Forecasting Customer Electricity Meter Installation Demand Using Autoregressive Integrated Moving Average (ARIMA)

Case Study : PT PLN (Persero) UP3 Kediri

Risky Dinal Ardianto - 23522305 (*Author*)
Program Studi Magister Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung
E-mail (gmail): 23522305@std.stei.itb.ac.id

***Abstract*— PT PLN Persero is an electricity supply company which is obliged to electrify all corners of Indonesia. Increasing customer electricity connection capacity is one of the challenges to support economic growth, besides that it can also provide the potential to increase PLN's income. Connecting customer electricity is one of the main business processes that is routinely carried out by PLN. The accuracy of forecasting demand for new electricity meter installations can affect the quality and speed of service. Currently, PT PLN (Persero) does not yet have a forecasting model that can scientifically calculate the demand for new customer pairs in order to increase the number of customers. So in this research, a forecasting model was built using the ARIMA, SARIMA methods. Based on the experimental results, it was found that the SARIMA model had the best RMSE value if an experiment was carried out using test data, while the ARIMA model had a better SMAPE value compared to other models.**

Keywords—Time Series; Forecasting; ARIMA; SARIMA; Electricity Connection Demand (key words)

I. INTRODUCTION (*HEADING 1*)

Electricity is a primary need for humans. All activities require electricity, especially for household needs, industry, business and public facilities. PT PLN (Persero) as an electricity provider has an obligation to electrify all Indonesian people. As time goes by, PLN is required to continue developing and growing in order to serve the Indonesian people needs. Increasing customer electricity connection capacity is one of the challenges to support economic growth, besides that it can also provide the potential to increase PLN's income. Electricity connection is the main business process of PLN UP3 Kediri. Customer electricity connection is divided into two categories, namely 1 phase customers and 3 phase customers. Single phase electricity connection is carried out routinely every day depending on the number of applicants who wish to install a new one. Single phase connection power is from 450VA to a maximum power of 5500VA.. Accuracy in planning can influence the level of quality and speed of services provided. By predicting the demand for new electricity connections, PLN can easily carry out material planning and workforce planning to assist the management in creating strategies action in order to improve services quality. PLN needs to be able to plan precisely the number of daily requests for new installations, so that execution planning can be carried out using human resources and supporting materials.

There are several things that can influence the number of new customer installment process, including the availability of materials and several external factors. These external factors include economic and political factors. In 2020, the whole world experienced a pandemic, which had an impact on the number of new pairs. This of course affects the plans that were set the previous year, so management needs to maneuver action so that the main data goals are achieved. Based on these problems, PLN needs tools that can plan and forecast the number of requests for new electricity installations. Accuracy in planning will affect the level of quality and speed of services provided. By predicting the demand for new electricity connections, PLN can easily carry out material planning and workforce planning for work teams so that it can assist management in creating strategies to accelerate connections in order to improve services.

There have been several previous studies that used ARIMA to forecast time series data, such as research conducted by (Wirdyacahya and Prastuti, 2022) used ARIMA to forecast cement demand, where the best model was using the ARIMA model ([1,12],1 ,[1,12]) with an RMSE accuracy value of 52,776 [7] Then (Shukla and Jharkharia, 2013) applied the ARIMA model to predict onion sales in grocery stores where the highest demand pattern occurs on Mondays [5]. (Falatouri et al., 2022) used Seasonal ARIMA / SARIMA to forecast demand for sales of vegetables and fruit in a three year period, namely January 2017 – December 2019. The model was applied to predict demand for certain vegetable products and produced the best SARIMA model for cucumber products [4].

The focus on this research is to create a forecasting model in accordance with the existing problem domain at PLN, namely related to the demand for new installations of single phase customer electricity. To overcome existing problems, PLN can do the same thing as previous research, by applying a demand forecasting model and utilizing historical time series data for the past several years. This research uses historical data taken over a 10 year period, January 2014 - December 2023. PLN has a scientifically tested forecasting model that can predict demand for new single-phase electricity installations efficiently and right on target. ARIMA / Auto Regressive Integrated Moving Average which a statistical mathematical method that looks at historical data over time to plan needs.

