KIEAE Journal, Vol. 20, No. 6, Dec. 2020, pp.21-30.

KIEAE Journal

Korea Institute of Ecological Architecture and Environment

106

Identification of Primary Factors Influencing Energy Consumption Patterns of Commercial and Residential Buildings

Gigih Rahmandhani Setyantho* · Seongju Chang**

* M.S. Student, Dept. Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology, South Korea (setyanthogr@kaist.ac.kr)
** Corresponding author, Invited Professor, Dept. Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology, South Korea (schang@kaist.ac.kr)

ABSTRACT

Purpose: Commercial and residential buildings are primary building types covering the majority of our built environment. Though, energy consumption pattern identification of those building types based on real-world data analysis has not been aggressively explored. Identification of primary factors influencing energy consumption patterns of commercial and residential buildings is the goal of this study. **Method:** CBECS and RECS data sets in the United States were used for commercial and residential building-energy-consumption pattern analysis. Multi-linear and seven machine learning algorithms are utilized to analyze building characteristics and end-use energy consumption patterns, e.g., cooling, heating, and water-heating. The SHAP value is utilized to describe influential factors in each energy consumption analysis model. **Result:** Ensemble algorithm yielded the lowest error rates compared with other algorithms. The calculated error rates also showed a lower level than the precedent studies performed on the CBECS and the RECS. Commercial building's cooling and heating energy consumption is more likely influenced by occupancy, while residential building's energy consumption is affected by equipment and climatic conditions. In the meantime, water-heating energy consumption shows noticeable dependency over the occupancy and climatic conditions for commercial and residential buildings. As a critical passive design element in the building, window has a more significant influence than overall insulation or roof finishing in identified residential buildings' energy consumption patterns.

KEYWORD

Building energy consumption
Heating
Cooling
Water heating
SHAP values

ACCEPTANCE INFO

Received Nov. 13, 2020
Final revision received Dec. 2, 2020
Accepted Dec. 7, 2020

1. Introduction

1.1. Research background

Building energy performance is typically affected by outdoor conditions, building characteristics, occupants, and operational specifics. Due to this complexity, predicting building energy consumption is a difficult task to deal with[1-3]. More than a decade ago, Perez-Lombard et al.(2008) argued that available building energy information was insufficient and not proportional to its importance[4].

Amasyali and El-Gohary(2018) attempted a building-energy studies review. It was found that 47%, 31%, 20% and 2% of the examined studies were related to total, cooling, heating, and lighting energy prediction, respectively. On the other hand, Do and Cetin(2018) suggested to work on the residential-building energy-consumption prediction since residential buildings have received less attention. Less attention happened because of the limited availability of residential building data sets[3].

Furthermore, identifying key variables to construct a prediction model also becomes an essential task. Choi et al. (2017) performed a quantitative analysis based on energy consumption big data sets for residential buildings in Seoul, Korea[5]. Analysis performed with Pearson's correlation techniques showed that the correlation between electricity consumption and trading prices is higher than residential area and trading prices[5]. Kim et al.(2017) investigated the correlation between buildings and urban planning factors to identify significant variables on family-income groups[6].

In the meantime, the United States Energy Information Administration (EIA) has been administering the Commercial Building Energy Consumption Survey (CBECS) and the Residential Energy Consumption Survey (RECS) for nationally representative sample buildings. These surveys collect data sets related to building characteristics, energy sources, and energy usages[7][8].

1.2. Precedent studies

Previous studies based on CBECS and RECS (references [9]-[15]) are summarized in Table 1. Robinson et al.(2015) presented a comparison between using the least variables or common variables (total area, number of floors, HDD, CDD, and principal building activity) and extended variables. Results show

pISSN 2288-968X, eISSN 2288-9698
http://dx.doi.org/10.12813/kieae.2020.20.6.021

Table 1. Comparison among precedent studies carried out with CBECS and RECS datasets

Author	Modeling technique	Variable evaluation	Target energy usage type
Yalcintas, et. al.(2007) [9]	• Multi-linear regression (MLR) and artificial neural network (ANN)	MLR	• Energy use intensity (EUI) - Electricity
Kaskhendikar, et. al(2010) [10]	• Linear regression and random forest (RF)	Regression trees	• Total EUI
Robinson, et. al.(2015) [11]	• XGBoost, linear regression, ridge regression, support vector regressor (SVR), bagging, RF, extra trees, linear SVR, AdaBoost, k-nearest neighbors (kNN) regression, multi-layer perceptron (MLP) regression, ElasticNet, lasso regression	Reduction in Gini impurity	• EUI – Major fuel
Lokhandwala, et. al.(2018) [12]	• Mean, generalized linear model (GLAM), multivariate adaptive regression splines (MARS), SVR, RF, neural networks	% Increase in mean square error (MSE)	• Cooling EUI – Electricity
Deng, et. al.(2018) [13]	• Linear regression, lasso regression, SVR, ANN, gradient boosting, RF	% Increase in MSE	• Total EUI, HVAC EUI, Plug loads EUI
Troup, et. al.(2019) [14]	• Multi-linear additive regression	-	• Total EUI, Heating EUI, Cooling EUI, Lighting EUI, Ventilation EUI
Kim, et. al.(2018) [15]	• Regression	Standard regression coefficient (SRCs)	• Total EUI

that the extended variables model's MAE is lower than common features model one[11]. Noh et al.(2017) proposed a data cube model combined with association rule mining for analyzing the CBECS 2012 data sets. They harnessed C5.0 decision tree algorithm as an effective and powerful classification method to represent overall building characteristics into ten variables representing the most influential factors for identifying building energy consumption patterns[17]. Different from other studies, Troup et al.(2019) and Kim et al.(2018) studied building energy consumption prediction concerning windows[14][15].

As for identifying influential variables contributing to a certain outcome, regression coefficient calculation used to be a commonly used method. The percentage of mean square error (MSE) has also been used to detect critical variables. But, both of these methods only show a single numerical value to represent the relation between input and the output variables[12][13]. In the meantime, the latest approach in the domain of interpretable machine learning is using SHAP value as a method to describe the accuracy of model prediction based on Shapley value and local surrogate model[18].

