Short Term Load Forecasting in CAISO Electricity Market.

Martin Gao, Daniel Lee, Paul Maina, and Zohaib Siddique

Department of Computer Science, Department of International Relations and Affairs, Tufts University, 161 College Ave, 02155 Medford, MA, USA; martin.gao@tufts.edu; daniel.c_lee@tufts.edu; paul.maina@tufts.edu; zohaib.siddique@tufts.edu

Abstract

Short-term load forecasting is a critical task for electricity suppliers and is increasing in importance with the emergence of deregulation in electricity markets and the advent of smart grids. Many operating decisions rely on load forecasts such as dispatch scheduling of generation capacity reliability analysis of the grid and maintenance planning for generators. This paper proposes the use of a two-stage K-means clustering for variable selection and then using decision trees and support vector regressors for day-ahead load forecasting in the CAISO electricity market. The clustering process is a variable selection technique to select representative cities with diverse weather and electrical consumption patterns with which to create the models. The availability of both city-level weather data and state-level energy demand data at an hourly resolution encourages a scientific approach to quantify the impact of weather on electric load. This is hugely beneficial to the modeling process because a list of cities with differing weather profiles and electrical consumption patterns can be selected. Once the cities are selected, the cities' weather variables such as temperature, relative humidity, wind speed, dew point, and cloud type are input into a training process to build the models. The models are then trained on hourly data from the CAISO Electricity market from 2016 to 2017 and tested on 2018 out-of-sample data. The results reflect the effectiveness of the proposed method as an average MAPE of 5.5% was obtained on the out-of-sample data.

Introduction

Electrical load forecasting is an essential input for the decision-making processes in the power industry. Over the past decades, numerous forecasting models have been built to forecast electricity parameters such as load and energy prices. Since CAISO opened its electricity market in 1998, short term load forecasting has been a vital process in providing information to system operators, planners, and energy traders to reduce the risk profile of decision-making and minimizing operating costs [1]. Load forecasting can be carried out at a short-term level and a long-term level. Short-term load forecasting is usually from one day to one week for day-ahead scheduling. Medium and long-term forecasts are often from one week to a year out and can be used to identify changing trends such as industrial and population growth [2]. The scope of this study is predicting hourly state energy demand for the day-ahead, which makes it a short-term load forecasting problem.

Weather is a driver in electricity consumption and has been used extensively in electrical load forecasting [3]. Though temperature is a frequently used weather variable, other variables such as humidity and wind speed have also been used. [4]. With load being distributed throughout the state, weather observations from a single city may inhibit load prediction accuracy for a wide geographical area [5] [6].

As a result, weather observations from multiple cities around the state of California are taken into consideration, but this leaves load forecasters with a specific question: how to use representative cities whose data is used to predict the hourly load for the state? Tripathi and Sahay looked at using

temperatures of a "high demand region" in an approach to forecast load in PJM and ISO-NE [6]. Our research differs from this approach and hopes to improve upon it by considering multiple cities. Additionally, forecasting load for PJM and ISO-NE requires demand forecasting over multiple states as the ISO's serve multiple states, whereas our research is focused only on the state of California. Hong et al. similarly, approaches the task of choosing representative weather data by proposing a comprehensive weather station selection and combination methodology based on in-sample fit error [7]. Dordonnat et al. approaches this problem in a similar manner but instead put-forth an innovative idea to have every weather station generate a unique forecast, implement an algorithm to combine forecasts, and choose the best combination [8]. Sobhani et al. considered multiple historically proposed methodologies and evaluated them based on comparison [9]. All the mentioned papers only look at weather data to forecast load and overlook city-level electrical consumption patterns by sector. Furthermore, the mentioned papers may not have considered variance of weather across cities because the cities were chosen primarily on performance and their forecasting ability against ISO demand. This may lead to a collection of regions that may not be representative of the different climates in the ISO. The methodology proposed in this study seeks to not only identify cities with different weather and electrical consumption patterns but also provide a diverse subset of cities for the load forecasting process in the state of California

Methodology

1.1 Data Inputs

The data consisted of three datasets and are displayed in table 1 below:

TABLE 1. Datasets used as input

Dataset and Resolution
Hourly Weather Data (City Level) [10]
Annual Electricity Consumption Patterns by Sector (City Level) [11]
Hourly CAISO Load Data (State-Level) [12]

A web scraping script was developed to scrape weather data from the national solar radiation database [13]. City-level weather data for cities in California were retrieved at hourly resolution from 2016 to 2018. The data had no missing fields and consisted of the following variables: city name, date, hour, temperature, relative humidity, dew point, wind speed, and cloud type. Data for some cities were not available. Out of 482 cities/municipalities in California, data were available for 451 cities.

Annual electricity consumption data at the city level was retrieved from the City Energy Data API at NREL.gov [11]. It provided annual electrical consumption data by sector – residential, industrial, and commercial. The dataset consisted of city-level consumption at Megawatts per hour (MWh). Like the previous dataset, the data was available for the 451 cities out of the total 482. The variables contained in this dataset were residential load (MWh), commercial load (MWh), and industrial load (MWh). Electrical consumption in the agriculture sector was not provided or specified in this dataset. A total load of each city was computed, which is the sum of all three sectors at the city level and three additional variables computed to represent the percentage of each sector out of the total load for each city.

Total load of city = City residential load + city industrial load + city commercial load. (1.1)

Percentage load of city per sector = Sector load of city/ Total load of city. (1.2)

The last dataset was the data of the response variable, the hourly energy demand data for the ISO/state of California. This data contained no missing fields or missing data.

Data exploration of the weather data set revealed correlations between the weather variables, as shown in Table 2 below.

TABLE 2. Correlation matrix of weather variables.

Correlations	Month	Day	Hour	Temperature	Relative Humidity	Cloud Type	Dew Point	Wind Speed
Month	1	0.01	0	0.19	-0.22	-0.14	-0.05	-0.06
Day	0.01	1	0	0	-0.02	-0.05	-0.03	0.02
Hour	0	0	1	0.11	-0.08	0	0.03	0.07
Temperature	0.19	0	0.11	1	-0.72	-0.27	0.2	0.15
Relative Humidity	-0.22	-0.02	-0.08	-0.72	1	0.3	0.46	-0.08
Cloud Type	-0.14	-0.05	0	-0.27	0.3	1	0.04	-0.03
Dew Point	-0.05	-0.03	0.03	0.2	0.46	0.04	1	0.07
Wind Speed	-0.06	0.02	0.07	0.15	-0.08	-0.03	0.07	1

Table 2 shows that the strongest correlations were -0.72 and 0.46 between temperature and relative humidity and between relative humidity and dew point, respectively. There were no other strong indications of correlation amongst the weather variables.

Data exploration of the city load data revealed a strong correlation between all the variables, as shown in Table 3 below.

TABLE 3. Correlation Matrix of Load by Sector

Correlations	Total Population	Residential Load	Commercial Load	Industrial Load
Total Population	1	0.99	0.95	0.87
Residential Load	0.99	1	0.95	0.85
Commercial Load	0.95	0.95	1	0.76
Industrial Load	0.87	0.85	0.76	1

Table 3 shows that the strongest correlations observed were between total population and residential load at 0.99, commercial load, and residential load at 0.95 and total population and industrial load at 0.87.

Given that load values within the sectors exhibited strong correlations, sector load values were calculated as a percentage of the total load. Weaker correlations between percentage sector load variables provided sufficient motivation to use percentage metrics vis a vis absolute load value. Table 4 displays the correlations after the sector values are computed as percentages of the total load.

TABLE 4. Correlation Matrix of Load by Sector (in Percentage)

Correlations	Total Population	% Residential Load from Total	% Commercial Load from Total	% Industrial Load from Total
Total Population	1.00	(0.08)	0.08	0.01
% Residential Load from Total	(0.08)	1.00	(0.38)	(0.58)
% Commercial Load from Total	0.08	(0.38)	1.00	(0.53)
% Industrial Load from Total	0.01	(0.58)	(0.53)	1.00

Table 4 shows that weaker correlations between variables are observed when using sector contribution to total load consumption. Correlation between percentage residential load and the total population was –0.08, and the strongest correlation was between the percentage of industrial load and percentage residential load at –0.58.

