Model for Predicting Electrical Energy Consumption Using ARIMA Method

Muhammad Ridwan Fathin[1] Yudi Widhiyasana[1] Nurjannah Syakrani[1,*]

[1]*Department of Computer and Informatics Engineering, Politeknik Negeri Bandung, Bandung, Indonesia*
**Corresponding author. Email: nurjannahsy@jtk.polban.ac.id*

ABSTRACT
The growth of the human population and technology has led to a rapid increase in electrical energy consumption. Excess electrical energy would be detrimental to the provider, whereas providing less would be detrimental to the consumers. One method for reducing these losses is to forecast the amount of electrical energy that must be available to meet demand. Prediction results can help with three types of decisions, depending on the prediction period: operational decisions (short-term), tactical decisions (medium-term), and strategic decisions (long-term). Short-term forecasts are less relevant given the urgency of the situation. This study aims to help electricity providers to make decisions by making medium and long-term predictions using the Auto-Regressive Integrated Moving Average (ARIMA) method. In the best order determination experiment, ARIMA (8,2,0) was found to be the best model with the smallest error. ARIMA (8,2,0) has an average percentage error of 5.3 percent based on the overall prediction results. There is no linearity between accuracy and prediction period in the prediction period experiment. According to the experimental results, the highest accuracy is obtained in the medium term (monthly) with a value of RMSE 753,983.98. As a result, based on the time period, ARIMA is the best for tactical decisions (medium-term) regarding electrical energy consumption.

Keywords: *Prediction, Electrical Energy Consumption, ARIMA, Prediction Period.*

1. INTRODUCTION

The growth of the human population and the development and application of technology has resulted in a rapid increase in electrical energy consumption. As a result, predicting electrical energy consumption is required when making electrical energy management decisions [1]. Predicting electricity consumption is critical for developing-country governments especially to improve energy efficiency. The 2017-2026 Power Generation Business Plan in Indonesia has been revised (RUPTL). This revision is necessary because electricity consumption in 2017 was lower than expected in the previous year, raising concerns about an oversupply [2]. Oversupply of electrical energy will result in overcapacity, which means wasted resources, whereas a scarcity of energy will result in higher operating costs for additional energy suppliers. As a result, modeling the prediction of electricity consumption with high accuracy becomes critical for minimizing losses [3].

The two tactical decisions (medium-term, i.e. weeks to months) are, for example, the construction of a transmission network in Indonesia, where the electricity provider can decide how to distribute electricity from backup power plants to meet electrical energy needs. The three operational decisions (short-term, i.e. hours to days) include, for example, providing other supporting facilities such as determining the number of staff to anticipate problems that typically arise when there is an increase in electricity consumption [4].

There are several methods for modeling predictions, one of which is the Auto-Regressive Integrated Moving Average (ARIMA) method. Y. Lu [5], V. Yakovyna and O. Bachkai [6,] and K. Sakulkitbanjong and C. Pongchavalit [7] compared ARIMA to other time series models for predicting renewable energy in America, Angular software failure, and fire risk, respectively. The three results show that ARIMA outperforms other time series methods, such as the exponential smoothing-additive model, the exponential smoothing-multiplicative model, and the seasonal dummy with GARCH methods, compared to the scientific work.

The accuracy of the prediction model is calculated by comparing the prediction results to the actual value. One method is to compute the Root Mean Squared Error (RMSE). The lower the RMSE value (close to zero), the more accurate the prediction model's predictions, and vice versa [2].

This study aims to help electrical energy providers make decisions by predicting electrical energy consumption in the medium and long term.

1.1. Auto-regressive Integrated Moving Average

The Auto-Regressive Integrated Moving Average (ARIMA) model is a combination of the Auto-Regressive (AR) and Moving Average (MA) models, with data that has been differentiated n times. This model, also known as the Box-Jenkins model, was developed by George Box and Gwilyn Jenkinson and is used to analyze and predict time series data. For time-series data, the ARIMA model is self-regresses with error correction via the moving average method. Regression is performed in the AR model between a variable at a specific time and the variable itself in the past (lag). The MA model is one of several analytical methods for determining the moving average of a variable over a given time period [6].

ARIMA(p,d,q) is a model in which p is the AR order value, d is the number of data differencing processes until the data reaches the stationary condition, and q is the MA order. In general, the AR model is as shown in equation (1) below.

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + e_t \tag{1}$$

With description:

- y_t : variable value at time t
- α_1 : AR coefficient; i : 1,2, ..., p;
- e_t : error value at time t
- p : order AR

In equation (1), the AR model depends on the value of previous observations.

