Effective Features to Predict Residential Energy Consumption Using Machine Learning

Conference Paper · June 2019

DOI: 10.1061/9780784482445.036

3 authors:

Yunjeong Mo
Michigan State University
13 PUBLICATIONS **83** CITATIONS
SEE PROFILE

Dong Zhao
Michigan State University
71 PUBLICATIONS **850** CITATIONS
SEE PROFILE

M. G. Matt Syal
Michigan State University
69 PUBLICATIONS **932** CITATIONS
SEE PROFILE

Effective Features to Predict Residential Energy Consumption Using Machine Learning

Yunjeong Mo, Ph.D., S.M.ASCE[1]; Dong Zhao, Ph.D., M.ASCE[2];
and Matt Syal, Ph.D., M.ASCE[3]

[1]Research Assistant, School of Planning, Design, and Construction, Michigan State Univ., East Lansing, MI 48824 (corresponding author). E-mail: moyunjeo@msu.edu
[2]Assistant Professor, School of Planning, Design, and Construction, Michigan State Univ., East Lansing, MI 48824. E-mail: dzhao@msu.edu
[3]Professor, School of Planning, Design, and Construction, Michigan State Univ., East Lansing, MI 48824. E-mail: syalm@msu.edu

ABSTRACT

Humans have a greater influence on energy consumption in residential buildings than other types of buildings. Although existing studies focus on how energy consumption is affected by building technologies and occupant demographics, few studies have incorporated the impact of occupant energy use patterns. The goal of this study is to identify the features that affect energy consumption in residential buildings and to measure their predictive performance. The researchers examined the impact of occupants' energy use behaviors and the energy use patterns of home appliances on home energy consumption. The patterns reflect on a combination of appliances, their use times and frequencies, and the configurations set by users. Data from the Residential Energy Consumption Survey (RECS) are analyzed to select features for prediction, using multiple machine learning algorithms including support vector machine (SVM) and random forest. The results provide a list of features that efficiently predict energy consumption in residential buildings. The selected 32 features achieve 98% of the prediction performance of that from the entire 271 features. This list of effective features can be used to improve the effectiveness of energy saving programs and to educate occupants about their energy use patterns. The relationship between occupants' behavior patterns and energy use patterns revealed from this study provides the groundwork for researchers to further explore the prediction of occupant behavior from energy consumption.

INTRODUCTION AND BACKGROUND

The residential sector accounts for 39% of the total electricity consumption in the United States, according to the U.S. Department of Energy (U.S.DOE 2017). Occupants have a greater impact on the energy consumption in residential buildings than in other types of buildings (Zhao et al. 2018). Energy consumption in individual household depends on various factors, including environmental conditions, building technology, resident demographics, Heating, Ventilation and Air Conditioning (HVAC) systems, appliances in the home (Zhao et al. 2017). Among the factors, the usage pattern of HVAC systems and appliances are more related to occupant behavior and energy costs of households (McCoy et al. 2018).

However, a comprehensive understanding of the features affecting home energy consumption is lacking, without which it is less likely to develop effective energy efficiency programs and provide relevant educational information to occupants. The goal of this research is to identify the features that effectively affect energy consumption in residential buildings as measured by their predictive performance. In particular, behavior-related features from appliances and their usage patterns are separately examined to see the effects of occupant behavior on energy consumption.

This paper consists of three sections: (1) the descriptions of data and variables; (2) the selection of critical features for electricity consumption; and (3) the measurement and comparison of predictive performance of the selected features.

DATA

The Residential Energy Consumption Survey (RECS) is a national energy survey for residential buildings conducted by the DOE Energy Information Administration (EIA). The survey has been conducted every three years since 1978 (Sanquist et al. 2012). The RECS collects households' energy data over one year with building, appliance, and occupant information. The RECS dataset is a good source of energy consumption and occupant behavior, which has been used by many studies. Sanquist et al. (2012) performed a lifestyle analysis of electricity consumption in residential buildings using the 2005 RECS data. Diao et al. (2017) identified and classified occupant behaviors with energy consumption outcomes. Aksanli et al. (2016) developed a residential energy modeling framework based on human activities to estimate the energy consumption in residential buildings.

