

Time Series Regression: Prediction of Electricity Consumption Based on Number of Consumers at National Electricity Supply Company

Mohammad Idhom¹, Akhmad Fauzi², Trimono Trimono¹, Prismahardi Riyantoko¹

¹ Department of Data Science, University of Pembangunan Nasional Veteran Jawa Timur, Indonesia

² Department of Management, University of Pembangunan Nasional Veteran Jawa Timur, Indonesia

Abstract – Electrical energy is one of the components of Gross Domestic Product that is able to encourage the economy because it has become a basic need of the community. To meet the increasing demand for electrical energy, the Indonesia National Electricity Providers (PLN) need to predict the amount of electrical power required based on the customer numbers to meet the demand for adequate electricity supply. This study aims to predict electric power based on electricity user customers using a time series regression model. The data used in this study are secondary data which get from PLN annual report in 2021. This study resulted in a finding of the best prediction model based on the Akaike Information Criterion (AIC) value, namely the time series regression model with the error value modeled by the AR(1) model, while the forecasting accuracy measure used the value MAPE of 9.77%. This means that the result of model prediction is highly accurate.

Keywords – Power electricity usage, number of consumers, time series regression, MAPE.

DOI: 10.18421/TEM123-39

<https://doi.org/10.18421/TEM123-39>

Corresponding author: Trimono Trimono,
Department of Data Science, University of Pembangunan
Nasional Veteran Jawa Timur, Indonesia


Email: trimono.stat@upnjatim.ac.id

Received: 28 April 2023.

Revised: 31 July 2023.

Accepted: 09 August 2023.

Published: 28 August 2023.

 © 2023 Mohammad Idhom, Akhmad Fauzi, Trimono Trimono & Prismahardi Riyantoko; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDeriv 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

1. Introduction

Gross Domestic Product (GDP) is an important factor in Indonesia's economic situation. One of the components of GDP which can encourage economic matters and increase the standard of living is electrical energy [19]. According to the Central Bureau of Statistics Indonesia (known as BPS Indonesia) in 2020, the electrical energy usage in Indonesia in 2019 is up to 188.342,41 GWh, while in 2019, it reaches up to 199,892.8 GWh. Thus, as being said, the usage of electrical energy is increasing up to 6.13% in 2019-2020. This number will be increasing along with the economic progress in Indonesia. Meanwhile, in East Java, the electricity usage in 2019 is reached 18,205.08 GWh and increasing to 19,596.00 GWh [3].

Most of the electrical energy comes from a State-owned Enterprise (SOE, known as BUMN), which is a PLN Company (Persero). PLN Company is an electricity company that does planning and performing electrical power with a long lead time, so it requires a long-term expansion plan of the electrical system. PLN needs to develop the system or operation to fulfil the consumption needs which are increasing every year [6]. This is because electricity becomes a basic need in common practice, so the number of PLN consumers is growing every month. The more numbers of consumers, the more power is being produced by PLN [17].

One of the methods which can be used to predict the occurrence in a future period based on data or previous periods is Time Series Analysis. Time series analysis is not only able to be done on univariate, but also for many variables (multivariate) [8]. On the multivariate model, it could be a bivariate data analysis and multivariate data. One of the multivariate models for time series analysis is the Time Series Regression Model [5]. In the analysis of time series, it can develop a displeasing residual assumption; one of them assumption of normality.

The cause of the abnormality is an outlier in residual, so it requires a procedure on detecting the outlier and the parameter must be reassessed by model [24]. After receiving the best outlier, it can be used for the prediction of future period.

A portion of past studies that investigate the modeling of time series regression, focusing specifically on the prediction of Jute Yarn demand in Bangladesh is presented in [2]. The result shows that this model gives a better approach on the Jute Yarn demand prediction. This model can be used to predict further Jute Yarn demand prediction which can assist investors to take best business strategy that needs to carry out. Model of Time Series Regression is used for modelling and predicting number of death cases caused by COVID-19 on several countries in the world [1]. The result shows that this model is effective to predict death cases with MAPE score of 2.23%. Other research conducted by [2], predicts numbers of airline passengers in USA by using Time Series Regression model. The result shows that the modeling is suitable for numbers of passenger in 1974, with MAPE score of 5%.

This research examines the application of time series regression model on power electricity usage (as dependent variable) and numbers of PLN consumers (as independent variable) with chosen study case in PT. PLN of Surabaya City, Indonesia between January 2015 – December 2020.

