Time Series Machine Learning: Implementing ARIMA and Hybrid ARIMA-ANN for Electricity Forecasting Modeling

Wahyu Wibowo[(✉)], Sarirazty Dwijantari, and Alia Hartati

Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
wahyu_w@statistika.its.ac.id

Abstract. The aims of this paper are to develop a linear and nonlinear model in time series to forecast electricity consumption of the lowest household category in East Java, Indonesia. The installed capacity in the lowest household customer category has various power, i.e. 450 VA, 900 VA, 1300 VA, and 2200 VA. ARIMA models are family of linear model for time series analysis and forecasting for both stationary and non-stationary, seasonal and non-seasonal time series data. A nonlinear time series model is proposed by hybrid ARIMA-ANN, a Radial Basis Function using orthogonal least squares. The criteria used to choose the best forecasting model are the Mean Absolute Percentage Error and the Root Mean Square Error. The ARIMA best model are ARIMA ([1, 2], 1, 0) $(0, 1, 0)^{12}$, ARIMA $(0, 1, 1)$ $(0, 1, 0)^{12}$, ARIMA $(0, 1, 1)$ $(0, 1, 0)^{12}$, ARIMA $(1, 0, 0)$ $(0, 1, 0)^{12}$ respectively. The ANN architecture optimum are ANN (2, 12, 1), ANN (1, 12, 1), ANN (1, 12, 1), and ANN (1, 12, 1). The best models are ARIMA ([1, 2], 1, 0) $(0, 1, 0)^{12}$, ARIMA $(0, 1, 1)$ $(0, 1, 0)^{12}$, ANN (1, 12, 1), and ANN (1, 12, 1) in each category respectively. Hence, the result shows that a complex model is not always better than a simpler model. Additionally, a better hybrid ANN model is relied on the choice of a weighted input constant of RBF.

Keywords: ARIMA · Hybrid ARIMA-ANN · Forecasting
Electricity consumption

1 Introduction

In the last decade, machine learning has growth rapidly as tool for making prediction. Basically, the most algorithms in machine learning are developed from classical statistical method. In the statistical perspective, machine learning frames data in the context of a hypothetical function (f) that the machine learning algorithm aims to learn. Given some input variables (Input) the function answers the question as to what is the predicted output variable (Output). For categorical data prediction, hundreds algorithm classifier can be applied in many software [1]. In the future, the feature of machine learning algorithm will be standardized, simple, specialized, composable, scalable yet cost effective in order to realize the goal of completely automated machine learning utilities, which will become an integral part of the modern software application [2].

One situation in making prediction by machine learning are dealing with the time series data. The input variables in time series data are observations from the time lag previously. There are many machine learning models for time series forecasting [3]. There is good paper to survey recent literature in the domain of machine learning techniques and artificial intelligence used to forecast time series data [4].

This paper aims to develop model for a linear and nonlinear model in time series machine learning to forecast electricity consumption of the lowest household category in East Java, Indonesia. ARIMA models are family of linear model for time series analysis and forecasting for both stationary and non-stationary, seasonal and non-seasonal time series data. A nonlinear time series model is proposed by hybrid ARIMA-ANN, a Radial Basis Function Neural Networks architecture within orthogonal least squares estimation. Radial Basis Fuction (RBF) neural networks is a class of feed forward network that use a radial basis function as its activation function [9]. The orthogonal least squares procedure choses radial basis function centers one by one in a rational way until an adequate network has been constructed [10].

The installed capacity in the lowest household customer category has various power, i.e. 450 VA, 900 VA, 1300 VA, and 2200 VA. Some papers previously have devoted to develop forecasting model for electricity. Two-level seasonal model based on hybrid ARIMA-ANFIS to forecast half-hourly electricity load in Java-Bali Indonesia [5]. The results show that two-level seasonal hybrid ARIMA-ANFIS model with Gaussian membership function produces more accurate forecast values than individual approach of ARIMA and ANFIS model for predicting half-hourly electricity load, particularly up to 2 days ahead. Meanwhile, the electricity forecasting for industrial category in East Java has been forecasted using ARIMA model [6].

2 Data and Method

The data is from PLN (Perusahaan Listrik Negara/National Electricity Company) as central holding company to serve and maintain supply and distribute electricity across Indonesia. The data is electricity consumption from the lowest household category (R1) in East Java and the time period is from January 2010–December 2016. The customer from the lowest household category is divided by various installed capacity, 450 VA, 900 VA, 1300 VA, and 2200 VA.

The forecasting method of individual series will be used by ARIMA model. ARIMA models, abbreviation of Autoregressive Integrated Moving Averaged, were popularized by George Box and Gwilym Jenkins in the early 1970s. ARIMA models are family of linear model for time series analysis and forecasting for both stationary and non-stationary, seasonal and non-seasonal time series data. There are four steps in ARIMA modelling are, model identification, parameter estimation, diagnostic checking and finally model is used in prediction purposes. There are model classification in the family of ARIMA [7].