1.2 One-stage and Two-stage clustering methodology

K-means clustering was used to identify subgroups in both the weather and load data. K-means clustering guarantees a level of convergence, is easy to deploy and scales to large data [14] [15].

The objective was to identify cities that had similar weather characteristics and electricity consumption patterns and group them together. Then, one representative city is chosen from each cluster, which ensures a diversity of weather and electricity consumption patterns across the chosen cities.

Pearson correlation was used as a similarity metric in the one-stage clustering as strong weather data correlation was expected amongst neighboring cities. In the second stage of clustering, as a result of low correlations between variables, Euclidean distance was used as the similarity metric.

In one-stage clustering, city-level time-series weather data was first scaled using a standardized scaler for accounting for the difference in feature units and then clustered based on the scaled data, thus assigning each city to a specific cluster [16].

A python package algorithm, tslearn, was used in the one-stage clustering because the package is suitable for handling weather data as a time-series dataset [17]. The goal of the one-stage of clustering is to group cities into clusters with similar weather patterns. The elbow method was used to compute the optimal number of clusters. The optimal number of clusters was decided at 5. This is displayed in Figure **1.** **Map 1** shows the geographical locations of the weather clusters.

Figure 1: Elbow method to visualize optimal clusters for one-stage clustering

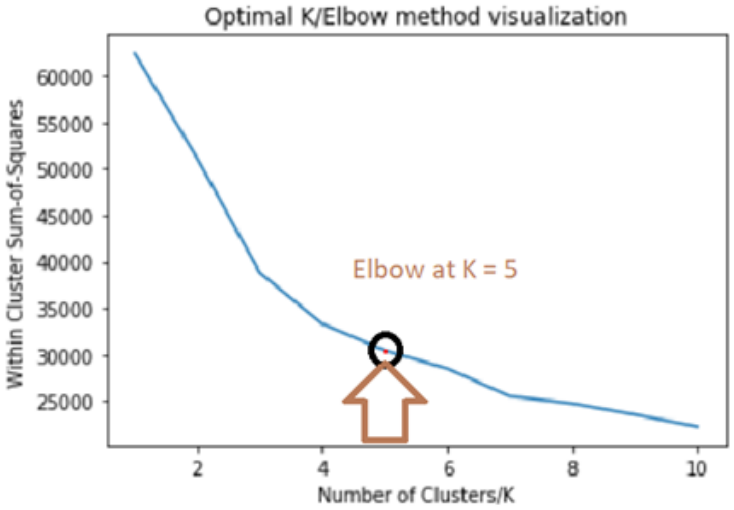
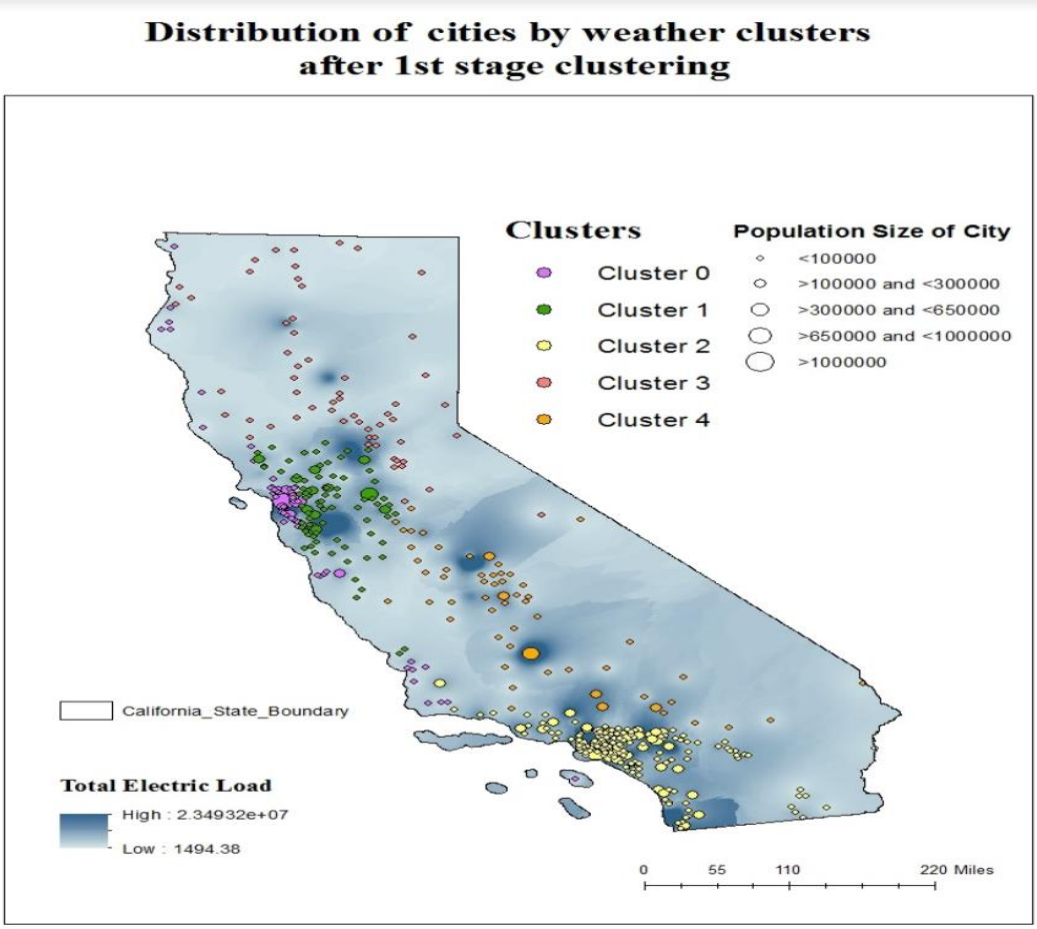


Figure 1: *The optimal number of clusters from one-stage clustering is 5.*

Map 1: City distribution by weather cluster after one-stage clustering



The two-stage clustering process groups cities further into sub-clusters of cities that have similar electricity load profiles. Electric load profile offers valuable insight into the productivity and economic profile of cities, which are critical drivers of electricity consumption. The elbow method is used to compute the optimal number of clusters. **Figure 2** displays the visualization of the elbow method for each weather cluster. **Map 2** shows the geographical locations of the cities in their sub-clusters. The clustering behavior and division into sub-clusters are shown in **Figure 3.**

