energies

Article

Investigating Primary Factors Affecting Electricity Consumption in Non-Residential Buildings Using a Data-Driven Approach

Sooyoun Cho [1], Jeehang Lee [2], Jumi Baek [1], Gi-Seok Kim [3] and Seung-Bok Leigh [1,*]

[1] Department of Architectural Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea; suyouncho@yonsei.ac.kr (S.C.); jumi100@yonsei.ac.kr (J.B.)

[2] Department of Bio and Brain Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Korea; jeehanglee@gmail.com

[3] Center for Sustainable Buildings, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea; giseok_kim@yonsei.ac.kr

* Correspondence: sbleigh@yonsei.ac.kr

Received: 6 September 2019; Accepted: 21 October 2019; Published: 24 October 2019

Abstract: Although the latest energy-efficient buildings use a large number of sensors and measuring instruments to predict consumption more accurately, it is generally not possible to identify which data are the most valuable or key for analysis among the tens of thousands of data points. This study selected the electric energy as a subset of total building energy consumption because it accounts for more than 65% of the total building energy consumption, and identified the variables that contribute to electric energy use. However, this study aimed to confirm data from a building using clustering in machine learning, instead of a calculation method from engineering simulation, to examine the variables that were identified and determine whether these variables had a strong correlation with energy consumption. Three different methods confirmed that the major variables related to electric energy consumption were significant. This research has significance because it was able to identify the factors in electric energy, accounting for more than half of the total building energy consumption, that had a major effect on energy consumption and revealed that these key variables alone, not the default values of many different items in simulation analysis, can ensure the reliable prediction of energy consumption.

Keywords: feature selection; prediction of energy consumption; electricity consumption; machine learning; non-residential buildings

1. Introduction

As carbon emissions and energy problems have become key issues in major cities worldwide [1], there has been an increasing awareness regarding the need for energy and carbon emission reductions in the urban development industry (including construction). After the Paris Climate Conference in 2015, in contrast to the arrangements under the Kyoto Protocol, cities around the world must now propose their own reduction goals, the target ranges must be expanded, and the established reduction goals must be actively presented. Therefore, the importance of energy reduction in the building industry, which accounts for at least 25% of all energy usage, must be recognized, and detailed stage-based energy reduction strategies have become an urgent requirement [2]. Energy consumption in the building sector has significantly increased in the last few decades. Energy is an essential part of our lives, and almost all things in some way are associated with electricity [3,4]. According to a report issued by the US Energy Information Administration (EIA), 28% growth in the global energy demand may occur until 2040 [5]. Because of improper usage, a tremendous amount of energy is wasted annually; thus, energy wastage can be avoided by the efficient utilization of energy. In addition, the third largest use of electricity among total energy sources [6] may have a direct impact on the aforementioned CO_2 savings. Most studies that have analyzed the energy performance of buildings have focused on an analysis or estimation of the energy consumption by each area of a building, on the quantitative prediction of energy savings, and on verification of the efficiency of heating and cooling systems. Rather than relying on actual, measured data, these studies have provided estimates by modeling the physical characteristics of buildings (e.g., building surfaces and volumes), the number and behavior of occupants, and the functioning of the heating/cooling systems [7].

An issue arising from such dynamic simulation is that the succession of estimates obtained from calculation formulas can cause an increasing discrepancy between the estimated and actual values of energy consumption, saving, and efficiency. This limitation has been pointed out in earlier studies [8]. Another drawback is that the reliability of the results of dynamic simulation depends on the skill of the expert who performs the simulation. Because an increasing number of buildings are operated automatically, the number of studies in which sets of data from buildings are analyzed by means of analytical tools initially devised for other fields is also increasing.