The MA model, in general, is as in equation (2) below.

$$y_t = e_t - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} \tag{2}$$

With description:

- y_t : variable value at time t
- θ_i : coefficient of moving average; i: 1,2, ..., q
- e_t : error value at time t
- q : order MA

The MA model is influenced by the current error value and the error value with a certain weight in the past, as shown in equation (2).

We get the ARIMA model in general from the AR model in equation (1) and the MA model in equation (2), with parameter descriptions in equations (2) and (3) as before.

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + e_t - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-} \tag{3}$$

The description is the same as the equation (1) and (2). In equations (3), the current value depends on or is influenced by several previous values, including the current error value and several previous error values. The ARIMA method has the following stages: stationary test, AR and MA value determination, and the best model [9].

1.1.1. Stationary Test

One of the requirements for using the ARIMA method is that the time series data be in a stationary state. The term "stationary" refers to the absence of a trend of data growth or decline. If the data is not in a stationary state, the differencing process can be used.

There are several methods for determining whether a time series data set is stationary, including summary statistics and statistical tests. Summary statistics, such as the average and variance of time series data, are used to determine whether or not there is a significant change in the average and variance values within a given time range. Other methods include statistical tests, which are used to determine stationary conditions in time series data. The Augmented Dickey-Fuller Test is a statistical test frequently used to determine stationary conditions (ADF test).

The ADF test is a statistical test that is also known as a unit root test. The unit root test is used to determine how strongly a time series data contains an element of trend. The p-value indicates the outcome of the ADF test. The p-value represents the probability that the data series is not stationary. If the p-value is 0.9622, the data series is 96.22 percent nonstationary. Stationary data series can also be identified by comparing the test-statistic value to the critical value (1 percent). If the test-statistic value exceeds the critical value (1%), the data series is not stationary, and vice versa.

If the data is not in a stationary state, the differencing process can be used to convert the time series data to a stationary state. The process of differencing is carried out in the same manner as in equation (4).

$$\nabla \quad {}^{d}y_t = y_t - y_{t-1} \tag{4}$$

With description:

- y_t : current value
- y_{t-i} : previous value
- $\nabla \ {}^{d}y_t$: differencing result
- d : value of order differencing

As in the equation, each value in the time series data is carried out by a differencing process (4). The differencing process can be repeated multiple times until the time series data becomes stationary. The order differencing value is represented by the number of differencing processes performed. The variable is commonly used to indicate order differencing [9].

1.1.2. Determination of AR and MA Value

Once the time series data has reached a stationary state, the initial estimated values of AR and MA can be determined. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) values can be used to calculate AR and MA values. ACF and PACF graphs can be created visually to help determine the AR and MA values. The ACF and PACF charts are depicted in Figure 1. In Figure 1, the vertical line of the ACF and PACF graphs represents the correlation coefficient value, which is known as lag. The horizontal line in Figure 1 represents a significant limit indicating whether or not the autocorrelation coefficient or lag is significant. The significant limit value is obtained from equation (7).

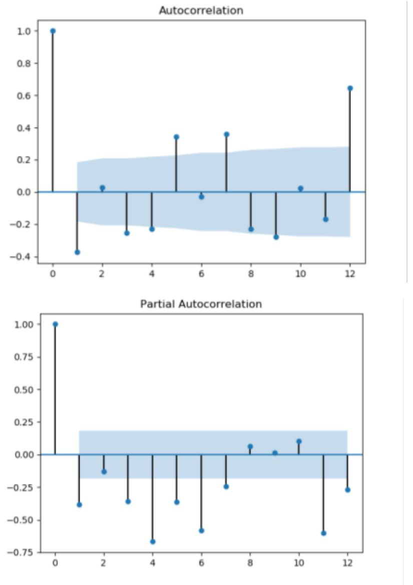

Figure 1 Example of ACF and PACF graph plots The Y-axis of the ACF and PACF graphs in Figure 1 represents the correlation coefficient value, known as lag. The ACF and PACF lag values are calculated using equations (5) and (6). The significant limit, represented by the X-axis, indicates whether or not there is an autocorrelation coefficient or a significant lag. The blue background indicates the significant limit value of the graph, and it is derived from equation (8).

