Journal of Building Engineering 44 (2021) 102891

Contents lists available at ScienceDirect

Journal of Building Engineering

journal homepage: www.elsevier.com/locate/jobe

Effective factors for residential building energy modeling using feature engineering

Yunjeong Mo [a,*], Dong Zhao [b]

[a] Construction Management Department, University of North Florida, 1 UNF Drive, Jacksonville, FL, 32224, USA
[b] School of Planning, Design and Construction, Department of Civil and Environmental Engineering, Michigan State University, 552 W Circle Drive, East Lansing, MI, 48824, USA

ARTICLE INFO

Keywords:
Residential energy use
Residential building
Machine learning
Feature selection
Energy use prediction
Census region

ABSTRACT

Given the improved understanding of sustainability, hundreds of factors are identified to have relevance to building energy efficiency. However, there is still a lack of knowledge about what factors play a significant role in energy consumption prediction for residential buildings. In the absence of this information, building energy consumption prediction would not be efficient. To tackle this problem, this study creates a feature engineering-based analytic framework to select effective factors for energy consumption prediction and assess their implications. Two application cases are reported to demonstrate the efficiency improvement of energy consumption prediction for residential buildings. The cases use the Residential Energy Consumption Survey database that contains more than 270 energy use-related factors about buildings and occupants in the United States. Data analysis from the two cases shows that selected features achieve 97–102% of prediction power while using 12–15% number of factors, largely reducing the dimensionality for energy prediction. The results also produce a list of significant features that are efficient predictors for residential energy modeling and evaluation at the national and regional levels. Examples of the selected features are the total number of rooms and full bathrooms, frequency of clothes dryer used, type of the housing unit, number of ceiling fans and television. The selected features explain energy use patterns and their relationships which help designer, contractors, and occupants better understand energy, behaviors, and the built environment. The resultant energy use patterns inform regional similarities, differences, and distinctive characteristics.

1. Introduction

Based on the U.S. Department of Energy [1], residential buildings account for 21% of the total energy consumption in the United States. Energy consumption prediction in residential buildings is important for designers and engineers but is difficult to achieve because energy consumption depends on diverse factors such as environmental conditions, building technology, resident demographics, occupant behavior, heating, ventilation and air conditioning (HVAC) systems, and appliance use [2]. Occupants in residential buildings have greater influences on energy consumption since they are more independent to control building systems such as HVAC and appliances [3,4]. Moreover, the occupant behavior in residential context is more complicated compared to that in commercial context due to the higher variability and uncertainty [5,6]. The stochastic nature of occupant behavior leads to various energy consumption [7].

A comprehensive understanding of factors that influence building energy consumption is necessary to design and build sustainable homes. However, the large number of technical and behavioral factors that are related to energy use in residential buildings prevent from identifying their relationships and increase the difficulty of building energy modeling. Existing models can address energy modeling for single residential buildings but are difficult to model residential buildings at the community or city scale. This is because, along with the increase of scale, the number of factors dramatically increases, resulting in expensive computational cost and low accuracy. Building energy modeling includes parameter calibration with optimization technique which is computationally intensive. This becomes more problematic as the number of factors requiring parameter calibration increases [8]. In the absence of solutions to address the problem, accurate and prompt energy consumption prediction in residential buildings would remain challenging.

* Corresponding author.
E-mail addresses: y.mo@unf.edu (Y. Mo), dzhao@msu.edu (D. Zhao).
Available online 17 June 2021
https://doi.org/10.1016/j.jobe.2021.102891
2352-7102/
Received 26 March 2021; Received in revised form 24 May 2021; Accepted 14 June 2021

The objective of this study is to create a feature engineering-based analytic framework that can identify effective factors to improve large-scaled energy consumption modeling in residential buildings. The effective factors reduce the dimensionality of parameters and improve the efficiency in energy modeling. This study also provides two application cases to demonstrate the framework's usage and validate effective factors' prediction power. In the two application cases, the Residential Energy Consumption Survey (RECS) database by DOE was analyzed: one at the national level and the other at the regional level. The effective factors identified from the two case advance the understanding of how the factors interact and collectively impact home energy consumption. Implications of adopting this framework are discussed to inform the architecture, engineering, and construction (AEC) industry about energy modeling improvement.

2. Background

2.1. Energy consumption modeling for residential buildings

Improvement in building thermal properties and technological advancement of energy efficient appliances help to reduce building energy consumption. Given the applications of new building technologies and stricter requirements to use energy efficient appliances, yet residential building energy consumption has not decreased [9]. This indicates a strong impact of occupant behavior and living style and emphasizes the significant role of occupant behavior in residential building energy modeling. Therefore, many theoretical models are developed to explain residential building energy consumption considering occupant behavior. Darby [10] stated that energy consumption can be reduced by up to 20% when efficient energy feedback is provided to residents. Wood and Newborough [11] reported that energy savings can increase more than 10% through more specific information strategies for occupants. Ouyang and Hokao [12] reported that occupant behavior improvement allow for an average of 14% energy savings in residential buildings.

Compared to buildings' physical properties, occupant behavior is more difficult to quantify and assess. Recent studies [13–16] have focused on energy uses of appliances in household to measure occupant behavior since appliance usage reflects on occupant behavior patterns at varying weekdays and times. However, limitations still exist in these models which underline pieces of relevant factors. More rational and systematic models are needed to explain relationships of occupant behaviors, HVAC systems, appliances, and the built environment, and to improve energy prediction for residential buildings.

Recent studies in building have used feature selection and feature engineering to explain building energy consumption. Zhang et al. [17] proposed a feature engineering method using the residential building energy data from the Pecan Street Project in Texas. They analyzed and ranked 124 features using exploratory data analysis, random forest, and principal component analysis. Zhang and Wen [18] proposed a feature selection procedure for commercial buildings using a wrapper method but this model only addressed energy modeling in a single building. Although these studies tried feature selection and feature engineering approaches, they overlooked the characteristics of residential buildings and occupant behavior data. There is still a lack of framework to guide feature selection for residential buildings based on the relationships among energy related factors, and thus a lack of understanding of effective factors for residential building energy prediction. Also, they less focused on reducing the dimensionality of the factors and evaluating the functionality of the selected factors. Thus, the framework presented in this study helps select effective factors to predict energy consumption in residential buildings and evaluate the efficiency and functionality of the selected factors.

