Determinants of Electricity Consumption of Energy-Vulnerable Group Using Ensemble Gradient-Boosting Algorithm

Hyunsoo Kim, Youngwoo Kwon, and Yeol Choi

ARTICLE HISTORY
Received 3 November 2021
Revised 5 April 2022
Accepted 21 July 2022
Published Online 21 September 2022

KEYWORDS
Energy vulnerability
Energy equity
Residential electricity consumption
Environmental factors
Machine learning
Ensemble gradient boosting

1. Introduction
As energy transformation progresses, universal energy access and sustainable resource management are becoming primary policy goals in several cities. In 2020, Korea’s final energy consumption was 233 million TOE (Tonnage of oil equivalent (TOE), reaching its eighth-highest standing globally (Korea Energy Agency, 2019b). Residential and commercial buildings accounted for 41.7 million TOE of energy, or 17.5% of the nation’s total consumption (Korea Energy Agency, 2019b), whereas the numbers were significantly higher in large metropolitan cities. In Busan, where this study was conducted, the total electricity consumed by the residential and commercial sectors was 12,420 GWh, accounting for up to 60% of the total electricity consumption (Yun and Park, 2017). Although modern building technologies and efficient energy systems can effectively reduce the electricity requirements for daily living, several households in cities still rely on old and deteriorated infrastructure, physically restricting access to an efficient and safe energy environment. More importantly, this inaccessibility to a sustainable energy environment is likely to considerably burden vulnerable families with limited incomes and resources.
Universal energy access ensures the provision of affordable and clean energy to deprived households. As one of the UN-defined sustainable development goals, universal energy access emphasizes the collective effort to impose sustainable energy solutions, i.e., providing clean, safe, and affordable energy sources for basic living needs. Although the agenda has been discussed at global and national levels, it offers valuable policy goals at city- and regional- levels to provide energy solutions to those with limited income and access to sustainable energy systems. Universal energy access is essential in Busan, where the energy burden gap between regions and social groups has been a critical issue and is continuously expanding. According to a survey conducted by Statistics Korea (2019), the total household expenditure reduced by 2.2% compared to the previous year, whereas the expenditure on household energy use increased by 0.88%. The energy...
expenditure constitutes ~12% of the total expenditure in an average household. The burden of energy expenditure is far more significant for low-income families (households earning less than 1 million KRW), exceeding 21% of the total expenditure (Statistics Korea, 2019).

As the energy burden on vulnerable families in a city grows, an empirical understanding of how cities can identify and characterize energy-vulnerable groups can provide valuable guidance on future urban and welfare policies. Accordingly, this study aims to characterize the relationship between residential energy use and environmental factors (economic, climatic, structural, and regional characteristics) to compare average and vulnerable households. Specifically, it focuses on electricity consumption at the building level and the environmental factors. Owing to the absence of detailed socioeconomic data at the building level, data were aggregated at the parcel level and prepared for analysis. Each parcel’s appraised value (housing price) was used as an economic indicator as household income data were not available at the geographic scale of interest. This study builds on the conventional understanding of regional energy consumption to acquire comprehensive information with regard to vulnerable families and their residential environments. The expected outcome of this study can help provide the necessary information for urban and regional energy policies to establish energy equity in cities.

The remaining of the paper is organized as follows. Section 2 reviews the related research findings, focusing on the energy consumption of vulnerable households. Section 3 explains the data and analytical models used to investigate the factors associated with household energy consumption. Based on the results, the comparison between average and vulnerable households is prepared and presented in Section 4. Next, the discussion of some notable results and policy implications are presented in Section 5. Finally, Section 6 presents a brief overview of the study, its limitations, and the concluding remarks.

2. Literature Review and Theory

2.1 Determinants of Household Energy Consumption

The determinants of energy use are manifested in several forms based on the characteristics of the energy source and the government’s energy policy. Various studies have been conducted based on economic and non-economic perspectives. Recent studies have considered regional perspectives and found that energy consumption and consumer behavior are determined by socioeconomic, built environment, and climatic factors (Azevedo et al., 2016; Kim et al., 2017; Kim and Jung, 2019). However, the stochastic modeling of household energy consumption at the regional level has been challenging owing to uncertainties, e.g., an end-user’s consumption behavior and environmental characteristics, which can significantly influence the outcome. Extensive data and advanced modeling techniques have efficiently helped mitigate this issue, allowing for creative approaches to investigate the determinants of energy consumption at urban and regional scales.

A conventional understanding of household energy use is that physical factors govern the final consumption, e.g., the volume of space required for heating and cooling, energy efficiency of housing, and the number of household members and appliances. Although larger buildings, less efficient energy systems, and more end-users consume additional energy in most cases, the combination of width, height, and energy source characterizes the consumption level differently. Few studies have found that taller buildings use more electricity than smaller ones with the same floor area (Kontokosta and Tull, 2017; Lim et al., 2019), whereas taller buildings with smaller units consume less energy for heating needs (Lee and Chae, 2008; Kim, 2013; Eum et al., 2018). The effect of structural elements can be extended based on the energy efficiency of a building, which can further break down to building design, efficiency of energy systems, construction materials, and environmental characteristics, e.g., regional and local climates (Uihlein and Eder, 2010; Kontokosta and Tull, 2017). In particular, older buildings are less efficient, specifically in terms of heat conservation (Mastrucci et al., 2014), suggesting that the renovation of old buildings with efficient wall- and roof-types can help improve insulation (Uihlin and Eder, 2010).