The ACF formula is:

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \qquad (5)$$

With description:

- rk : autocorrelation coefficient on lag-k
- k : time difference
- n : number of observations
- $\bar{x}$: average observation
- x_t : observation at time t
- x_{t+k} : observations at time t+k, k = 1, 2, 3, …

The PACF formula is:

$$\pi_k = \begin{cases} 1, k = 0 \\ r_1, k = 1 \\ \dfrac{r_k - \sum_{t=1}^{k-1} \pi_{k-1,t} \times r_{k-t}}{1 - \sum_{t=1}^{k-1} \pi_{k-1,t} \times r_{k-t}} \end{cases} \qquad (6)$$

With description:

- π_k : partial autocorrelation coefficient on lag-k
- r_k : autocorrelation coefficient on lag-k

The value $\pi_{k,t}$ is obtained through equation (7).

$$\pi_{k,t} = \pi_{k-1,t} - \pi_k \pi_{k-1,k-t} \qquad (7)$$

With description:

- $\pi_{k,k} = \pi_k$

The significant limit formula is:

$$v = \pm \frac{1.96}{\sqrt{N}} \qquad (8)$$

With description:

- v : critical value or significant value
- N: the amount of observation data used

The AR and MA values are determined by the number of lags that are outside the significant limit. For example, in Figure 1 of the ACF graph, there are three significant lags (outside the significant limit), so the initial MA value is three, whereas the initial AR value from the PACF graph is one. The amount of lag that is significant for the AR order value is typically used for variables and to indicate the MA order [9].

1.1.3. Determination of the Best Model

After determining the initial estimated values, these values can be used to generate several possible ARIMA model order estimates. As a result, several ARIMA model order combinations can be created, and each order model can be compared to determine the best ARIMA model order [9]. The best order model is one that produces a model with the smallest error in its prediction results. The Root Mean Squared Error (RMSE) value can be used to evaluate the error. The equation shows the formula for calculating the RMSE value (9).

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N} \quad (y_t - \quad)^2}{N}} \qquad \widehat{y}_t \tag{9}$$

With description:

- y_t : actual time series value at time
- $\widehat{y}_t$: the predicted value at time
- N : amount of data used for prediction evaluation

The RMSE value indicates the model's accuracy in predicting the actual value. The lower the RMSE value (close to zero)is, the more accurate the prediction model's predictions, and vice versa [9].

2. DATASET

The hourly electricity consumption data from Kaggle.com [8] is aggregated into monthly form, yielding 164 data points. The convenience sampling method [10] divides the dataset into training and test data based on time intervals. The training data to test data ratio is seven to three. The training data collected ranges from October 2004 to May 2014, with 116 data points, while the test data collected ranges from May 2014 to July 2018, with 48 data points.

3. EXPERIMENTS AND RESULTS

This research employs a quantitative approach as well as experimental research techniques. This study includes two experiments. The first experiment is the best order model experiment, which is used to generate the best order model. The second experiment is a time-limited experiment that is carried out to generate predicted electrical energy consumption data using the ARIMA model for each prediction period (one month, one semester, one year, two years, and four years).

This study's variables are the independent variable and the dependent variable. The RMSE value represents the model's accuracy. During the execution of this experiment, the relationship between the variables is shown in Figure 2.

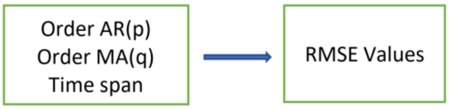

Figure 2 Independent and Dependent Variable

Order AR(p) is the sum of past values (lag) that significantly influence future values. The AR order value (p) ranges from 0 to the initial estimated value shown in the Partial Auto-correlation Function (PACF) graph. The graph makes it easier to see the relationship between future and past values that are outside of the significant limit determined by equation (6). Assumed that eight lags cross the significant limit, resulting in an initial estimated value of the AR order (p) of eight. Experiments were conducted to test all possible AR orders (p) ranging from zero to the maximum value (the initial estimated value obtained from the PACF graph) in order to obtain an AR order value (p).

Order MA(q) is the number of error values in past values (lag) that significantly impact future values. The error value is calculated by subtracting the current value from the moving average value of q. The obtained error value is expected to be an error correction of the AR prediction results to increase the accuracy of the ARIMA method prediction results. MA (q) has an order value ranging from zero to the initial estimated value shown in the Auto-correlation Function (ACF) graph. The graph makes it easier to see the relationship between future and past values that are outside of the significant limit determined by equation (5). Assumed eight lags cross the significant limit, resulting in an initial estimated value of the MA (q) order of eight. Experiments were carried out to try all possible AR orders (p) ranging from zero to the maximum value (the initial estimated value obtained from the PACF graph) in order to obtain an AR order value (p) that gave prediction results with the highest accuracy value.

