Data Science and Engineering (DSE) Record, Volume 3, issue 1.

Development of Electricity Consumption Forecasting Model for Campus-Scaled Buildings Using Machine Learning

Patcharapol Yasamut [1] and Pree Thiengburanathum [2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] College of Arts, Media and Technology, Chiang Mai University, Chiang Mai, Thailand
patcharapol_y@cmu.ac.th

Abstract. The demand for electricity in buildings on a national and international scale is currently rising rapidly. Building electricity usage can be decreased by using a forecasting model. It can reduce utility costs not just for one building but also throughout a whole region. According to literature review, machine-learning and deep-learning techniques have been used in previous studies on forecasting electricity consumption. However, there is a dearth of research into the use of clustering to predict electricity consumption in tropical regions such as Thailand or any of the countries in Southeast Asia. In this project, we present new research for hourly forecasting building energy usage. 1-hour interval electricity consumption data is collected from nineteen buildings for a year and five months by smart meters. 1-hour interval weather data including PM 10, PM 2.5, temperature, and humidity collected is also collected from one building. The analysis of the cross correlation between weather data and electricity consumption indicated that that there was a weak correlation between weather and electricity consumption data. Vector Auto Regression (VAR), Vector Auto-Regressive Moving Average (VARMA), Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) models were used to develop the forecasting models as the baseline models. The SVR model can outperform the other models with the lowest RMSE validation scores on training dataset. The hyperparameters of SVR models were optimized to maximize forecasting accuracy on training dataset. To reduce time consuming for training and optimizing the models, the k-Shape clustering approach is used to analyse electricity consumption into pattern groups. We used the centroid of each cluster as a representation of the cluster's electricity consumption data in order to forecast the electricity consumption of buildings within the cluster. The result of comparing the forecasting performance of SVR with and without clustering technique by using t-test indicated that there is no statistically significant evidence that the forecasting performance of SVR model with and without clustering technique are different at P-values of 0.7258.

Keywords: Building Energy Consumption, Outlier Detection, Clustering, Forecasting

1 Introduction

The United Nations created Sustainable Development Goal 7 (SDG 7) [1] as one of 17 SDGs. The SDG7 targets must be met, including providing everyone with reliable, cheap, and modern energy services. It should also contribute to raising the worldwide proportion of renewable energy, doubling energy efficiency gains, supporting clean energy research and investment, and expanding and promoting energy services for developing nation.

Thailand's energy generation by fuel type from 1995 to 2019 is depicted in Fig.1. In 2019, total electricity generation (the pink line) increased to 212,000 gigawatt hours (GWh). And when electricity generation in 2019 is compared to five years ago in 2014, we see an increase of approximately 14%. Increased electricity generation implies an increase in Thailand's electricity consumption as well. Additionally, the trend toward renewable energy (light blue line) tends to increase following 2015. Energy forecasting may be critical in ensuring supply reliability and in assisting with future planning and investment in additional capacity.

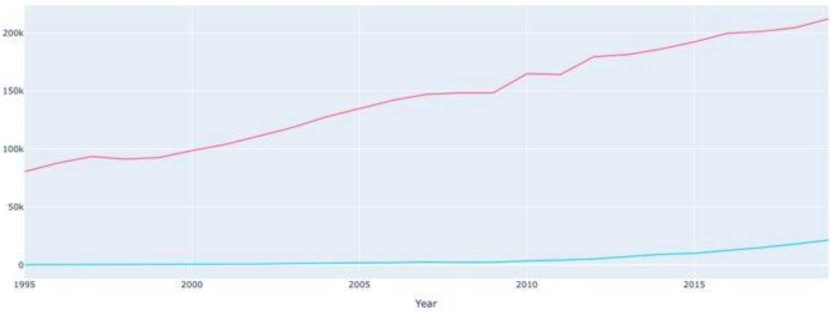

Fig.1 Electricity Generation in Thailand from 1995 to 2019

Regarding a survey of studies published in the last five years examining energy prediction methodologies, we conducted a literature review. The review often emphasized black-box methodologies (e.g., SVM [2], artificial neural networks (ANN) [3], Tree-based models[4], and other machine learning algorithms) utilized in their researches. We found that the majority of the research focused on calculating power consumption in a single building. Furthermore, there are studies that deal with estimating electricity usage in several buildings, such as three buildings [5], ten buildings [6], and fifty-seven buildings [7]. However, there is a dearth of research into the use of clustering to predict electricity consumption in tropical regions such as Thailand or any of the countries in Southeast Asia.

In this project, we present a new research paradigm for hourly forecasting building energy usage. 1-hour interval electricity consumption data is collected from nineteen buildings for a year and five months by smart meters. 1-hour interval weather data including PM 10, PM 2.5, temperature, and humidity collected is also collected from one building. The analysis of the correlation between weather data and electricity consumption is conducted in this study. To forecast the electricity consumption, VAR, VARMA, SVM and MLP are used as the baseline models. The best performance of baseline model is selected to improve its forecasting performance by optimizing hyperparameters. Moreover, the k-Shape clustering approach is used to classify electricity consumption into pattern groups and to use the centroid of each cluster as a representation of the cluster's electricity consumption data in order to forecast the electricity consumption of buildings within the cluster. Finally, the performance of the hyperparameter-optimized model is compared to that of the model trained using the centroid of clusters.