- Model Autoregressive (AR)

  A pth-order of autoregressive or AR (p) model can be written in the form,

$$\dot{Z}_t = \varphi_1 \dot{Z}_{t-1} + \ldots + \varphi_p \dot{Z}_{t-p} + a_t \tag{1}$$

- Model Moving Averages (MA)

  A qth-order of moving averages or MA (q) model can be written in the form

$$\dot{Z}_t = a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \tag{2}$$

- Model Autoregressive Moving Average (ARMA)

  A pth order of Autoregressive and qth order of Moving Average or ARMA (p, q) model can be written in the form,

$$\dot{Z}_t = \varphi_1 \dot{Z}_{t-1} + \ldots + \varphi_p \dot{Z}_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \tag{3}$$

- Model Autoregressive Integrated Moving Average (ARIMA)

  By taking difference to the original, a pth order Autoregressive of and qth order of Moving Average, notated by ARIMA (p, d, q), the model will be

$$\phi_p(B)(1-B)^d \dot{Z}_t = \theta_0 + \theta_q(B)a_t \tag{4}$$

- Model Seasonal Autoregressive Integrated Moving Average (SARIMA)

  Seasonal ARIMA usually are notated by ARIMA (p, d, q) (P, D, Q)^S. The general form of Seasonal ARIMA can be written as,

$$\Phi_P(B^S)\varphi_p(B)(1-B)^d(1-B^S)^D \dot{Z}_t = \theta_q(B)\Theta_Q(B^S)a_t \tag{5}$$

Hybrid ARIMA-ANN is hybridisation of ARIMA and radial basis function (RBF) neural networks. RBF is a class of feed forward network that use a radial basis function as its activation function and a traditionally used for strict interpolation in multidimensional space [8], composed by three layers i.e. an input layer, a hidden layer, and an output layer. The input layer applies an orthogonal least squares procedure to choose the radial basis function centers in the input layer and a weighted input constant [9]. The hidden layer of RBF is nonlinear, whereas the output layer is linear. RBF network employs to modelling time series is $y(t)$ as the current time series value. The idea is to use the RBF network [10]

$$\widehat{y}(t) = f_r(x(t)) \tag{6}$$

as the one step-ahead predictor for $y(t)$, where the inputs to the RBF network

$$x(t) = \left[y(t-1)\ldots y(t-n_y)\right]^T \tag{7}$$

are past observation of the series. The nonlinearity $\phi(\cdot)$ within the RBF network is chosen to be the gaussian function (8) and the RBF network model writes as (9)

$$\phi(v) = \exp\left(-v^2/\beta^2\right) \tag{8}$$

$$f_r(x) = \lambda_0 + \sum_{i=1}^{n_r} \lambda_i \phi(\|\mathbf{x} - \mathbf{c}_i\|) \tag{9}$$

The weighted input constant of RBF is proposed by using measure of dispersion, it applied to build a radial basis function in the hidden layer are semi interquartile range, inter quartile range, range, and standard deviation [11].

$$Q = Q_3 - Q_1/2 \tag{10}$$

$$IQR = Q_3 - Q_1 \tag{11}$$

$$Range = Max - Min \tag{12}$$

$$s = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \bigg/ n - 1} \tag{13}$$

The criteria used to choose the best forecasting model are the Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE) [7].

$$MAPE = \frac{\sum_{t=1}^{n} \frac{|Z_t - \hat{Z}_t|}{Z_t}}{n} \times 100\% \tag{14}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(Z_t - \hat{Z}_t\right)^2} \tag{15}$$

3   Results

The short description regarding the electricity consumption is presented in Table 1. The table shows that the largest contributor to total consumption is from 900 VA power capacity as well as the smallest is from 2200 VA category. This is caused by the customer number of 900 VA is the largest and the customer number of 2200 VA is the smallest. Only around 9.84% the percentage of customer is from 1300 VA and 2200 VA.

Before starting the modeling procedure, electricity consumption data is divided into two parts, namely data in-sample and out-sample. The in-sample data is used to determine the modeling, while the out-sample data is used for model selection. The in-sample data uses data from January 2010 to December 2015, while the out-sample data uses data from January 2016 through December 2016.

Table 1. Description of electricity consumption of each category

Capacity	Mean	SD	Minimum	Maximum
450 VA	298.346.202	34.628.908	183.345.937	366.022.852
900 VA	349.990.409	75.226.313	184.317.229	474.736.024
1300 VA	101.155.026	12.484.667	61.474.737	120.808.796
2200 VA	67.034.305	8.526.496	41.580.243	79.661.957

Figure 1 shows the time series plot of the electricity consumption for each category. Obviously, the pattern of each series is trend and non-stationary. From the Fig. 1, it is not clear whether there is seasonal pattern or not for each series data. So, the monthly boxplot in Fig. 2 is presented to help to identify the seasonal pattern. By observing the line between monthly means of each box plot, each series looks have the similar pattern.

