Interrelated Patterns of Electricity, Gas, and Water Consumption in Large-Scale Buildings

Ali Movahedi*[a] and Sybil Derrible[a]

[a]*Complex and Sustainable Urban Networks(CSUN) Laboratory, University of Illinois at Chicago, 842 W Taylor Street, Chicago, IL 60607, United States*

Abstract

As cities keep growing worldwide, so does the demand for key resources such as energy (electricity and gas) and water that residents consume. Meeting the demand for these resources can be challenging and requires an understanding of their consumptions patterns. In this work, we apply XGBoost (Extreme Gradient Boosting) to predict and analyze water and energy consumption in large-scale buildings in New York City. For this, the New York City's local law 84 extensive dataset was merged with the Primary Land Use Tax Lot Output (PLUTO) dataset as well as with other socio-economic databases. Specifically, we developed three models: electricity, gas, and water consumption. Seven major lessons were learnt in terms of interrelationships between electricity, gas, and water consumption. In particular, water and gas consumption are highly interrelated with one another (often because gas is used for water heating). Furthermore, electricity consumption is affected by building type, and electricity and water consumption are particularly interrelated in nonresidential buildings. Overall, the knowledge gained from the models and from the SHAP analysis can help planners, engineers, and policymakers develop more effective strategies and help them manage the demand for energy and water in large-scale buildings.

Keywords: urban sustainability; water consumption; electricity consumption; gas consumption; machine learning

*Corresponding author. Email Addresses: amovah2@uic.edu and derrible@uic.edu

1. Introduction

As the world population keeps growing, so does the demand for key energy and resources. Moreover, most of this growth is happening in cities that already account for over 67% of energy use worldwide (Hobley, 2019). As a result, providing the necessary resources and energy for a growing population has become one of the most significant challenges for engineers, planners, and policymakers, while needing to account for local specificities (Mohareb et al., 2016). Specifically, building energy consumption represents a significant portion of total energy consumption and carbon emissions worldwide (Dalia, 2014). Buildings in the United States (US), for instance, contribute to 69% of electricity use, 39% of energy use, and more than 29% of total U.S. CO_2e emissions (Becerik-gerber et al. 2013; USEPA 2017). Similarly, despite being challenging (Ahmad and Derrible, 2015), the importance of studying water consumption trends is essential, especially considering the fact that 36% of the world's population experiences water scarcity (Von Grebmer et al., 2012). In this context, understanding building energy and water consumption is fundamental to help prepare cities meet future demand in energy and resources. In this respect, the literature on building energy and water consumption is relatively rich. Nonetheless, existing studies tend to study energy and water separately, without looking at the interrelationships between the two.

Building upon previous works, the main contribution of this article is to gain a fundamental understanding of the interrelationships between electricity, gas, and water consumption patterns. In the literature, an important knowledge gap exists to better understand how interconnected and interdependent the demand for and supply of urban infrastructure systems (Derrible, 2017; 2018). This article further leverages advances in Machine Learning (ML) to uncover meaningful patterns between the consumption of electricity, gas, and water in buildings. While ML techniques are generally less interpretable than traditional approaches that are based on strong predetermined assumptions (e.g., Kashani et al., 2019; Parsa et al., 2019a), to partly alleviate this issue, model-agnostic methods can be adopted to help interpret what ML models have essentially "learned." Various model-agnostic measures exist to interpret ML models such as feature importance (FI), partial dependence (PD), and individual conditional expectation (ICE) (Goldstein et al., 2013; Hastie et al., 2009). This article adopts the SHAP (SHapley Additive exPlanation) package (Lundberg and Lee, 2017), built in Python and previously used to study building energy consumption (Padopoulos and Kontokosta 2019).

More specifically, this article focuses on investigating the interrelationships between electricity, gas, and water use patterns in large-scale buildings using ML. Toward this goal, we use XG Boost to model resource consumptions—that is, electricity, gas, and water—and SHAP to interpret the learned ML models. The specific objectives of this article are to:

- Describe and analyze the performance of XG Boost to model electricity, gas, and water consumption,
- Investigate the role of specific variables for each model separately using SHAP,
- Study the interrelationships between electricity, gas, and water consumption in large-scale buildings using SHAP.

In the next section, we first briefly review the ML literature on building energy and water consumption. Then, we describe the datasets used in this study and we explain the methodological approach adopted. In each model, we predict either electricity, gas, or water based on the other two variables; for example, we use both water and gas consumption to predict electricity consumption. Furthermore, we include variables related to land use and socioeconomic variables that are commonly included in energy and resource consumption modeling. Subsequently, we use SHAP to interpret the results, identify the impact of individual variables, and study the interrelationships between electricity, gas, and water consumption.

2. Literature Review

The literature on building energy and water consumption is relatively abundant. The body of work on building energy and resource use is significantly aided by energy disclosure laws that enable the provision of open-source datasets (Athanassiadis, 2020) such as the New York local law 84 dataset that is used in this study and that notably contains information on water, electricity, and gas consumption in buildings with a total floor area greater than 50,000 square feet. Kontokosta and Tull (2017) applied machine learning to predict building energy use in New York City. Kontokosta and Jain (2015) used weighted robust regression and geographically weighted regression models to investigate and understand the determinants of water consumption in urban multi-family housing in New York City. Hsu (2015) demonstrated the application of penalized regression on commercial office and multifamily buildings energy consumption data set in New York City. Hsu (2015b) compared integrated clustering methods in terms of predicting building energy consumption and found that clusterwise regression gives more accurate predictions while K-means gives more stable clusters. Ma and Cheng (2016) analyzed the impact of 171 different kinds of features on regional energy use intensity (EUI) of residential buildings in New York City and identified the top 20 influential features. Meng et la. (2017) evaluated the impact of benchmarking policies on saving energy in the buildings in New York City. Using support vector machine, Jain et al. (2014) developed a model to forecast building energy consumption and applied the model to data from multi-family residential buildings in New York City. Finally, leveraging extreme gradient boosting and shapely additive explanation (used in this study as well), Papadopoulos and Kontokosta (2019) developed an interpretable and robust method to grade energy performance of residential buildings in New York City.

In general, many different factors are found to affect building energy consumption, including building physical properties and occupants behavior (Ashouri et al., 2019; Li et al., 2019), which makes energy consumption prediction challenging (Kwok and Lee, 2011). In particular, socioeconomic attributes can greatly affect energy and water consumption behavior. Several studies show that household income, household members age, education, and the percentage of female occupants affect building energy consumption (Pérez-Lombard et al., 2008; Tso and Yau, 2007). Other studies show that similar variables impact water consumption (Christenson et al., 2006; Kenney et al., 2008; Kontokosta and Jain, 2015; Wentz and Gober, 2007). Besides socioeconomic attributes, land use has also shown to have a significant impact on energy and water consumption (Pérez-Lombard et al., 2008). For example, energy and water consumptions patterns differ significantly depending on their type, from official to residential and commercial buildings, and (Pérez-Lombard et al., 2008) found that the average energy use intensity in offices is twice that of residential buildings. In this work, we look both at the effects of socioeconomic and land use factors on electricity, gas, and water consumption.