2 Literature Review

2.1 Clustering techniques used in electricity consumption forecasting

Dong, Z., et al. [7] created ensemble learning models in 2021 to forecast electricity consumption in an office building using hourly meteorological data and energy consumption data. The ensemble learning model is divided into two layers. The first layer was a hybrid of ANN and SVR models, the outputs of which were utilized to train an Multiple linear regression (MLR) in the second layer. They also used decision tree model to classify electricity consumption data into patterns then employed an ANN model, an SVR model, and the ensemble learning model to forecast for each pattern then compare the accuracy score of each model. ANOVA analysis and the Tukey-HSD test revealed significant differences between the four patterns. Furthermore, the ensemble learning model using pattern division technique outperforms the other models by 17.7%, 16.1%, 15.4%, 15.8%, and 15.6% of coefficient of the variation of the root mean square error (CV(RMSE)) with 20%, 40%, 60%, 80%, and 100% data availability, respectively.

Yang J. et al. [6], in 2017, employed the clustering technique to divide energy consumption data into categories. In the experiment, 1-hour interval electricity consumption data from 10 buildings gathered over four months were clustered into three clusters using the best performing clustering approach obtained by comparing the k-shape and Dynamic time warping (DTW) algorithms. The clustering result is compared using seven Cluster Validity Indexes (CVIs). The k-shape algorithm outperforms the DTW algorithm. Then, using the representative building for each cluster as a forecasting model, each building in the cluster is forecasted. As a consequence, when compared to the forecast model without clustering, the clustering technique greatly increases the SVR model's forecasting accuracy for seven buildings.

Chen, Y., et al. [9], in 2017, forecasted the hourly electricity consumption of a hotel and a mall. The mall building is classified as a stationary operated building since its electricity consumption is similar; electricity consumption increases fast at 10 a.m. and decreases rapidly at 19 p.m. every day and season. The hotel, on the other hand, is classified as a non-stationary operated building, and its electricity consumption varies according to the season. They clustered the similar electricity consumption patterns using the Fuzzy c-means clustering method, then forecasted the electricity consumption using the SVR model combined with multi-resolution wavelet decomposition. mean absolute percentage error (MAPE) measured the forecast accuracies of stationary and non-stationary series at 3.82 percent and 9.72 percent, respectively.

Fan, C., et al. [10], in 2014, developed ensemble models for predicting next-day energy usage and peak power demand using a data mining technique. This technique is separated into three stages. First, using an outlier identification technique that includes feature extraction, clustering analysis using entropy-weighted k-means (EWKM) algorithm, and the generalized extreme studentized deviate (GESD) to remove the anomalous daily energy consumption patterns. Second, the influence input features are selected using the recursive feature elimination (RFE) approach. Finally, an ensemble model was developed and compared to eight prediction models using a genetic algorithm to optimize model weights. The results show that the outlier detection method is effective in identifying abnormal daily energy consumption profiles, the RFE method can significantly reduce computation load while improving model performance, and the prediction accuracy of the ensemble models measured by MAPE is 2.23 percent and 2.85 percent for the next-day energy consumption and peak power demand, respectively.

2.2 Forecasting models used in electricity consumption forecasting

Shapi, M., et al. [11] developed forecasting models for electricity consumption in 2021 using Microsoft Azure's cloud-based machine learning platform. They employed three machine learning algorithms: k-Nearest Neighbor (k-NN), SVM, and ANN. The forecasting model will take electrical power data, such as power factor, voltage, and current, as input feature variables. Energy demand data was collected in 1-minute intervals from June to December 2018. When evaluating three models, the SVM outperformed the others in terms of accuracy when calculating average electricity consumption demand.

Park, S. K., et al. [12] studied the quantitative study of the influencing variables on the performance of a ground source heat pump system and forecasted hourly performance for more accurate cost savings projections in 2018. They used MLR model to evaluate the relationship between input and output variables, and an ANN model to provide high prediction accuracy, which is a more complicated algorithm than the MLR model. They collected 29 variables at 1-minute intervals from November 2016 to March 2017. The ANN provided a forecast accuracy of 1.75 percent based on the coefficient of variation of root mean squared error (CVRMSE), which is higher than the MLR's prediction accuracy of 3.56 percent.

Ahmad, M. W., et al. [13], in 2017, compared the performance of random forest (RF) and ANN models in predicting building energy consumption. They predicted electricity consumption hourly using 5-minute intervals of data obtained from an HVAC system and a combination of 30-minute and 1-hour intervals of meteorological data. With root-mean-square errors (RMSE) of 4.97 and 6.10, respectively, ANN outperformed RF.

3 Data and Methodology

3.1 Data

This study combined two different datasets, a smart meter and a weather dataset, to build forecasting models. The smart meter dataset is obtained in CSV format from the website of Ching Mai University's Energy Research and Development Institute (ERDI).

The weather dataset is available in excel format from the Chiang Mai University Climate Change Data Center (CMU CCDC) website.

