Chapter 11

Electricity Load Forecasting Using Data Mining Technique

Intan Azmira binti Wan Abdul Razak, Shah bin Majid,
Mohd Shahrieel bin Mohd. Aras and Arfah binti Ahmad

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/48657

1. Introduction

Accurate load forecasting is become crucial in power system operation and planning [1-3]; both for deregulated and regulated electricity market. Electric load forecasting can be divided into three categories that are short term load forecasting, medium term load forecasting and long term load forecasting. The short term load forecasting predicts the load demand from one day to several weeks. It helps to estimate load flows that can prevent overloading and hence lead to more economic and secure power system . The medium term load forecasting predicts the load demand from a month to several years that provides information for power system planning and operations. The long term load forecasting predicts the load demand from a year up to twenty years and it is mainly for power system planning [1].

A variety of methods including neural networks [2], time series [1], hybrid method [3,4] and fuzzy logic [5] have been developed for load forecasting. The time series techniques have been widely used because load behaviour can be analyzed in a time series signal with hourly, daily, weekly, and seasonal periodicities. Besides, it is able to deal with non stationary data to reflect the variation of variables [4].

However, for a huge power system covering large geographical area such as Peninsular Malaysia, a single forecasting model for the entire Malaysia would not satisfy the forecasting accuracy; due to the load and weather diversity[6]. Thus, this research will cater these conditions whereby five models of SARIMA (Seasonal ARIMA) Time Series [7,8] were developed for five day types.

2. Problem statement

Electric load forecasting is very important in power system operation such as during start-up and shut-down schedules of generating units as well as for overhaul planning [2] and spot market energy pricing [4]. In normal working condition, system generating capacity should meet load requirement to avoid adding generating units and importing power from the neighbouring network [9].

This research applied ARIMA time series approach to forecast future load in Peninsular Malaysia. Time series method that was introduced by Box and Jenkins is a sequence of data points that measured typically at successive times and time intervals [10].

3. Data mining with SARIMA time series

Before proceeding the forecasting process, load data need to be analyzed. Table 1 shows the average maximum and minimum demand, average energy and peak hour per day within a week. From the analysis, it can be concluded that the load characteristic among the days in a week is different. The average energy for Monday is slightly lower compared to Tuesday, Wednesday and Thursday. On the other hand, the average energy for those three days is fairly around 255MWh so that they can be clustered in a category. The average energy for Friday shows the lowest value within weekdays while the energy used for weekend is much lower than the consumption on weekdays. Comparing energy consumed on weekend, there is more consumption on Saturday rather than Sunday. Hence, the forecast will be conducted based on five day types that are:

Type 1 : Monday
Type 2 : Tuesday, Wednesday, Thursday
Type 3 : Friday
Type 4 : Saturday
Type 5 : Sunday

Day	Average Maximum Demand (MW)	Average Minimum Demand (MW)	Average Energy (MWh)	Peak Hour
Monday	12 442	7 842	249.06	3.00 – 4.30 pm
Tuesday	12 484	8 526	254.89	
Wednesday	12 508	8 565	255.95	
Thursday	12 436	8 543	255.03	
Friday	11 884	8 463	246.23	
Saturday	10 718	8 122	227.26	11.30am – 12.00pm
Sunday	10 116	7 605	211.01	8.00 – 9.00 pm