1.3. Research objectives

Few studies have addressed the prediction modeling for residential building energy consumption and yet identifying influential variables is crucial for better understanding what variables determine observable energy consumption profiles in residential and commercial buildings. In this perspective, we performed analysis on commercial and residential building energy consumption patterns to find influential factors embedded in the CBECS and the RECS data sets utilizing SHAP value calculation. Through analyzing multiple years of survey data sets, we also tried to track down the differences between the 1990s and the 2010s in influential factors affecting building energy consumption. This research evaluated commercial and residential building energy consumption data sets with the following objectives:

a. How to predict energy consumption (cooling, heating, and water-heating) with higher accuracy based on selected features of different building types?

b. What variables could explain the differences between commercial and residential building energy consumption patterns based on each SHAP values?

c. What are the influential building energy-consumption factors of the 1990s compared to those of the 2010s?

To the best of our knowledge, this research is the first attempt to analyze the RECS data sets for residential building energy consumption prediction; compared influential variables for energy consumption predictions that embedded in the CBECS and the RECS data sets based on SHAP value as the variable significance indicator in such predictions.

2. Materials and methods

This research's overall framework is shown in Fig. 1. Each data set was classified into three different regression models based on the end-use energy consumption types (heating, cooling, and water-heating). Each model went through multi-linear regression to find significant variables/features based on each variable's p-value. Significant variables become candidate independent variables for the energy consumption prediction modeling based on selected machine learning algorithms. After this process, the variable importances were evaluated based on

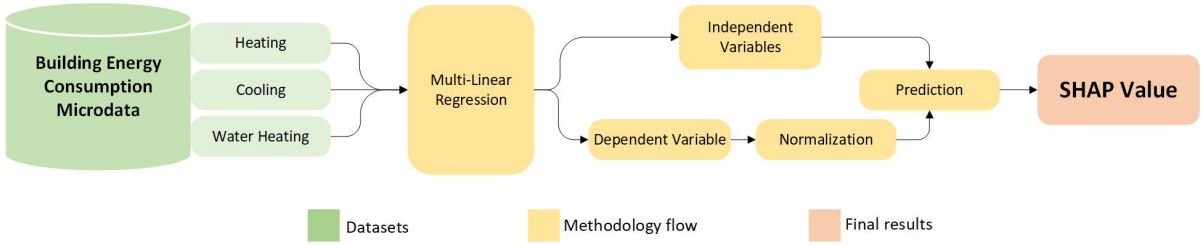


Fig. 1. Overview of the adopted research procedure

the SHAP value calculation. Among those variables that were turned out to be important in each data set, the comparisons were performed over the 1990s and the 2010s to find temporal changes reflected in significant variable set differences. The differences between the significant variables for commercial and residential building energy consumption predictions were also traced.

2.1. Commercial Buildings Energy Consumption Survey (CBECS) microdata

CBECS were conducted in two phases: building survey and energy supplier survey that non-energy related data sets were collected. The first CBECS was conducted in 1979, and the most recent one was the CBECS 2012[7]. In this research, we used the CBECS 1992 and the CBECS 2012. The CBECS 1992 included 6,734 buildings' data with 945 variables. Differently, 6,720 buildings' data and 1,120 variables were listed in the CBECS 2012. The detailed information for each variable are made available on the CBECS website[7].

2.2. Residential energy consumption survey (RECS) microdata

Like the CBECS, the RECS is also conducted through interviews to collect data on housing units' energy characteristics, energy usage patterns, and household demographics. With the latest RECS 2015, additional data from energy suppliers were added to estimate energy cost, usages for heating, cooling, appliances, and other end-users[8]. For this research, we used the RECS 1993 and the RECS 2015. The RECS 1993 includes 7,111 buildings' data with 826 variables, while the RECS 2015 contains 5,686 buildings' data with 760 variables. The detailed explanations for each variable are available on the RECS site[8].

2.3. Data processing

We performed a series of multi-linear regressions to explore influential variables as the potential KPIs(Key performance indicators) to predict building energy consumption and to identify each variable's significance level in each end-use energy consumption model[13]. The statistical significance for each end-use energy consumption model was tested using a p-value with $\alpha = 0.05$; this test was done to all variables, including categorical variables, since those variables were also converted into numerical values.

Table 2. Hyperparameter values of the selected machine learning algorithms

Algorithm	Hyperparameter
Ridge regression	• α = 15.5 – 19.8
Lasso regression	• α = 5e-05 – 0.0008
Elastic net regression	• α = 0.0001 – 0.0007 • L1_ratio = 0.8 – 1
Support vector regressor	• C = 1 – 70 • γ = 0.00001 – 0.001

Further analysis of the variables was done by calculating the missing values in each variable across all data sets. In this work, we took average value (for numerical variables) and mode value (for categorical variables) to impute each variable's missing values. Afterward, output or target data normalization was completed for each model to obtain higher accuracy. Data normalization was done through logarithmic transformation and exponential function was deployed to the numerical results as an antilog from the precedent normalization.

In this research, we tested seven different machine learning algorithms to predict building energy consumption profiles. Those algorithms were ridge regression, lasso regression, elastic net regression, support vector regressor (SVR), light gradient boosting machine (LightGBM), gradient boosting, and XGBoost [13][19]. Table 2. shows the hyperparameter value for those selected machine learning algorithms which include the constant (α), ElasticNet mixing parameter (L1_ratio), regularization parameter (C) and kernel coefficient (γ).

K-Fold cross-validation (kFold =10) were also performed for more accurate predictions. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were adopted as the error rate evaluation metrics.

2.4. SHAP value

Shapley additive explanation (SHAP) value is an indicator to interpret the prediction acquired through complex model algorithms which could explain individual variable. SHAP value is a combination between the Shapley value (which computes

based on the cooperative game theory model) and the Local interpretable model-agnostic explanation (LIME) model (which focuses on training local surrogates to explain individual predictions)[18][20]. SHAP value has been proved to be consistent in both global and local interpretations.

We calculated the SHAP values by using the XGBoost algorithm to describe the influential factors. The twenty highest significant variables based on their SHAP value were plotted in the SHAP summary plot, which describes comparable importances of specific variables in each prediction instance.

3. Analytical results

3.1. Multi-linear regression analysis

CBECS and RECS data sets include a large number of variables. In order to reduce the number of variables, we performed a multi-linear regression analysis. Significant variables in the CBECS 1992, CBECS 2012, RECS 1993, and RECS 2015 were listed in Appendix A. Explanation of each variable is accessible on the EIA website[7][8].

3.2. Prediction metrics

Prediction metrics are described in Table 3., Table 4., and Table 5. Each building energy model was evaluated by calculating the average and the standard deviation of the evaluative metrics (RMSE and MAE) drawn from each fold, since 10 K-fold cross-validations were performed. Bold numbers in the tables indicate the lowest RMSE in each model, whereas italic numbers represent the lowest MAE in each model.