From a modeling perspective, various types of approaches have been adopted to model energy and water consumption. Beyond some of the approaches already mentioned, Suganthi and Samuel (2012) reviewed several modeling algorithms used to model energy consumption, including regression models, econometric models, and artificial systems. Donkor et al. (2014) reviewed the performance of econometric models, Artificial Neural Networks (ANN), and simulation or scenario-based modeling on water consumption. Recently, machine learning (ML) approaches have gained significant interest (Bourdeau et al., 2019; Seyedzadeh et al., 2019; Yu et al., 2016; Froemelt et al., 2019) and tend to perform better than traditional statistical regression modeling techniques thanks to their machine-based algorithmic computation. In particular, Decision Tree (DT) (Badhrudeen et al., 2020), ANN, ensemble methods (e.g., bagging and boosting methods), Support Vector Machine (SVM) (Parsa et al., 2019b), and regularized linear regression (e.g., ridge and lasso regression) tend to be the most popular ML approaches applied in the previous studies (Wei et al., 2018). For instance, Ahmad et al. (2014) forecasted building electricity consumption by using ANN and SVM, and Li et al. (2015) predicted building electricity consumption using ANNs and principle component analysis (PCA). Deng et al. (2018) compared several ML techniques to estimate commercial building energy consumption.

Furthermore, various studies (including some already mentioned) compare the performance of different methods. Papadopoulos et al. (2018a) evaluated the performance of random forest (RF), extremely randomized tree (RT), and gradient boosted regression tree (GBRT) to predict heating and cooling loads of buildings and found out GBRT is outperforming RF and RT. Robinson et al. (2017) also compared several techniques to estimate commercial building energy consumption and found that Gradient Boosting Machine (GBM) regression performed well. Lee and Derrible (2019) reviewed the performance 12 different techniques (four traditional and eight machine learning techniques) to predict household water consumption under two different scenarios (one with only a selected few variables generally available to utilities and one with an extensive number of variables) and found that gradient boosting performed best in both scenarios. Finally, the application of extreme gradient boosting and SHAP in modeling energy consumption in buildings has been utilized by Papadopoulos and Kontokosta (2019) as mentioned previously.

3. Data and Methodologies

3.1 Data Description

This article mainly uses three datasets: (1) the New York City's Local Law 84 Water and Energy benchmarking dataset 2017 (NYC's local law 84), (2) the Primary Land Use Tax Lot Output (PLUTO) dataset, and (3) the American Community Survey (ACS) from the US Census Bureau. NYC's local law 84 applies to buildings with a gross floor area larger than 50,000 square feet and it mainly contains information about the consumption of water, electricity, and gas in buildings, as well as several other building characteristics such as building age, gross floor area, and building usage type. PLUTO dataset files contain three types of data: tax lot characteristics, building characteristics, and geographic/political/administrative districts. Notably, the PLUTO dataset provides detailed land-use information at the tax lot level in New York City. In addition, the ACS dataset is used to incorporate socioeconomic information available at the census tract level including gender proportion, age distribution, education level, and median income.

The original dataset consists of a total of 11,746 buildings. Nonetheless, it contains many erroneous entries since the data are reported by individual building owners. In terms of cleaning, first, all rows that were missing values for either electricity, gas, or water consumption were removed since these values are essential for this work. Second, the dataset includes 937 buildings that use steam from the New York City large district steam system. All were removed since energy consumption from steam can be significantly large (e.g., for building heating) but 937 is too low to consider steam as a fourth dependent variable and most cities do not possess a district steam system. Finally, a traditional cleaning process was applied to remove all errors, outliers, duplicated data points, and other buildings with missing data.

The final cleaned dataset included 6,527 buildings and fifteen variables. A list of the variables collected and a summary of their statistics are shown in Table 1.

Table 1 Descriptive statistics of the merged dataset

Variable	Definitions	Mean	Std	Max	Min
Electricity	Electricity consumption per square meter per year $[kWh/m^2/y]$.	49	66.95	769	1.5
Gas	Gas consumption per square meter per year $[kWh/m^2/y]$.	298	307.75	2736.56	0
Water	Water consumption per square meter per year $[L/m^2/y]$	4239	3070.68	17670.23	87.02
Residential, Office, Commercial	Dummy variables to identify the type of the major part of a building.	—	—	—	—
Age	Age of a building [y].	71	31.48	191	2
More 60	The percentage of people older than 60 years old in a census tract.	18.64	8.09	58.7	1.7
Under 9	The percentage of people younger than 9 years old in a census tract.	10.91	4.84	41.7	0.8
Female Percentage	The percentage of female in a census tract.	52.7	4	72	31
Median Income	Median income of people in a census tract.	71253	40715	233529	9740
Building Density	The total gross area of buildings in a census tract divided by census tract area $[m^2/m^2]$.	1.65	1.35	10.47	0.05
Residential Area	The total gross area of residential buildings in a census tract divided by census tract area $[m^2/m^2]$.	0.98	0.58	4.46	0.004
Office Area	The total gross area of offices in a census tract divide by census tract area $[m^2/m^2]$.	0.3	0.74	6.27	0
Factory Area	The total gross area of factories in a census tract divide by census tract area $[m^2/m^2]$.	0.014	0.085	1.59	0
No High School Diploma	The percentage of people who did not get high school degree.	13.6	12.06	57.9	0
College Education	The percentage of people who got bachelor degree or more in a census tract.	53	26.92	93.9	5

3.2 Methodologies

To model and understand building energy consumption, this article mainly uses extreme gradient boosting (XG Boost). Boosting—as an ensemble method—is developed to enhance the modeling performances of decision tree (DT), executed through a repetitive machine-based algorithm and effectively controlling the bias-variance trade-off (Friedman et all., 2001). Specifically, a DT has a structure similar to a tree with a root node (topmost node), internal nodes, and leaf nodes (end nodes). DT algorithms use relatively simple rules to start from a root node and split up each node into two or more branches, going through internal nodes, and ending with the leaves. As "weak learners," DT tend to have a low performance and be unstable such that a small change in the data can make a large change in the tree structure. To overcome the weakness of DT, XG Boost models uses an ensemble of weak consecutive decision trees where each tree learns from the previous tree and tries to improve the model. XG Boost is detailed in the next section.

3.2.1 Extreme Gradient Boosting - XGBoost

XGBoost is an advanced version of gradient boosting machine. In this section, we briefly explain the main concepts and equations behind it, but the reader is referred to Chen and Guestrin (2016) for more details.

In a dataset with *n* samples, each independent variable x_i can have m features, $x_i \in \mathbb{R}^m$. Corresponding to each of these variables, we define a dependent variable y_i, $y_i \in \mathbb{R}$. By using independent variables x_i, a tree ensemble model *f* uses *K* additive functions to predict the dependent variable such that:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{1}$$

where $\hat{y}_i$ is the predicted value of y_i. Moreover, here, F is the space of regression trees and f_k corresponds to an independent tree structure with leaf scores. Our goal is to minimize the following function:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

where

$$\Omega(f) = \gamma \mathrm{T} + \frac{1}{2}\lambda \parallel \omega \parallel^2 \tag{3}$$

Here l is a loss function that measures the difference between the predicted and the target values, Ω is a term that penalizes the complexity of the model, T is the number of leaves, and ω_i is the score on i-th leaf.

For a fixed structure $q(x)$, the optimal value for ω_j^* and its corresponding values is:

$$\omega_j^* = -\frac{\sum_{i \in I_j} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}_i{}^{t-1})}{\sum_{i \in I_j} \partial^2_{\hat{y}^{t-1}} l(y_i, \hat{y}_i{}^{t-1}) + \lambda} \tag{4}$$

$$\tilde{\mathcal{L}}^t(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}_i{}^{t-1}))^2}{\sum_{i \in I_j} \partial^2_{\hat{y}^{t-1}} l(y_i, \hat{y}_i{}^{t-1}) + \lambda} + \gamma \mathrm{T} \tag{5}$$

Here $\hat{y}_i{}^t$ is the predicted value of *i*-th instance at the *t*-th iteration and define $I_j = \{i | q(x_i) = j\}$ where q represents the structure of a tree, thus I_j is the instance set of j.