Clearly that the time series plot shows that the data is non-stationary and have seasonal pattern. Therefore, it is need to make the data become stationary before constructing the forecasting model. For this purpose, it will be taken $d = 1$ for non-seasonal, $D = 1$ for seasonal and $S = 12$ for seasonal. It can be shown that the data become stationary as shown in Annex 1.

Model Identification. As mentioned previously, the first step in ARIMA modeling is model identification based on stationary series data. By using Autocorrelation (ACF) and Partial Autocorrelation (PACF) plot, the tentative models can be proposed, especially order the ARIMA model. The ACF and PACF of each series data set is shown in Annex 1 and the tentative best models of each series is presented in Table 2.

Model estimation. Then, based on tentative ARIMA model, the parameter of each model will be estimated. The coefficient in each model, for both seasonal and non-seasonal, must satisfy the significant criteria (Annex 2).

Model checking. This involve the residual examination by using Ljung-Box test to check whether white noise or not and checking the p-value of the coefficient, then the significant model can be determined. Table 2 shows residual of the candidate best models already meet the white noise properties.

The best model. The best model would be selected by identifying model with the minimum MAPE and RMSE. Table 2 also shows the MAPE and RMSE of each best candidate model. Therefore, it can be identified easily, which model is the best. Table 3 present the best model of each series and its model specification. All the best models are seasonal models. These results correspond to the initial hypothesis at the identification stage that the best model is likely to be seasonal.

Forecasting. Based on the best model, then it will be computed forecasting electricity consumption of each individual series for 2017 year. Table 4 presented the forecasting of each series as well as Fig. 3 shows the interval forecasting of each individual series.

The forecasting result in Table 4 corresponds to the short data description as presented Table 1. The forecasting capacity 900 VA is highest in each month as well as forecasting of 2200 VA is lowest. In addition, the forecast results also generate an
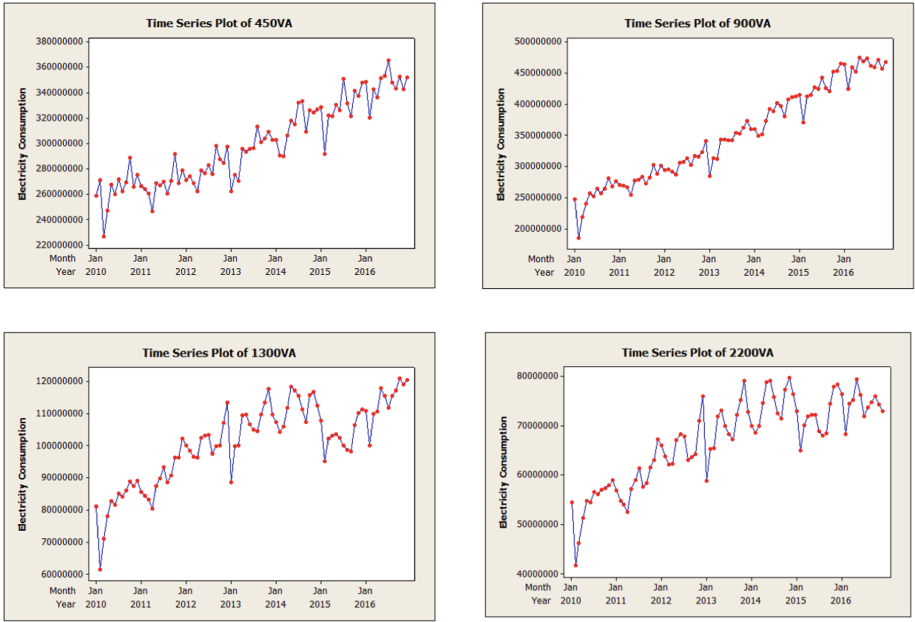
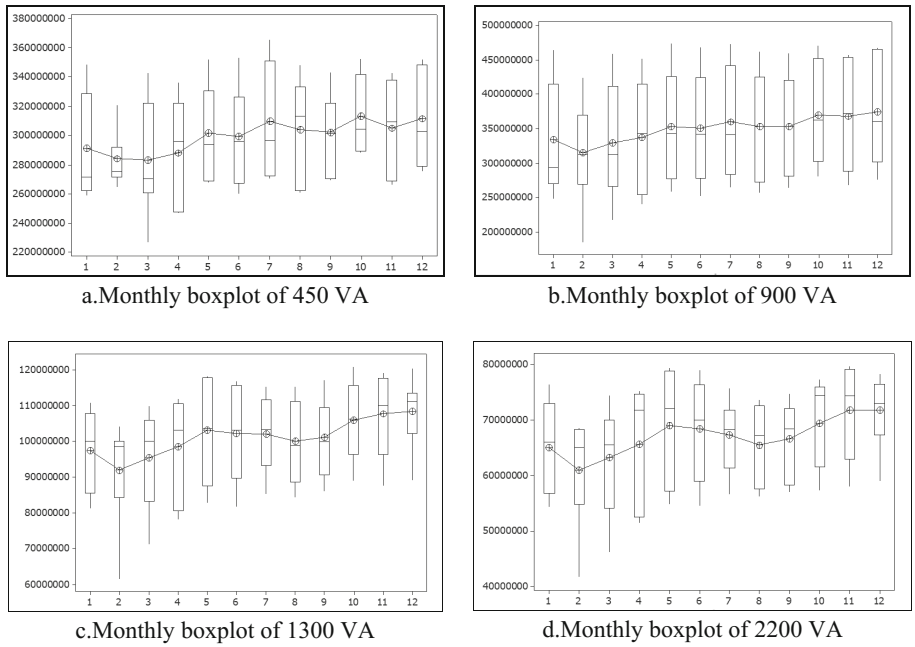**Fig. 1.** Individual time series plot

