Forecasting Residential Energy Consumption Using Support Vector Regressions

Xiaoou Monica Zhang, Katarina Grolinger, Miriam A. M. Capretz
Electricity and Computer Engineering, Western University
London, Ontario, Canada, N6A 5B9
Email: {xzha28, kgroling, mcapretz}@uwo.ca

***Abstract*—With the development of smart electricity metering technologies, huge amount of consumption data can be retrieved on daily and hourly basis. Energy consumption forecasting facilitates electricity demand management and utilities load planning. Most of the prior researches are focused on commercial customers or residential building-level energy consumption, or use behavioral and occupancy sensor data to experiment on individual household's electrical consumption. This paper investigates fifteen anonymous individual household's electricity consumption forecasting using support vector regression(SVR) modelling approach, applied to both daily and hourly data granularity. The electricity usage dataset was collected from fifteen households by London Hydro, a local utility company, from 2014 to 2016. Exploratory data analysis (EDA) is adopted for data visualization and feature selection. Our analysis demonstrates that forecasting residential energy consumption by weather, calendar and Time-of-usage price is feasible and reliable with sufficient accuracy for some individual residential uses in either daily or hourly prediction.**

I. Introduction

Canadian households consume 1.4 million Tera-joules of energy in their homes in 2013, up 7.2% from 2011. Households use electricity for cooling, heating and power appliances and electronics. Electricity accounts for 44.6% of the total energy consumed by those Canadian residential customers [1]. The accelerated development of smart metering technologies and the Green Button initiative[2] enables measuring, collecting and presenting electricity consumption information for residential customers. Applications of Home Area Network and Home Energy Management System (HEMS) and Demand Response (DR) have brought a new focus on individual household. The premise and potential of residential energy consumption prediction have gradually been recognized by governments and research institutes. Smart meters, as part of Advance Metering Infrastructure (AMI), are being installed in households at increasing rate. modelling and forecasting household electricity consumption facilitates utility industry to optimize city energy load planning and provide personalized residential services [3].

Electricity usage at the individual household level shows high variance, since it relies on users' lifestyle, occupancy behavior, building charcteristics, weather and calendar information [4**?** , 5]. Feeding machine learning models with the data from energy smart meters and other relevant factors to infer the energy consumption for the next days and hours is known as sensor-based forecasting approach [6]. Prior research [7–11] has established the prediction accuracy of sensor-based approaches on the forecasting of commercial or residential in building level. In addition, these studies have investigated which machine learning techniques perform well at modelling commercial consumption. However, unlike the regularity in workplace with aggregated electric consumptions on routine schedules, more irregularity is foreseen in residential electrical consumption. Most households exhibit low base loads (0-500W) [12] and load profiles for appliances, which have relatively high power consumption, such as air conditioning, clothes watcher and drying, pool pumps, electric heaters, etc., are highly dynamic.

Some targeted residential energy measures are taken on residential buildings [6, 8, 13] with specific datasets on multi-family residential building level of aggregation or sensor data about user behaviors and occupancy. However, in the real application, utility companies' knowledge about the customers is generally limited to the billing address, smart meter ID and contract account's basic information [8].

Conducting detailed surveys to acquire detailed customer profile might be typically expensive impractical, time-consuming and facing low customer participation [14]. In this case, forecasting single-family's electrical consumption using the raw data collected from commonly deployed smart meters in household's home with weather variables[15, 16] and calendar information[13, 17, 18]) is a cost-effective approach to gain insights into residential customers and optimize energy efficiency programs.

Traditional utility prices involve a set rate per kilowatt-hour, fluctuating during the Summer and Winder. "Time-of-usage (TOU)" rate plan is a sliding rate scale structured according to on-peak, mid-peak and off-peak times of day [19]. In Ontario, TOU rates are defined by Ontario Energy board [20] as shown Table I, and are mandatory for residential customers across the province. Our study also examines the impact of TOU pricing scheme on residential electrical usage.

In this work, we conduct experiments on a dataset of fifteen randomly-selected residential users from a local utility company. It is three-year's (2014-2016) hourly electricity consumption data from anonymous households, with unknown dwelling properties, occupation, nor household's socio-economic status. Our study reveals the competency of support vector regression model, which is the most popular machine learning approach for business energy prediction, in the ap-

X. M. Zhang, K. Grolinger, M. A. M. Capretz, Forecasting Residential Energy Consumption: Single Household Perspective, Proc. of the IEEE International Conference on Machine Learning and Applications, 2018.

TABLE I: Ontario Electricity Time-of-use Price Table [20]

From	To	Summer Rate (May-Oct. Weekdays)	Winter Rate (Nov.-Apr. Weekdays)
7:00AM	11:00AM	mid-peak rate 9.5 cents/kWh	on-peak rate 13.2 cents/kWh
11:00AM	5:00PM	on-peak rate 13.2 cents/kWh	mid-peak rate 9.5 cents/kWh
5:00PM	7:00PM	mid-peak rate 9.5 cents/kWh	on-peak rate 13.2 cents/kWh
7:00PM	7:00AM	off-peak rate 6.5 cents/kWh	off-peak rate 6.5 cents/kWh

plication field of residential sector. Due to complexity of individual behaviors, the prediction accuracy varies among those single-families.

The remainder sections of the paper is organized as follows: Section II gives an overview of related researches in the area of energy modelling and prediction; Section III analyzes the most popular algorithm for energy forecasting and how to evaluate the accuracy; Section IV describes the methodology of machine learning approach in detail; Section V presents the implementation and experimental results and the evaluation of the algorithm's correctness; Section VI presents our conclusions and future directions.

II. Related Work

The number of research studies in energy consumption, including annual consumption for various uses, characteristics impacting energy usage and consumption prediction, is dramatically increased with the development of smart meters and other data collection methods. Many studies have explored machine learning approaches for modelling electrical consumption, applied in both commercial and residential sectors.