Figure 2: Elbow method to visualize optimal clusters

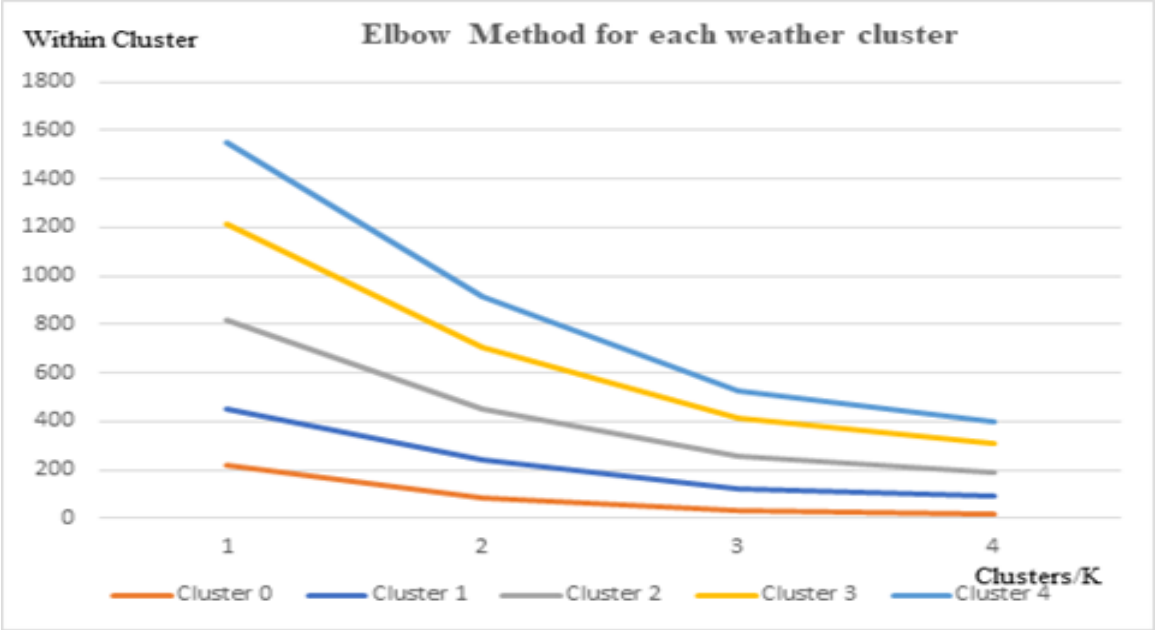
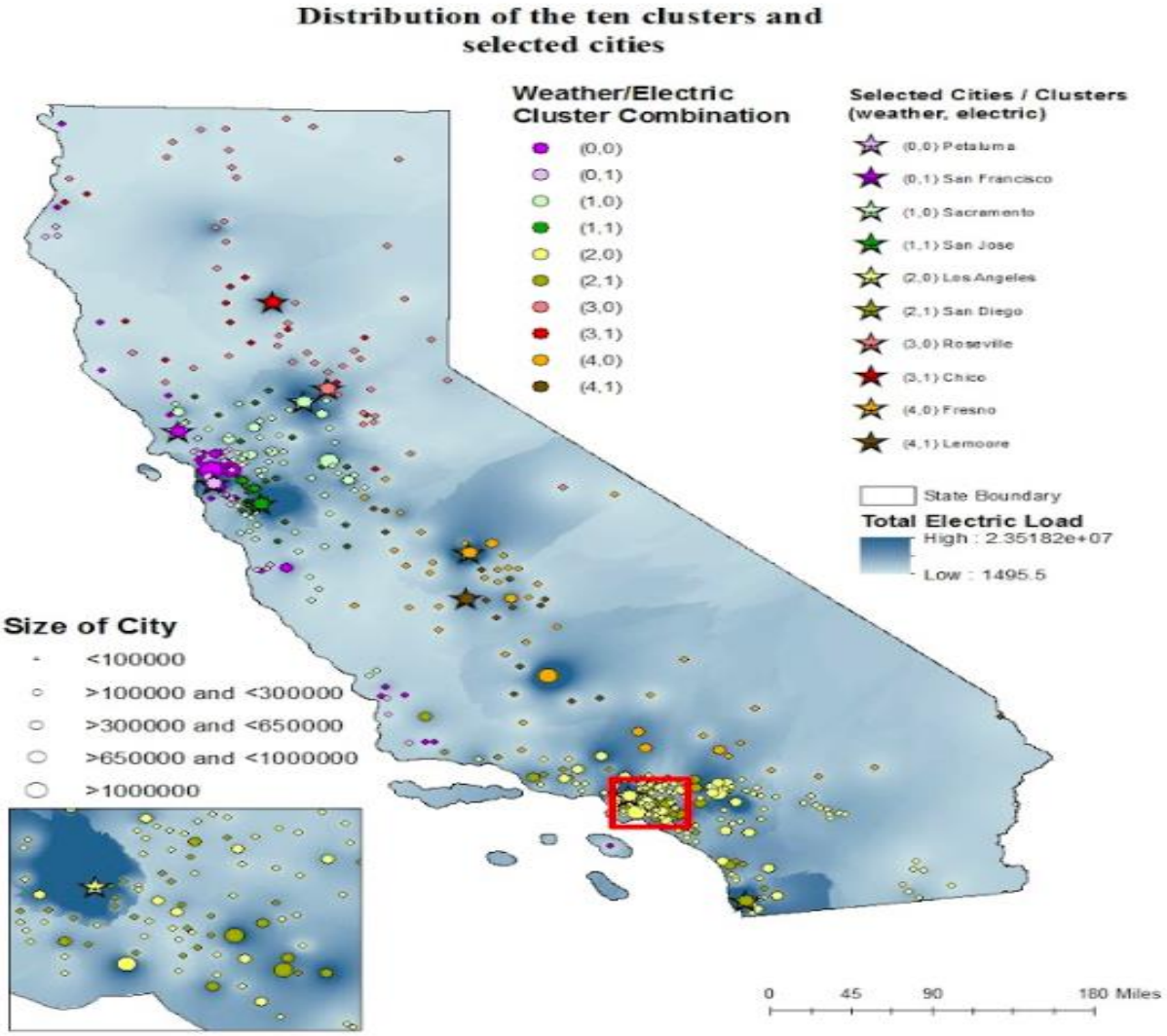


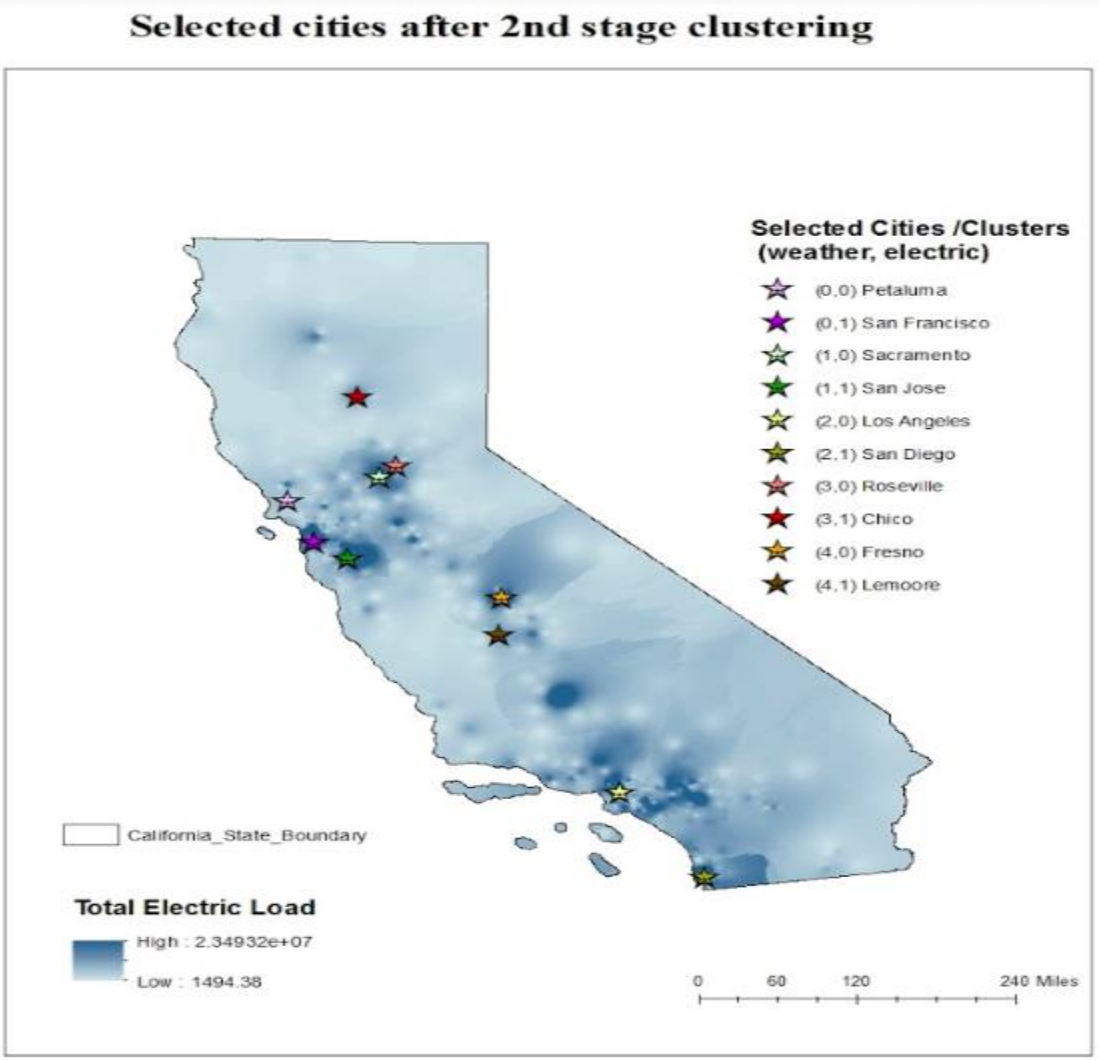
Figure 2*. The Optimum Number of sub-clusters across all five weather clusters is two.*

Once sub-clusters are generated, the selection of representative cities is based on the city with maximum electricity consumption (total load) amongst cities of the sub-cluster. Therefore, from the ten sub-clusters, ten cities are selected, and their weather features are used as input into the regression models. **Map 2** shows the segmentation of the cities based on weather and electric load clustering and **Map 3**, the final cities selected.

