Transdisciplinary Engineering for Complex Socio-technical Systems
K. Hiekata et al. (Eds.)

doi:10.3233/ATDE190139

Improvement of Building Electricity Load Prediction Accuracy Using Hybrid k-Shape Clustering EMD Based Support Vector Regression

Irene KARIJADI[a,1], Shuo-Yan CHOU[a], Anindhita DEWABHARATA[a] and Tiffany Hui-Kuang YU[b]
[a] *Department of Industrial Management, National Taiwan University of Science and Technology, Taiwan, ROC*
[b] *Department of Public Finance, Feng Chia University, Taiwan, ROC*

Abstract. An accurate electricity load prediction is important to optimizing building electricity load performance. However, building electricity load prediction is complex due to many influencing factors. This study develops a hybrid algorithm that combines clustering approach, empirical mode decomposition, and support vector regression to develop a prediction model for building electricity load. *k*-shape clustering is used to extract similar building electricity load pattern, and empirical mode decomposition is employed to decompose electricity load data into several Intrinsic Mode Functions (IMF). Finally, a prediction model using support vector regression is built for each IMF individually, and the prediction result of all IMFs is combined to obtain an aggregated output of electricity load. Numerical testing demonstrated that the proposed method can accurately predict the electricity load in the building.

Keywords. Energy, Time Series, Prediction, Data Mining

Introduction

Buildings contribute for nearly 40% of the total global energy consumption, and it has exceeded the other major sectors, such as industrial and transportation [1]. With the growth in population and increasing demand for energy consumption, have made energy efficiency as one of the primary concern in our society nowadays. It is essential to improve building energy efficiency since it brings environmental benefit and reduced energy cost in the building. Building electricity load prediction is one of the solutions to improve energy efficiency. By predicting the electricity load, the building manager can appropriately manage the building operational strategies to have an intelligent and efficient system. When the prediction shows such peak or irregular consumption in the near future, the system can give an early warning to the system, and further action can be taken, such as turning off the unnecessary appliance, or shift the peak load using peak load shifting technique.

[1] Corresponding Author, Department of Industrial Management, National Taiwan University of Science and Technology, No 43 Section 4, Keelung Rd., Taipei 10607, Taiwan, ROC; Email: irenekarijadi92@gmail.com

Many factors influence energy consumption in the building, such as weather condition, building structure, building characteristic, and occupant behavior. Due to this complex situation, predicting electricity load accurately is quite challenging. In the past decades, researchers have developed lots of prediction methods for building electricity load, include engineering, statistical and artificial intelligence methods. Support Vector Regression (SVR) is one of the artificial intelligence technique, and it has been used in building electricity load prediction for a long time since it is capable of non-linear time series modeling [2]. Dong *et al*. [3] utilized SVR to predict monthly electricity load in the tropical region, and it has better performance than other approaches such as neural network and genetic programming. Li *et al.* [4] applied SVR for the prediction of hourly cooling demand in Guangzhou, China. Other studies also applied SVR to building electricity load prediction [5],[6] with the prediction accuracy showed better results than other prediction methods. These previous researches demonstrate that SVMs can perform well in predicting building electricity load.

Generally, the single prediction method is widely used for AI-based building energy use prediction [7]. The model is developed using basic data-mining algorithms. However, improvement in the building electricity load prediction accuracy is crucial for scheduling or planning electricity load in the building. As a result, hybrid prediction models have been proposed by many researchers to improve building electricity load prediction accuracy. Chen *et al.* [8] introduced a hybrid SVR model combined with wavelet decomposition to predict the hourly electric demand intensity in a mall and a hotel building. Their result indicates that the introduction of wavelet decomposition can improve the prediction accuracy for hotel building. Yang *et al.* [9] performed *k*-shape clustering to discover building electricity load pattern and further implemented into an SVR model to enhance building energy forecasting accuracy. Qiu *et al.* [10] proposed an ensemble method composed of Empirical Mode Decomposition (EMD) and deep belief network (DBN) to predict load demand. The EMD algorithm decomposes a load time series signal into several intrinsic mode functions (IMFs) and residual. Then, a DBN model is trained separately for each of the extracted IMF and residual, and the accuracy of the prediction model is improved compared with single DBN model.

