ICIC Express Letters
Part B: Applications
Volume **11**, Number **6(tentative)**, June **2020**

ICIC International ©2020 ISSN 2185-2766
pp. ICICIC2019-SS15-03

VACCINE PREDICTION SYSTEM USING ARIMA METHOD

Julian Satya Sahisnu[1], Friska Natalia[1,*], Ferry Vincenttius Ferdinand[2]
Sud Sudirman[3] and Chang Seong Ko[4]

[1]Department of Information Systems
Universitas Multimedia Nusantara
Jl. Scientia Boulevard, Gading Serpong, Tangerang, Banten 15811, Indonesia
*Corresponding author: friska.natalia@umn.ac.id

[2]Department of Mathematics
Universitas Pelita Harapan
Jl. M. H. Thamrin Boulevard, 1100 Lippo Village, Tangerang 15811, Indonesia

[3]Department of Computer Science
Liverpool John Moores University
Liverpool L3 3AF, UK

[4]Department of Industrial and Management Engineering
Kyungsung University
309, Suyeong-ro, Nam-gu, Busan 48434, Korea

Received December 2019; accepted March 2020

Abstract. *Indonesia is a country that runs health programs in the form of primary compulsory immunizations for children aged 0-11 months. According to the Health Law, Number 36 of 2009 states that every child has the right to receive primary immunization by the provisions to prevent the occurrence of diseases that can be avoided through immunization. The Indonesian government are also obliged to provide complete immunization to every baby and child by the implementation of immunization contained in the Minister of Health Regulation Number 42 of 2013. The purpose of this study is to predict vaccine stock for immunization needs, and the government can use the application to determine vaccine stock requirements for each clinic so that there is no shortage or excess stock. This prediction can ensure that immunization coverage is well distributed. We can help parties who organize primary immunization activities by making predictions and forecasting results based on the R application. In addition, the application can provide information in the form of predictive analysis. The method used in measuring predictions is ARIMA (Auto-Regressive Integrated Moving Average) to calculate the prediction of immunization.*

Keywords: Inventory forecasting, ARIMA, Immunization, Time series, Data visualization, Vaccine-preventable diseases

1. **Introduction.** Vaccine-Preventable Diseases (VPDs) are infectious diseases for which an effective preventive vaccine exists. While most VPDs are practically kept under control in developed countries, they still pose a significant risk to a population's health in most developing countries. In Indonesia, for example, the country's Ministry of Health estimated that VPD is responsible for between two to three million deaths every year. A vaccination program is considered as the most cost-effective and the most efficient way to improve the general health of a population by preventing VPDs and to avoid vaccine-preventable deaths.

Vaccination, or immunization, is a way to increase one's immunity to VPDs by injecting an agent that resembles the disease-causing microorganism. This agent is often made from weakened or killed forms of the microorganism, its toxins or one of its surface proteins to

DOI: 10.24507/icicelb.11.06.xxx

train the body's immune system so that it can fight the microorganism when it encounters them [1]. Vaccination is a disease prevention measure as opposed to a disease treatment measure hence it is often administered to babies and small children. Several necessary types of vaccinations are given to children under 18 months; these include Measles, Polio, BCG (Bacillus Calmette-Guerin), HBo, and DTPHBHib (Diptheria, Tetanus, Pertussis, Hepatitis B, Pneumonia and Meningitis) [2].

The demand for vaccines can vary significantly due to birth rates, the occurrence of outbreaks, or manufacturing problems causing stock shortages. This problem provides a rationale for proper inventory management to be implemented to prevent stock shortages [3]. It is a means to predict the availability of a commodity over a time period in the future based on historical data using various qualitative and quantities approaches [4]. Demand forecasting systems can help prevent stock shortages when changes in vaccine demand for example due to natural disasters, changes in labor demand, or political unrest. Without a demand forecasting system, vaccine stock managers may find it difficult to predict changes in demand at vaccination sites [5]. Auto-Regressive Integrated Moving Average (ARIMA) is a popular method used by researchers to predict medicines stock levels at multiple health centers to prevent shortages and oversupply [6]. The method has also been applied to predicting other types of stock level including agricultural commodity [7], and to predicting the number of visitors requiring emergency attention, where the latter has been shown to be sufficiently able to predict the number of visitors to the hospital and can be employed to assist in the decision making process [8]. In [9], drug inventory cost in a hospital is reduced by representing the demand for antibiotics drugs based on diabetic foot ulcer patients' condition by adapting a Markov Decision Process (MDP) and subsequently use the model to determine the drug's appropriate inventory.