Table 1. Load data analysis within a week

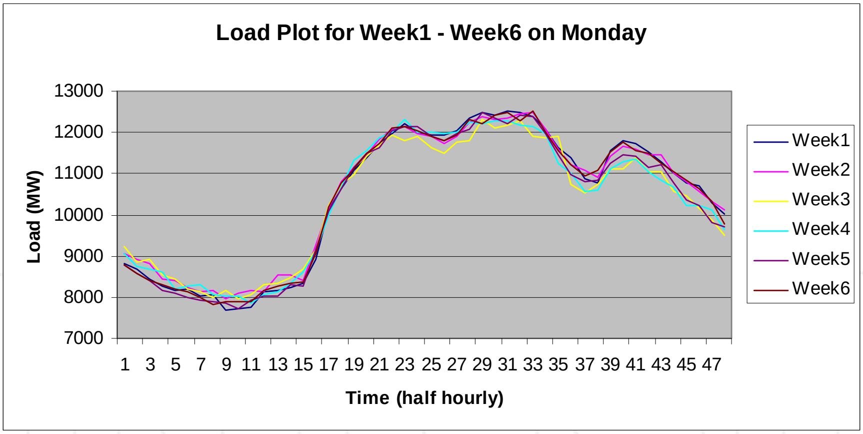


Figure 1. Load Plot for Monday

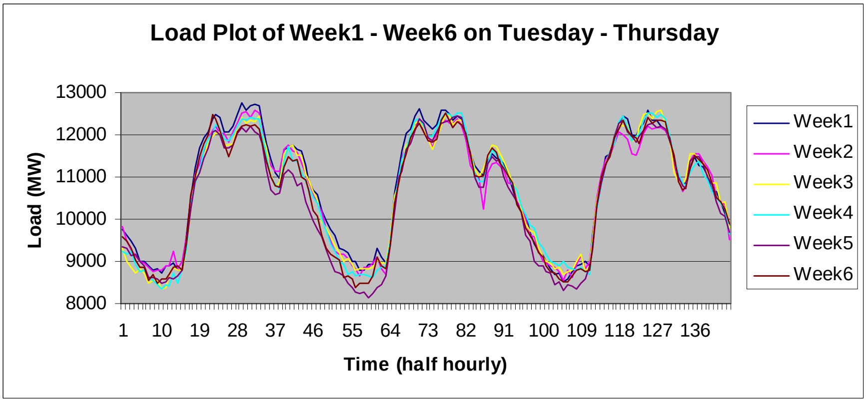


Figure 2. Load Plot for Tuesday, Wednesday and Thursday

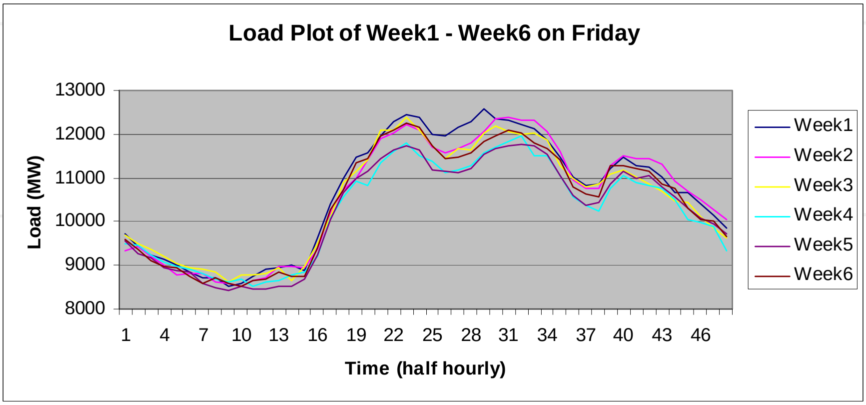


Figure 3. Load Plot for Friday

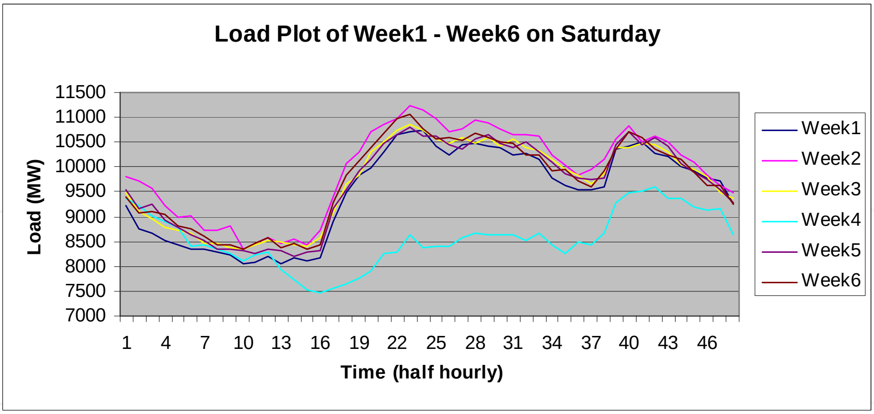


Figure 4. Load Plot for Saturday

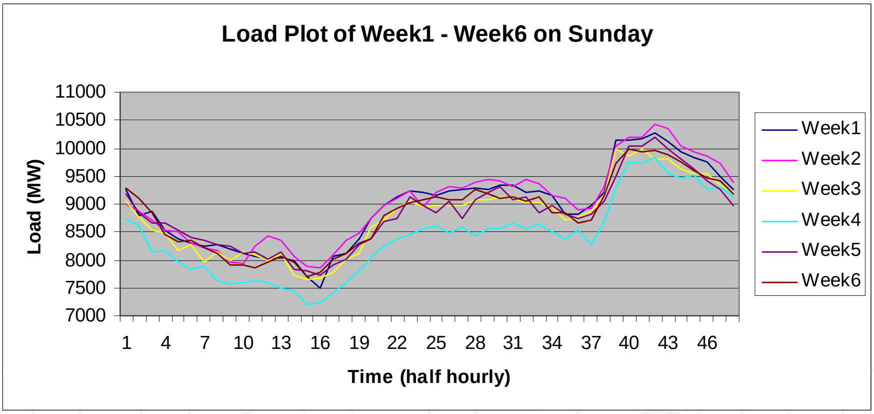


Figure 5. Load Plot for Sunday

Apart from that, load plot for each day types can be observed as in Figure 1-5. Their characteristic for certain time interval can be simplified as in Table 2 (a) and (b). Referring Table 2(a) for weekday, load consumption is decreasing from time 20.00 till 00.00 and 00.30 till 04.00 where people are having some rest or sleeping at night. However, starting 04.00 till 17.00 the load consumption is increasing because people start using home appliances and go to work. The load consumption for 17.00 till 19.00 shows slight decrease as people come back to home. The next an hour show the load consumption increasing where people spend some time watching television or having a dinner. However, there are bit differences of people activities during weekend that affect load consumption.

Time	Monday	Tuesday – Thursday	Friday
00.30 – 04.00	9 100 – 8 000	9 800 – 8 800	9 300 – 8 500
04.00 – 17.00	8 000 – 12 500	8 800 – 12 600	8 500 – 12 300
17.00 – 19.00	12 500 – 11 100	12 600 – 11 100	12 300 – 10 800
19.00 – 20.00	11 000 – 11 700	11 100 – 11 700	10 800 – 11 600
20.00 – 00.00	11 700 – 10 100	11 700 – 10 200	11 600 – 10 000

(a) Weekday

Time	Saturday	Sunday
00.30 - 08.00	9 800 – 8 400	9 000 – 7 900
08.00 – 12.00	8 400 – 11 300	08.00 – 16.00: 7 900 – 9 400 16.00 – 18.00: 9 400 – 8 900
12.00 – 18.00	11 300 – 9 800	
18.00 – 21.00	9 800 – 10 600	8 900 – 10 400
21.00 – 00.00	10 600 – 9 500	10 400 – 9 400

(b) Weekend

Table 2. Load consumption per day (MW)

Five models of SARIMA were developed in Minitab which represents the five day types. ARIMA; Autoregressive Integrated Moving Average involves the filtering steps in constructing the ARIMA model until only random noise remains. ARIMA model can be classified as seasonal or non-seasonal model. The series with seasonal repeating pattern is categorized as seasonal model or seasonal ARIMA (SARIMA) while the series with random series or no seasonal repeating trend is called as non-seasonal pattern. At least four or five seasons of the data are needed to fit the SARIMA model. Instead, ARIMA modeling identifies an acceptable model by some steps which are differencing, autocorrelation and partial autocorrelation functions. A non-seasonal ARIMA model is known as an ARIMA (p, d, q) model while a seasonal ARIMA model is named as ARIMA (P, D, Q) model where P or P is the number of autoregressive term (AR), d or D is the number of non-seasonal differences and q or Q is the number of lagged forecast errors in the prediction equation (MA). Appropriate ARIMA model is determined by identifying the p, d, q and P, D, Q parameters [10].

During modelling an ARIMA, the first step is determining whether the series has a trend or not. Trend analysis determines the seasonality and stationary. The second step is determining period for the seasonal model; by plotting spectral plot in MATLAB or ACF from Minitab. Usually the period is already known and it can be seen from ACF but spectral plot will prove that assumption. The third step involved is data transformation (if any) by Box-Cox plot; depending on the value of λ as shown in Table 3.

Value of λ	Transformation
-1.0	$\frac{1}{x_t}$
-0.5	$\frac{1}{\sqrt{x_t}}$
0	$\ln x_t$
0.5	$\sqrt{x_t}$
1.0	X_t

Table 3. Box-Cox Transformation

The last step is identifying the p, d, q and P, D, Q parameters. It started by determining the order of differencing needed to stationarize the series [10]. Normally the lowest order of differencing leads time series to fluctuates around a well-defined mean value and the spikes of ACF and PACF decays fairly rapidly to zero. After chosen appropriate order of differencing, AR and MA terms are then identified to determine whether the AR and MA terms are needed to correct any autocorrelation that remains in the differenced series.

Apart from that, the best fit of the model must meet these specifications:

a. -1.96 ≥ t-value ≥ 1.96
b. The lowest standard deviation
c. Chi-Square at Lag-12 is acceptable
d. -1 ≤ Parameter's coefficient ≤ 1

Some equations related to ARIMA model are shows in (1) to (4).

The order of d can be expressed in terms of the backshift operator B as:

$$\nabla^d = (1 - B)^d \tag{1}$$

The seasonal backshift operator;

$$B^S z_t = z_{t-S} \tag{2}$$

Where
S = seasonal period,
Z_t = transformed data at time t

The seasonal difference operator;

$$\nabla_S^D = (1 - B^S)^D \tag{3}$$

Combining (1) and (3) yields:

$$Y_t = (1 - B)^d (1 - B^S)^D z_t \tag{4}$$

Where Y_t = differenced data at time t

4. SARIMA modelling

4.1. ARIMA model for monday

The load data on Monday for six weeks had been plotted by trend analysis. Figure 6 shows that the data is seasonal and non-stationary so the period of the data must be identified. It can be done by plotting spectral plot in MATLAB as shown in Figure 7.