In this study, *k*-shape clustering techniques and EMD decomposition method are used to improve the accuracy of building electricity load prediction using SVR. Currently, there is limited research on the improvement of building electricity load prediction accuracy considering the contribution of combination *k*-shape-EMD processing that could make.

1. Literature Review

1.1. Support Vector Regression

SVM is a learning system using a high dimensional feature space. A version of a SVM for regression has been proposed in 1997 by Vapnik, Steven Golowich, and Alex Smola [11]. This method is called support vector regression (SVR).

The idea of SVR is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped via a nonlinear function. The goal of SVR is to find a function f(x) that has at most ε deviation from the actual obtained target y_i for all the training data, and at the same time is as flat as possible [12]. Training the SVR means solving:

$$\text{Min } \frac{1}{2}\|w^2\| \tag{1}$$

$$\text{Subject to } \begin{cases} y_i - \langle w, x_i \rangle - b \le \varepsilon, \\ \langle w, x_i \rangle + b - y_i \le \varepsilon, \end{cases} \tag{2}$$

Where x_i is a training sample with target value y_i. The inner product plus intercept $\langle w, x_i \rangle + b$ is the prediction for that sample, and ε is a free parameter that represents as a threshold.

1.2. k-Shape Clustering

k-shape clustering first introduced by Paparrizos [13]. This algorithm used a normalized cross-correlation algorithm to take into account the shapes of time series data. *k*-shape groups time series into clusters based on their shape similarities. The study shows an outperforming in terms of partitional, hierarchical and spectral clustering than other approaches.

k-shape is a partitional clustering method that is based on an iterative refinement procedure similar to the one used in *k*-means. Unlike the *k*-means approach, *k*-shape uses both a different distance measure and a different method for centroid computation. In every iteration, *k*-shape performs two steps:

- In the assignment step, *k*-shape relies on the Shape Based Distance (SBD) measures to update the cluster membership. Geometric mean of autocorrelation of each time series data is considered and used to divide the coefficient normalization as shown in equation 3

$$SBD\ (\vec{x}, \vec{y}) = 1 - \ max_w \left(\frac{CC_w(\vec{x}, \vec{y})}{\sqrt{R_0(\vec{x}, \vec{x}).R_0(\vec{y}, \vec{y})}} \right) \tag{3}$$

- In the refinement step, the cluster centroids are updated to reflect the changes in cluster memberships in the previous step.

The algorithm repeats these two steps until either no change in cluster membership occurs or the maximum number of iterations allowed is reached.

1.3. Empirical Mode Decomposition

In order to improve the prediction performance, the use of Empirical Mode Decomposition (EMD) method have gained interest in the recent years [14] [15]. EMD offers a new way to deal with non-linear and non stationary data. EMD algorithm works by decomposing a non-stationary time series signal into a set of Intrinsic Mode Functions (IMFs) along with a residue which are more stable component and can be more easily modeled to improve prediction accuracy. The step by step procedure of EMD algorithm is given in Algorithm 1, as shown in table 1

Table 1. EMD Algorithm.

Algorithm 1 Empirical Mode Decomposition (EMD) Algorithm
Step 1. Given a time series signal $S(t)$, identify local maxima and minima.
Step 2. Calculate upper $S_u[t]$ and lower $S_l[t]$ envelope by interpolation of local maxima and minima.
Step 3. Compute the mean of upper and lower envelopes. $m_t = \frac{S_u[t]+S_l[t]}{2}$
Step 4. Subtract mean from time series signal $h(t) = S(t) - m_t$
Step 5. Repeat Step 2 to 4 until one of the stopping criteria is reached: • $S(t)$ reaches zero • Max. Number of iterations reached • Both constraints are satisfied
Step 6. Treat h(t) as new IMF and calculate the residual signal $r(t)$ as $r(t) = S(t) - h(t)$
Step 7. Use as new IMF and repeat step to 6, until all IMFs are obtained