Although the physical characteristics at the building level can help in fundamentally understanding the overall energy consumption, the relationship can be mainly influenced by the behavioral aspects of end-users. Uncertainties caused by user behavior factors can be compensated by incorporating socioeconomic and environmental factors on a regional scale, aggregated in blocks and neighborhood units. Among environmental factors, land use and development density have been the dominant determinants of the overall energy consumption. Development density is positively correlated with energy use intensity, and mixed-land use with additional commercial activities consumes more energy than residences only (Abbasabadi and Ashayeri, 2019). In terms of the type of residential land use, areas with a higher density of single-family households consume more energy than those with multi-family households (Ewing and Long, 2008; See et al., 2012), indicating that an increase in residential density can decrease the overall energy use in an urban environment. These results align with the statistical analysis findings in Korea, where population density is negatively correlated with electricity consumption (Jung et al., 2015; Lim et al., 2019). The study conducted by Li et al. (2018) found that high-density residential neighborhoods consume more electricity than low-density neighborhoods in the summer season owing to the urban heat island effect.

While the magnitude may vary, socioeconomic factors (sex, age, education level, housing tenure, and income) are considered to be significant determinants of final consumption (Abrahamse and Steg, 2011). Based on several studies, among the aforementioned factors, income has consistently been a significant driving force of energy consumption (Pachauri and Jiang, 2008; Kuusela et al., 2015; Mashhoodi and Timmeren, 2018). Household income is substantial because it is closely related to a household’s solvency for energy consumption and the ability to transition to clean and efficient energy systems. Ouedraogo (2006) found that income is the only significant factor among other socioeconomic factors,
which include frequency of cooking, size of households, level of education, and housing tenure, based on the choice of clean energy sources. The author stated that low-income households tend to resist transitioning to clean energy sources, regardless of the government’s energy policy, because their access to these energy sources is insufficient. While income is a significant factor in most cases, Azevedo et al. (2016) found that the effect of income is more evident in moderate-income households than that in high-income households. The authors suggested that the living environment condition and daily pattern of energy use are relatively consistent across moderate-income households, whereas high-income households tend to have a large deviation in energy use owing to their sufficient economic solvency.

The proportion of energy expenditure to income can be a critical indicator of energy affordability. Previous studies have found that the proportion of energy expenditure to income was relatively high among low-income households, indicating inequity in energy affordability (Jeong and Park, 2013; An et al., 2014; Jung et al., 2015). A statistical report of Korea demonstrated that most of the low-income households rely on low-efficiency energy sources, e.g., kerosene and coal (briquettes), and have limited access to more efficient sources, e.g., natural gas, owing to old energy infrastructure and limited income. Previous studies revealed that the household income to energy expenditure nexus varies across income segments (Boardman, 1991; Abrahamse and Steg, 2011; Jeong and Park, 2013; Kim and Lim, 2015). In the high-income quintile, the physical factors of housing had a significant but slight effect on energy consumption. In low-income brackets, physical factors, e.g., type of housing and year of construction, significantly affect the overall consumption, indicating that it is closely related to the deterioration level of housing and the insufficiency of their energy systems. These results indicate the need for differentiated alternative arrangements, depending on the income bracket, to effectively analyze energy affordability.

2.2 Characterization of the Energy-Vulnerable Households

As the study attempts to investigate the household energy consumption and draw a comparison between the average and vulnerable households, obtaining a set of guidelines to characterize the vulnerable is essential. While there are many ways to determine energy vulnerability, energy poverty is the most widely used term for characterizing the energy-vulnerable households. Energy poverty can be defined as households that cannot afford adequate energy consumption (Korea Energy Agency, 2019a). While governments and energy agencies provide different definitions and standards for energy poverty, energy expenditure and household income have been used as indicators to characterize them (Boardman, 1991; International Energy Agency, 2015).

While identifying energy poverty can rely on a single indicator, setting the minimum income or expenditure level necessary for basic living cannot encompass low-income families overburdened with energy expenditure. A comprehensive and inclusive approach is needed to identify families referred to as energy-vulnerable groups. One widely used approach is the ten percent rule index (TPRI), which suggests that households in the lower 30% income bracket whose energy expenditure exceeds 10% of the household income can be considered in the energy-vulnerable group (Boardman, 1991). A recent study (Villalobos et al., 2021) proposed an alternative to the monetary poverty identification method using the perception-based multidimensional energy poverty index and suggested that multiple definitions can help improve the identification of energy-vulnerable groups.

Currently, the Korean government considers four indicators to identify energy-vulnerable groups (Yun and Park, 2017). The first indicator (the most common) is the expenditure-to-income ratio of 10%, similar to the TPRI. The second indicator recognizes those who cannot afford the minimum energy expenditure, which constitutes 70% of the average energy expenditure per household member. The third indicator is based on the income and household composition, identifying households below 40% of the median income and having an elderly or infant family member. The fourth indicator considers the minimum affordability level by comparing the income after the energy cost (household income minus the energy expenditure) to the minimum standard (the minimum cost of living minus the minimum lighting and heating expenses). These indicators have been used to implement an array of energy welfare policies, including the energy voucher program. However, most welfare programs are not accessible to households requiring such policy support because the identification processes are complex and have not been sufficiently promoted in public.