2.2. Factors related to energy use in residential buildings

The DOE's Energy Information Administration (EIA) produce and manage the RECS national database of energy consumption. The RECS energy use data are collected every three years since 1978 [14]. The RECS database includes annual household energy consumption data and associated 270 factors such as energy fuel type, building geometry information, household demographics, and appliance information [15]. EIA merges data from energy suppliers and residential units to estimate energy use for appliances, heating, cooling, and other end devices [19]. In particular, a subset of RECS factors reflect occupant behavior patterns, including geographic location, household equipment and appliances, family structure, income, and local electricity price [14]. Therefore, the RECS provides a comprehensive and exclusive set of factors related to energy use in residential buildings. The factors include energy consumption, occupant behavior, and building technologies, which are useful for energy consumption modeling.

The RECS provides a reliable source of data for researchers to study building energy efficiency and develop building energy models. Sanquist, Orr, Shui and Bittner [14] used the 2005 RECS data to perform lifestyle analysis of energy consumption in residential buildings with multivariate statistical techniques. They identified five lifestyle factors that were associated with behavior patterns: air conditioning, use of laundry usage, use of personal computer, climate zone, and use of TV. The lifestyle factors explain 40–54% of variance in energy consumption when considering household and market characteristics such as income, access to natural gas, and local electricity price. Diao, Sun, Chen and Chen [15] identified and classified occupant behavior with energy consumption outcomes. They extracted occupant features of five typical house types in New York State from the 2009 RECS data. The features include number of occupants, number of rooms, floor area and heated area and number of windows in heated area for the house types of single family (detached), single family (attached), apartment (2–4 units), apartment (5+ units), and mobile home. The information was applied to the behavior clusters from the ATUS by mapping the demographic information of the ATUS and the RECS. Aksanli, Akyurek and Rosing [16] proposed a residential energy modeling method based on human-activities to estimate the energy consumption in residential buildings. They extracted appliance-related parameters from the RECS including the types, numbers, and frequency of usage, and associated them with specific actions and activities. They grouped the activities based on the demographic information of the occupants, such as age, gender, employment status, and number of household members. They aimed to capture the use activities based on the probabilistic time-series nature which is depending on demographic variables and time variables (time of a day, day of a week etc.).

Nevertheless, these studies analyze only a part of the RECS's variables, for example, electricity appliances at home, building technologies, occupant behavior, or demographic information, and they did not provide a holistic view of all related features about energy prediction. This study includes all aspects of the RECS variables and implements a comprehensive and effective framework for residential energy consumption prediction.

2.3. Relationships among energy consumption, occupant, and building

The characteristics of both occupants and the building affect energy consumption. Previous studies defined the associations of building technology, resident behavior, and energy consumption [2], and explained the interaction relationship between building physics, weather condition, and occupant behavior [17]. Accordingly, the relationships among energy consumption, occupant and building were explained as described in Fig. 1 in the preliminary study [20]. In the study, energy use-related factors are categorized into two groups: occupant and building. Each group has two sub-groups. The two sub-groups for the occupant group are demographic (static) information

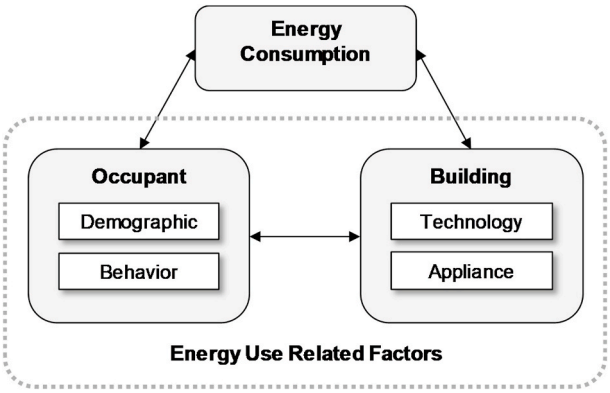

Fig. 1. Interactive relationships among energy consumption, occupant, and building.

such as age and gender, and the behavior (dynamic) information such as how often the occupant uses a certain appliance. The two sub-groups for the building group are technology (static or system-related) information such as building envelope, and appliance (dynamic and occupant behavior-related) information such as TV, refrigerator. The interactive relationship and categorization are appropriate to analyze the RECS data. The four sub-groups help to analyze the contextual meaning of the selected features and suggest strategies for applying the selected features to further studies or policy makings.

3. Framework development

Based on the literature about energy consumption, occupant, and building (Fig. 1), an analytic framework was developed to reduce dimensionality of predictors and assess extracted features. The core of the framework is machine learning, a data driven technology. Machine learning involves searching a large space of possible hypotheses to find one that best fits the observed data and any prior knowledge held by the learner. Machine learning is concerned with answering questions such as what is the best strategy for choosing a useful next training experience, what is the best way to reduce the learning task to one or more function approximation problems, and so on [21]. The framework allows to analyze occupant and building data and improve the efficiency of energy consumption modeling. Fig. 2 displays this analytic framework to select factors for energy prediction and assess their consequences for energy saving. Unlike conventional analytic processes that input all energy use-related factors into energy consumption prediction or select factors by experience and interest, the analytic framework integrates feature engineering and feature assessment processes to reduce the dimensionality of energy use-related factors. Four steps of the features engineering process are explained in the following sub-sections.

The analytic framework was demonstrated and validated via two application cases of the RECS datasets: (1) a national level dataset and (2) four regional level datasets to evaluate the efficiency of the selected features. Then, the features were assessed based on the four factor sub-groups: appliance, behavior, demographic, and technology. The application cases are described in Fig. 3.