In practice, it is impossible to calculate this value for all possible tree structures q to determine which split should be chosen. Instead, the following formula is used to evaluate the split candidates:

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L}\partial_{\hat{y}^{t-1}}l\left(y_i,\hat{y}_i^{\,t-1}\right)\right)^2}{\sum_{i\in I_L}\partial^2{}_{\hat{y}^{t-1}}l\left(y_i,\hat{y}_i^{\,t-1}\right)+\lambda}+\frac{\left(\sum_{i\in I_R}\partial_{\hat{y}^{t-1}}l\left(y_i,\hat{y}_i^{\,t-1}\right)\right)^2}{\sum_{i\in I_R}\partial^2{}_{\hat{y}^{t-1}}l\left(y_i,\hat{y}_i^{\,t-1}\right)+\lambda}-\frac{\left(\sum_{i\in I}\partial_{\hat{y}^{t-1}}l\left(y_i,\hat{y}_i^{\,t-1}\right)\right)^2}{\sum_{i\in I}\partial^2{}_{\hat{y}^{t-1}}l\left(y_i,\hat{y}_i^{\,t-1}\right)+\lambda}\right]-\gamma \quad (6)$$

where I_L and I_R are the instances sets of left and right nodes after the split and $I = I_L \cup I_R$ (Chen and Guestrin, 2016).

In the context of building electricity, gas, and water consumption, x_i represents the independent variables of a building (see Table 1) and y_i represents one of the three consumption values. For example for electricity consumption, y_i is electricity consumption and x_i includes all other m variables that are used to predict electricity consumption.

3.2.2 Model Parameters and Selection

XGBoost models can be defined by selecting specific values for several parameters, some of which are briefly explained in this section. The parameters should be selected to maximize the model performance on a given dataset and in such a way to prevent overfitting and unnecessary complexity. Overfitting happens when a model starts to learn the noises and random fluctuations present in the data and sees them as meaningful facts or concepts. XGBoost uses multiple parameters, which makes parameter tuning relatively difficult. Such parameters include number of boosting iterations, maximum depth, random subsampling, learning rate, lambda, and alpha. The number of boosting iterations in XG boost refers to the number of consecutive trees that are fitted. The maximum depth (max_depth) of the tree indicates the maximum number of splits; too high a maximum depth can sometimes cause overfitting. Randomly subsampling corresponds to a specific ratio of the training dataset in each iteration prior to growing trees; the same technique of subsampling can be applied to columns and in each iteration a ratio of columns is used to grow the trees. Learning rate (eta) is used to shrink the weights at each step and make the model more robust; essentially, the learning rate reduces the impact of each individual tree and let future trees improve the model (Chen and Guestrin, 2016). The parameters "lambda" and "alpha" respectively are L2 and L1 regularization term on weights, and by increasing them the model becomes more conservative. The following values represent the final parameters we used for the models after the cross-validation process (explained in the next section): number of boosting iterations: 500; maximum tree depth: 8; subsample ratio of training dataset in each split: 0.8; subsample ratio of columns in each split: 0.5; lambda: 1.5; alpha: 0.2; learning rate: 0.015.

3.2.3 Model Execution and Validation

In this work, the hold-out method is used to split the data into two groups: 80% for training and 20% for testing. In addition, we evaluate the model through 10-fold cross validation (CV). Specifically, we randomly split the training dataset only into ten folds. Then, in each step, we set a unique fold as our test data set and use the other nine folds to train a model and assess its performance on the test fold. We repeat the process so that all ten folds are using for testing once. The performance of each of the ten models during the CV process is calculated. Specifically, we look for a stable performance across all ten models to ensure that the technique used (i.e., XG

Boost) is meaningful. The combined average performance measures are calculated as well. If the CV process is satisfactory, the model is retrained on the full 80% of the data used as the training set and tested on the test set.

In terms of performance measure, in this work, we use goodness of fit R^2, defined as:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{7}$$

where y_i is the actual value of a data point, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of all the actual values.

3.2.4 Model Interpretation

In this article, we further use SHAP (Shapley Additive exPlanations) values to interpret the outputs. SHAP is based on game theory (Strumbelj and Kononenko, 2014) and local explanations (Ribeiro et al., 2016), and it unifies ideas from them in order to represent a locally accurate and consistent feature attribution (Lundberg and Lee, 2017). Let us assume a game v. In the game, there is a group *N* (with *n* members) with a specific output and the value of the output is $v(N)$. The idea behind shapley values is the value of the production $v(N)$ should be divided among the *n* members based on their marginal contributions (Shapley, 1953). In XGBoost, *n* represents the features (i.e., the independent variables) and the value of the production is the target value (i.e., the dependent variable). Therefore, we want to know how much each feature contributes (ϕ_i: attribution of feature *i*) in the final prediction of our model. There are four axioms that help us to fairly identify their contributions (Shapley, 1953):

- Efficiency axiom: the sum of the feature contributions is equal to the final prediction, such that $\sum_{i \epsilon N} \phi_i = v(N)$
- Symmetry axiom: if two features *i* and *j* have the same marginal contribution when we add them to any feature combinations for prediction, they should receive the same contribution from the final prediction, such that if $v(S \cup \{i\}) = v(S \cup \{j\})$ where $S \subset N \ and \ i, j \ \notin S$, then $\phi_i(v) = \phi_j(v)$
- Dummy axiom: if by adding a feature to any combinations of features, the prediction will not be changed, the feature will not receive any contribution from the final prediction, such that if $v(S \cup \{i\}) = v(S)$ where $S \subset N$ and $i \ \notin S$, then $\phi_i(v) = 0$
- Linearity axiom: if the model for prediction divided into two parts, then the total contribution of a feature is equal to the sum of the feature contribution in each of them, such that if $v(S) = v_1(S) + v_2(S)$ for all S, then $\phi(v) = \phi(v_1) + \phi(v_2)$

Based on these axioms, shapley values are determined by the following formula:

$$\phi_i = \sum_{S \subseteq N\{i\}} \frac{|S|!(n-|S|-1)!}{n!} [v(S \cup \{i\}) - v(S)] \tag{8}$$

In the additive feature attribution method, there is a linear function of binary variables g defined as:

$$g(z^{'}) = \phi_i + \sum_{i=1}^{M} \phi_i z_i^{'} \tag{9}$$

where $z' = 1$ when a feature observed, otherwise it is 0, and M is the number of input features (Lundberg and Lee, 2017).

4. Results

Figure 1 shows the results of the 10-fold cross validation process for the electricity, gas, and water consumption models. The overall low variation of CV-prediction accuracy for all models indicates that the performance of the models is consistent.

Figure 2 shows the actual versus the predicted electricity, gas, and water consumption values using XG Boost. Moreover, the goodness of fit R^2 of the three models are 0.80, 0.60, and 0.64 for electricity, gas, and water consumption respectively. The higher accuracy of the electricity consumption model is expected considering the independent variables selected. Indeed, gas consumption depends heavily on whether gas is used for building heating or not; in New York City, a significant number of large-scale buildings use steam for heating (not studied here) instead of gas. As a result, our model predicts the same gas consumption value for two buildings with the exact same characteristics, even if one uses gas for heating and the other uses electricity, hence the lower accuracy. As for water consumption model, water consumption also depends on other types of variables that could not be included in this study, such as the frequency and duration of taking a shower / bath, whether a unit has a washing machine and dishwater or not, and so on, hence the lower accuracy.