Considering the need for improved methods to identify energy vulnerability, few academic studies have explored empirical approaches to define, measure, and characterize energy-vulnerable groups. Owing to difficulties in acquiring income and other socioeconomic data at the household level, most studies have relied on national surveys comprising various types of information that indicate the cost burden, behavior, and conception of energy use. Apparently, the cost burden on energy expenditure is substantial in income-vulnerable households, suggesting the need for detailed information on consumption behavior at the household level for low-income families (Jin et al., 2009; Jeong and Park, 2013). Lee (2015) surveyed low-income households in Cheongju, Korea, and found that most respondents experienced hardships with an energy cost burden. The study showed that 10% of the respondents experienced an electricity cutoff; 6% experienced a gas cutoff; 70% experienced reduced other basic expenditures, e.g., food and healthcare, to pay their energy bill; and 40% had not been able to afford energy cost for at least a month. A similar study conducted by Kim and Lim (2015) compared households with and without an elderly member and found that those with an elderly member exhibited a higher expenditure to income ratio, implying that they were more vulnerable to energy poverty.

Identifying energy vulnerability is a crucial step toward welfare policy for vulnerable groups, especially for low-income families with elderly, infant, and disabled members. The unaffordability of essential heating and cooling poses a health concern as it can lead...
to heat- and cold-related illnesses and fatality. Similarly, access to safe and efficient energy infrastructure can prevent fire and the inhalation of toxic gases from unsafe heating systems. Thus, the identification of energy vulnerability requires a multidimensional approach that considers socioeconomic and environmental factors.

3. Methodology

3.1 Data
This study used the monthly building electricity consumption data for Busan, Korea, provided by the Ministry of Land, Infrastructure, and Transport (MOLIT) (MOLIT, 2021a). The consumption data across 12 months (January–December, 2018) were collected and aggregated at the parcel level to link with the socioeconomic data geographically. A total of 86,024 parcels were prepared and used for analysis.

Currently, there is no clear definition for the energy-vulnerable group, and the definitions considered by the government rely on household income data, which are not available. Considering this limitation, the definition of the energy-vulnerable group used in this study used housing price as a proxy for income. Based on the indicators, households below the median price that consumed more electricity than the median level were identified as energy-vulnerable groups (Fig. 1).

The dependent variable is the electricity consumption per unit space (m$^2$), calculated as the total monthly consumption per parcel divided by the total floor space of the building(s). Considering the findings from previous studies on energy consumption at household and regional levels, economic, climate, building, and land use characteristics were selected and analyzed. For the economic characteristics, the housing price was used as a proxy

![Fig. 1. Definition of the Energy-Vulnerable Group](image)

Table 1. Variable Description

| Category                  | Variable                | Unit       | Description                                                                 | Source                                                                 |
|---------------------------|-------------------------|------------|------------------------------------------------------------------------------|------------------------------------------------------------------------|
| Dependent variable        | Electricity consumption | kWh/m$^2$  | Monthly average electricity consumption per unit area                        | Building data open portal (Ministry of Land (2021a), Infrastructure, and Transport (2021)) |
| Economic characteristics   | Housing price           | Price 1 million KRW/m$^2$ | Appraised housing price per unit area                                      | Korea Appraisal Board (2021)                                         |
| (EC)                      | Surface temperature     | °C         | Monthly average surface temperature measured with Landsat 7 ETM+ for each parcel | US Geological Survey                                                   |
| Climate characteristics   | Housing type            | Dummy      | 1 = apartment 0 = detached housing                                          | Building data open portal (Ministry of Land (2021b), Infrastructure, and Transport) |
| (CC)                      | Roof type               | Roof       | 1 = reinforced concrete 0 = others                                        |
| Building characteristics   | Wall type               | Wall       | 1 = reinforced concrete 0 = others                                        |
| (BC)                      | Household               | #          | Number of households                                                       |
|                           | Building age            | Years      | Building age in 2018                                                        |
|                           | Floors                  | #          | Number of overground floors of a building(s)                               |
| Land-use characteristics  | Land cover              | #          | Impervious surface index in the 50-m radius of a building                  | Environmental spatial information service (Ministry of Environment (2021)) |
| (LC)                      | Residential zone        | Dummy      | 1 = residential zone 0 = others                                           |
|                           | Commercial zone         | Commercial | 1 = Commercial zone 0 = others                                             |
|                           | Building density        | Cluster    | Local Moran’s index of buildings                                           | National spatial data infrastructure portal (Ministry of Land (2021c), Infrastructure, and Transport) |
variable to replace the level of household income. It represents the housing price per squared meter, annually appraised by Korea Appraisal Board (2021). To represent the climate characteristic, the study utilized satellite imagery (Landsat 7) provided by the United States Geological Survey (USGS, 2021) to estimate the monthly average of the surface temperature. This approach allowed access to more accurate records of the microclimate of the research subjects compared to the air temperature measurements from a few weather stations scattered across the city. The study considered several building characteristics obtained from the building data (MOLIT, 2021b), including the housing type, construction material, number of households in a building, building age, and the number of floors. Lastly, a set of land use characteristics includes land cover, land use by zoning, and building density. The land cover variable represents the proportion of impervious surfaces within 50 meters radius of each building, estimated using a land cover map (Ministry of Environment, 2021). The land use by zoning ordinance (MOLIT, 2021c) was utilized to determine whether a household is located on residential or commercial land. The local Moran’s Index was calculated to represent the relative building density of the households’ location. All variables were obtained and processed using ArcGIS (version 10.3). Table 1 lists the variables used in this analysis.

3.2 Analytical Model

The objective of the analysis was to produce an empirical relationship among the dependent variable, electricity consumption, and independent variables comprising various environmental factors. The stability of the variables was evaluated prior to the analysis, including the distribution, normality, and correlation among independent variables. A preliminary understanding of the dataset and its relationship suggested that nonlinear modeling could improve the fitting process and provide sufficient explanatory power. Thus, this study employed a semi-log model as a base model.