Fig. 2. Monthly boxplot individual time series

Table 2. Summary of the tentative model performance

Series	The tentative model	Residual	MAPE	RMSE
450 VA	ARIMA ([1, 2], 1, 0) (0, 1, 0)^{12}	White noise	1.78	7429122
	ARIMA (0, 1, 1) (0, 1, 0)^{12}	White noise	1.83	7586711
		White noise	3.38	22262659
900 VA	ARIMA (1, 1, 0) (0, 1, 0)^{12}	White noise	2.87	18347725
	ARIMA (0, 1, 1) (0, 1, 0)^{12}	White noise	3.24	21346976
	ARIMA ([1, 2], 1, 0) (0, 1, 0)^{12}			
1300 VA	ARIMA (1, 1, 0) (0, 1, 0)^{12}	White noise	11.69	14206993
	ARIMA (0, 1, 1) (0, 1, 0)^{12}	White noise	14.42	17186928
2200 VA	ARIMA (1, 0, 0) (0, 1, 0)^{12}	White noise	5.21	4240813
	ARIMA (1, 0, 1) (0, 1, 0)^{12}	White noise	6.13	4838444

Table 3. The best model if individual series.

Series	ARIMA best model	Model
450 VA	([1, 2], 1, 0) (0, 1, 0)^{12}	$(1 - \varphi_1 B - \varphi_2 B^2)(1 - B)^1(1 - B^{12})^1 \dot{Z}_t = a_t$
900 VA	(0, 1, 1) (0, 1, 0)^{12}	$(1 - B)^1(1 - B^{12})^1 \dot{Z}_t = (1 - \theta_1 B)a_t$
1300 VA	(1, 1, 0) (0, 1, 0)^{12}	$(1 - B)^1(1 - B^{12})^1 \dot{Z}_t = (1 - \theta_1 B)a_t$
2200 VA	(1, 0, 0) (0, 1, 0)^{12}	$(1 - \varphi_1 B)(1 - B^{12})^1 \dot{Z}_t = a_t$

Table 4. Forecasting of each individual series

Month	400 VA	900 VA	1300 VA	2200 VA
January	356.437.843	487.224.399	121.158.305	72.574.265
February	325.622.443	446.809.645	110.309.109	65.479.913
March	348.720.894	481.860.192	120.259.136	72.502.790
April	342.423.412	474.746.479	121.042.694	73.766.180
May	357.343.364	497.542.040	128.142.514	78.339.198
June	359.222.035	491.795.524	125.976.680	75.561.174
July	371.806.140	495.713.077	122.120.258	71.297.066
August	353.792.327	484.458.465	125.804.791	73.248.833
September	349.300.011	482.198.764	127.441.154	74.437.918
October	358.477.706	493.836.646	131.192.699	75.684.615
November	348.381.888	479.827.444	129.456.987	74.149.104
December	357.870.168	490.326.505	130.716.836	72.879.403

increasing trend from the previous years. Then, Fig. 4 presented the time series plot of forecasting results along with the upper and lower limits to be better understand the variation from month to month, what's going down and what month is up. Interestingly, forecasting in February resulted in minimum values.
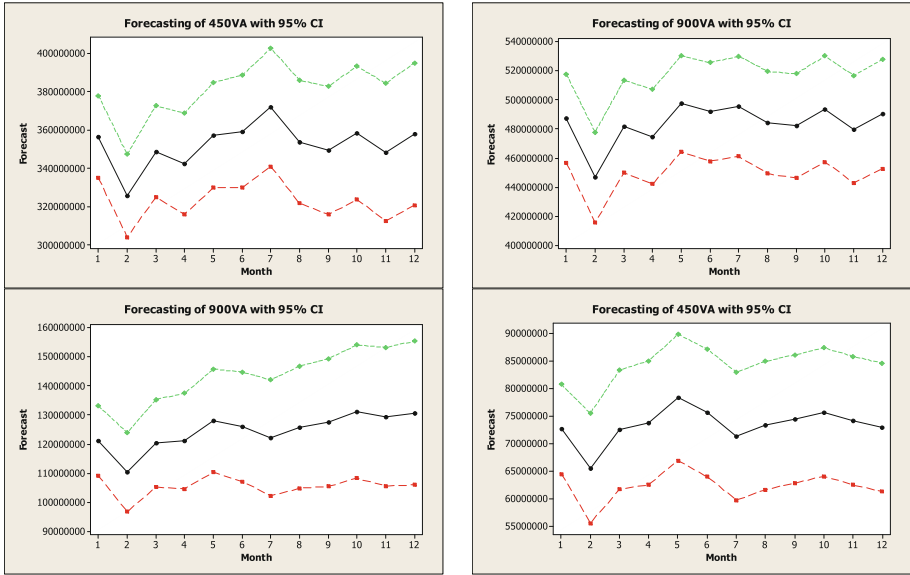
Fig. 3. Interval forecasting of each individual series