2. Theoretical Framework

This section will explain the theories that will be used in predicting electrical consumption in Surabaya city using a time series regression model. the explanation will start from the Cross Correlation Function (CCF), then Phrewhitening, time series regression analysis, and the last is Performance Measurement Model

2.1. Cross Correlation Function (CCF)

The first step before doing modeling for time series regression is detecting and measuring power relation between variable X and Y by using Cross Correlation Function (CCF). Cross correlation can be calculated by using equation [12]:

$$r_{XY}(k) = \frac{C_{XY}(k)}{\sqrt{C_{XX}(0)C_{YY}(0)}} \quad (1)$$

$$r_{YX}(k) = \frac{C_{YX}(k)}{\sqrt{C_{XX}(0)C_{YY}(0)}} \quad (2)$$

k is time lag, $r_{XY}(k)$ is the cross correlation between variable X and Y on lag k . $C_{XY}(k)$ is the cross covariance between X and Y on lag k . $k = 0, \pm 1, \pm 2, \pm 3, \dots$

2.2. Prewhitening

Prewhitening is used to obtain a white noise series [16]. For example, if series X is being modeled as ARIMA process (p, d, q) , it can be defined as follows:

$$\varphi_p(B)(1 - B)^d X_t = \theta_q(B)e_t \quad (3)$$

By rearranging the equation term on (3) so it can differentiate series X_t into series e_t , as follows:

$$e_t = \frac{\varphi_p(B)(1 - B)^d}{\theta_q(B)} X_t \quad (4)$$

This series e_t is gone through prewhitening series X_t . Steps of prewhitening on series X_t shall be applied to series Y_t . Series Y_t which has been through prewhitening action is called series β_t , with [14]:

$$\beta_t = \frac{\varphi_p(B)(1 - B)^d}{\theta_q(B)} X_t \quad (5)$$

2.3. Linear Regression Analysis

Linear Regression Analysis is statistical analysis which discusses the linear correlation on independent variable with dependent variable [9]. It is aimed to determine value of dependent variable in certain condition of independent variable. According to [7], regression model with k regressor can be determined using the following formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon \quad (6)$$

With variable Y is the dependent variable and variable X is the independent variable, β_0 is the intercept, and β_j is regression coefficient, with $j = 1, 2, \dots, k$, and ε is errors. The probability model for linear regression is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k + \varepsilon \quad (7)$$

In this research, parameter value of regression model is determined by method of Ordinary Least Square (OLS). The choice of this method is to carry some merits, which are, “the statistical method reveals information about cost structures and distinguishes between different variables’ roles in affecting output. The adjustment turns the OLS into a ‘frontier’ approach” [10]. After getting parameter estimation, the next step is to examine assumption testing and hypothesis testing. This examined assumption consists of normality, homoscedasticity, non-autocorrelation, and residual non-multicollinearity. The hypothesis testing is aimed to examine the suitable modeling and the significant impact on each parameter on the modeling [13].

2.4. Time Series Modeling

Time series is a form of data collected based on a certain time sequence. The basic idea of time series is current observation (Z_t), depending on several previous observations [23]. The purposes are to comprehend and define certain mechanism, to predict value in the future, and to optimize control system [20].

Autoregressive Integrated Moving Average (ARIMA) Model

ARMA model is aimed to describe stationary time series, while ARIMA is aimed to describe non-stationary time series with variation of homogeneity [11]. The implementation of stationary time series modeling is rarely used. Therefore, it is required to carry out differencing process to make the stationary. Common modeling of order autoregressive p , differential order d , and order of moving average order q (ARIMA (p, d, q)) are the results of compiling between model AR(p) and MA(q) modeling by stationed non-stationary process [4], where d is the order of differentiate. Differentiate is a proses of subtraction between period data t with previous data period. By equation modeling of ARIMA (p, d, q) can be arranged as follows [15]:

$$\phi_p(B)(1 - B)^d Z_t = \theta_0 + \theta_q(B)a_t \tag{8}$$

with:

$$\begin{aligned} \phi_p(B) &= (1 - \phi_1 B - \dots - \phi_p B^p) \\ \theta_q(B) &= (1 + \theta_1 B + \dots + \theta_q B^q) \end{aligned}$$

If $d = 0$, model of equation (8) is stationary time series modeling.