Map 2: City distribution after two-stage clustering



Map 3: Cities selected after second-stage clustering



Map 3 *shows the cities selected after the second stage of clustering. The weather variables of the selected cities are used to train the models.*

Figure 3: Cluster characteristics by percent load by sector

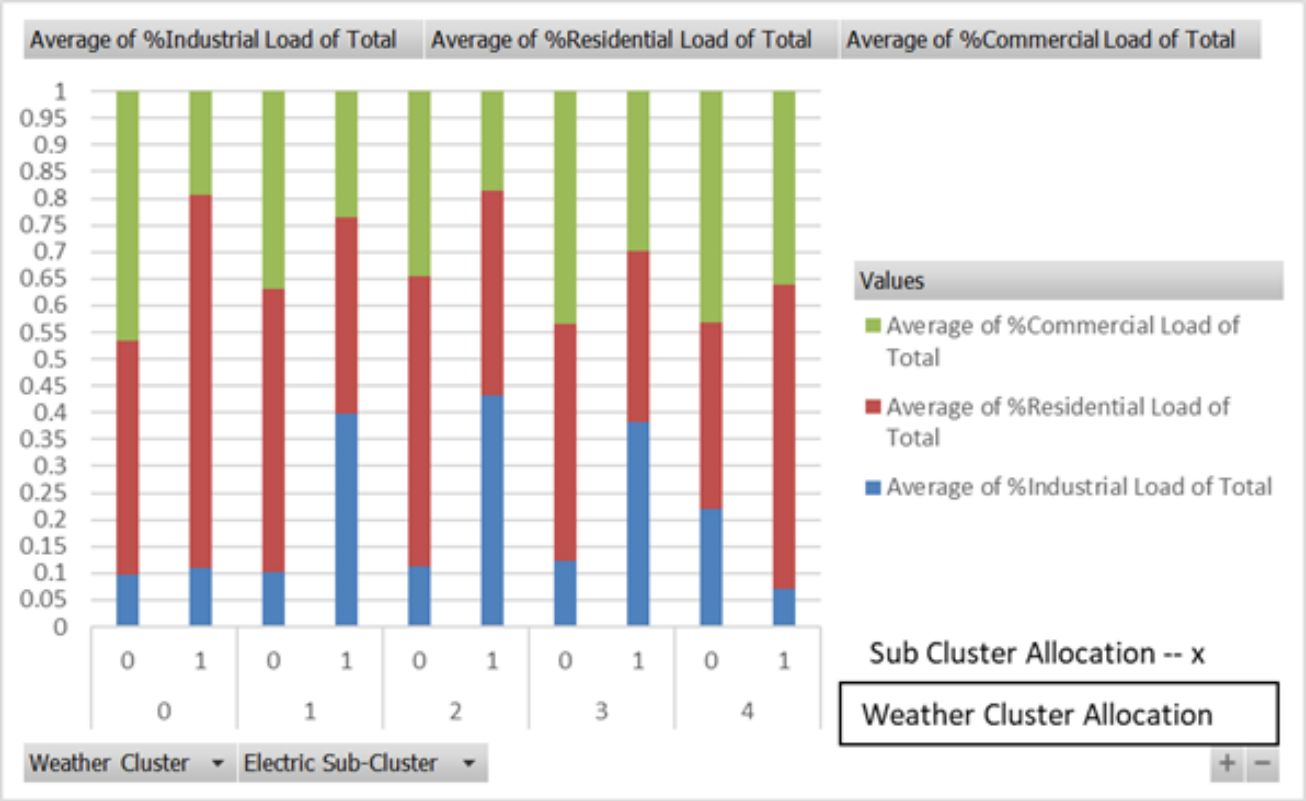


Figure 3 *shows the average percentage of load statistics by sector for the cities in those sub-clusters. The algorithm seems to subgroup the cities based on percentage residential. For example, in weather cluster 0, cities with lower percentage residential load are grouped to subgroup 0, while cities with higher residential percentage are grouped to subgroup 1.*

1.3 Regression Models

The models used in this study were decision trees, random forests, and support vector machines. The models were trained on hourly data from 2016 - 2017 and tested on out-of-sample 2018 data.

For comparison, the three models were trained on three city selection models, namely:

1. Training on data from the city of Los Angeles. Los Angeles was chosen for a lone model because it has the highest electric load in the state. Using one city is similar to Sahay and Tripathi's study, where one high demand region is used [6].
2. Training on selected cities from a one-stage clustering (clustering only on weather data). There are a total of five cities, one from each cluster.
3. Training on selected cities from the two-stage clustering (weather and electricity consumption statistics). There are a total of ten cities, one from each sub-cluster.

Decision trees and random forests were preferred because they perform better than multilinear regression on data with nonlinearity. In addition, features do not have to be scaled [18]. The disadvantage of a decision tree is that it is not as robust in prediction performance because trees are sensitive to small changes in data [19]. Although tuning a decision tree can be helpful, random forests provide more robust performance because it randomly selects a subset of predictors at every split. As a result, random forests help reduce the correlation between features, leading to less overfitting [18].

Support vector machines have a better predictive performance than multilinear regression on data with nonlinearity. Support Vector Machine adds structural constraints to find an optimal plane to classify the data. Since the weather and load data cannot be separated by a linear hyperplane, the radial based kernel was used to map the data into a higher-dimensional space [20]. Consequently, the algorithm minimizes overfitting and bias.

For the load forecasting process, the weather variables defined below were used (list 1):

1. Temperature
2. Relative Humidity
3. Dew Point
4. Wind Speed
5. Cloud Type

We also constructed time-based categorical variables to account for seasonality (list 2):

6. Hour of the day
7. Day of the week
8. Week of the year
9. Month

After specific cities were selected from the clustering methodologies, all the variables in lists 1 and 2 were used for variable selection to training the models. For example, in the two-stage clustering methodology, as 10 cities were chosen from the sub-clusters, then 54 variables would be included in the training and test set, out of which 50 of them would represent the weather variables' (Temperature, Relative Humidity, Dew Point, Wind Speed, Cloud Type) for the selected cities and the remaining would be the time-based categorical variables described above. The weather variables chosen are shown in the tables below:

TABLE 5. Representative cities and variables from one-stage clustering

Cluster #	City Name	Weather Variables selected for specific city
1	San Francisco	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
2	San Jose	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
3	Los Angeles	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
4	Roseville	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
5	Fresno	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed

Table 5. The cities selected from one-stage clustering and the variables of the cities used to train the models.

TABLE 6. Representative cities and variables from two-stage clustering

#	City Name	Weather Variables selected for specific city
1	Petaluma	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
2	San Francisco	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
3	Sacramento	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
4	San Jose	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
5	Los Angeles	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
6	San Diego	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
7	Roseville	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
8	Chico	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
9	Fresno	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed
10	Lemoore	Temperature, Relative Humidity, Cloud Cover, Dew Point, Wind Speed

Table 6. The cities selected from two-stage clustering and the variables of the cities used to train the models.

1.4 Model Parameters

For the support vector machines, the radial base function kernel was used. In addition four Cs of 0.1, 1, 10, 100, and 1000 and four gammas of 1, 0.1, 0.01, 0.001, 0.0001 were tested using a randomized search to find optimal hyperparameters. On the other hand, decision trees also used a randomized search process to identify the optimal number for the maximum leaf nodes, maximum depth, and maximum features for the tree. Finally, the random forest model used the same search method to generate the optimal maximum leaf nodes, number of estimators, maximum features, maximum depth, minimum sample leaf, and minimum sample split.