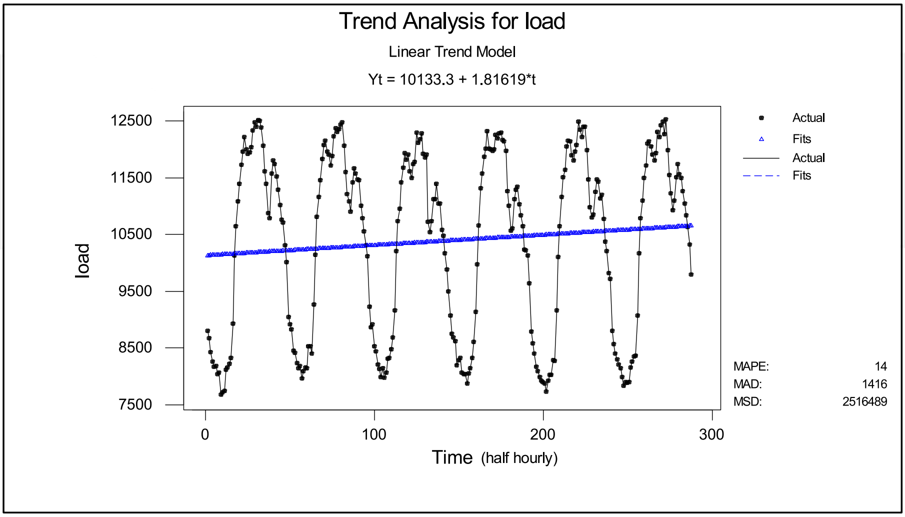


Figure 6. Trend analysis for Monday

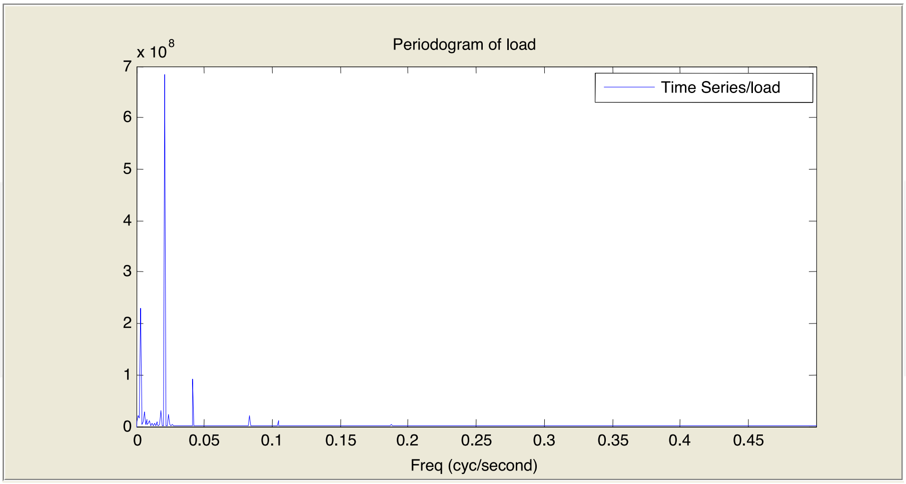


Figure 7. Spectral plot

Figure 7 shows that the graph had no aliasing or crossing on x-axis; meaning that the data is suitable for an analysis. The period is determined by;

$$T = 1/f \qquad (5)$$

Where T = period, and f = frequency

From Figure 8, the frequency was 0.0208 thus the period was approximately 48. This value was determined based on the half hourly load and is valid for all day types. Since the data was not stationary, the actual data must be transformed; depending on the value of λ.

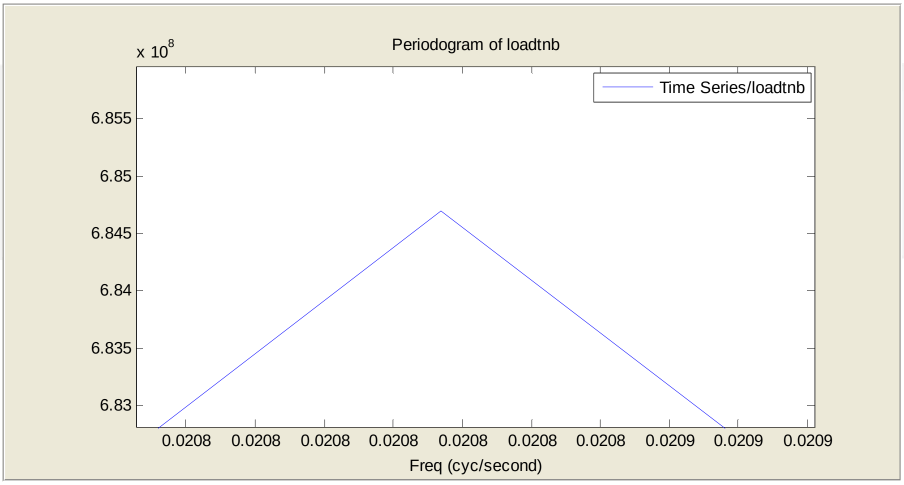


Figure 8. Enlarged view of spectral plot

Figure 9 shows the Box-Cox Plot for Monday where the value of λ = 0.562 so the rounded value is 0.5. Hence the actual data was transformed to $\sqrt{X_t}$.

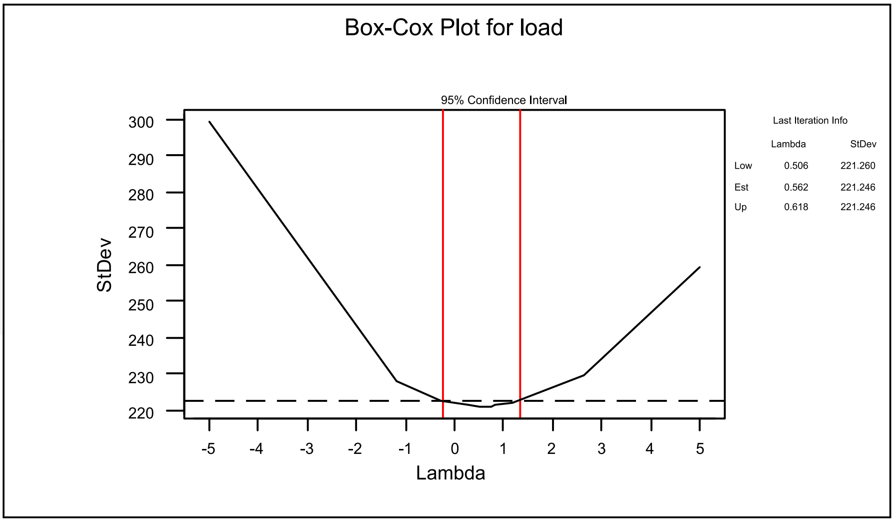
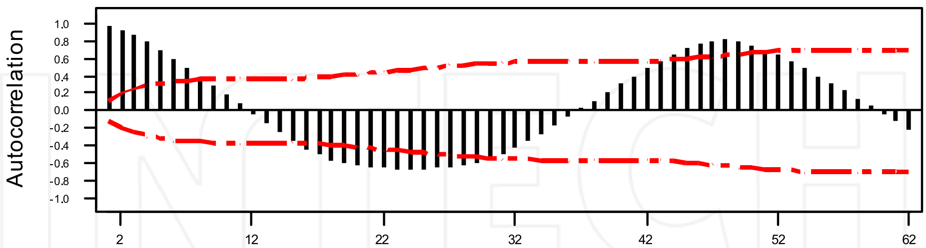


Figure 9. Box-Cox plot for Monday

Figure 10 show bad ACF (sine-cosines' phenomenon) and PACF when all parameters are zero. It is important to ensure that all the spikes are within the boundary to be a stationary model. Then the ARIMA parameters were identified and the selected model was ARIMA $(2,1,1)(0,1,1)_{48}$.