Table 3. shows the prediction results for the cooling model in both commercial and residential buildings. In commercial-building cooling-energy prediction (the CBECS 1992 and the CBECS 2012), the lowest RMSE and MAE are marked by either LightGBM, GradBoost, or XGBoost. Though, the lowest RMSE and MAE in residential cooling-energy prediction are made only by XGBoost.

Table 4. indicates prediction outcomes for the heating model. Similar to the cooling prediction case, the lowest RMSE and MAE for commercial heating-energy prediction were yielded by either LightGBM, GradBoost, or XGBoost. However, in the residential heating-energy prediction case, the lowest RMSE is marked by XGBoost, while GradBoost makes the lowest MAE.

Table 5. shows the prediction results for the water-heating model. The lowest RMSE and MAE for commercial water-heating-energy prediction were made by either LightGBM, GradBoost, or XGBoost. In this model, the lowest RMSE and MAE were generated by GradBoost.

Furthermore, the consistency of each model is evaluated through its standard deviation. As represented by the italicized numbers in Table 3., Table 4. and Table 5., most of lowest standard deviations were ascribed to the ensemble algorithm, XGBoost, specifically. In contrast with other data sets, regularization- regression (Lasso and ElasticNet) shows reliable performance with the RECS 2015 data set. In the cooling, heating, and water-heating RECS 2015 models, the lowest RMSE and standard deviation were generated by either Lasso or ElasticNet.

3.3. SHAP values in CBECS dataset

SHAP values are presented in summary plot graphs. In these plots, 20 variables that show the highest SHAP values are illustrated. For each variable, there are two outcomes for each evaluation point: SHAP value and variable value. SHAP value is defined by the positive or negative value on the x-axis, as shown in Fig. 2. and Fig. 3. Each dot's color defines the variable value, for instance pink color represents higher variable value whereas blue color indicates lower variable value.

Fig. 2. shows the SHAP summary plot for each end-use energy

Table 3. Prediction metrics for cooling energy analysis model (mean ± std. deviation calculated from each fold cross-validation)

Algorithm	CBECS 1992		CBECS 2012		RECS 1993		RECS 2015	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Ridge regression	11.1522 ±1.7803	4.9442 ±0.3188	21.0493 ±2.8031	7.9548 ±0.7045	3.1318 ±0.3308	1.7958 ±0.0536	1.8824 ±0.0378	1.5993 ±0.0195
Lasso regression	11.1482 ±1.7905	4.8992 ±0.3124	21.0366 ±2.7421	7.9531 ±0.6986	3.1223 ±0.3186	1.7769 ±0.0515	1.8779 *±0.0377*	1.5945 ±0.0190
Elastic Net regression	11.1468 ±1.7883	4.9009 ±0.3126	21.0392 ±2.7471	7.9530 ±0.6981	3.1225 ±0.3185	1.7773 ±0.0517	1.8779 *±0.0377*	1.5945 ±0.0190
SVR	16.0414 ±3.7637	5.1243 ±0.5186	25.7204 ±4.8724	5.6634 ±0.6330	2.9614 ±0.3315	1.4937 ±0.0437	1.8114 ±0.0535	1.5280 ±0.0173
LightGBM	6.1448 ±1.4609	2.6555 ±0.1893	**14.2661** ±2.6137	4.9070 ±0.4691	1.7116 ±0.0723	1.4038 ±0.0219	1.7225 ±0.0594	1.4658 ±0.0184
GradBoost	6.2478 ±1.5644	**2.4044** ±0.1577	15.2933 ±3.6301	**4.3966** *±0.4541*	1.5175 ±0.0601	1.2969 ±0.0159	1.7058 ±0.0639	1.4437 ±0.0212
XGBoost	**5.8483** *±1.3678*	2.5014 *±0.1471*	14.2851 *±2.5634*	4.8830 ±0.4582	**1.4688** *±0.0499*	**1.2680** *±0.0145*	**1.6905** ±0.0589	**1.4263** *±0.0149*

Table 4. Prediction metrics for heating energy analysis model (mean ± std. deviation calculated from each fold cross-validation)

Algorithm	CBECS 1992		CBECS 2012		RECS 1993		RECS 2015	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Ridge regression	27.8235 ±4.7698	11.1148 ±1.2897	28.5069 ±4.7570	10.3066 ±0.9881	5.2683 ±0.8103	2.3010 ±0.1139	3.7425 ±0.3615	2.1969 ±0.0758
Lasso regression	27.8294 ±4.7702	11.1275 ±1.2914	28.5041 ±4.7572	10.3161 ±0.9898	5.2598 ±0.8086	2.2955 ±0.1122	3.7411 *±0.3615*	2.1963 ±0.0758
Elastic Net regression	27.8289 ±4.7716	11.1268 ±1.2922	28.5054 ±4.7570	10.3160 ±0.9897	5.2596 ±0.8086	2.2953 ±0.1122	3.7412 ±0.3616	2.1964 ±0.0759
SVR	14.2928 ±2.7733	4.3194 ±0.4532	30.3556 ±6.6161	7.5224 ±0.9125	5.0027 ±0.8091	1.9283 ±0.1033	3.5214 ±0.4056	1.9159 ±0.0722
LightGBM	9.9605 ±1.3977	3.9490 ±0.2542	**15.2504** ±2.4363	5.7705 *±0.4834*	3.6388 ±0.3816	1.8426 ±0.0508	2.9132 ±0.4618	1.8035 ±0.0654
GradBoost	10.2897 ±1.5870	**3.4616** ±0.2413	16.0412 ±2.9311	**5.3306** ±0.5266	3.7688 ±0.3999	**1.7312** ±0.0451	2.9947 ±0.3956	**1.7530** *±0.0589*
XGBoost	**9.8023** *±1.2922*	3.8190 *±0.2195*	15.6038 *±2.4053*	5.8237 ±0.4997	**3.5310** *±0.3480*	1.8006 *±0.0444*	**2.8980** ±0.4090	1.8072 ±0.0601