Different from existing studies that focused on a part of RECS variables, this study incorporates the full RECS variables (e.g. electrical appliances at home, building technologies, occupant behaviors, and occupant demographic information). Specifically, the year of 2015 RECS microdata is used, which contains household characteristics, energy consumption and expenditures data from 5,686 household instances. After removing imputation flags, replicate weights, and irrelevant features from the initial features, the remaining 272 features are grouped by their characteristics as described in Table 1. Appliance, Behavior, Technology, and Demographic categories are used as inputs in feature selection and machine learning (ML) algorithms to predict total electricity consumption in kWh, which is a numeric variable. As expected from the number of features, the RECS has detailed data about appliances, building technology, occupant demographics, but has less detailed data about occupant behavior.

Table 1. Categories of RECS Data

Category	Feature Examples	Count
Appliance	Appliances, Lighting, Internet, Number, Size, Type, Age, Fuel type for appliances, Energy star appliances	81
Behavior	Frequency, Duration, Number of days/months used, Heating/cooling temperature set-point, Dishwasher, washer, dryer temperature and cycle setting, Smart meter data check	32
Technology	Building envelope, HVAC, Water heater, Fuel type for Tech, Thermostat, Light controller, Sensor, Smart meter install, Building audit, Pool, Hot tub	117
Demographic	Occupant/family characteristics, Who pays bill, Receive/participate in home energy assistance program	41
kWh	Electricity usage in kWh	1

METHODS

The methodology follows the ML features selection and algorithm selection process. Features are selected from different categories, and the selected features are used to predict total energy consumption using various ML algorithms. The efficiency of the selected features is evaluated by comparing the prediction performance of the selected features and the prediction

performance of all features together. The categories for feature selection and energy consumption prediction are as follows: (1) All, (2) Appliance, (3) Behavior, (4) Technology, (5) Demographic, and (6) Appliance + Behavior.

Feature Selection

Feature selection is the process of selecting a subset of features to be used for model construction in ML and statistics. It is also called attribute selection, variable selection, or variable subset selection James et al. (2013). It aims to find faster and more cost-effective predictors, improve the prediction performance of the predictors, and help researchers understand the underlying process better (Guyon and Elisseeff 2003). In this study, Correlation-based Feature Selection (CFS) with Greedy Stepwise method is used for the feature selection. It evaluates the worth of a subset of features by considering the single predictive ability of each feature and the degree of redundancy between them. Greedy Stepwise performs a greedy search forward or backward through the feature subset. It stops when the addition or deletion of any remaining features results in a decreased performance evaluation (Hall 1998).

First, feature selection is performed for all features to identify the most critical and efficient features for predicting energy consumption among all possible features in the RECS data. Then, feature selection is performed for each category to find the most efficient features in each category. In case only limited features of data are available from the real-world datasets, this feature selection will be useful to find the most efficient features for predicting energy consumption from a limited dataset. Finally, feature selection is performed for the combination of the Appliance and Behavior categories. Appliance and Behavior are determined by the occupants more than Technology or Demographic, and they reflect the behavioral patterns of the occupants. Thus, the effect of this combination of features is examined separately.

Algorithm Selection

Linear Regression, Support Vector Machine, Random Forest, M5P Trees, and M5 Rules are used to test RECS features for predicting electricity energy consumption. (1) Linear Regression is one of the most common algorithms for numeric prediction, and thus it is used as the baseline. (2) Sequential minimal optimization (SMO) regression implements the Support Vector Machine (SVM) for regression. It produces a model that can be expressed with support vectors and can be applied to nonlinear datasets using kernel functions (Witten et al. 2016). Its predictive performance is influenced by the kernels and parameter settings, and Radial Basis Function (RBK) kernel with C value 1 and gamma value 0.01 is used in this study. (3) Decision trees and rules work more naturally with nominal features, but they can be extended to numeric features by combining with numeric-value tests into the decision tree or rule-induction scheme, with pre-discretization of numeric features into nominal ones (Witten et al. 2016). Random Forest is an ensemble learning method that builds a randomized decision tree in each iteration of the training, and outputs the mean prediction of the individual trees (Breiman 2001; Witten et al. 2016). (4) M5P Trees combines a conventional decision tree with the possibility of linear regression at the nodes (Quinlan 1992; Wang and Witten 1996). (5) M5 Rules generate a decision list for regression problems with a divide-and-conquer approach by constructing a model tree using M5 and developing the best leaf into a rule in each iteration (Holmes et al. 1999). Correlation Coefficient and Root Mean Squared Error (RMSE) are used to evaluate performance (Witten et al. 2016).

