The 6th International Seminar on Science and Technology (ISST) 2020
July 25th 2020, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Hybrid Double Seasonal ARIMA and Support Vector Regression in Short-Term Electricity Load Forecasting

Kinanti Hanugera Gusti, Irhamah, and Heri Kuswanto
Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya
e-mail: kinantihanugera@gmail.com

Abstract—**Forecasting is the main purpose of time series modelling. In short-term forecast, data can be predicted for a half hour-ahead. A half hour-ahead prediction faced with overlapping data series patterns risk. On the other hand, time series model can be analyzed with a linier or nonlinier approach. In this paper, we proposed the combination (hybrid) liner and nonlinier model for modelling the short-term electricity load in East Java. A half-hour electricity load forecasting is needed for real time controlling and short-term maintenance schedulling. However, the main problem of modelling time series data is determining linier or nonlinier time patterns. In short-term electricity load forecast, it depend on the moment of time (i.e weekdays, weekend, public holidays, joint holidays or religious holiday, etc) and the electricity load classification. In this analysis, we developed the Double Seasonal ARIMA (DSARIMA), Support Vector Regression (SVR), and hybrid DSARIMA-SVR. The DSARIMA model belong to linier model based on a well-known Box-Jenkins methodology. The SVR model belong to nonlinier model and the hybrid model is a mixing of linier and nonlinier models. The models are evaluated using Root Mean Square Error (RMSE) and Symmetric Mean Absolute Percentage Error (MAPE). The result shows that the accuracy of hybrid DSARIMA-SVR models are superior to the other individual models.**

Keywords—**Forecast, DSARIMA, SVR, Hybrid, Load.**

I. INTRODUCTION

ACCURACY is the most important thing in time series forecasting as the main point of decision making. Over a few decades, an extensive time series forecasting models have been expanded to achieve the accuracy forecast performance. The purpose of this method is to find historical pattern patterns and extrapolate these patterns into the future so that the results can be used as a reference for forecasting future values. The time series forecasting method is divided into two parts. First, forecasting model based on statistical mathematical approach such as AR, MA, ARIMA, and SARIMA. Second, forecasting model based on artificial intelligence. The ARIMA method is known have very good accuracy for short-term forecasting and for non stationary time series linear data, but when a long period forecasting data is done the accuracy tend to be flat in average value [1].

Although the ARIMA model is quite flexible, identifying more complex models requires more experience and linear assumptions of the modeled data are considered unsuitable for modeling complex nonlinear time series. Methods with an articial intelligence approach have the ability to accurately predict nonlinear pattern data compared to the ARIMA method. In general, support vector machine build a hyperplane or set of hyperplane in the dimensions of high space or limited space, which can be used for classification, regression, or other tasks. The SVR method produces more accurate forecasting result among compared to other artificial intelligence approach, such as Neural Network (NN), because using principle of structural risk minimization by minimizing the bound of generalization error to overcome overfitting [2].

The main concept of SVR is to maximize the margin around the hyperplane and to obtain data points that become the support vectors. Although the SVR method has advantages in terms of accuracy, but these advantages depend on the choice of optimal parameter values. Fahmi and Sofyan studied at forecasting household electricity consumption in Aceh using significant lags from significant PACF lag of ARIMA as input selection of Forward Neural Networks (FFNN) [3]. Riyani et al studied forecasting daily sales of men clothes using significant PACF lag of ARIMA and ARIMAX as feature selection by choosing lags as input on SVR [4]. Other studied regarding forecasting Crude Palm Oil (CPO) use PACF lag data as input selection on SVR [5].

The case study in this research is a half hour electricity load. A half-hour electricity load forecasting is needed for real time controlling and short-term maintenance schedulling. In short-term electricity load forecast, it depend on the moment of time (i.e weekdays, weekend, public holidays, joint holidays or religious holiday, etc) and the electricity load classification, so it is difficult to determine seasonal patterns. The complicated of characteristic pattern is possible because of the overlap recurring patterns and are generated by linear and nonlinear processes [3]. In this study, we propose the combination model for modelling electricity forecast.The idea behind this approached is the compelling evidence of most time series data is probably generated by linear and non-linear processes.

Overlapping time patterns are also found in Khusna and Suhartono’s research namely Double Seasonal ARIMA (DSARIMA) for short-term electricity load data. In addition, different from researches before we will use DSARIMA as linear process, the SVR model with significant lags PACF as feature selection, and SVR model with significant lags based on DSARIMA as hybrid linear and nonlinear model [5].

