FORECASTING DATA USING BOX-JENKINS PROCEDURE: A CASE STUDY FOR UNEMPLOYED PEOPLE IN MALAYSIA

Fadila Amira Razali [1*], Nur Fatihah Haron [2]

[1]*College of Computing, Informatics and Media*
Universiti Teknologi MARA UiTM Pahang, 26400 Bandar Tun Abdul Razak Jengka, Pahang Malaysia

[2]*College of Computing, Informatics and Media*
Universiti Teknologi MARA UiTM Pahang, Raub Campus, 27600 Raub, Pahang Malaysia

[*]*Corresponding author: fadila962@uitm.edu.my*

Abstract

In forecasting and time series analysis, the Box-Jenkins approach has been widely employed as it is among the most effective methods to forecast business and non-business data. This paper focuses on modelling and applying the Box-Jenkins model to forecast how many unemployed people there are in Malaysia. Hence, the time series data of unemployed people in Malaysia between January 2010 and November 2022 were used for this study's purpose. There are three main stages in constructing the best model: model estimation and validation, identification, as well as model application. In the first stage, the actual data on unemployed persons in Malaysia was found to be non-stationary, with a rising trend pattern. As a result, the data series was subjected to the Box-Jenkins model with first-order differencing. Hence, three models were exhibited, which are Autoregressive Integrated Moving Average (ARIMA) (1, 1, 0), ARIMA (2, 1, 0) as well as ARIMA (3, 1, 0) by referring to both time plot and Autocorrelation Function (ACF) plots. The most appropriate model was determined in the second stage, which is ARIMA (3, 1, 0) by comparing the Bayesian Information Criterion (BIC), Root Mean Square Error (RMSE) as well as Mean Absolute Percentage Error (MAPE). Finally, the best model was applied to forecast unemployment, and an increasing trend in the number of unemployed was estimated to occur for the following year.

Keywords: ARIMA models, Box-Jenkins, unemployed

Introduction

Unemployment is one of the most concerning issues worldwide. Besides being one factor contributing to the current market's economic problem, it also brings considerable social consequences, including in Malaysia. Although the rate of unemployed persons in Malaysia continued to decline, as reported by the Department of Statistics Malaysia (DOSM, 2022), there are still a few states in Malaysia with a high rate of unemployment recorded during the third quarter of 2022, which include Pahang, Selangor and W.P. Putrajaya. According to research by Atif and Ishak (2015), economic globalisation has a positive and significant long-term influence on minimizing unemployment in Malaysia. Hence, policymakers in Malaysia should encourage economic globalization to sustain the country's low unemployment rate.

Many researchers have studied the trend of the unemployment rate worldwide. Then, several forecasting methods were used to model the unemployment rate. The most common method which had been applied was the Box-Jenkins model. Dobre and Alexandru (2008)

also modelled the evolution of the monthly unemployment rate in Romania during the period of 1998 – 2007 using Box-Jenkins, and the study found that to forecast the unemployment rate between January as well as February 2008, the most suitable model was Autoregressive Integrated Moving Average (ARIMA) (2, 1, 2). Besides, Stanley and Edwin (2016) have also applied the Box-Jenkins procedure to forecast Malaria cases in Zambia. The results indicated that this case would continue to occur shortly if no proper intervention measures were implemented and initiated on time. Research conducted by Nikolaos and Paraskevi (2018) proved that using one or a combination of the Box-Jenkins model could provide the best forecast for the monthly unemployment rate in US between January and July 2017. In addition, the ARIMA (2, 1, 2) has been found to be the most optimal model, showing an increment in Malaysia's unemployment rate (Suraya *et al*., 2018). Furthermore, according to the study conducted by Norliana *et al*. (2021), the Box-Jenkins model, which was ARIMA (2, 1, 3), revealed that the smoothing methods had generated the smallest value of all error measures. This includes the double exponential smoothing as well as Holt's model that has been applied to forecast the unemployment rate in Malaysia.

This research aims to identify the most fitted Box-Jenkins model to forecast unemployed people in Malaysia for the following years by applying an unemployed person's data in Malaysia from January 2010 until November 2022 to develop the models.

Materials and Methods

The monthly unemployment data in Malaysia from January 2010 and November 2022 were used for the study, and this data has been acquired from the Department of Statistics Malaysia (DOSM) website. Additionally, the analysis for the model building has been conducted using two software: Microsoft Excel and IBM SPSS Version 23.