To reduce the impact of individual user's stochasticity in energy consumption, the majority of energy modelling studies work on aggregated level in either residential buildings [8, 11, 13, 16, 17, 21–23] or commercial buidings [6, 7, 9, 10, 24–27]. For example, Jain and Smith presents energy modelling results on multi-family residential buildings [8] , with single-family's consumption aggregated into building level. They built energy forecasting model using support vector regression on different data granularity, and the results indicated the most effective models were built with hourly consumption at floor level with a CV value of 2.16 and a standard error of 28%. Y. Liu and W. Wang proposed a forecast engineer [11] based on sliding window empirical mode decomposition (SWEMD) and Elman Neural Network (IENN) to predict electricity load of building level consumers. The research of Katarina Grolinger et al. [7] demonstrates both Neural Network(NN) and Support Vector Machines(SVM) are accurate in the "consumption prediction for event-organizing venues" using energy consumption data and event-related attributes.

F. Zhang and C. Deb investigated an institutional building's energy consumption dataset using weighted Support Vector Regression model [24], which was used to forecast half-hourly and daily electrical consumption for the same building.

In Fazil Kaytez's paper [25], the result of LS-SVM model is a quick prediction and also more accurate than the result of both traditional regression analysis and Artificial Neural Networks (ANN).

In the area of individual residential energy forecasting, prior research studied on highly detailed dataset [13, 28] to predict household's electrical loads including demographic survey information, dwelling properties, appliance ownership, and occupancy detection, which are typical Bottom-up approaches [29]. The research focusing on residential buildings by Richard E. Edwards and Joshua New [13] studies a special research data set: three homes with 140 sensors collecting human behaviors on opening/closing refrigerators, using ovens as well as occupancy patterns. It presents results for commercial consumption prediction and residential consumption prediction that NN-based methods are the best methods for commercial consumption prediction, while Least Squares Support Vector Machines(LS-SVM) perform better on predicting residential consumption hourly. The prediction accuracy comes from decomposing of electrical usage behaviors by a large number of sensors in the experiments, and can hardly be replicated for large-scale and cost-sensitive field application. Li and Dong built occupancy prediction models [21] based on occupancy sensors experimented in four residential houses, to predict occupancy presence using Markov model.

Without deploying behavioral sensors, other works focused on behavior pattern clustering and stochastic simulation approach[14, 16, 22] to recognize and simulate occupant behavioral patterns with clustering algorithm.

Regarding the selection of features, there are four main categories of factors are defined by Bechel et al., dramatically affecting the electricity consumption: Weather and location, dwelling characteristics, appliance and electronics stock, and occupancy and behavior [14]. To be more practical and feasible, more recent researches have looked into ways to utilize timporal variables, such as calendar information (hour of a day, day of a week, week of a year, holidays, weekends, weekdays, season, etc.)[9, 30] and weather variables and forecasts (weather condition, temperature, humidity, irradiation and wind speed, etc.), which can be obtained from local or regional weather stations [7, 31].

Wide range of modelling techniques have been used to predict electricity load, including Neural Networks (ANN) [11, 13], support vector machine (SVM) [13, 21, 24], autoregressive integrated moving average (ARIMA) models [32], regressions models [32], clustering techniques [14] and empirical mode decomposition (EMD) [11]. Neural networks have been extensively used for industrial electricity forecasting, while Support vector regression (SVR) have been successfully used in solving nonlinear regression and time series problems [24], is considered as an emerging technique and performed best in some residential energy experiments [13].

In this study, Support Vector Regression(SVR) is used for the prediction of fifteen households' residential electricity consumption. These households are anonymous residents in London Ontario, with a home-installed smart meter for electricity measurement.

III. Background

This section introduces the scenarios of machine learning approach Support Vector Regression and performance measures.

A. Support Vector Regression

Support vector machines (SVM) [33] are supervised machine learning models used for classification and regression problems. SVMs essentially consist of kernel and optimizer algorithm. Kernel divides ono-linear data into high-dimensional space and makes data linearly separable. The learning takes place in the feature space, and the data points only appear inside dot products. The optimizer algorithm is applied to solve the optimization problem. Since SVM seeks to minimize an upper bound of the generalization error consisting of the sum of the training error and a confidence level, it shows superiority compared with commonly used empirical risk minimization (ERM) principle, which only minimize the training error. Thus, SVM usually achieves higher generalization performance than other machine learning techniques. Support vector regression (SVR) is a version of SVM, which is a non-linear regression model that looks at the extremes of data sets and draw a decision boundary (or a hyperplane) to solve function fitting problems. A non-linear regression with epsilon intensive band is presented in Fig. 1. Sometimes data sets are linearly non-separable and have to be mapped onto an N-dimensional space and an (N-1)-dimensional separating hyperplane need to be found. However, the process is computationally expensive. A suitable kernel trick could significantly reduce the computational cost.

The relationship between inputs $x_1, x_2, ..., x_n$ and output Y is determined as:

$$Y = W\varphi(x) + b \tag{1}$$

$\varphi(x)$ is a kernel function, and RBF, Polynomial, Linear kernels are used in our implementation.

$$k(x, y) = exp(-\frac{\|x-y\|^2}{2\sigma^2}) \tag{2}$$

The probability of making mistakes by the trained model using training data is minimized by minimizing the following convex criterion function:

$$\frac{1}{2}\|w\|^2 + C\frac{1}{N}\sum_{i=1}^{N}\xi_i + \xi_i^* \tag{3}$$

with the following constraints:

$$y_i - w^T\varphi(x_i) - b \le \epsilon + \xi_i$$
$$w^T\varphi(x_i) + b - y_i \le \epsilon + \xi_i^*$$

In the above equations, w is a weight vector and C is the cost of making an error. ξ_i and ξ_i^* are the residuals beyond the ϵ boundary, which are shown in Fig. 1.

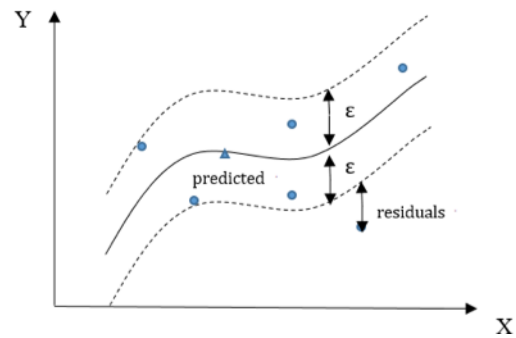


Fig. 1: Non-linear Regression with epsilon intensive band.