As a comparison, Papadopoulos and Kontokosta (2019) get a R^2 of 0.31 using XGBoost to predict energy use intensity (EUI) of New York City large-scale buildings. Our results are significantly higher specifically because of we consider electricity, gas, and water consumption together, suggesting that these three variables are strong interrelated; more on interrelationships later.

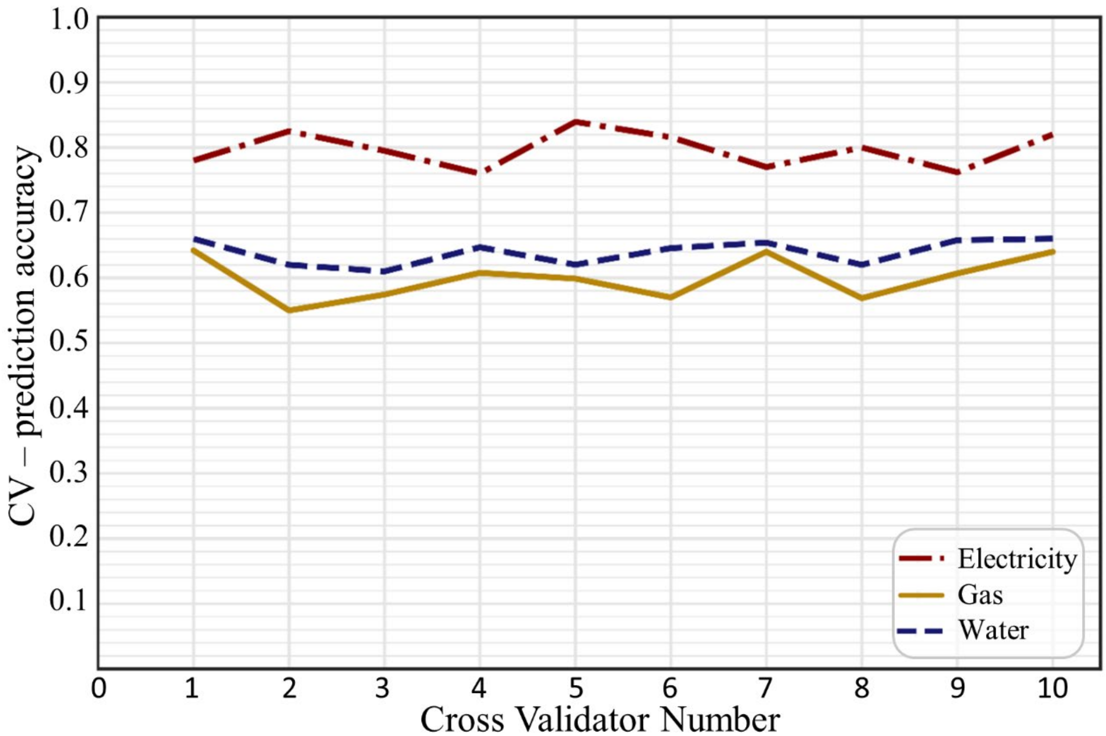


Figure 1. Ten-fold cross-validation results for electricity, gas, and water consumption models.

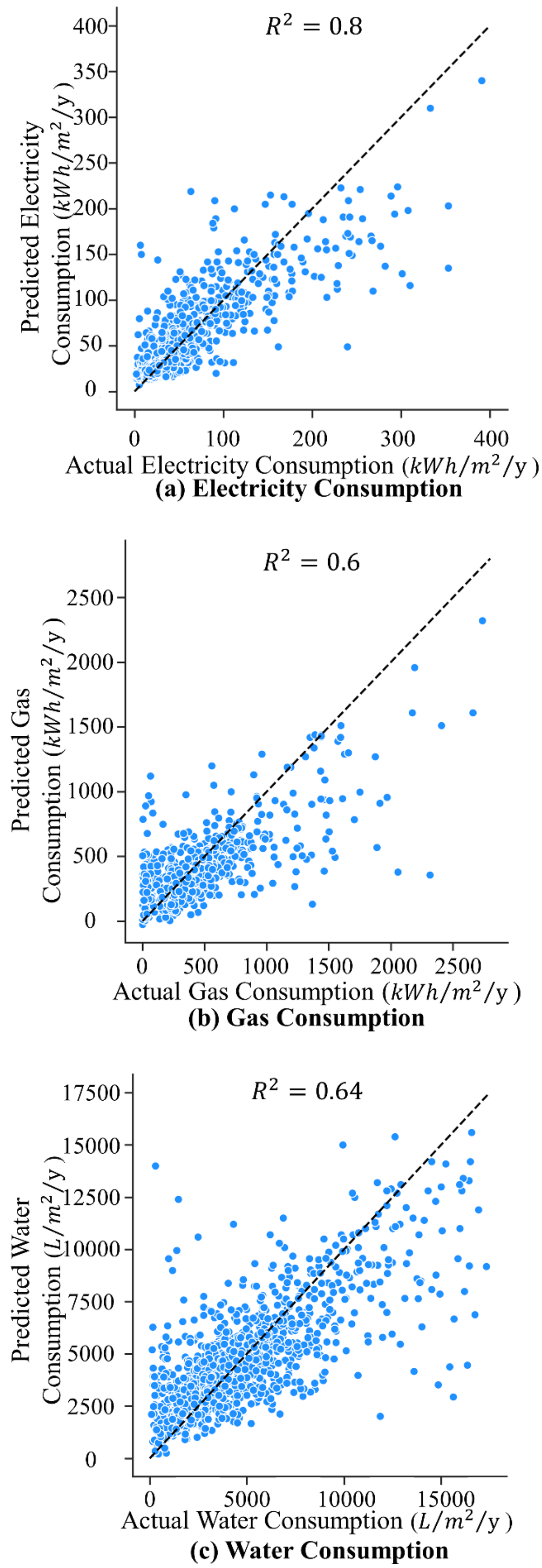


Figure 2. Actual vs predicted values for the electricity, gas, and water consumption models

Figure 3 shows the SHAP summary plots for predicting electricity, gas, and water consumption. The *x*-axis shows the impact of features on the outcomes based on the SHAP values. Features are sorted based on their impact, thus the variable at the top has the highest impact. The color represents the feature value, with red for high values and blue for low values. The vertical dispersion for each feature corresponds to the data points with the same SHAP values. If red points are mostly present on the positive side of SHAP values, it means by increasing the value of the independent variable, the dependent variable increases as well.

Figure 3.a shows that the most important variable when it comes to electricity consumption is whether a building is a residential building or not, and that by changing the land use type of a building from residential to something else, electricity consumption would increase. The three next important features are water consumption, gas consumption, and building age. Water and gas consumption have an overall a positive effect on electricity consumption, which means that a building that consumes more water and gas is likely to consume more electricity as well. Here, it is likely that water consumption acts as a proxy for building occupancy (which was not available as an independent variable)—an office building, for example, will have more occupants consuming both more electricity and water. A positive and significant relationship between electricity and water is therefore expected. In contrast, building age shows to have a negative relationship with electricity consumption. In other words, newer buildings tend to consume more electricity, which might reflect a higher usage of high energy consuming appliances such as air conditioners (EIA 2005).

For gas consumption (Figure 3.b), water and electricity consumption have the highest impacts, and by increasing water and electricity consumption, gas consumption generally increases. The third most important feature is building density, and by increasing building density, gas consumption is likely to decrease. This result is probably linked to the fact that higher building density suggests the presence of larger buildings that might have lower *U*-values and thus require less energy per unit area for heating.