The Box-Jenkins approach is generally synonymous with the ARIMA modelling. ARIMA combines the Autoregressive/Integrated/Moving Average model. Box and Jenkins introduced this method in the year 1976. Mixed ARIMA is the common term used. The formula for the model is ARIMA (p, d, q), in which 'p' refers to the order of the lagged dependant value in the Autoregressive (AR), 'q' denotes the order of the periods that lag in the Moving Average (MA), while 'd' denotes the number of times a variable must differ to ensure stationarity (Lazim, 2016).

There are three main stages in ARIMA Model development. Model identification, estimation, and evaluation are all included, as well as model application. **Figure 1** below illustrates the implementation of the three main stages.

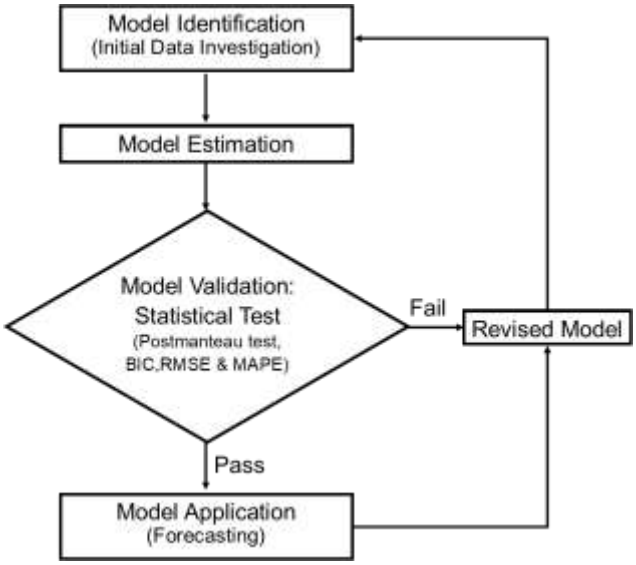

Figure 1 Stages in ARIMA model building

Model Identification

At this stage, a preliminary data analysis is done to see if any four time series components—trend, seasonal, irregularities, as well as cyclical are present based on the fluctuation of the time plots. Also, trend components are present if there is an increasing and decreasing movement, while seasonality exists when the time series shows regular fluctuation within a specific period (Douglas *et al*., 2011). Meanwhile, any pattern showing an up-and-down movement around a given unspecified time is identified as cyclical. Additionally, if the time chart has outliers or random shock events, an irregularity component is said to be present (Douglas *et al*., 2011).

The ACF and PACF has been displayed as well as observed in order to verify the stationarity of the data series to identify the model. The data series had to be stationary for the Box-Jenkins method to work, which was one of its assumptions. A data series is said to be stationary if there is no upward and downward movement over time, indicating the trend component. The data series can achieve stationary by taking the differencing or specifically removing the trend (Lazim, 2016).

Next, model identification can be applied by observing the ACF and PACF plot for the potential order of p, d, and q parameters. However, identifying the p and q parameters is somewhat tricky. Therefore, it is worth considering several possible models and applying several statistical test procedures to determine the appropriate fitted model (Lazim, 2016).

Model Estimation, Validation and Diagnostic Checking

The ideal model for use in forecasting will be identified in this part. Furthermore, the data series had previously been divided into both evaluation and estimation parts. This process was used to make sure the estimation model was accurate when used with the actual population data. The part was divided by taking 70% of data for estimation part and 30% for validation part.The estimation was taken from January 2010 until December 2018; meanwhile, the evaluation was taken from January 2019 until November 2022. The diagnostic checking of the random error and Bayesian Information Criterion (BIC) were checked on the estimation part for all models. The smallest value of BIC is said to have the best estimation model (Lazim, 2016). In the Box-Jenkins model, it is assumed that the error terms should be white noise or not correlated to each other (Lazim, 2016). Therefore, the

portmanteau test of Box-Pierce Q statistics was conducted to check the assumptions with the hypothesis given are:

H_0 : The errors are white noise (random)

H_1 : The errors are not white noise (not random)