II. METHOD

A. Autoregressive Integrated Moving Average (ARIMA)

ARIMA or Autoregressive Integrated Moving Average is a time series model that can be used on data that has seasonal or non-seasonal patterns. The models commonly used in BoxJenkins ARIMA are the AR (autoregressive), MA (moving average), and ARMA (autoregressive-moving average) models [6]. The autoregressive process with order p AR(p) follows the following equation:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + at \quad (1)$$

Meanwhile, the MA (Moving Average) model is used to explain an event where an observation at time t is expressed as a linear combination of a number of residuals[6]. The form of the MA model equation can be written as follows:

$$Z_t = at - \Theta_1 a_{t-1} - \Theta_2 a_{t-2} - \cdots + \Theta_q a_{t-q} \quad (2)$$

Meanwhile, the combined AR and MA (ARMA) model or with ARMA notation (p,q) can be written as follows:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + at - \Theta_1 a_{t-1} - \ldots \Theta_q a_{t-q} \quad (3)$$

The ARIMA model (p,d,q) can be written mathematically using the following equation:

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

B. Seasonal ARIMA (SARIMA)

Electricity connection demand pattern exhibits periodic fluctuations that align with seasonal changes. These fluctuations can be effectively managed using the SARIMA model, which incorporates seasonal terms into the ARIMA model. Generally, the SARIMA model is denoted as SARIMA (p,d,q)(P,D,Q)S [8]. Here, 'p' represents the autoregressive order, while 'P' denotes the seasonal autoregressive order. The differences are indicated by 'd' for regular differences and 'D' for seasonal differences [8]. The terms 'q' and 'Q' represent the moving average (MA) order and the seasonal moving average (SMA) order, respectively. 'S' stands for the length of the seasonal cycle . The seasonal component of the model is illustrated by the seasonal lag in the autocorrelation function plot. The SARIMA model can be described as follows:

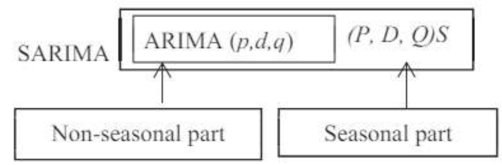

Fig. 1 SARIMA

III. TECHNIQUES ADOPTED

Four important steps are necessary to identify the correct forecasting model: stationarity testing, model identification, model fitting, and finally, performance evaluation. In this paper, the ARIMA and SARIMA models are considered, and their performances are compared. Historical data taken from the AP2T system is data on applications for new installations from 2014 to 2023 originating from all ULPs in UP3 Kediri. UP3 Kediri has 11 unit areas. Where during this period visualization is carried out first.

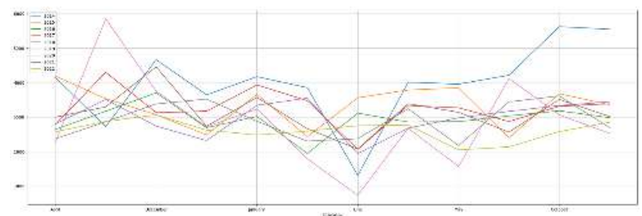

Fig. 2 : Distribution of the movement of new installation applications as of 2020.

If we look at Fig. 2, the trend in 2020, there was a sharp decline in April 2020, which coincided with the COVID-19 pandemic during that period. However, there was an increase in August 2020, marking the highest trend in nine years. In Figure 7, we can also see the trend from 2014 to 2022, where the data shows a seasonal pattern due to continuous fluctuations over certain periods.