The buildings' electricity consumption data was downloaded from the website in CSV format. There are 42 files available that can be downloaded in total. The "energyConsumption" column identifies the accumulated value of electricity use. The *"timeIn"* column specifies the date and time when the data was collected as shown in Fig.2 (left). The weather dataset contains PM 10, PM 2.5, temperature, and humidity measurements as shown in Fig.2 (right) taken at a one-hour interval at the faculty of mass communication building.

	timeIn	energyConsumtion
0	13/3/2018 16:50	253480.720
1	13/3/2018 17:50	253547.936
2	13/3/2018 18:50	253595.568
3	13/3/2018 19:50	253637.712
4	13/3/2018 20:50	253680.352
...	...	...
90747	21/8/2021 11:03	1233582.976
90748	21/8/2021 11:18	1233586.816
90749	21/8/2021 11:33	1233590.912
90750	21/8/2021 11:48	1233594.880
90751	21/8/2021 12:04	1233599.232

	pm10	pm2.5	temp	humid	timestamp
13581	54	44	30.0	27	2020-02-07 13:00:00
13580	53	43	30.0	27	2020-02-07 14:00:00
13579	44	36	30.0	24	2020-02-07 15:00:00
13578	45	37	31.0	24	2020-02-07 16:00:00
13577	43	36	31.0	32	2020-02-07 17:00:00
...	...	...	...	...	...
4	2	1	29.0	66	2021-09-25 23:00:00
3	1	1	29.0	67	2021-09-26 03:00:00
2	1	1	29.0	68	2021-09-26 04:00:00
1	4	3	29.0	69	2021-09-26 05:00:00
0	1	1	29.0	68	2021-09-26 06:00:00

Fig.2 Electricity consumption and weather datasets

3.2 Methodology

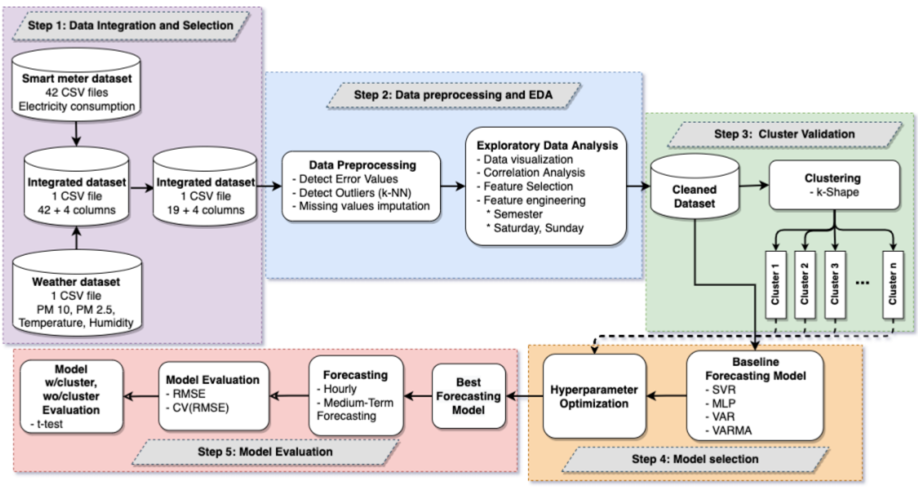

Fig.3 The framework overall design

The research framework is depicted in Fig.3. It is divided into five steps. The first step is to integrate and select data. The smart meter and weather datasets will be downloaded from websites. On the website, 42 smart meter datasets are available. Each smart meter dataset includes a single CSV file containing each building's electricity consumption. Then, in the same timestamps, combine electricity consumption datasets with a weather dataset that includes PM 10, PM 2.5, temperature, and humidity. As a result, the integrated dataset will have 46 columns, 42 from the smart meter dataset, and 4 from the weather dataset. Some smart meter datasets have too many missing values to conduct the experiment. Finally, the integrated datasets include 23 columns with missing values that are less than 20% of the total value, 19 columns from electricity consumption datasets, and 4 columns from the weather dataset.

Data preprocessing and exploratory data analysis are the next steps. The raw dataset must be cleaned first by detecting extreme values, also known as error values, from data collection. Then, detecting outliers in electricity consumption, which occur as a result of humans consuming abnormally large amounts of electricity. Then, using feature engineering, create new feature columns to help the models determine whether the day is a holiday, weekend, exam day, or semester day. Following completion of data preprocessing, the cleaned dataset was used to investigate the relationship between input features and electricity consumption in order to eliminate unneeded input features. Because reducing the number of input variables has the potential to reduce modeling computational costs while improving model performance in some cases.

The third step is to cluster multiple buildings into groups using the whole cleaned data from the second step using k-Shape methods, which will inform us of the number of clusters, and which cluster each building should be assigned to. Then each cluster will be used to train the forecasting models in the next step.

The fourth step will divide each cluster of buildings' electricity data into a training and testing dataset. The forecasting models will be constructed using the training dataset. And the forecasting models will be evaluated using the testing dataset. Four base line forecasting models (MLP [17], SVR [16], VAR [14], and VARMA [14]) will be used to forecast each cluster of building electricity consumption data in order to evaluate the models' performance prior to tuning the hyperparameters. After that, perform hyperparameter optimization to determine the optimal set of forecasting model hyperparameters. Then, the most effective forecasting model is chosen from the three tuned forecasting models to forecast unseen data from the testing dataset.

Finally, model evaluation is the last step. The performance of the most effective forecasting models with and without clustering technique will be compared. The t-test score is used to determine whether the forecasting model with clustering is more effective than the forecasting model without clustering. The error between predicted and actual values is calculated using evaluation metrics including RMSE, and CV(RMSE).

This research used Python for programming which Pandas library is used for manipulating the data and data imputation, Pycaret library is used for anomaly dectection, Statmodel and sklearn library are used for building forecasting models .