Next, on the evaluation part, the forecast value was obtained for each model to evaluate the model and identify the closes fitted values with the actual values by measuring the Root Mean Squared Error (RMSE) as well as Mean Absolute Percentage Error (MAPE). RMSE is considered the best measurement in comparing forecasting performance. In contrast, MAPE is the most widely used among practitioners and has a unit-free measure compared to other error measures. The RMSE as well as MAPE are defined as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t}^{n} e_t^{\,2}}{n}} \tag{1}$$

$$\mathrm{MAPE} = \sum_{t=1}^{n} \frac{\left|\left(e_t / y_t\right) \times 100\right|}{n} \tag{2}$$

where

$e_t = y_t - \hat{y}_t$

y_t is the actual observed value in time *t*

$\hat{y}_t$ is the fitted value in time *t*

$\left|\left(e_t / y_t\right) \times 100\right|$ denotes as the absolute percentage errors calculated on the fitted values

n is the effective data points *t*

The error measure above was utilized to compare as well as select the best model by observing the smallest value in the error measure. The best-selected model was then applied for the whole actual data for the part of the model application, which is forecasting.

Results and Discussion

Figure 2 below shows the time series data for Malaysia's unemployment numbers between January 2010 and November 2022, which has been utilized for this study. Moreover, based on the observation, a time plot of Malaysia's unemployment numbers revealed a growing trend, a non-seasonal pattern, and considerable changes. This indicates that the actual data on unemployment numbers are not stationary. Furthermore, the ACF and PACF are plotted to collect more conclusive evidence on their fixed condition.

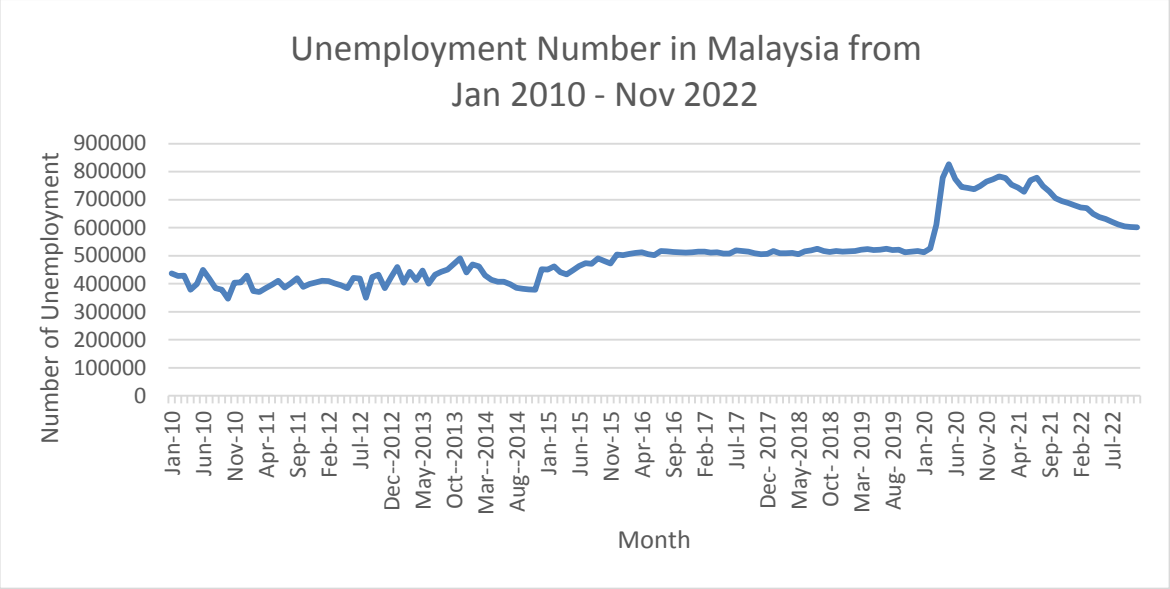

Figure 2 Time series plot of unemployment number in Malaysia from Jan 2010 until Nov 2022

The results were supported by the ACF graph in **Figure 3(a)**, which displayed a significantly decaying pattern. Moreover, in the Box-Jenkins model, it is necessary to have static data. Therefore, it is necessary to apply the differencing to achieve stationary in data series. The PACF graph in **Figure 3(b)** showed a rather significant spike in the beginning of the lag, then followed by lesser spikes. According to this, the stationary data series can be obtained days after the first differencing. Moreover, the time plot exhibits no regular fluctuation, whereas the ACF plot exhibits no wave-like pattern. This denotes that the data series has no seasonal variation, which prevented the need to differentiate the data series by season. However, the figure above depicts a situation in which Malaysia's unemployment rate has increased between January 2020 and May 2020. This situation occurred due to the outbreak of the Covid-19 endemic in Malaysia.