Figure 3.c shows the results for water consumption. We can see that gas and electricity consumption have the highest impact in the positive direction, which means here again that buildings with higher electricity and gas consumption also generally consume more water. Here again, electricity likely acts as a proxy for building occupancy, and the positive interrelationship between electricity and water is expected. Furthermore, using the same dataset, Kontokosta and Jain (2015) also found a strongly relationship between electricity and gas consumption (in the form of energy use intensity) and water consumption. The third feature is college education (i.e., percentage of people with a college education) and it is in the opposite direction, which means that buildings with more educated people consume proportionally less water. This feature likely relates to the presence of median income as well in the seventh place, similarly captured in Kontokosta and Jain (2015) here as well. Interestingly, we also see that building age has a negative impact on water consumption; i.e., newer buildings tend to consume comparatively more water. Finally, Figure 3.b and 3.c shows the most important variable for gas consumption model is water consumption, and the most important parameter for water consumption is gas consumption. Water and gas consumption are highly interrelated, which can be partly explained by the fact that gas is often used for water heating; thus, a higher hot water consumption translates directly to a higher gas consumption.

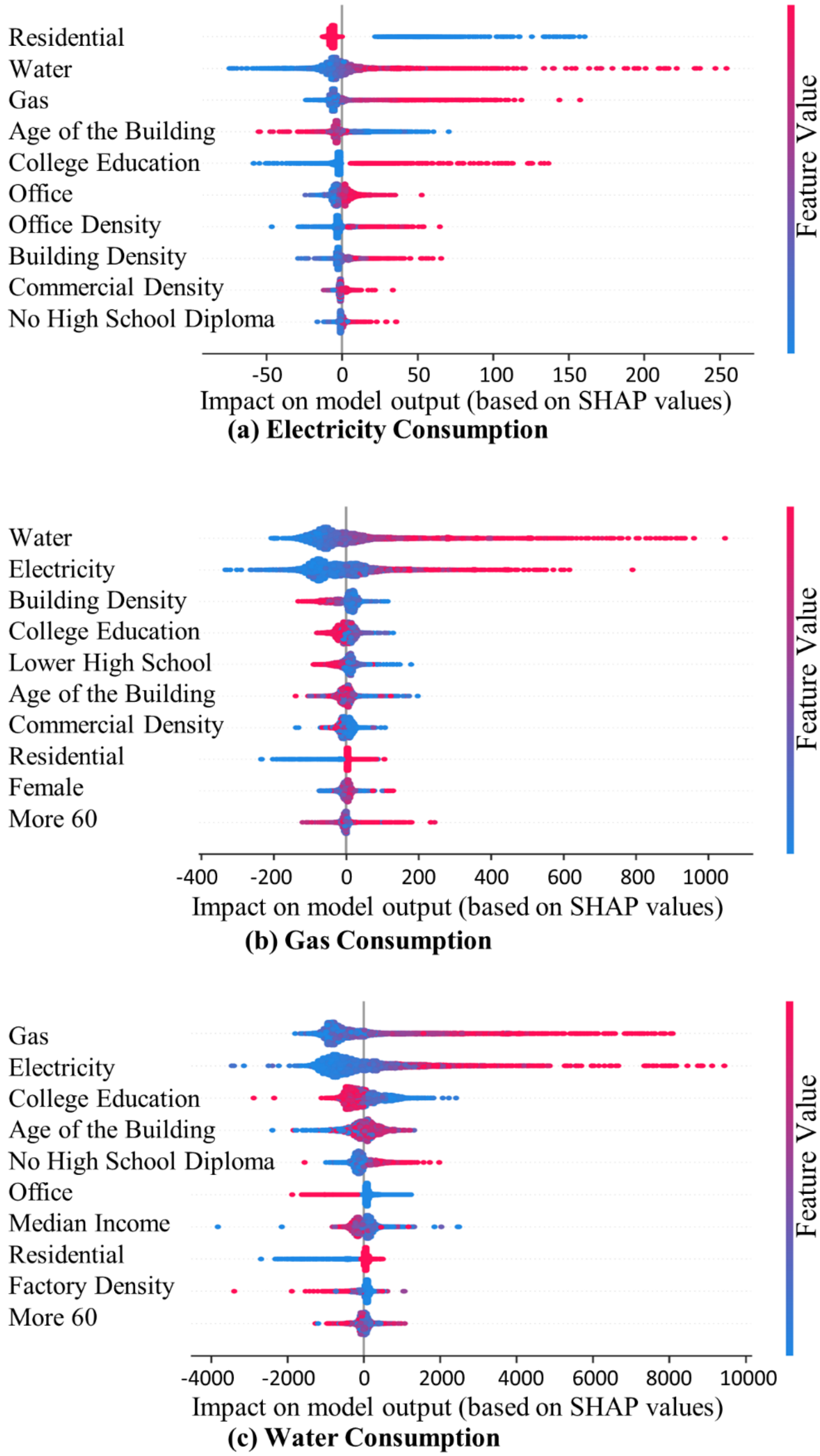
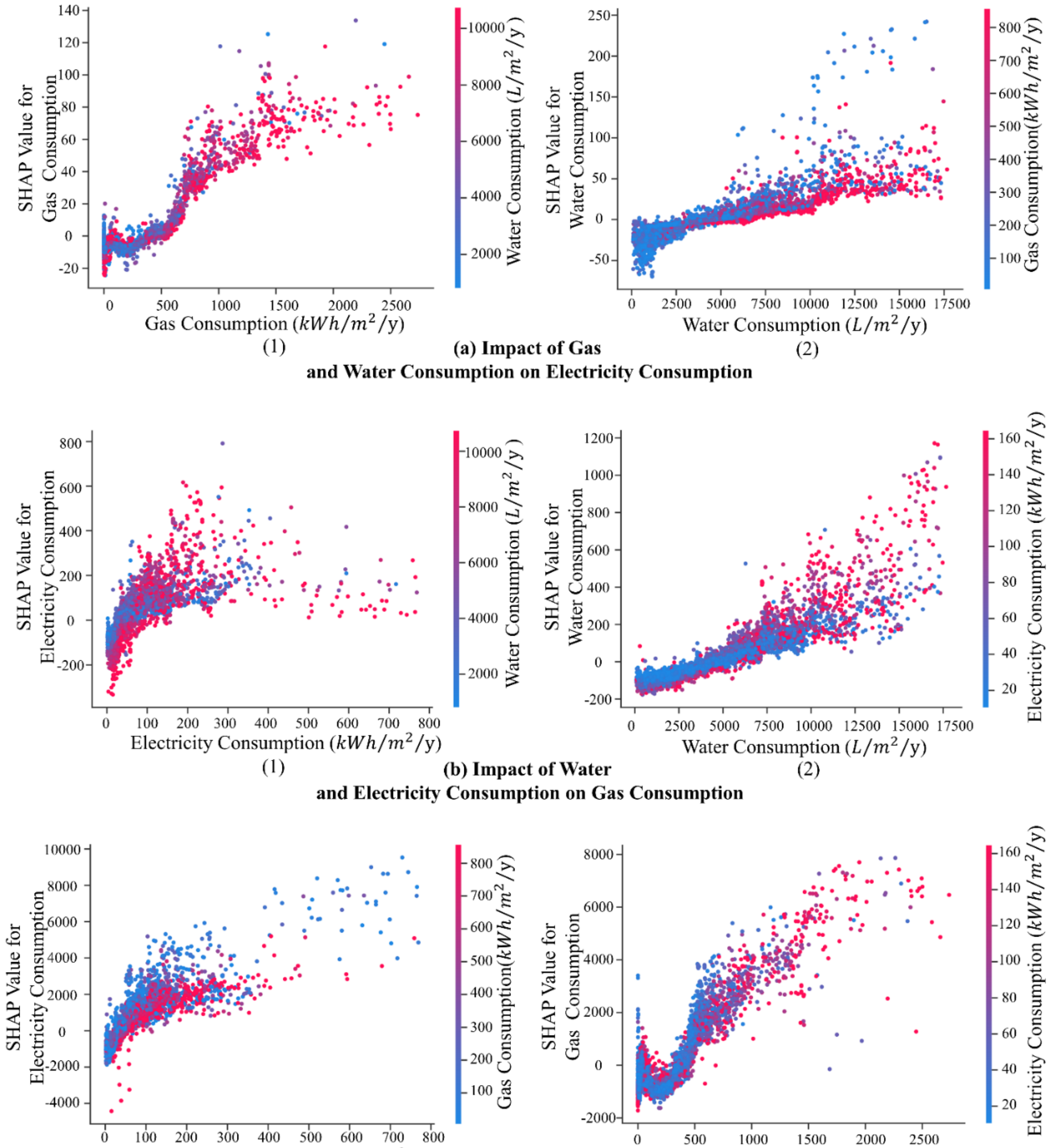