B. Performance Measures

To determine the performance of the prediction model, two metrics are adopted for the validation: the mean absolute percentage of error (MAPE).

The MAPE metric is a widely-adopted measure of prediction accuracy in electricity predictions studies [7, 24, 28]. It generally represents accuracy as a percentage, defined by the formula:

$$MAPE(\%) = \frac{100}{n}\sum_{j=1}^{n}\left|\frac{y_j - \hat{y_j}}{y_j}\right| \tag{4}$$

where y_j denotes the actual electricity consumption of household j, $\hat{y_j}$ denotes the predicted consumption, and n is the number of observations.

IV. Methodology

This work uses machine learning approach: support vector regression modelling for electricity forecasting, both hourly and daily granularities are investigated with evaluated accuracy. This section introduces the data set and how the prediction model is built with several variants and data exploratory analysis. Fig 2 illustrates the framework of individual household electrical forecasting.

A. Data preprocessing

In this paper, we study on the energy data from London Hydro, the electricity service provider in London, Ontario, Canada. The dataset incorporates hourly electricity consumption of fifteen households from 2014 to 2016. The raw dataset from smart meter reading contains electricity consumption data measured in kilowatts-hour (kWh) for households on a time scale of one hour. We obtain the weather and humidity historical hourly data from Canadian Government Official Website as impacted factors to residential energy consumption.

1) Data Cleansing: The original electricity consumption data contains 534,966 lines of hourly meter reading value for fifteen households in 3 years, however, there exists some missing or disordered hours. To avoid undesirable impacts to the prediction model, we replace those invalid or missing consumption values by the average electrical consumption value of the previous and the following hours. There are also some invalid weather conditions or temperature/humidity data

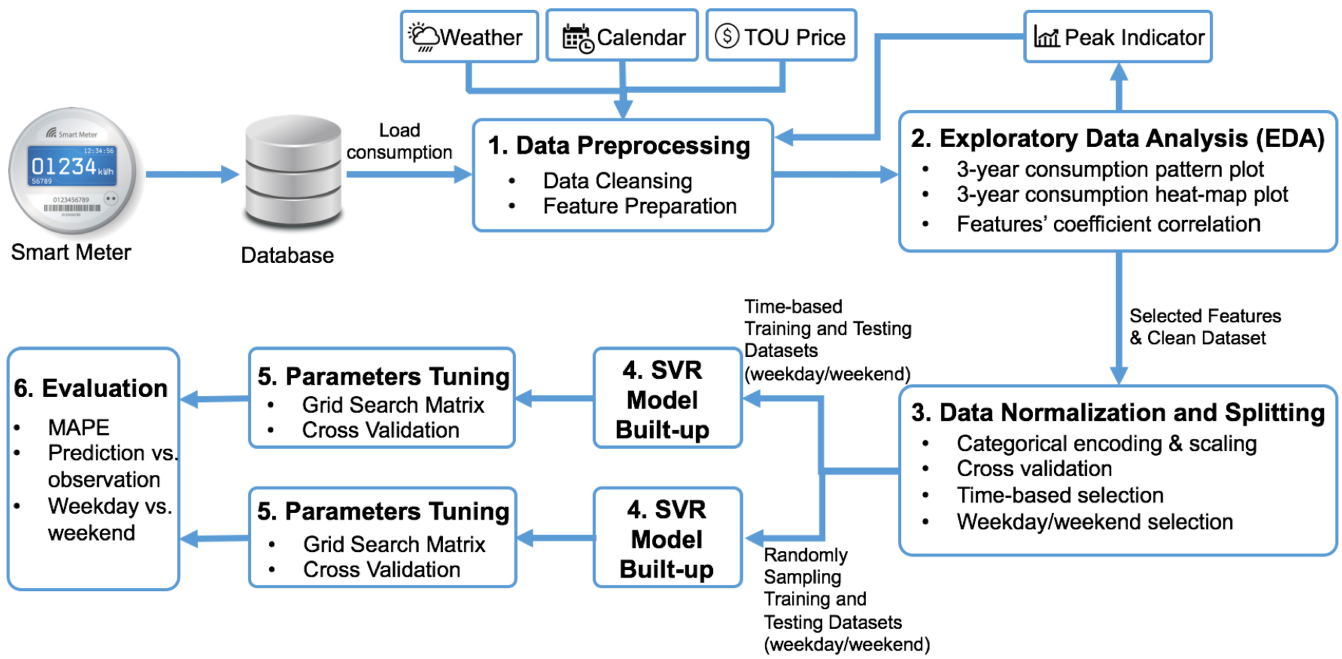


Fig. 2: Framework of individual household electricity forecasting

in the dataset downloaded from the Governmental Website. To tackle the missing hourly temperature and humidity, an average temperature/humidity value of the previous and following hours is calculated as a substitute for the missing hour; for the invalid weather condition (Snow, Cloudy, Clear, etc.), we refill the cell by the last hour's weather condition.