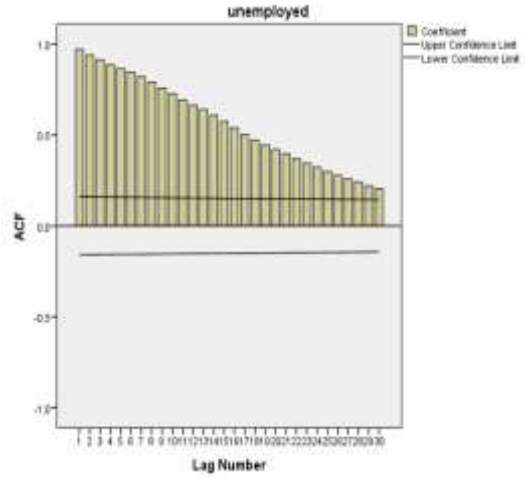

Figure 3(a) ACF Graph of unemployment in Malaysia

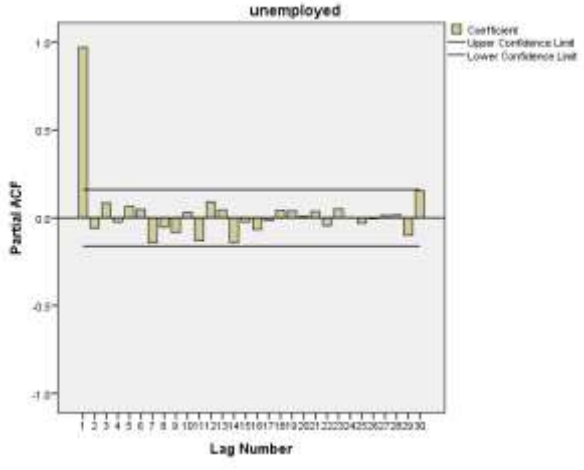

Figure 3(b) PACF Graph of unemployment in Malaysia

The original data series was made stationary by first-order differencing, known as a simple procedure. The first-order difference, z_t, was determined by using the formula below:

$$z_t = y_t - y_{t-1} \tag{3}$$

The ACF as well as PACF correlograms of $Z_t = y_t - y_{t-1}$ were subsequently plotted, as seen in **Figures 4(a)** and **4(b)**. The fading pattern has disappeared from the ACF plot, and spikes at different delays are now visible. After the first order differencing, ACF and PACF displayed an irregular pattern with various spikes; an ARIMA (p, d, q) model is the one that captured it best.