1) Data Integration

We combined 42 buildings' electricity consumption datasets and a weather dataset into a single dataframe by creating an dataframe and filling values into its columns. The reasons are buildings’ electricity consumption and weather data were collected at different time period. the electricity consumption data has 1-hour, 15-minute, and 5-minute intervals in each dataset. The data on electricity consumption was not collected at the appropriate time (slightly delay). Then, we filled the values from the closest time in each dataset of power usage and weather have been put into the empty dataframe. we iterated through all of the timestamps in the empty dataframe's "Time" column in Fig.4, comparing them to the "timeIn" and "timestamp" columns of the electricity consumption and weather datasets in Fig.2, respectively.

	Time	pm10	pm2.5	temp	humid	accba	...	rech_society	rech_argo	erdi	test_animal	vet	food_indus
0	2020-02-07 00:00:00	NaN	NaN	NaN	NaN	1250305.79	...	NaN	53724.34	333785.89	1928848.90	NaN	1328596.86
1	2020-02-07 01:00:00	NaN	NaN	NaN	NaN	1250324.61	...	NaN	53726.22	333800.83	1928923.26	NaN	1328677.25
2	2020-02-07 02:00:00	NaN	NaN	NaN	NaN	1250342.53	...	NaN	53728.20	333814.78	1928994.18	NaN	1328755.58
3	2020-02-07 03:00:00	NaN	NaN	NaN	NaN	1250360.83	...	NaN	53730.15	NaN	1929065.60	NaN	1328833.28
4	2020-02-07 04:00:00	NaN	NaN	NaN	NaN	NaN	...	NaN	53732.06	333840.06	1929135.62	NaN	1328911.23
...	...	...	...	...	...	...	...	...	...	...	...	...	...
13484	2021-08-21 20:00:00	4.0	3.0	32.0	61.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
13485	2021-08-21 21:00:00	3.0	2.0	31.0	60.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
13486	2021-08-21 22:00:00	3.0	2.0	31.0	60.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
13487	2021-08-21 23:00:00	3.0	2.0	31.0	61.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
13488	2021-08-22 00:00:00	2.0	2.0	31.0	61.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN

Fig.4 The dataframe after filling values

Then, columns were checked how many it has missing values that account for more than 20% of the total number of values and remove those columns.

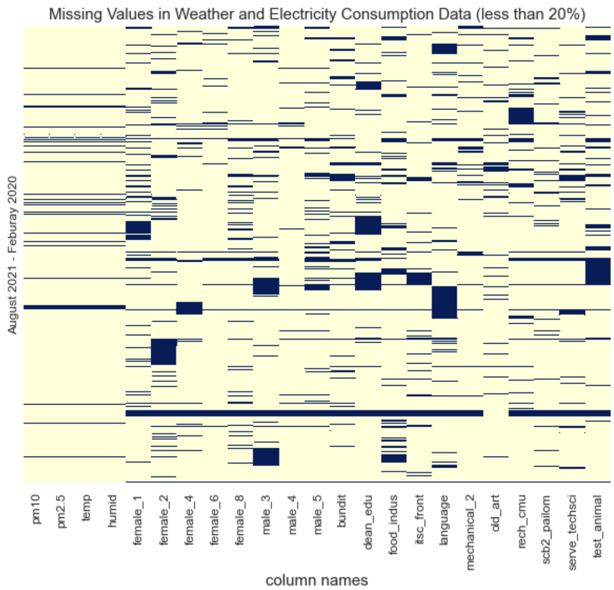

Fig.5 Columns with missing values that account for less than 20% of total rows

2) Data preprocessing

In this study, we removed extreme values and outliers in the dataset. Extreme values may occur as a result of smart meter data collecting errors. Electricity consumption data is normally collected by smart meters as cumulative data, which is the sum of values over time. Outliers are abnormal values in a dataset that do not match the normal distribution and occur from people using unusual electricity consumption which is not from the error of smart meter collecting the data. Both can have an impact on the predictive model's performance. As a result, we manually removed extreme values and used k-NN algorithms to detect and remove outliers [15] [19]. Finally, we imputed missing values by using linear interpolation method. The cleaned dataset contains 13489 rows and 23 columns including 4 columns for weather data and 19 columns for electricity consumption

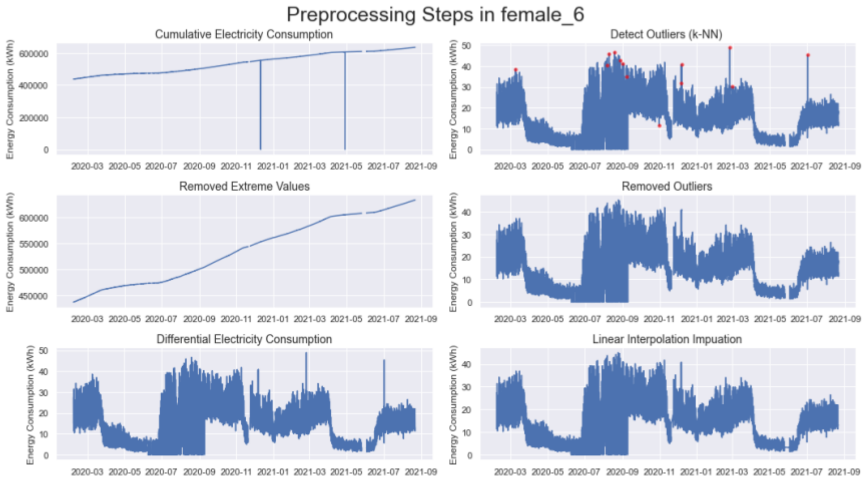

Fig.6 Data preprocessing steps

3) Clustering analysis

The k-shape technique is used to cluster data on electricity consumption [6] [18]. The distance of each building's electricity consumption to its cluster centroid is compared to 1 to 6 clusters as shown in Fig.7 (left). The elbow method, which chooses the curve's elbow as the number of clusters to utilize, is used to select how many clusters should the data be clustered. The graph shows that the distance between data and centroids does not diminish rapidly after three clusters. Therefore, the electricity consumption was clustered into three clusters as shown in Fig.7 (right)