2) Feature preparation: The cleaned and consistent dataset is reorganized and additional new features are generated as follows:

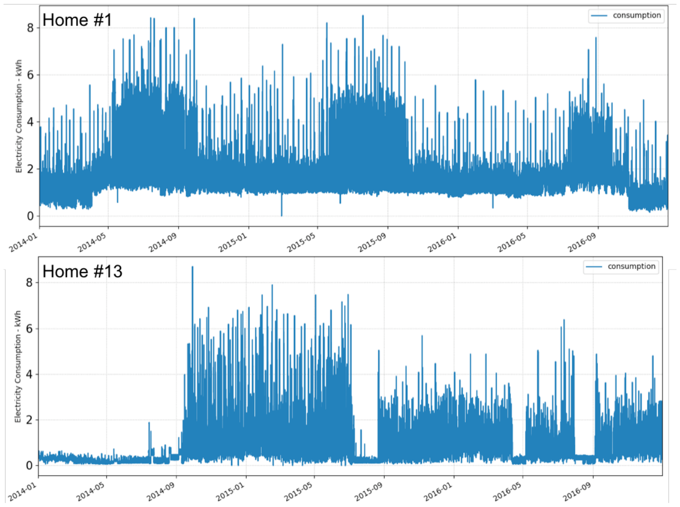
- Weather Condition: We aggregate over 20 weather conditions in the original version to seven main weather conditions: "Cloudy", "Clear", "Fog", "Haze", "Rain","Snow" and "Ice".
- Temperature: Temperature is a factor used to predict hydro usage because Air Condition (AC) consumption takes a large proportion in all electricity consumption, especially in summer, as well as other residential electrical applications.
- Humidity. We retrieve the relative humidity data from government of Canada website, which represents the measurement of water vapor relative to the temperature of the air as a percentage.
- Hour of the Day: 1 to 24. Hour of the day is an important attribute to forecast energy consumption. The consumption in midnight is highly possible less than the consumption at 7:00 pm, which could also be confirmed by Integrated Energy Mapping Strategy [34].
- Day of the Week: 1 to 8 for weekdays, weekends and long weekends. Monday through Sunday is represented by number 1 to 7, and 8 represents long weekends in Ontario. We use this factor to detect the usage pattern in a national holiday.
- Week of the Year: 1 to 52 to represent the week of the year.
- Month: 1 to 12. We implement month as a factor to complement temperature and reduce the possibility of making prediction errors.
- Season: Spring, Summer, Fall and Winter. We use season as an attribute to detect the relationship between season and hydro usage and to improve the forecasting accuracy.
- Price: The electricity time-of-usage (TOU) rate [20] is listed in Table I and it could influence the electrical usage for the households.

B. Exploratory data analysis

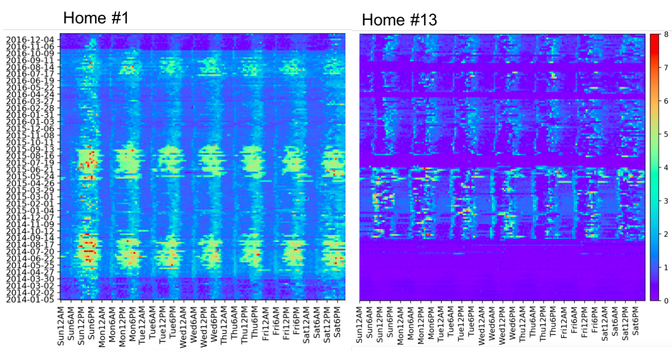
Comparing with commercial electricity consumption, the most significant characteristic of residential electricity usage is that some families live regular lives, which can be reflected from the similarities in consumption patterns. However, some families are irregular in energy usage. It leads to various consumption patterns among different families.

Investigating the correlation between the input features and the forecasted loads [16] facilitate choosing the features which exhibit significant correlation to energy consumption for a particular household, thus the computational time for parameters tuning is reduced with a subset of significantly correlated variables.

1) Consumption Patterns: The stochasticity in family member's behaviors results in various electricity consumption patterns. Fig. 3(a) shows two households' electricity usage trends and pattern respectively, during Jan. 1 2014 to December 31, 2016. In the figure, X-axis denotes time, and Y-axis stands for electricity consumption.



(a) 3-year hourly consumption pattern



(b) 3-year hourly consumption heat-map

Fig. 3: Two households three years hourly consumption

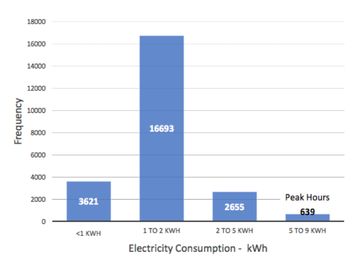
There is a visually detectable pattern for home#1, whose electricity usage ramps up during summer time and relatively lower in winter. It is typical residential behaviors in Canada due to the air conditioning usage in high-temperature days, which boosts the electricity usage during May to October. For Winter heating system, gas is more widely used for residential homes other than electricity in most Ontario houses and apartments, for the sake of lower cost. Therefore, there is less electricity usage from October to May of next year. However, the electricity usage pattern is not easily to be recognized for home#2. Less regularity in electricity consumption pattern can be deducted from the home#2 hourly trend.

For a better data visualization, detailed 2-D consumption heat-maps are plotted for those two households as shown in Fig. 3(b). In the figure, X-axis denotes the repetitive presented hours in a week from Sunday to Saturday with a step of 6 hours, and Y-axis denotes the days from 2014 to 2016. Each hourly consumption is represented by a color square according to a predefined "rainbow" color bar with colors from warm to cold. The more consumption for a particular hour has, the warmer the color will be. For home#1, the periodical appearance of high usage during daytime in Summers months is evident, especially in the year 2014 and 2015. By contrast, less regularity can be identified from the consumption data visualization for home#2.

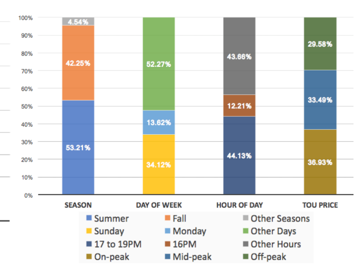
Residential characteristics including household's socio-economic status, its dwelling properties, employment status and even number of persons living in the household is relevant to the residential electricity usage hour-by-hour [14]. For example, rental residence with changing tenants and undetermined vacancy can hardly be predicted in the electricity consumption.

2) Peak Consumption Indicators: A residential customer using multiple appliances simultaneously in a certain hour produces in an accumulated consumption peak. Those peaks with three to five times of average hourly consumption contribute significantly to the MAPE errors.