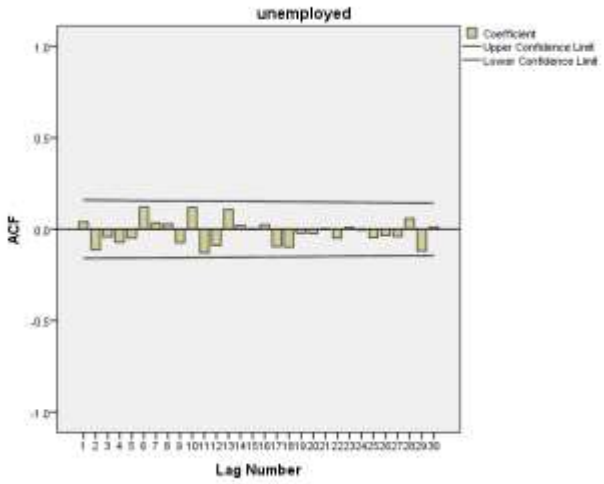

Figure 4(a) ACF graph of unemployment in Malaysia after first differencing

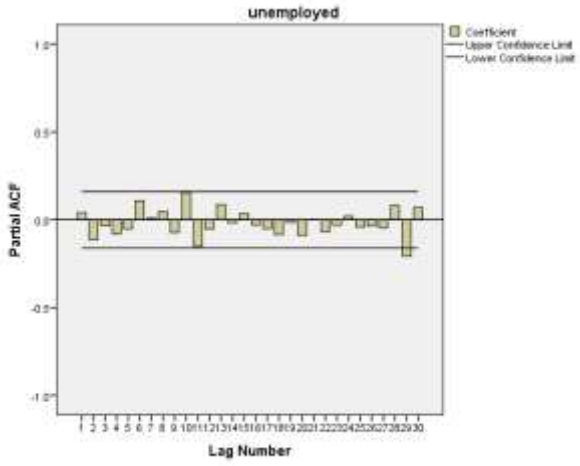

Figure 4(b) PACF graph of unemployment in Malaysia after first differencing

The next stage in this analysis is to identify the model. Identifying the suitable models for the fitted data series involved the ACF and PACF graphs from **Figure 4(a)** and **Figure 4(b)**. Since ARIMA (p, d, q) was chosen as the best suggested Box-Jenkins model, the parameters p as well as q have been determined by observing the significant spike in the ACF as well as PACF graphs. The ACF graph showed no significant spike exceeding the two standard error lines. Therefore, parameter q will be set using the ACF graph as well as equal to 0. In contrast, the PACF graph showed several significant spikes at lag 10, 11 and 29. Thus, the parameter p will be equal to 3. Meanwhile, the parameter d will be identified by the number of differencing. Since the differencing was applied for one time, p should be equal to 1. Hence, there are three ARIMA models that can be deduced by the following three models ARIMA (1, 1, 0), ARIMA (2, 1, 0) and ARIMA (3, 1, 0).

In well-fitted models, residuals that have been obtained are usually expected to have the property of white noise, which is a random error. Therefore, the Ljung-Box Q was provided to establish the well-fitted model with the stationary condition of the residual. The portmanteau test for each of the three models used for the estimating section is displayed in Table 1. Except for the first model, all subsequent models that have been applied to the data series matched the assumption of random error over time by accepting the null hypothesis as a result of attaining smaller calculated values than the corresponding tabulated values, according to the portmanteau test. However, only one model is needed among the three well-specified models. ARIMA (3, 1, 0) is the best model based on the smallest Q statistic. Nevertheless, the ARIMA (3, 1, 0) model does not show the lowest value of the Bayesian Information Criterion (BIC) compared to the ARIMA (2, 1, 0) model, as presented in **Table 1**. Therefore, since both models have the property of white noise, the best model has been decided by comparing the error measures of the evaluation part, as shown in **Table 2**.

Table 1 Summary of Portmanteau test using Ljung-Box Q

Statistics	**Model**		
	ARIMA (1,1,0)	ARIMA (2,1,0)	ARIMA (3,1,0)
Calculated Q	34.005	20.508	19.624
df	17	16	15
Tabulated Q*(0.05)	27.58	26.29	24.99
P-value	0.08	0.198	0.187
Decision	Reject Ho	Accept Ho	Accept Ho
Conclusion	The error is not random	The error is random	The error is random
BIC	20.165	20.127	20.173

Based on a comparison of the error measures in **Table 2**, ARIMA (3, 1, 0) obtained the smallest value of error measures for MSE as well as MAPE, with a very small difference when compared to another model. Additionally, as seen in **Table 1**, the value of the Q statistic for ARIMA (3, 1, 0) similarly had the lowest value. And as a result, it can be said that ARIMA (3, 1, 0) is the most accurate model for forecasting the number of unemployed people in Malaysia.

Table 2 Error measures for evaluation part

ERROR MEASURE	ARIMA (1,1,0)	ARIMA (2,1,0)	ARIMA (3,1,0)
RMSE	161526.8249	159974.5927	158706.8399
MAPE	17.82063735	17.57684862	17.38160644

Additionally, the model has been applied to the historical data to forecast the number of unemployed people in Malaysia for the upcoming month until the year 2023. ARIMA (3, 1, 0) was found to be the best Box-Jenkins model that fitted the time series data of unemployment in Malaysia. The historical data is shown in **Figure 5**, together with its fitted and projected values for 2023. Additionally, The plots also showed that the number of unemployed people in Malaysia had been estimated to increase gradually.