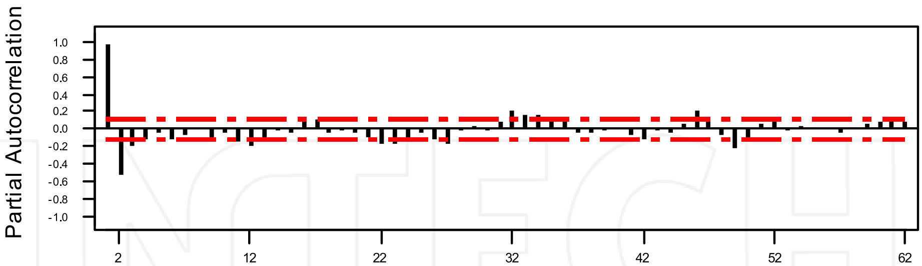
Autocorrelation Function for d0 D0



Lag	Corr	T	LBQ	Lag	Corr	T	LBQ	Lag	Corr	T	LBQ	Lag	Corr	T	LBQ	Lag	Corr	T	LBQ
1	0.98	16.57	277.31	16	-0.45	-2.36	1457.14	31	-0.52	-1.85	3359.97	46	0.76	2.47	4298.02	61	-0.13	-0.37	5794.05
2	0.93	9.24	528.78	17	-0.52	-2.68	1541.13	32	-0.44	-1.57	3424.26	47	0.80	2.54	4518.19	62	-0.22	-0.62	5812.18
3	0.86	6.80	746.52	18	-0.58	-2.91	1645.09	33	-0.36	-1.27	3466.91	48	0.81	2.52	4746.75				
4	0.78	5.37	926.57	19	-0.62	-3.03	1765.22	34	-0.27	-0.94	3490.73	49	0.79	2.42	4967.23				
5	0.69	4.35	1068.78	20	-0.65	-3.08	1897.12	35	-0.17	-0.61	3500.69	50	0.75	2.25	5166.45				
6	0.60	3.52	1174.48	21	-0.67	-3.06	2036.54	36	-0.08	-0.27	3502.61	51	0.70	2.05	5338.03				
7	0.49	2.79	1247.06	22	-0.68	-3.00	2180.29	37	0.02	0.07	3502.74	52	0.63	1.83	5479.66				
8	0.39	2.14	1292.28	23	-0.68	-2.93	2326.72	38	0.11	0.39	3507.03	53	0.56	1.60	5590.48				
9	0.28	1.53	1316.26	24	-0.68	-2.86	2474.92	39	0.21	0.72	3521.35	54	0.48	1.36	5672.23				
10	0.18	0.95	1325.67	25	-0.68	-2.78	2622.91	40	0.30	1.04	3551.77	55	0.39	1.11	5727.82				
11	0.07	0.38	1327.16	26	-0.68	-2.68	2768.93	41	0.39	1.36	3603.74	56	0.31	0.86	5761.74				
12	-0.04	-0.22	1327.65	27	-0.67	-2.58	2911.02	42	0.48	1.65	3681.80	57	0.22	0.62	5779.39				
13	-0.15	-0.81	1334.66	28	-0.65	-2.45	3045.37	43	0.56	1.92	3789.80	58	0.13	0.38	5785.97				
14	-0.26	-1.39	1355.16	29	-0.62	-2.28	3167.60	44	0.64	2.14	3929.11	59	0.05	0.13	5786.77				
15	-0.36	-1.92	1395.02	30	-0.57	-2.08	3273.68	45	0.70	2.33	4099.26	60	-0.04	-0.12	5787.45				

ACF Plot for ARIMA $(0,0,0)(0,0,0)_{48}$

Partial Autocorrelation Function for d0 D0



Lag	PAC	T	Lag	PAC	T	Lag	PAC	T	Lag	PAC	T	Lag	PAC	T
1	0.98	16.57	16	0.11	1.79	31	0.08	1.29	46	0.20	3.33	61	0.07	1.27
2	-0.53	-9.01	17	0.11	1.95	32	0.20	3.47	47	0.12	2.02	62	0.08	1.31
3	-0.19	-3.29	18	-0.06	-0.97	33	0.16	2.64	48	-0.08	-1.35			
4	-0.12	-2.05	19	-0.02	-0.34	34	0.15	2.52	49	-0.23	-3.93			
5	-0.05	-0.78	20	-0.06	-1.01	35	0.08	1.39	50	-0.15	-2.50			
6	-0.14	-2.37	21	-0.10	-1.64	36	0.08	1.33	51	0.05	0.82			
7	-0.09	-1.51	22	-0.18	-2.97	37	-0.05	-0.81	52	0.12	2.04			
8	0.01	0.21	23	-0.17	-2.90	38	-0.06	-1.00	53	-0.03	-0.54			
9	-0.10	-1.65	24	-0.16	-2.77	39	-0.03	-0.57	54	0.02	0.28			
10	-0.06	-1.08	25	-0.06	-1.00	40	0.00	0.05	55	-0.01	-0.12			
11	-0.15	-2.58	26	-0.12	-1.97	41	-0.07	-1.20	56	-0.01	-0.17			
12	-0.21	-3.60	27	-0.17	-2.97	42	-0.13	-2.28	57	-0.05	-0.86			
13	-0.09	-1.53	28	-0.03	-0.46	43	-0.02	-0.38	58	-0.01	-0.21			
14	-0.02	-0.28	29	0.02	0.37	44	-0.06	-1.03	59	0.05	0.82			
15	-0.06	-1.00	30	-0.02	-0.27	45	0.05	0.84	60	0.06	1.09			

Figure 10. ACF and PACF for Monday

Figure 11-12 shows good ACF and PACF where the spikes decay fairly rapidly to zero. There was strong autocorrelation at lag-48 that shows the period of the data. All these steps were repeated for other day types.