2. Methodology and Experiment

2.1. Datasets desription

The electricity load datasets from OpenEI [16] were used to evaluate the performance of the proposed model. 4 retail buildings and 2 office buildings in United states were selected as case study buildings. The electricity load data from 1 January 2010 to 30 December 2010 is used in this study. Data is divided into training and test data sets. The electricity load data with respect to time for each building are shown in Figs.1-2, respectively. The training data set is used to develop the models while the test data set is used to evaluate the prediction performance of the models. In this study, the first 70% of the data will be used as training dataset and the remaining 30% of the data will be used for the test data set. The detailed decripton of these datasets are given in Table 2.

2.2. Proposed method

The framework of the proposed method is illustrated in figure 3. In this proposed method, data preprocessing is first applied to the original data. After data preprocessing, *k*-shape clustering is performed to find out the groups of days in the electricity data that follows similar hourly consumption patterns. "*dtwclust*" package in R [17] is used to implement *k*-shape clustering algorithm. Then, Empirical mode decomposition is applied to decompose the original data. Finally, SVR prediction model is built for each IMF individually. Detail of EMD-SVR is depicted in figure 4. The EMD is completed by applying the"*Rlibeemd*" package in R [18]. While for SVR prediction model, the "*e1071"* package in R [19] is used to predict the electricity load.

Table 2. Dataset Descriptive Summary.

	Building 1	Building 2	Building 3	Building 4	Building 5	Building 6
Type	Retail	Retail	Retail	Retail	Office	Office
Granularity	15 Minute	15 Minute	15 Minute	15 Minute	1 Hour	1 Hour
Unit	Kilo Watt	Kilo Watt	Kilo Watt	Kilo Watt	Kilo Watt	Kilo Watt
Maximum	457.9	504.6	529.7	558.1	92.5	119
Minimum	42.4	7.4	32.3	77.6	8.4	14.8
Mean Value	261.8	282.2	266.0	334.7	27.04	34.08
Missing	25	15	83	0	0	1

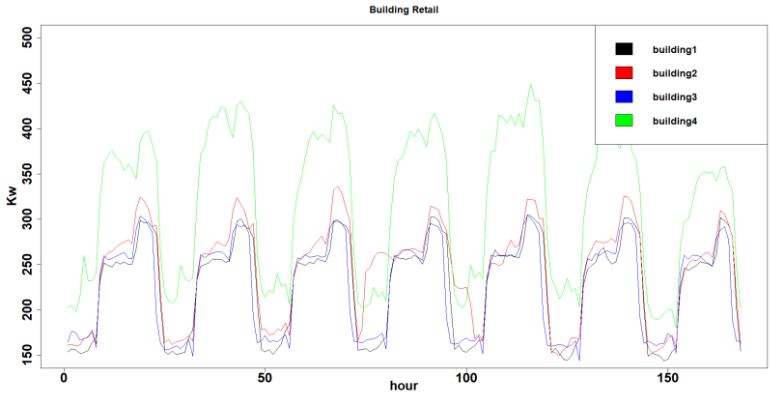

Figure 1. One week observation for retail building.

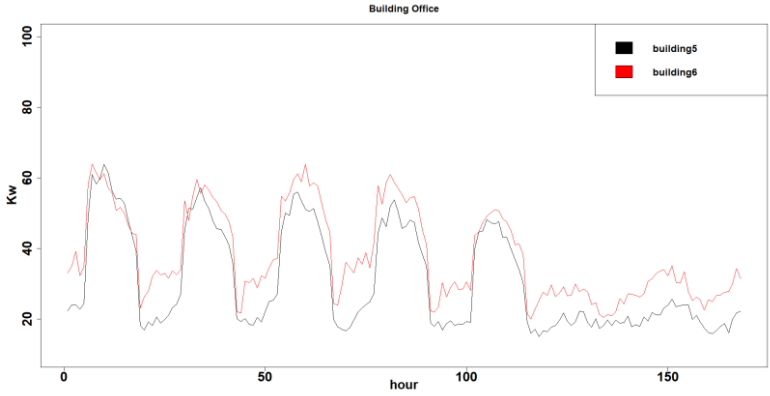

Figure 2. One week observation for office building.