Table 2. Selected Features from All

Category	Feature	Description
Appliance	NUMFRIG	Number of refrigerators used
	SIZRFRI1	Size of most-used refrigerator
	ICE	Through-the-door ice on most-used refrigerator
	SIZFREEZ	Size of most-used freezer
	OVEN	Number of separate ovens
	TVCOLOR	Number of televisions used
	NUMCFAN	Number of ceiling fans used
Behavior	MONPOOL	Months swimming pool used in the last year
	DRYRUSE	Frequency clothes dryer used
	TVONWE1	Most-used TV usage on weekends
Technology	UATYP10	Census 2010 Urban Type
	TYPEHUQ	Type of housing unit
	NCOMBATH	Number of full bathrooms
	TOTROOMS	Total number of rooms in the unit, excluding bathrooms
	UGASHERE	Natural gas available in neighborhood
	POOL	Heated swimming pool
	FUELTUB	Fuel used for heating hot tub
	FUELHEAT	Main space heating fuel
	AIRCOND	Air conditioning equipment used
	COOLTYPE	Type of air conditioning equipment used
	CENACHP	Central air conditioner is a heat pump
	FUELH2O	Fuel used by main water heater
	FUELH2O2	Fuel used by secondary water heater
	ELWARM	Electricity used for space heating
	ELWATER	Electricity used for water heating
	ELFOOD	Electricity used for cooking
	FOWATER	Fuel oil used for water heating
	CLIMATE_REGION	Building America Climate Zone
Demographic	NHSLDMEM	Number of household members
	NUMADULT	Number of household members age 18 or older
	NOACBROKE	Unable to use cooling equipment in the last year because equipment was broken/ could not afford repair or replacement
	PERIODNG	Number of days covered by Energy Supplier Survey gas billing data/used to calculate annual consumption and expenditures

RESULT

Effective Features of Energy Consumption Prediction

The feature selection results suggest key features for predicting energy consumption. Among all 271 features, 32 features are selected as summarized in Table 2. From the Appliance category, the number and size of refrigerators and freezers, and the number of ovens, televisions, and

ceiling fans are selected. This means that the usage patterns of these appliances are good predictors of the total electricity consumption in a residential building. From the Behavior category, duration of swimming pool usage, frequency of clothes dryer usage, and duration of TV usage on weekends are selected. From the Technology category, climate, the location and type of the house, and Heating, Ventilation and Air Conditioning (HVAC)-related features including fuel type are selected. From the Demographic category, the number of total household members and number of adults are selected. Two economic features were also selected from this category: if a household could not afford repair or replacement of broken cooling equipment, and the number of days covered by Energy Supplier Survey natural gas billing data. While these features indicate the economic status of the household, they indirectly provide cooling and heating information about the household as well. In summary, HVAC, refrigerator/freezer and TV, climate, location, house type, and number of household members are the main features for electricity consumption prediction in residential buildings.

- When selecting features only from the Appliance category, 19 out of 81 features are selected. Refrigerators and freezers, cooling appliances, TVs, computers, smartphones, and light bulbs are the main features to predict electricity consumption from the Appliance category.
- When selecting features only from the Behavior category, 9 out of 32 features are selected. The months when swimming pools and hot tubs were used in the last year, oven, dishwasher, clothes dryer, TV usage on weekends or weekdays, water temperature for dishwasher rinse cycles, and the set-point temperature on summer nights are selected as important features for predicting electricity consumption.
- When selecting features only from the Technology category, 17 out of 117 features are selected. They include climate zone, census urban type, housing type, numbers of rooms and bathrooms, presence of a heated or unheated swimming pool, types of air conditioning equipment, and fuel types for space heating, cooling, water heating, and cooking. It is noticeable that the features selected focus mainly on HVAC and fuel types for heating, cooling, and cooking compared to features about building envelopes, which implies that HVAC-related features predict electricity consumption more efficiently.
- When selecting features only from the Demographic category, 6 out of 41 features are selected. They are: if the house is owned or rented, numbers of household members and household adults, and features indicating the economic status of the household, including it they participated in home energy assistance programs, if they could afford to repair or replace broken cooling equipment, and the number of days covered by energy supplier survey billing data.
- When selecting features from the Appliance and Behavior categories, 23 out of 113 features are selected. Previously, 19 features had been selected from Appliance only, and 9 features had been selected from Behavior only. When combining these 2 categories, use of coffee maker, number of satellite boxes are excluded, and refrigerator size is added from the Appliance category. Also, oven, dishwasher, clothes dryer usage, water temperature for dishwasher rinse cycles, the set-point temperature on summer nights are excluded, and clothes washer usage is added in Behavior category.