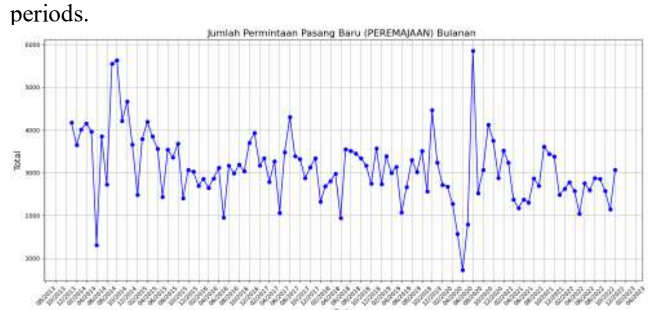

Fig. 3 : visualization of time series data plots for new installation applications

A. Stationarity Test

Before modeling, it is necessary to check the stationarity of the data. To see whether stationarity can be identified in data, you can use the unit root test or Augmented Dickey Fuller (ADF).

To see whether the data is stationary or not, use the following formula:

$$\tau = |\hat{\gamma} / (SE(\hat{\gamma}))|$$

$$H0 : \gamma = 0 \text{ (not stationer)}$$

$$H1 : \gamma \neq 0 \text{ (Stationer)}$$

Time series data needs to be decomposed to determine trends, seasonality and residual as can be seen in figure 3. Based on this decomposition, the data shows an up and down trend but tends to continue to decline from 2014 to 2022.

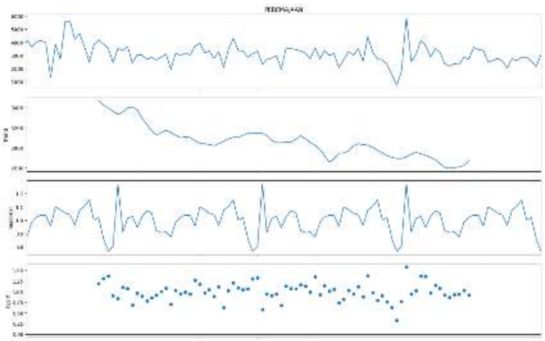

Fig. 3 : Decomposation Tren, Seasonal, Residual

The results of the augmented Dickey-Fuller (ADF) stationarity test provide ADF statistics of -7.745774 with a p-value of 0.000000. In general, the ADF test results show that the data tested is stationary, which means the data does not have a unit root and has a constant mean, variance and covariance over time. Because the data is stationary, there is no need for differentiation of the data.

B. *Model Identification*

The first step is to identify possible patterns in the data in order to establish an appropriate ARIMA model. The ARIMA model can be determined by analyzing ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) graphs from time series. The autocorrelation function (ACF) is a linear relationship between observations Yt and observations $Yt{-}k$ The partial autocorrelation function (PACF) is useful for measuring the relationship between values of the same variable, assuming that the influence of all other time lags is fixed or constant.

The next step is to identify the model using ACF and PACF with a critical limit of 0.25 which can be seen from the stationary ACF and PACF diagrams. The results of ACF and PACF identification can be seen in Figure 9 and Figure 10.

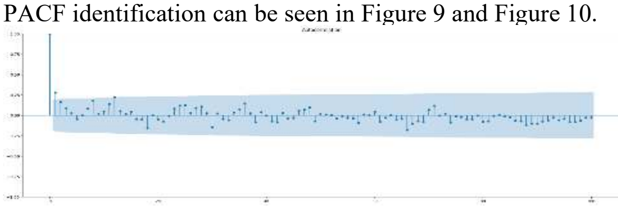

Fig. 4 : Autocorrelation (ACF)

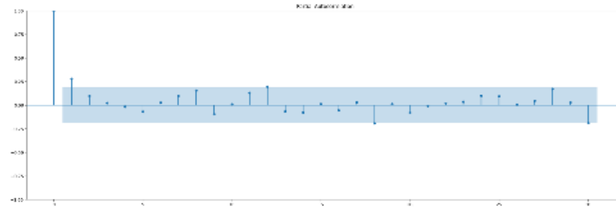

Fig. 5 : Partial Autocorrelation (PACF)

Based on the ACF plot, it shows 2 significant autocorrelations, while the PACF plot shows 2 significant partial autocorrelations. So the Autoregression model AR(1) and Moving Average MA(1) are set.