In this paper, we describe the development of a vaccine prediction system using the ARIMA method and its implementation as an R-based application. As a case study, we apply this system to predicting vaccine stock in South Tangerang City, a regency west of Jakarta. The remainder of the paper is organized as follows. Section 2 points the problem statement and research methodology. Section 3 describes the implementation and result. Section 4 presents the conclusion and future research area.

2. Problem Statement and Research Methodology.

The research work described in this paper aims to develop the vaccine prediction system for each health center in the area of South Tangerang City. The object of the research used in this study is vaccine stock history data at the South Tangerang City Health Center between 2013 and 2017. The method used in this study is ARIMA (Auto-Regressive Integrated Moving Average) method. In the ARIMA method, there are steps in carrying out calculations. A process in carrying out the ARIMA method. There is identification of data to find out whether there are outliers in the data, then stationary test to find out differencing requirements on data, auto-correlation test to determine AR and MA values, fit ARIMA model to test the most appropriate ARIMA model used, and model evaluation by calculating success rates towards the model [10].

Outliers can be determined by defining the first quartile (Q_1), third quartile (Q_3), and interquartile range (IQR). Any data that is lower than $Q_1 - 1.5IQR$ and greater than $Q_3 + 1.5IQR$ is an outlier. Stationary test is done by doing the Augmented Dickey-Fuller [11] test $Z_t = \rho Z_{t-1} + a_t$, where $\rho = 1$ represents a model that fits stationary data and where $\rho \neq 1$ otherwise.

Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) will be used to determine ARIMA form. The ACF formula is $r_k = \frac{\sum_{t=k+1}^{n}(Y_t-\bar{Y})(Y_{t-k}-\bar{Y})}{\sum_{t=1}^{n}(Y_t-\bar{Y})^2}$ and the PACF formula is $\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1}\phi_{k-1,j}\rho_{k-j}}{1-\sum_{j=1}^{k-1}\phi_{k-1,j}\rho_j}$. While the general form of ARIMA (p, d, q) is $W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \cdots + \phi_p W_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$ with $W_t = \nabla^d Y_t$.

Results from the stationarity test will be used to determine d, ACF to determine p, and PACF to determine q [6,12,13].

3. **Implementation and Result.** There are several tools that can be used to implement data science processes and visualize the results. Data analytic softwares such as Power BI, Tableau and R software, for example, have been used to develop an interactive dashboard to cluster the tourism objects in Bali [14] and to cluster and visualize the data taken from river stations in Tangerang regency area [15,16]. In this paper, we use the R programming language to implement the predictions and visualization of the vaccines. R is a popular language in the data science community to conduct research because it is well developed, simple, and effective [17]. R Studio is used particularly in this research because the program has a library forecast that can assist in making predictive models. This forecast library has an ARIMA function that can do ARIMA mathematical calculations so that it can support this research. An R-based application will be built using the Shiny package to help create a vaccine prediction system application and create an ARIMA model to represent the results. The application development process is divided into four stages.

Stage 1: Data preparation. In this stage, data cleaning is carried out using Power BI tools to provide visualization of the data that is owned and use R Studio to clear data to calculate predictive values.

Stage 2: R model development. In this stage, the prediction model is developed using the R programming language using R Studio software. In the ARIMA method, several steps must be done before making a prediction. The first step is to plot existing data to test whether there are outliers in the data. If there are outliers in the vaccine data they can interfere with the process of calculating predictions. Therefore, it is necessary to delete these outliers. The mean and spread of the data before and after removing are shown as a box plot in Figures 1 and 2.

