Proceedings of the ASME 2018 Power and Energy Conferece
PowerEnergy2018
June 24-28, 2018, Lake Buena Vista, FL, USA

PowerEnergy2018-7461

ACCURATE AND DATA-LIMITED PREDICTION FOR SMART HOME ENERGY MANAGEMENT

Baris Aksanli
Electrical and Computer Engineering Department
San Diego State University
San Diego, California 92182
Email: baksanli@sdsu.edu

ABSTRACT

Residential energy applications have become an important domain of cyber-physical systems. These applications provide significant opportunities for end-users to reduce their electricity costs and for the utilities to balance their supply and demand in the most effective way. One of the most important applications is predicting the total energy usage of a house. However, an accurate time-series prediction may require significant amount of data, e.g. per appliance energy consumption values, that need costly installations, data storage units, and computation and communication devices. In this paper, we propose a framework that uses a forward-selection-based input filtering mechanism for residential prediction applications. Our framework can effectively reduce the amount of data required for residential energy prediction without sacrificing prediction performance. We demonstrate that 94% of the houses can leverage our method, which leads to up to 80% reduction in required data, greatly reducing the system cost and overhead.

INTRODUCTION

Residential smart grid applications have recently gained a lot of attention due to advancements in cyber-physical systems (CPS) (sensor technology, smart meters, smart appliances, etc.) and the Internet of Things (IoT) (advanced networking capabilities). One important target of these applications is to achieve efficient energy consumption. Efficient energy consumption has recently become one of the major global concerns, due to the growing energy demands and lack of natural resources to meet them [1] [2]. To support this vision in residential power grids, utilities are replacing the electromechanical meters with smart meters [2]. "In 2014, U.S. electric utilities had about 58.5 million advanced (smart) metering infrastructure (AMI) installations. About 88% were residential customer installations" [3].

Residential energy management applications target to control the individual loads in a house (such as washer, dishwasher, dryer, heating, ventilation, lighting and air conditioning (HVAC) units, etc.) to reduce the energy consumption or the electricity cost. Different than other domains (such as commercial or industrial buildings) controlling residential loads require explicit human participation. Residential applications rely on various sensors and smart meter data to make sure that the control decisions are aligned with the requirements of household members. Although the IoT provides significant leverage to provide data, it also leads to an important problem: dramatically increased number of available variables [4]. While the applications may make use of all these available inputs provided by various sources, they cannot rely on them due to network congestion, device failures, inadequate infrastructure, etc. in order to maintain the operation.

Venkatesh et al. [5] provide a modular approach for context aware IoT applications to address the increased input problem. They identified that the current approaches are very inefficient and introduce significant redundancy. They also pointed out how the complexity of such applications increases with the increase in the number of inputs. This introduces many scalability issues while dealing with a large amount of sensor data in a heterogeneous application environment. Most of the energy management applications are designed to make fast real time decisions within

the network (interconnect of smart devices). These systems are generally based on light energy efficient hardware (compute constrained mobile and embedded CPU's) and often times running multiple applications in parallel. This makes the amount of inputs required for training such systems practically challenging. Thus, we identify the inherent need for reducing the number of inputs for residential applications. This reduction in the number of inputs will reduce the computation required by the application and hence improves the its efficiency.

In our work, we mainly consider residential customer installations, with smart meters and smart appliances. The goal of smart meter deployment in residential buildings is to 1) monitor the real-time energy consumption in a house and 2) manage the controllable loads (such as smart appliances, HVAC, lighting) to increase the efficiency of the system. The power consumption data for the appliances represents the usage (of appliances) patterns/habits of a user which can be used to characterize total energy consumption at certain periods in time. It is also important to note that individual appliance power consumption values may have different relationship with the total power consumption of the house e.g:if a microwave oven is turned on at a particular time it might represent that the resident will potentially use other appliances soon to increase the total power consumption. But in case of a light, if a light bulb is 'on' that may not represent increased total energy consumption in the house. We use TESLA [6] method that to predict time series data based on multiple inputs. This algorithm produces highly accurate and efficient predictions with proper input.