Effective Algorithm for Energy Consumption Prediction

The energy consumption prediction is performed for each feature category. The prediction's performance between all features and the selected features are compared using different

algorithms. As summarized in Table 3, SVM shows the best performance for the features from all categories. The correlation coefficient is 0.8024 with all 271 features, and 0.7848 with the selected 32 features. It is notable that using the selected features, which are only 12% of the total number of features, still achieves 98% of the predictive performance reached with all of the features. This implies that the selected features are an efficient way to predict electricity consumption.

- For the Appliance category, SVM shows the best performance, with a correlation coefficient of 0.6369 with all of 81 Appliance features and 0.5945 with the selected 19 features. The selected features achieve 93% of baseline predictive performance with 23% of the number of Appliance category features.
- For the Behavior category, SVM shows the best performance with a correlation coefficient of 0.5719 with all 32 Behavior features, and M5P Trees shows the best performance with a correlation coefficient of 0.5444 with the selected 7 features. The selected features achieve 95% of baseline predictive performance with 28% of the number of features.
- For the Technology category, SVM shows the best performance with a correlation coefficient of 0.7492 with all 117 Technology features, and 0.7240 with the selected 17 features. The selected features achieve 97% of baseline predictive performance with 15% of the number of features
- For the Demographic category, M5 Rules shows the best performance with a correlation coefficient of 0.5815 with all 41 Demographic features, and 0.5316 with the selected 6 features. Unlike the other categories, SVM is not the best algorithm for this category. The performance difference is not big when using all features (M5 Rules is 0.5815 and SVM is 0.5734, which is 98.6% of M5 Rules), but it shows greater differences with the selected features (M5 Rules is 0.5316 and SVM is 0.5085, which is 95.7% of M5 Rules). Using M5 Rules, the selected features achieve 92% of baseline predictive performance with 15% of the number of features.
- For the Appliance and Behavior categories together, SVM shows the best performance with a correlation coefficient of 0.6831 with all 113 features, and 0.6429 with the selected 23 features. The selected features achieve 94% of baseline predictive performance with 20% of the number of features.

Table 3. Algorithm Performance with All Features

All Features	All (271)		Selected (32)	
Algorithm	Cor.Coeff.	RMSE	Cor.Coeff.	RMSE
SVM	0.8024	4222.66	0.7848	4404.92
Linear Regression	0.7853	4367.89	0.7826	4444.12
Random Forest	0.7825	4533.78	0.7809	4405.09
M5P Trees	0.7794	4418.81	0.7755	4451.09
M5 Rules	0.7794	4418.81	0.7637	4550.76

SVM shows the best performance for most of the categories except for the Demographic category (all features and selected features) and the Behavior category (selected features). In the Demographic category, SVM reaches 99% of its performance with all features and 96% of its performance with the selected features compared to M5 Rules. In the Behavior category, SVM reaches 99% of its performance with the selected features compared to M5P Trees. Thus, the performance of the SVM algorithm in each category is compared (Table 4). The best

performance is 0.8024 with all 271 features. However, it is meaningful that the selected 32 features from all categories still achieve 98% of the all-feature baseline performance. This performance is even better with 117 Technology features (correlation coefficient 0.7492) or 113 Appliance and Behavior features (correlation coefficient 0.6831).

The selected features generally achieve more than 90% of the performance achieved with all features (Table 4 and Figure 1). Given the number of features and the performance achieved, this demonstrates that the selected features are efficient to predict electricity consumption.

Table 4. Performance Comparison by Different Features

Features (SVM)	Cor.Coeff.		# Features		Ratio (Selected/All)	
	All	Selected	All	Selected	#Features	Cor.Coeff.
All Features	0.8024	0.7848	271	32	12%	98%
Technology	0.7492	0.7240	117	17	15%	97%
Appliance + Behavior	0.6831	0.6429	113	23	20%	94%
Appliance	0.6369	0.5945	81	19	23%	93%
Demographic	0.5734	0.5085	41	6	15%	92%
Behavior	0.5719	0.5414	32	9	28%	95%

