

Short-term Electricity Load Forecasting in Thailand: an Analysis on Different Input Variables

Su Wutyi Hnin¹ and Chawalit Jeenanunta¹

¹School of Management Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand, 12120
E-mail:d6022300203@g.siit.tu.ac.th, chawalit@g.siit.tu.ac.th

Abstract. This paper suggests a support vector regression model to make short-term load forecasting in Thailand by different training inputs. The primary objective of this paper is to describe the importance of data pre-processing and the external factors for accurate forecasting. The Electricity Generating Authority of Thailand (EGAT) provides the half-hourly electricity load demand. For numerical analysis, a dataset of net peak load of Thailand for a period of weeks from January 2016 to December 2017 is selected. The historical load demand is filtered for each day by Local regression filtering technique. After filtering the data, the effectiveness of input variables is important for accurate performance. Mean absolute percentage error (MAPE) is used to evaluate the model performance. By comparing the three models, model which considerate the temperature and seasonal factors enhances the model performance.

1. Introduction

According to the economy growth, the improved life-style, and the usage of electricity demand is increasing year by year. As a consequence, load forecasting plays an important role in power energy management. There are three types of electricity load forecasting in order to time horizon. Short-term load forecasting is to forecast one hour to a week and forecasting a month up to a year is mid-term load forecasting as well as long-term is over a year. Basically, long-term load forecasting (LTLF) and mid-term load forecasting (MTLF) intend for planning new transmission line, and power station. Short-term load forecasting (STLF) plays a key factor in planning and maintenance, load flow analysis, and operating strategies in power system [1]. The main target in load forecasting has been on STLF because it is a vital thing for the day to day operation of the power utility system. There are three basic steps before forecasting: exploring the data, data pre-processing, and developing the forecasting model. The electricity load demand is non-linear and correlation with many exogenous variables, involving the weather conditions, seasonality, trend and calendar effect. The data must be accurate. However, the data collected from many resources is often noises, full of errors, outliers, and duplicates. Therefore, pre-processed and cleaned data is needed in order to the data structure [2]. In the past few decades, the bulk of research has concerned electricity load forecasting. The techniques applied in STLF can be classified into two groups, namely classical group and artificial intelligence based group. The classical methods are widely supported for forecasting areas, which include regression [3], exponential smoothing [4], ARIMA models [5] and gray forecasting model [6]. However, there are some limitations and difficult to solve the non-linear data. Since artificial intelligence based techniques are able to solve nonlinearity problem, it has been popular in these days. Some of those are neural network [7], fuzzy logic [8] and support vector regression [9]. By introducing structural risk minimizing principle, support vector regression (SVR) gives better generalization capability.



According to the ability of solving nonlinearity problems, SVR became an efficient model among machine learning algorithms [10].

The aim of this study is to improve the STLF models for net peak load demand in Thailand. The paper structures as follows. Section 2 reviews the analysis of the data. Methodology is described in section 3. Then, the preparation of the data is examined in section 4. In section 5, the forecasting results are discussed and the last session concludes the findings and presents the future research.

2. Data analysis

The first thing is to observe the behavior of the electricity load demand. There are some relations among load demand and other external factors, such as seasonality, temperature, and holidays. The following are the observations of the demand described in each section.

2.1. Behavior of load demand

The data are collected as half an hour per day. Figure 1 shows the net peak load of Thailand 2017 and November, 2017. By simple analysis, it is easy to discover that there are some properties of the electricity load. It has seasonality: high demand in the summer while low demand in the winter. In order to the load pattern, it has the relation between weather condition and electricity load in the seasons. Then, the daily load pattern of each day has different demand. The load in weekdays is higher than that of weekend (Saturday and Sunday). Further, the load on Saturday is slightly higher than that on Sunday. Moreover, there can be seen low load pattern especially the load demand in April and December due to the Songkran festival and end year holidays.