Hybrid ARIMA-ANN. The standard ARIMA model will be compared to Hybrid ARIMA-ANN. This model is built by the hybridization of ARIMA and RBF. The input is determined by the order of autocorrelation or partial autocorrelation of ARIMA model. The hidden layer is composed by the measure of dispersion constant as the width, determined by the seasonality pattern, and proceed by the orthogonal least squares. The measure of dispersion is determined by semi interquartile range (SIR), interquartile range (IQR), range, and standard deviation. The output layer is linear process as a univariate forecasting.

The architecture optimum. Based on Table 5, the architecture optimum is built by the measure of dispersion constant and proceed by the orthogonal least squares method. The architecture optimum is shown in Annex 2 (Table 6).

Forecasting. Based on the architecture optimum, it will be computed a univariate forecasting electricity consumption of each individual series in 2017. Table 7 presents the forecasting of each series as well as Fig. 4 shows the interval forecasting.

Comparation model. Based on the best model in each series, the best models are ARIMA ([1, 2], 1, 0) (0, 1, 0)^{12}, ARIMA (0, 1, 1) (0, 1, 0)^{12}, ANN (1, 12, 1), and ANN (1, 12, 1) respectively. It shows that a complex model is not always better than a simpler model. Hence, a better hybrid ANN model is relied on the choice of a weighted input constant of RBF. In addition, hybrid model forecasting results have a smaller variation compared to the ARIMA model. This can be seen in Fig. 4. Time series plot of forecasting results from month to month shows small changes. In addition, forecasting results also show a different pattern with the original data plot. If the time series of the original data in general shows an upward trend pattern, then the forecasting result of the hybrid model shows the pattern tends to go down as in forecasting 2200 VA.
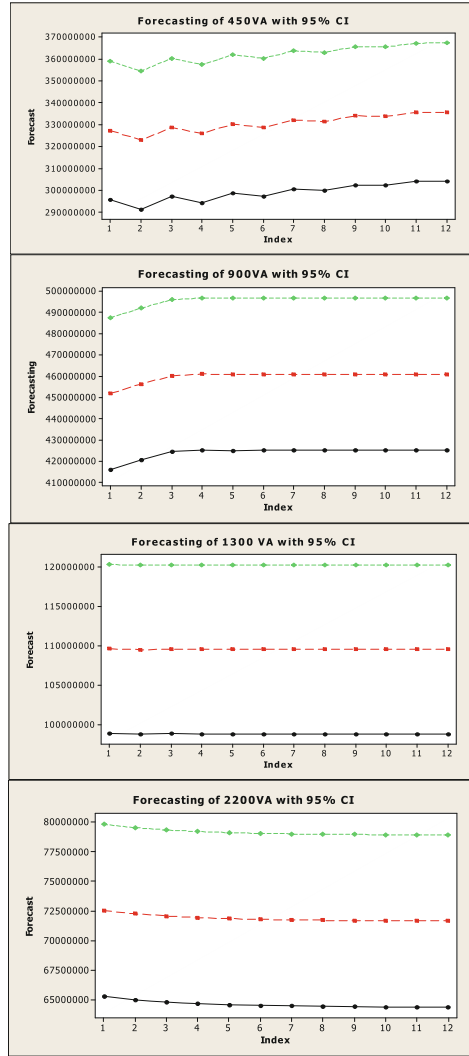
Fig. 4. Interval forecasting of each individual series

Table 5. The measure of dispersion of radial basis function

Capacity	SIR	IQR	Range	St. Deviation
450 VA	22929807.38	45859614.75	124845858	27678858.64
900 VA	53617084	107234168	281078109	64456310.27
1300 VA	9375176.38	18750352.75	56820592	11994479.18
2200 VA	6892877.25	13785754.5	38081714	8544743.80