Take home#1 for example, for the period of three years (2014-2016, totally 26304 hours), 77.2% of hours consume less than two kWh electricity (Fig 4(a)). The average hourly consumption of 3 years is 1.68 kWh, and median hourly consumption is 1.35 kWh. In this case, those hours with electricity usage greater than five kWh is defined as peak consumption hours. Fig. 4(b) illustrates the relevant factors among those 639 peak usage hours. For seasonal impact, 95.5% of peak hours appear in both Summer and Fall. Nearly one-third of actual peak hours are in Sunday and followed by Monday (13.62%). Regarding "hour of the day", the hours of 17-19PM contribute the most electricity usage in a day. "TOU Price" represents the Time of Usage pricing scheme defined by Ontario Energy Board as described in Table I. The "on-peak", "mid-peak" and "off-peak" priced hours have similar proportions in peak electricity consumption.



(a) Hourly consumption distribution in three years



(b) Peak consumption hours distribution in three years

Fig. 4: A household consumption and peak hours distribution

To capture peak consumption hours for residential users, a new feature named "peak index" is added to fit the prediction model for this household by calculating the peak factors (peak seasons, top days in a week, and top hours in a day) for each hour and align a corresponding weight to those hours.

C. Prediction Model

The prediction model is designed to work with SVR on both hourly and daily data granularities for every household. For each data granularity, one observation is associated with one energy reading. Other features discussed before are added to the energy reading data set. To explore the accuracy on daily granularity, the Green Button hourly data are aggregated as follows to obtain the daily electricity consumption:

$$C_d = \sum_{i=1}^{24} C_{h_i} \tag{5}$$

where C_{h_i} is the electrical consumption for the i^{th} hour in a day. To train the prediction model, splitting time series into chronological sets and checking for parameters stabilities over time is proved to be useful for business buildings energy consumption prediction [7, 13]. However, irregularity and uncertainty over time for some residential customers makes consecutive training set less accurate, compared with random sampling. Hence, for every household both consecutive time splitting and random sampling are applied for the residential electricity consumption prediction, and the method with less errors will be selected.

For the machine learning regression model SVR, parameters in the prediction model have to be resolved in the learning phase. With a nonlinear kernel function, two basics parameters should be determined in advance: the cost C, which denotes the penalty for errors greater than ϵ as shown in Fig. 1 and the nonlinear kernel coefficient γ. Combinations of model parameters constitute a model configuration, among which the best parameters need to be chosen for best prediction accuracy. Grid search method is adopted for tuning the model configuration. In the implementation, a set of C and γ parameters are formed by assembling a grid of search parameters. Once the optimized parameters are resolved, the prediction error could be validated by model assessment. The SVR prediction module was implemented in Python language.

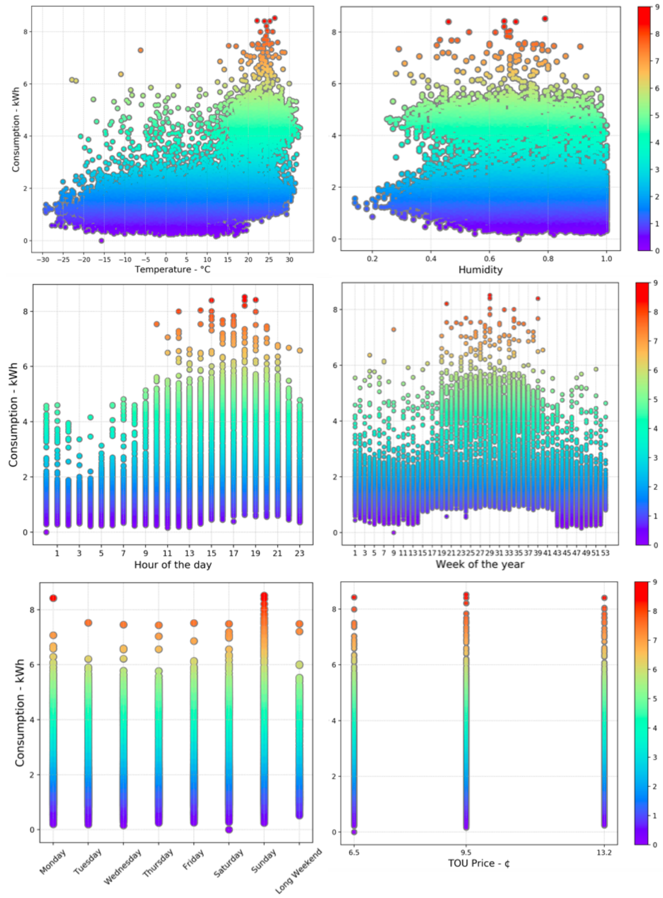
Since each household requires SVR parameters tuning for both chronological splitting and random sampling in hourly and daily data granularities to get the best modelling configuration, there are at least $15 \times 2 \times 2 = 60$ times of grid search tuning needed for the 15 households. A small cluster of servers is used, and each server has 24 Intel Xeon CPUs, 96GB memory and external storage to enable fast data reading during execution. To fasten the SVR parameters' cross-validation, the parallelized computing was implemented in Python to maximize the concurrently running of vacant CPUs.

V. Evaluation

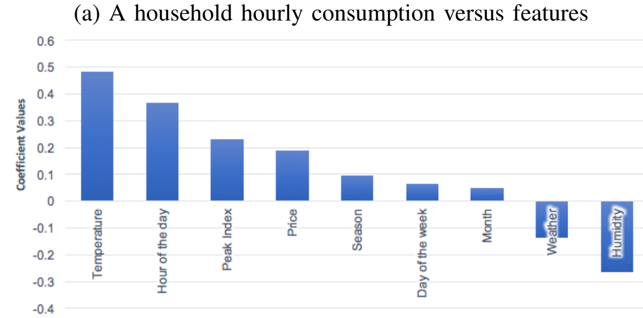
The proposed method has been experimented using residential electricity consumption data provided by London Hydro in over three years. Data exploratory analysis as described in section IV is conducted for all 15 households to visualize whether households has settled habits and routine in daily life or not. Then the SVR model is applied to each household respectively on different data granularities.

A. Feature Coefficients

More exploratory analysis is made for each household to analyze the relationship between those important factors and the household?s electrical consumption.