3.1. Step 1: data preparation

Features represent the energy use-related factors in the dataset. To prepare for feature selection, all available features were examined in the dataset. The data preparation shall first exclude irrelevant features such as data collection related flags and then group the remaining energy use-related features into the four sub-groups (Fig. 1). Depending on the analysis and comparison, the whole dataset can be divided into multiple sub-datasets.

3.2. Step 2: Feature selection

The feature selection is a machine learning process that aims to find faster and more cost-effective predictors. Selected features are effective predictors that improve prediction performance and help researchers better understand the underlying process [22]. This process is also called attribute selection, variable selection, or variable subset selection [23]. This study employs the Correlation-based Feature Selection (CFS) approach with Greedy stepwise technique. CFS evaluates the worth of a subset of features by considering the predictive ability of each feature in the subset and the degree of redundancy between the features. Feature redundancy is a tendency that values of features are changing systematically with their category membership, and the evaluation formula of CFS puts more merits to feature subsets with a low level of redundancy [24,25]. Thus, CFS selects a subset of features that individually correlated well with the class or dependent variable, but have little intercorrelation. The correlation between two features A and B can be measures with the symmetric uncertainty as follows.

$$U(A,B) = 2\,\frac{H(A)+H(B)-H(A,B)}{H(A)+H(B)}$$

where H is the entropy function, which is based on the probability associated with each attribute value. H (A, B) is the joint entropy of A and B, which is calculated from the joint probabilities of all combinations of values of A and B [26]. Greedy stepwise performs a greedy search forward or backward through addition or deletion of any remaining features until the optimal performance is reached [25]. This research used Weka (3-8-0) for feature selection in the following application cases.

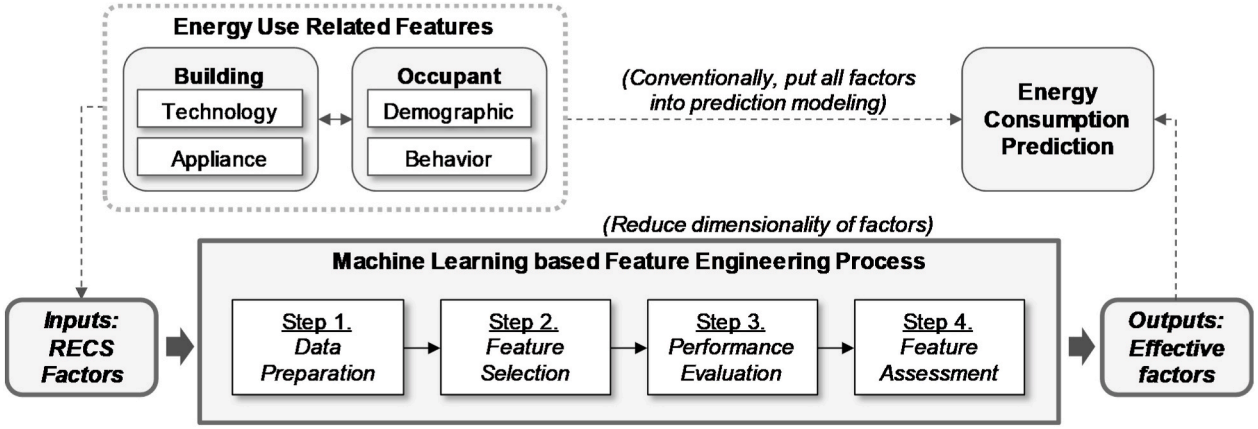

Fig. 2. Feature engineering-based analytic framework.

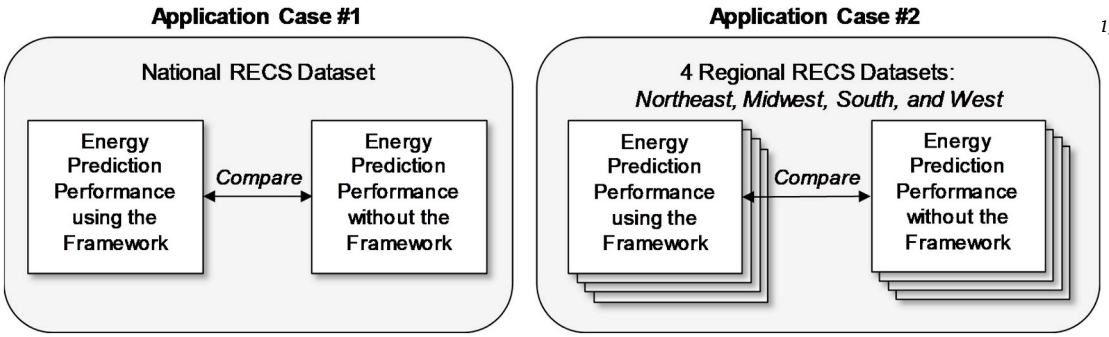

Fig. 3. Application cases of the framework.

3.3. Step 3: performance evaluation

The performance evaluation is a computational process to examine the prediction power of selected features. Then, prediction power is compared between the whole features (270) and selected features using value engineering principle (Value = Function/Cost) [27]. The function refers to the prediction power measured by correlation coefficient and the cost refers to the number of feature reduction.

Several studies suggested the Support Vector Machine (SVM) algorithm for energy consumption prediction assessment [28–32]. Also, our preliminary study found SVM generated the better performance on the data than multiple other popular algorithms such as Linear Regression, Random Forest, M5P Trees, and M5 Rules. In the study, those five algorithms were compared to examine the performance of the features using the whole nation data. SVM showed the best performance with the highest correlation coefficient and the lowest root mean squared error (RMSE) values in both cases [33]. The finding is consistent with many studies that showed SVM is suitable to predict building energy consumption. Ahmad et al. [28] reviewed several machine learning algorithms for building electrical energy consumption forecasting and they explained that SVM was one of the widely used and accurate methods in this field. Dong et al. [29] applied SVM to analyze building energy consumption in tropical region. Wei et al. [30], Deb et al. [31], and Edwards et al. [32] compared machine learning methods for building energy prediction and SVM was explained as one of the suitable methods.