II. METHODS

A. Double Seasonal ARIMA

Generally short-term electricity load data have a double seasonal pattern [5]. The ARIMA model that suitable for

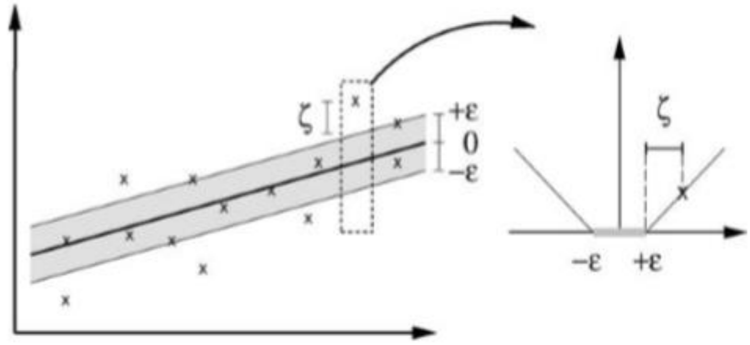


Figure 1. Boundary Illustration in SVR.

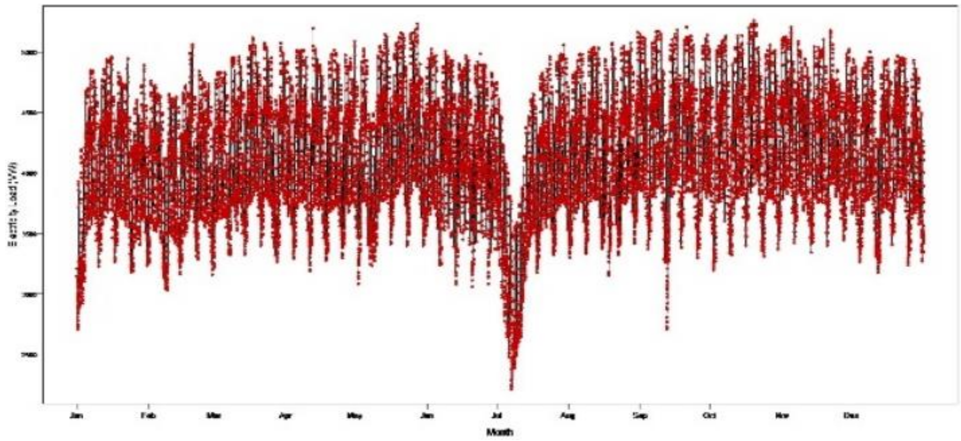


Figure 2. Time Series Plot of Electricity Load.

Table 1.
Structure Data

Index	Day	Month	Electricity Load	Data
1	1	January	$Y_{i,1}$	*Training*
⋮	⋮	⋮	⋮	⋮
1	365	November	$Y_{i,17549}$	*Training*
1	365	November	$Y_{i,17520}$	*Training*
17521	1	December	$Y_{i,17521}$	*Testing*
⋮	⋮	⋮	⋮	⋮
24768	730	December	$Y_{i,24768}$	*Testing*

short-term electricity load forecasting is multiplicative double seasonal ARIMA or ARIMA $(p,d,q)(P_1,D_1,Q_1)^{S1}(P_2,D_2,Q_2)^{S2}$. Mathematically, this ARIMA model can be written with the following equation

$$\phi_p(B)\Phi_{P_1}(B^{S_1})\Phi_{P_2}(B^{S_2})(1-B)^d(1-B^{S_1})^{D_1}(1-B^{S_2})^{D_2}Z_t$$
$$=\theta_q(B)\Theta_{Q_1}(B^{S_1})\Theta_{Q_2}(B^{S_2})a_t \quad (1)$$

B. Support Vector Regression

SVM is method used for classification, but the principle of the method can be developed in regression and forecasting methods. For example, there is i training data $(\boldsymbol{x}_i, y_i)$ with input data $\boldsymbol{x}=\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_P\}$, we get the the following regression function

$$f(\boldsymbol{X})=\langle\boldsymbol{\omega},\phi(\boldsymbol{X})\rangle+b \quad (2)$$

In order to obtain the regression function as thin as possible, the solution obtained to minimize the following object functions

$$\min\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{T}(\xi_i+\xi_i^*) \quad (3)$$

Where the loss insentitive function is defined as follows

$$|y-f(\boldsymbol{X},\boldsymbol{\omega})|_\varepsilon=\begin{cases}0, & |y-f(\boldsymbol{X},\boldsymbol{\omega})|\le\varepsilon\\ |y-f(\boldsymbol{X},\boldsymbol{\omega})|-\varepsilon, & lainnya\end{cases} \quad (4)$$

Illustration of the ε boundary shown in Figure 1. Only observations outside the shades area or outside boundary ε are given a constanta C, as well as any deviation inseide shaded area will be given a zero value. The optimization of problem (3) is solved using dual formulation by forming primal lagrangian as in equation (5) and continued to optimize using dual lagrangian in equation (6). SVR parameter is done by KKT optimization so that the general equation is obtained as equation (7).