2.5. The Best Regression Model Selection

The Best Regression Model Selection is used to obtain the most significant model for prediction. The best regression model selection is chosen based on the model by the smallest value of AIC. The formula of AIC described as follows [21]:

$$AIC = n \ln(\hat{\sigma}_a^2) + 2(p + q) \tag{9}$$

with:

- $\hat{\sigma}_a^2$: estimated residual variance
- n : numbers of observation
- p, q : order in ARIMA model

2.6. Performance Measurement Model

Performance measurement model can be done by using Mean Absolute Percentage Error (MAPE). MAPE is a method commonly used to evaluate the forecasting value by considering the impact of numbers of actual value [18].

The measurement of MAPE is stated below:

$$MAPE = \left(\frac{1}{L} \sum_{t=1}^L \left| \frac{S_t - \hat{S}_t}{S_t} \right| \right) \times 100\% \tag{10}$$

with:

- S_t : Actual value on t -th period
- \hat{S}_t : Predictive value on t -th period
- L : sample size

A good model is its accuracy of prediction is high when the actual value and predictive value have small number of differences [22]. The following table shows accuracy scale of predictive results based on MAPE value:

Table 1. Accuracy Scale on MAPE Evaluation

MAPE value	Accuracy Scale
< 10%	Highly accurate
11% - 20%	Good forecast
21% - 50%	Reasonable forecast
>51%	Inaccurate forecast

3. Results and Discussion

This research consists of two variables, which are the number of customers (as independent variable/ X) and the number of electric power consumptions (as dependent variable/ Y) of PT. PLN in Surabaya City. For each variable, the data used are monthly data starting from May 2016 – May 2021 (61 months). A time series plot of customer and electric power consumption is described below:

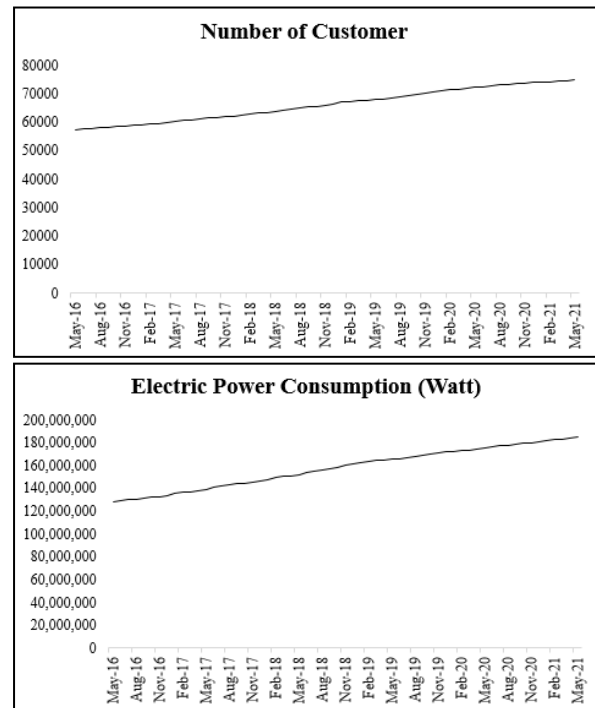


Figure 1. Time series plot for numbers of customer and electric power consumption

Based on time series plot in Figure 1, we can determine that numbers of customer and electric power consumption is ascending periodically, with various changes in value. For the next step, to observe the characteristics of data in detail, it can be determined by value of descriptive statistics.

Table 2. Descriptive statistics of numbers of customer and electric power consumption

	Customers	Power (Watt)
Minimum	57,216	127,770,027
Maximum	74,754	184,868,383
Median	66,064	157,870,488.44
Standard Deviation	5,535.52	17,377,844.83
Skewness	0.03	-0.15
Kurtosis	-1.36	-1.28

Based on Table 2, the average consumption of electrical power is 157,870,488.44 Watt per month. Skewness value on both variables shows that every data has asymmetrical distribution curve.

Electrical power consumption modeling and prediction based on numbers of customer by time series regression model requires in-sample and out-sample data. In-sample data is used to build a model and out-sample data is used to validate a model. For each variable, in-sample data is determined in total of 56 data from May 2016 to December 2020.

Before using time series regression model, it needs an examination if the modeling variable has significant correlation. By using correlation test, we obtain the result for correlation value as follow:

Table 3. Correlation value customer (X) and electrical power consumption (Y)

	X	Y
X	1	0.961
Y	0.961	1

Table 4. Result of Assumption Test of Linear Regression Model

	Normality	Homoscedasticity	Non-autocorrelation
Regression model	✓	✓	×

On the assumption test, the non-autocorrelation assumption is not met. This shows that there is autocorrelation between residual models. Therefore, time series regression model is needed to do

The correlation value that is formed between variables X and Y is 0.961, so it can be shown that both variables have a strong correlation. The more value X has, the more value Y goes.