SVM was originally developed as a binary classifier that define the hyperplane that divide a dataset into two separate classes with the gap between them as wide as possible [34]. This maximum-margin hyperplane concept applies to classification, but SVM have been developed for numeric prediction. Sequential minimal optimization (SMO) regression implements the SVM to produce a model that can be expressed with support vectors and can be applied to nonlinear datasets using kernel functions. This support vector regression function can be explained as follows [26].

$$x = b + \sum_{i \text{ is support vector}} \alpha_i \, \mathrm{a(i)} \bullet \mathrm{a}$$

where x is the outcome, b and α_i are parameters that have to be determined by the learning algorithm. The vector a is a test instance, and the vector a(i) are the support vectors that are selected members of the training set. The term a(i) • a indicates the dot product of the test instance with one of the support vectors, which is explained as follows.

$$\mathrm{a(i)} \bullet \mathrm{a} = \sum_j a(i)_j a_j$$

To evaluate the performance of the selected features Radial Basis Function kernel with C value 1 and gamma value 0.01 were used. This research used Python 2.7 (Scikit-learn package) for the performance evaluation in the following application cases.

3.4. Step 4: Feature assessment

The feature assessment evaluates the implications from selected features based on the four sub-groups explained in Fig. 1. Descriptive analysis and qualitative methods are used to assess whether selected features make sense to inform energy use, e.g., the frequency of features per sub-group. The characteristics of the selected features of a sub-group are compared with the ones of the other sub-groups and further strategies regarding efficient energy usage in residential buildings can be assessed.

In this study, three types of error analysis are recommended to improve the process at the feature level, algorithm level, and data level. During the feature selection, each of the selected features and their correlation values with various feature combinations can be examined to select the most effective features and reduce the dimensionality. During the performance evaluation, different values of hyperparameters can be tested when running SVM and the suitable values should be selected. During the case studies, prediction performance on different scales of data can be analyzed, e.g., at the national level and regional level, to avoid overfitting and underfitting.

4. Application case 1: national RECS dataset

4.1. Data preparation

This case used the 2015 RECS dataset that included 736 initial features and covered 5686 U S. household instances. The dataset was the recent round of data collection, i.e. the 14th iteration of the RECS program (EIA 2018). This study used 272 features for data analysis after removing imputation flags, replicate weights, and irrelevant features. Within the 272, the dependent variable was KWH (Electricity usage in kWh) and the rest were used as predictor variables. REGIONC (Census region) was reserved for dataset preparation, resulting in a total of 270 predictors. Based on the characteristics, the features were grouped by appliance, behavior, demographic, and technology (Table 1). (1)

Table 1
Categories of Energy use-related Features.

Category	Feature Examples	#Feature
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
Demographic	Occupant/family characteristics, Who pays bill, Receive/participate in home energy assistance program	41
Technology	Building envelope, HVAC, Water heater, Fuel type for Tech, Thermostat, Light controller, Sensor, Smart meter install, Building audit, Pool, Hot tub	116
Total		270

Appliance features are the appliances that are directly related to energy consumption and utility bills. (2) Behavior features are occupants' appliance and HVAC system usage patterns or settings that are related to energy consumption. (3) Demographic features are occupants' demographic characteristics. (4) Technology features include building technologies, HVAC systems, fuel types, and climate. In the first application case, the whole national RECS dataset was used.

4.2. Feature selection

Table 2 lists the 36 selected features from the 270 original RECS features in the nation dataset. They were categorized according to the characteristics: appliance, behavior, demographic, and technology. Among the selected features, there were 9 appliance features, 3 behavior features, 5 demographic features, and 19 technology features. CFS is a correlation-based feature selection method and higher absolute correlation between the feature and the dependent variable means stronger relation between them. Table 3 lists the top 10 features with higher correlation values among the selected features, which are the main features to contribute to the prediction power of the model. Table 4 summarizes all the selected features from nation and regions. Two technology features, the type of housing unit and the fuel oil used for water heating, were only selected in the nation dataset.

Table 2
Effective features selected from all features in national dataset.

Category	Feature Code	Description
Appliance	COMBODVR	Number of cable or satellite boxes with DVR
	ICE	Through-the-door ice on most-used refrigerator
	NUMCFAN	Number of ceiling fans used
	NUMFREEZ	Number of separate freezers used
	NUMFRIG	Number of refrigerators used
	OVEN	Number of separate ovens
	SIZFREEZ	Size of most-used freezer
	SIZRFRI1	Size of most-used refrigerator
	TVCOLOR	Number of televisions used
Behavior	DRYRUSE	Frequency clothes dryer used
	MONPOOL	Months swimming pool used in the last year
	TVONWE1	Most-used TV usage on weekends
Demographic	NHSLDMEM	Number of household members
	NOACBROKE	Unable to use cooling equipment in the last year because equipment was broken and could not afford repair or replacement
	NUMADULT	Number of household members age 18 or older
	NUMCHILD	Number of household members age 17 or younger
	PERIODNG	Number of days covered by Energy Supplier Survey natural gas billing data and used to calculate annual consumption and expenditures
Technology	AIRCOND	Air conditioning equipment used
	CENACHP	Central air conditioner is a heat pump
	CLIMATE_REGION_PUB	Building America Climate Zone
	COOLTYPE	Type of air conditioning equipment used
	ELFOOD	Electricity used for cooking
	ELWARM	Electricity used for space heating
	ELWATER	Electricity used for water heating
	FOWATER	Fuel oil used for water heating
	FUELH2O	Fuel used by main water heater
	FUELH2O2	Fuel used by secondary water heater
	FUELHEAT	Main space heating fuel
	FUELTUB	Fuel used for heating hot tub
	NCOMBATH	Number of full bathrooms
	POOL	Heated swimming pool
	TOTROOMS	Total number of rooms in the housing unit, excluding bathrooms
	TYPEHUQ	Type of housing unit
	UATYP10	Census 2010 Urban Type
	UGASHERE	Natural gas available in neighborhood
	UGWATER	Natural gas used for water heating

Table 3
List of top ten features with high correlation values in national dataset.