Figure 3. Contribution of the most important features in each consumption.

To further study the interrelationships between electricity, gas, and water consumption, Figure 4 shows plots with the value of a feature on the *x*-axis and its SHAP value on the *y*-axis by changing a specific feature in the model. Considering the fact that SHAP values indicate a feature's impact on a change in the model output, this type of figures shows the impact of one feature on the model output, while the change in the second feature is displayed by changing the color of the data points from blue to red. While these plots are not trivial to analyze, they offer an effective means to look at the interrelationships between three features.

Figure 4.a shows the impact of water and gas consumption on electricity consumption. In Figure 4.a.2, using the electricity consumption model, positive SHAP values can be seen for buildings with average to high water consumption, and data points tend to be blue, thus have a low gas consumption. This means that the interrelationship between electricity and water is particularly strong in buildings with low gas consumption but with average to high water and electricity consumption. This can be seen in Figure 4.c.1 as well that applies the water consumption model, and we can see that in buildings with average to high electricity consumption, SHAP values are higher in buildings with low gas consumption values. Buildings with average to high electricity consumption and low gas consumption likely use electricity for space heating and water heating. Considering the relationship between water consumption and water heating, buildings with average to high electricity consumption probably use electricity for water heating, which can be why electricity and water are more significantly interrelated.

Figure 4.b illustrates the impact of electricity and water consumption on gas consumption. Specifically, figure 4.b.1 shows a peak SHAP value for an electricity consumption value of about 300 $kWh/m^2/y$, which suggests that electricity consumption is increasingly interrelated with gas consumption up to a point, after which the interrelationship decreases. Moreover, the changing SHAP values between buildings with high and low water consumption suggests that electricity consumption has a higher impact on gas consumption for buildings with high water consumption. Figure 4.b.2 not only shows a direct relationship between water and gas consumption, it shows that this relationship becomes more important as water consumption increases since SHAP values increase simultaneously. Figure 4.c.2 echoes this observation as after a gas consumption of about 200 $kWh/m^2/y$, an increase in gas consumption almost directly leads to an increasing SHAP value.

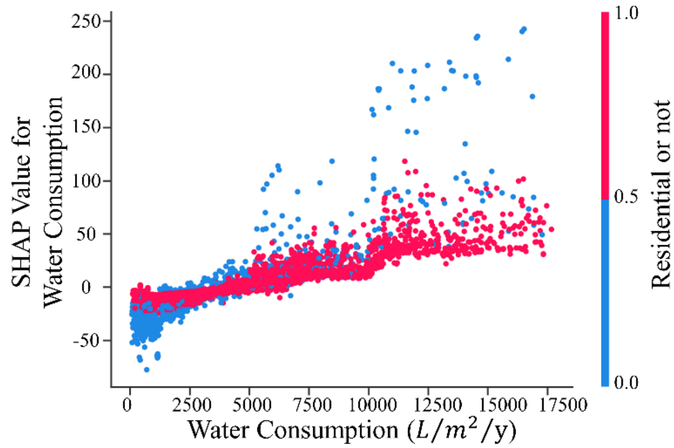


(a) Impact of Gas and Water Consumption on Electricity Consumption

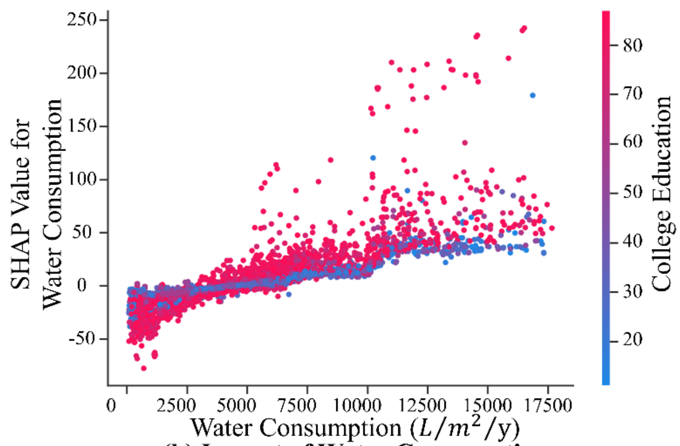
(b) Impact of Water and Electricity Consumption on Gas Consumption

(c) Impact of Electricity and Gas Consumption on Water Consumption

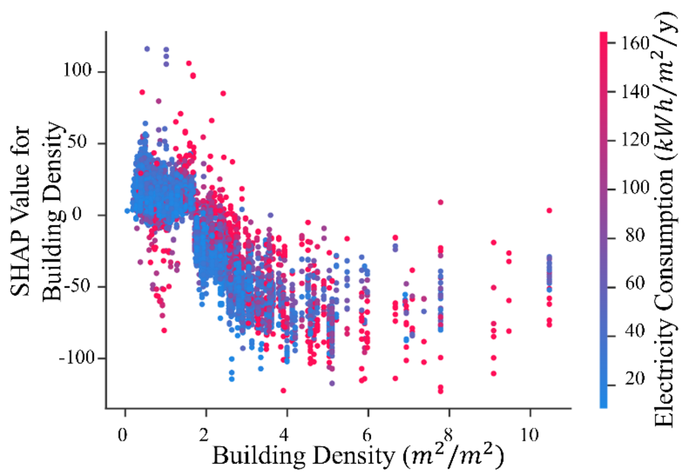
Figure 4. Interrelationships between electricity, gas, and water consumption



(a) Impact of Water Consumption and Building Type on Electricity Consumption



(b) Impact of Water Consumption and College Education on Electricity Consumption



(c) Impact of Building Density and Electricity Consumption on Gas Consumption

Figure 5. Impact of two features on the model output

We can then further study the impact of other features on electricity gas, and water consumption. Figure 5.a displays the results of water consumption as a feature to determine its impact on electricity consumption when the building type is changed from residential to non-residential. In the figure, the red points are the residential buildings and the blue points are the nonresidential buildings. For nonresidential buildings (blue points), by changing water consumption, SHAP values change more significantly than for residential buildings (red points). Essentially, the figure shows that water consumption has more of an impact on electricity consumption in nonresidential buildings (e.g., in office buildings). This is interesting because it suggests that water consumption not only captures elements of building occupancy (as mentioned before), it also captures some elements of consumption behavior (as building type has an impact on electricity consumption). Furthermore, Figure 5.b shows the impact of college education and water consumption on electricity consumption. Here, the figure shows that by changing water consumption, the SHAP values change more for the red points than the blue points, where the red points represent buildings with a higher percentage of people that have a college education. Therefore, in buildings with more educated people, higher water consumption patterns tend to translate into higher electricity consumption as well.