C. *Model Fitting*

Determining the forecasting model was carried out using total autocorrelation with lag 1 and partial autocorrelation 1 to produce the ARIMA(1,1,1) model and the SARIMA(1,1,1) model. The model was trained with the historical data in Python programming language and several statistical libraries to perform ARIMA and SARIMA modeling as we can see in Fig 6 and fig. 7.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:             PEREMAJAAN   No. Observations:                  108
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -862.430
Date:                Mon, 03 Jun 2024   AIC                           1730.860
Time:                        11:41:33   BIC                           1738.879
Sample:                    01-01-2014   HQIC                          1734.111
                         - 12-01-2022
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2046      0.093      2.205      0.027       0.023       0.387
ma.L1         -0.9450      0.041    -23.309      0.000      -1.024      -0.866
sigma2      5.747e+05   5.92e+04      9.711      0.000    4.59e+05    6.91e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.11   Jarque-Bera (JB):                79.50
Prob(Q):                              0.74   Prob(JB):                         0.00
Heteroskedasticity (H):               0.98   Skew:                             0.80
Prob(H) (two-sided):                  0.94   Kurtosis:                         6.91
===================================================================================
```

Fig. 6 : Model ARIMA(1,1,1)

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                         PEREMAJAAN   No. Observations:                  108
Model:             SARIMAX(1, 1, 1)x(1, 1, 1, 12)   Log Likelihood                -763.487
Date:                            Mon, 03 Jun 2024   AIC                           1536.974
Time:                                    11:43:05   BIC                           1549.743
Sample:                                01-01-2014   HQIC                          1542.134
                                     - 12-01-2022
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1560      0.092      1.702      0.089      -0.024       0.336
ma.L1         -0.9999      0.164     -6.079      0.000      -1.322      -0.677
ar.S.L12       0.0599      0.184      0.326      0.745      -0.300       0.420
ma.S.L12      -0.8871      0.337     -2.631      0.009      -1.548      -0.226
sigma2      4.671e+05   3.51e-07   1.33e+12      0.000    4.67e+05    4.67e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.20   Jarque-Bera (JB):                42.75
Prob(Q):                              0.65   Prob(JB):                         0.00
Heteroskedasticity (H):               1.13   Skew:                             0.81
Prob(H) (two-sided):                  0.74   Kurtosis:                         5.85
===================================================================================
```

Fig 7 : Model SARIMA (1,1,1)(1,1,1)

D. *SARIMA Parameter Significance Test*

To ensure that the selected SARIMA model parameters are appropriate, significant tests were carried out with several parameter scenarios. In table 1, the test scenarios for significant parameters.

Model	Parameter	Koefisien	2,5%	97,5%	p-Value	Signifikan
SARIMA (1,1,1)	AR(1)	0.1560	-0.024	0.336	0.089	Tidak Signifikan
	MA(1)	- 0.9999	-1.322	-0.677	0.000	Signifikan
	AR.S (12)	0.0599	-0.300	0.420	0.745	Tidak Signifikan
	MA.S(12)	-0.8871	-1.548	-0.226	0.009	Signifikan
SARIMA (2,1,1)	AR(1)	0.1906	-0.024	0.008	0.041	Signifikan
	MA(1)	- 0.9998	-1.258	-0.742	0.000	Signifikan
	AR.S (12)	-0.2021	-0.641	0.237	0.367	Tidak Signifikan
	AR.S (24)	-0.2894	-0.620	0.041	0.086	Tidak Signifikan
	MA.S(12)	-0.5744	-1.005	-0.144	0.009	Signifikan
SARIMA (1,1,2)	AR(1)	0.1834	-0.005	0.372	0.056	Signifikan
	MA(1)	-0.9993	-9.544	7.545	0.819	Signifikan
	AR.S (12)	-0.8417	-1.081	-0.603	0.000	Tidak Signifikan
	MA.S(12)	0.1266	-8.474	8.728	0.977	Tidak Signifikan
	MA.S(24)	-0.8730	-8.254	6.508	0.817	Signifikan
SARIMA (2,1,2)	AR(1)	0.2040	0.005	0.403	0.044	Signifikan
	MA(1)	-0.9840	-1.246	-0.722	0.000	Signifikan
	AR.S (12)	0.0877	-1.180	1.356	0.892	Tidak Signifikan
	AR.S (24)	-0.3468	-0.699	0.005	0.053	Signifikan
	MA.S(12)	-0.8744	-2.232	0.484	0.207	Tidak Signifikan