$$L_{SVM}=\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{k=1}^{T}(\xi_i+\xi_i^*)-\sum_{k=1}^{T}(\eta_i\xi_i+\eta_i^*\xi_i^*)$$
$$-\sum_{i=1}^{T}a_i\left(\varepsilon+\xi_i-y_i+\langle\boldsymbol{\omega},\phi(\boldsymbol{X}_i)\rangle+b\right)$$
$$-\sum_{i=1}^{T}a_i^*\left(\varepsilon+\xi_i^*-y_i+\langle\boldsymbol{\omega},\phi(\boldsymbol{X}_i)\rangle+b\right) \quad (5)$$

$$\max_{\alpha,\alpha^*}Q=-\frac{1}{2}\sum_{i,l=1}^{T}(\alpha_i-\alpha_i^*)(\alpha_i-\alpha_l^*)\langle\phi(\boldsymbol{X}_i),\phi(\boldsymbol{X}_l)\rangle$$
$$-\varepsilon\sum_{i=1}^{T}(\alpha_i-\alpha_i^*)+\sum_{i=1}^{T}y_i(\alpha_i-\alpha_i^*) \quad (6)$$

$$f(\boldsymbol{X})=\sum_{i=1}^{T}(\alpha_i-\alpha_i^*)K(\boldsymbol{X},\boldsymbol{X_i})+b \quad (7)$$

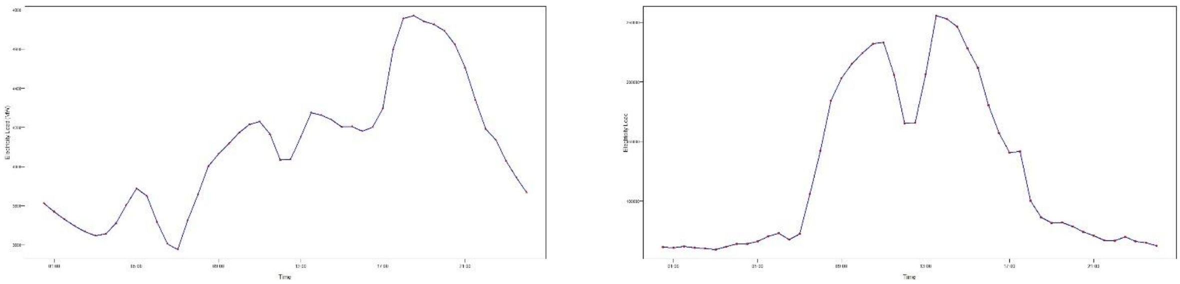

(a) (b)

Figure 3. (a) Mean of Electricity Load; (b) Variance of Electricity Load.

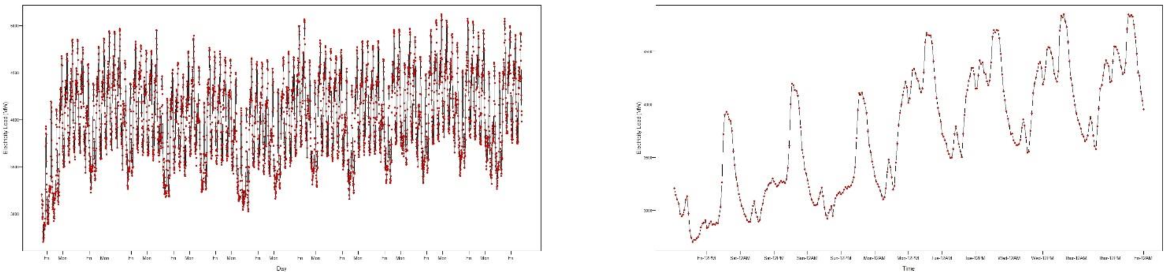

(a) (b)

Figure 4. (a) Weekly Seasonal Plot of Electricity Load in East Java; (b) Daily Seasonal Plot of Electricity Load in East Java.

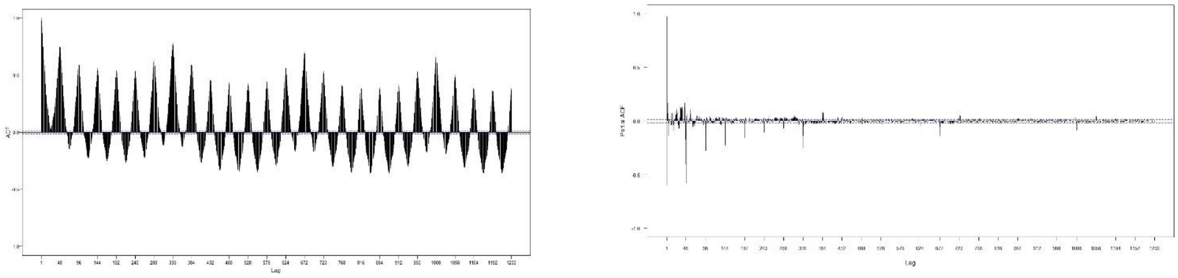

(a) (b)

Figure 5. (a) ACF Plot of Electricity Load; (b) PACF Plot of Electricity Load.