Table 6. The tentative hybrid ARIMA-ANN architecture performance

Series	The architecture	MAPE	RMSE	Dispersion
450 VA	ANN (2, 12, 1)	7.232	28900205.8	SIR
	ANN (2, 12, 1)	10.846	47637343.31	IQR
	ANN (2, 12, 1)	7.252	30138899.10	Range
	ANN (2, 12, 1)	7.321	30605083.83	Standard deviation
900 VA	ANN (1, 12, 1)	79.75	500393532.39	SIR
	ANN (1, 12, 1)	187.58	1261928434.30	IQR
	ANN (1, 12, 1)	4.52	25172771.32	Range
	ANN (1, 12, 1)	104.89	671564249.29	Standard deviation
1300 VA	ANN (1, 12, 1)	5.13	6916929.11	SIR
	ANN (1, 12, 1)	8.73	18485362.53	IQR
	ANN (1, 12, 1)	5.99	9735898.68	Range
	ANN (1, 12, 1)	6.19	9738836.02	Standard deviation
2200 VA	ANN (1, 12, 1)	3.68	3465369.36	SIR
	ANN (1, 12, 1)	3.43	3290203.42	IQR
	ANN (1, 12, 1)	3.26	3119319.11	Range
	ANN (1, 12, 1)	3.55	3376712.69	Standard deviation

Table 7. Forecasting of each individual series

Month	450 VA	900 VA	1300 VA	2200 VA
January	327.264.333	451.794.490	109.590.586	72.538.369
February	322.862.608	456.310.202	109.485.466	72.249.617
March	328.646.077	460.195.002	109.513.222	72.055.395
April	325.904.722	461.057.594	109.506.386	71.922.917
May	330.278.944	460.815.162	109.508.103	71.831.855
June	328.665.855	460.901.946	109.507.674	71.768.973
July	332.091.946	460.872.634	109.507.781	71.725.467
August	331.347.847	460.882.490	109.507.754	71.695.299
September	333.965.777	460.879.290	109.507.761	71.674.361
October	333.826.763	460.880.314	109.507.759	71.659.827
November	335.596.047	460.879.546	109.507.760	71.649.721
December	335.709.623	460.880.186	109.507.760	71.642.705

4   Conclusion and Further Research

There are four individual series of electricity consumption, 450 VA, 900 VA, 1300 VA and 2200 VA. The time series machine learning to forecast each series is ARIMA model and the best model is ARIMA $([1, 2], 1, 0) (0, 1, 0)^{12}$, ARIMA $(0, 1, 1)$ $(0, 1, 0)^{12}$, ARIMA $(1, 1, 0) (0, 1, 0)^{12}$, and ARIMA $(1, 0, 0) (0, 1, 0)^{12}$ respectively. It is also noted that the largest contribution to household electricity is from 900 VA category. Meanwhile, the hybrid ARIMA-ANN forecasting generates an architecture

and an architecture optimum i.e. ANN (2, 12, 1) using semi interquartile, ANN (1, 12, 1) using range, ANN (1, 12, 1) using semi interquartile, and ANN (1, 12, 1) using range respectively. The comparation model shows that there is an inconsistent model because a complex model doesn't work better in two series of electricity consumption i.e. the 450 VA series and the 900 VA series.

If the electricity consumption of each category is summed up, then it will produce the total household electricity consumption. This time series data is called hierarchical time series data. The alternative method can be used for this type of time series data is hierarchical time series model by applying the top-down, bottom up and optimal combination approach [12]. These approaches could be considered to be applied to this data in the next research.

Annex 1. ACF and PACF Plot
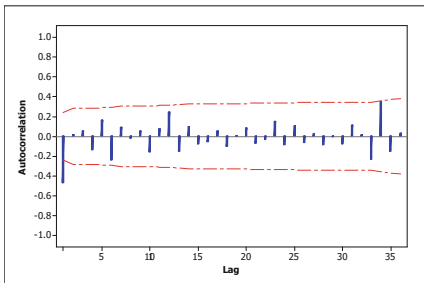
a.1. Plot ACF *Differencing* Lag 1 of 450VA
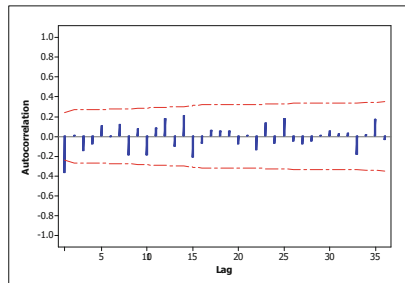
b.1 Plot ACF *Differencing* Lag 1 of 900VA
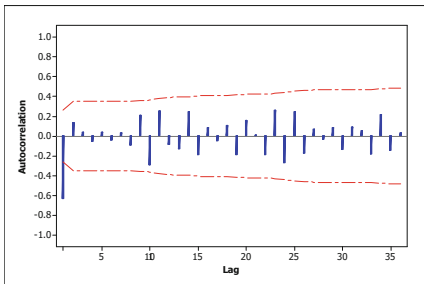
a.2. Plot ACF *Differencing* Lag 12 of 450VA
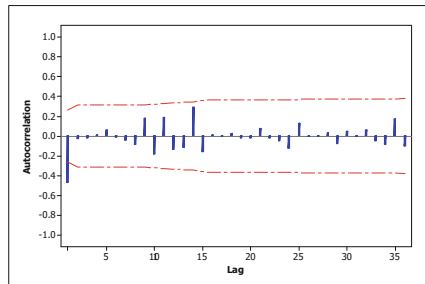
b.2. Plot ACF *Differencing* Lag 12 of 900VA
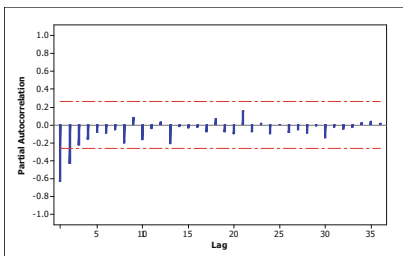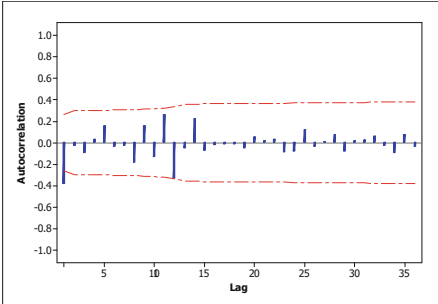