Finally, we present a framework to reduce the number of inputs required for the residential energy prediction application. Our framework is based on ordinary least square regression and forward selection to reduce the number of inputs for a given application. Our method identifies the multicollinearity within the input data set and selects the inputs that provide the most useful information for the application. In this setup, individual appliance power consumption values are used as inputs to prediction. We first determine the correlation between all the inputs and the output (total power consumption). From this correlation information, we determine the best input data, that would not decrease the quality of prediction below a certain threshold. Our framework also makes sure that it chooses the set of best possible unique inputs that results in prediction error below a given threshold. We test our framework on a publicly available data set provided by Pecan Street Inc. [7], that includes disaggregated energy consumption values for hundreds of houses.

Experimental results verify that up to 94% of the analyzed houses can leverage our method, i.e. they can reduce the amount of data required for prediction while meeting target error rates. In the meantime, we achieve up to 80% data reduction. Finally, using our detailed appliance selection analysis, we can define a common set of appliances that can potentially be used for total house energy consumption prediction across multiple houses.

RELATED WORK

Residential energy prediction has been an important research topic for several years. Traditionally, it has been an important problem mostly for the electricity providers, so that they could balance the supply and the demand in the electrical grid [8]. As a result, most of the residential energy modeling and estimations were performed in large-scale [9]. In addition, there were country- or region-specific (e.g. U.S. [10], Europe [11], China [12], etc.) and application-specific (e.g. climate change [13], sustainability [14]) studies as well. With the recent technological advancements (such as the Internet of Things, etc.), modeling, analyzing and controlling the energy consumption of a single house has become very important [15], creating the Smart Home domain, which is a very good CPS example, with physical devices, such as appliances, controlled by small microcontrollers, forming the cyber part [16]. Thus, predicting the energy consumption of a single house accurately, rather than thousands of houses together, has become a necessity.

Energy prediction of a house requires time-series data analysis. Several research projects have proposed a variety of methods. These include using linear regression on historical data [17], support vector regression [18], ARIMA [19], etc. All these studies assume that the houses are already equipped with smart meter infrastructure to collect data or rely on manual user surveys [20]. Furthermore, other researchers argue that analyzing the energy consumption at the appliance level (disaggregated energy data) would produce more accurate results and help us understand the patterns and trends in energy consumption better [21], [22]. These studies can further use data dimensionality reduction methods to mitigate the large data problem, such as principal component analysis [23], or independent component analysis [24]. Although these methods can reduce the amount of data to be fed into prediction methods, they still require the full set of initial data available to calculate the reduced set, i.e. they do not completely eliminate the need for all the available data.

All of these methods require significant installation costs (per appliance installation costs), as well as computation and communication infrastructure to process the large amounts of data to be generated by individual devices. Furthermore, although smart meter deployment is very common, it is still not as widespread as desired, and the data generated by those smart meters are not easily accessible even by the household members creating the data. This necessitates further deployment for smart home applications that require residential energy prediction. In this paper, we develop a framework, which can choose the most relevant data required to predict the entire house energy consumption. We can achieve significant data reduction, that translates into reduced installation and infrastructure cost. Additionally, using this framework, we can identify common appliances that can be used across multiple houses, decreasing the initial analysis overhead to be performed on individual houses.

DATA REDUCTION FRAMEWORK

The main contribution of this paper is the input filtering framework for the residential energy prediction. The main idea is to determine and leverage the correlation between the input variables (individual appliance power consumption data) and the output (overall energy consumption). We use Pecan Street database [7] as our main data source mainly because 1) it has data for several different houses and 2) it provides disaggregated data corresponding to individual appliances.

In our current framework, we leverage TESLA (Taylor Expanded Analog Forecasting Algorithm) [6], a statistical learning model that can be fully generalized, as the prediction algorithm. It provides efficient model generation: $O(n^{\alpha})$, where n is the number of inputs and α is the function order of the Taylor expansion. The generic function of this expansion is established as follows:

$$\begin{matrix} \sum_{i=o}^{n} C_i x_i & (1^{st} order) \\ \sum_{i=0}^{n}\sum_{j=0}^{i} C_{ij} x_i x_j & (2^{nd} order), \quad etc. \end{matrix}$$

where C_{ij} represents coefficients learned with observations, and $x_0 = 1$ (the constant factor). The resulting equation is $Ax = B$, where A is the matrix of input observations; x is the vector of coefficients, and B is the vector of output observations, each entry correlating with the corresponding row of A, and solved by least squares estimation. Higher function orders are able to represent more accurate correlations between input variables, but they require exponentially more training samples with respect to α for example, 1^{st}-order (linear) functions only require n samples, whereas 2^{nd}-order functions require n^2 samples.