Finally, figure 5.c shows the impact of building density and electricity consumption on gas consumption. Despite some significant noise, we can see that SHAP values are mostly positive and do not vary significantly until a building density of 2 m^2/m^2. We can then see a significant decrease in SHAP value from approximately 0 to -75 for building densities ranging from 2 to 4 m^2/m^2, after which SHAP values become stable again. As mentioned, gas is often used as the primary source of heating, and higher building densities corresponds to larger buildings that tend to have lower *U*-values. We note that the rate of SHAP values reduction mostly stop after a building density of 4 m^2/m^2. Moreover, the noise suggests that building density is one out of many important variables that impact gas consumption (others include building insulation material, resident behavior, and so on (Haas, Auer, and Biermayr 1998; Guerra Santin, Itard, and Visscher 2009). In contrast, we do not observe any pattern between electricity consumption and building density using the gas consumption model.

To summarize the results, in this work, we found that:

- water and gas consumption tend to be strongly related to electricity consumption, but electricity consumption is slightly less sensitive to water and gas consumption. While this observation is interesting, we must acknowledge that it could be an artifice of the models since many important variables were not available, which also explains the lower accuracy of the water and gas consumption models. For example, data on building and water heating technology (i.e., electricity, gas, or steam) were not provided.
- water and gas consumption tend to be significantly interrelated, which is likely because gas is often used for water heating.
- electricity and water consumption are particularly interrelated in buildings with low gas consumption, again likely electricity is likely used for water heating in buildings with low gas consumption.
- electricity and gas consumption are interrelated up to a point after which they are not as interrelated, likely linked with the technology used for heating

- in buildings with low electricity consumption, water and gas consumption are negatively interrelated; these buildings tend to be warehouses and other building types where an increase in gas consumption per square meter has no impact on water consumption.
- water consumption has more of an impact on electricity consumption in nonresidential buildings; this suggests that water consumption not only acts as a proxy to building occupancy, but it also captures some elements of consumption behavior.
- after an initial threshold, increasing building density generally lowers gas consumption significantly, but only until up to a point after which the relationship plateaus; this suggests that building intensification is effective only after a certain threshold and only up to a specific building density.

5. Conclusion

Understanding the consumption patterns and interrelationships between energy and water consumption in buildings is paramount to be able to meet future demand in a world with a growing population. In this work, we looked at building energy and water consumption data in New York city and merged this data with the Primary Land Use Tax Lot Output (PLUTO) dataset as well as with other socio-economic data from the American Community Survey. The final dataset allowed us to develop three models: (1) electricity, (2) gas, and (3) water consumption.

The results show that the all three models perform relatively well, especially the electricity consumption model; gas is used heavily for heating and building heating technology was not available, and water consumption depends on other variables that were not available. Next, we determined the importance of each variable in each model, and we calculated how the variation in these features could affect energy and water consumption by using SHAP. Subsequently, we used SHAP to analyze how changing two features at the same time could affect the energy and water consumption. In the end, we listed seven major lessons learnt from this study such that electricity and water consumption are particularly interrelated in buildings with low gas consumption.

Arguably, the largest limitations of this study are the fact that only large-scale buildings are included and the specific results are not generalizable beyond the context of New York Cities. Ideally, future work would include small-scale buildings (ideally single detached homes) located in different climate zones to fully investigate interrelationships between electricity, gas, and water consumption. Nonetheless, the knowledge gained from the models and from the SHAP analysis can provide some insights to help develop data-driven policies, especially in cities with large-scale buildings, and evaluate different policy alternatives to find out how effectively energy and water consumption can be managed in a sustainable way.

6. Acknowledgments

This research is partly supported by the National Science Foundation (NSF) CAREER award #155173.

References

Ahmad, A.S., Hassan, M.Y., Abdullah, M.P., Rahman, H.A., Hussin, F., Abdullah, H., Saidur, R., 2014. A review on applications of ANN and SVM for building electrical energy consumption forecasting. Renew. Sustain. Energy Rev. 33, 102–109. https://doi.org/10.1016/j.rser.2014.01.069

Ahmad, N., Derrible, S., 2015. Evolution of Public Supply Water Withdrawal in the USA: A Network Approach. J. Ind. Ecol. 19, 321–330. https://doi.org/10.1111/jiec.12266

Ashouri, M., Haghighat, F., Fung, B.C.M., Yoshino, H., 2019. Development of a ranking procedure for energy performance evaluation of buildings based on occupant behavior. Energy and Buildings 183, 659–671. https://doi.org/10.1016/j.enbuild.2018.11.050

Athanassiadis, A., 2020. Urban Metabolism and Open Data: Opportunities and Challenges for Urban Resource Efficiency, in: Hawken, S., Han, H., Pettit, C. (Eds.), Open Cities | Open Data: Collaborative Cities in the Information Era. Springer Singapore, Singapore, pp. 177–196. https://doi.org/10.1007/978-981-13-6605-5_8

Badhrudeen, M., Naranjo, N., Movahedi, A., Derrible, S., 2020. Machine learning based tool for identifying errors in CAD to GIS converted data, in: Ha-Minh, C., Dao, D.V., Benboudjema, F., Derrible, S., Huynh, D.V.K., Tang, A.M. (Eds.), CIGOS 2019, Innovation for Sustainable Infrastructure. Springer Singapore, Singapore, pp. 1185–1190.

Becerik-gerber, B., Siddiqui, M., Brilakis, I., El-anwar, O., 2013. Civil Engineering Grand Challenges: Opportunities for Data Sensing, Information Analysis, and Knowledge Discovery. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000290

Bourdeau, M., Zhai, X. qiang, Nefzaoui, E., Guo, X., Chatellier, P., 2019. Modeling and forecasting building energy consumption: A review of data-driven techniques. Sustainable Cities and Society 48, 101533. https://doi.org/10.1016/j.scs.2019.101533

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System 785–794.

Christenson, M., Manz, H., Gyalistras, D., 2006. Climate warming impact on degree-days and building energy demand in Switzerland. Energy Convers. Manag. 47, 671–686.

Dalia, S., 2014. Residential energy consumption trends, main drivers and policies in Lithuania 35, 285–293. https://doi.org/10.1016/j.rser.2014.04.012

Deng, H., Fannon, D., Eckelman, M.J., 2018. Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata. Energy Build. 163, 34–43. https://doi.org/10.1016/j.enbuild.2017.12.031

Derrible, S., 2018. An approach to designing sustainable urban infrastructure. MRS Energy Sustain. 5, E15. https://doi.org/10.1557/mre.2018.14

Derrible, S., 2017. Urban infrastructure is not a tree: Integrating and decentralizing urban infrastructure systems. https://doi.org/10.1177/0265813516647063

Donkor, E.A., Mazzuchi, T.A., Soyer, R., Alan Roberson, J., 2014. Urban Water Demand Forecasting: Review of Methods and Models. J. Water Resour. Plan. Manag. 140, 146–159. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000314

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. Springer series in statistics New York.