(a) A household hourly consumption versus features



(b) Correlation coefficients of features for the household

Fig. 5: Relationship between features and consumption

Fig. 5(a) visualizes the correlations between a household's three-year hourly consumption and features like "temperature", "humidity", "hour of the day", "week of the year", "day of the week" and "TOU price". From the figure, the "temperature" has an obvious positive correlation with energy usage: the hourly consumption increases along with the rising temperature. By contrast, "humidity" doesn?t have positive correlation with consumption. For "hour of the day", the hours of 2AM to 6AM account for the lower electricity usage, and hours in the afternoon always consume more. Averagely, week 19 to week 39 in a year have more electrical hourly consumption, which is correspondingly in the season of

Summer and Fall. During a week, Sunday has the most energy usage compared with other days. For the features prepared for the prediction, the correlation coefficients with the hourly consumption of the same household are calculated, as shown in Fig. 5(b). “temperature”, “hour of the day” and “peak index” are the top three important factors with the top three high coefficients. Those variables with negative coefficient reveal the inverse relationships with hourly consumption for the households. It is noted that household’s coefficients between features and consumption differ from each other. For example, a certain proportion of Ontario residents use electricity for heating during Winter other than Gas, which results in a remarkable electrical consumption at frozen days when the temperature is low. Thus, non-positive correlation between temperature and consumption can be found for those families.

B. Experiments and results

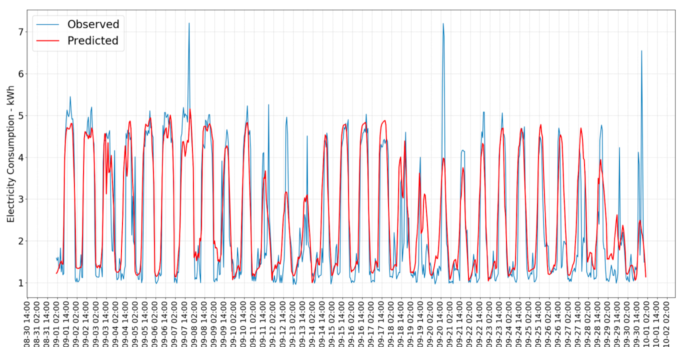


Fig. 6: A household hourly prediction vs actual observation

Fig. 6 illustrates the hourly actual electricity usage and the predicted energy consumption values obtained by SVR model for one month (e.g. September 2015) for home#1 mentioned in section IV. It is noticed that the prediction curve follows the fluctuation of the actual electricity usage during the day and night on an hourly basis, except for several peak usage hours. This occurs due to the fact that there are random variations for an individual home user.

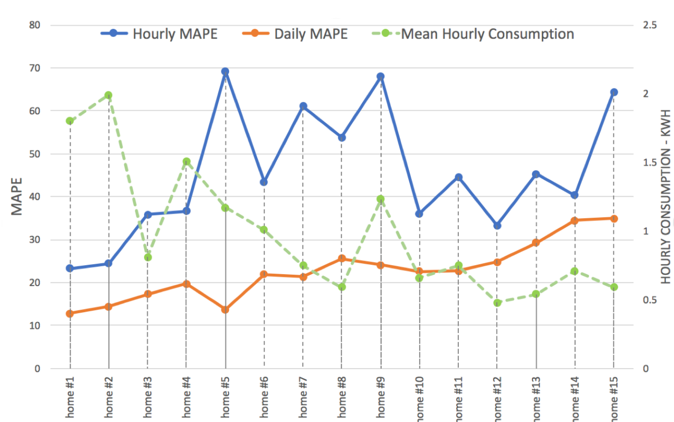


Fig. 7: Fifteen households’ consumption MAPE and mean hourly consumption

TABLE II: Fifteen households’ SVR modelling evaluation results and categories

Home No.	MAPE Hourly	MAPE Daily	Accuracy Category	Data Splitting Method
#1	23.31	12.78	Good hourly and daily accuracy	Time-based splitting
#2	24.42	14.46		
#3	35.82	17.30	Weak hourly, better daily accuracy	
#4	36.65	19.66		
#5	69.17	13.72	Poor hourly, better daily accuracy	
#6	43.41	21.89		
#7	61.07	21.31		
#8	53.81	25.61		
#9	67.96	24.14		
#10	36.05	22.55	Weak hourly, better daily accuracy	Randomly sampling
#11	44.63	22.66		
#12	33.33	24.8		
#13	45.33	29.31		
#14	40.34	34.49	Poor hourly and daily accuracy	
#15	64.38	34.95		

The evaluation metrics for all 15 households is shown in Table II. Although these 15 households’ consumption dataset is a random sample from London’s residential customers energy data, it represents several categories classified by families’ regularities in electrical consumption, as indicates by the column of “Accuracy categories”. Most of the households have acceptable daily consumption MAPE (under 30), while the accuracy for hourly consumption fluctuates dramatically. The first two households (home#1 and #2) having both hourly and daily MAPE under 30 show regularities in the activities of electrical consumption over time. It is noticed that these two families also have the highest hourly average consumptions (Fig. 7) during three years. The families with ID from #3 to #9 have weaker hourly prediction accuracy, but present regularity if hourly usage is aggregated into one day’s consumption. Those households have similarities in their consumption patterns over time, and gain better accuracy in terms of time-based training and testing subsets splitting. By contrast, home #10 to #15 demonstrate less regularity on continuous time period, and can be better forecasted on randomly sampling splitting.