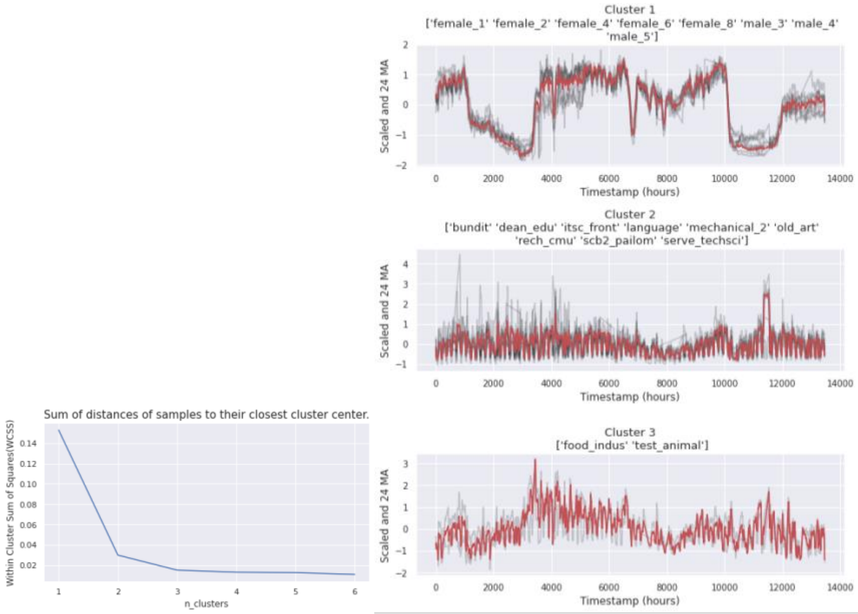

Fig.7 Sum of distances of samples to their closest cluster center (left) and the electricity consumption data is grouped into three clusters (right)

4) Time Series Cross Validation

Cross-validation is a resampling technique that avoids overfitting and selection bias by training and testing a model on different datasets. It is also used when optimizing hyperparameters of forecasting models. This study divides the data into 5 different datasets using 5 folds cross validation. Furthermore, because time series data have a temporal dependence between observations, the cross-validation method cannot randomly pick and assign values to the training and testing sets. The training set includes only of observations that occurred before the test set observation. Therefore, no future observations can be included into the forecast. Fig.8 illustrates 5 folds of time series cross validation.

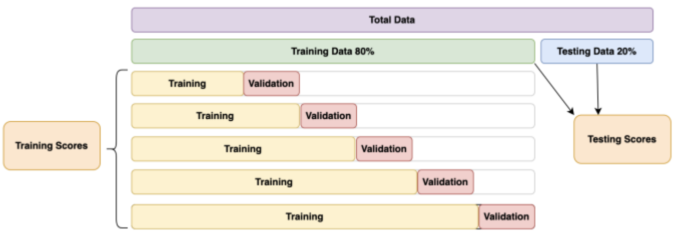

Fig.8 Time Series Cross Validation

4 Results

4.1 Cross Correlation between electricity consumption and weather data

Cross correlation is a measure of similarity between two series that tracks the movements of two or more sets of time series data relative to one another. The Pearson's correlation function is used to calculate the correlation between weather data and electricity consumption data for each lag time lag between -4 and 4 as shown in Fig.. The results showed that there was a weak correlation between PM2.5 and electricity consumption. Likewise, the temperature is not strongly impact to the electricity consumption directly because not every building's electricity consumption has the same direction of correlation. The dormitory buildings had a negative correlation while the office and laboratory buildings had a positive correlation to ambient temperature. Since temperature is variant from time in the day and occupants in the buildings have time as activity specification such as students leave their dormitory to go to the class and users in the office and laboratory buildings consume electricity in daytime where the ambient temperature begins to hot. On the other hand, when the ambient temperature drops in the evening and at night, students go back to their dorms and people in office and laboratory buildings leave the buildings.