C. *Best Model Criteria*

The model selection is resulted using testing criteria by comparing the Root Mean Square Error (RMSE) and Symmetric Mean Absolute Percentage Error (SMAPE) as follows:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_t^2} \qquad (8)$$

$$s\,MAPE = \frac{100\%}{T}\sum_{t=1}^{T}\frac{|Y_t-\hat{Y}_t|}{(|Y_t|+|\hat{Y}_t|)/2} \qquad (9)$$

where *t* is the amount of data.

D. *Methodology*

This research analyze the short-term (a half hour) electricity load in East Java from January 2016 until December 2016. The data is divided into training data from January 2016 to November 2016 and December 2016 data as testing data. The load data used is recorded at 20kV substation (distribution section) so that the load is not classified into various type. Input and Output variables are describes as follows:

1) *Ouput Series*

Z_t: a half hour electricity load in East Java.

2) *Input Series*

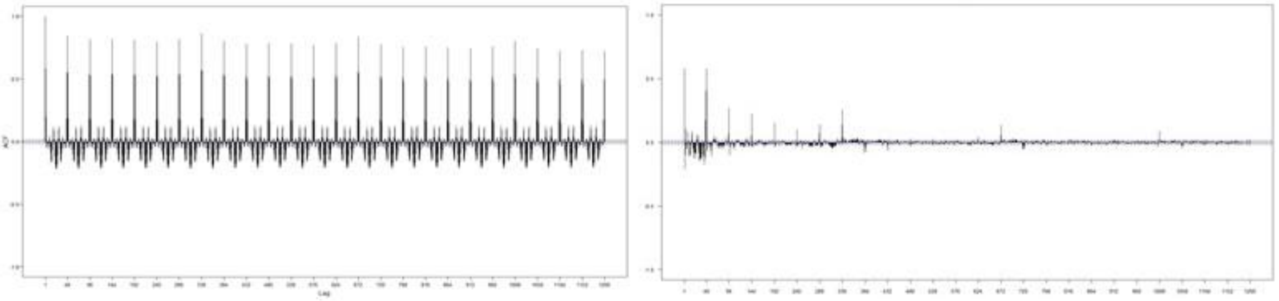
X : significant lag of PACF plot; significant lag of PACF from DSARIMA

Table 1 shows the data structure used. The step used to analyze the data in this research is described as follows:

1. Describing the characteristics of a half-hour electricity load in East Java using time series plot.
2. Checking stasionarity of training data (checking stationary in mean).
3. Identifying the order of ARIMA based on ACF and PACF.
4. Modelling ARIMA based on the result of step 3.
5. Calculating the performance of proposed method using RMSE and SMAPE.
6. Using significant lag from PACF data plot model as input SVR.
7. Using significant lag from PACF of ARIMA model as input SVR.
8. Calculating the performance model of step 6 and 7.
9. Comparing the results.

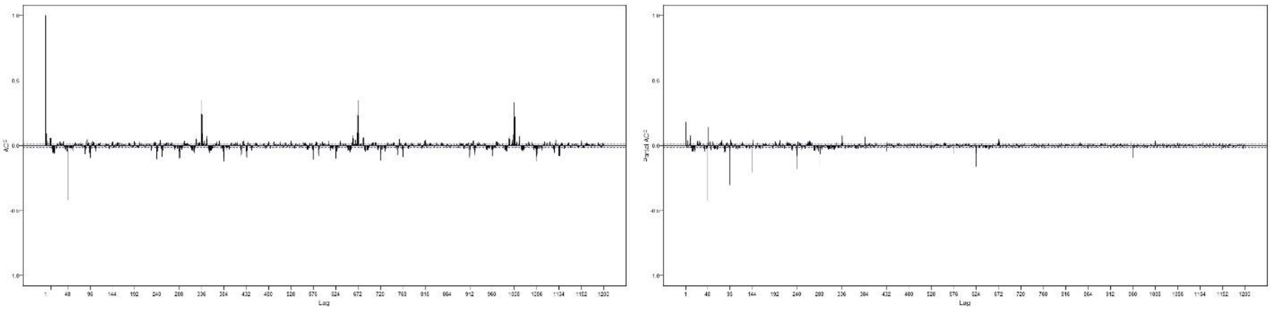
III. RESULTS AND DISCUSSION

The characteristics of electricity load data in East Java are obtained through descriptive statistical analysis by exploring information in data without making inference. Figure 2 shows the demand for electricity load plots recorded per half hour from January 1 to December 31 2016. There is up and down trend pattern that explains the non-stationary mean data. The up and down trend pattern is explained in Figure 3. Based on