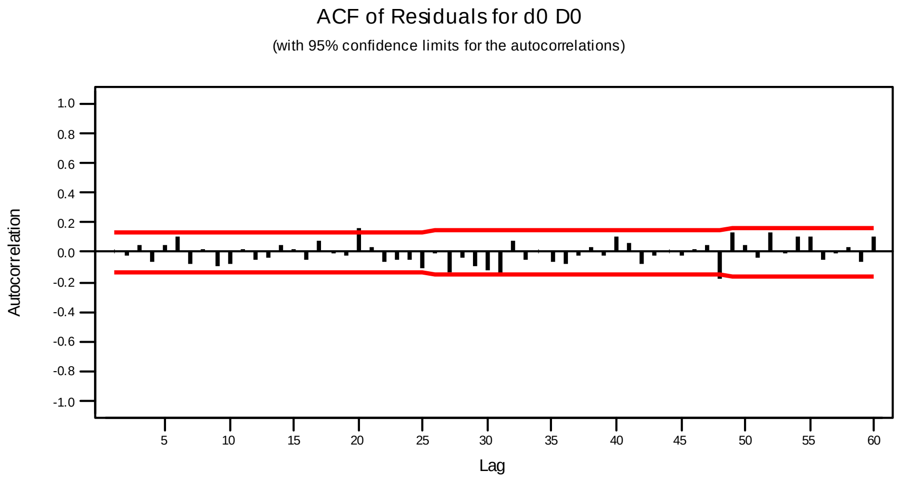


Figure 11. ACF Plot for ARIMA $(2,1,1)(0,1,1)_{48}$ on Monday

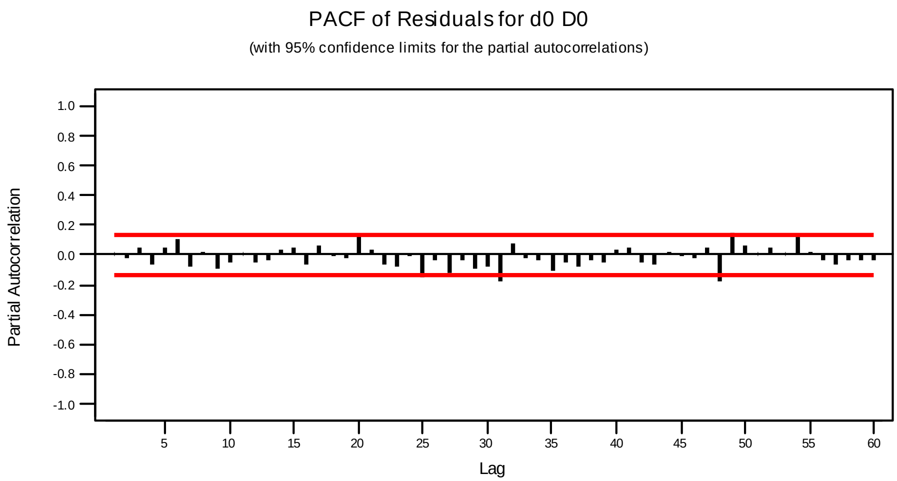


Figure 12. PACF Plot for ARIMA $(2,1,1)(0,1,1)_{48}$ on Monday

4.2. ARIMA model for Tuesday, Wednesday and Thursday

The steps taken for modelling ARIMA for this second model were repeated as for Monday. The trend analysis for Tuesday, Wednesday and Thursday was plotted followed by Box-Cox plot. The value of λ is 0.45 thus the rounded value is 0.5. After the data had been transformed to $\sqrt{X_t}$, the fitted ARIMA model was ARIMA $(1,1,1)(0,1,1)_{48}$.

Figure 13 and 14 show good ACF and PACF for selected ARIMA model where less spikes were found outside the boundary.

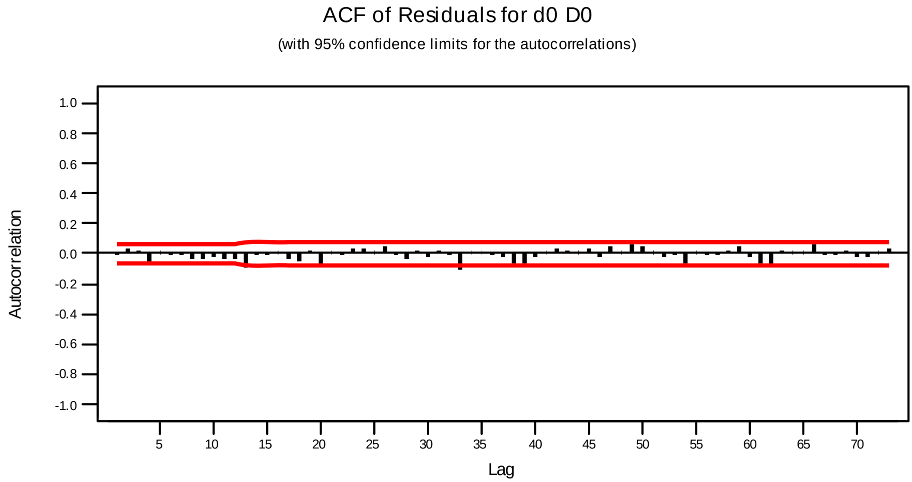


Figure 13. ACF Plot for Tuesday, Wednesday and Thursday

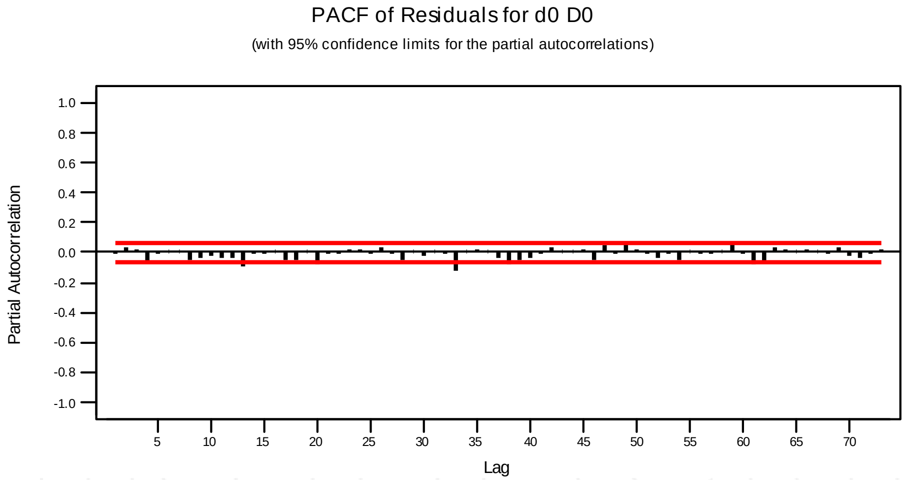


Figure 14. PACF Plot for Tuesday, Wednesday and Thursday

4.3. ARIMA model for Friday

The steps taken for modelling ARIMA for this third model were repeated as for two previous models. The trend analysis for Friday was plotted followed by Box-Cox plot. The trend analysis showed that the data is seasonal and non-stationary thus it must be transformed. Box-Cox plot showed that the value of λ is -0.112 and the rounded value is 0. The data was transformed to $ln\ X_t$ and the selected model is ARIMA $(0,1,1)(0,1,1)_{48}$.

Figure 15 and 16 show good ACF and PACF for Friday model with no spikes outside the boundary.