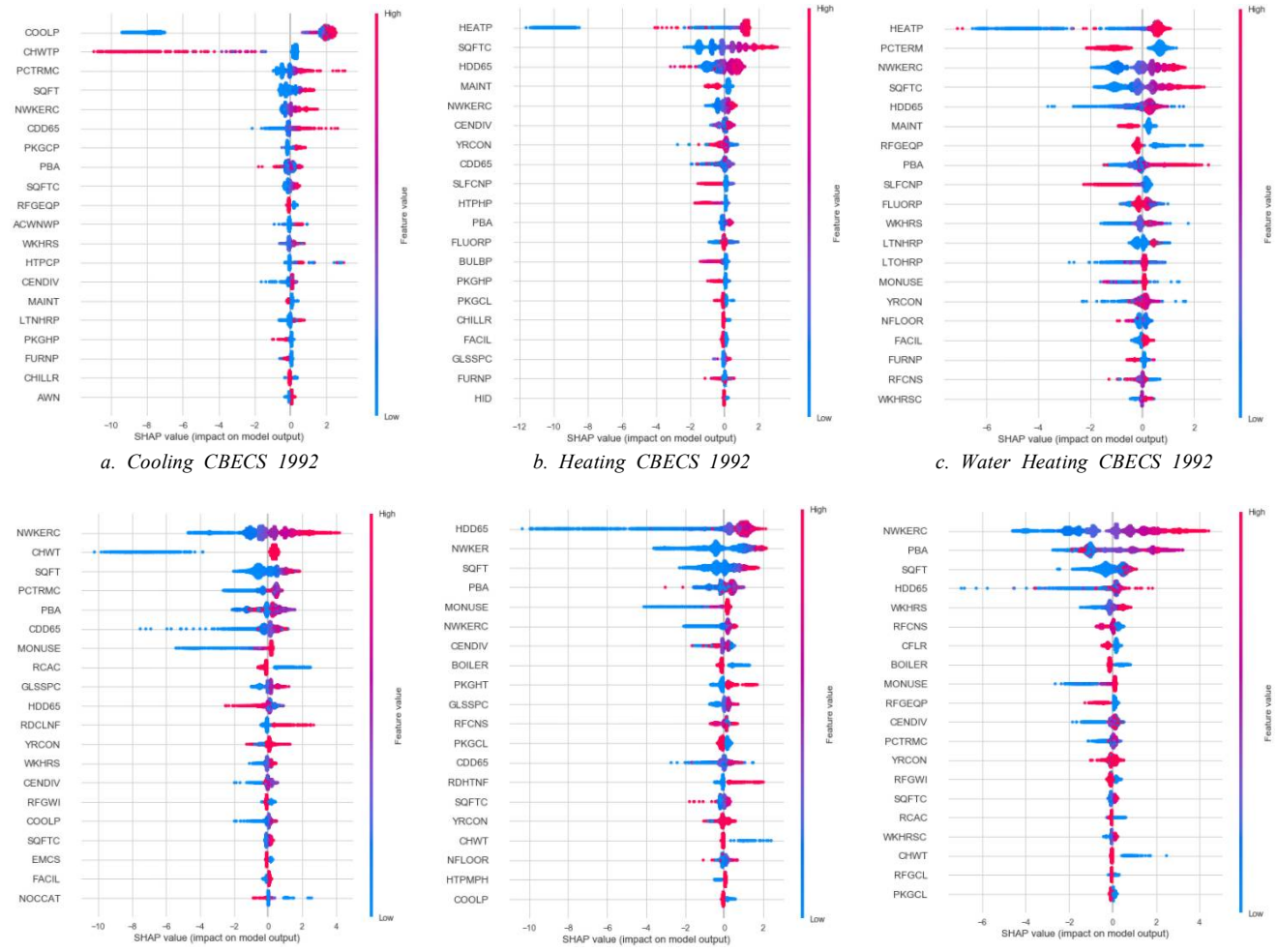
Table 5. Prediction metrics for water heating energy analysis model (mean ± std. deviation calculated from each fold cross-validation)

Algorithm	CBECS 1992		CBECS 2012		RECS 1993		RECS 2015	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Ridge regression	41.1601 ±7.1351	14.9803 ±1.6295	20.1728 ±2.4120	8.9688 ±0.6863	2.6986 ±0.2679	1.7570 ±0.0319	1.8705 ±0.1488	1.4506 ±0.0173
Lasso regression	41.1673 ±7.1336	14.9845 ±1.6280	20.1734 ±2.4119	8.9714 ±0.6866	2.6939 ±0.2670	1.7498 ±0.0318	1.8704 ±0.1491	1.4506 ±0.0171
Elastic Net regression	41.1656 ±7.1324	14.9841 ±1.6280	20.1740 ±2.4120	8.9713 ±0.6869	2.6939 ±0.2670	1.7498 ±0.0318	1.8704 ±0.1490	1.4506 ±0.0171
SVR	44.5340 ±10.9480	9.3098 ±1.4614	19.5053 ±2.8780	7.7244 ±0.7691	2.6552 ±0.2797	1.6761 ±0.0349	1.8571 *±0.1473*	1.3999 *±0.0133*
LightGBM	29.0626 ±4.7902	9.4846 ±0.8866	**13.6551** *±1.9819*	5.8858 *±0.4841*	2.5098 *±0.2603*	1.6497 *±0.0307*	1.8473 ±0.1481	1.4168 ±0.0178
GradBoost	29.5442 ±5.7027	**8.2447** ±0.8358	13.8251 ±2.1747	**5.5691** ±0.4925	**2.5082** ±0.2671	**1.6089** ±0.0330	**1.8435** ±0.1481	**1.3897** ±0.0136
XGBoost	**28.0291** *±4.6407*	8.9564 *±0.7646*	13.7955 ±2.0623	5.8718 ±0.5422	2.5118 ±0.2620	1.6525 ±0.0308	1.8652 ±0.1515	1.4196 ±0.0175

consumption model built upon CBECS data sets. With CBECS 1992 dataset, the cooling model illustrated in Fig. 2.(a) shows a set of influential variables including the highest five variables, namely, cooling percentage (COOLP), cooling percentage by district chilled water (CHWTP), number of computers (PCTRMC), total building area (SQFT), and number of workers category (NWKERC). CHWTP variable shows a negative correlation with cooling energy, which the higher value of CHWTP significantly reduces cooling energy consumption. PCTRMC, SQFT, and NWKERC variables show the same characteristics for commercial cooling energy consumption patterns; higher variable values induce higher positive SHAP value, impacting the corresponding model output. COOLP variable also shows a positive correlation, yet the lower value of COOLP would significantly generate a lower negative SHAP value than other CBECS 1992 variables.

Conversely, in the cooling model CBECS 2012 model shown in Fig. 2.(d), COOLP variables do not show the importance level as the same variable in the CBECS 1992 model. Instead, NWKERC marks the most significant variable in cooling-energy consumption in the CBECS 2012. District chill water (CHWT) also becomes one of the influential variable factors, which the lower value of CHWT (indicates district chilled water piped in) generates a negative correlation to cooling energy consumption. This evidence also supports the ground basis on CHWTP in the CBECS 1992.