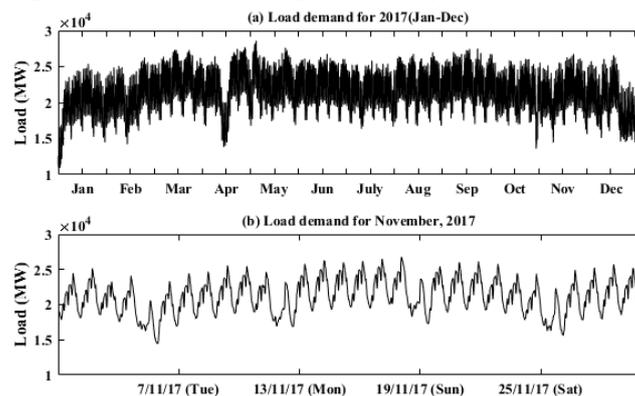


Figure 1: The load demand curve for 2017 and November, 2017

2.2. Weather influence

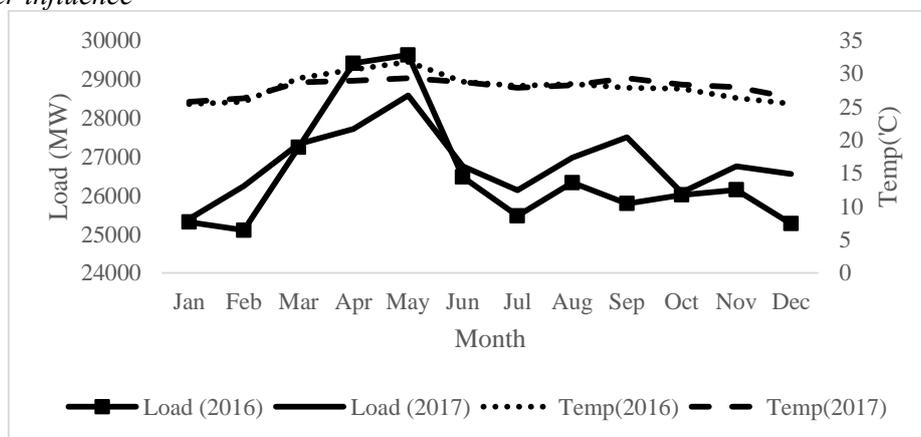


Figure 2: The yearly load demand and temperature for 2016 and 2017

Weather conditions always play an essential role in the electricity forecasting. Temperature, rainfall, humidity and some special condition like storms included as weather conditions for forecasting. However, these considerations are regarded on different situation in order to different location. Temperature has the higher effect than other weather conditions in Thailand. The correlation coefficient between the load and temperature is 0.5. It is clearly seen from figure 2 that the higher the temperature, the higher the electricity load demand. It is estimated that cities the size of Bangkok may need about 2 GWs of additional electricity for increasing of 1 degree Celsius according to the increased demand for air conditioning [11].

3. Methodology

3.1. Support vector regression (SVR)

Support vector machine (SVM) is a supervised learning technique which is originally approach from Vapnik's statistical learning theory [12]. It was first introduced for classification then extended for regression. The main idea of support vector regression (SVR) is to apply the non-linear mapping function $f(x)$ and to minimize the error, individualizing hyper plane which maximizes the margin. The training data $L = \{(x_i, y_i)\}_{i=1}^T$, where x_i is the input data and y_i is the associated output value of x_i , the non-linear problem can be defined by a regression function.

$$f(x) = w\phi(x) + b \quad (1)$$

Where, w represents the regression coefficients and b is the bias term. The empirical structural risk function can be described as:

$$R = \frac{1}{2} \|w\|^2 + C \left(\sum_{i=1}^l \xi_i + \xi_i^* \right) \quad (2)$$

Subjected to the constraints; $y_i - wx_i - b \leq \varepsilon + \xi$, $wx_i + b - y_i \leq \varepsilon + \xi_i^*$, $\xi_i^* \geq 0$, $\xi_i \geq 0$

Where, ε is the width of the loss function, C denotes the tuning parameter between the training data set and model complexity. ξ_i^* and ξ_i are the slack variables with non-negative values to ensure feasible constraints. There are three kinds of kernel function used in SVR. The radial based function (RBF) is widely used in SVR as it relies on only the parameter γ to be tuned upon the data. The RBF kernel function is regarded as $K(x, x_i)$.

$$K(x, x_i) = \exp \gamma \|x - x_i\|^2 \quad (3)$$

Where, x and x_i are the inputs in the respective dimensions and γ is the width of the RBF function.