We use TESLA in our study for its versatility and ease of model generation, but other prediction algorithms can be used as well. TESLA assumes that there is a relation between inputs (individual appliance power consumption values) and the output (total power consumption of the house), but this relation is not known and it does not make any assumptions on this relation. It tries to approximate this relation by its Taylor expansion. Next, we are going to show how we are reducing the number of inputs required to predict the output value by using the correlation between inputs and the output value.

Correlation Calculation

To quantify the correlation of inputs to the output values, we use the **VIF** (variance inflation factor) coefficient. It provides the measure of increase in variance of the regression predictor variable due to collinearity. To determine this we use the ordinary least squares (OLS) linear regression model to predict a certain variable using others.

$$VIF = \frac{1}{1-R^2}$$

Here, R^2 is calculated from the OLS model [25]. Its value is a fraction between 0.0 and 1.0. A value of 0.0 denotes that there is no linear relationship between X and Y and a value of 1.0 denotes that these two variables are perfectly correlated. We use the VIF value to determine the collinearity of the inputs. For example, VIF = 1 indicates no correlation. VIF values higher than 10 show very high correlation. Generally, VIF values higher than 5 depict strong correlation, as shown in [26] and [25].

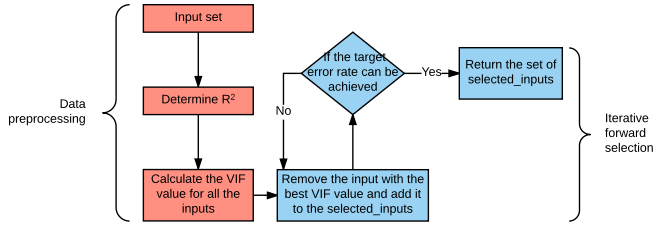


FIGURE 1: Input filtering algorithm given a set of input variables (appliance power consumption) and ground truth (total power consumption). Red parts show the data preprocessing and the blue parts show the iterative decision making processes.

Forward Selection

After determining the correlation among the input variables, we select the best input(s) (the input with highest correlation with the output- total energy consumption, using the calculated VIF values) for energy prediction. If the prediction error after selecting this input is within an acceptable threshold, we do not need to perform forward selection. If the error has not yet converged to a desired value (or range), we perform forward selection to add another input to our selected set (to reduce the overall error when estimating the output). The next input to be added to the selected set, the goal is to find the variable that can add the largest possible missing variance (in the predictor). We use a forward selection based scheme iteratively determine this input variable. The next subsection explains this iterative process in more detail.

Iterative Framework

The main framework using correlation calculation and forward selection is shown as a flow chart in Figure 1. We assume that our system is provided with a set of input variables (appliance energy consumption data) and some ground truth output (total energy consumption data). Given this 'input set' (list of all input variables available), we start by determining the correlation factor R^2. We then select the input with the highest correlation value (R^2) and pass this into a set which we label as 'selected set'. We refer to the remaining set of inputs as the 'test set'.

The 'test_set' and the 'selected_set' are then passed to the forward selection method. The forward selection method takes all the input variables present in the 'selected_set' and adds one input from the 'test_set' at a time. We label this temporary set as 'temp_set' and the input taken from the 'test_set' as 'test_input'.

Data: Individual appliance power data = all_input
Result: Selected appliances subset = $selected\_input$
set $threshold\_value$;
for *each appliance* $\in$ all_input **do**
  calculate VIF value;
end
$current\_error = \infty$;
$selected\_input = \emptyset$;
$test\_set$ = all_input;
while $current\_error > threshold\_value$ **do**
  $test\_input = \arg\max_{appliance} VIF\ values$;
  $selected\_input = selected\_input + test\_input$;
  $test\_set = test\_set$ - $test\_input$;
  $current\_error$ = prediction error with $selected\_input$;
end
return $selected\_input$;

Algorithm 1: Main iterative framework of our method

This 'test_input' is the one that has the highest VIF value in the 'test_set'. The method then determines the prediction error using the 'temp_set' to obtain the output (total power consumption of the house). We calculate the error as 'Normalized Mean Absolute Error' for this prediction and save the results to compare with the threshold value. This method finds the most relevant input variable, with whose addition to the 'selected_set', we can get high reduction in the prediction error (here prediction error is calculated with respect to the 'selected_set'). Now both 'test' and 'selected' sets are returned to the main framework. Here, we determine if the NMAE is higher than some specified threshold (we use different error threshold values - see the Results section). If the error is higher than the threshold, the 'selected' and the 'test sets are again passed to the forward selection method for the selection of the next input. In summary, if the error is higher than some required or acceptable threshold, our framework keeps adding more input variables, so that the new input would provide some additional information about the output value to reduce the prediction error. Algorithm 1 outlines the steps of our method.