a.3. Plot PACF of 450VA
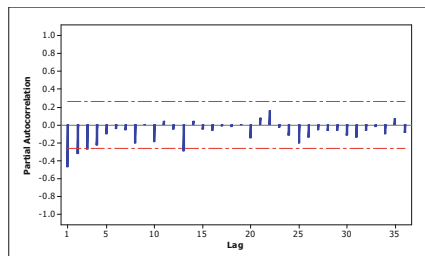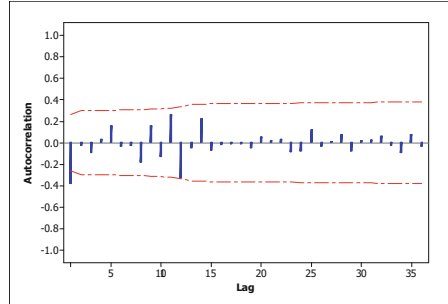

b.3. Plot PACF of 900VA
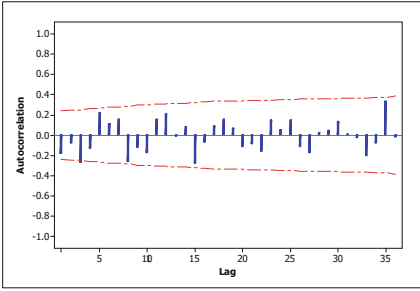
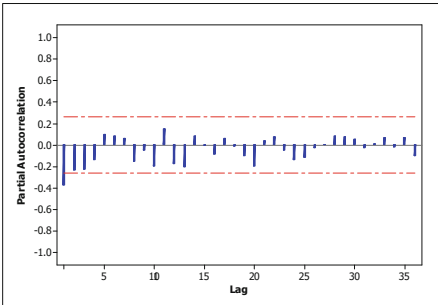**c.1.** Plot ACF *Differencing* Lag 1 of 1300VA
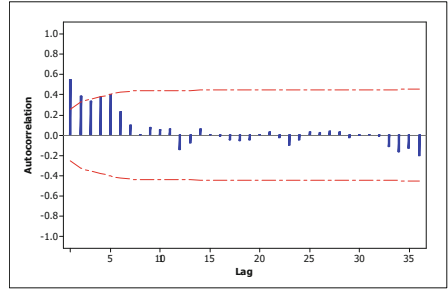
d.1. Plot ACF *Differencing* Lag 12 of 2200VA
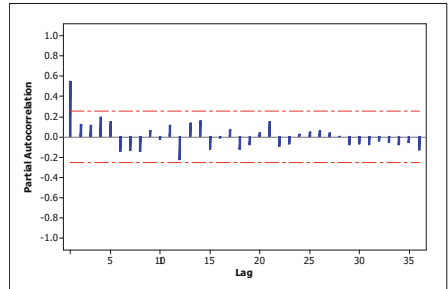

c.2. Plot ACF *Differencing* Lag 12 of 1300VA

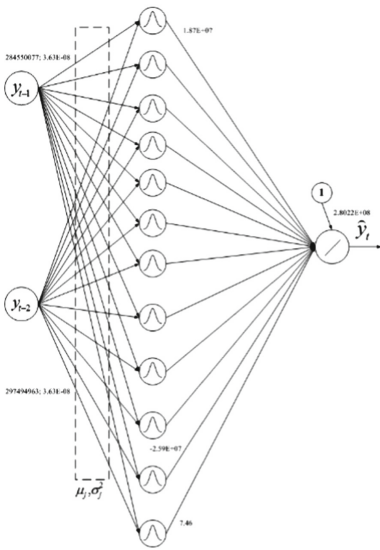**d.2.** Plot ACF *Differencing* Lag 12 of 2200VA