3.2. R-loess

Loess is a special case for linear regression which is locally estimated scatter plot smoothing. It considered the neighboring points within a span to smooth the data. The span is calculated for each point by the regression weight function.

$$w_i = \left(1 - \left| \frac{x - x_i}{d(x)} \right|^3 \right)^3 \quad (4)$$

Where,

x = the predictor value associated with the responded value to be smooth

x_i = the nearest neighbors of x as defined by the span

$d(x)$ = the distance along the abscissa from x to the most distant predictor value within the span

The weighted value of the surrounded data points determines the smoothed value if there is no influence of the outside points of the span. The outliers of the dataset are along with the loess curve. So, the robust smoothing is modified to loess (r-loess). The idea is to eliminate the effect of the outliers on loess. R-loess can put zero weights on the outliers and non-influence on the data to be smoothed [13]. The bisquare function gives the weights of the r-loess.

$$w_i = \begin{cases} (1 - (r_i/6MAD)^2)^2, & |r_i| < 6MAD, \\ 0, & |r_i| \geq 6MAD, \end{cases} \quad (5)$$

Where,

r_i = the residual of the i^{th} data point produced by the regression smoothing procedure,

MAD = the median absolute deviation of the residuals,

MAD = median ($|r_i|$)

The spread out of the residual is measured by MAD . If the value of r_i is small in comparison with $6MAD$, the robust weight is zero and the related data point is not in the smooth calculation. The range of the span is between 0 and 1. The greater the value of the span, the smoother is the fitted curve.

4. Data preparation

4.1. Smoothing of the temperature and electricity load demand

First, the load and temperature data is separated into seven group according to the day such as Monday, Tuesday etc. Then, each group is organized by each time period: 00:00 a.m., 01:00 a.m., etc. The data between 2016 and 2017 is selected to train and test the model.

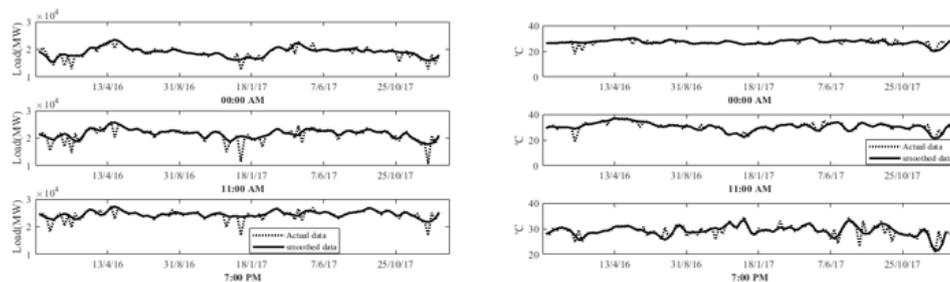


Figure 3: The load and temperature curve (2016-2017) before smoothing and after smoothing by r-loess

Therefore, the load data from 2015 December and 2018 January for data smoothing is taken because an r-loess span of seven is used to calculate the smooth data. The total data points of each period for each group are 112 data points. An r-loess span of seven defines the data which is smoothed by using a span of 6.25% of the total number of the data points. Figure 3 show the load demand curve and temperature curve for before and after smoothing at 00:00 a.m., 11:00 a.m., and 7:00 p.m. from December, 2015 to January, 2017.

4.2. Training and testing data of segmentation

The load data is organized into seven groups as Monday, Tuesday and so on. Forecasting for Monday is that the testing and training data includes only Monday load. Therefore, there are the total of seven groups for testing and training. The walk-forward testing routine [14] is used for testing and training the data. Testing 1 dataset is meant by training 52 datasets. For that reason, the data is sliding forward for the rest 51 dataset and implement the same process. The testing dataset slides 1 dataset forward (1-52 testing dataset), the SVR model is performed with the new sliding training dataset as shown in figure.