Feature	Correlation	Description
TOTROOMS	0.4117	Total number of rooms in the housing unit, excluding bathrooms
DRYRUSE	0.3962	Frequency clothes dryer used
NCOMBATH	0.3862	Number of full bathrooms
NUMCFAN	0.3708	Number of ceiling fans used
TYPEHUQ	−0.3687	Type of housing unit
TVCOLOR	0.3664	Number of televisions used
PERIODNG	−0.3532	Number of days covered by Energy Supplier Survey natural gas billing data and used to calculate annual consumption and expenditures
NUMFRIG	0.3424	Number of refrigerators used
ELWATER	0.3244	Electricity used for water heating
FUELH2O	0.3220	Fuel used by main water heater

4.3. Performance evaluation

Table 5 shows the comparison of prediction performance of nation dataset pre and post feature selection. The results show that the 36 features selected from the nation dataset provide 0.77 correlation coefficient, indicating that they use only 13% of the total features to reach 97% of the prediction performance with all features. When comparing pre-value and post-value, the value improvement is 725% in this dataset.

4.4. Feature assessment

In the nation dataset, the selected features cover all the four subgroups. Among the appliance features, refrigerators, freezers, and TVs, and ceiling fans are good predictors of energy consumption in residential buildings. The selected appliance features include the number of refrigerators used, the size of most-used refrigerator, if the most-used refrigerator has the through-the-door ice, the number of separate freezers used, the size of most-used freezer, the number of separate ovens, the number of televisions used, the number of cable or satellite boxes with DVR, and the number of ceiling fans used. Among the occupant behavior and demographic features, family composition, financial status, use patterns of swimming pool (if they have), clothes dryer, and TV are good predictors of energy consumption. The selected behavior and demographic features include the number of household members, the number of household members age 18 or older, the number of household members age 17 or younger, 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, the months swimming pool used in the last year, the frequency clothes dryer used, the most-used TV usage on weekends. Among the technology features, the house type and size and the fuel types for heating and cooling are key predictors of energy consumption. The selected technology features include the type of housing unit, the number of rooms and of full bathrooms, the fuel type for space heating, water heater, and heating hot tub, the type of air conditioner type, if electricity is used for space heating, water heating and cooking, urban type of the housing unit location, and Building America Climate Zone.

5. Application case 2: regional RECS datasets

5.1. Data preparation

In the second application case, the whole national dataset was divided into four sub-datasets, Northeast (R1), Midwest (R2), South (R3), and West (R4) to identify regional similarities and discrepancies.

Table 4
List of selected features in four regions.

Count	Code	Description	R0	R1	R2	R3	R4
4	NUMFRIG	Number of refrigerators used	X	X	X	X	X
	SIZRFRI1	Size of most-used refrigerator	X	X	X	X	X
	FUELTUB	Fuel used for heating hot tub	X	X	X	X	X
	FUELH2O2	Fuel used by secondary water heater	X	X	X	X	X
	ELWARM	Electricity used for space heating	X	X	X	X	X
	ELFOOD	Electricity used for cooking	X	X	X	X	X
3	COMBODVR	Number of cable or satellite boxes with DVR	X	X	X	X	
	DESKTOP	Number of desktop computers		X	X	X	
	MONPOOL	Months swimming pool used in the last year	X	X	X	X	
	NHSLDMEM	Number of household members	X	X	X	X	
	NUMADULT	Number of household members age 18 or older	X	X	X	X	
	NCOMBATH	Number of full bathrooms	X	X	X	X	
	COOLTYPE	Type of air conditioning equipment used	X	X	X		X
	ELWATER	Electricity used for water heating	X	X	X		X
	MICRO	Microwave oven used		X		X	X
	SIZFREEZ	Size of most-used freezer	X	X		X	X
	NUMFREEZ	Number of separate freezers used	X		X	X	X
	TVCOLOR	Number of televisions used	X		X	X	X
	POOL	Heated swimming pool	X		X	X	X
	RECBATH	Hot tub			X	X	X
	UGASHERE	Natural gas available in neighborhood	X		X	X	X
2	EQUIPAUX	Secondary space heating equipment used		X	X		
	SIZRFRI2	Size of second most-used refrigerator		X		X	
	DRYRUSE	Frequency clothes dryer used	X	X		X	
	NUMCHILD	Number of household members age 17 or younger	X	X		X	
	BEDROOMS	Number of bedrooms		X		X	
	USENOTMOIST	Number of months dehumidifier used in last year		X			X
	FUELHEAT	Main space heating fuel	X	X			X
	CENACHP	Central air conditioner is a heat pump	X	X			X
	FUELH2O	Fuel used by main water heater	X	X			X
	USENG	Natural gas used		X			X
	OVEN	Number of separate ovens	X		X	X	
	PERIODNG	Number of days covered by Energy Supplier Survey natural gas billing data and used to calculate annual consumption and expenditures	X		X	X	
	NGPAY	Who pays for natural gas			X		X
	UGWATER	Natural gas used for water heating	X		X		X
	AIRCOND	Air conditioning equipment used	X			X	X
	LGTOUTNUM	Number of light bulbs installed outside the home				X	X
1	COFFEE	Coffee maker used		X			
	ICE	Through-the-door ice on most-used refrigerator	X	X			
	NUMCFAN	Number of ceiling fans used	X	X			
	NUMTABLET	Number of tablet computers		X			
	LGTIN4	Number of inside light bulbs turned on at least 4 h a day		X			
	HHSEX	Respondent sex		X			
	DIVISION	Census Division		X			
	OTHROOMS	Number of other rooms		X			
	YEARMADERANGE	Range when housing unit was built		X			
	DRYRFUEL	Fuel used by clothes dryer			X		
	UPRTFRZR	Door arrangement of most-used freezer			X		
	TEMPGONE	Winter temperature when no one is home during the day			X		
	NOACBROKE	Unable to use cooling equipment in the last year because equipment was broken and could not afford repair or replacement	X		X		
	CLIMATE_REGION_PUB	Building America Climate Zone	X		X		
	METROMICRO	Housing unit in Census Metropolitan Statistical Area or Micropolitan Statistical Area			X		
	TOTROOMS	Total number of rooms in the housing unit, excluding bathrooms	X		X		
	UATYP10	Census 2010 Urban Type	X		X		
	APPOTHER	Other small appliance used				X	
	DRYER	Have clothes dryer in home				X	
	NUMSMPHONE	Number of smart phones				X	
	STOVE	Number of separate cooktops				X	
	TOAST	Toaster used				X	
	TVONWE1	Most-used TV usage on weekends	X			X	
	ENERGYASST14	Received home energy assistance in 2014				X	
	FUELPOOL	Fuel used for heating swimming pool				X	
	LPOTHER	Propane used, other than for space heating, water heating, or cooking				X	
	MORETHAN1H2O	More than one water heater				X	
	NHAFBATH	Number of half bathrooms				X	
	SOLWATER	Solar used for water heating				X	
	SWIMPOOL	Swimming pool				X	
	LGTINNUM	Number of light bulbs installed inside the home					X
	LOCRFRI2	Location of second most-used refrigerator					X
	NOTMOIST	Dehumidifier used					X
	SCALEG	Frequency of keeping home at unhealthy temperature					X