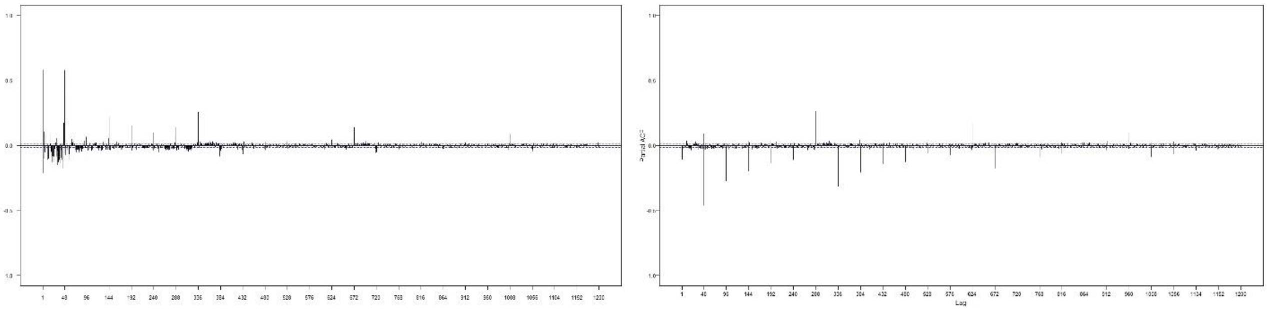
(a) (b)

Figure 6. (a) ACF Plot First Lag Differencing; (b) PACF Plot First Lag Differencing.



(a) (b)

Figure 7. (a) ACF Plot 48th Lag Differencing; (b) PACF Plot 48th Lag Differencing.



(a) (b)

Figure 8. (a) ACF Plot 336th Lag Differencing; (b) PACF Plot 336th Lag Differencing.

Figure 3(a), It can be seen that the lowest average electricity consumption occurs at 07:00 along with the commencement of community activities to work outside the home. The average electricity load consumption reaches its highest peak at 18:30 along with the end of industrial activities and activities outside the home. After that, the average electricity load gradually drops along with the inclusion of community and agency breaks and agencies that do not operate at night. In addition, based on Figure 3(b) it is known that during the daytime electricity consumption has a high variance. This relates to consumption patterns that are dominated by the industrial sector. Whereas in the morning and at night, electricity consumption tends not to be varied.

The characteristics of seasonal patterns from the data are shown in Figure 4 which shows the amount of electricity consumption from January 1 to March 31, 2016. This illustrates the weekly seasonal pattern where on Saturday and Sunday the consumption of electricity loads is lower compared to active day, because it is contributed by industrial consumption which operates on that day. Meanwhile, Figure 4 also shows the magnitude of electricity consumption on January 1 to January 7, 2016, which illustrates the pattern of low electricity consumption at night until early morning and then an increase in the morning to evening. This phenomenon shows the alleged existence of daily seasonal patterns in the data. Based on the information above it is estimated that there are daily variations (per 48 hours) and weekly variations (per 336 hours) which will then be used in modeling.

The characteristics of seasonal patterns from the data are shown in Figure 4 which shows the amount of electricity consumption from January 1 to March 31, 2016. This illustrates the weekly seasonal pattern where on Saturday and Sunday the consumption of electricity loads is lower compared to active day, because it is contributed by industrial consumption which operates on that day. Meanwhile, Figure 4 also shows the magnitude of electricity consumption on January 1 to January 7, 2016, which illustrates the pattern of low electricity consumption at night until early morning and then an increase in the morning to evening. This phenomenon shows the alleged existence of daily seasonal patterns in the data. Based on the information above it is estimated that there are daily variations (per 48 hours) and weekly variations (per 336 hours) which will then be used in modeling.

A. Double Seasonal ARIMA

Identification of the stasionary data will be identified by the ACF and PACF plot presented by Figure 5.

Table 2.
Estimated ARIMA Parameters

	Model 1		Model 2	
	Est	p-value	Est	p-value
θ_{11}	-0.052	<.000	-0.053	<.000
θ_{12}	-0.029	0.000	-0.027	0.000
θ_{31}	-0.230	0.001	-	-
$\Theta_{1(48)}$	0.707	<.000	0.708	<.000
$\Theta_{2(336)}$	0.786	<.000	0.786	<.000
$\Theta_{3(672)}$	-0.031	0.000	-0.031	0.000
ϕ_{28}	-0.012	0.010	-0.019	0.017
ϕ_{31}	-0.240	0.000	-	-
ϕ_{47}	0.078	<.000	0.080	<.000

Table 3.
Performance of Grid-Search SVR

epsilon	cost	gamma	RMSE
1E-04	1E-04	0.01	8883.432
1	1E-04	0.01	8838.269
1E-04	10	0.01	85.481
1	10	0.01	1166.597
1E-04	1E-04	50	9069.867
1	1E-04	50	9044.719
1E-04	10	50	9064.944
1	10	50	8887.404

Table 4.
Performance of Grid-Search SVR

epsilon	cost	gamma	RMSE
1E-04	1E-04	0.01	8883.432
1	1E-04	0.01	8838.269
1E-04	10	0.01	85.481
1	10	0.01	1166.597
1E-04	1E-04	50	9069.867
1	1E-04	50	9044.719
1E-04	10	50	9064.944
1	10	50	8887.404

Table 5.
Algorithm Comparison

	SVR		DSARIMA-SVR	
	RMSE	SMAPE	RMSE	SMAPE
Training	8.685	0.000	23.517	0.003
Testing	85.800	0,0162	58.737	0.011

Figure 6 shows the data is not stationary in the mean because of the ACF plot drops very slowly and Figure 6 shows the highest significant PACF lag in lag 1 so that the data will be differencing in lag 1.