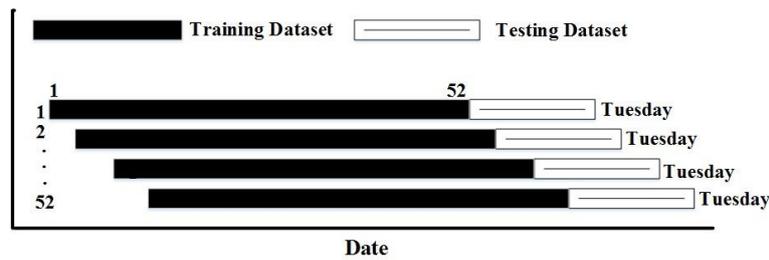


Figure 4: Walk-forward testing routine

4.3. Model specification

The testing data is from January 2017 to December 2017 by training 2016 the whole year. There are three models to forecast the load demand of 2017.

Model 1:

$$F_t^d = a_1 L_t^{d-7} + a_2 L_{t-1}^{d-1} + a_3 L_{t-2}^{d-1} + a_4 L_t^{d-1} \quad (6)$$

Model 2:

$$F_t^d = a_1 L_t^{d-7} + a_2 L_{t-1}^{d-1} + a_3 L_{t-2}^{d-1} + a_4 L_t^{d-1} + a_5 T_t^{d-1} \quad (7)$$

Model 3:

$$F_t^d = a_1 L_t^{d-7} + a_2 L_{t-1}^{d-1} + a_3 L_{t-2}^{d-1} + a_4 L_t^{d-1} + a_5 T_t^{d-1} + a_6 SI^m + a_7 Jan \text{ or not} + \dots \quad (8)$$

$$+ a_{18} Dec \text{ or not}$$

Where

F_t^d = Forecasted load of day d at period t ,

L_t^{d-7} = Previous week same day load of day d at period t ,

L_t^{d-1} = Yesterday load of day d at period t ,

T_t^{d-1} = Yesterday temperature of day d at period t ,

SI^m = Monthly seasonal index (monthly load divided by yearly load)

Jan or not = 0 or 1 {e.g if it is January-> only January is 1, others are 0}

5. Result and Discussion

The comparison analysis of the three models is described in this section. All the data of three models are filtered by r-loess and perform STL on different input training variables. The performance of the monthly and yearly MAPE for three models is shown in table 1. The average for each column is the MAPE performance for each month in 2017. The yearly MAPE is to select the best one among the three models. The yearly average of the model 3 is 2.57 while model 1 is 3.32 and model 2 is 2.87. So, considering the yearly MAPE, model 3 performs well in comparison with the other models. The minimum and the maximum monthly average of the model 3 is 1.84 (January) and 4.82 (December). Both model 1 and model 2 give the maximum monthly average MAPE in December which is 5.74 and 5.45 respectively. The minimum monthly average is 2.13 (July) for model 1 and that of the model 2 is 2.02 (November).

The forecasting performance in December, during which the lowest temperature of the year are examined, are usually the worst performance of the year due to the lowest temperature and lower load demand. Since there is a correlation between temperature and load demand, seasonality and load demand, model 3 performs better than other months which the training input variables are included the

seasonality and temperature. Seeing that Thailand is a tropical climate region, electricity load demand is sensitive to temperature and seasonality.

Table 1: Comparison of MAPE for the three models

	Model 1	Model 2	Model 3
JANUARY	3.71	2.29	1.84
FEBRUARY	4.55	2.83	3.16
MARCH	3.47	3.24	2.62
APRIL	4.53	4.11	3.06
MAY	3.49	3.12	2.96
JUNE	2.84	2.66	2.40
JULY	2.13	2.10	2.03
AUGUST	2.22	2.06	1.93
SEPTEMBER	2.73	2.39	2.05
OCTOBER	2.17	2.14	2.08
NOVEMBER	2.30	2.02	1.93
DECEMBER	5.74	5.45	4.82
Total Average	3.32	2.87	2.57