1.5 Error Measurement

As error measurement statistics play a critical role in tracking forecast accuracy, monitoring for exceptions, this study utilized MAPE (Mean Absolute Percentage Error) as the metric. MAPE measures the size of the error in percentage terms [21]. It is calculated as the average of the unsigned percentage error, as shown in the picture below [6]:

$$MAPE\ [\%] = \frac{1}{N}\sum_{i=1}^{N}\frac{|L_A^i - L_F^i|}{L_A^i} \times 100$$

In the above equation, L_A is the actual load, L_F is the load forecast, and N is the number of data points. The MAPE is calculated for every hourly prediction for all models.

For this paper, the models are trained with CAISO Electricity market data from the years 2016 to 2017 and tested on data from 2018.

Model Results

In this study, each of the following models—decision tree regression, support vector, and random forest —were trained on the two-stage clustering (clustering by weather and electricity consumption data), the one-stage clustering (clustering only by weather) and lastly on Los Angeles weather variables for comparison analysis. Since the two-stage clustering generated 10 sub-cluster and therefore 10 cities selected, the inputs into the two-stage clustering model were greater than the one-stage model, which had 5 cities, which could contribute to a lower MAPE for the two-stage clustering model. **Table 7** shows the average MAPE for the test set. The graphs below also provide a monthly breakdown of the MAPE by month.

TABLE 7: MAPEs of model type and clustering method

Clustering Method	MAPE Random Forest	MAPE Support Vector Regression	MAPE Decision Tree
One High Demand Region (LA)	7.79%	12.29%	10.57%
One-Stage Clustering	7.67%	10.34%	9.21%
Two-Stage Clustering	5.57%	8.13%	6.30%

Table 7 shows that models trained after two-stage clustering have the lowest MAPE

Figure 4: MAPE by model and month on LA

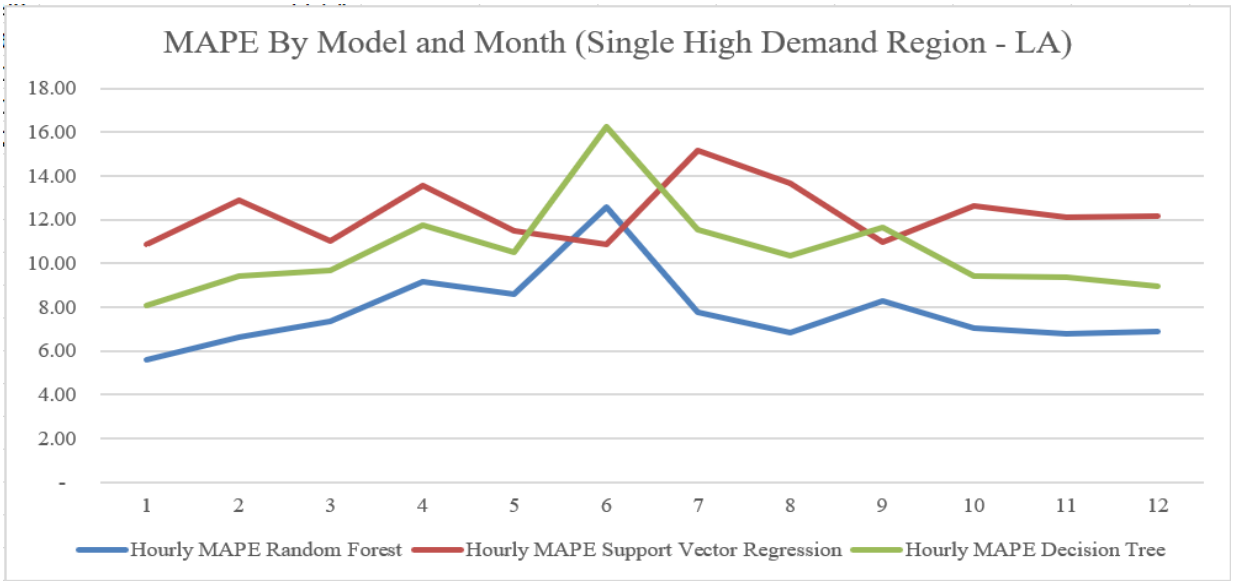


Figure 5: MAPE by model and month from first stage clustering.

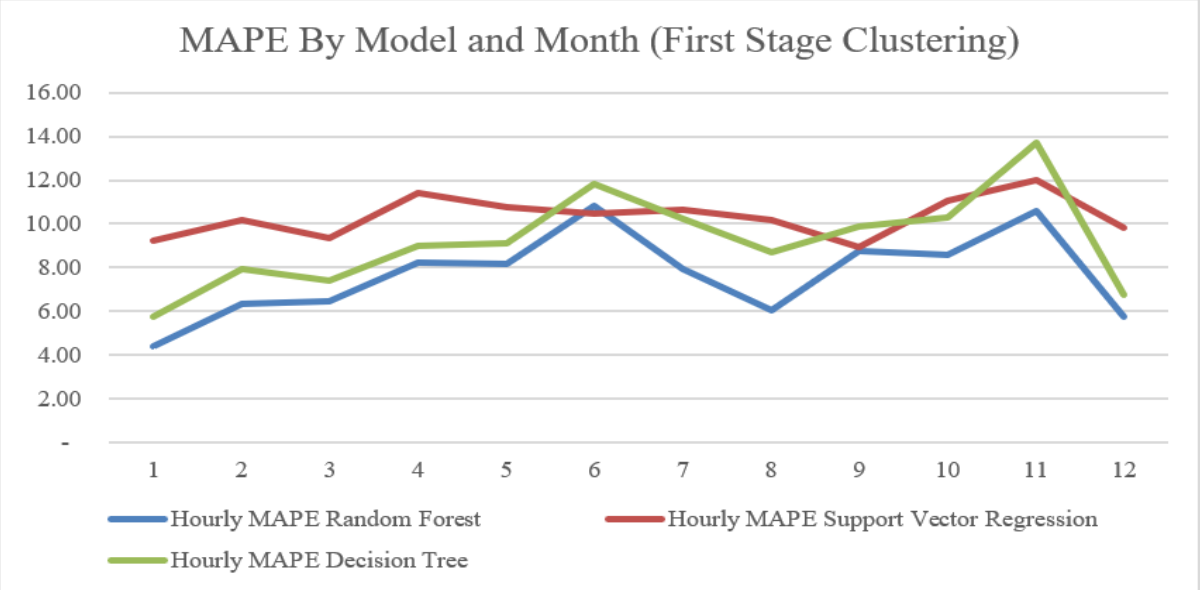
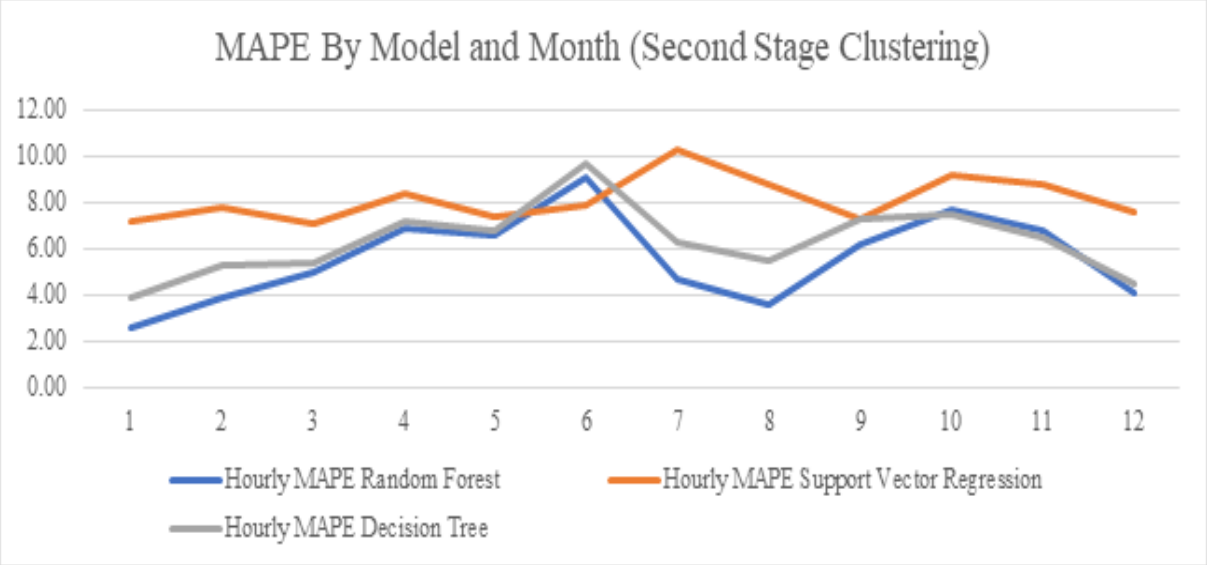


Figure 6: MAPE by model and month from two-stage clustering



From **Table 7** and **Figures 4, 5,** and **6**, the two-stage clustering methodology results in a better MAPE across all models. Random forest models generate the lowest MAPE of all three algorithms, and the random forest model with two-stage clustering produces the lowest MAPE (5.57%) out of all models.