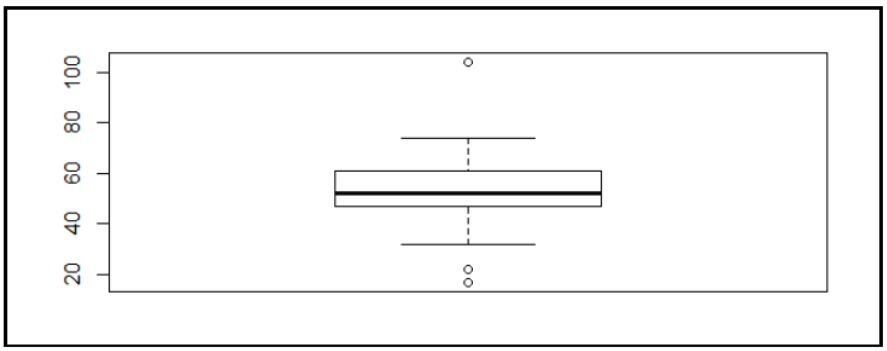

FIGURE 1. Mean and spread of the data with outlier

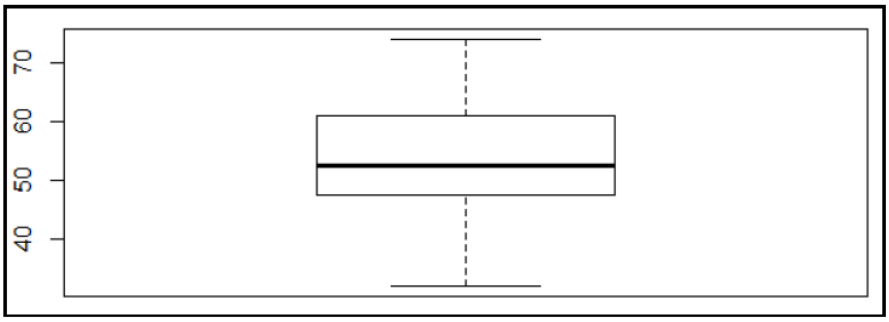

FIGURE 2. Mean and spread of the data after outlier removal

To use the ARIMA method, there is a requirement that the data must be stationary; therefore, in the next stage, stationary testing must be performed. If the results were found to be not stationary, a differencing process needs to be carried out. This process is not needed if the data were found to be stationary. The stationary test data is carried out using the Augmented Dickey-Fuller (ADF) test with a standard hypothesis.

$H_0 = p$-value > 0.05, time series is not stationary

$H_1 = p$-value < 0.05, stationary time series

Based on the result of stationary testing so that it can be a requirement to use the ARIMA model. The p-value obtained is 0.0841. Since the p-value is greater than 0.05, we can conclude that there is not enough evidence to reject hypothesis H_0; therefore, we can say that the polio 1 data in Serpong 1 health center is non-stationary data in Figure 3.

```
> adf.test(Serpong1$`Polio 1`)

        Augmented Dickey-Fuller Test

data:  Serpong1$`Polio 1`
Dickey-Fuller = -3.2781, Lag order = 3, p-value = 0.0841
alternative hypothesis: stationary
```

FIGURE 3. Non-stationary data

The process of differentiation uses the diff() function. Differentiation is done once this is because the data used is monthly. The following results obtained are 0.01; therefore, hypothesis H_0 is rejected and we can conclude that polio 1 data in Serpong 1 health center is stationary data and can be used in the next step. The value of d in ARIMA is number 1 shown in Figure 4.

```
> adf.test(diff(Serpong1$`Polio 1`,1))

        Augmented Dickey-Fuller Test

data:  diff(Serpong1$`Polio 1`, 1)
Dickey-Fuller = -6.057, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary
```

FIGURE 4. Stationary data

Stage 3: Application development. In this step the application development uses a Shiny package that is run from R Studio. Based on the data, source code is used to create global parameters so that it can be run across applications. This global parameter uses the "read.xlsx" function to call the data to be used in the form of data that has an Excel format. The next step in the ARIMA prediction calculation process is to create a model to determine the values of AR and MA or p and q. The determination of AR and MA values can use the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) methods in Figures 5 and 6.

The lag value on ACF is 1 while the lag value on PACF is 1 and 16 so it can be concluded that the ARIMA model obtained is ARIMA (1, 1, 1) and ARIMA (1, 1, 16). Then, checking the ARIMA model uses the forecast () function. Figure 7 and Figure 8 show the results of calculating the ARIMA model (1, 1, 16) and the ARIMA model (1, 1, 1).

This application is displaying vaccine data in graphical form so that it is easy to read by users. The graph used is in the form of a straight line and use the "highchartoutput"