Besides linear correlation, which is determined by Pearson correlation, we need to examine a strong cross correlation between X and Y. If there is a strong cross-correlation, it needs a whitening process to get an unbiased modeling result.

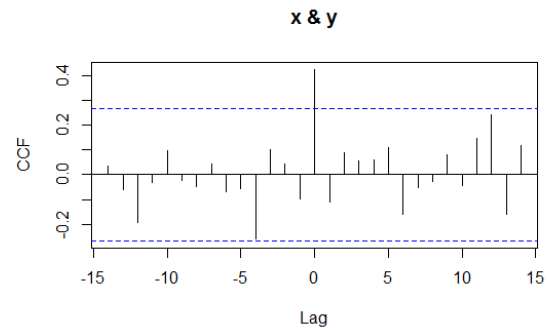


Figure 2. Cross correlations customers (X) and electrical power consumption (Y)

Based on CCF plot in Figure 2, it shows that a significant correlation value between variable X and Y is occurred when lag is 0, which means there are no strong cross correlations, so that does not require prewhitening process.

As it does not require prewhitening process, the analysis can be preceded to steps of linear regression model. Based on estimation value by using software R, we obtain the value result $\beta_0 = -49.730.693$ and $\beta_1 = 3.143$. So, the linier regression model formed is:

$$\hat{Y}_t = -49,730,693 + 3.143X_t + e_t \quad (11)$$

where Y_t is electrical power consumption of PLN (watt) period t , and X_t is numbers of costumer period t . R^2 (Coefficient of determination) in total 0.9906, means that 99.06% of total of electrical power consumption is affected by the numbers of costumer. On hypothetical testing on model compatibility through F testing, at the significant level $\alpha = 5\%$, the regression model is suitable to apply. Then, on t testing to examine the significance of model parameter, we get the conclusion that parameter affects model significantly on the level $\alpha = 5\%$. On the assumption test, the results are presented in Table 4:

modeling on the residual model. The first procedure on time series regression modeling is examining stationaries on residual data.

The results using the Augmented Dickey-Fuller Test is provided in the table below:

Table 5. Results of ADF test on residual data residual data of regression model

Variable	Hypothesis	t-statistics	Sig level (α)	p-value
Residual regression model	H_0 : residual data is not stationary H_1 : residual data is stationary	-5.401	5%	0.0421

The condition to reject or accept H_0 is based on the p-value. If the p-value is less than α , it can be determined that H_0 is rejected. Based on Table 5, the amount of p-value is 0.0421, more significant than α . So the test condition is to reject H_0 , or in other words, residual data of the regression model has been stationary.

The second step after examining stationery is determining the order of the time series model for the residual data model through ACF and PACF plots. Based on ACF, the plot is disconnected after the first lag. Therefore, the time series regression model for residual model, which is potential is ARMA (1, 0), ARMA (0,1), and ARMA (1,1). The results of parameter significance, diagnostic test, and value of AIC for these three models are:

Table 6. Results of modeling test and value of AIC

Model	Significance parameter test	White noise	Homoscedasticity	Normality	AIC
ARMA (1,0)	All parameter is significant	✓	✓	✓	1530.21
ARMA (0,1)	All parameter is significant	✓	✓	✓	1584.58
ARMA (1,1)	Parameter θ_1 is not significant		Not being measured		Not being measured

On the test of parameter significance, one of the model ARMA (1,1) is not significant, so other testing are not in need to be examined. For other two models which are significant, result of diagnostic test, involving tests of white noise, homoscedasticity, and normality give the end result that those models are passed the test. To choose the best model for prediction, it will be determined by the value of AIC. Based on AIC value, model ARMA (1,0) is chosen as the best model because it has the smallest value of AIC. The realization of ARMA (1,0) for residual modeling based on regression model is stated as below:

$$\begin{aligned} e_t &= \phi_1 e_{t-1} + a_t \\ e_t &= 0.851 e_{t-1} + a_t \end{aligned} \quad (12)$$

with, e_t and e_{t-1} is error regression model on the formula (13) for period t and $t-1$. a_t is error of time series model on ARMA (1,0).