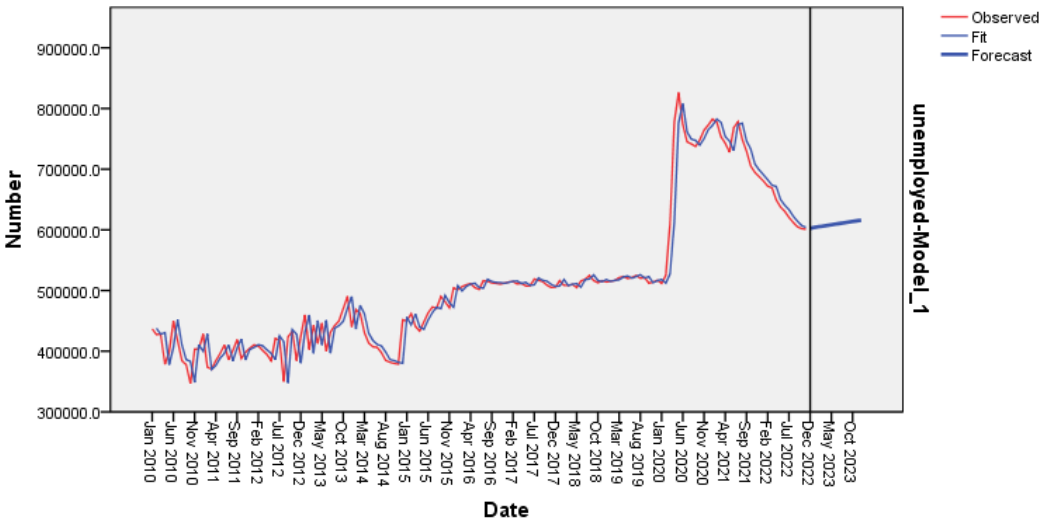

Figure 5 Time plot of actual, fitted, and forecast values for unemployment in Malaysia

Conclusion

In order to choose an ARIMA model as well as proceed to forecast, the Box-Jenkins modeling technique was presented in this study. Furthermore, The secondary data from January 2010 until November 2022 from the Department of Statistics Malaysia were considered for analysis. In the model identification process, ARIMA (1, 1, 0), ARIMA (2, 1, 0) and ARIMA (3, 1, 0) were obtained and then compared to identify the appropriate model to be used for forecasting. The results indicated that ARIMA (3, 1, 0) with the first-order difference was the best Box-Jenkins model based on the comparison of errors. Hence, it was used for forecasting the number of unemployed persons in Malaysia. As a result, the number of unemployed people in Malaysia has been anticipated to show an upward trend for the following year. Based on the findings, the graduated students should be aware of the increasing number of unemployed and be prepared for the future situation. The government must also take measures to manage the unemployment numbers and protect the overall economy. Studying additional factors, such as inflation, population growth, as well as irregularities, is advised because this paper primarily focused on the modeling procedure that could significantly affect the number of unemployment in Malaysia.

Ethics Statement

The research does not require research ethics approval.

Acknowledgement

The facilities used to support this research were provided by Universiti Teknologi MARA Pahang Branch, for which the author is grateful.

Conflict of interests

The authors have declared no conflict of interest exists.

References

Atif, A., & Ishak, Y. (2016). The impact of economic globalisation on unemployment: The

Malaysian experience. *The Journal of International Trade & Economic Development.* http://dx.doi.org/10.1080/09638199.2016.1151069.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis:forecasting and control*. San Fracisco: Holden-day.

Department of Statistics Malaysia Official Portal. [online] Dosm.gov.my. Available at: < https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=439&bul_id=b2U1a3J2VnRuaFNpNmY0Mlh4aUdIdz09&menu_id=Tm8zcnRjdVRNWWlpWjRlbmtlaDk1UT09> [Accessed 10 February 2023].

Dobre, I., & Alexandru, A. A. (2008). Modelling unemployment rate using Box-jenkins procedure. *Journal of Applied Quantitative Methods*, 3(2), 156–166.

Douglas, C. Montgomery, Cheryl, L., & Jennings, M. K. (2011). *Introduction to Time Series Analysis and Forecasting (illustrate)*. John Wiley & Sons.

Lazim, M. A. (2016). *Introductory Business Forecasting a Practical Approach 3rd Edition*. Kuala Lumpur, MY: UiTM Press.

Nikolaos, D., & Paraskevi, K. (2018). Forecasting unemployment rates in USA using Box-Jenkins methodology. *International Journal of Economics and Financial Issues*, *8*(1), 9-20.

Norliana, M. L., Nurul, L. N. M. R., N. Izzati, I., N. Azzarina, M., N. Rasyida, M. R., Fatin, A. H., & Hanafi, I. (2021). Comparative study of Smoothing Methods and Box-Jenkins model in forecasting unemployment rate in Malaysia. *GADING Journal of Science and Technology*, 4(1).

Stanley, J., & Edwin, M. (2019). Modelling epidemiological data using Box-Jenkins procedure. *Open Journal of Statistics,* 6*,* 295-302.

Suraya, F. R., M. Firdaus, Hazmi, U., M. Khairi, & Amirul, Z. (2018). Prediction of the unemployment rate in Malaysia. *International Journal of Modern Trends in Social Sciences*, 1(4), 38 – 44.