The heating (Fig. 2.(b)) and water-heating (Fig. 2.(c)) model for CBECS 1992 data set share the same variables such as heating percentage (HEATP), total building area category (SQFTC), heating degree days (HDD65), number of workers category (NWKERC) and HVAC regular maintenance (MAINT). Other variables show a positive correlation to the heating and water heating model; while, the MAINT variable has the opposite tendency. A high variable value of MAINT (indicates no regular



a. Cooling CBECS 1992 *b. Heating CBECS 1992* *c. Water Heating CBECS 1992*

d. Cooling CBECS 2012 *e. Heating CBECS 2012* *f. Water Heating CBECS 2012*

Fig. 2. SHAP values for CBECS datasets

HVAC maintenance) would generate a lower negative SHAP value for heating and water-heating energy consumption. Surprisingly enough, in the water-heating model illustrated in Fig. 2.(c), computer usage (PCTERM) becomes one of the most significant variables. The high variable value of PCTERM (indicates computer usage in the building) also yields lower water-heating energy consumption. It is found that a higher PCTERM variable value indicated 'No computer usage' in the CBECS 1992 annotations. In fact, the absence of computer usage also indicated fewer workers (occupancy) in commercial buildings. This leads to lower water-heating energy consumption.

In the heating (Fig. 2.(e)) and water-heating (Fig. 2.(f)) model of CBECS 2012, NWKER (or NWKERC), HDD6, and SQFT were marked as common influential variables. Also, for the CBECS 2012, building activity importance level is higher than the CBECS 1992 case. It is also found that HDD65 variable value in CBECS 2012 (Fig. 2.(e) and 2.(f)) yields higher SHAP values than in the CBECS 1992 case (Fig. 2.(b) and 2.(c)). The low variable value of HDD65 in the CBECS 2012 yields a more significant low negative-SHAP value than HDD65 in CBECS 1992 case.

3.4. SHAP values in RECS dataset

Fig. 3. shows the SHAP summary plot for each end-use energy consumption model built upon RECS data sets. In Fig. 3.(a), the cooling-energy influential factors in the RECS 1993 are the usage of air-conditioning (AIRCOND), air-conditioning usage behavior (USECENAC), number of individual air-conditioning (NUMBERAC), cooling degree days (CDD65), and age of air-conditioning (AGECENAC). AIRCOND and CDD65 show a positive correlation with cooling energy consumption. Though, a higher variable value of NUMBERAC would generate the lower negative SHAP value. This case happens because RECS 1993 questionnaire put annotate '99' to indicate 'not applicable' choice for interviewees. The AGECENAC variable also has the same annotation as the NUMBERAC.

The bias annotation problem identified in RECS 1993 case is not repeated in RECS 2015 case; instead of annotating with the

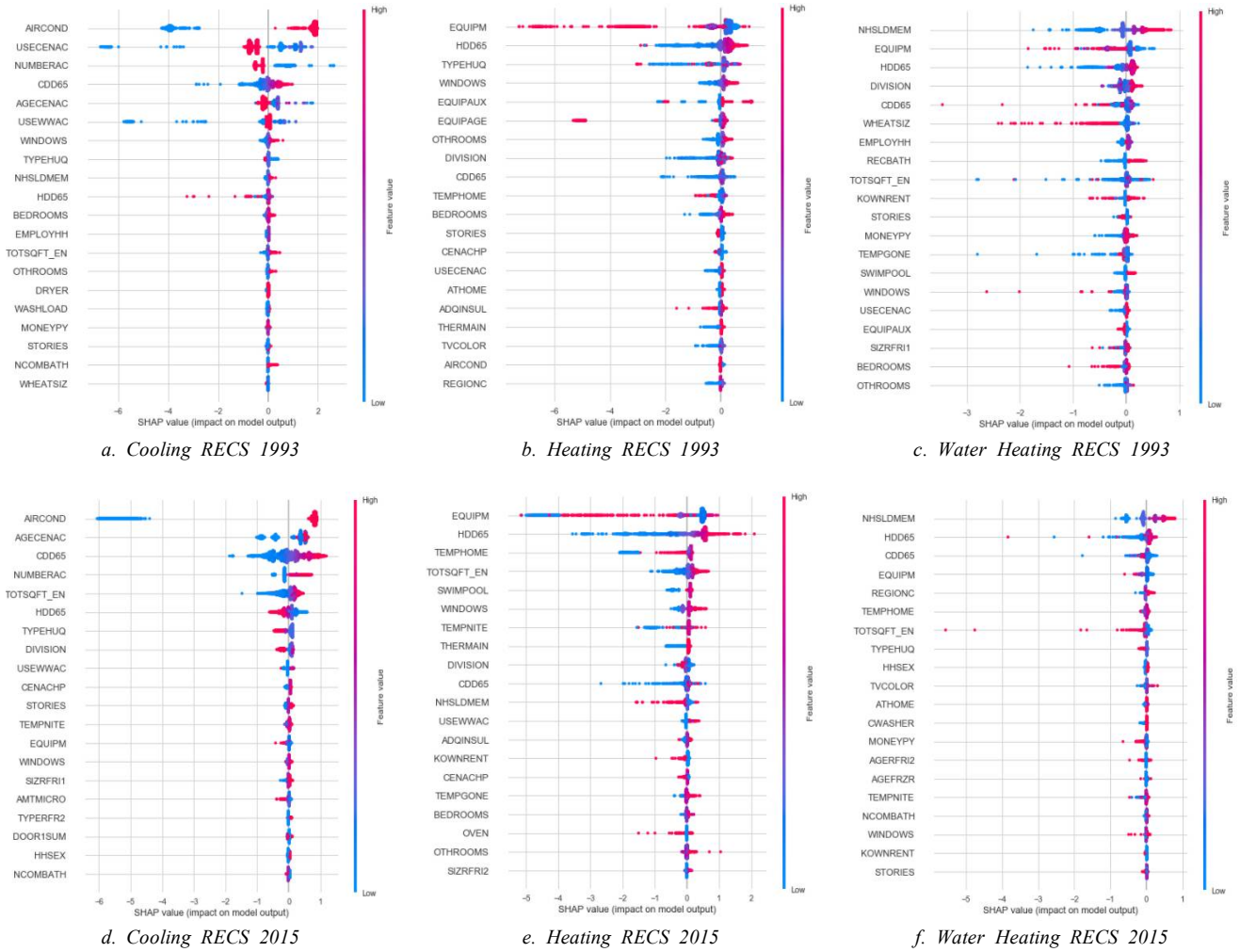
'99' for 'not applicable' choice, RECS 2015 annotates it with the '−2' value. Fig. 3.(d) shows SHAP values for RECS 2015 with the AIRCOND, AGECENAC, CDD65, and NUMBERAC are turned to be significantly influential variables.