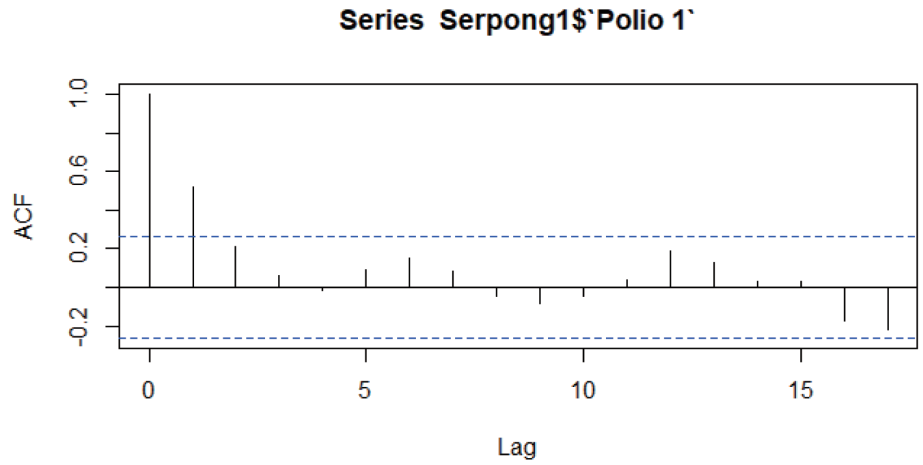

FIGURE 5. Auto-correlation function

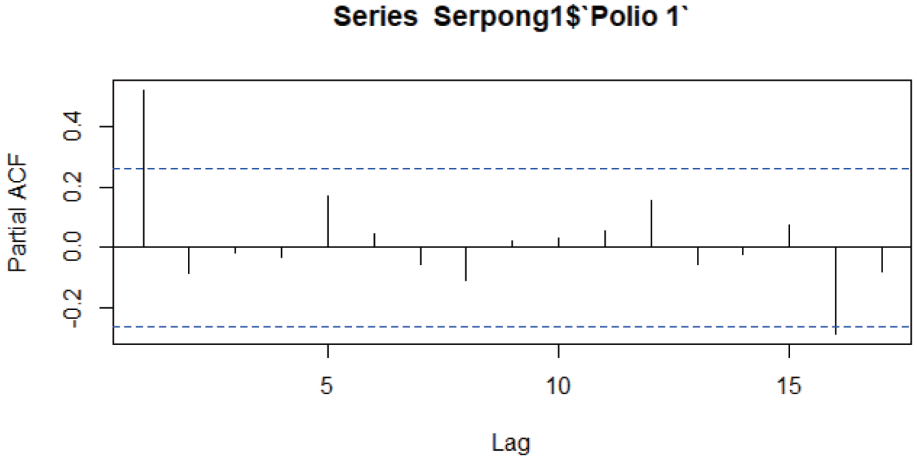

FIGURE 6. Partial auto-correlation function

Forecasts:

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
57	51.10099	43.47396	58.72801	39.43646	62.76551
58	48.24068	39.48023	57.00112	34.84273	61.63862
59	53.83185	44.76413	62.89957	39.96397	67.69973
60	52.02613	42.88924	61.16302	38.05246	65.99980
61	44.26248	35.11693	53.40803	30.27556	58.24939
62	51.01498	41.66730	60.36266	36.71893	65.31103
63	50.13209	40.56876	59.69541	35.50624	64.75793
64	50.21605	40.08558	60.34652	34.72284	65.70927
65	50.05902	39.77567	60.34236	34.33200	65.78603
66	52.17499	41.88093	62.46905	36.43159	67.91839

FIGURE 7. ARIMA (1, 1, 16) results

function to produce the output of the chart we want and also use the "renderhighchart" function to display visualizations for all types of vaccines in the R programming language using R Studio software. The result can be seen in Figure 9.

The "plotoutput" function in the source code serves to display the output in the form of a plot so that it can assist users in reading the predicted results. This source code is used on the server to make ARIMA forecast predictions according to the model previously

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
57	53.12728	43.22634	63.02822	37.98510	68.26946
58	52.68592	41.32780	64.04404	35.31517	70.05666
59	52.46271	40.61430	64.31111	34.34213	70.58328
60	52.34982	40.28708	64.41256	33.90145	70.79819
61	52.29273	40.11042	64.47504	33.66149	70.92396
62	52.26386	39.99923	64.52848	33.50673	71.02099
63	52.24925	39.91854	64.57997	33.39106	71.10745
64	52.24187	39.85269	64.63105	33.29425	71.18948
65	52.23813	39.79426	64.68201	33.20687	71.26939
66	52.23625	39.73964	64.73285	33.12434	71.34815

FIGURE 8. ARIMA (1, 1, 1) results

FIGURE 9. Visualization of vaccines data

made. Figure 10 shows the result of a prediction of a vaccine that displays predictive values and plots.

Stage 4: Verification of forecast results. Before the ARIMA model was used it was necessary to test the accuracy of the model to determine the percentage level of errors in the use of the model. MAPE (Mean Absolute Percentage Error) is a method to test the error level of a model with a percentage value to be easy to read.