From the test results, there are no SARIMA parameters that overall have a significant value. So in this work the parameter selection remains in SARIMA(1,1,1) based on the parameter selection in ACF and PACF

IV. PERFORMANCE AND EVALUATION

Based on the diagnostic results for the ARIMA(1,1,1) and SARIMA(1,1,1) models, the results are given from the standardized residual plot, residual histogram, Residual ACF which can be seen in figures 8 and 9. The residual plot shows that the model can produce The residual is stationary with a mean of zero, but there are several large peaks that need attention because they can indicate outliers or high variability at some points in time. Meanwhile, the autocorrelation correlogram value is within the significant limits (shown by the blue area), which indicates that there is no significant autocorrelation in the residuals.

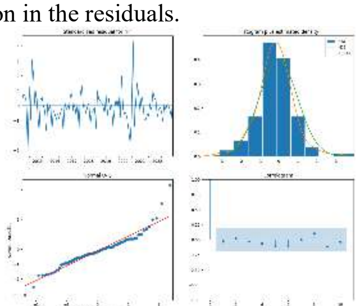

Fig. 8 : ARIMA(1,1,1) Diagnostic Test

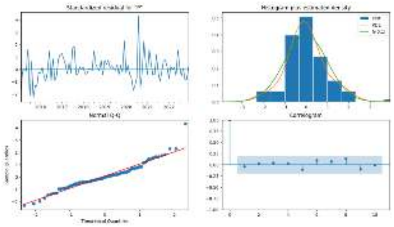

Fig. 9 : SARIMA(1,1,1) Diagnostic Test

Testing was carried out with the models discussed previously, namely the ARIMA(1,1,1) and SARIMA(1,1,1) models which were used to predict demand for new electricity installations by comparing test data in the 2023 period. In Figure 13a you can see the forecasting results using ARIMA(1,1,1) model, while in Figure 13b the SARIMA(1,1,1) model is used.