EXPERIMENTAL RESULTS

This section first introduces the experimental setup to evaluate our input filtering mechanism and then presents the results. The prediction algorithm we use, i.e. TESLA, takes disaggregated residential power consumption data and produces total power consumption estimations. The prediction accuracy gets better if there are more data available at per appliance level. We use Pecan Street [7] database to obtain this disaggregated data (per appliance or plug energy consumption values). The data set consists of power data, along with their corresponding timestamps, for various set of appliances found in hundreds of houses. The data contains raw power consumption values for 15-minute time intervals. The number of samples vary from house to house, but generally there are more than 80,000 per appliance.

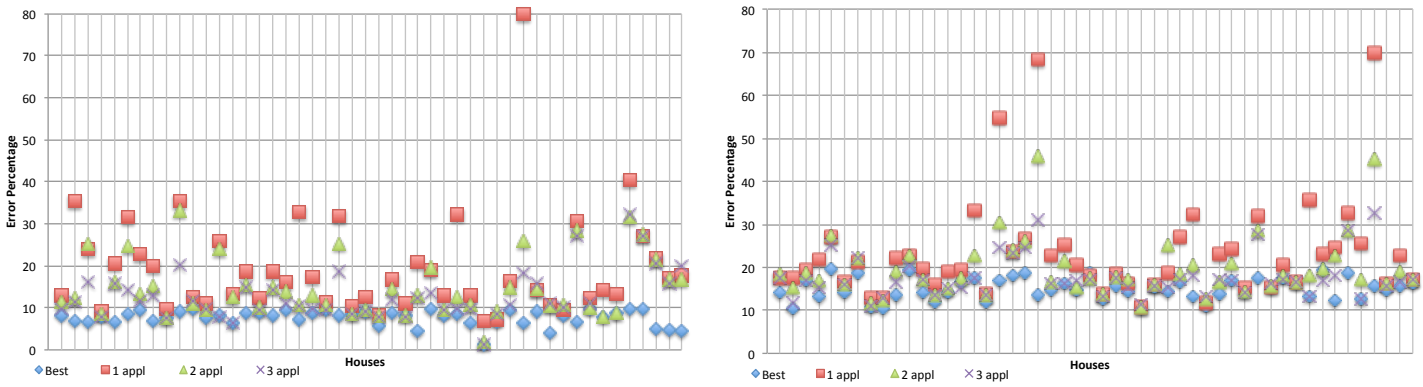
However, not all houses have power consumption data for a variety of appliances, which severely affect the baseline prediction performance (where we use all the available appliance data for prediction for the best performance). Thus, we first make an initial analysis on this database to identify a number of houses that have data for as many individual appliances as possible, so that TESLA can give accurate results. We collect individual appliance/plug energy consumption data for the year 2014, and then apply TESLA for various training and test durations (we use 1-7 days for training interval and 1-8 weeks for testing). After this analysis, we identify two sets of houses: 1) $Set10$: The best prediction error is less than 10%. The number of houses in this category is 48. 2) $Set20$: The best prediction error is between 10% and 20%. The number of houses in this category is 185. Together, the number of houses we analyze is 233. Finally, we set the error threshold as 10%, for $Set10$ and 20% for $Set20$ for the iterative forward selection algorithm. This way, our goal is to get as close as possible to the best prediction performance using the least amount of appliance data possible.

We report our results in three categories: 1) We compare the error performance of the best prediction vs. our forward selection-based method that selects up to 3 appliances. 2) We analyze the amount of possible data reduction using our forward selection-based method. This is important in terms of reducing the installation overhead, data storage overhead and communication overhead. 3) We investigate the selected appliances across the houses. Our goal is to identify if there are any common appliances that are selected by the majority of the houses.