With technological advancement, the efficiency and accuracy of quantitative analyses have dramatically evolved in the recent decade. Among the advanced methodologies, machine learning algorithms have become a prominent candidate for handling a large amount of data and producing highly accurate predictability. Several algorithms are capable of solving regression problems, and they have been proved efficient in various research topics, including energy consumption and resource allocation and management in an urban environment (Kim, 2017; Abbasabadi and Ashayeri, 2019; Tirkolaee et al., 2020, 2021). Thus, a machine learning technique was employed in the modeling process. Specifically, a regression gradient boosting technique was used, which is known for its high performance in nonlinear regression modeling environments. The technique belongs to the ensemble model and is an algorithm combining several decision trees (Fig. 2).

The algorithm exhibits high accuracy by dissecting the singular tree of variation into multiple trees, creating an “ensemble” of weak predictive models. The main advantage of the model is its ability to decrease the variance of a single estimate by combining several estimates from an ensemble of different models, thereby exhibiting high stability (Chen and Guestrin, 2016; Wang et al., 2019). The ensemble model was classified into bagging and boosting depending on the weight designation in the learning processes, and the boosting algorithm was considered adequate for the analysis as it was optimized for error reduction. In addition to the predictive accuracy, the boosting algorithm’s flexible mechanism is another advantage, allowing for optimization on different loss functions and with several hyperparameters. It also has lenient data requirements, as it often works well with categorical and numerical values without data pre-processing and effectively handles missing data. On the other hand, one of the main disadvantages of the algorithm is its risk of overfitting caused by the extensive iteration process of error minimization (Chen and Guestrin, 2016; Bentéjac et al., 2021). To mediate the risk of overfitting, the final output was compared with the results of the base model (semi-log model).

This study used XGBoost developed by Chen and Guestrin (2016), which is known for its high performance among the libraries
of the ensemble algorithm. Compared to the traditional gradient boosting machines, the primary advantage of XGBoost is its efficiency, which comes from highly flexible model specification through hyperparameter settings and parallel processing that can effectively utilize computing resources. Another advantage of this method is that it produces intuitively interpretable outputs when applied to regression problems.

4. Analysis Results

4.1 Descriptive Statistics

Prior to modeling, the variables were evaluated to ensure stability. The descriptive statistics of the continuous and discrete variables are listed in Tables 2 and 3, respectively.

### Table 2. Descriptive Statistics of the Continuous Variables

| Category    | Mean | Standard deviation | Min | Max |
|-------------|------|--------------------|-----|-----|
| Dependent variable | **ELEC** | 17.77 | 15.73 | 0.00096 | 83.33 |
| Independent variable | **Economical characteristics** | Price | 1.06 | 0.59 | 0.05 | 11.30 |
| | **Climate characteristics** | Temperature | 24.47 | 2.10 | 11.27 | 34.86 |
| | **Building characteristics** | Household | 21.81 | 162.17 | 1 | 7,374 |
| | | Year | 29.84 | 10.61 | 1 | 89 |
| | | Floor | 3.82 | 3.72 | 1 | 72 |
| | | Area | 22.92 | 180.16 | 0.06 | 9,963.27 |
| | **Location characteristics** | Cover | 4.34 | 1.15 | 1 | 5 |
| | | Cluster | 0.29 | 0.2 | 0 | 1 |

### Table 3. Descriptive Statistics of the Dummy Variables

| Category      | Frequency | % | Mean |
|---------------|-----------|---|------|
| Dummy variables | Building characteristics | Type | 15,774 | 76.84 | 0.23 |
| | | 1 | 4,755 | 23.16 |
| | Roof | 0 | 980 | 4.77 | 0.95 |
| | | 1 | 19,549 | 95.23 |
| | Wall | 0 | 5,749 | 28.00 | 0.72 |
| | | 1 | 14,780 | 72.00 |
| | Location characteristics | Residential | 4,491 | 21.88 | 0.78 |
| | | 1 | 16,038 | 78.12 |
| | Commercial | 0 | 16,512 | 80.43 | 0.20 |
| | | 1 | 4,017 | 19.57 |

### Table 4. Semi-Log Model Results

| Category     | Average household | Energy-vulnerable households |
|--------------|-------------------|-----------------------------|
|              | Unstandardized coefficients | β | p | Unstandardized coefficients | β | p |
| Intercept    | B | SE | 0.177 | 2.779 *** | 0.111 | 0.000 |
| EC Price     | 0.464 *** | 0.012 | 0.304 | 0.000 | 0.124 *** | 0.046 | 0.038 | 0.007 |
| CC Temperature | -0.029 *** | 0.004 | -0.064 | 0.000 | -0.005 | 0.004 | -0.019 | 0.218 |
| BC Type      | -0.632 *** | 0.101 | -0.454 | 0.000 | -0.960 *** | 0.079 | -0.634 | 0.000 |
| BC Roof      | -0.233 *** | 0.030 | -0.057 | 0.000 | -0.067 | 0.053 | -0.015 | 0.207 |
| BC Wall      | -0.257 *** | 0.016 | -0.128 | 0.000 | -0.088 *** | 0.025 | -0.044 | 0.000 |
| BC Household | 0.000 | 0.000 | 0.029 | 0.380 | 0.000 *** | 0.000 | 0.118 | 0.001 |
| BC Year      | 0.010 *** | 0.001 | 0.117 | 0.000 | 0.005 *** | 0.001 | 0.070 | 0.000 |
| BC Floor     | 0.064 *** | 0.004 | 0.259 | 0.000 | 0.028 *** | 0.004 | 0.163 | 0.000 |
| BC Cover     | 0.076 *** | 0.007 | 0.091 | 0.000 | 0.005 | 0.007 | 0.012 | 0.440 |
| BC Residential | 0.031 | 0.043 | 0.013 | 0.475 | -0.045 | 0.052 | -0.035 | 0.384 |
| BC Commercial | 0.094 | 0.049 | 0.039 | 0.053 | 0.019 | 0.055 | 0.015 | 0.726 |
| BC Cluster   | 0.014 *** | 0.001 | 0.201 | 0.000 | 0.003 *** | 0.001 | 0.061 | 0.000 |
| R²           | 0.256 | Adj. R² | 0.256 | R² | 0.344 | Adj. R² | 0.342 |
The overall fitness of the model subjected to the energy-vulnerable households was better than that subjected to the average households, as observed by the r-squared values.