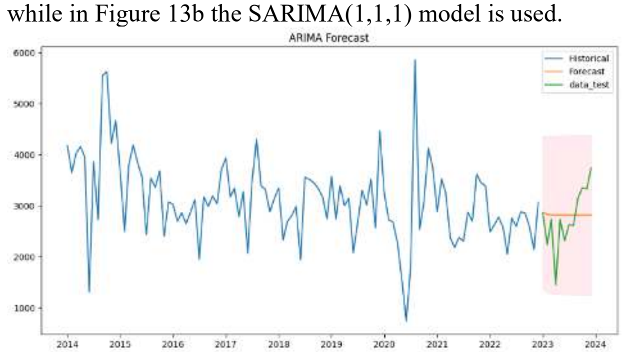

Fig. 10 : Forecasting Results ARIMA (1,1,1) on testing data

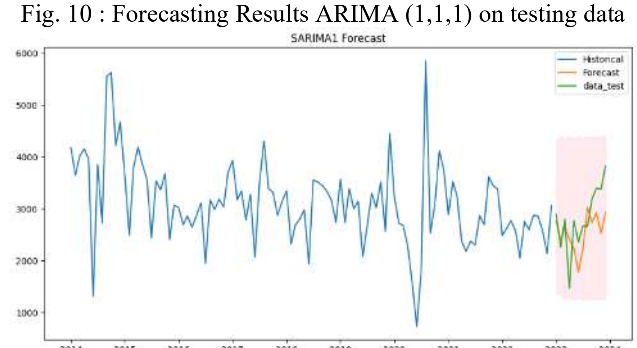

Fig. 10 : Forecasting Results SARIMA (1,1,1) on testing data

The evaluation results show that forecasting on testing data produces better accuracy than train data as indicated by smaller SMAPE, RMSE and MAE values than test data. Meanwhile, the comparison between ARIMA and SARIMA shows that the forecast accuracy value data is better using ARIMA according to the test results in table 2.

Model	Jenis Data	SMAPE	RMSE	MAE
SARIMA (1,1,1)	data train	19.87	901.6	616.5
	data test	18.28	537.14	465
ARIMA (1,1,1)	data train	19.56	799.58	595.34
	data test	16.766	582.33	443.90

Table 2 : Test forecasting accuracy

V. CONCLUSION AND RECOMENDATION

In research conducted for a case study of forecasting customer electricity connection demand, two models (ARIMA and SARIMA) were used separately to test historical data for the past 10 years. The results of testing on test data obtained different results when compared with testing on train data. For both models, testing the test data provides better results in terms of SMAPE, RMSE and MAE indicators. For future improvements, you can use other machine learning methods and consider external aspects as supporting variables

REFERENCES

[1] Aguilar Madrid, E., Antonio, N., 2021. Short-Term Electricity Load Forecasting with Machine Learning. Information 12, 50. https://doi.org/10.3390/info12020050

[2] Chandramitasari, W., Kurniawan, B., Fujimura, S., 2018. Building Deep Neural Network Model for Short Term Electricity Consumption Forecasting, in: 2018 International Symposium on Advanced Intelligent Informatics (SAIN). Presented at the 2018 International Symposium on Advanced Intelligent Informatics (SAIN), pp. 43–48. https://doi.org/10.1109/SAIN.2018.8673340

[3] De Gooijer, J.G., Hyndman, R.J., 2006. 25 years of time series forecasting. International Journal of Forecasting 22, 443–473. https://doi.org/10.1016/j.ijforecast.2006.01.001

[4] Falatouri, T., Darbanian, F., Brandtner, P., Udokwu, C., 2022. Predictive Analytics for Demand Forecasting – A Comparison of SARIMA and LSTM in Retail SCM. Procedia Computer Science, 3rd International Conference on Industry 4.0 and Smart Manufacturing 200, 993–1003. https://doi.org/10.1016/j.procs.2022.01.298

[5] Shukla, M., Jharkharia, S., 2013. Applicability of ARIMA Models in Wholesale Vegetable Market: An Investigation. IJISSCM 6, 105–119. https://doi.org/10.4018/ijisscm.2013070105

[6] Wei, W., 2006. Time Series Analysis: Univariate and Multivariate Methods, 2nd edition, 2006.

[7] Wirdyacahya, B.S., Prastuti, M., 2022. Peramalan Permintaan Semen di PT. XYZ Menggunakan Time Series Regression dan ARIMA. JSSITS 11, D96–D101. https://doi.org/10.12962/j23373520.v11i1.63222

[8] Falatouri, T., Darbanian, F., Brandtner, P., Udokwu, C., 2022. Predictive Analytics for Demand Forecasting – A Comparison of SARIMA and LSTM in Retail SCM. Procedia Computer Science, 3rd International Conference on Industry 4.0 and Smart Manufacturing 200, 993–1003. https://doi.org/10.1016/j.procs.2022.01.298

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 12 Juni 2024

Risky Dinal Ardianto - 23522305