Unlike heating and water-heating models built upon the CBECS dataset, heating and water-heating in RECS dataset do not share the same influential factors. The main space heating equipment type (EQUIPM) variable becomes the only common influential factor for heating and water-heating energy consumption in the RECS 1993 and RECS 2015.

In the heating model of RECS 1993 (Fig. 3.(b)), influential factors are EQUIPM, HDD65, type of housing unit (TYPEHUQ), number of windows (WINDOWS), secondary space heating (EQUIPAUX), and age of main space equipment (EQUIPAGE). HDD65, WINDOWS, and EQUIPAUX are positively correlated with heating energy consumption. In this model, bias annotations are also found in EQUIPAGE that yields pink dots in the lower negative SHAP values (Fig. 3.(b)).

In the heating model of RECS 2015, the EQUIPM, HDD65, and WINDOWS variables turn out to be influential factors (Fig. 3.(e)). Other than that, indoor-temperature related variables become influential factors such as daytime indoor home temperature during winter (TEMPHOME) and night time indoor home temperature during winter (TEMPNIGHT). Also, swimming pool ownership influences the heating energy consumption where a lower variable value (indicates not having a swimming pool) would generate a lower negative SHAP value. Housing unit variables also influence the heating energy consumption, such as total square footage (TOTSQFT_EN) that is positively related.

In the water-heating model of RECS 1993 (Fig. 3.(c)) and RECS 2015 (Fig. 3.(f)), there are some common influential variables in both years. The number of household members (NHSLDMEM), location (census region (REGIONC) and census division (DIVISION)), EQUIPM, CDD65, and HDD65 are those common influential factors in both years. Interestingly, this is the only model that yields both CDD65 and HDD65 as influential factors. CDD65 is the factor that negatively correlated with water-heating energy consumption. On the other hand, HDD65 is positively correlated with water-heating energy



a. Cooling RECS 1993 *b. Heating RECS 1993* *c. Water Heating RECS 1993*

d. Cooling RECS 2015 *e. Heating RECS 2015* *f. Water Heating RECS 2015*

Fig. 3. SHAP values for RECS datasets

Table 6. Findings on the variables' trend differences for each energy model

Model	Variable	1992	2012	Model	Variable	1993	2015
CBECS	Cooling	COOLP (+) NWKERC (+)	COOLP (×) NWKERC (↑)	RECS	Cooling	-	-
	Heating and Water-heating	PBA (+) HDD65 (+)	PBA (↑) HDD65 (↑)		Heating	TEMPHOME (×) TEMPNITE (×) SWIMPOOL (×) NHSLDMEM (×)	TEMPHOME (+) TEMPNITE (+) SWIMPOOL (+) NHSLDMEM (-)
	Water-heating	PCTERM (-)	PCTERM (×)		Water-heating	TOTSQFT_EN (?)	TOTSQFT_EN (-)

(+) : Positively related, **(-) :** Negatively related, **(↑) :** Higher influence than precedent year, **(×) :** Unavailable factor, **(?) :** Scattered values

consumption.

In addition to the water-heating model of RECS 2015 (shown in Fig. 3.(f)), TEMPHOME variable is an influential one even though it is not so significant. The total building area (TOTSQFT_EN) variable in the water-heating model also significantly impacts water-heating energy consumption. A high variable value of TOTSQFT_EN would significantly lower water-heating energy consumption (Fig. 3.(f)). In contrast, the low variable value of TOTSQFT_EN in the RECS 1993 scatters along with the SHAP value without any trends detected (shown in Fig. 3.(c)).

4. Discussion

Machine-learning-based building-energy consumption predictions with SHAP value calculation were performed to identify the most influential factors impacting energy consumption and the differences reflected by survey years and building types. Results show that the ensemble algorithms possess advantages in lowering both RMSE and MAE. These results also confirm the outcome of a previous study by Robinson et al. (2015)[11]. We also found that RMSE and MAE that are marked in this study show the lowest error rate than other precedent studies.

Regarding its consistency, we also identified that the ensemble algorithm has better consistency than other algorithms. Meanwhile, in the RECS dataset, the regularization-regression algorithm (Lasso and ElasticNet) shows fairly reliable performance in error rate and standard deviation. We could infer that the regularization-regression algorithm could work well on a complete dataset since there is no missing data on the RECS 1993 and RECS 2015.

As being reflected on SHAP values, influential factors on energy consumption are also revealed. Table 6. shows the findings on the variables' trend differences in each energy model. Lokhandwala et al.(2018) found that the CBECS 2012 cooling-energy consumption increases as a function of non-climatic conditions relative to the CBECS 2003[12]. In reference to that statement, our study revealed that neither CDD65 nor HDD65 is marked as the source for the highest cooling-energy SHAP value. Fig. 2.(d) shows that the lower value of CDD65 in the CBECS 2012 significantly decreases cooling-energy consumption compared to CDD65 in CBECS 1992, which only slightly impacts the output.

The comparisons of SHAP values between the 1990s data set and the 2010s data set give some fascinating insight. Based on the heating energy model of RECS 2015, a higher number of household members (NHSLDMEM) would decrease heating energy consumption, which did not happen in the RECS 1993. Also, in the RECS 1993, square footage (TOTSQFT_EN) was positively related to water-heating energy consumption, which means a more extensive building area would consume more. Reversely, TOTSQFT_EN was negatively related to water-heating energy consumption in the RECS 2015. In addition, Fig. 2. shows that the number of worker factors in the CBECS 2012 becomes more influential than its SHAP values in the CBECS 1992.

In commercial buildings, common variables that influence cooling and heating energy consumption are the number of workers and the occupancy variables. On the other hand, residential energy consumption is more influenced by its HVAC appliances and climatic conditions. Only residential water-heating energy consumption is significantly influenced by NHSLDMEM. Furthermore, we also identified that residential buildings' windows are more influential than other passive design factors (i.e., insulation and roof).

5. Conclusion

This study investigated commercial and residential building-energy consumption prediction models based on the CBECS and the RECS data sets collected in the United States. Several selected machine learning algorithms were tested and acquired outstanding error rate reduction compared with precedent studies. SHAP value was also utilized to illustrate influential factors for building energy consumption. Influential factors such as climatic condition and occupancy show differences in commercial and residential buildings between the

1990s and the 2010s.