Housing price, housing type, wall type, building age, number of floors, and development density consistently exhibited significant results for both models, whereas the surface temperature, roof type, number of households, and impervious land cover were significant for one of the models. As expected, housing price was a significant factor for electricity consumption, regardless of energy vulnerability. The magnitude of the effect was larger for the average households, implying that the variability of electricity consumption associated with income is wider for the average households compared with the vulnerable households. According to the estimates, a unit increase (1 million KRW) in housing prices equates to an increase in consumption by 46% and 12% for average and energy-vulnerable households, respectively. Considering that the average consumption in the two groups was different, the difference in the estimated increase in consumption affected by the change in income was even larger between the groups. The housing type exhibited a negative effect, indicating that apartments use more electricity in both groups. Interestingly, the effect of housing type was larger than that of any other building type considered in the analysis. The effect was dominant in the case of the energy-vulnerable groups, estimating that detached housing would nearly halve consumption compared with apartment housing. Regarding the wall type, the reinforced concrete wall reduced consumption by 25% and 8% for the average and vulnerable households, respectively. Building age and the number of floors were found to be positive factors for electricity consumption in both cases, meaning that older and higher buildings use more electricity. The development density of the surroundings was the only consistent factor among land-use variables in the models. The effect was dominant in the case of the energy-vulnerable households. It was estimated that an increase of 1°C would decrease the electricity consumption of an average household by ~3%. Furthermore, the reinforced concrete wall would decrease the consumption by 23%, and a unit increase in the impervious surface in the surroundings would increase the consumption by 7% for the average household.

4.3 Ensemble-Boosting (via XGBoost) Model Results
As an alternative approach using a traditional regression technique, this study conducted another set of analyses using a machine learning algorithm, which was expected to produce improved fitness to characterize the relationship between electricity consumption and environmental characteristics. Although the performance of machine learning algorithms is enhanced by the computing power that enables articulated arithmetic operations following a predefined logic, it is primarily governed by determining the optimal specifications known as hyperparameters.

Similar to other algorithms, XGBoost possesses an array of hyperparameters that specify internal learning and prediction processes. To identify the optimal setting for the final model, this study tested different combinations of major hyperparameters, e.g., learning rate, maximum depth, gamma, and subsample rate. Specifically, the arrays of hyperparameters used to determine the optimal setting are as follows: learning rate = [0.01, 0.005, 0.001], maximum depth = [4, 5, 6, 7, 8, 9, 10], gamma = [0, 0.01, 0.1, 1], and subsample = [0.1, 0.2, 0.3, 0.7, 0.8, 0.9, 1.0]. Other minor hyperparameters used in the model optimization were determined using the xgb_model.fit package provided by the algorithm library. The r-squared values produced for the trained and predicted regressions were used as indicators to evaluate the fitness of the final model. Based on this process, the optimized hyperparameters were identified as follows: learning rate = 0.01,
maximum depth = 7, gamma = 0, and subsample = 0.7.

The summary of the final outputs of the average household is listed in Table 5, and the best performance indicated by the highest r-squared value (0.98 for the trained and 0.97 for the tested) was observed at ~9,000 iterations.

The results show the F-scores (feature importance) of the variables at each iteration, which outlines each variable’s weight assignment based on the extent of predictability reduction if the variable is excluded from the model. Although the values indicate the overall importance of the variables, it is difficult to determine the direction in which each independent variable influences the dependent variable. Therefore, this study employs the method proposed by Shapley and Roth (1988) to estimate the direction of the effects considering the SHAP value, which evaluates the impact of each feature. The SHAP value summarizes the magnitude and direction of each variable by estimating the change in average caused by the presence of a feature.

Observing the F-score of each independent variable, the housing price, number of floors, total floor area, and building density were important factors for predicting electricity consumption in Busan. This result was consistent with the output of the semi-log model. While slightly less in magnitude, the relative importance of building age and surface temperature was prominent as well.

Figure 3 shows the distribution of the SHAP values with each variable’s relative importance (F-score) in the final prediction. By observing the color scheme of the F-score and SHAP values, the outstanding magnitude and direction of each variable’s effect on the analysis of electricity consumption can be identified. The estimated values were tabulated for ease of comparison. According to the results, the increasing factors of electricity consumption for the average household were the number of floors, housing prices, building density, building age, and land cover. Similarly, the decreasing factors were surface temperature, wall type, and land use (commercial). Although the output does not indicate statistical significance, as in the semi-log model, the pattern of magnitude, direction, and distribution of relative importance portrays notable results that were not delivered in the semi-log model. In the case of the number of floors, the distribution of SHAP values and the magnitude of relative importance exhibited a clear pattern, indicating that the effect is likely to be associated with low-, middle-, and high-rise buildings. Another notable result was obtained with regard to the housing price, i.e., a long tail of low SHAP values toward the lower price end, indicating a significant reduction in electricity consumption in lower-priced housing.