By combining the formula (11) and (12), we can obtain final model of time series regression to predict number of electrical power consumption (Y) based on numbers of customer (X) on PT. PLN of Surabaya City as follows:

$$\hat{Y}_t = -49,730,693 + 3.143X_t + 0.851e_{t-1} + a_t$$

Furthermore, after obtaining the best time series regression model, we can use the model to estimate electricity consumption in the following periods. The following are the prediction results obtained for the period January 2021 – August 2021:

Table 7. Power Electricity Usage (GWh) predictions:

Period	Electrical Power Consumption (GWh)
January 2021	178,335,533
February 2021	179,333,533
March 2021	180,189,733
April 2021	180,741,683
May 2021	181,785,783
June 2021	182,918,683
July 2021	183,472,133
August 2021	184,177,183
September 2021	184,868,383

The calculation of the accuracy of the electrical power consumption prediction based on PLN customers using a time series regression model is carried out using the MAPE method. The MAPE value for MAPE calculation results is 9.77% which indicates that the prediction accuracy is very good.

The results of predictions that are very accurate can justify that the time series regression model is one of the right models to predict the amount of electric power consumption based on changes in the number of customers, especially in the Surabaya area. We highly recommend this model for use by stakeholders related to the management of energy resources as well as for researchers who will study predictions of electricity consumption, because this model is relatively simple and can provide accurate prediction results.

4. Conclusion

Based on the analysis that has been carried out in section 3, the best time series regression models that can be used to predict electric power consumption are:

$$\hat{Y}_t = -49,730,693 + 3.143X_t + e_t$$

where,

$$e_t = 0.851e_{t-1} + a_t$$

with, e_t and e_{t-1} is error from regression model on period t and $t-1$, a_t is error of time series model on ARMA (1,0). Therefore, using these two equations, final model is created as presented below:

$$Y_t = -49,730,693 + 3.143X_t + 0.851e_{t-1} + a_t.$$

The MAPE value for the final model is 9.77% or it can be interpreted that the prediction accuracy is highly accurate.

From the results obtained from the previous chapter, the best prediction model based on the AIC value, namely the time series regression model with an error value modeled by the AR(1) model, while the measure of forecasting accuracy uses a MAPE value of 9.77%. This implies that predictions of electricity consumption in the Surabaya can be predicted accurately using this model. This research can be a reference for stakeholders to prepare the right strategy in the future related to the amount of electricity consumption, so that it will reduce the possibility of a shortage of electricity supply. The suggestion for future research is that if the available data does not meet the assumptions of the time series regression model, then the analysis can be carried out using alternative models such as GARCH or asymmetric time series models.

Acknowledgements

This work is supported by Department of Data Science, Faculty of Computer Science, University of Pembangunan Nasional Veteran Jawa Timur Indonesia.

References:

- [1]. Andrade, M. G., Achcar, J. A., Conceição, K. S., & Ravishanker, N. (2021). Time Series Regression Models for COVID-19 Deaths. *Journal of Data Science*, 19(2), 269–292. Doi: 10.6339/21-JDS991
- [2]. Arumsari, M., & Dani, A. (2021). Peramalan Data Runtun Waktu menggunakan Model Hybrid Time Series Regression – Autoregressive Integrated Moving Average. *Jurnal Siger Matematika*, 2(1). Doi: 10.23960/jsm.v2i1.2736
- [3]. BPS Indonesia. (2021). *2021 Indonesia Statistics*. BPS Indonesia Publisher.
- [4]. Fauziyah, E., Ispriyanti, D., & Tarno, T. (2021). Pemodelan Dan Peramalan Indeks Harga Saham Gabungan (Ihsg) Menggunakan Arimax-Tarch. *Jurnal Gaussian*, 10(4), 595–604.
- [5]. Jain, G., & Mallick, B. (2017). A Study of Time Series Models ARIMA and ETS. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2898968>
- [6]. Johan, S., & Ginting, A.M. (2022). Determinasi Konsumsi Listrik Di Indonesia. *Media Ekonomi*, 30(1), 106–117. Doi: 10.25105/me.v30i1.10662
- [7]. Kumari, K., & Yadav, S. (2018). Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences*, 4(1), 33. Doi: 10.4103/jpcs.jpcs_8_18
- [8]. Lee, Y. W., Tay, K. G., & Choy, Y. Y. (2018). Forecasting Electricity Consumption Using Time Series Model. *International Journal of Engineering & Technology*, 7, 218. Doi: 10.14419/ijet.v7i4.30.22124
- [9]. Li, Y., Dai, S., & Niu, D. (2017). Electricity consumption forecast of hunan province using combined model based on multivariate linear regression and BP neural network. In *2017 7th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2017)*, 651–655. Atlantis Press.
- [10]. Luo, X. (2017). A Comparison of Three Estimation Methods In Linear Regression Analysis. In *2016 4th International Conference on Machinery, Materials and Information Technology Applications*, 498–502. Atlantis Press.
- [11]. Maruddani, D. A. I., & Abdurakhman. (2021). Delta-Normal Value at Risk Using Exponential Duration with Convexity for Measuring Government Bond Risk. *DLSU Business & Economics Review*, 31(1), 72–80.
- [12]. Pavlič, K., & Parlov, J. (2019). Cross-Correlation and Cross-Spectral Analysis of the Hydrographs in the Northern Part of the Dinaric Karst of Croatia. *Geosciences*, 9(2), 86. Doi: 10.3390/geosciences9020086
- [13]. Rahardjo, S. S., & Sanusi, R. (2019). Linear Regression Analysis on the Determinants of Hypertension Prevention Behavior. *Journal of Health Promotion and Behavior*, 4(1), 22–31. Doi: 10.26911/thejhp.2019.04.01.03