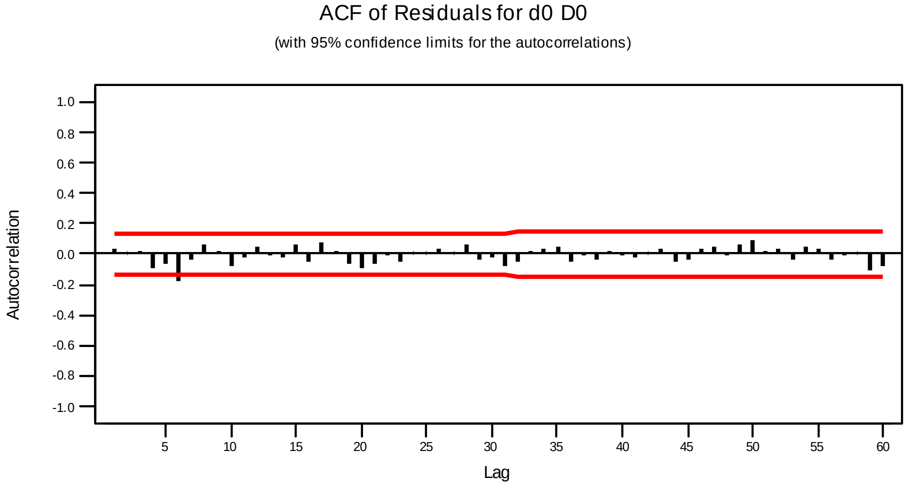


Figure 15. ACF Plot for Friday

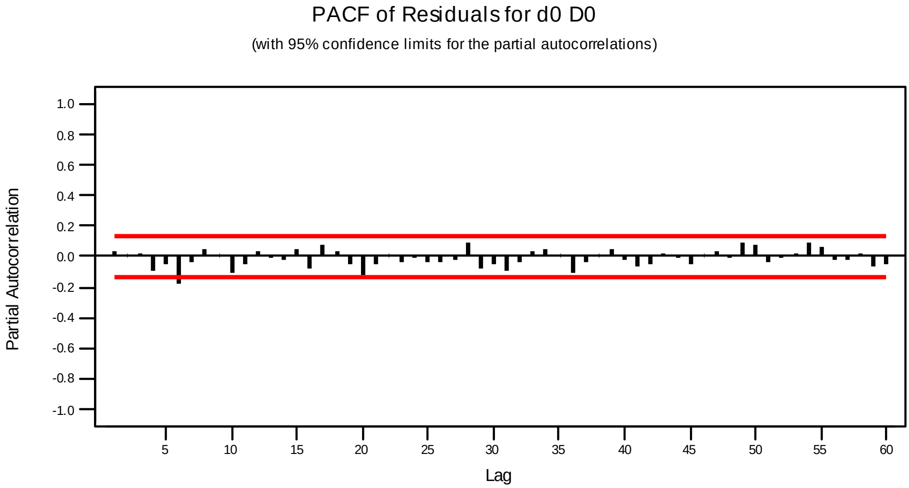


Figure 16. PACF Plot for Friday

4.4. ARIMA model for Saturday

The steps taken for modelling ARIMA for this fourth model were repeated as for three previous models. The trend analysis for Saturday was plotted followed by Box-Cox plot. The trend analysis showed that the data is seasonal and non-stationary thus it must be transformed. Box-Cox plot showed that the value of λ is 0.113 and the rounded value is 0.

After the actual data had been transformed to $ln\ X_t$, the selected model was ARIMA (2,1,1)(0,1,1)$_{48}$.

Figure 17-18 show good ACF and PACF for Saturday with ARIMA model selected.

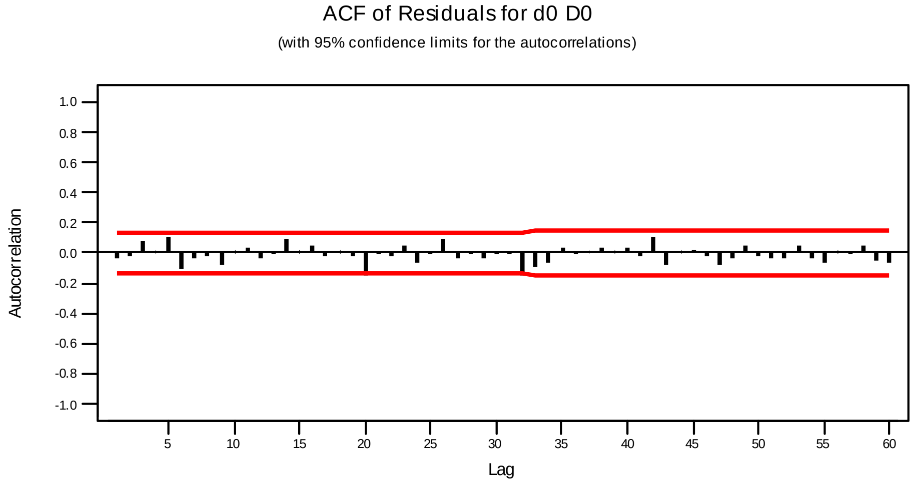


Figure 17. ACF Plot for Saturday

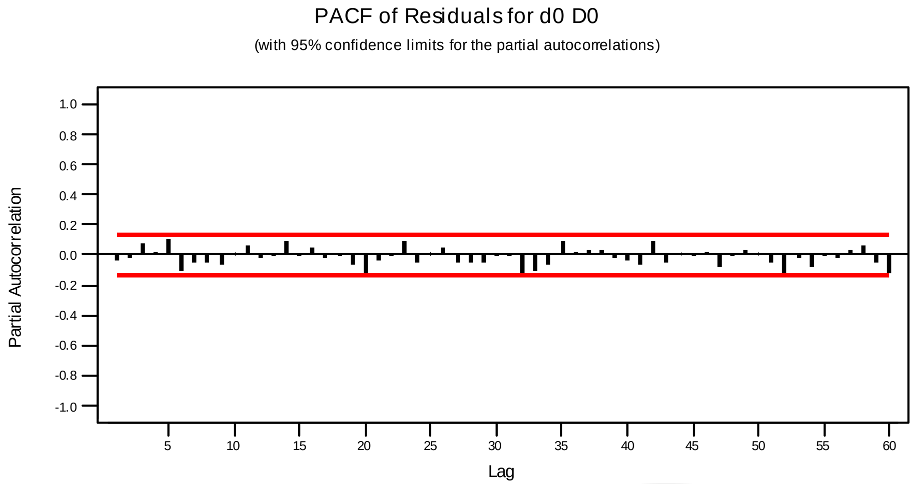


Figure 18. PACF Plot for Saturday

4.5. ARIMA model for Sunday

The steps taken for modelling ARIMA for this fifth model were repeated as for four previous models. The trend analysis for Sunday was plotted followed by Box-Cox plot. The trend analysis showed that the data is seasonal and non-stationary thus it must be transformed. Box-Cox plot showed that the value of λ is 0.225 and the rounded value is 0. After the actual data had been transformed to $ln\ X_t$, the selected model was ARIMA $(0,1,1)(0,1,1)_{48}$.

Figure 19-20 show good ACF and PACF for the fitted model. The plots show less spikes outside the boundary after a differencing and good selection of p, P, q and Q.