A summary of the output details obtained using the algorithm for the energy-vulnerable group is listed in Table 6. The maximum predictability was achieved at ~5,000 iterations with the r-squared values of 0.99 and 0.92 for the trained and tested datasets, respectively.

In the case of the energy-vulnerable groups, the relative importance of the building density, surface temperature, and the number of households was prominent. Whereas the building age, wall type, and roof type were not as evident as in the previous case. As shown in Fig. 4, the distribution of the SHAP values for the vulnerable households is not as wide as that for the average household, suggesting that the variability of the effects is narrower for the energy-vulnerable households. The direction of the effects for the vulnerable households is not as clear as the output for the average households, as indicated by the distribution of

Fig. 3. SHAP Output for the Average Household
5. Discussion

Similar to the findings in previous household energy consumption literature, the income level was a dominant factor in energy consumption at the household level. The effect was statistically significant for both groups, whereas the magnitude and relative importance decreased in the energy-vulnerable groups owing to the reduction in sample size at the lower end. A long-tail of SHAP values toward the lower end for the average households was observed, indicating that the negative effect of income became more significant toward the lower end of the consumption.

Table 6. Summary of the XGBoost Model Output for the Energy-Vulnerable Household

| Variables    | n_estimators | 1,000       | 2,000       | 3,000       | 4,000       | 5,000       |
|--------------|--------------|-------------|-------------|-------------|-------------|-------------|
|              | n_estimators | Price       | Temperature | Type        | Roof        | Wall        | Household   | Year        | Floor       | Cover       | Residential | Commercial  | Cluster      |
|              |              | 0.1894 ± 0.0065 | 0.1937 ± 0.0058 | 0.1566 ± 0.0098 | 0.0005 ± 0.0001 | 0.0040 ± 0.0005 | 0.1578 ± 0.0033 | 0.1394 ± 0.0048 | 0.1499 ± 0.0095 | 0.0269 ± 0.0015 | 0.0049 ± 0.0004 | 0.0113 ± 0.0013 | 0.2130 ± 0.0061 | 0.858   | 0.959   | 0.986   | 0.995   | 0.998   |
|              |              | 0.2182 ± 0.0081 | 0.1943 ± 0.0060 | 0.1558 ± 0.0048 | 0.0004 ± 0.0001 | 0.0029 ± 0.0003 | 0.1542 ± 0.0037 | 0.1430 ± 0.0060 | 0.1380 ± 0.0024 | 0.0280 ± 0.0024 | 0.0039 ± 0.0001 | 0.0076 ± 0.0004 | 0.2366 ± 0.0108 | 0.861   | 0.903   | 0.925   |

Fig. 4. SHAP Output for the Energy-Vulnerable Household

F-scores. With regard to building density and housing price, the distribution of high (red) and low (blue) F-scores are more dispersed along the positive and negative directions than that in the previous case, implying that the effect of the variables weakened when the energy-vulnerable group was considered. Moreover, this result is consistent with the findings of the semi-log model. Another evidence is the discrete pattern in the F-score distribution, as observed for the housing type. This pattern can be interpreted as the effect of the housing type as a clear determinant for energy-vulnerable households, as observed in the semi-log model.
The result indicates the presence of the lowest-income groups who consume much less electricity than the rest when other factors are controlled. This finding requires further investigation, with additional divisions of income categories below the median in future studies.

Moreover, the effect of housing type was statistically significant in both groups; however, it was more prominent for the energy-vulnerable groups, indicating that vulnerable families living in detached households are more likely to be burdened with energy expenditure (Yun and Park, 2017). This result conflicts with the findings of previous studies that identified apartment buildings as an energy-efficient housing type. However, this result is likely to be associated with housing prices and the housing profile of the study area, given that most high-income households are apartments in Busan (Kim et al., 2019). Moreover, the majority of detached households in Busan are older and equipped with less efficient energy systems compared to apartments. The effects of the roof and wall types supported this result, given that apartments in cities are predominantly built with reinforced concrete.

The effect of surface temperature was different from the expectation as its relative importance and statistical significance decreased when the energy-vulnerable households were considered. This result may imply that vulnerable households are less sensitive to temperature than average households, requiring more concrete evidence by considering the seasonal variability of electricity consumption (Santamouris, 2016). Moreover, considering that climate-related illnesses occur at extreme temperatures, energy vulnerability can be better characterized by investigating the effect of temperature in the hottest and coldest months, suggesting the need for further investigation regarding seasonal differences with regard to the effect of temperature.

The building density was a minor but a positive factor of electricity consumption, indicating that households in more densely developed areas tend to use more energy. Although in previous studies, the heat island effect was the increasing factor of residential electricity consumption, this study cannot conclude on the effect because seasonal variability has not been considered. Instead, it can be assumed that families in denser areas, e.g., city centers, have a higher income level and solvency for energy consumption associated with higher consumption. While this study provided preliminary but limited information, the density-consumption relationship requires to be analyzed in detail with regard to income and consumption behavior based on location and urban density.

Finally, the performance of the machine-learning algorithm was outstanding based on the analytical methodologies used in the study. It is advantageous to generate comprehensive information using computing power and visual aid, e.g., the SHAP value chart. The study demonstrated that the computing power and accuracy of the algorithm could be powerful in monitoring household energy consumption and forecasting energy demand, which provides critical information for developing and managing an efficient distribution system. The accurate predictions of energy consumption can also help set viable visions and goals for making sustainable cities (Wang et al., 2019). Moreover, the traditional regression technique was efficient and straightforward in providing intuitive information regarding the relationship. The findings from the study conveyed the information necessary for making energy policies for vulnerable households. As the study found a set of influential factors of the energy consumption of vulnerable households, it can be used in developing energy welfare policies such as housing renovation (efficiency improvement) programs and energy voucher programs. The information can also be used for identifying the energy-vulnerable regions needing priorities in energy infrastructure management and planning.