Figure 3. Flowchart of the proposed methodology.

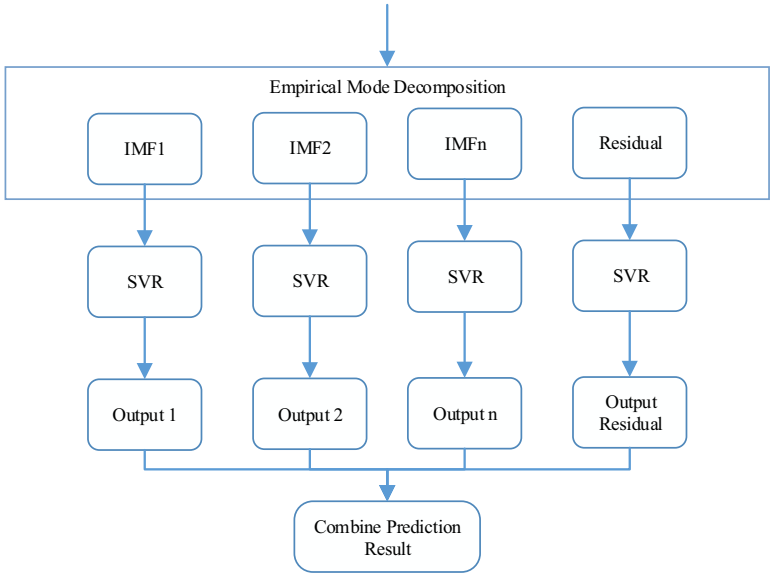

Figure 4. Flowchart of data decomposition and prediction.

- **Data preprocessing**
  Data preprocessing is important for data-driven approach, since real world data tends to be incomplete, noisy and inconsistent. Data preprocessing in this study include data cleaning and data aggregation. In data cleaning, we replace the missing value using linear interpolation method [20]. Data collected at 15-min interval are aggregated into hourly because at 15-min interval data are often to be found noisy and fluctuative. While for daily data, it exhibit less variation and noise than the hourly data,but it still provides information about trend. Therefore, hourly prediction is selected as the preferred time-scale in this study

- **Data clustering**
  k-shape clustering is used for the recognition of similar days pattern in each building. To choose the most suitable number of clusters, a method based on the silhouette index [21] is applied, where a higher silhouette coefficient score relates to a model with better defined clusters. The number of clusters ranges from 2 to 11 and the highest silhoutte coefficient is selected as the number of cluster as shown in figure 5. The chosen cluster number for each building is shown in Table 3. Clustering result is utilized as a feature in prediction model to improve the accuracy of prediction result. Figure 6 shows the cluster results for building 6 and each cluster represents a typical daily consumption pattern of this building.

Table 3. The number of cluster for each building.

	Building 1	**Building 2**	**Building 3**	**Building 4**	**Building 5**	**Building 6**
Cluster	3	4	2	2	2	4

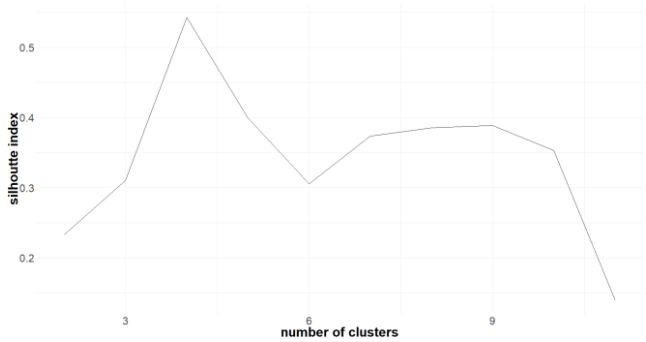

Figure 5. Silhoutte index with different number of cluster for building 6, cluster number = 4 is selected.