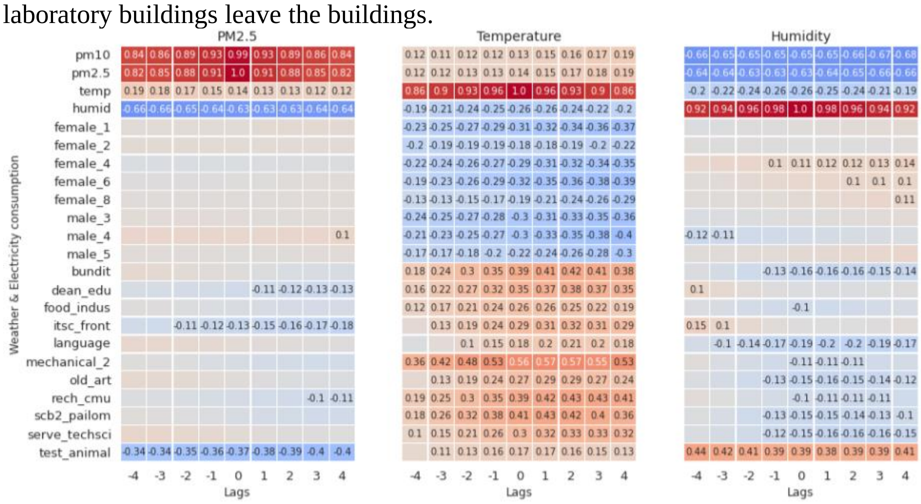

Fig.9 Cross-correlation of PM2.5, temperature and humidity to other columns

4.2 Model algorithm selection

The result of cross correation analysis indicated that PM10 and PM2.5 have a strong correlation with each other. They also have similar correlation to buildings electricity consumption. As a result, no PM10 data was used to train the forecasting models. The forecasting models were trained using PM2.5, temperature, and humidity data.

Table 1 shows training scores (the average and standard deviation of 5-fold cross validation scores) as well as testing scores (the results of evaluating models forecast on unseen data). In terms of training scores, the SVR model outperformed the VAR, VARMA, and MLP models. It indicated that the SVR model can forecast better in different size of training set and different time of testing set. Therefore, the SVR model is selected to develop electricity consumption forecasting model in further.

Table 1. Validation and testing scores of baseline forecasting models (RMSE)

Building	Training Scores (Validation)				Testing Scores			
Names	VAR	VARMA	SVR	MLP	VAR	VARMA	SVR	MLP
female_1	9.5 ± 2.3	9.6 ± 2.2	***8.0 ± 4.0**	14.9 ± 5.9	10.5	10.5	9.7	***6.3**
female_2	8.0 ± 1.7	8.0 ± 1.8	***6.6 ± 3.2**	8.7 ± 2.3	6.38	6.47	7.79	***5.5**
female_4	9.3 ± 2.9	9.3 ± 2.7	***8.0 ± 4.7**	11.5 ± 4.3	8.10	8.54	7.63	***5.1**
female_6	10.2 ± 3.4	10.2 ± 3.4	***8.1 ± 4.9**	13.9 ± 5.5	7.8	7.7	8.2	***5.5**
female_8	11.9 ± 3.7	11.7 ± 3.4	***9.8 ± 5.8**	15.3 ± 5.7	12.4	12.3	10.7	***9.6**
male_3	11.1 ± 3.1	11.1 ± 3.1	***8.9 ± 5.10**	12.8 ± 3.9	9.42	9.46	***6.1**	12.2
male_4	11.1 ± 3.3	11.1 ± 3.27	***9.2 ± 4.08**	11.8 ± 4.5	9.9	10.0	8.0	***5.4**
male_5	9.6 ± 3.2	9.6 ± 3.2	***7.8 ± 4.57**	14.0 ± 8.3	7.9	8.1	10.0	***6.5**
bundit	30.5 ± 7.3	30.6 ± 7.3	***28.1 ± 13.0**	38.2 ± 11.5	34.7	34.7	***31.3**	36.0
dean_edu	31.3 ± 7.8	31.2 ± 7.9	***28.8 ± 12.7**	41.9 ± 13.0	38.4	38.4	***35.7**	40.2
food_indus	60.5 ± 8.7	60.4 ± 8.9	***56.2 ± 17.4**	78.1 ± 22.4	60.2	60.5	***53.7**	69.8
itsc_front	25.9 ± 7.1	26.0 ± 7.0	***25.9 ± 8.6**	33.3 ± 14.9	30.6	30.7	26.1	***25.9**
language	19.1 ± 4.4	18.8 ± 4.4	***18.6 ± 4.9**	42.0 ± 21.1	18.3	18.3	***16.1**	27.5
mechanical_2	7.1 ± 0.9	7.1 ± 0.9	***5.3 ± 2.5**	8.8 ± 3.5	8.5	8.6	***6.0**	8.2
old_art	24.4 ± 4.2	24.5 ± 4.0	***21.8 ± 9.0**	28.8 ± 7.0	21.4	21.5	***16.4**	31.9
rech_cmu	18.0 ± 4.7	18.0 ± 4.6	***16.3 ± 8.3**	42.3 ± 29.5	26.0	26.0	***25.1**	28.6
scb2_pailom	72.0 ± 9.7	71.9 ± 9.7	***66.4 ± 18.3**	87.3 ± 39.7	97.7	97.7	***92.7**	104.0
serve_techsci	4.1 ± 0.7	4.1 ± 0.7	***3.7 ± 0.9**	7.4 ± 3.1	7.6	7.6	7.4	**6.7**
test_animal	47.0 ± 17.9	46.8 ± 18.3	***42.9 ± 22.6**	96.8 ± 73.5	***38.2**	38.3	43.5	64.3

4.3 Hyperparameter optimization

To optimize SVR hyperparameters, it is required to search through a large number of C and gamma values in order to find the one that yields the lowest RMSE score. Deeper searching hyperameters from large scale to super-fine scale of C and gamma is shown in heatmaps in Fig.9. Furthermore, validation curve and learning curve are also employed to monitor how well the SVR model perform while it is trained. The validation curve as shown in Fig.10 is a diagnostic tool to show good the sensitivity between to changes in machine learning model's accuracy with change in C and gamma hyperparameter of the SVR model. Learning curve as shown in Fig.11 plots the RMSE scores and model fitting times (right) over varying numbers of training samples.