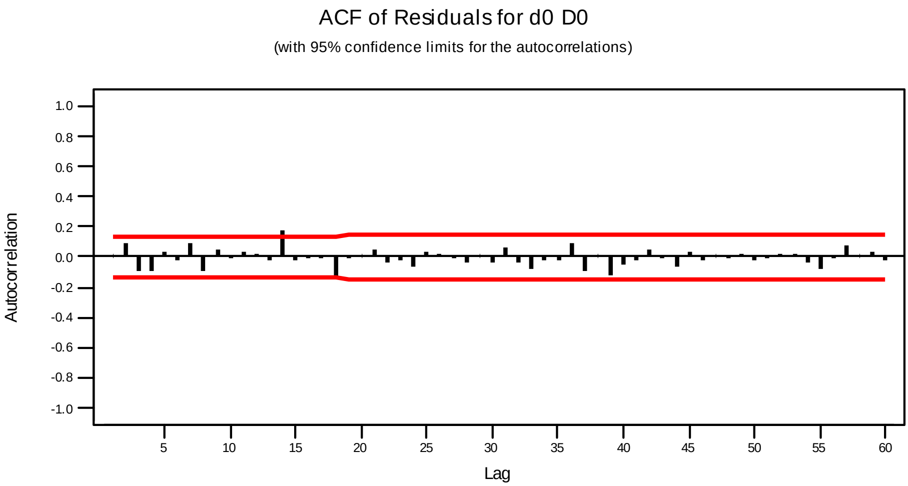


Figure 19. ACF Plot for Sunday

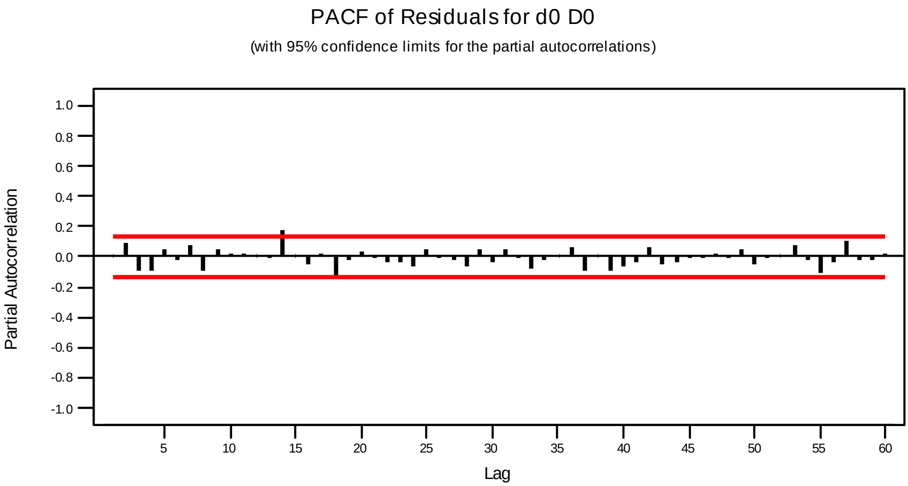


Figure 20. PACF Plot for Sunday

5. Result and analysis

The forecasting was held for 48 points that represent a day ahead for each day types. Table 4-8 show model specifications for all day types. Referring to t-values for all models, they satisfied the condition -1.96 ≥ t-value ≥ 1.96. Besides, good standard deviations shown for all models as well as Chi-Square at Lag-12 are also acceptable. The parameters' coefficients also fulfil the condition within the range of -1 and 1.

Parameters' Coefficient	t-value	Standard Deviation	Chi-Square at Lag-12	DF
AR 1 -0.3879 AR 2 -0.2675 MA 1 0.3717 SMA 48 08382	-2.98 -2.86 2.86 13.87	82.9448	10.9	8

Table 4. Model Specification for Monday

Parameters' Coefficient	t-value	Standard Deviation	Chi-Square at Lag-12	DF
AR 1 0.1962 MA 1 0.6848 SMA 48 0.9120	3.23 15.20 41.80	226.319	9.8	9

Table 5. Model Specification for Tuesday, Wednesday and Thursday

Parameters' Coefficient	t-value	Standard Deviation	Chi-Square at Lag-12	DF
MA 1 0.4878 SMA 48 0.6243	8.73 11.01	0.0355914	14.4	10

Table 6. Model Specification for Friday

Parameters' Coefficient	t-value	Standard Deviation	Chi-Square at Lag-12	DF
AR 1 0.4083 AR 2 0.3241 MA 1 0.6491 SMA 48 0.7511	2.33 5.16 3.59 13.48	0.0632711	10.1	8

Table 7. Model Specification for Saturday

Parameters' Coefficient	t-value	Standard Deviation	Chi-Square at Lag-12	DF
MA 1 0.5746 SMA 48 0.6823	10.81 11.77	0.0596651	11.9	10