Error Performance of Our Method

This section compares the error performance of the best prediction vs. our forward-selection-based prediction for the houses in $Set10$ and $Set20$. Figure 2 outlines the main results, where Figure 2a shows the results for $Set10$ and Figure 2b shows them for $Set20$ (note that in Figure 2b, we show only a subset of all available houses - 50 out of 185 - for clarity of the figure). In both figures, y-axis represents the error percentage values and each vertical line crossing the x-axis corresponds to a house. For both graphs, each 4-tuple (shapes: blue diamond, red square, green triangle, and purple cross) placed on a vertical line represents a house. Below is the explanation of the each value in a 4-tuple:

1. Blue diamond: Best prediction error value that uses all available appliance data. As a result all the blue diamond points are below 10% in Figure 2a and in between 10% and 20% in Figure 2b.
2. Red square: The prediction error value where forward-selection chooses 1 appliance



(a) Best error value is $< 10\%$

(b) Best error value is $< 20\%$

FIGURE 2: Error performance with different numbers of selected appliances

Set10 – 48 houses					Set20 – 185 houses				
	1 appl	2 appl	3 appl	Total		1 appl	2 appl	3 appl	Total
# houses that can achieve 10% error	6	7	8	21	**# houses that can achieve 20% error**	72	30	27	129
Percentage	12.50	14.58	16.67	43.75	**Percentage**	39.13	16.30	14.67	70.11
# houses that can achieve 15% error	22	9	6	37	**# houses that can achieve 25% error**	117	31	14	162
Percentage	45.83	18.75	12.50	77.08	**Percentage**	63.59	16.85	7.61	88.04
# houses that can achieve 20% error	31	7	5	43	**# houses that can achieve 30% error**	143	21	9	173
Percentage	64.58	14.58	10.42	89.58	**Percentage**	77.72	11.41	4.89	94.02

TABLE 1: Number of houses that can reach different target error rates in *Set*10 and *Set*20

3. Green triangle: The prediction error value where forward-selection chooses 2 appliances
4. Purple cross: The prediction error value where forward-selection chooses 3 appliances

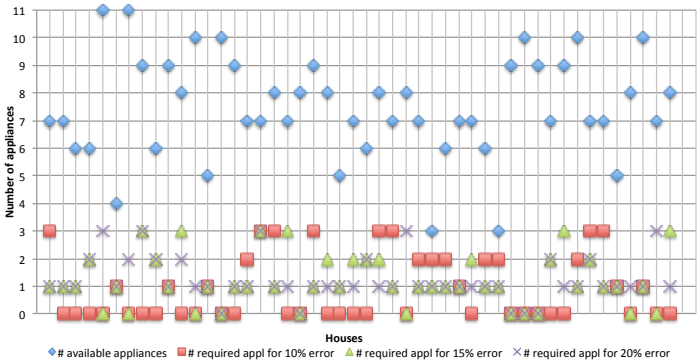
The figure shows that our forward-selection method can choose the most relevant appliance data for predicting the total energy consumption of the individual houses in several cases. For example, if we choose the target error rate as 10% for the houses in *Set*10, by choosing up to 3 appliances with our method, 45% of the houses can achieve error less than 10%. We also analyzed different target error rates. For the houses in *Set*10, if we choose target error rate as 15%, by choosing 3 appliances, 77% of the houses can achieve the target error rate. Similarly, if we set the target as 20%, 90% of the houses can achieve the target error rate. Table 1 outlines these results for houses in *Set*10 and *Set*20. Note that, the houses using 1 appliance vs. 2 appliances vs. 3 appliances are mutually exclusive, i.e. if a house can meet the target error rate with 1 appliance, that house is not considered further for 2 or 3 appliances. In the table, we can see that more houses achieve their target error values. In *Set*20, the percentage of houses that can reach the target error rate goes up to 94%.

We can say that our forward-selection-based method is effective in terms of finding the most relevant appliance data to predict the overall house energy consumption.

Data Reduction using Our Method

In this section, we demonstrate how much data reduction is possible using our method. Data reduction is important for cyber-physical system applications because more data requires 1) more devices installed to collect that data - increased device cost, 2) more data storage to store the data - increased device cost, 3) more complex prediction and machine learning methods - increased computation cost, and 4) more communication bandwidth to send the data to computation points - increased communication cost. Thus, our method targets to reduce the amount of data without sacrificing too much of the prediction performance.