Figure 7 shows the ACF plot having a repeating pattern. This is reinforced by the PACF plot which has a high significance at a multiple of 48, so the data must be differencing at lag 48 with the aim of the data being stationary.

In Figure 8, there is still unstability data. The ACF data plot looks repeated in multiples of 336. in addition, the highest significance of PACF lag at multiple of 336 indicates that data needs to be performed differencing at lag 336.

The possibility of model being formed is,

$$ARIMA([28,31,47],1,[11,12,31])(0,1,1)^{48}(0,1,2)^{336}$$

$$ARIMA([28,47],1,[11,12])(0,1,1)^{48}(0,1,2)^{336}$$

Furthermore, after the model is obtained, the estimated parameter values will be calculated and the results are presented in Table 2.

Both models have significant parameters, because the p-value of estimate parameter under the tolerance error, so forecasting can be done and compared with testing data. The procedure produces RMSE values of 618.169 and 646.465 with sMAPE values of 13.76 and 14.48 for each model 1 and model 2. The best model is model 1 which can be written as follows

$$(1-\phi_{28}B^{28}-\phi_{31}B^{31}-\phi_{47}B^{47})(1-B)(1-B^{48})$$
$$(1-B^{336})(1-B^{672})Z_t$$
$$=(1-\theta_{11}B^{11}-\theta_{12}B^{12}-\theta_{31}B^{31})(1-\Theta B^{48})$$
$$(1-\Theta B^{336})(1-\Theta B^{672})a_t$$

Forecasting results for 1488 half-hour ahead based on the model obtained are visualized in Figure 9.

Figure 9 shows the plot between the testing data (black line) and the forecast result (red line). This indicates that the estimation results obtained by the DSARIMA model are still far from the original data and there is a possibility that there are other patterns that cannot be captured, so an analysis of artificial intelligence approach will be tried in the hope that a better estimation result can be produced.

B. SVR

Modeling using SVR as a nonliner process uses significant PACF data lag values based on stationary data such as Figure 8. The PACF plot will be cut at 1200 lags to obtain 180

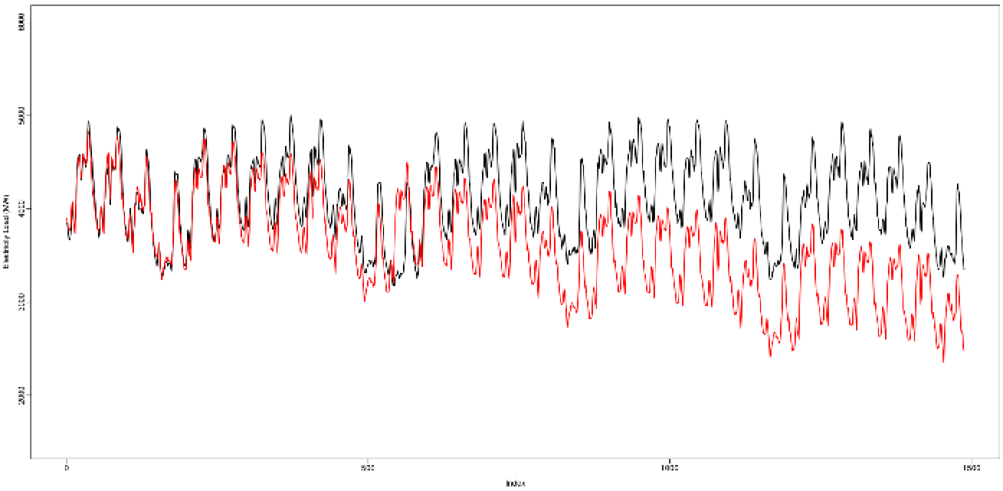


Figure 9. Plot Actual vs Predicted DSARIMA.