Random forest with two-stage clustering generates the best performance because it does not consider all the predictors; hence it would not favor the strongest predictor out of the group of predictors [19]. As a result, random forests produce decorrelated trees and are essentially a collection of decision trees whose results are aggregated. As other models do not work to carry out feature selection, it is likely that other regression models show a higher MAPE.

Two-stage clustering chooses the greatest number of representative cities, and a greater number of city data could contribute to the lower MAPE. Ten representative cities are chosen from two-stage clustering, compared to five representative cities with one-stage clustering and only one representative city when we only use Los Angeles.

The top 20 features by importance from the two-stage clustering and the tree models are shown in **Figure 7** and **Figure 8** below:

Figure 7: Feature importance from two-stage clustering with random forests

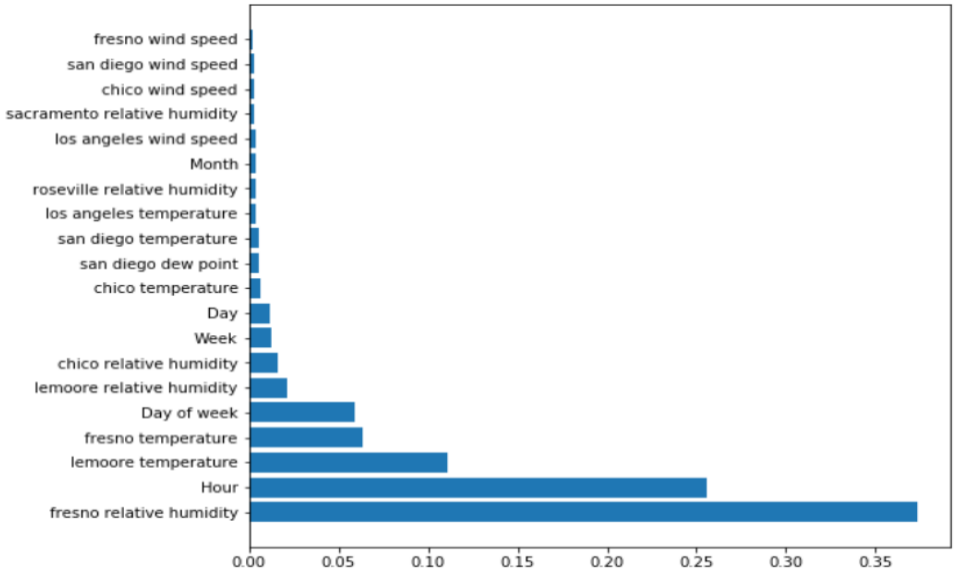
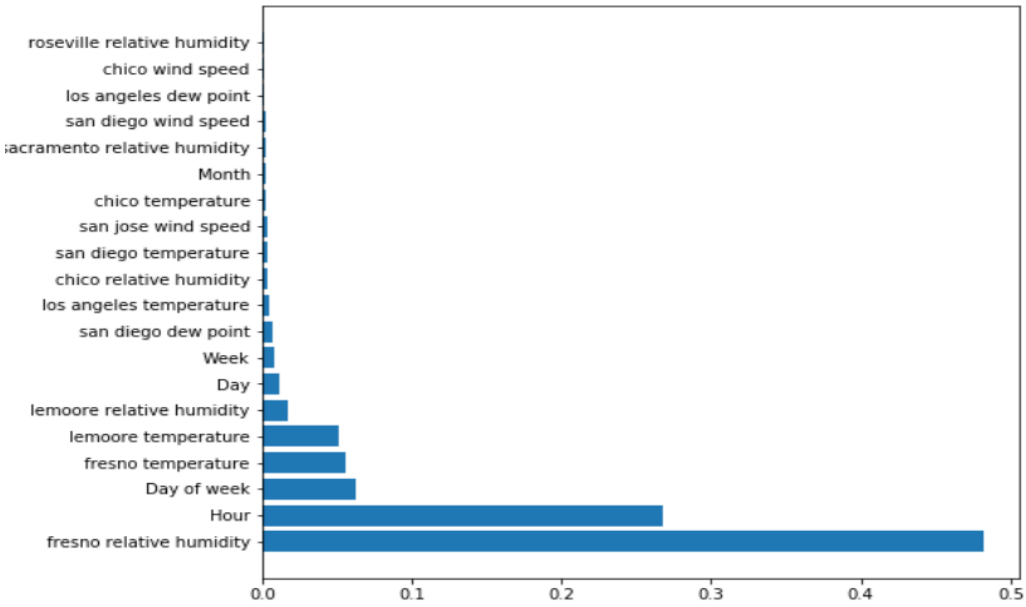


Figure 8: Feature importance from two-stage clustering and decision tree.



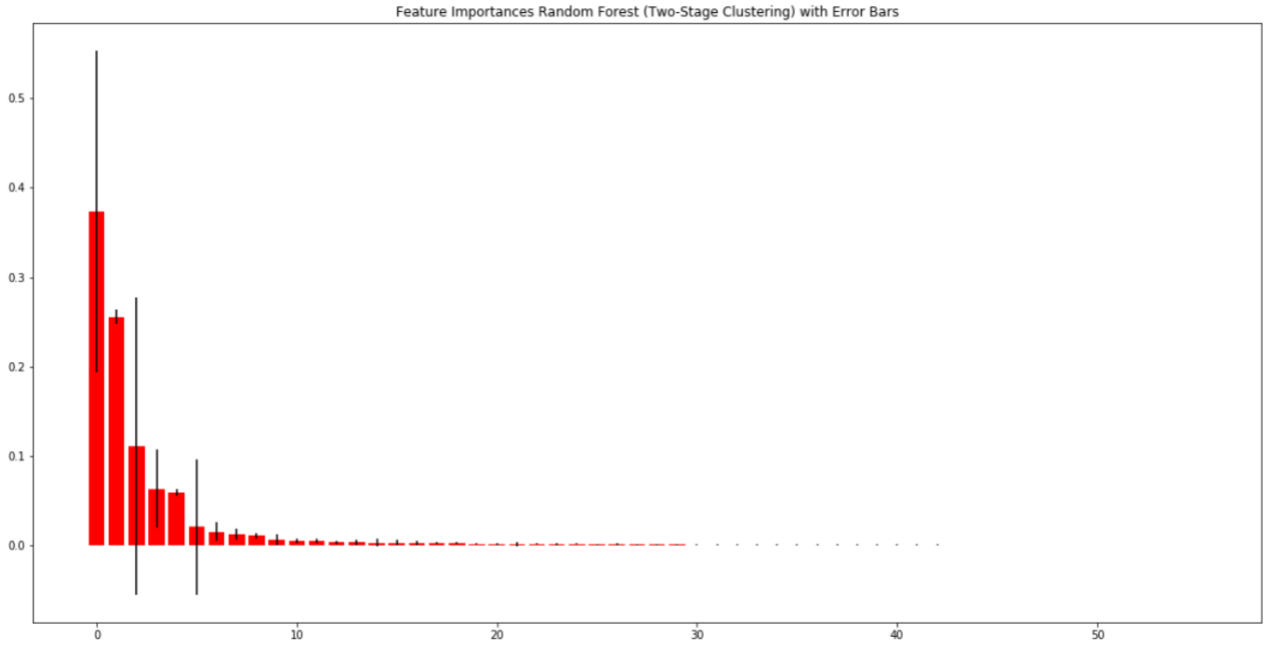


Figure 8*. Feature importance of random forest with error bar*

From the graphs, both the tree and random forest models shared similar feature importance in the top five. The inclusion of time variables also improved the models due to the presence of seasonality in the models. Regarding weather variables, relative humidity and temperatures in specific locations,

especially around Fresno and Lemoore, are considered more important than other city weather variables.