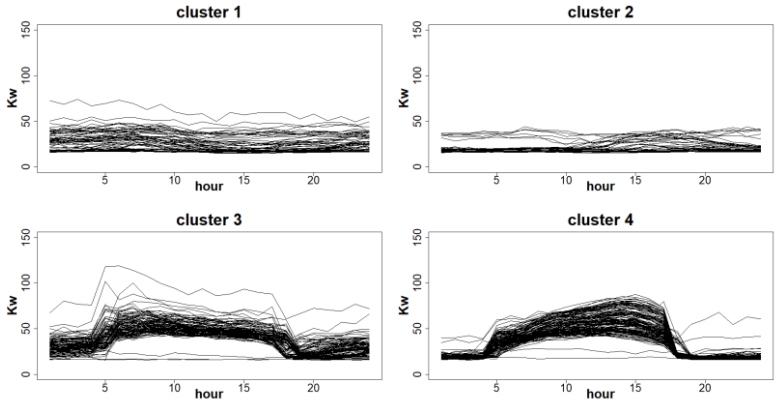

Figure 6. Clustering results for building 6 (4 clusters).

- **Data decomposition and prediction**
  The original data is decomposed by Empirical Mode Decomposition into several IMF and one residue. EMD offers a way for trend extraction which represent as residual (last IMF). The residual is used in the prediction model and is predicted using proposed approach which described in figure 4. After being decomposed by EMD, SVR model is trained on each IMF and residue as shown in figure 7. For the SVR prediction model, the model input includes the previous hourly electricity load (time lagged), and cluster information as extra features. Autocorrelation function (ACF) is applied to determine appropriate number of time lagged used for prediction. We choose time-lagged which have autocorrelation value greater than 0.5. Time lagged results are summarized in Table 4. Prediction result of all IMFs and residue is combined together by summation to obtain an aggregated output of electricity load. Output of the model is the prediction of consumption values for the next hour.

Table 4. Time-lagged result for each building.

	Building 1	**Building 2**	**Building 3**	**Building 4**	**Building 5**	**Building 6**
time-lagged	(t-1), (t-2), (t-3)	(t-1), (t-2), (t-3)	(t-1), (t-2), (t-3),(t-4)	(t-1), (t-2), (t-3), (t-4)	(t-1), (t-2), (t-3), (t-4)	t-1), (t-2), (t-3)

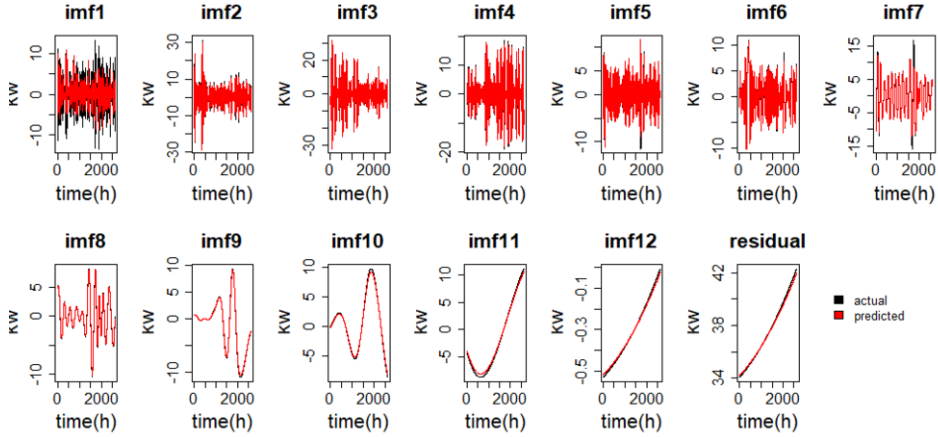

Figure 7. Decomposition and prediction results (building 6).

2.3. Performance Criteria

We evaluate the prediction accuracy by calculating three different statistical metrics, the mean absolute percentage error (MAPE), the root mean square error (RMSE), and the mean absolute error (MAE). The definitions of MAPE, RMSE, and MAE are expressed as equations (4–6):

$$MAPE\ \% = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t' - y_t}{y_t}\right| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t' - y_t)^2} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t' - y_t| \tag{6}$$

Where y_t' is the predicted value, y_t is the actual value and n is the number of data points in the time series.