There are eight plots to show the average monthly MAPE for each day and total average for each day in 2017. The plots show the model 3 (solid line) compared with the other two models: model 1 (dotted line), model 2 (dashed line). The first plot is Monday monthly MAPE. Normally Monday MAPE is difficult to forecast because it is the first day of working day. The forecasting model includes yesterday load demand in input variables. So the influence of the difference of load demand between Monday and Sunday is large since the load demand on Sunday is significantly low compared to the other days. However, model 3 which includes the temperature and dummy variables is effective to overcome this problem. It can be seen that the MAPE of Monday of model 3 in April is considerably better than the other two models. On the other hand, monthly MAPE of Tuesday and Thursday are almost the same in all models except January and February. The three models go the same pattern for Wednesday in every month but go slightly difference for December. Nevertheless, the total average of Wednesday monthly MAPE in model 3 is slightly higher than other models as the December MAPE of Wednesday is 6.908. Thursday's monthly MAPE of model 3 performs well especially in December. However, the MAPE of Thursday and Friday in May is moderately high as the holidays of May are on Thursday and Friday. Load forecasting on holiday is not easy to handle. Weekend's load forecasting is normally insufficient compared to weekdays'. But, some companies and industries are still opened on Saturday in Thailand. Therefore, the load demand on Saturday is not too low like that of Sunday. The value of the monthly MAPE of Saturday on model 3 is not more than 3 except December. Due to many factors affecting on December, the performance of model 3 on Thursday and Saturday are poor compared to model 1 and model 2. Since the load demand trains on the separated day (training on the same day), it is strongly sufficient on Sunday.

February 14 (Tuesday) and April 23 (Sunday) are selected for illustration because the temperature values in these days are minimum and maximum temperature in 2017. It is like a hard case for load forecasting. Figure shows the actual electricity load demand compared with all three forecasting models. According to figure 6, all three models follow the actual demand. The MAPE of model 1 on February 14 is 7.74 while that of model 2 and model 3 are 3.01 and 2.37. As for April 23, model 1 is 7.39, model 2 is 6.54, and model 3 is 2.67. Considering temperature variables and seasonal variables gives accurate forecast.

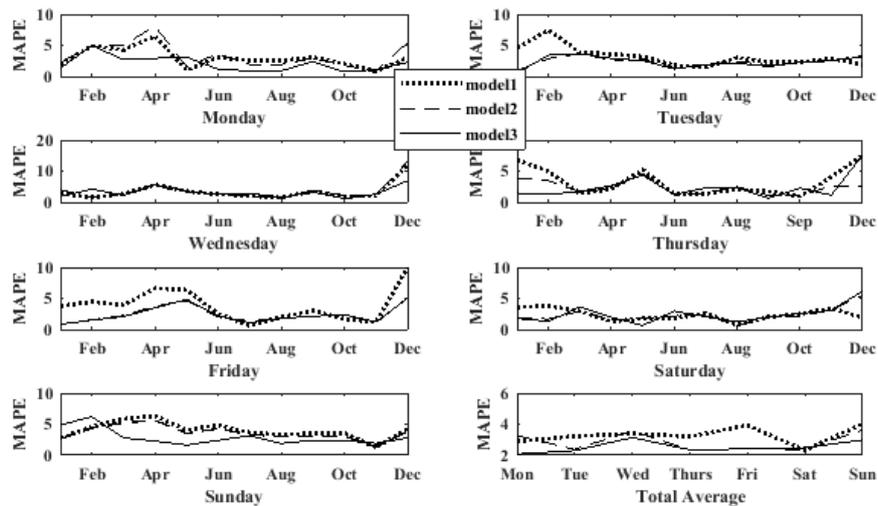


Figure 5: Average monthly MAPE for each day and total average of each day in 2017