Another interesting finding in the MAPE charts was that the MAPEs seem to be high during the spring and early summer months, possibly because of increasing behind-the-meter solar measures during these seasons. As the models in this study do not consider any forecast of solar generation nor account for the increasing dependence on photovoltaic systems, it would support the view that the models fail to catch this important parameter.

In addition, the hours of the day during which solar generation is expected to be the highest, show higher MAPEs as well. The table below displays this finding and a visual comparison of how the model overpredicts the load curve during these hours.

TABLE 8: MAPE by month and hour

	MAPE by Month and Hour for Two-Stage Clustering + Random Forest											
Hour	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
0	2.17	3.63	2.68	3.79	2.83	3.9	5.25	4.29	4.34	3.05	7.32	2.25
1	2.65	3.49	3	4.41	3.33	3.64	4.37	2.58	4.83	3.01	7.76	2.07
2	2.05	3	3.38	5.28	4.28	4.79	4.33	2.53	5.51	3.63	8.5	2.42
3	2.05	2.93	3.92	5.78	4.75	5.6	3.9	2.03	6.3	4.05	8.97	2.75
4	2.4	3.67	2.81	4.65	3.73	5.44	3.39	2.17	5.13	2.96	7.23	2.49
5	1.82	3.17	2.19	3.99	3.24	5.57	2.9	2.46	4.4	2.52	5.37	2.76
6	1.96	3.35	2.05	3.14	3.02	6.01	4.52	3.27	3.67	2.1	3.77	3.75
7	3.06	3.57	2.38	5.23	5.07	8.2	3.34	2.1	3.92	3.29	4.85	2.93
8	2.52	3.76	2.44	6.22	5.5	8.58	4.52	2.07	5.02	4.31	4.5	2.99
9	2.18	4.73	3.39	8.25	6.74	10.39	5.35	2.14	6.5	6.42	5.39	3.89
10	2.91	6.08	5.24	9.96	8.38	12.61	6.14	2.35	9.17	7.14	6.76	5.8
11	3.76	7.2	6.95	11.67	10.49	14.68	6.4	2.97	10.79	8.78	7.38	7.64
12	4.1	8.09	8.78	12.94	11.74	15.55	6.27	2.78	9.96	9.5	7.54	8.66
13	3.36	8.99	9.77	12.51	11.63	15.92	7.12	3.74	9.25	10.75	9.11	8.3
14	3	9.98	10.75	11.91	12.08	18.62	8.58	6.02	12.57	19.62	17.28	6.52
15	2.62	8.72	9.64	10.57	11.55	16.54	8.93	7.09	11.07	19.5	16.89	3.99
16	2.14	2.57	8.06	8.97	9.32	12.53	9.79	8.21	9.75	10.1	10.75	3.02
17	2.21	2.97	8.08	9.52	11.87	11.41	8.8	7.34	7.49	7.23	6.74	2.79
18	2.27	2.57	5.45	5.22	6.54	8.94	8.04	6.55	5.68	3.64	6.69	2.82
19	2.02	2.95	2.49	3.05	4.12	6.16	6.46	5.03	4.66	3.36	8.5	3.6
20	2.36	3.29	2.56	1.91	2.35	4.46	5.36	4.55	4.65	4.88	7.01	3.82
21	1.74	3.27	3.03	3.17	3.64	4.35	4.51	2.63	6.54	6.71	7.39	2.62
22	1.51	3.55	2.73	3.62	3.48	4.79	5.57	3.57	8.35	8.4	6.38	2.08
23	1.82	3.03	3.22	4.3	4.06	6.2	8.03	5.7	10.11	8.57	7.01	2.64

Table 8: *Note the increase in MAPE in month 4 all through 6*

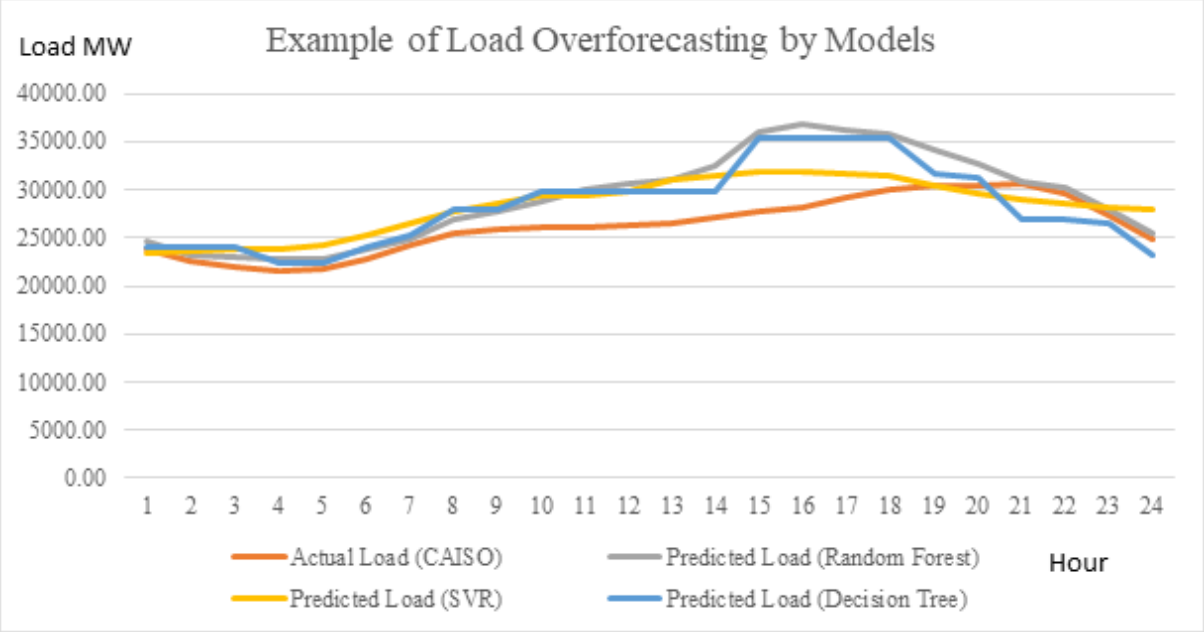


Figure 9. *On the particular day shown in the graph, it looks like the Support Vector model did not over-forecast as much as the others, but the distinction between the red curve (actual load) and the other models is clear. This means that the models are missing something during these hours and thus, resulting in the models over-forecasting possibly because of the duck curve phenomenon* [22].

Conclusion

From the models trained and tested within this study, results show that the two-stage clustering methodology, then a random forest model produces the best MAPE of all possible models. An annual average MAPE of 5.5% signals the viability of such models. The second stage of clustering (for choosing representative cities) decreased the error of the models. Training a random forest model on only one high demand region (LA) with no clustering had a MAPE of 7.79%. Using one-stage clustering on weather variables, the MAPE decreased to 7.67%. Finally, with further clustering using electric consumption statistics (two-stage clustering), MAPE decreased to 5.57%. This trend of decreased error with more stages of clustering is also consistent for the support vector regression and decision tree models. The significant reduction in error still signals the efficacy of the clustering method used.