In the ARIMA model (1, 1, 1) the results are 11.32606 if the percentage is made to 11.32%, meaning that the value has a success rate of 89.68%. Figure 11 shows the calculation result using the MAPE (Mean Absolute Percentage Error) for the ARIMA (1, 1, 1) model.

Figure 12 shows the results of the ARIMA model (1, 1, 16) used are obtained at 8.625089 if it is made in a percentage number to 8.62%, meaning that the value means that it has a success rate of 91.38% so that the ARIMA model (1, 1, 16) can be said to be better than the model ARIMA (1, 1, 1) and can be used to make predictions.

Before this study, the available vaccines often experienced excess or lack of stock so that when there was a shortage of vaccine stock, the South Tangerang City Health Service could not provide immunization. After conducting vaccine prediction research using the ARIMA method with the ARIMA model (1, 1, 16) obtained stock results minimum that can be a reference by the South Tangerang City Health Office to provide vaccine stock

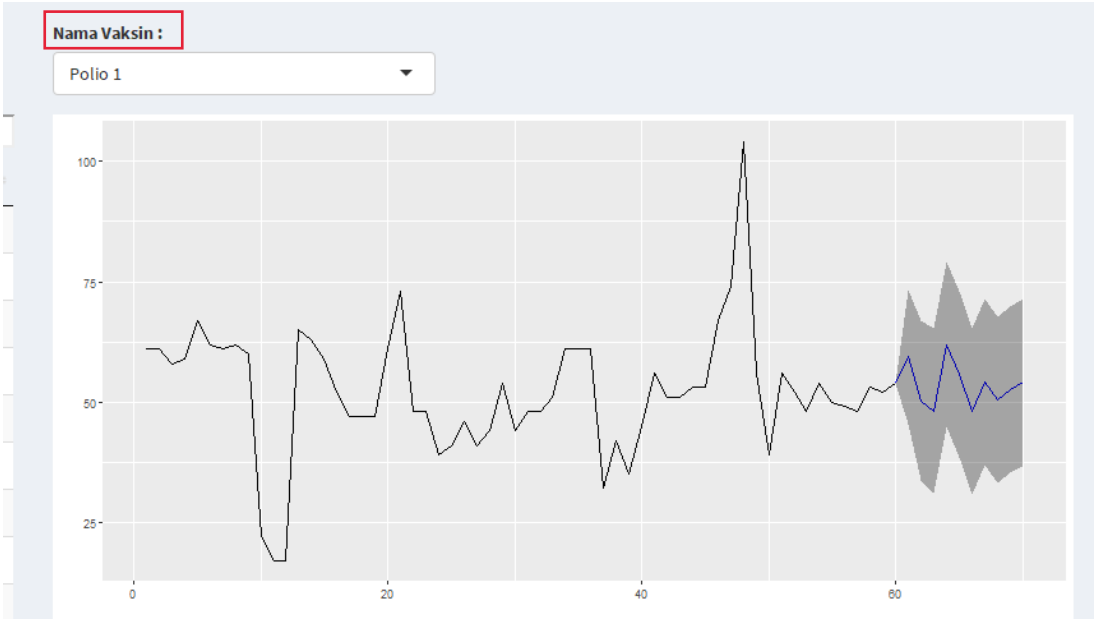

FIGURE 10. Prediction of vaccine in clinic center

```
Error measures:
                     ME      RMSE       MAE        MPE      MAPE
Training set -1.13444 7.655956 5.503708 -4.238831 11.32606
```

FIGURE 11. Error measures of ARIMA (1, 1, 1)

```
Error measures:
                      ME      RMSE       MAE       MPE      MAPE
Training set -0.8489744 5.447732 4.229008 -2.797556 8.625089
```

FIGURE 12. Error measures of ARIMA (1, 1, 16)

requirements for 10 consecutive months. The South Tangerang City Health Office can provide vaccine stock requirements using the vaccine prediction application and the Health Service can view vaccine stock data through the graphics provided by the application.