(*continued on next page*)

Table 4 (*continued*)

Count	Code	Description	R0	R1	R2	R3	R4
	TVONWD1	Most-used TV usage on weekdays					X
	WASHLOAD	Frequency clothes washer used					X
	SDESCENT	Respondent is Hispanic or Latino					X
	BACKUP	Back-up generator					X
	WALLTYPE	Major outside wall material					X

Table 5
Performance comparison of selected features by region.

Dataset	Correlation Coefficient (Function)			#Features (Cost)			Value (Function/Cost)		
	Pre	Post	Post/Pre	Pre	Post	Post/Pre	Pre	Post	Post/Pre
Nation	0.7975	0.7704	97%	270	36	13%	0.002954	0.021400	725%
Region									
Northeast	0.7624	0.7765	102%	270	35	13%	0.002824	0.022186	786%
Midwest	0.7687	0.7536	98%	270	32	12%	0.002847	0.023550	827%
South	0.7305	0.7243	99%	270	40	15%	0.002706	0.018108	669%
West	0.7875	0.7650	97%	270	33	12%	0.002917	0.023182	795%

The regions in RECS data are aligned with the U.S. Census Bureau's definition of regions (Table 6).

5.2. Feature selection

Table 4 summarizes the 76 selected features from regions (R1, R2, R3, and R4). The selected features were sorted by the count of the overlapped feature across the four regions. The results from the whole nation dataset (R0) were also listed as a baseline for comparison. Results showed 6 overlapped features from all the four regions, indicating the cross-region similarity. These features include the energy usage for space heating and cooking, fuel type for secondary water heater and hot tub, and the number and size of refrigerator are good predictors for four regions. These features are consistent with the national dataset (R0).

5.3. Performance evaluation

Table 5 lists the comparison of prediction performance by region pre and post feature selection. The results show that the 32 selected features from the Midwest dataset have the highest improvement of an 827% value, followed by 795% improvement for West, 786% improvement for Northeast, and 669% improvement from South. Most of the improvements are higher than the improvement using the whole data (725%), indicating that the narrowed and region-specific feature selection may generate a higher prediction efficiency. The 35 features selected from the Northeast dataset provide 0.78 correlation coefficient, indicating that they use only 13% of the total features to reach 102% of the prediction performance with all features. Similarly, 32 selected features from Midwest (correlation coefficient = 0.75) use 12% of the total features to reach 98% of the prediction performance; the 40 selected features from South (correlation coefficient = 0.72) use 15% of the total features to reach 99% of the prediction performance; and the 33 selected 33 features from West (correlation coefficient = 0.76) use 12% of the total features to reach 97% of the prediction performance.

Table 6
U.S regions, divisions and states.

Region	Code	#Instances	Division	State
Northeast	R1	794	New England	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont
			Mid-Atlantic	New Jersey, New York, Pennsylvania
Midwest	R2	1327	East North Central	Illinois, Indiana, Michigan, Ohio, Wisconsin
			West North Central	Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota
South	R3	2010	South Atlantic	Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, District of Columbia, West Virginia
			East South Central	Alabama, Kentucky, Mississippi, Tennessee
			West South Central	Arkansas, Louisiana, Oklahoma, Texas
West	R4	1555	Mountain	Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming
			Pacific	Alaska, California, Hawaii, Oregon, Washington

5.4. Feature assessment

The selected features reveal unique energy use patterns in each region. (1) In the Northeast, appliances (e.g., coffee maker, tablet computers, refrigerator through-the-door ice), demographics (e.g., gender), building technology (e.g., years), and behavior (e.g., interior lighting usage) are highly correlated with total energy consumption. Compared to other three regions where heated tub or swimming pool are good predictors for energy consumption, lighting usage, small appliance usage show a higher prediction power due to relatively longer night hours and diverse population and lifestyle. (2) In the Midwest, appliances (e.g., fuel type of clothes dryer and freezer door type), behavior (e.g., winter temperature setting when unoccupied), demographics (the respondent's unaffordable financial status), building technology, especially environmental features (e.g., climate zone by Building America, if the housing is located in urban or suburban, metropolitan or micropolitan area) are highly correlated with total energy consumption. Compared to other three regions where microwave, size of most used freezer are good predictors for energy consumption, the location of the housing unit regarding climate and urbanicity show a higher prediction power. (3) In the South, appliances (e.g., clothes dryer, smart phone, separate cooktops, toaster, and other small appliances), behavior (e.g., TV usage on weekends), demographics (e.g., home energy assistance), and building technology (e.g., fuel type for pool, propane use, additional water heater, solar energy use) are highly correlated with total energy consumption. Compared to other three regions where air conditioning type, electricity use for water heating are good predictors for energy consumption, unique energy types, extra use of water heating and swimming pool use show a higher prediction power. (4) In the West, appliances (e.g., number of light bulbs inside the home, additional