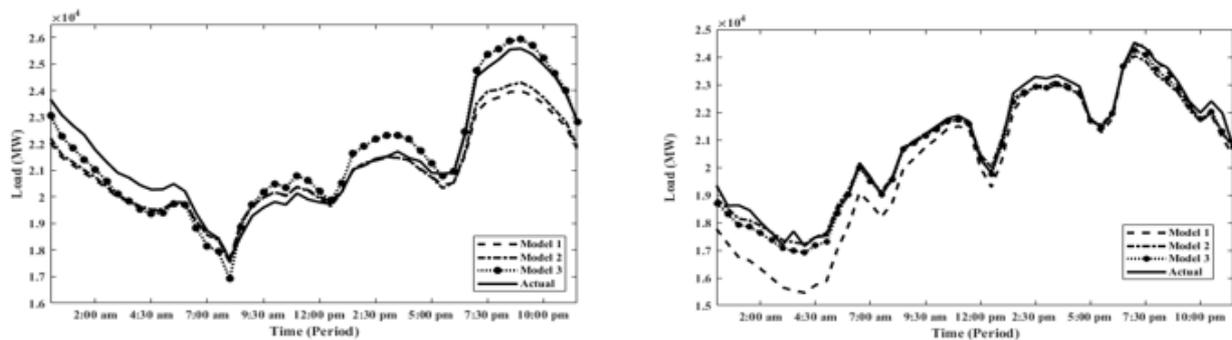


Figure 6: Actual demand and forecasted demand of three model on 23rd April and 14th February, 2017

6. Conclusion

The main contribution of the paper is to propose the importance of data preprocessing before forecasting and the effectiveness of selecting external variables. An SVR based model has been built for comparative study. There are three models which have been trained and tested on the dataset from EGAT (Electricity Generating Authority in Thailand). Among all three models, choosing the appropriate variables enhance the model performance. It can be seen that temperature and seasonality are strongly influenced on the load forecasting of Thailand. Model 3 can generate forecasting distributions incorporating the randomness from the model and exogenous variables. The study has found that the performance on December is still weak. The data are smoothed by r-loess for each period of the day. Furthermore, there are some outliers left which are long holidays as the window span is seven for the data point. Therefore, future research is to find a new way to get better performance on December.

7. Acknowledgement

This research is partially supported by the Logistics and Supply Chain Systems Engineering Research Unit (LogEn), Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU), Thailand. Data used in this research is provided by EGAT. Therefore, we acknowledge their support for completing this research successfully.

8. References

- [1] E. A. Feinberg and D. Genethliou, "Load forecasting," in *Applied mathematics for restructured electric power systems*: Springer, 2005, pp. 269-285.
- [2] M. MathWorks, "the Mathworks," *Inc., Natick, MA*, 1992.
- [3] A. D. Papalexopoulos and T. C. Hesterberg, "A regression-based approach to short-term system load forecasting," *IEEE Transactions on Power Systems*, vol. 5, no. 4, pp. 1535-1547, 1990.
- [4] W. Christiaanse, "Short-term load forecasting using general exponential smoothing," *IEEE Transactions on Power Apparatus and Systems*, no. 2, pp. 900-911, 1971.
- [5] M. Cho, J. Hwang, and C. Chen, "Customer short term load forecasting by using ARIMA transfer function model," in *Energy Management and Power Delivery, 1995. Proceedings of EMPD'95., 1995 International Conference on*, 1995, vol. 1, pp. 317-322: IEEE.
- [6] J.-f. Zhang, Y.-a. Wu, and J.-j. Wu, "Application of gray system theory in load forecasting [J]," *Electric Power Automation Equipment*, vol. 5, p. 005, 2004.
- [7] C. Jeenanunta and K. D. Abeyrathna, "Combine Particle Swarm Optimization with Artificial Neural Networks for Short-Term Load Forecasting."
- [8] M. B. Mbarek and R. Feki, "Using fuzzy logic to renewable energy forecasting: a case study of France," *International Journal of Energy Technology and Policy*, vol. 12, no. 4, pp. 357-376, 2016.
- [9] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199-222, 2004.
- [10] M. Barman, N. D. Choudhury, and S. Sutradhar, "A regional hybrid GOA-SVM model based on similar day approach for short-term load forecasting in Assam, India," *Energy*, 2018.
- [11] C. Greacen and A. Palettu, "Electricity sector planning and hydropower in the Mekong Region," *Democratizing water governance in the Mekong region*, pp. 93-125, 2007.
- [12] S. R. Gunn, "Support vector machines for classification and regression," *ISIS technical report*, vol. 14, pp. 85-86, 1998.
- [13] W. Wettayaprasit, N. Laosen, and S. Chevakidagarn, "Data filtering technique for neural networks forecasting," in *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization*, 2007, pp. 225-230: World Scientific and Engineering Academy and Society (WSEAS).
- [14] L.-J. Cao and F. E. H. Tay, "Support vector machine with adaptive parameters in financial time series forecasting," *IEEE Transactions on neural networks*, vol. 14, no. 6, pp. 1506-1518, 2003.