Table 8. Model Specification for Sunday

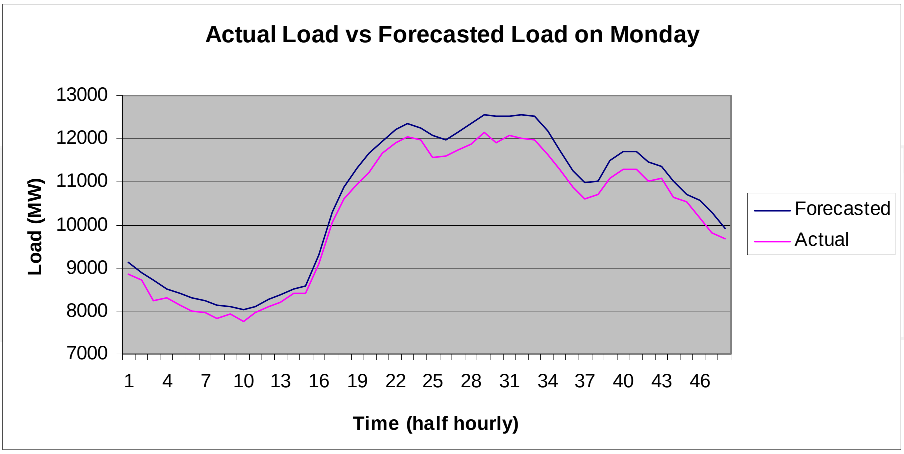


Figure 21. Actual load vs. forecasted load on Monday

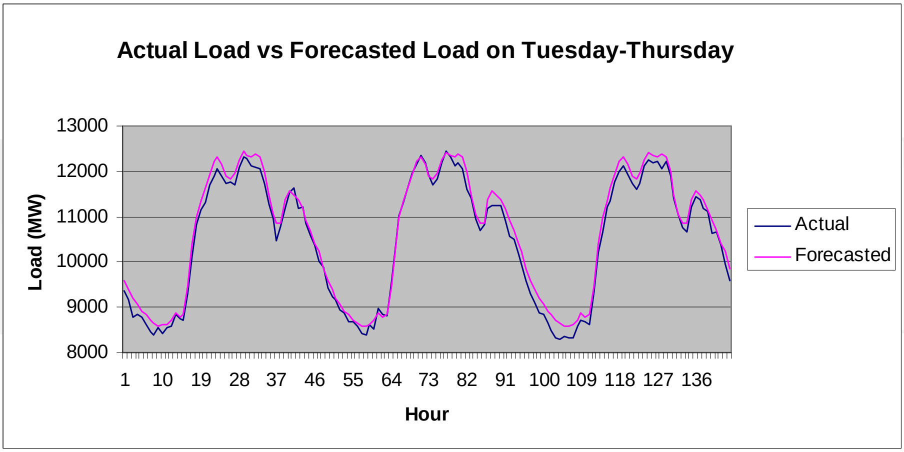


Figure 22. Actual load vs. forecasted load on Tuesday, Wednesday and Thursday

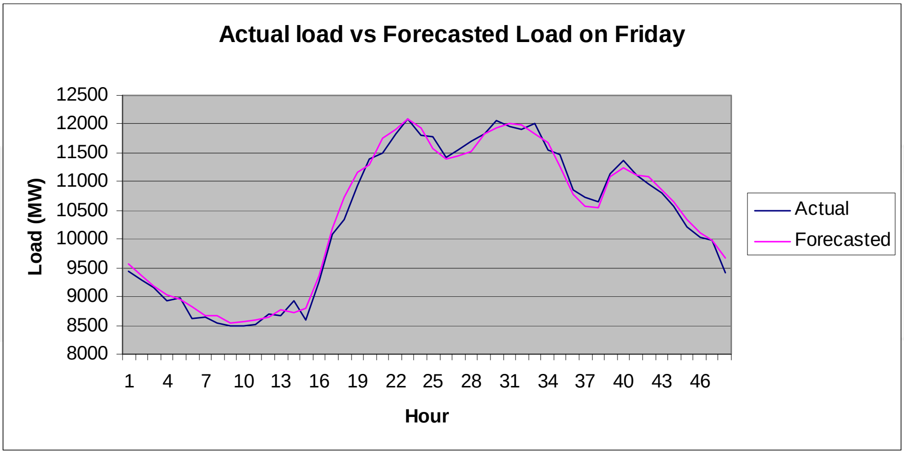


Figure 23. Actual load vs. forecasted load on Friday

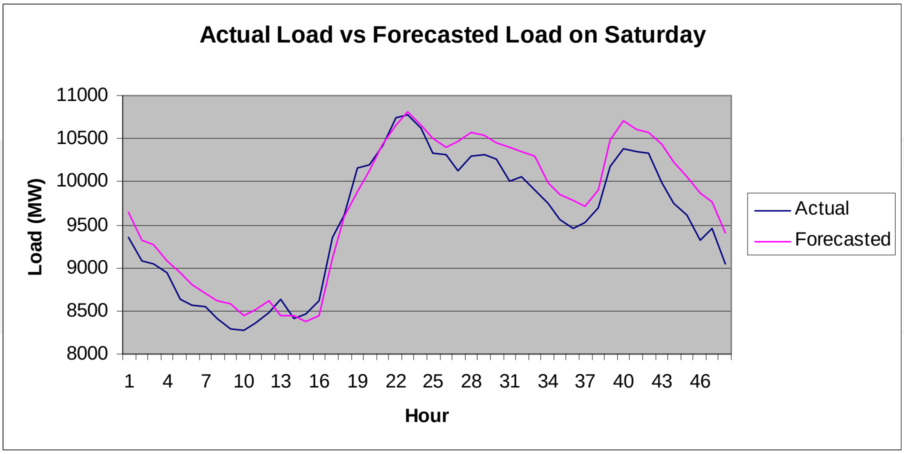


Figure 24. Actual load vs. forecasted load on Saturday

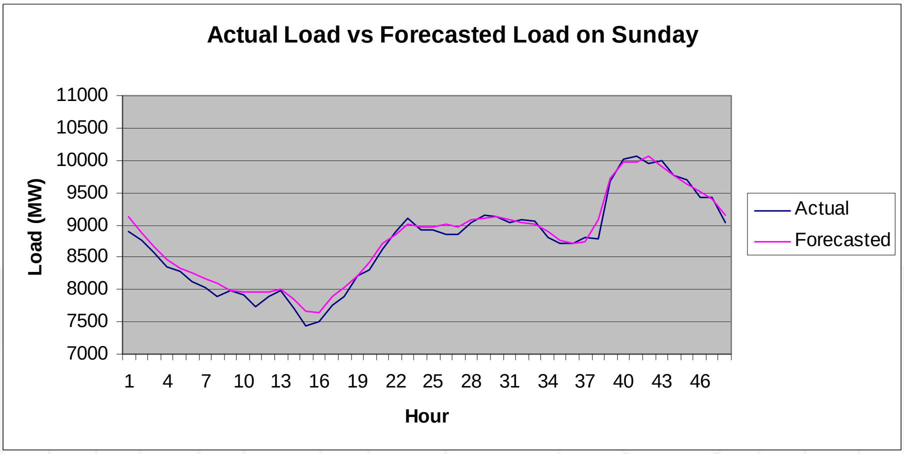


Figure 25. Actual load vs. forecasted load on Sunday

Figure 21-25 show the plots of forecasted load vs. actual load. The forecasted load plot are seems to be close as actual load plot. Mean Absolute Percentage Error (MAPE) for all day types were calculated as in (7):

$$\text{MAPE (\%)} = \frac{1}{N}\left[\frac{\left|Z'_t - x_t\right|}{x_t}\right] \times 100\% \qquad (6)$$

Where $Z't$ = Forecasted Load,
Xt = Actual Load
N = Forecasting number

Table 9 shows the ARIMA models and their MAPEs for all day types. It can be seen that the difference order (d and D) for all models is 1which is the lowest order and the best selection. The result is considered as accurate when the MAPE is lower than 1.5% as shown for Tuesday –Thursday, Friday and Sunday models. The higher MAPE for Monday and Saturday models may caused by load or weather fluctuation.

Day	ARIMA Model	MAPE
Monday	$(2,1,1)(0,1,1)_{48}$	3.26064%
Tuesday -Thursday	$(1,1,1)(0,1,1)_{48}$	1.62094%
Friday	$(0,1,1)(0,1,1)_{48}$	1.11833 %
Saturday	$(2,1,1)(0,1,1)_{48}$	2.41944 %
Sunday	$(0,1,1)(0,1,1)_{48}$	1.07158 %

Table 9. Forecasting result for all day types

6. Conclusion

From the data analysis, load data was clustered to five day types and hence five models of SARIMA are designed. Each forecasting model is developed for each day except for Tuesday, Wednesday and Thursday which clustered as a model. Forecasting method is held by Time Series - SARIMA where it is one of data mining methods which require enough experience on determining its parameter (p,d,q,P,D,Q). Sometimes it is needs for trial and error during identifying the parameters. However, the MAPEs obtained for each day types were ranging from 1% to 3%. This new approach had improved the accuracy of forecasting compared to traditional approach of ARIMA that use only a model for all days in a week.

7. Further research

Additional input variables can be included in the forecasting process such as weather data, customers' classes and event day; instead of only the load data. Besides, other methods may be implemented such as Neural Network, Fuzzy Logic as well as hybrid method [11].

Author details

Intan Azmira binti Wan Abdul Razak*,
Mohd Shahrieel bin Mohd. Aras and Arfah binti Ahmad
Faculty of Electrical Engineering, UTeM, Malacca, Malaysia

Shah bin Majid
Faculty of Electrical Engineering, UTM, Johor, Malaysia

Acknowledgement

I wish to express my gratitude to honorable University (**Universiti Teknikal Malaysia Melaka- UTeM**) especially to Faculty of Electrical Engineering for give the financial as well as moral support. My special thanks also fall to Mr. Fuad Jamaluddin from Utility of Malaysia for his valuable advice and help during completion of this research. Also to all my research members that give full commitments and cooperation.

8. References

[1] I. Azmira, A. Razak, S. Majid, and H.A. Rahman, "Short Term Load Forecasting Using Data Mining Technique," *Energy Conversion and Management*, 2008, pp. 139-142.
[2] "Application of Pattern Recognition and Artificial Neural Network to Load Forecasting in Electric Power System," *Pattern Recognition*, 2007.
[3] P. Qingle and Z. Min, "Very Short-Term Load Forecasting Based on Neural Network and Rough Set," *Network*, 2010, pp. 1132-1135.

* Corresponding Author

[4] J.C. Hwang and C.S. Chen, "Customer Short Term Load Forecasting by Using Arima Transfer Function Model," *Electrical Engineering*, pp. 317-322.

[5] B. Ye, N.N. Yan, C.X. Guo, and Y.J. Cao, "Identification of Fuzzy Model for Short-Term," *Evolution*, 2006, pp. 1-8.

[6] S. Fan, Y.-kang Wu, W.-jen Lee, and C.-yin Lee, "Different Geographical Distributed Loads," *Systems Research*, 2011, pp. 1-8.

[7] Y.H. Kareem and A.R. Majeed, "Sulaimany Governorate Using SARIMA .," *Building*, 2006, pp. 1-5.

[8] "Robust Estimation of Sarima Models : Application to Short-Term Load Forecasting Yacine Chakhchoukh , Patrick Panciatici Versailles , France," *Signal Processing*, 2009, pp. 77-80.

[9] J.K. Basu, D. Bhattacharyya, and T.-hoon Kim, "Use of Artificial Neural Network in Pattern Recognition," *Engineering*, vol. 4, 2010, pp. 23-34.

[10] H.L. Willis, *Power Distribution Planning Reference Book*, North Carolina,USA: Marcel Dekker, Inc., 2004.

[11] L. Xuemei, D. Lixing, and D. Yuyuan, "Hybrid Support Vector Machine and ARIMA Model in Building Cooling Prediction," *Built Environment*, 2010.