In comparison to other studies, no other load forecasting studies were carried out in CAISO for the test year 2018. One source of reference is in Kaur and colleagues' research on load forecasting, which tested on the year 2014 [23]. The models use an ensemble of kernel-based gaussian processes where model parameters were selected using a genetic algorithm. Those models provided a 2.2% MAPE, which is better than the results of this study. However, the lower MAPE may also be a result of lower behind-the-meter generation present in California at the time, which this study hypothesizes as one of the main sources of error during this study.

There are some biases that may have affected the models.

Behind-the-meter solar generation: In California, Solar (Photovoltaic) PV and Solar Thermal generation account for approximately 14.1 GW of the capacity, of which approximately 9.4 GW is from Solar PV [24] [25]. California also exhibits an annual average capacity factor for Solar PV generation at about 28.4% [26]. This energy is not accounted for by the grid as CAISO currently does not measure behind-the-meter generation [27]. Thus, it is difficult to estimate the impact of actual behind-the-meter generation. The models in this study are also prone to over-predicting afternoon electrical load. With a solar mandate going to effect in 2020 in California, which requires a PV system as a power source for new construction homes, behind-the-meter solar generation will be an increasingly significant source of error if not taken into account in the models [28].

Microclimates in cities: The resolution of both the weather and electric load data was the city level. Some cities in California are large and exhibit multiple microclimates. For example, San Francisco is a city that is often studied for the environmental impacts of its differing microclimates [29] [30]. However, each city is treated homogeneously as a whole in our analysis. Large cities with microclimates can be broken up into more refined geographical regions.

Finally, there are limitations regarding transferring this model to other states because they have

- Different climate profiles.
- Might be geographically smaller states.

This constraint could possibly be overcome by re-tuning the models and by considering different weather variables and solar generation as well. For example, it can be logical to consider the amount of snowfall in the east coast states. It is possible to improve on the results by breaking up large cities into more refined geographical regions by gathering data and training models with higher resolution.

Based on the models, there is room for improving feature engineering for future work. While the random forest model approximates the process of feature selection through the random selection of features, the other models are lacking in this aspect and could be improved upon.

Incorporating behind-the-meter solar generation data into the forecasting model may lead to increased forecasting accuracy. Wildfires are prevalent in the state of California and may play a role in load forecasting. There were 8527 known wildfires in the year 2018 and 6872 known wildfires in 2019; thus, further research can be carried out by adding and implementing wildfire data into the models [31].

Bibliography

[1] California ISO, "“CaliforniaISO- News.”," [Online]. Available: http://www.caiso.com/about/pages/news/default.aspx. [Accessed 20 April 2020].

[2] K. Daware, "ElectricalEasy," [Online]. Available: https://www.electricaleasy.com/2016/06/types-of-electrical-loads.html.

[3] J. Xie, Y. Chen, T. Hong and T. Laing, "Relative Humidity for Load Forecasting Models," *IEEETrans. Smart Grid*, vol. 9, pp. 191-198, 2018.

[4] S. Fan and R. Hyndman, "Short-Term Load Forecasting Based on a Semi-Parametric Additive Model," *IEEE Trans. Power Syst.* , vol. 27, p. 134–141 , 2012.

[5] R. Nedellec, J. Cugliari and Y. Goude, "Electric load forecasting and backcasting with semi-parametric models," *Int. J. Forecast.* , no. 30, p. 375–381, 2014.

[6] K. B. S. a. M. M. Tripathi, "Day ahead hourly load forecast of PJM electricity market and ISO New England market by using artificial neural network," *IEEE Innovative Smart Grid Technologies-Asia*, pp. 1-5, 2013.

[7] T. Hong, P. Wang and L. White, "Weather station selection for electric load forecasting," *Int. J.Forecast*, no. 31, p. 286–295, 2015.

[8] V. Dordonnat, S. Koopman, M. Ooms, A. Dessertaine and J. Collet, "An hourly periodic state space model for modelling French national electricity load," *Int. J. Forecast*, no. 24, pp. 566-587, 2008.

[9] A. C. S. S. C. L. T. H. Masoud Sobhani, "Combining Weather Stations for Electric Load Forecasting," *Energies*, no. 12, 2019.

[10] National Renewable Energy Laboratory, "NSRDB Data Downloads," [Online]. Available: https://developer.nrel.gov/docs/solar/nsrdb/. [Accessed April 2020].

[11] National Renewable Energy Laboratory,, "Electricity and Natural Gas APIs," [Online]. Available: https://developer.nrel.gov/docs/cleap/elec_and_nat_gas/. [Accessed April 2020].

[12] California ISO, "Historical EMS Hourly Load Data Available," [Online]. Available: https://www.caiso.com/Documents/HistoricalEMSHourlyLoadDataAvailable.html. [Accessed April 2020].

[13] NREL, "API Instructions - NSRDB," [Online]. Available: https://nsrdb.nrel.gov/data-sets/api-instructions.html. [Accessed April 2020].

[14] D. M. J. Garbade, "Towards Data Science," 13 Sept 2018. [Online]. Available: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1. [Accessed April 2020].

[15] "k-Means Advantages and Disadvantages," [Online]. Available: https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages . [Accessed April 2020].

[16] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[17] "tslearn.clustering.TimeSeriesKMeans," [Online]. Available: https://tslearn.readthedocs.io/en/latest/gen_modules/clustering/tslearn.clustering.TimeSeriesKMeans.html . [Accessed April 2020].

[18] S. Yildirim, "Decision Trees and Random Forests — Explained," 11 Feb 2020. [Online]. Available: https://towardsdatascience.com/decision-tree-and-random-forest-explained-8d20ddabc9dd. [Accessed April 2020].

[19] W. H. T. James, "Elements of Statistical Learning," 2017, p. 318.

[20] S. Bhattacharyya, "Support Vector Machine: Kernel Trick; Mercer's Theorem," 19 Dec 2018. [Online]. Available: https://towardsdatascience.com/understanding-support-vector-machine-part-2-kernel-trick-mercers-theorem-e1e6848c6c4d. [Accessed April 2020].

[21] "Forecasting 101: A Guide to Forecast Error Measurement Statistics and How to Use Them," [Online]. Available: https://www.forecastpro.com/Trends/forecasting101August2011.html. [Accessed March 2020].

[22] NREL, "Ten Years of Analyzing the Duck Chart," 26 Feb 2018. [Online]. Available: https://www.nrel.gov/news/program/2018/10-years-duck-curve.html. [Accessed April 2020].

[23] "U.S. economy and electricity demand growth are linked, but relationship is changing," 22 March 2013. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=10491. [Accessed April 2020].

[24] California Energy Comission, "California Solar Energy Statistics and Data," [Online]. Available: https://ww2.energy.ca.gov/almanac/renewables_data/solar/index_cms.php. [Accessed April 2020].

[25] US Energy Information Administration, "Southwestern states have better solar resources and higher solar PV capacity factors," 12 June 2019. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=39832. [Accessed April 2020].

[26] US Energy Information Administration, "Solar photovoltaic capacity factors in the United States between 2014 and 2017, by select state," Statistica, [Online]. Available: https://www.statista.com/statistics/1019796/solar-pv-capacity-factors-us-by-state/. [Accessed April 2020].

[27] California Energy Commission, "Improving Solar and Load Forecasts by Reducing Operational Uncertainty," San Diego, CA, 2019.

[28] "An overview of the California solar mandate," Energy Sage, [Online]. Available: https://news.energysage.com/an-overview-of-the-california-solar-mandate/. [Accessed April 2020].

[29] "2019 California Wildfires," Center for Disaster Philanthropy, 10 Oct 2019. [Online]. Available: https://disasterphilanthropy.org/disaster/2019-california-wildfires/. [Accessed May 2020].

[30] T. S. H. W. John Zacharias, "Spatial Behavior in San Francisco's Plazas: The Effects of Microclimate, Other People, and Environmental Design," *Environment and Behavior*, vol. 36, no. 5, 2004.

[31] C. S. M. A. Meredith Martin, "Survival is not enough: The effects of microclimate on the growth and health of three common urban tree species in San Francisco, California," *Urban Forestry & Urban Greening*, vol. 19, no. 1 September 2016,, pp. 1-6.

[32] H. T. P. C. F. C. Amanpreet Kaur, "Ensemble re-forecasting methods for enhanced power load prediction," 2014.