Figure 3 demonstrates the main results of this section, where Figure 3a shows the results for *Set*10 and Figure 3b shows them for *Set*20 (note that in Figure 3b, we show only a subset of all available houses - 50 out of 185 - for clarity of the figure). In both figures, y-axis represents the the number of appliances and each vertical line crossing the x-axis corresponds to a house. For both graphs, each 4-tuple (shapes: blue diamond, red square, green triangle, and purple cross) placed on a vertical line represents a house. Below is the explanation of the each value in a 4-tuple:



(a) Best error value is $< 10\%$

(b) Best error value is $< 20\%$

FIGURE 3: Number of appliances required for different error percentage targets

1. Blue diamond: Total number of appliances for which power data is available in a particular house.
2. Red square: The number of appliances required for 10% (Figure 3a) and 20% (Figure 3b) target error rate
3. Green triangle: The number of appliances required for 15% (Figure 3a) and 25% (Figure 3b) target error rate
4. Purple cross: The number of appliances required for 20% (Figure 3a) and 30% (Figure 3b) target error rate

A sample reading from the figures is as follows: The first vertical line in Figure 3a illustrates that the respective house has a total of 7 appliances for which time-series power consumption data is available. If set the target error rate for this house as 10%-15%-20%, our method identifies 3-1-1 appliances, respectively, i.e. we can use these subsets of appliances to achieve the target error rate, rather than using all possible appliances. Thus the data required for prediction can be reduced by 57%, 86%, and 86% for 10%, 15%, and 20% target error rates, respectively. In both figures if any data point is placed at 0 (zero) number of appliances, it means that the respective target error rate is not achievable for that house (i.e. the data reduction is 0%).

These figures also support the results shown in Table 1. For example, for the houses in *Set*10, 43 houses can reach the set error target rates. Thus, in Figure 3a, we have 5 houses, where all the data points of those houses (except the total appliance numbers) are placed at 0. For most of the houses, we can see that the number of available appliances range from 7 to 11, which means that our forward-selection-based method can reduce this number to 1-3 without sacrificing the prediction performance.

Table 2 summarizes the results of this study. We see that 66% data reduction on average is possible in the worst case (with the lowest target error rate), and in the best case, 80% average data reduction is possible (with a little increased target error rate).

Set10 – data reduction percentage		**Set20 – data reduction percentage**	
with 10% target error	66.66	*with 20% target error*	70.78
with 15% target error	76.78	*with 25% target error*	77.26
with 20% target error	79.87	*with 30% target error*	79.91

TABLE 2: Average data reduction percentage values for different target error rates in *Set*10 and *Set*20

Selected Appliance Analysis

In this section, we analyze the selected appliances across the houses in *Set*10 and *Set*20. For this analysis, we selected up to 3 appliances for all the houses in *Set*10 and *Set*20 using our forward-selection-based method. Table 3 summarizes the results of this study, where Table 3a lists the selected appliances for *Set*10 and Table 3b shows the selected appliances for *Set*20. In both tables, the first column lists the list of appliances that are selected by at least one house (note that we preserve the naming notation from Pecan Street database). The next six columns show the number of houses (and their percentage to the total size of the respective set) that select a specific appliance as the first, second, and third choice. Finally, the last two columns show the total number of houses (and their percentage to the total) that select a specific appliance as one of the first, second or third choice (i.e. it is a summary column).

We see that in Table 3a, the three most selected appliances are furnace, dryer, and refrigerator. Table 3b produces similar results: dryer, furnace and refrigerator. This means that for the majority of the houses furnace, dryer, and refrigerator power consumption data are enough to predict the power consumption of the entire house. This is because these appliances can also explain the usage of other appliances in the house, i.e. the usage of other appliances shows high correlation with these selected appliances. In other words, if we want to select a common set of appliances to instrument in a variety of houses, in order to pre-

Appliance	Appliance1		Appliance2		Appliance3		Sum per appliance	Total %
	Count	%	Count	%	Count	%		
furnace1	19	39.58	6	12.50	4	8.70	29	60.42
dryer1	7	14.58	12	25.00	6	13.04	25	52.08
diningroom1	1	2.08	0	0.00	0	0.00	1	2.08
car1	5	10.42	3	6.25	0	0.00	8	16.67
refrigerator1	6	12.50	4	8.33	5	10.87	15	31.25
oven1	1	2.08	5	10.42	5	10.87	11	22.92
dishwasher1	2	4.17	1	2.08	8	17.39	11	22.92
poolpump1	5	10.42	0	0.00	0	0.00	5	10.42
clotheswasher1	1	2.08	3	6.25	7	15.22	11	22.92
livingroom1	1	2.08	7	14.58	1	2.17	9	18.75
microwave1	0	0.00	3	6.25	4	8.70	7	14.58
bedroom1	0	0.00	4	8.33	4	8.70	8	16.67
bathroom1	0	0.00	0	0.00	2	4.35	2	4.17