refrigerator, dehumidifier), behavior (e.g., unhealthy temperature setting at home, TV usage on weekdays, frequency of clothes washer use), demographics (e.g., if the respondent is Hispanic/Latino or not), and building technologies (e.g., back-up generator, exterior wall type) are highly correlated with total energy consumption. Compared to other three regions where cable or satellite boxes, desktop, swimming pool use, number of household members and grown-ups, and number of bathrooms are good predictors for energy consumption, building materials, respondent's race/ethnicity, lighting, and temperature settings show a higher prediction power.

6. Discussion

The selected features improve the modeling efficiency for residential buildings when considering variables. Fig. 4 exhibits that the selected features achieved more than 97% of the prediction performance while using less than 15% of total features. It demonstrates that the analytic feature engineering framework is efficient to predict energy consumption for residential buildings. Also, the selected features resulting from the feature engineering framework improve the resource efficiency for energy modeling and computation and highlight unique patterns of energy use for American households.

The selected features have implications for better understanding the energy use patterns between regions. (1) In Northeast and Midwest, secondary space heating equipment usage is a predictor, suggesting that the two regions are colder where secondary space heating equipment is necessary and useful for energy prediction. (2) In Northeast and South, the number of children in households, the frequency of clothes dryer usage, number of bedrooms, and size of the secondary refrigerator (if they use) are predictors, suggesting that the two east coastal regions have a bigger family size and need extra room, dryer use, and refrigerator. (3) In Northeast and West, dehumidifier usage, fuel type for space and water heating, and air conditioner type are predictors, suggesting that the two regions contain diverse types of building across a long range of years that their equipment varies. (4) In Midwest and South, the number of separate ovens and the number of days covered by Energy Supplier Survey natural gas billing data are predictors, suggesting that cocking at home is important for families in these two regions. (5) In Midwest and West, who pays for natural gas and natural gas usage for water heating are predictors, suggesting that less natural gas as an alternate energy source is available for homes in the two west coastal regions. (6) In South and West, air conditioning use and the number of light bulbs installed outside the home are predictors, suggesting that the two southern regions have a higher temperature and a longer time for night activities.

7. Conclusion

In this study, an analytic machine learning-based feature engineering framework has been developed to select and assess factors to improve energy prediction efficiency by reducing the dimensionality of the factors. Based on the value engineering principle, the two implications cases show that the selected features achieve considerable improvement in efficiency to predict energy consumption in residential buildings.

This study contributes to residential building energy modeling and evaluation. (1) The presented analytic framework adds values on residential energy prediction process by improving the efficiency. The process combined feature selection and energy prediction algorithms together and applied them to the domain of residential energy data analysis. The predictor reduction helps to enable more economic and feasible energy data collection. In the energy management system (EMS) market, an enormous number of sensors are installed, or numerous questions are included in the survey to collect energy consumption related data. The results from this research suggest the most efficient factors to be included in those sensor installations or surveys, and reduce the required number of sensors or questions. (2) The results provide a list of efficient features for predicting energy consumption in residential buildings. The features effectively explain energy use patterns and their relationships which help residential designer, constructors, and occupants understand energy, behaviors, and the built environment. The resultant energy use patterns inform regional similarities, differences, and distinctive characteristics. The specific regional characteristics help develop optimal energy policies strategies to promote energy savings. Scholars and construction managers can use this framework to produce suggestions for region-specific decision making, for example, in regional retrofit projects. In addition, this framework can be extended to other energy survey or sensor-measured datasets and research domains.

A limitation of this study is related to RECS dataset. The RECS uses the four census regions rather than more detailed regional features at state or county level, or ASHRAE climate zones. In the RECS, behavior- and time-related activities are less detailed than appliances, building technology, demographic information, and energy use. The RECS dataset lacks HVAC use behavior such as thermostat set-point temperature for heating or cooling seasons. The dataset limitation might lose important energy-usage predictive features. In the future, state or county level geographical data and behavior-related features can be further examined in detail by using datasets from smart meters and home sensors.

Author contributions

Yunjeong Mo: Conceptualization, Methodology, Software,

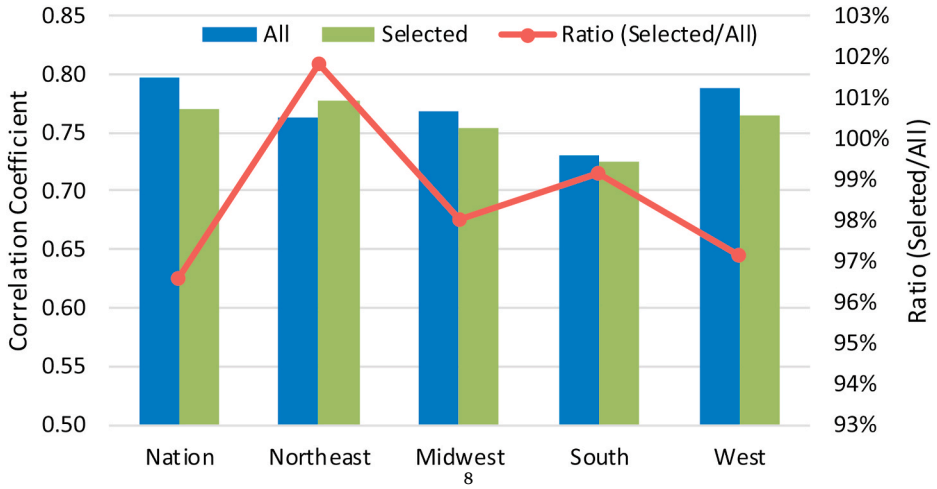

Fig. 4. Performance comparison of selected features by region.

Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. Dong Zhao: Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported by the National Science Foundation (NSF) through Grant #2046374. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the researchers and do not necessarily reflect the views of NSF.

References

[1] U.S.Doe, How Much Energy Is Consumed in U.S. Residential and Commercial Buildings?, 2017. https://www.eia.gov/tools/faqs/faq.php?id=86&t=1.
[2] D. Zhao, A.P. McCoy, J. Du, P. Agee, Y. Lu, Interaction effects of building technology and resident behavior on energy consumption in residential buildings, Energy Build. 134 (2017) 223–233.
[3] D. Zhao, A.P. McCoy, P. Agee, Y. Mo, G. Reichard, F. Paige, Time effects of green buildings on energy use for low-income households: a longitudinal study in the United States, Sustainable cities and society 40 (2018) 559–568.
[4] A.P. McCoy, D. Zhao, T. Ladipo, P. Agee, Y. Mo, Comparison of green home energy performance between simulation and observation: a case of Virginia, United States, Journal of Green Building 13 (3) (2018) 70–88.
[5] R.K. Jain, K.M. Smith, P.J. Culligan, 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, Appl. Energy 123 (2014) 168–178.
[6] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renew. Sustain. Energy Rev. 81 (2018) 1192–1205.
[7] M. Schweiker, A. Wagner, The effect of occupancy on perceived control, neutral temperature, and behavioral patterns, Energy Build. 117 (2016) 246–259.
[8] J. Chen, X. Gao, Y. Hu, Z. Zeng, Y. Liu, A meta-model-based optimization approach for fast and reliable calibration of building energy models, Energy 188 (2019) 116046.
[9] S. Chen, W. Yang, H. Yoshino, M.D. Levine, K. Newhouse, A. Hinge, Definition of occupant behavior in residential buildings and its application to behavior analysis in case studies, Energy Build. 104 (2015) 1–13.
[10] S. Darby, The effectiveness of feedback on energy consumption, A Review for DEFRA of the Literature on Metering, Billing and direct Displays 486 (2006), 2006.
[11] G. Wood, M. Newborough, Dynamic energy-consumption indicators for domestic appliances: environment, behaviour and design, Energy Build. 35 (8) (2003) 821–841.
[12] J. Ouyang, K. Hokao, Energy-saving potential by improving occupants' behavior in urban residential sector in Hangzhou City, China, Energy Build. 41 (7) (2009) 711–720.
[13] O.G. Santin, L. Itard, H. Visscher, The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock, Energy Build. 41 (11) (2009) 1223–1232.
[14] T.F. Sanquist, H. Orr, B. Shui, A.C. Bittner, Lifestyle factors in US residential electricity consumption, Energy Pol. 42 (2012) 354–364.
[15] L. Diao, Y. Sun, Z. Chen, J. Chen, Modeling Energy Consumption in Residential Buildings: A Bottom-Up Analysis Based on Occupant Behavior Pattern Clustering and Stochastic Simulation, Energy and Buildings, 2017.
[16] B. Aksanli, A.S. Akyurek, T.S. Rosing, User Behavior Modeling for Estimating Residential Energy Consumption, Springer, 2016, pp. 348–361. Smart City 360°.
[17] C. Zhang, L. Cao, A. Romagnoli, On the feature engineering of building energy data mining, Sustainable cities and society 39 (2018) 508–518.
[18] L. Zhang, J. Wen, A systematic feature selection procedure for short-term data-driven building energy forecasting model development, Energy Build. 183 (2019) 428–442.
[19] Eia, Residnetial Energy Consumption Survey (RECS), 2018. https://www.eia.gov/consumption/residential/about.php.
[20] Y. Mo, Occupant Behavior Prediction Model Based on Energy Consumption Using Machine Learning Approaches, Michigan State University, East Lansing, MI, 2018.
[21] T.M. Mitchell, Machine learning and data mining, Commun. ACM 42 (11) (1999) 30–36.
[22] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (Mar) (2003) 1157–1182.
[23] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, Springer, 2013.
[24] J.H. Gennari, P. Langley, D. Fisher, Models of incremental concept formation, Artif. Intell. 40 (1–3) (1989) 11–61.
[25] M.A. Hall, Correlation-based Feature Subset Selection for Machine Learning, Computer Science The University of Waikato, Hamilton, New Zealand, 1999.
[26] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.
[27] R. Cooper, Target Costing and Value Engineering, Routledge, 2017.
[28] A. Ahmad, M. Hassan, M. Abdullah, H. Rahman, F. Hussin, H. Abdullah, R. Saidur, A review on applications of ANN and SVM for building electrical energy consumption forecasting, Renew. Sustain. Energy Rev. 33 (2014) 102–109.
[29] B. Dong, C. Cao, S.E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy Build. 37 (5) (2005) 545–553.
[30] L. Wei, W. Tian, E.A. Silva, R. Choudhary, Q. Meng, S. Yang, Comparative study on machine learning for urban building energy analysis, Procedia Engineering 121 (2015) 285–292.
[31] C. Deb, F. Zhang, J. Yang, S.E. Lee, K.W. Shah, A review on time series forecasting techniques for building energy consumption, Renew. Sustain. Energy Rev. 74 (2017) 902–924.
[32] R.E. Edwards, J. New, L.E. Parker, Predicting future hourly residential electrical consumption: a machine learning case study, Energy Build. 49 (2012) 591–603.
[33] Y. Mo, D. Zhao, M. Syal, Effective Features to Predict Residential Energy Consumption Using Machine Learning, Computing in Civil Engineering 2019: Smart Cities, Sustainability, and Resilience, American Society of Civil Engineers Reston, VA2019, pp. 284-291..
[34] H. Sun, H.V. Burton, H. Huang, Machine learning applications for building structural design and performance assessment: state-of-the-art review, Journal of Building Engineering (2020) 101816.