2.4. Performance Result

The prediction performances of the proposed method are shown in Table 5 and Figure 8. It shows the prediction accuracy is improved after the application of EMD and *k*-shape clustering. Based on the obtained experimental results, we conclude that EMD can decompose the load data pattern into sequent elements, with higher and lower frequencies. It becomes a useful technique to help SVR prediction model to deal well with the tendencies of each component and it becomes easier to predict for each component. In the meanwhile, *k*-shape clustering is used to group load pattern based on their shape characteristics. Therefore, *k*-shape is able to determine days of the week with similar energy consumption pattern. Information about days of the week with similar energy consumption profiles is further utilized in the prediction model. Hence, the daily and weekly pattern are captured and the prediction accuracy is enhanced. The combination of *k*-shape clustering-EMD along with SVR can generate better prediction model.

3. Conclusion and Future Research

Building electricity load prediction can lead to a better decision in strategic planning of building energy management system. In the recent years, various methods including statistical approach and Artificial Intelligence based approaches were developed to accurately predict electricity load. In this study, a hybrid method using *k*-shape clustering, Empirical Mode Decomposition and SVR is proposed to predict hourly electricity load in the building. Based on the experimental results, the proposed approach (*k*-shape and EMD) can be utilized to deal efficiently with nonlinear features of electricity load data and can be used to improve SVR prediction accuracy.

As a future work, we are planning to investigate the effect of the proposed algorithm using others prediction method (such as ANN,random forest, etc) and add different granularity (15-minutely, hourly and daily) in order to figure out the modeling adaptability.

Table 5. Summary of prediction results.

Building	Method	Training			Testing		
		MAPE (%)	RMSE	MAE	MAPE (%)	RMSE	MAE
1	Pure SVR	5.0532	22.0902	12.8291	4.9507	21.9704	12.8898
	k-shape EMD-SVR	3.437	11.5409	8.5272	3.6855	12.9588	9.3834
	Improvement	1.6162	10.5493	4.3019	1.2652	9.0116	3.5064
2	Pure SVR	5.1882	22.8917	13.92919	5.8214	24.1941	14.5668
	k-shape EMD-SVR	3.5524	12.7601	9.2927	4.2162	14.1118	10.4878
	Improvement	1.6358	10.1316	4.63649	1.6052	10.0823	4.079
3	Pure SVR	5.2563	25.2151	13.0902	4.9548	18.4348	11.1619
	k-shape EMD-SVR	3.8916	13.3174	9.5964	3.2248	9.7187	7.2495
	Improvement	1.3647	11.8977	3.4938	1.73	8.7161	3.9124
4	Pure SVR	5.7336	27.5195	18.0166	5.711	23.7274	16.62
	k-shape EMD-SVR	3.7818	15.4517	11.7168	4.6795	17.1247	12.761
	Improvement	1.9518	12.0678	6.2998	1.0315	6.6027	3.859
5	Pure SVR	9.4915	4.5748	2.2452	11.1276	5.6156	2.9845
	k-shape EMD-SVR	7.719	2.6153	1.722	8.9237	3.2287	2.1142
	Improvement	1.7725	1.9595	0.5232	2.2039	2.3869	0.8703
6	Pure SVR	8.8205	5.213	2.6982	10.0436	6.592	2.6892
	k-shape EMD-SVR	6.6239	3.13	2.0025	7.2934	3.4401	2.3212
	Improvement	2.1966	2.083	0.6957	2.7502	3.1519	0.368

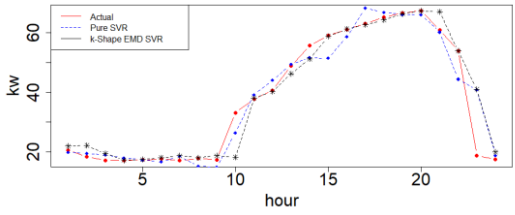

Figure 8. Prediction Results on test dataset Building 6 (1 day observation).