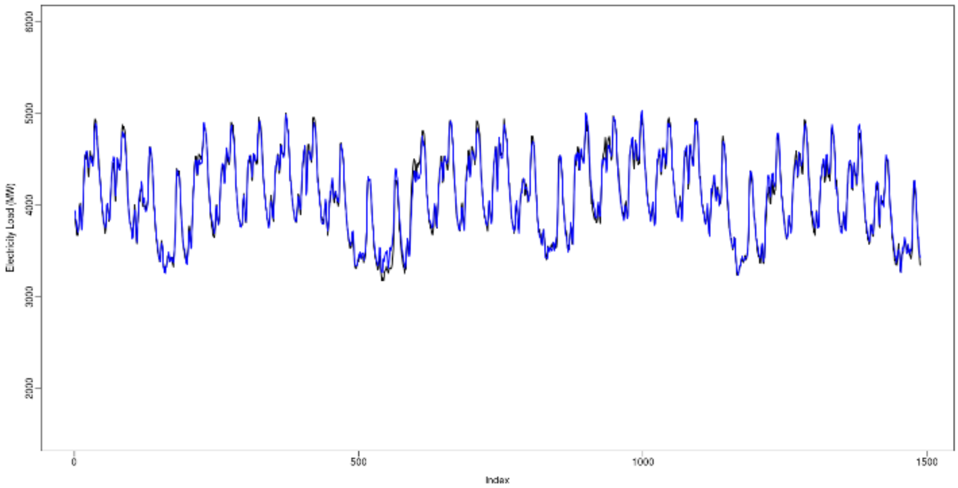


Figure 10. Actual vs Predicted SVR.

significant lags and will be used as input variables. Initially, as much as the largest lag of initial data will be empty (NAN). So that we will get 14837 training data with 180 variable inputs.

The selection of optimal parameters is the most important thing in the SVR method in order to obtain accurate forecast results. The parameter determination will be done using Grid Search by dividing the range of parameters to be optimized into the grid and crossing all points to get the optimal parameters. These various possible parameters will be evaluated using the minimum RMSE criteria. Range of parameter used are {(1e-2,1),(1e-4,10),(1e-2,50)} for ε,C, and σ^2. Based on minimum RMSE criteria, the optimal parameters are 1e-4,10,1e-2 for ε,C, and σ^2. The performance of the model obtained is the RMSE value of 85.481 and SMAPE of 0.016 and the details are shown at Table 3. The result of SVR model forecast are shown in Figure 10.

Based on Figure 10 shows the plot between the testing data (black line) and the forecast result (blue line) of SVR. The estimation results obtained by the SVR as a nonlinear process have quite high accuracy because the forecast results are close to the original data.

C. *Double Seasonal ARIMA – SVR*

Modeling using SVR as a hybrid linear and nonliner process uses significant PACF data lag values based on DSARIMA model is different from the previous procedure. Significant PACF lag on DSARIMA obtained by describing the model eqaquations. Based on the obtained model, 111 significant lags were obtained with the largest value being 1163. So that we will get 14869 training data with 111 variable inputs.

As well as SVR method as a nonlinear process, the range of parameters to be used is equal to $\{(1e-2,1),(1e-4,10),(1e-2,50)\}$ for ε, C, and σ^2. Based on minimum RMSE criteria, the optimal parameters are 1e-2,10,1e-2 for ε, C, and σ^2. The performance of the model obtained is the RMSE value of 58.74 and SMAPE of 0.011 and the details are shown in Table 4. The result of SVR model forecast are shown in Fig 11.

Based on Figure 11 shows the plot between the testing data (black line) and the forecast result (blue line) of DSARIMA-SVR. The estimation results obtained by the DSARIMA-SVR as a hybrid linear and nonlinear process have quite high accuracy because the forecast results are close to the original data.

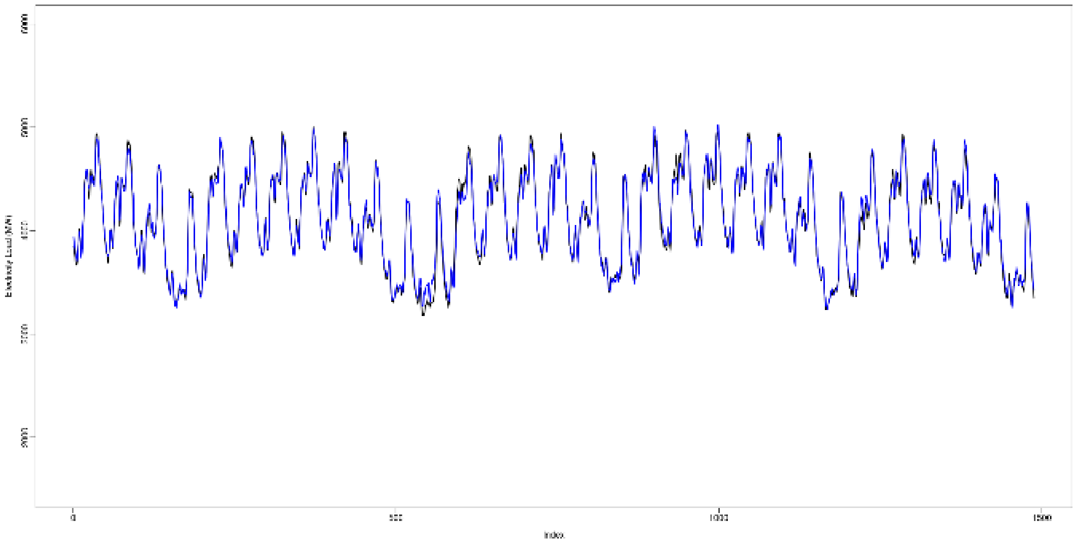


Figure 11. Actual vs Predicted SVR.