c.3. Plot PACF of 1300VA
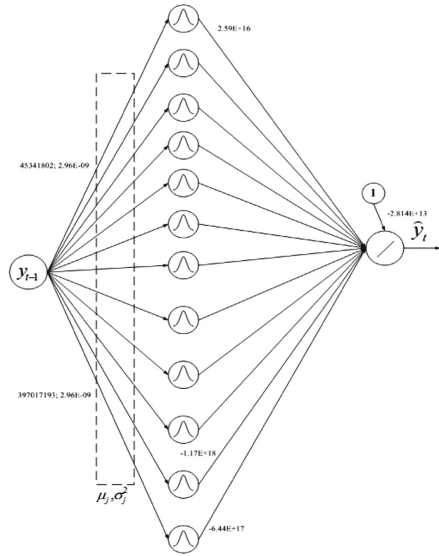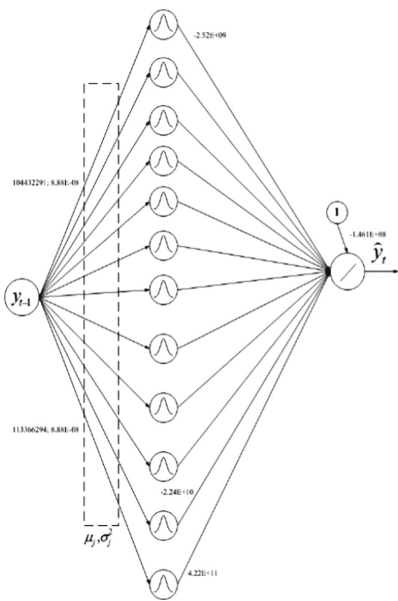

d.3. Plot PACF of 2200VA

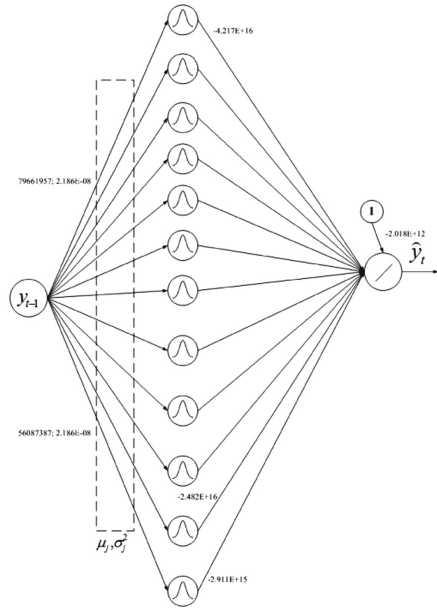# Annex 2. The Architecture Optimum of Hybrid ARIMA-ANN

a. ANN (2,12,1) of 450 VA category

b. ANN (1,12,1) of 900VA category

c. ANN (1,12,1) of 1300 VA category

d. ANN (1,12,1) of 2200 VA category

References

1. Delgado, M.F., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. **15**, 3133–3181 (2014)
2. Cetinsoy, A., Martin, F.J., Ortega, J.A., Petersen, P.: The past, present, and future of machine learning APIs. In: JMLR: Workshop and Conference Proceedings, vol. 50, pp. 43–49 (2016)
3. Ahmed, N.K., Atiya, A.F., Gayar, N.E., Shishiny, H.E.: An empirical comparison of machine learning models for time series forecasting. Econom. Rev. **29**(5–6), 594–621 (2010)
4. Krollner, B., Vanstone, B., Finnie, G.: Financial time series forecasting with machine learning techniques: a survey. In: ESANN 2010 Proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning, Bruges (Belgium), 28–30 April 2010. D-Side Publishing (2010). ISBN 2-930307-10-2
5. Suhartono, Puspitasari, I., Akbar, M.S., Leem, M.H.: Two-level seasonal model based on hybrid ARIMA-ANFIS for forecasting short-term electricity load in Indonesia. In: IEEE Proceeding, International Conference on Statistics in Science, Business and Engineering (ICSSBE) (2012). https://doi.org/10.1109/ICSSBE.2012.6396642
6. Saputri, I.A.: Forecasting of Electricity in Industrial Sector at PT PLN East Java: Final Assignment, June 2016. Department Of Statistics, Faculty of Mathematics and Natural Sciences (2016)
7. Cryer, J.D., Chan, K.: Time Series Analysis With Application in R, 2nd edn. Springer Science Business Media, New York (2008). https://doi.org/10.1007/978-0-387-75959-3
8. Sivanandam, S.N., Sumathi, S., Deepa, S.N.: Introduction Neural Networks Using Matlab 6.0. Tata McGraw-Hill, New Delhi (2006)
9. Beale, M.H., Demuth, H.B., Hagan, M.T.: Neural Network Toolbox User's Guide R2011b. The MathWorks Inc., Natick (2011)
10. Chen, S., Cowan, C.F., Grant, P.M.: Orthogonal least squares learning algorithm for radial basis function networks. IEEE Trans. Neural Netw. **2**(2), 302–309 (1991)
11. Mann, P.S.: Statistics for Business and Economics. Wiley, Hoboken (1995)
12. Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., Shang, H.L.: Optimal combination forecasts for hierarchical time series. Comput. Stat. Data Anal. **55**, 2579–2589 (2011)