(a) *Set*10 - 48 houses

Appliance	Appliance1		Appliance2		Appliance3		Sum per appliance	Total %
	Count	%	Count	%	Count	%		
furnace1	59	32.07	19	10.56	15	8.62	93	50.27
dryer1	52	28.26	40	22.22	5	2.87	97	52.43
diningroom1	0	0.00	3	1.67	0	0.00	3	1.62
car1	24	13.04	5	2.78	0	0.00	29	15.68
refrigerator1	5	2.72	18	10.00	38	21.84	61	32.97
oven1	13	7.07	19	10.56	16	9.20	48	25.95
dishwasher1	6	3.26	14	7.78	27	15.52	47	25.41
poolpump1	10	5.43	2	1.11	0	0.00	12	6.49
clotheswasher1	3	1.63	16	8.89	27	15.52	46	24.86
livingroom1	5	2.72	17	9.44	18	10.34	40	21.62
microwave1	1	0.54	20	11.11	19	10.92	40	21.62
bedroom1	3	1.63	4	2.22	7	4.02	14	7.57
bathroom1	3	1.63	3	1.67	2	1.15	8	4.32

(b) *Set*20 - 185 houses

TABLE 3: Selected appliances analysis for the houses in *Set*10 and *Set*20

dict their entire house power consumption, these three appliances (furnace, dryer, and refrigerator) constitute a common set.

CONCLUSION

Smart homes have recently become an important cyber-physical system (CPS) domain by adding various sensor and computational devices to traditional homes. These smart homes leverage the advancements in technology (such as sensor technology, smart meters, smart appliances) and provide several applications to their users, such as automated residential energy management, increased security, etc. From these applications, residential energy management is particularly important due to the distributed nature of the grid and the considerable portion of residential domain in overall energy demand. It is important for both users and the electricity providers to accurately predict the individual energy consumption. The users require prediction to manage their loads in a more cost and energy efficient way. The electricity providers, on the other hand, need prediction values to better balance the supply and demand in the electrical grid. However, accurate prediction may require significant amount of data. It might not always be possible to rely on the existence of such data due to network congestion, device failures, inadequate infrastructure, etc. Furthermore, although the technology has made it easier to connect even to the smallest device, there are numerous houses that do not even have the necessary infrastructure to place smart meters. Thus, this application needs to be implemented in a way that requires minimal data availability. In this paper, we present a method to decrease the amount of necessary data to perform accurate energy consumption prediction. Our method finds the most relevant and useful subset of data within the input domain. We model the inherent correlation among user input variables and implement an approach using least square regression and forward selection to select the most relevant variables. One advantage of our method is that, we do not require all the data within the input domain to be available at the same time in order to apply filtering. To test our method, we use a public database that has data for hundreds of residential homes. Our evaluation shows that our method can reduce the number of variables required for the prediction application effectively, where up to 94% of the houses can meet target error rates. Furthermore, while doing so, we can reduce the required data by 80%. This data reduction can translate into significant reduction in device costs, data storage costs, computation and communication costs. And finally, we show a comprehensive analysis about the selected appliance data across the houses investigated. Using these results, we can define a common appliance set that can be used across the houses.

REFERENCES

[1] Howarth, R., and Bringezu, S., 2009. "Biofuels & environmental imp-acts: scientific analysis and implications for sustainability". *Policy Brief Series, UNESCOSCOPE-UNEP*.

[2] Khan, A., Nicholson, J., Mellor, S., Jackson, D., Ladha, K., Ladha, C., Hand, J., Clarke, J., Olivier, P., and Plötz, T., 2014. "Occupancy monitoring using environmental & context sensors and a hierarchical analysis framework.". In BuildSys@ SenSys, pp. 90–99.

[3] U.S Energy Information Administration, 2015. How many smart meters are installed in the united states, and who has them? http://www.eia.gov/tools/faqs/faq.cfm?id=108&t=3. "[Online; accessed 21-June-2016]".