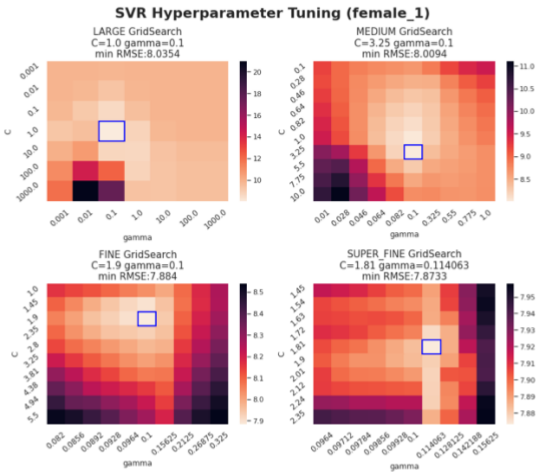

Fig.9 An example of large to super-fine scale of hyperparameter optimization

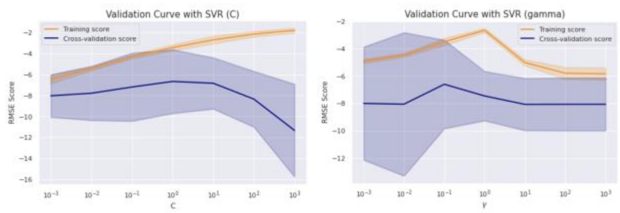

Fig.10 RMSE with change in gamma and C (Validation curve)

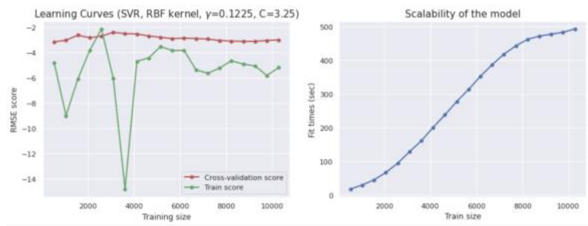

Fig.11 RMSE and training time with change in training data size (Learning curve)

4.4 Evaluating

Table 2 showed the result of comparing the forecasting performance of baseline, tuned without clustering, and tuned with clustering SVR models. The fifteen average RMSE scores of tuned models without clustering were lower than those of the baseline model. Consequently, the tuned models perform better in forecasting than the baseline models for validation scores. However, the four validation scores of tuned SVR models trained by centroids of clusters were lower than those of tuned SVR models without clustering. In term of testing scores, unlikely, the baseline SVR models can outperform the tuned SVR models on six of nineteen datasets. However, tuned SVR models without clustering still outperformed other models with 10 lowest RMSE scores on testing dataset.

Table 2 Validation and testing scores of baseline, tuned without clustering SVR, tuned with clustering SVR models (RMSE)

Building Names	Validation Scores			Testing Scores		
	Baseline SVR	Without Cluster Tuned SVR	With Cluster Tuned SVR	Baseline SVR	Without Cluster Tuned SVR	With Cluster Tuned SVR
female_1	8.0 ± 4.0	***7.9 ± 3.6**	8.5 ± 4.8	***9.7**	10.2	13.3
female_2	6.6 ± 3.2	***6.5 ± 2.6**	6.8 ± 3.5	***7.8**	10.1	9.3
female_4	8.0 ± 4.7	***7.7 ± 3.9**	7.9 ± 4.4	***7.6**	10.6	8.9
female_6	8.1 ± 4.9	8.0 ± 4.6	***7.8 ± 4.8**	8.2	***8.0**	10.7
female_8	9.8 ± 5.8	***9.4 ± 5.2**	9.9 ± 6.7	***10.7**	13.5	14.1
male_3	8.9 ± 5.1	8.7 ± 5.9	***8.0 ± 3.9**	6.1	***6.1**	9.3
male_4	9.2 ± 4.1	9.0 ± 3.7	***7.7 ± 4.1**	8.0	***5.3**	11.3
male_5	7.8 ± 4.6	***7.4 ± 3.4**	8.1 ± 4.9	***10.0**	12.3	11.9
bundit	28.1 ± 13.0	***25.4 ± 11.5**	27.5 ± 11.2	31.3	***28.6**	31.6
dean_edu	28.8 ± 12.7	***25.7 ± 10.8**	29.2 ± 10.8	35.7	***31.7**	33.6
food_indus	56.2 ± 17.4	***52.1 ± 16.8**	58.7 ± 19.1	***53.7**	54.3	58.1
itsc_front	25.9 ± 8.6	***22.2 ± 8.3**	22.6 ± 8.0	26.1	***23.0**	27.8
language	18.6 ± 4.9	17.0 ± 4.8	***17.0 ± 4.1**	16.1	***16.3**	15.7
mechanical_2	5.3 ± 2.5	***5.2 ± 2.4**	6.4 ± 2.0	6.0	***5.7**	8.0
old_art	21.8 ± 9.0	***19.7 ± 7.0**	21.4 ± 6.5	16.4	17.8	***15.6**
rech_cmu	16.3 ± 8.3	***15.5 ± 7.4**	15.8 ± 6.8	25.1	25.3	***23.1**
scb2_pailom	66.4 ± 18.3	***54.6 ± 22.7**	61.9 ± 18.4	92.7	***72.6**	89.4
serve_techsci	3.7 ± 0.9	***3.4 ± 0.7**	3.6 ± 0.9	7.4	6.9	***6.7**
test_animal	42.9 ± 22.6	***41.6 ± 20.6**	41.7 ± 17.7	43.5	***41.9**	45.2

The purpose of the hypothesis test is to investigate whether the performance of tuned SVR models trained with centroids ($RMSE_{centriods}$) has a statistically significant relationship with the performance of tuned SVR models trained with each building's electricity consumption ($RMSE_{building}$). As shown in Table 3, the results of hypothesis testing reveal that $RMSE_{centriods}$ is not statistically significantly more than or less than $RMSE_{building}$.