6. Conclusions

In this study, the relationship between household electricity consumption and various environmental factors was characterized to compare average and energy-vulnerable households in Busan, Korea. This study has adopted a new approach of defining energy-vulnerable households that can be used at the regional level by considering the income and electricity expenditure levels. In addition to the traditional regression method (semi-log model), an advanced machine-learning technique (ensemble gradient boosting algorithm, XGboost) was employed to improve the explanatory power of the final results. A pair of relationships identifying the determinants of household electricity consumption in the average and energy-vulnerable households was produced using the analytical methods.

As presented in this study, the absence of household income data has been a critical limitation in regional energy consumption studies. Although this study relied on a proxy variable and arbitrary definition of the energy-vulnerable group, using actual income data can allow future studies to use a better definition and more sophisticated approach to characterize them. Furthermore, the study relied on the r-squared values to determine the overall fitness of each model, and the values cannot adequately compare the performance of the methodologies because of the complexities associated with the internal fitting processes in machine learning. With these limitations, this study expanded the conventional understanding of household energy consumption by constructing a microscale urban model based on monthly electricity consumption and detailed environmental factors at the parcel level. The findings provide the necessary information to achieve universal energy access in cities, including an improved way of identifying the target group for future energy welfare programs. Furthermore, future policy implications derived from this study can guide sustainable energy transition by leading energy infrastructure renovation and district and community energy programs.

Acknowledgments

This study was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2020R1A2C1013045).
 References

Abbasabadi N, Ashayeri M (2019) Urban energy use modeling methods and tools: A review and an outlook. Building & Environment 161:106270, DOI: 10.1016/j.buildenv.2019.106270

Abrahamse W, Steg L (2011) Factors related to household energy use and intention to reduce it: The role of psychological and socio-demographic variables. Human Ecology Review 18(1):30-40, https://www.jstor.org/stable/24707684

An YS, Kim KJ, Lee SI (2014) An empirical research on the difference of energy consumption according to the housing and regional characteristics of Seoul. Journal of Korea Planning Association 49(3):175-194, DOI: 10.17208/jkpa.2014.06.49.3.175

Azevedo JA, Chapman L, Muller CL (2016) Urban heat and residential electricity consumption: A preliminary study. Applied Geography 70:59-67, DOI: 10.1016/j.apgeog.2016.03.002

Bentéjac C, Csörgő A, Martínez-Muñoz GA (2021) Comparative analysis of gradient boosting algorithms. Artificial Intelligence Review 54:1937-1967, DOI: 10.1007/s10462-020-09896-5

Boardman B (1991) Fuel poverty: From cold homes to affordable warmth. UNKNO: London

Chen T, Guerstrin C (2016) Xgboost: A scalable tree boosting system Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, August 13-17, New York, NY, USA, DOI: 10.1145/2939672.2939785

Eun MR, Hong WH, Lee JA (2018) Deriving factors affecting energy usage for improving apartment energy consumption evaluation. The Journal of the Architectural Institute of Korea Structure & Construction 34(7):27-34, DOI: 10.5659/JAIK_SC.2018.34.7.27

Ewing R, Rong F (2008) The impact of urban form on US residential energy use. Housing Policy Debate 19(1):1-30, DOI: 10.1080/10511482.2008.9521624

Haebner GM, Hamilton I, Chalabi Z, Shipworth D, Oreszczyn T (2015) Explaining domestic energy consumption – the comparative contribution of building factors, socio-demographics, behaviours and attitudes. Applied Energy 159:589-600, DOI: 10.1016/j.apenergy.2015.09.028

International Energy Agency (2015) Energy Poverty: How to make modern energy access universal? OECD/IEA, Paris, https://www.iea.org/publications/energy-poverty-how-make-modern-energy-access-universal

Jeong YK, Park KS (2013) Analysis of energy consumption expenditure by household characteristics. KEEI Publication No.13-11, Korea Energy Economics Institute, Korea

Jin SH, Park EC, Hwang IC (2009) Research and analysis on the actual condition of energy consumption in low-income households. The Seoul Institute, 1-160

Jung J, Yi C, Lee S (2015) An integrative analysis of the factors affecting the household energy consumption in Seoul. Journal of Korea Planners Association 50(8):75-94, DOI: 10.17208/jkpa.2015.12.508.75

Kim MK (2013) An estimation model of residential building electricity consumption in Seoul. Seoul Studies 14(2):179-192

Kim KY (2017) Priority decision for energy selection using fuzzy TOPSIS. New & Renewable Energy 13(3):73-84, DOI: 10.7849/knsre.2017.9.13.3.073

Kim KJ, An Y, Lee S (2017) 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):253-267, DOI: 10.17208/jkpa.2017.10.52.5.233

Kim MS, Jung SW (2019) Analysis of the factors affecting energy consumption by the physical elements and the household features in residential buildings. Journal of the Korean Housing Association 30(1):13-25, DOI: 10.6107/JKHA.2019.30.1.013

Kim H, Lee G, Lee J, Choi Y (2019) Understanding the local impact of urban park plans and park typology on housing price: A case study of the Busan metropolitan region, Korea. Landscape & Urban Planning 184:1-11, DOI: 10.1016/j.landurbplan.2018.12.007

Kim HN, Lim MY (2015) The study of fuel poverty of the elderly households in South Korea. The Korean Association for Environmental Sociology 19(2):133-164