The prediction period influences the prediction model's accuracy. In this study, the prediction period (**time span**) is one month, one semester, one year, two years, and four years. This value influences the amount of model correction. The model is corrected once a month for a total of one thousand and two hundred training data points, so the model knowledge correction process is repeated one thousand and two hundred times. The knowledge correction process is repeated 200 times over the course of the semester. The knowledge correction process is repeated a hundred times over the course of a year. It is hoped that this experiment can yield information about the effect of the amount of knowledge correction process (prediction period) on the prediction model's accuracy.

The RMSE values represent the prediction results' accuracy. The RMSE value is calculated using equation (9).

3.1. Best Order Determination Result

The RMSE calculation results from each combination of order models according to the scenario specified in the experimental design are included in the best order determination experiment.

The data used in this article are not in a stationary state. The differencing process can be used to convert the time series data to a stationary state. The process of differencing is carried out in the same manner as in equation (4). We get d = 2 because we need two different processes. The experimental results using monthly training data are shown in Table 1.

Table 1. RMSE for each order of ARIMA

No	Model	RMSE
1	ARIMA (0,2,1)	1213038
2	ARIMA (0,2,2)	1278000
3	ARIMA (0,2,3)	1227733
...	...	...
69	ARIMA (8,2,5)	1041266
70	ARIMA (8,2,6)	939378.5
71	ARIMA (8,2,7)	868519.6

Table 2 shows the three best order models with the smallest RMSE value from all the ARIMA model combinations in Table 1.

Table 2. The three best model orders based on RMSE

No	Model	RMSE
1	ARIMA (8,2,0)	747794.1
2	ARIMA (8,2,1)	767223.7
3	ARIMA (7,2,0)	771630.4

The three order models in Table 2 are the results of an experiment to determine the best order model for the experiment to use for the longest period of time.

3.2. Maximum Duration of Experiment Results

The RMSE value represents the evaluation results of the prediction of electrical energy consumption in the best order determination experiment. Predictions are made based on the prediction period: one month, one semester, one year, two years, and four years. The ARIMA model is used to predict each prediction period based on the results of determining the best order model, namely ARIMA (8,2,0), ARIMA (8,2,1), and ARIMA (7,2,0).

Experiments predicting the next month yielded the lowest RMSE in the ARIMA model (8,2.0). The RMSE for the ARIMA model (8,2,0) is 753.983.98. The graph plot of the prediction results from the ARIMA model that has been made for the next month can be seen in Figure 3.

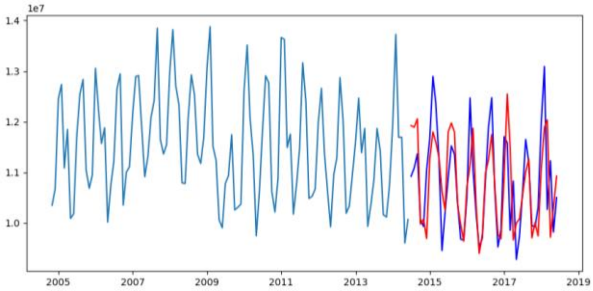

Figure 3 Graph of predictions for the next month

The light blue color represents training data, the blue color represents test data, and the red color represents predicted data. The Y-Axis represents the total amount of energy consumed, while the X-Axis represents the time of year.

According to the graph plot in Figure 3, the graph plot of the predicted results can follow the pattern of the graph plot of the test data.

In comparison, the ARIMA model (8,2,1) is used to forecast the next four years, yielding the highest RMSE value of 830993.63546. Figure 4 depicts a graph plot of the prediction results from the ARIMA model for the next month.

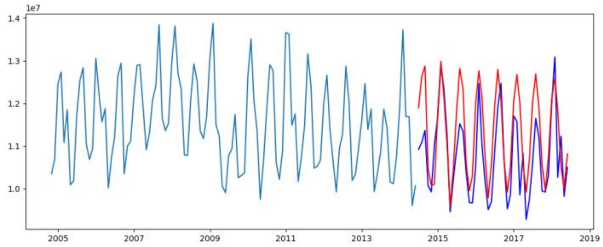

Figure 4 Graph of predictions for the next 4 years

The graph plot in Figure 4 shows that there is a significant difference between the prediction results and the actual results. The graph plot pattern forms a seasonal pattern with the same length and height. When compared to other prediction periods, this produces the highest RMSE value.