Table 3 Result of T-test between the results with/without clustering for each building

Null Hypothesis ($\boldsymbol{H_0}$)	Alternative Hypothesis ($\boldsymbol{H_1}$)	P-value	Results
$\boldsymbol{RMSE_{centriods} = RMSE_{building}}$	$RMSE_{centriods} \neq RMSE_{building}$	0.7258	**Fail to reject $\boldsymbol{H_0}$**
$\boldsymbol{RMSE_{centriods} \geq RMSE_{building}}$	$RMSE_{centriods} < RMSE_{building}$	0.3629	**Fail to reject $\boldsymbol{H_0}$**
$\boldsymbol{RMSE_{centriods} \leq RMSE_{building}}$	$RMSE_{centriods} > RMSE_{building}$	0.6371	**Fail to reject $\boldsymbol{H_0}$**

References

1. Sustainable Development Goals: United Nations Development Programme. Sustainablevelopment Goals | United Nations Development Programme. (n.d.). https://www.undp.org/sustainable-development-goals#affordable-and-clean-energy.
2. Shapi, M. M., Ramli, N. A., & Awalin, L. J. (2021). Energy consumption prediction by using machine learning for smart building: Case study in Malaysia. Developments in the Built Environment, 5. doi:https://doi.org/10.1016/j.dibe.2020.100037
3. Kim, M. K., Kim, Y. S., & Srebric, J. (2020). Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: Artificial neural network vs. linear regression. Sustainable Cities and Society, 62. doi:https://doi.org/10.1016/j.scs.2020.102385
4. Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy and Buildings, 147, 77-89. doi:https://doi.org/10.1016/j.enbuild.2017.04.038
5. Chae, Y. T., Horesh, R., Hwang, Y., & Lee, Y. M. (2016). Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. *Energy and Buildings*, 111, 184-194. doi:https://doi.org/10.1016/j.enbuild.2015.11.045
6. Yang, J., Ning, C., Deb, C., Zhang, F., Cheong, D., Lee, S. E., ... Sekhar, C. (2017). K-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. Energy and Buildings, 146, 27-37. doi:https://doi.org/10.1016/j.enbuild.2017.03.071
7. Dong, Z., Liu, J., Liu, B., Li, K., & Li, X. (2021). Hourly energy consumption prediction of an office building based on ensemble learning and energy consumption pattern classification. Energy and Buildings, 241. doi:https://doi.org/10.1016/j.enbuild.2021.110929
8. Yang, J., Ning, C., Deb, C., Zhang, F., Cheong, D., Lee, S. E., ... Sekhar, C. (2017). K-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. Energy and Buildings, 146, 27-37. doi:https://doi.org/10.1016/j.enbuild.2017.03.071
9. Chen, Y., Tan, H., & Berardi, U. (2017). Day-ahead prediction of hourly electric demand in non-stationary operated commercial buildings: A clustering-based hybrid approach. Energy and Buildings, 148, 228-237. doi:https://doi.org/10.1016/j.enbuild.2017.05.003
10. Fan, C., Xiao, F., & Wang, S. (2014). Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Applied Energy, 127, 1-10. doi:https://doi.org/10.1016/j.apenergy.2014.04.016
11. Shapi, M. M., Ramli, N. A., & Awalin, L. J. (2021). Energy consumption prediction by using machine learning for smart building: Case study in Malaysia. Developments in the Built Environment, 5. doi:https://doi.org/10.1016/j.dibe.2020.100037
12. Park, S. K., Moon, H. J., Min, K. C., Hwang, C., & Kim, S. (2018). Application of a multiple linear regression and an artificial neural network model for the heating performance analysis and hourly prediction of a large-scale ground source heat pump system. Energy and Buildings, 165, 206-215. doi:https://doi.org/10.1016/j.enbuild.2018.01.029

13. Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy and Buildings, 147, 77-89. doi:https://doi.org/10.1016/j.enbuild.2017.04.038
14. Harvey, A.C. (1990). ARIMA Models. In: Eatwell, J., Milgate, M., Newman, P. (eds) Time Series and Statistics. The New Palgrave. Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-20865-4_2
15. Dang, T. T., Ngan, H. Y. T., & Liu, W. (2015). Distance-based K-nearest neighbors outlier detection method in large-scale traffic data. 2015 IEEE International Conference on Digital Signal Processing (DSP). https://doi.org/10.1109/icdsp.2015.7251924
16. M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428.
17. S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, and classification," in IEEE Transactions on Neural Networks, vol. 3, no. 5, pp. 683-697, Sept. 1992, doi: 10.1109/72.159058.
18. Paparrizos, J., & Gravano, L. (2016). K-shape. ACM SIGMOD Record, 45(1), 69–76. https://doi.org/10.1145/2949741.2949758
19. Xu, H., Zhang, L., Li, P., & Zhu, F. (2022). Outlier detection algorithm based on K-nearest neighbors-local outlier factor. Journal of Algorithms & Computational Technology, 16, 174830262210781. https://doi.org/10.1177/17483026221078111