A more significant influence of climatic conditions is expected in future years. The number of workers in the 2010s also increasingly influence energy consumption in commercial buildings compared with the 1990s. Commercial buildings' cooling and heating energy are highly influenced by occupancy patterns based on our study, while residential buildings' cooling and heating energy are highly affected by equipment and climatic conditions. In the meantime, water-heating energy consumption is influenced by occupancy and climatic condition in both commercial and residential buildings.

In this research, we limited our study to CBECS and RECS data sets considering the data sets' data availability and size. Further study might need to relate the results inferred from the calculated SHAP values with various social and economic conditions. This process would strengthen the merit for machine-learning-based building-energy consumption prediction and stimulate potential beneficiaries to guarantee more informed and controlled energy consumption behaviors in the future based on climatic, social, and economic dimensions.

Acknowledgement

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure, and Transport (Grant 20PIYR-B153277-02).

This work is also supported by Smart City R&D project of the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant 20NSPS-B150126-03).

Reference

[1] H. Zhao and F. Magoulès, A review on the prediction of building energy consumption, Renew. Sustain. Energy Rev., 16(6), 2012, pp. 3586–3592.

[2] K. Amasyali and N. M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renew. Sustain. Energy Rev., vol. 81, 2018, pp.1192–1205.

[3] H. Do and K. S. Cetin, Residential Building Energy Consumption: a Review of Energy Data Availability, Characteristics, and Energy Performance Prediction Methods, Curr. Sustain. Energy Rep., 5(1), 2018, pp.76–85.

[4] L. Pérez-Lombard, J. Ortiz, and C. Pout, A review on buildings energy consumption information, Energy Build., 40(3), 2008, pp.394–398.

[5] 최준우 외 4인, 주거용 건물의 에너지 실사용량의 불확실성을 내포한 설명변수 인자에 대한 빅데이터 분석 기반의 정량화 방법-서울지역의 공동주택을 중심으로J, 생태환경건축학회지, 17(3), 2017, pp.75-81. // (J. W. Choi et. al., The Method of Quantitative Analysis Based on Big Data Analysis for Explanatory Variables Containing Uncertainty of Energy Consumption in Residential Buildings : Focused on Apartment in Seoul Korea, KIEAE Journal, 17(3), 2017, pp.75–81.)

[6] 김기중, 안영수, 이승일, 소득격차를 고려한 조건에서 건물과 도시계획 요소가 건물에너지 소비에 미치는 영향요인 분석, 국토계획, 52(5), pp.253–267, 2017. // (K.J. Kim, Y.S. An, S.I. Lee, Analysis of Influencing Factors of Building and Urban Planning on Building Energy Consumption Considering Income Gap – Focused on electricity consumption on August in Seoul, Journal of Korea Planning Association, 52(5), 2017, pp.253–267.)

[7] Energy Information Administration (EIA)- Commercial Buildings Energy Consumption Survey (CBECS), https://www.eia.gov/consumption/commercial/ (accessed Jun. 11, 2019).

[8] Residential Energy Consumption Survey (RECS) - Energy Information Administration, https://www.eia.gov/consumption/residential/index.php (accessed Jun. 11, 2019).

[9] M. Yalcintas and U. Aytun Ozturk, An energy benchmarking model based on artificial neural network method utilizing US Commercial Buildings Energy Consumption Survey (CBECS) database, Int. J. Energy Res., 31(4), 2007, pp.412–421.

[10] A. Kaskhedikar, T. A. Reddy, and G. Runger, Use of random forest algorithm to evaluate model-based EUI benchmarks from CBECS database, ASHRAE Trans., 121(1), 2015, pp.17-28.

[11] C. Robinson et al., Machine learning approaches for estimating commercial building energy consumption, Appl. Energy, vol. 208, 2017, pp.889–904.

[12] M. Lokhandwala and R. Nateghi, Leveraging advanced predictive analytics to assess commercial cooling load in the U.S., Sustain. Prod. Consum., vol. 14, 2018, pp.66–81.

[13] H. Deng, D. Fannon, and M. J. Eckelman, Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata, Energy Build., vol. 163, 2018, pp.34–43.

[14] L. Troup, R. Phillips, M. J. Eckelman, and D. Fannon, Effect of window-to-wall ratio on measured energy consumption in US office buildings, Energy Build., vol. 203, 2019, p.109434.

[15] S. Kim and E. Shim, Statistical Analysis of Window Impacts on Cooling and Heating Energy Use in Single Family Residence Based on Climate Regions in U.S.A., KIEAE Journal, 18(6), 2018, pp.5–11.

[16] P. A. Mathew, L. N. Dunn, M. D. Sohn, A. Mercado, C. Custudio, and T. Walter, Big-data for building energy performance: Lessons from assembling a very large national database of building energy use, Appl. Energy, vol. 140, 2015, pp.85–93.

[17] B. Noh, J. Son, H. Park, and S. Chang, In-Depth Analysis of Energy Efficiency Related Factors in Commercial Buildings Using Data Cube and Association Rule Mining, Sustainability, 9(11), 2017, p.2119.

[18] S. M. Lundberg and S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in Advances in Neural Information Processing Systems 30, Long Beach, 2017, p.10.

[19] J. Han, M. Kamber, J. Pei. Data Mining Concepts and Techniques, 3rd ed., Massachusetts: Morgan Kaufmann, 2012.

[20] C. Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, https://christophm.github.io/interpretable-ml-book/, 2020.