4. **Conclusion.** We have proposed a novel strategy to predict the required vaccine stock level for immunization purposes. Our methodology utilizes a stationary test data using Augmented Dickey-Fuller method to calculate the values of the auto-correlation function and partial auto-correlation function. The result is used to design and validate an appropriate ARIMA model. In our experiment, we apply our proposed methodology to predicting the required vaccine stock level using the data from South Tangerang Health Service in the Serpong Area 1 in South Tangerang, Indonesia during January 2018 – October 2018. We found that in this case, the ARIMA method (1, 1, 16) is an ARIMA model to be used. This model is then implemented in R Studio and used to make predictions on the required vaccine stock level to anticipate the stock shortages before they occur. The accuracy of the ARIMA method (1, 1, 16) model is calculated using the mean absolute percentage error and is found to be 91.38%. The finding of our work is significant because of the novelty in applying the different methodologies and combining them to make the successful prediction of vaccine stock level for immunization purposes. In the future, we plan to apply this methodology to a larger scale such as a provincial Health Services in Jakarta, Banten and West Java provinces to assess the scalability of the overall strategy.

Acknowledgments. All data provided by South Tangerang City Health Service and all computations are performed in Big Data Laboratory of Universitas Multimedia Nusantara.

REFERENCES

[1] The World Health Organization, *Vaccines*, Available at https://www.who.int/topics/vaccines/en/, Accessed on 15-Sep-2019.
[2] NHS UK, *Vaccinations*, Available at https://www.nhs.uk/conditions/vaccinations/, Accessed on 15-Sep-2019.
[3] C. Mekel, S. P. D. Anantadjaya and L. Lahindah, Stock out analysis: An empirical study on forecasting, re-order point and safety stock level at PT Combiphar, Indonesia, *RIBER Rev. Integr. Bus. Econ. Res.*, vol.3, no.1, pp.52-64, 2014.
[4] B. Brunaud, J. M. Laíinez-Aguirre, J. M. Pinto and I. E. Grossmann, Inventory policies and safety stock optimization for supply chain planning, *AIChE J.*, vol.65, no.1, pp.99-112, 2019.
[5] L. E. Mueller et al., The impact of implementing a demand forecasting system into a low-income country's supply chain, *Vaccine*, vol.34, no.32, pp.3663-3669, 2016.
[6] C.-Y. Cheng, K.-L. Chiang and M.-Y. Chen, Intermittent demand forecasting in a tertiary pediatric intensive care unit, *J. Med. Syst.*, vol.40, no.10, p.217, 2016.
[7] H. Wu, H. Wu, M. Zhu, W. Chen and W. Chen, A new method of large-scale short-term forecasting of agricultural commodity prices: Illustrated by the case of agricultural markets in Beijing, *J. Big Data*, vol.4, no.1, 2017.
[8] W.-C. Juang, S.-J. Huang, F.-D. Huang, P.-W. Cheng and S.-R. Wann, Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in southern Taiwan, *BMJ Open*, vol.7, no.11, p.e018628, 2017.
[9] B. A. Sheldon, M. L. Mahadevan, E. A. Kumar et al., Demand forecasting and inventory cost reduction for the antibiotics for inpatients in hospital using Markov decision process, *International J. Res. Anal. Rev.*, vol.6, no.1, pp.554-564, 2019.
[10] R. Adhikari and R. K. Agrawal, An introductory study on time series modeling and forecasting, *arXiv Prepr. arXiv1302.6613*, 2013.
[11] W. A. Fuller, *Introduction to Statistical Time Series*, John Wiley & Sons, 2009.
[12] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications: With R Examples*, Springer, 2017.
[13] S. Aziz, A. Sayuti and M. Mustakim, Penerapan Metode ARIMA Untuk Peramalan Pengunjung Perpustakaan UIN Suska Riau, *Seminar Nasional Teknologi Informasi Komunikasi dan Industri*, pp.186-193, 2017.
[14] S. Monica, F. Natalia and S. Sudirman, Clustering tourism object in Bali province using k-means and x-means clustering algorithm, *2018 IEEE the 20th International Conference on High Performance Computing and Communications; IEEE the 16th International Conference on Smart City; IEEE the 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pp.1462-1467, 2018.
[15] F. Natalia, Y. Eko, F. V Ferdinand, I. M. Murwantara and C. S. Ko, Interactive dashboard of flood patterns using clustering algorithms, *ICIC Express Letters, Part B: Applications*, vol.10, no.5, pp.413-418, 2019.
[16] F. Natalia Ferdinand, Y. Soelistio, F. Vincenttius Ferdinand and I. Murwantara, Cluster-based water level patterns detection, *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol.17, pp.1376-1384, 2019.
[17] W. N. Venables, D. M. Smith, R. D. C. Team et al., *An Introduction to R*, Network Theory Limited, 2009.