Froemelt, A., Buffat, R., Hellweg, S., 2019. Machine learning based modeling of households: A regionalized bottom-up approach to investigate consumption-induced environmental impacts. Journal of Industrial Ecology jiec.12969. https://doi.org/10.1111/jiec.12969

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E 2013. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. ArXiv13096392 Stat.
Guerra Santin, O., Itard, L., Visscher, H., 2009. The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock. Energy Build. 41, 1223–1232. https://doi.org/10.1016/j.enbuild.2009.07.002
Haas, R., Auer, H., Biermayr, P., n.d. The impact of consumer behavior on residential energy demand for space heating 11.
Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction, 2nd ed. ed, Springer series in statistics. Springer, New York, NY.
Hobley, A., 2019. Will gas be gone in the United Kingdom (UK) by 2050? An impact assessment of urban heat decarbonisation and low emission vehicle uptake on future UK energy system scenarios. Renew. Energy 142, 695–705. https://doi.org/10.1016/j.renene.2019.04.052
Hsu, D., 2015a. Identifying key variables and interactions in statistical models of building energy consumption using regularization. Energy 83, 144–155. https://doi.org/10.1016/j.energy.2015.02.008
Hsu, D., 2015b. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. Applied Energy 160, 153–163. https://doi.org/10.1016/j.apenergy.2015.08.126
Jain, R.K., Smith, K.M., Culligan, P.J., Taylor, J.E., 2014. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. Appl. Energy 123, 168–178. https://doi.org/10.1016/j.apenergy.2014.02.057
Kashani, H., Movahedi, A., Morshedi, M.A., 2019. An agent-based simulation model to evaluate the response to seismic retrofit promotion policies. International Journal of Disaster Risk Reduction 33, 181–195. https://doi.org/10.1016/j.ijdrr.2018.10.004
Kavousian, A., Rajagopal, R., Fischer, M., 2013. Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. Energy 55, 184–194. https://doi.org/10.1016/j.energy.2013.03.086
Kenney, D.S., Goemans, C., Klein, R., Lowrey, J., Reidy, K., 2008. Residential water demand management: lessons from aurora, colorado 1 44. https://doi.org/10.1111/j.1752-1688.2007.00147.x
Kontokosta, C.E., Jain, R.K., 2015. Modeling the determinants of large-scale building water use: Implications for data-driven urban sustainability policy. Sustain. Cities Soc. 18, 44–55. https://doi.org/10.1016/j.scs.2015.05.007
Kontokosta, C.E., Tull, C., 2017. A data-driven predictive model of city-scale energy use in buildings. Applied Energy 197, 303–317. https://doi.org/10.1016/j.apenergy.2017.04.005
Kwok, S.S.K., Lee, E.W.M., 2011. A study of the importance of occupancy to building cooling load in prediction by intelligent approach. Energy Convers. Manag. 52, 2555–2564. https://doi.org/10.1016/j.enconman.2011.02.002
Lee, D., and Derrible, S. (2019). "Predicting Residential Water Consumption: Modeling Techniques and Data Perspectives". Journal of Water Resources Planning and Management, in press

Li, J., Yu, Z. (Jerry), Haghighat, F., Zhang, G., 2019. Development and improvement of occupant behavior models towards realistic building performance simulation: A review. Sustainable Cities and Society 50, 101685. https://doi.org/10.1016/j.scs.2019.101685

Li, K., Hu, C., Liu, G., Xue, W., 2015. Building's electricity consumption prediction using optimized artificial neural networks and principal component analysis. Energy Build. 108, 106–113. https://doi.org/10.1016/j.enbuild.2015.09.002

Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions.

Ma, J., Cheng, J.C.P., 2016. Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests. Applied Energy 183, 193–201. https://doi.org/10.1016/j.apenergy.2016.08.096

Mashhoodi, B., van Timmeren, A., 2018. Local determinants of household gas and electricity consumption in Randstad region, Netherlands: application of geographically weighted regression. Spat. Inf. Res. 26, 607–618. https://doi.org/10.1007/s41324-018-0203-1

Meng, T., Hsu, D., Han, A., 2017. Estimating energy savings from benchmarking policies in New York City. Energy 133, 415–423. https://doi.org/10.1016/j.energy.2017.05.148

Mohareb Eugene, Derrible Sybil, Peiravian Farideddin, 2016. Intersections of Jane Jacobs' Conditions for Diversity and Low-Carbon Urban Systems: A Look at Four Global Cities. J. Urban Plan. Dev. 142, 05015004. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000287

Papadopoulos, S., Azar, E., Woon, W.-L., Kontokosta, C.E., 2018. Evaluation of tree-based ensemble learning algorithms for building energy performance estimation. Journal of Building Performance Simulation 11, 322–332. https://doi.org/10.1080/19401493.2017.1354919

Papadopoulos, S., Kontokosta, C.E., 2019. Grading buildings on energy performance using city benchmarking data. Applied Energy 233–234, 244–253. https://doi.org/10.1016/j.apenergy.2018.10.053

Parsa, A.B., Kamal, K., Taghipour, H., Mohammadian, A. (Kouros), 2019a. Does security of neighborhoods affect non-mandatory trips? a copula-based joint multinomial-ordinal model of mode and trip distance choices. Presented at Transportation Research Board 98th Annual Meeting, Washington DC, United States

Parsa, A.B., Taghipour, H., Derrible, S., Mohammadian, A. (Kouros), 2019b. Real-time accident detection: Coping with imbalanced data. Accident Analysis & Prevention 129, 202–210. https://doi.org/10.1016/j.aap.2019.05.014

Pérez-Lombard, L., Ortiz, J., Pout, C., 2008. A review on buildings energy consumption information. Energy Build. 40, 394–398. https://doi.org/10.1016/j.enbuild.2007.03.007

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. Presented at the the 22nd ACM SIGKDD International Conference, ACM Press, San Francisco, California, USA, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

Robinson, C., Dilkina, B., Zhang, W., Guhathakurta, S., Brown, M.A., Pendyala, R.M., 2017. Machine learning approaches for estimating commercial building energy consumption 208, 889–904. https://doi.org/10.1016/j.apenergy.2017.09.060

Seyedzadeh, S., Pour Rahimian, F., Rastogi, P., Glesk, I., 2019. Tuning machine learning models for prediction of building energy loads. Sustainable Cities and Society 47, 101484. https://doi.org/10.1016/j.scs.2019.101484

Shapley, L.S., 1953. A value for n-person games. Contrib. Theory Games 2, 307–317.
Strumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions 647–665. https://doi.org/10.1007/s10115-013-0679-x
Suganthi, L., Samuel, A.A., 2012. Energy models for demand forecasting - A review. Renew. Sustain. Energy Rev. 16, 1223–1240. https://doi.org/10.1016/j.rser.2011.08.014
Tso, G.K.F., Yau, K.K.W., 2007. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy 32, 1761–1768. https://doi.org/10.1016/j.energy.2006.11.010
US Energy Information Administration. US household electricity report 2005. Tech. Rep.. Washington, DC: US EIA; 2005.
USEPA. (2017). "Inventory of U.S. Greenhouse Gas Emissions and Sinks" ⟨https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks 1990-2015⟩ (2017).
Von Grebmer, K., Ringler, C., Rosegrant, M.W., Olofinbiyi, T., Wiesmann, D., Fritschel, H., Badiane, O., Torero, M., Yohannes, Y., 2012. Global HunGer Index.
Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., Han, M., Zhao, X., 2018. A review of data-driven approaches for prediction and classification of building energy consumption. Renew. Sustain. Energy Rev. 82, 1027–1047. https://doi.org/10.1016/j.rser.2017.09.108
Wentz, E.A., Gober, P., 2007. Determinants of Small-Area Water Consumption for the City of Phoenix, Arizona 1849–1863. https://doi.org/10.1007/s11269-006-9133-0
Yu, Z. (Jerry), Haghighat, F., Fung, B.C.M., 2016. Advances and challenges in building engineering and data mining applications for energy-efficient communities. Sustainable Cities and Society 25, 33–38. https://doi.org/10.1016/j.scs.2015.12.001