TABLE III: Home #1 consumption MAPE comparison between all days and weekdays

Home #1	All Days: Time-based Splitting	All Days: Randomly Sampling	Weekdays: Time-based Splitting	Weekdays: Randomly Sampling
MAPE	23.31	24.91	22.01	23.94

As analyzed before, more variability in electricity consumption has been seen in weekends other than weekdays. Table III shows and example of MAPE comparison between the whole dataset and the subset of weekdays for the house-

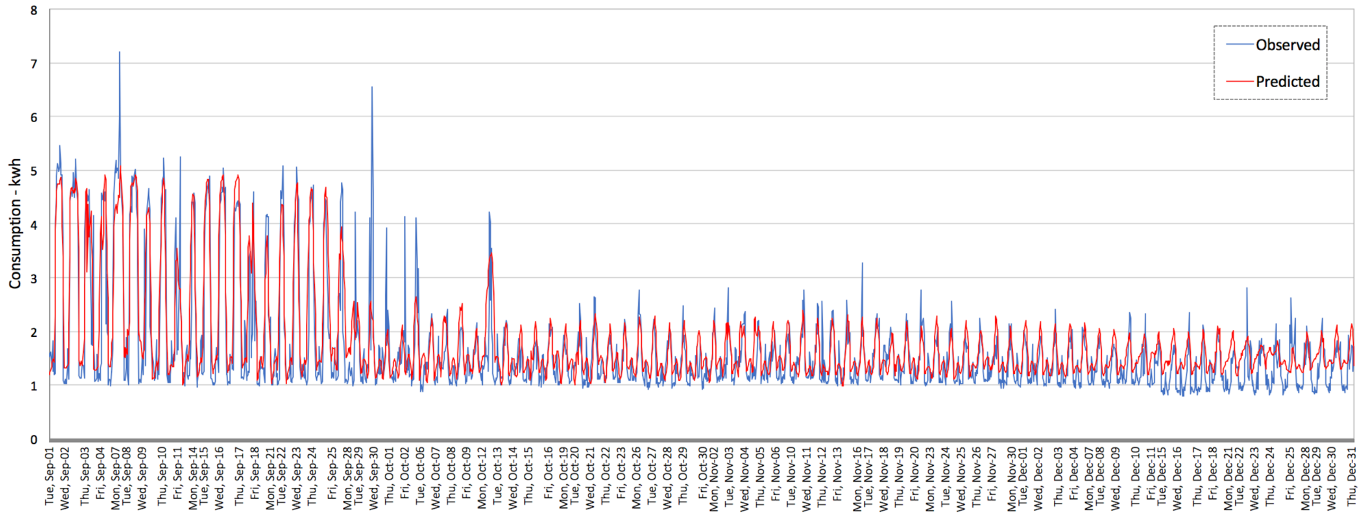


Fig. 8: A household weekday hourly prediction over four months

hold #1. The forecasting for weekdays enjoys lower MAPE errors for both time-based splitting and randomly sampling in processing the training and testing datasets. Fig. 8 visualizes the weekday electricity consumption prediction for the home #1 in 4 months. Seasonal change from September to October has brought substantial energy decrease, which led to non-perfectly matched prediction curve, however, for the majority of hours, fewer errors can be seen for the weekdays usage forecasting from October to December 2015.

VI. Conclusion

Smart metering technologies boost the data analytics in energy management, and create possibilities for new energy services. This study explores the accuracy of machine learning SVR modelling approach applying on residential customers. Unlike the large number of studies considering commercial building or multi-family residential buildings, this work has studies single-families with electrical usage collected by a smart meter installed by local utility company.

The development of smart meters and data collection technologies provides the possibility to collect detailed electricity consumption data. We used hourly electricity consumption data, detailed weather data in this study. This study confirmed the influence of elements, including weather condition, temperature, humidity holiday, hour of the day, day of the week, month, season and price of electricity, on residential electricity consumption and provided a quite efficient and accurate approach to make the prediction. We also revealed the correlation coefficients of those features mentioned above that hour of the day and temperature have the most significant impact on electricity consumption in our case. Different methods are used to split the dataset into training and testing subset. For the households with similarity in electrical consumption over time, consecutive time-based splitting works better than randomly sampling the data, however, for those who lacks regularity in the hourly electricity usage, sampling the whole dataset irrelevant to time and using 20% of the sampling as testing sub-dataset outperforms time-based approach.

Because of the stochasticity for single residential customers, daily data granularity achieves better prediction results than hourly data for all the 15 households. Aggregating hourly consumption to daily is an effective way to mitigate the impact of randomness in hourly behaviors of family members. The lowest MAPE error for one of the fifteen household is 12.78 for daily prediction and 23.31 for hourly prediction, and it reduces to 22.01 (hourly) if only weekdays are counted for the same family.

The model would reveal a better prediction result if occupancy of the house can be detected and added as a feature. Future work will explore the possibilities of revealing more user social and behavioral characteristics from the data, and use those data as an input to the consumption prediction model.

Our study uses one Machine Learning approach, Support Vector Regression (SVR), because of the restriction of time and computational hardware. Multiple models containing different attributes should be tested in the future.

Acknowledgment

The authors would like to thank London Hydro for supplying valuable anonymous single household data, Green button platform and industry knowledge used in this study.

References

[1] S. Canada, "Households and the environment survey: Energy use, 2013," http://www.statcan.gc.ca/daily-quotidien/160318/dq160318d-eng.htm, Mar 2016.

[2] U. D. of Energy, "Green button," https://energy.gov/data/green-button.

[3] Q. Sun, H. Li, Z. Ma, C. Wang, J. Campillo, Q. Zhang, F. Wallin, and J. Guo, "A comprehensive review of smart energy meters in intelligent energy networks," *IEEE*

Internet of Things Journal, vol. 3, no. 4, pp. 464–479, 2016.

[4] A. Tascikaraoglu, A. Boynuegri, and M. Uzunoglu, "A demand side management strategy based on forecasting of residential renewable sources: A smart home system in turkey," *Energy and Buildings*, vol. 80, pp. 309–320, 2014.

[5] P. Lusis, K. R. Khalilpour, L. Andrew, and A. Liebman, "Short-term residential load forecasting: Impact of calendar effects and forecast granularity," *Applied Energy*, vol. 205, pp. 654–669, 2017.

[6] H.-x. Zhao and F. Magoulès, "A review on the prediction of building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586–3592, 2012.

[7] K. Grolinger, A. L'Heureux, M. A. Capretz, and L. Seewald, "Energy forecasting for event venues: Big data and prediction accuracy," *Energy and Buildings*, vol. 112, pp. 222–233, 2016.

[8] R. K. Jain, K. M. Smith, P. J. Culligan, and J. E. Taylor, "Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy," *Applied Energy*, vol. 123, pp. 168–178, 2014.