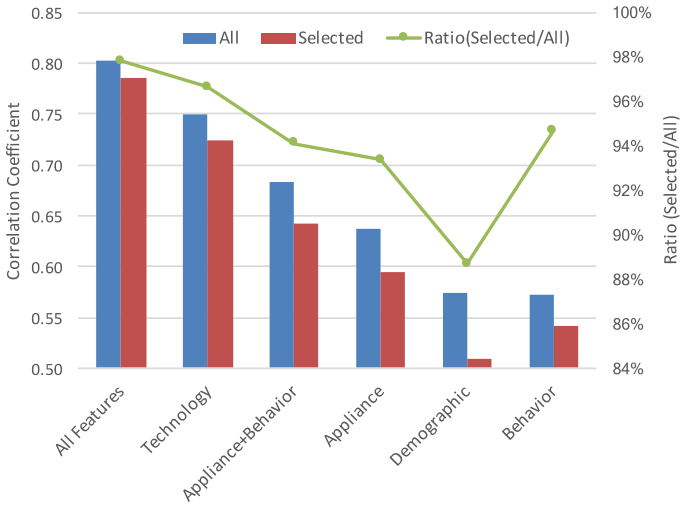

Figure 1. Performance Comparison by Different Features

CONCLUSION

Findings show that the selected 32 features reach the 78% of accuracy when predicting home energy consumption and remain 98% of prediction power of the whole 271 features. The selected features are refrigerator, freezer, oven and television from the Appliance category, TV, cloth dryer, and swimming pool usage from the Behavior category, housing type, number of rooms, energy types for heating and cooling, and climate zone from Technology category, and number of household members and number of young (under 18 years old) household members from the Demographic category.

This study provides a list of efficient features for predicting energy use in residential buildings. The features can effectively explain energy use patterns and their relationships which can be used to uncover technical and behavioral patterns behind. The findings help residential occupants understand their energy use patterns and behaviors. Furthermore, the relationships

between behavior-related features and energy usage provide the groundwork to predict occupant behavior from energy consumption data.

A limitation of this study is related to the RECS database where the data of behavior- and time-related activities are less detailed than the data of appliances, building technology, demographic information, and energy usage. Also, the RECS is lack of HVAC use behavior, such as thermostat set-point temperature during heating or cooling seasons, which might be one of the important energy-usage predictive features. In the future, behavior-related features can be further examined in detail by using datasets with more detailed behavior information.

REFERENCES

Aksanli, B., Akyurek, A. S., and Rosing, T. S. "User behavior modeling for estimating residential energy consumption." *Proc., Smart City 360°*, Springer, 348-361.

Breiman, L. (2001). "Random forests." *Machine learning*, 45(1), 5-32.

Diao, L., Sun, Y., Chen, Z., and Chen, J. (2017). "Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation." *Energy and Buildings*.

Guyon, I., and Elisseeff, A. (2003). "An introduction to variable and feature selection." *Journal of machine learning research*, 3(Mar), 1157-1182.

Hall, M. A. (1998). "Correlation-based feature subset selection for machine learning." *Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*.

Holmes, G., Hall, M., and Prank, E. "Generating rule sets from model trees." *Proc., Australasian Joint Conference on Artificial Intelligence*, Springer, 1-12.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, Springer.

McCoy, A. P., Zhao, D., Ladipo, T., Agee, P., and Mo, Y. (2018). "Comparison of green home energy performance between simulation and observation." *Journal of Green Building*, 13(3), 70-88.

Quinlan, J. R. "Learning with continuous classes." *Proc., 5th Australian joint conference on artificial intelligence*, World Scientific, 343-348.

Sanquist, T. F., Orr, H., Shui, B., and Bittner, A. C. (2012). "Lifestyle factors in US residential electricity consumption." *Energy Policy*, 42, 354-364.

U.S.DOE (2017). "How much energy is consumed in U.S. residential and commercial buildings?", <https://www.eia.gov/tools/faqs/faq.php?id=86&t=1>.

Wang, Y., and Witten, I. H. (1996). "Induction of model trees for predicting continuous classes."

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

Zhao, D., McCoy, A. P., Agee, P., Mo, Y., Reichard, G., and Paige, F. (2018). "Time effects of green buildings on energy use for low-income households: A longitudinal study in the United States." *Sustainable cities and society*, 40, 559-568.

Zhao, D., McCoy, A. P., Du, J., Agee, P., and Lu, Y. (2017). "Interaction Effects of Building Technology and Resident Behavior on Energy Consumption in Residential Buildings." *Energy and Buildings*, 134, 223-233.