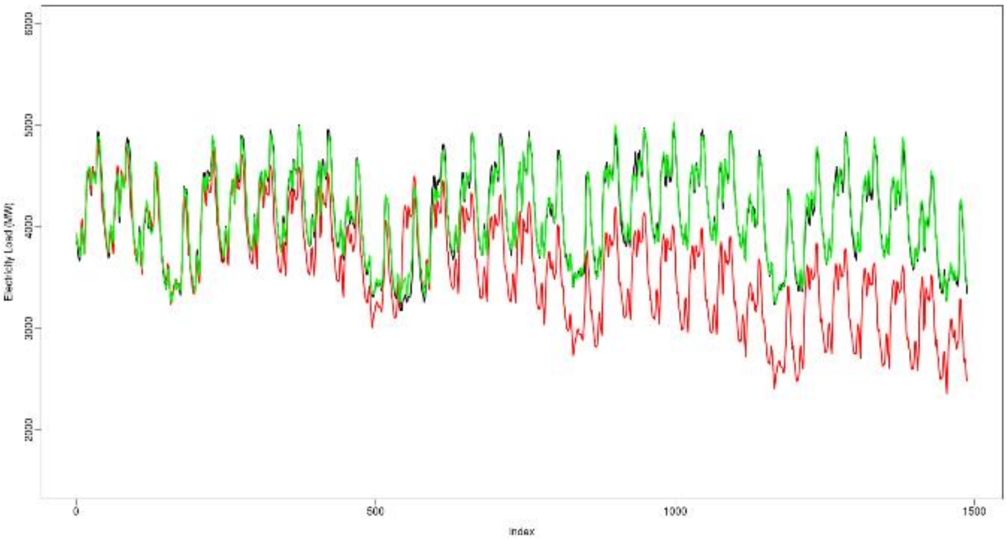


Figure 12. Comparison The Result of Estimation.

D. *Algorithm Comparison*

After obtaining the forecast results and the goodness of the model from each method, the next step is to compare the goodness of the model between methods. The best model criteria used is minimum RMSE. A summary of the comparison of algorithms is presented in Table 5.

Based on Table 5 it can be seen that artificial approach gets better result that ARIMA. That is because there is a nonlinear pattern that cannot be captured by the ARIMA model. The DSARIMA-SVR method has better accuracy because the RMSE values obtained in the testing data are smaller compared to SVR as a nonlinear process. An interesting thing to discuss is that using significant PACF lag as an input selection is not complicated especially the accuracy results obtained are quite good, especially in testing data. However, compared to the input selection using ARIMA is far better than PACF because the input lag that is used is relevant to the data pattern. The use of PACF data lag as a variable will be at risk in cases of overfitting. Models will tend to predict good data used as modeling but not good for forecasting. A plot comparison between the three methods is presented in Figure 12. The black line shows the original data, the red line estimation results of the DSARIMA, the blue line estimation results of the SVR, while the green line shows estimation results of the DSARIMA-SVR.

IV. CONCLUSION

Short term electricity load data in East Java area is known to have Double Seasonal ARIMA models $ARIMA([28,31,47],1,[11,12,32])(0,1,1)^{48}(0,1,2)^{336}$which if the model is described there are 111 significant lags. Whereas there are 180 significant lags based on stationary PACF plots. Method performance based on the artificial intelligence approach has better accuracy results compared to the statistical mathematical approach. The best input selection is based on the significant PACF model of ARIMA model, while the use of PACF lag in the linear model will be at risk of overfitting.

REFERENCES

[1] D. Wiyanti and R. Pulungan, "Peramalan Deret Waktu Menggunakan Model Fungsi Basis Radial (RBF) dan Auto Regressive Integrated Moving Average (ARIMA)," *Jurnal MIPA,* vol. 35, no. 2, pp. 175-182, 2012.

[2] M. Amin and M. Hoque, "Comparison of ARIMA and SVM for Short-Term Load Forecasting," pp. 205-210, 2019.

[3] F. Fahmi and H. Sofyan, "2017," *International Conference on Electrical Engineering and Informatics (ICELTICs)* , pp. 97-102, 2017.

[4] D. Riyani, D. D. Prastyo and Suhartono, "Input Selection in Support Vector Regression for Univariate Time Series Forecasting," *AIP Conference Proceedings* , vol. 020105, pp. 1-6, 2019.

[5] Khusna Hidayatul and Suhartono, "Pendekatan Percentila Error Bootstrap pada Model Double Seasonal Holt-Winters, Double Seasonal ARIMA, dan Naive untuk Peramalan Beban Listrik Jangka Pendek Area Jawa Timur-Bali," *JURNAL SAINS DAN SENI ITS,* vol. 4, no. 1, pp. 2337-3520, 2015.