Appendix A. Energy consumption modeling input variables selected from CBECS and RECS data sets

CBECS 1992			CBECS 2012			RECS 1993			RECS 2015		
Cooling	Heating	Water Heating	Cooling	Heating	Water Heating	Cooling	Heating	Water Heating	Cooling	Heating	Water Heating
ACWNWP	BLDSHP	BOILP	BOILER	ACWNWL	ACWNWL	AGECENAC	ADQINSUL	ADQINSUL	ADQINSUL	ADQINSUL	ADQINSUL
AWN	BULB	CHWTP	CDD65	BOILER	BLDSHP	AGERFRI1	AGERFRI2	AGERFRI2	AGECENAC	AIRCOND	AIRCOND
BULBP	BULBP	FACIL	CENDIV	BULB	BOILER	AIRCOND	AIRCOND	AIRCOND	AIRCOND	BEDROOMS	ATHOME
CDD65	CDD65	FDSEAT	CHWT	CDD65	BULB	BEDROOMS	AMTMICRO	BEDROOMS	AMTMICRO	CENACHP	BEDROOMS
CENDIV	CENDIV	FLUORP	COOLP	CENDIV	CENDIV	CDD65	ATHOME	CDD65	BEDROOMS	DIVISION	CDD65
CFLRP	CFLR	FURNP	EMCS	CHILLR	CFLR	DRYER	BEDROOMS	CENACHP	CDD65	DOOR1SUM	CENACHP
CHILLR	CHILLR	HCBED	EVAPCL	CHWT	CHWT	EMPLOYHH	CDD65	CWASHER	CENACHP	EQUIPM	EQUIPAUX
CHWTP	CHWTP	HDD65	FACIL	COOLP	COOLP	EQUIPAGE	CENACHP	DISHWASH	DIVISION	HDD65	EQUIPM
COOLP	FACIL	HEATP	GLSSPC	ECN	FACIL	HDD65	CWASHER	DIVISION	DOOR1SUM	HHSEX	HDD65
ECN	FLUORP	LTNHRP	HDD65	FACIL	FLUOR	NCOMBATH	DISHWASH	DWASHUSE	EQUIPM	NHSLDMEM	HHSEX
EMCS	FURNP	LTOHRP	HID	GLSSPC	FLUORP	NHAFBATH	DIVISION	EMPLOYHH	HDD65	NUMMEAL	ICE
EVAPP	GLSSPC	MAINT	MONUSE	HDD65	HDD65	NHSLDMEM	DOOR1SUM	EQUIPM	NUMBERAC	POOL	KOWNRENT
FACIL	HDD65	MONCON	NOCCAT	HEATP	HEATP	NUMBERAC	DRYER	HDD65	SIZRFRI1	SIZRFRI1	MICRO
FDSEAT	HEATP	MONUSE	NWKERC	HID	LTOHRP	OTHROOMS	DWASHUSE	KOWNRENT	STORIES	SWIMPOOL	MONEYPY
FLUORP	HID	NFLOOR	OTCLEQ	HTPMPH	MONUSE	SDESCENT	EQUIPAGE	MONEYPY	TOTSQFT_EN	TEMPHOME	NHSLDMEM
FURNP	HTPHP	NWKERC	OTLT	MAINT	NOCCAT	STORIES	EQUIPAUX	NHAFBATH	TYPEHUQ	TEMPNITE	NUMBERAC
GLSSPC	HTPMPC	OTLT	PBA	MONUSE	NWKERC	TYPEHUQ	EQUIPM	NHSLDMEM	USEWWAC	THERMAIN	NUMFREEZ
HEATP	HTPMPH	PBA	PCTRMC	NFLOOR	OTCLEQ	USECENAC	HDD65	NUMFRIG	WINDOWS	TOTSQFT_EN	NUMFRIG
HIDP	LODGRM	PCTERM	RCAC	NOCCAT	OTLT	USEWWAC	NHAFBATH	POOL	DISHWASH	WINDOWS	NUMMEAL
HTPCP	MAINT	PKGHP	RDCLNF	NWKER	PBA	WINDOWS	NHSLDMEM	RECBATH	THERMAIN	ICE	POOL
HTPHP	MONCON	RDHTNF	RDLTNF	NWKERC	PCTERM	DWASHUSE	NUMFRIG	REGIONC	OVEN	NUMCFAN	REGIONC
LODGRM	NRSBED	RDLTNF	RFGWI	OTCLEQ	PCTRMC	MONEYPY	OTHROOMS	SDESCENT	NUMCFAN	USEWWAC	SIZRFRI1
LTNHRP	NWKERC	RFCNS	SLFCON	OTLT	PKGCL	WASHLOAD	POOL	SIZRFRI1	NCOMBATH	SIZRFRI2	SIZRFRI2
LTOHRP	OTCLEQ	RFGEQP	SQFT	PBA	PKGHT	DISHWASH	RECBATH	STORIES	TEMPNITE	AGECENAC	STORIES
MAINT	OTCLP	RFGWIN	SQFTC	PKGCL	RCAC	WHEATSIZ	REGIONC	SWIMPOOL	HHSEX	ATHOME	SWIMPOOL
MONUSE	PBA	RWSEAT	WKHRS	PKGHT	RFCNS	THERMAIN	SIZRFRI1	THERMAIN	WHEATSIZ	OTHROOMS	TEMPHOME
NFLOOR	PKGCL	SLFCNP	WKHRSC	RDHTNF	RFGCL	WHEATAGE	STORIES	TVCOLOR	OTHROOMS	TYPERFR2	THERMAIN
NWKERC	PKGHP	SQFTC	YRCON	RDLTNF	RFGEQP	NUMMEAL	SWIMPOOL	TYPERFR2	TVCOLOR	CDD65	TVCOLOR
PBA	REGION	WKHRS		REGION	RFGOP	EQUIPAUX	TVCOLOR	USECENAC	TYPERFR2	TEMPGONE	TYPEHUQ
PCTRMC	RFGEQP	WKHRSC		RFCNS	RFGWI	TOTSQFT_EN	TYPEHUQ	WASHLOAD	STOVE	NHAFBATH	TYPERFR1
PKGCP	SLFCNP	YRCON		RFGCL	SLFCON	TVCOLOR	TYPERFR2	WHEATSIZ		USECENAC	TYPERFR2
PKGHP	SQFTC	YRCONC		RFGEQP	SQFT	DOOR1SUM	USECENAC	OTHROOMS		REGIONC	WINDOWS
RCACP	YRCON			RFGOP	SQFTC		USEWWAC	USEWWAC		TYPERFR1	CWASHER
RDHTNF	YRCONC			SQFT	WKHRS		WINDOWS	AMTMICRO		KOWNRENT	UPRTFRZR
RDLTNF				SQFTC	WKHRSC		THERMAIN	TEMPHOME		NUMFRIG	NCOMBATH
RFGEQP				YRCON	YRCON		WASHLOAD	ATHOME		OVEN	SDESCENT
RWSEAT				YRCONC	YRCONC		NUMBERAC	EQUIPAUX			EMPLOYHH
SLFCNP							TYPERFR1	TYPERFR1			AGERFRI2
SQFT							MICRO	WINDOWS			USEWWAC
SQFTC							SIZFREEZ	TEMPGONE			TEMPNITE
WKHRS							TEMPHOME	DOOR1SUM			AGEFRZR
WKHRSC							EMPLOYHH	TOTSQFT_EN			RECBATH
YRCON								DRYER			TOTSQFT_EN