- [14]. Razavi, S., & Vogel, R. (2018). Prewhitening of hydroclimatic time series? Implications for inferred change and variability across time scales. *Journal of Hydrology*, 557, 109–115.
Doi: 10.1016/j.jhydrol.2017.11.053
- [15]. Reddy, J. R., Ganesh, T., Venkateswaran, M., & Reddy, P. (2017). Forecasting of monthly mean rainfall in Coastal Andhra. *International Journal of Statistics and Applications*, 7(4), 197-204.
Doi: 10.7324/IJCRR.2017.9244
- [16]. Sanusi, W. (2017). Trend analysis of rainfall frequency of Makassar. *Scientific Pinisi Journal*, 3(1), 10–16.
- [17]. Sari, V. K. S. (2023). Does electricity consumption influence economic growth in Indonesia? *Jurnal Ekonomi Dan Pembangunan*, 30(1), 47–55.
Doi: 10.14203/JEP.30.1.2022.47-55
- [18]. Setiyowati, E., Rusgiono, A., & Tarno, T. (2018). Model Kombinasi Arima Dalam Peramalan Harga Minyak Mentah Dunia. *Jurnal Gaussian*, 7(1), 54–63.
- [19]. Szustak, G., Dąbrowski, P., Gradoń, W., & Szewczyk, Ł. (2021). The Relationship between Energy Production and GDP: Evidence from Selected European Economies. *Energies*, 15(1), 50.
Doi: 10.3390/en15010050
- [20]. Tan, Y.-F., Ong, L.-Y., Leow, M.-C., & Goh, Y.-X. (2021). Exploring Time-Series Forecasting Models for Dynamic Pricing in Digital Signage Advertising. *Future Internet*, 13(10), 241.
Doi: 10.3390/fi13100241
- [21]. Tarno, T., Maruddani, D. A. I., Rahmawati, R., Hoyyi, A., Trimono, & Munawar. (2020). ARIMA-GARCH Model and ARIMA-GARCH Ensemble for Value-at-Risk Prediction on Stocks Portfolio. *Preprints*, 1–20.
- [22]. Tarno, T., Rusgiono, A., Warsito, B., Sudarno, S., & Ispriyanti, D. (2018). Pemodelan Hybrid Arima-Anfis Untuk Data Produksi Tanaman Hortikultura Di Jawa Tengah. *MEDIA STATISTIKA*, 11(1), 65–78.
Doi: 10.14710/medstat.11.1.65-78
- [23]. Tarno, T., Trimono, T., Maruddani, D. A. I., Wilandari, Y., & Utami, R. S. (2022). Risk Assessment Of Stocks Portfolio Through Ensemble Arma-Garch And Value At Risk (Case Study: Indf.Jk And Icbp.Jk Stock Price). *MEDIA STATISTIKA*, 14(2), 125–136. Doi: 10.14710/medstat.14.2.125-136
- [24]. Udo Moffat, I. & Akapan, E.A. (2017). Identification and Modeling of Outliers in a Discrete - Time Stochastic Series. *American Journal of Theoretical and Applied Statistics*, 6(4), 191.
Doi: 10.11648/j.ajtas.20170604.14