[4] Perera, C., Zaslavsky, A., Christen, P., and Georgakopoulos, D., 2014. "Context aware computing for the internet of things: A survey". *IEEE Communications Surveys & Tutorials,* ***16***(1), pp. 414–454.

[5] Venkatesh, J., Chan, C., Akyurek, A. S., and Rosing, T. S., 2016. "A modular approach to context-aware iot applications". In 2016 IEEE First International Conference on Internet-of-Things Design and Implementation (IoTDI),

IEEE, pp. 235–240.
[6] Akyurek, B. O., Akyurek, A. S., Kleissl, J., and Rosing, T. Š., 2014. "Tesla: Taylor expanded solar analog forecasting". In Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on, IEEE, pp. 127–132.
[7] Pecan Street Inc., 2016. Dataport.
[8] Lund, H., Andersen, A. N., Østergaard, P. A., Mathiesen, B. V., and Connolly, D., 2012. "From electricity smart grids to smart energy systems–a market operation based approach and understanding". *Energy,* ***42***(1), pp. 96–102.
[9] Aydinalp, M., Ugursal, V. I., and Fung, A., 2003. "Modelling of residential energy consumption at the national level". *International Journal of Energy Research,* ***27***(4), pp. 441–453.
[10] Ewing, R., and Rong, F., 2008. "The impact of urban form on us residential energy use". *Housing policy debate,* ***19***(1), pp. 1–30.
[11] Balaras, C. A., Gaglia, A. G., Georgopoulou, E., Mirasgedis, S., Sarafidis, Y., and Lalas, D. P., 2007. "European residential buildings and empirical assessment of the hellenic building stock, energy consumption, emissions and potential energy savings". *Building and environment,* ***42***(3), pp. 1298–1314.
[12] Crompton, P., and Wu, Y., 2005. "Energy consumption in china: past trends and future directions". *Energy economics,* ***27***(1), pp. 195–208.
[13] Isaac, M., and Van Vuuren, D. P., 2009. "Modeling global residential sector energy demand for heating and air conditioning in the context of climate change". *Energy policy,* ***37***(2), pp. 507–521.
[14] Darby, S., et al., 2006. "The effectiveness of feedback on energy consumption". *A Review for DEFRA of the Literature on Metering, Billing and direct Displays,* ***486***(2006).
[15] Venkatesh, J., Aksanli, B., Junqua, J.-C., Morin, P., and Rosing, T. S., 2013. "Homesim: Comprehensive, smart, residential electrical energy simulation and scheduling". In Green Computing Conference (IGCC), 2013 International, IEEE, pp. 1–8.
[16] Han, D.-M., and Lim, J.-H., 2010. "Smart home energy management system using ieee 802.15. 4 and zigbee". *IEEE Transactions on Consumer Electronics,* ***56***(3).
[17] Fumo, N., and Biswas, M. R., 2015. "Regression analysis for prediction of residential energy consumption". *Renewable and Sustainable Energy Reviews,* ***47***, pp. 332–343.
[18] Jain, R. K., Smith, K. M., Culligan, P. J., and Taylor, J. E., 2014. "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,* ***123***, pp. 168–178.
[19] Suganthi, L., and Samuel, A. A., 2012. "Energy models for demand forecastinga review". *Renewable and sustainable energy reviews,* ***16***(2), pp. 1223–1240.
[20] US Energy Information Administration, 2009. Residential Energy Consumption Survey.
[21] Kolter, J. Z., and Johnson, M. J., 2011. "Redd: A public data set for energy disaggregation research". In Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA, Vol. 25, pp. 59–62.
[22] Barker, S., Mishra, A., Irwin, D., Cecchet, E., Shenoy, P., and Albrecht, J. "Smart*: An open data set and tools for enabling research in sustainable homes". *SustKDD'12*.
[23] Wold, S., Esbensen, K., and Geladi, P., 1987. "Principal component analysis". *Chemometrics and intelligent laboratory systems,* ***2***(1-3), pp. 37–52.
[24] Hyvärinen, A., Karhunen, J., and Oja, E., 2004. *Independent component analysis*, Vol. 46. John Wiley & Sons.
[25] Nachtsheim, C. J., Neter, J., Kutner, M. H., and Wasserman, W., 2004. "Applied linear regression models". *McGraw-Hill Irwin*.
[26] Hair, J. F., 2009. "Multivariate data analysis".