Table 4 shows the experimental results for the maximum period. Table 4 uses the term exp code to refer to the experiment code. In the ARIMA model, it can be seen that a period of one month produces the smallest RMSE value, with an RMSE value of 753.983.98. (8.2.0). Table 5 shows a comparison of the prediction results for the next month with the actual data.

Table 4. Prediction evaluation results based on time period

Exp code	Time Period	ARIMA Model	RMSE
E01	1 month	ARIMA (8,2,0)	753,983.98
E02		ARIMA (8,2,1)	777,726.365
E03		ARIMA (7,2,0)	777,003.972
E04	6 month	ARIMA (8,2,0)	789,233.032
E05		ARIMA (8,2,1)	807,056.316
E06		ARIMA (7,2,0)	831,978.453
E07	1 year	ARIMA (8,2,0)	769,902.043
E08		ARIMA (8,2,1)	789,509.259
E09		ARIMA (7,2,0)	820,745.287
E10	2 years	ARIMA (8,2,0)	755,061.562
E11		ARIMA (8,2,1)	763,411.132
E12		ARIMA (7,2,0)	833,791.775
E13	4 years	ARIMA (8,2,0)	835,320.898
E14		ARIMA (8,2,1)	830,993.635
E15		ARIMA (7,2,0)	951,646.05

Table 5. Comparison of actual data and predicted data

Actual Data(MWh)	Predicted Data(MWh)	Error (%)
10,921,235.00	11,923,331.00	9.18
11,086,809.00	11,891,391.00	7.26
11,367,056.00	12,058,903.00	6.09
10,069,273.00	9,977,628.00	0.91
...		
10,265,918.00	12,034,518.00	17.23
11,228,646.00	9,718,775.00	13.45
9,820,256.00	10,292,981.00	4.81
10,503,052.00	10,926,426.00	4.03

From Table 5, it can be seen the comparison of the actual data and the predicted results, with an error of up to millions of MWh and an error of up to 10%. Based on the calculation of the overall prediction results, ARIMA (8.2.0) has an average error percentage of 5.3% or, in other words, has an accuracy of 94.7%. In Indonesia, in 2019, all generating systems have a reserve margin of at least 30% [2].

4. CONCLUSION

Hourly Energy Consumption data from Kaggle.com is accumulated per month, per semester, per year, per 2 years, per 4 years used for predictions with the medium and long term combined for experiments using the ARIMA model. From the research results on the maximum prediction period, it can be concluded that the ARIMA method is the maximum to support tactical decisions (medium-term) that is monthly with an RMSE value of 753.983.98 or in percentage accuracy reaching 94.7%. These results indicate that ARIMA is maximum for medium-term predictions, namely predictions for the next one month.

REFERENCES

[1] K. A. Al-zahra, K. Moosa and B. H. Jasim, “A comparative Study of Forecasting the Electrical Demand in Basra City using Box-Jenkins and Modern Intelligent Techniques,” Iraq J. Electrical and Electronic Engineering, vol. 11, 2015.

[2] PT. PLN, "Rencana Usaha Penyediaan Tenaga Listrik (RUTPL) 2016-2025," Direktorat Perencanaan Korpora, Jakarta, 2016.

[3] F. Kaytez, M. C. Taplamacioglu, E. Cam and F. Hardalac, “Forecasting Electricity Consumption: A Comparison of Regression Analysis, Neural Networks And Least Squares Support Vector Machines,” Electrical Power and Energy Systems, vol. 67, pp. 431-438, 2015.

[4] F. S. Purnomo, "Penggunaan Metode ARIMA (Auto-Regressive Integrated Moving Average) untuk Prakiraan Beban Konsumsi Listrik Jangka Pendek (Short Term Forecasting)," 2015.

[5] Y. Lu, “Time Series Forecasts of Renewable Energy Consumption in the United States,” 2018.

[6] V. Yakovyna and O. Bachkai, “The Comparison of Holt-Winters and Box-Jenkins Methods for Software Failures Prediction,” 2017.

[7] K. Sakulkitbanjong and C. Pongchavalit, “Time Series Analysis and Forecasting of Forest Fire Weather,” 2017.

[8] R. Mulla, “Hourly Energy Consumption,” 30 August 2018. [Online]. Available: https://www.kaggle.com/robikscube/hourly-energy-consumption. [Accessed 30 January 2019].

[9] M. I. Fauzan, "Analisis Pemrosesan Paralel dalam Mendukung Layanan Prediksi Cuaca," 2018.

[10] Z. Reitermanov´a, “Data Splitting,” WDS’10 Proceedings of Contributed Papers, vol. Part I, p. 31–36, 2010.