[9] S. S. Abdelkader, K. Grolinger, and M. A. Capretz, "Predicting energy demand peak using m5 model trees," in *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on*. IEEE, 2015, pp. 509–514.

[10] D. B. Araya, K. Grolinger, H. F. ElYamany, M. A. Capretz, and G. Bitsuamlak, "An ensemble learning framework for anomaly detection in building energy consumption," *Energy and Buildings*, vol. 144, pp. 191–206, 2017.

[11] Y. Liu, W. Wang, and N. Ghadimi, "Electricity load forecasting by an improved forecast engine for building level consumers," *Energy*, vol. 139, pp. 18–30, 2017.

[12] B. Yildiz, J. Bilbao, J. Dore, and A. Sproul, "Recent advances in the analysis of residential electricity consumption and applications of smart meter data," *Applied Energy*, 2017.

[13] R. E. Edwards, J. New, and L. E. Parker, "Predicting future hourly residential electrical consumption: A machine learning case study," *Energy and Buildings*, vol. 49, pp. 591–603, 2012.

[14] C. Beckel, L. Sadamori, T. Staake, and S. Santini, "Revealing household characteristics from smart meter data," *Energy*, vol. 78, pp. 397–410, 2014.

[15] Y.-H. Hsiao, "Household electricity demand forecast based on context information and user daily schedule analysis from meter data," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 1, pp. 33–43, 2015.

[16] T. K. Wijaya, M. Vasirani, S. Humeau, and K. Aberer, "Cluster-based aggregate forecasting for residential electricity demand using smart meter data," in *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE, 2015, pp. 879–887.

[17] S. Humeau, T. K. Wijaya, M. Vasirani, and K. Aberer, "Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households," in *Sustainable Internet and ICT for Sustainability (SustainIT), 2013*. IEEE, 2013, pp. 1–6.

[18] M. Rossi and D. Brunelli, "Electricity demand forecasting of single residential units," in *Environmental Energy and Structural Monitoring Systems (EESMS), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–6.

[19] M. Hinterstocker, P. Schott, and S. von Roon, "Evaluation of the effects of time-of-use pricing for private households based on measured load data," in *European Energy Market (EEM), 2017 14th International Conference on the*. IEEE, 2017, pp. 1–6.

[20] O. E. Board, "Electricity rates," https://www.oeb.ca/rates-and-your-bill/electricity-rates.

[21] Z. Li and B. Dong, "A new modeling approach for short-term prediction of occupancy in residential buildings," *Building and Environment*, vol. 121, pp. 277–290, 2017.

[22] L. Diao, Y. Sun, Z. Chen, and J. Chen, "Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation," *Energy and Buildings*, vol. 147, pp. 47–66, 2017.

[23] M. R. Biswas, M. D. Robinson, and N. Fumo, "Prediction of residential building energy consumption: A neural network approach," *Energy*, vol. 117, pp. 84–92, 2016.

[24] F. Zhang, C. Deb, S. E. Lee, J. Yang, and K. W. Shah, "Time series forecasting for building energy consumption using weighted support vector regression with differential evolution optimization technique," *Energy and Buildings*, vol. 126, pp. 94–103, 2016.

[25] F. Kaytez, M. C. Taplamacioglu, E. Cam, and F. Hardalac, "Forecasting electricity consumption: a comparison of regression analysis, neural networks and least squares support vector machines," *International Journal of Electrical Power & Energy Systems*, vol. 67, pp. 431–438, 2015.

[26] B. Yildiz, J. Bilbao, and A. Sproul, "A review and analysis of regression and machine learning models on commercial building electricity load forecasting," *Renewable and Sustainable Energy Reviews*, vol. 73, pp. 1104–1122, 2017.

[27] C. Robinson, B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M. A. Brown, and R. M. Pendyala, "Machine learning approaches for estimating commercial building energy consumption," *Applied Energy*, vol. 208, pp. 889–904, 2017.

[28] B. Dong, Z. Li, S. M. Rahman, and R. Vega, "A hybrid model approach for forecasting future residential electricity consumption," *Energy and Buildings*, vol. 117, pp. 341–351, 2016.

[29] A. Grandjean, J. Adnot, and G. Binet, "A review and an analysis of the residential electric load curve models,"

Renewable and Sustainable Energy Reviews, vol. 16, no. 9, pp. 6539–6565, 2012.

[30] F. Rodrigues, C. Cardeira, and J. M. F. Calado, "The daily and hourly energy consumption and load forecasting using artificial neural network method: a case study using a set of 93 households in portugal," *Energy Procedia*, vol. 62, pp. 220–229, 2014.

[31] R. Miller, L. Golab, and C. Rosenberg, "Modelling weather effects for impact analysis of residential time-of-use electricity pricing," *Energy Policy*, vol. 105, pp. 534–546, 2017.

[32] K. Gajowniczek and T. Zabkowski, "Electricity forecasting on the individual household level enhanced based on activity patterns," *PloS one*, vol. 12, no. 4, p. e0174098, 2017.

[33] V. N. Vapnik and S. Kotz, *Estimation of dependences based on empirical data*. Springer-Verlag New York, 1982, vol. 40.

[34] B. Gilmour, "City of london: Integrated energy mapping strategy (l-iems)," Canadian Urban Institute, Tech. Rep., 2011.

[35] J. Hosking, R. Natarajan, S. Ghosh, S. Subramanian, and X. Zhang, "Short-term forecasting of the daily load curve for residential electricity usage in the smart grid," *Applied Stochastic Models in Business and Industry*, vol. 29, no. 6, pp. 604–620, 2013.

[36] A. Tascikaraoglu and B. M. Sanandaji, "Short-term residential electric load forecasting: A compressive spatio-temporal approach," *Energy and Buildings*, vol. 111, pp. 380–392, 2016.

[37] S. Rathi, K. Rawal, Y. Pawar, D. Zalte, S. Shivarkar, and B. Students, "Short term power demand forecasting," *International Journal of Engineering Science*, vol. 3423, 2016.