Kontokosta CE, Tull C (2017) A data-driven predictive model of city-scale energy use in buildings. Applied Energy 197:303-317, DOI: 10.1016/j.apenergy.2017.04.005

Korea Appraisal Board (2021) Annual appraisal of housing price retrieved from Open Data Portal, https://data.go.kr/data/3073746/fileData.do

Korea Energy Agency (2019a) 2019 knowing the energy right. Korea Energy Agency (KEA), Ulsan, Korea (in Korean)

Korea Energy Agency (2019b) 2019 energy statistics handbook. Korea Energy Agency (KEA), Ulsan, Korea

Kuusela P, Norros I, Weiss R, Sorasalmi T (2015) Practical lognormal framework for household energy consumption modeling. Energy & Buildings 108:223-235, DOI: 10.1016/j.enbuild.2015.09.008

Lee HJ (2015) Low-income households’ experiences and perception of home energy cost burdens in Cheongju, South Korea. Journal of the Korean Housing Association 26(5):79-87, DOI: 10.6107/JKHA.2015.26.5.079

Lee KH, Chae CU (2008) Estimation model of the energy consumption under the building exterior conditions in the apartment housing – focused on the maintenance stage. The Journal of the Architectural Institute of Korea Planning & Design 24(9):85-92

Li C, Song Y, Kaza N (2018) Urban form and household electricity consumption: A multilevel study. Energy & Buildings 158:181-193, DOI: 10.1016/j.enbuild.2017.10.007

Lim J, Huh I, Kang M (2019) A study on impact of apartment complex’s spatial characteristics on electricity consumption: Evidence of Seoul, Korea. Seoul Studies 20(3):19-37

Mashhoodi B, Van Timmeren A (2018) Local determinants of household gas and electricity consumption in Randstad region, Netherlands: Application of geographically weighted regression. Spatial Information Research 26(6):607-618, DOI: 10.1007/s41324-018-0203-1

Mastrucci A, Baume O, Stazi F, Leopold U (2014) Estimating energy savings for the residential building stock of an entire city: A GIS-based statistical downscaling approach applied to Rotterdam. Energy & Buildings 75:358-367, DOI: 10.1016/j.enbuild.2014.02.032

Ministry of Environment (2021) Land cover map retrieved from Environment Spatial Information Service, https://egis.me.go.kr/api/land.do

MOLIT (Ministry of Land, Infrastructure, and Transport) (2021a) Energy consumption data retrieved from Building Energy Information Services, https://data.go.kr/data/15054212/openaoi.do

MOLIT (Ministry of Land, Infrastructure, and Transport) (2021b) Building data retrieved from Open Data Portal, https://data.go.kr/data/15029175/openaoi.do

MOLIT (Ministry of Land, Infrastructure, and Transport) (2021c) Landuse
data retrieved from the National Spatial Data Infrastructure Portal, https://data.nsdi.go.kr/dataset/12659

Ouedraogo B (2006) Household energy preferences for cooking in urban Ouagadougou, Burkina Faso. *Energy Policy* 34(18):3787-3795, DOI: 10.1016/j.enpol.2005.09.006

Pachauri S, Jiang L. (2008) The household energy transition in India and China. *Energy Policy* 36(11):4022-4035, DOI: 10.1016/j.enpol.2008.06.016

Pérez-Lombard L, Ortiz J, Pout C (2008) A review on buildings energy consumption information. *Energy & Buildings* 40(3):394-398, DOI: 10.1016/j.enbuild.2007.03.007

Santamouris M (2016) Innovating to zero the building sector in Europe: Minimising the energy consumption, eradication of the energy poverty and mitigating the local climate change. *Solar Energy* 128:61-94, DOI: 10.1016/j.solener.2016.01.021

Seo HC, Hong WH, Nam GM (2012) Characteristics of electric-power use in residential building by family composition and their income level. *Journal of the Korean Housing Association* 23(6):31-38, DOI: 10.6107/JKHA.2012.23.6.031

Shapley LS, Roth AE (1988) The shapley value: Essays in honor of Lloyd S. Shapley. Cambridge University Press

Statistics Korea (2019) 2018 household survey data - Expenditure section

Tirkolaee EB, Abbasian P, Weber GW (2021) Sustainable fuzzy multi-trip location-routing problem for medical waste management during the COVID-19 outbreak. *Science of The Total Environment* 756(20):143607, DOI: 10.1016/j.scitotenv.2020.143607

Tirkolaee EB, Mahdavi I, Esfahani MMS, Weber GW (2020) A robust green location-allocation-inventory problem to design an urban waste management system under uncertainty. *Waste Management* 102:340-350, DOI: 10.1016/j.wasman.2019.10.038

Uihlein A, Eder P (2010) Policy options towards an energy efficient residential building stock in the EU-27. *Energy & Buildings* 42(6):791-798, DOI: 10.1016/j.enbuild.2009.11.016

USGS (2021) Landsat 7 satellite imagery retrieved from Earth Explorer, https://www.earthexplorer.usgs.gov

Villalobos C, Chávez C, Uribe A (2021) Energy poverty measures and the identification of the energy poor: A comparison between the utilitarian and capability-based approaches in Chile. *Energy Policy* 152:112146, DOI: 10.1016/j.enpol.2021.112146

Wang R, Lu S, Li Q (2019) Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings. *Sustainable Cities & Society* 49:101623, DOI: 10.1016/j.scs.2019.101623

Yun TY, Park KS (2017) Energy poverty estimation and analysis of energy consumption characteristics. KEEI Publication No. 17-02. Korea Energy Economics Institute, Ulsan, Korea