F. L. (2017). Forecasting the Quarterly Production of Rice and Corn in the Philippines: A Time Series Analysis. *Journal of Physics: Conference Series*, **755**, 011001.