References

[1] L. Pérez-Lombard, J. Ortiz, and C. Pout, A review on buildings energy consumption information, *Energy and buildings,* vol. 40, pp. 394-398, 2008.

[2] H.-x. Zhao and F. Magoulès, A review on the prediction of building energy consumption*, Renewable and Sustainable Energy Reviews*, vol. 16, pp. 3586-3592, 2012.

[3] B. Dong, C. Cao, and S. E. Lee, Applying support vector machines to predict building energy consumption in tropical region, *Energy and Buildings*, vol. 37, pp. 545-553, 2005.

[4] Q. Li, Q. Meng, J. Cai, H. Yoshino, and A. Mochida, Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks, *Energy Conversion and Management*, vol. 50, pp. 90-96, 2009.

[5] B.-J. Chen and M.-W. Chang, Load forecasting using support vector machines: A study on EUNITE competition 2001, *IEEE transactions on power systems*, vol. 19, pp. 1821-1830, 2004.

[6] M.-G. Zhang, Short-term load forecasting based on support vector machines regression, *2005 International Conference on Machine Learning and Cybernetics*, 2005, pp. 4310-4314.

[7] Z. Wang and R. S. Srinivasan, A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models, *Renewable and Sustainable Energy Reviews*, vol. 75, pp. 796-808, 2017.

[8] Y. Chen and H. Tan, Short-term prediction of electric demand in building sector via hybrid support vector regression, *Applied energy*, vol. 204, pp. 1363-1374, 2017.

[9] J. Yang, C. Ning, C. Deb, F. Zhang, D. Cheong, S. E. Lee, et al., k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement, *Energy and Buildings*, vol. 146, pp. 27-37, 2017.

[10] X. Qiu, Y. Ren, P. N. Suganthan, and G. A. Amaratunga, Empirical mode decomposition based ensemble deep learning for load demand time series forecasting, *Applied Soft Computing*, vol. 54, pp. 246-255, 2017.

[11] V. Vapnik, S. E. Golowich, and A. J. Smola, Support vector method for function approximation, regression estimation and signal processing, *Advances in neural information processing systems*, 1997, pp. 281-287.

[12] A. J. Smola and B. Schölkopf, A tutorial on support vector regression, *Statistics and computing*, vol. 14, pp. 199-222, 2004.

[13] J. Paparrizos and L. Gravano, k-shape: Efficient and accurate clustering of time series, in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 2015, pp. 1855-1870.

[14] G.-F. Fan, S. Qing, H. Wang, W.-C. Hong, and H.-J. Li, Support vector regression model based on empirical mode decomposition and auto regression for electric load forecasting, *Energies*, vol. 6, pp. 1887-1901, 2013.

[15] Y. Ren, P. N. Suganthan, and N. Srikanth, A novel empirical mode decomposition with support vector regression for wind speed forecasting, *IEEE transactions on neural networks and learning systems*, vol. 27, pp. 1793-1798, 2016.

[16] J.Langevin, 2016, *OpenEI*, [Online]. Available: https://openei.org/datasets/dataset/activity/consumption-outdoor-air-temperature-11-commercial-buildings , accessed Juni, 8 2019.

[17] A. Sarda-Espinosa, 2018, *dtwclust: Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance,* R package version 5.5.1, https://CRAN.R-project.org/package=dtwclust

[18] J. Helske, P. Luukko, 2018, *Rlibeemd: Ensemble empirical mode decomposition (EEMD) and its complete variant (CEEMDAN)*, R package version 1.4.1, https://github.com/helske/Rlibeemd.

[19] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel and F. Leisch, 2018, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien,* R package version 1.7-0. https://CRAN.R-project.org/package=e1071

[20] J O. Cokluk and M. Kayri, The Effects of Methods of Imputation for Missing Values on the Validity and Reliability of Scales, *Educational Sciences: Theory and Practice*, vol. 11, pp. 303-309, 2011.

[21] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53-65, 1987/11/01/ 1987