Machine learning algorithms have recently been used in academic fields including medicine, as well as industry [9]. Estimates obtained through machine learning have contributed to resolving the discrepancy between estimated and measured values [10]. In addition, such algorithms provide objective results that are independent of the operator's skills (i.e., the same estimate is produced by any operator when the same process is performed). This research aims to identify the major variables that affect the energy consumption in highly energy-demanding buildings, using the clustering method. With respect to the prediction of major variables through the machine learning algorithm, an estimate that takes all variables of the dynamic simulation and all parameters into account can be expected to eliminate the discrepancy between estimated and actual values which arises from the use of calculated values. It can also be expected to enhance the reliability of such estimates (regardless of the operator performing the analysis).

This study focused on electric energy consumption in office buildings [11], which has been increasing [12], and identified the factors contributing to electric energy consumption. Energy consumption prediction based on traditional dynamic simulation methods produces values predicted by default values, which cannot exactly reflect the physical conditions prior to a building's design or construction, e.g., building area, window-to-wall ratio, envelope thermal efficiency, and number of occupants. This study identified which factors in the measured data of electric energy consumed by users in an actual building had the largest effect on electric energy consumption. The variables that could predict electric energy consumption most accurately were identified among other data measured from an office building using data-driven clustering, and it was confirmed that the combinations of these major variables predicted an electric energy consumption similar to that which occurred when all kinds of information were used together. Based on the assumption that only the electric energy of the building is used, it is sufficient to measure only the major sensors and measuring instruments related to the electric energy consumption derived by clustering. This is expected to greatly reduce the cost and time in conventional energy or consumption analysis. The key significance of this study is that a comparison of the results predicted using all information and the results predicted using only major variables to test the research methods showed no significant difference.

Two additional procedures have also been adopted for validation of the results to improve the reliability of the study. The significance of our study lies in the ability of the proposed method to identify the fundamental cause of excessive energy consumption in buildings with diverse functions, not only in office buildings, which are the primary subject of the study.

The rest of the paper is organized as follows. Related work is given in Section 2, and a detailed explanation of the limitations of previous studies that have analyzed the building energy performance using engineering methods for energy building analyses is first presented. Then, after a review of the recent literature on building energy analysis conducted by machine learning, the novel approach used in this study is presented. In Section 3, for 11 non-residential office buildings located across South Korea, the clustering analysis method of machine learning is employed to identify the major variables that affect the energy usage. Section 4 shows the results of the main variables of building electricity energy usage with features selected by machine learning. Section 5 validates the prediction results using only the main variables and the prediction results using all variables in two different ways. The paper's conclusion and future work are discussed in Section 6.

2. Related Works

Methodologies for building energy analyses have been developed over the last 50 years. Because the results obtained by these methods vary, based on their suitability, accuracy, sensitivity, and purpose [13], it is important to identify the correct method for research purposes, target applications, and the environment. Over the last 20 years, steady-state data-based simulation and dynamic simulation-based engineering analysis methods have typically been used for building energy analyses in Korea [14]. These methods estimate the energy usage from the environmental conditions of a building and physical data such as the building envelope, building heat-cooling-ventilation system, and a thermodynamic equation. They also predict the building's energy consumption and performance of the equipment system [15]. Dynamic simulation programs such as BLAST, ESP, EnergyPlus, ESP-r, and TRNSYS are widely used, as they are accurate and can be utilized without restrictions of purpose and usage.

However, there are disadvantages to using these programs, as usage becomes more complex as the number of variables increases. In contrast, data-driven (machine learning-based) prediction method approaches in conjunction with machine learning techniques use data covering the building's entire history to predict the amount of energy that will be used in the future under detailed, but limited, conditions. However, the applicability of such methods is often hindered by malfunctions and defects in the equipment system [16]. Popular algorithms that are used for such methods include linear regression, artificial neural networks, and support vector regression [17]. It has been reported that the prediction accuracy can be improved using ensemble-algorithms, which have an improved accuracy and reliability compared with a single-algorithm prediction model under the same building energy usage and system performance conditions.

2.1. Engineering Analysis Method and Its Threshold

Dynamic simulations, which are widely used in industry as well as in academia, use a calculation method that treats heat movement by considering indoor and outdoor conditions and assumes that heat movement occurring indoors will vary with time. BLAST, EnergyPlus, ESP-r, and TRNSYS are some of the main dynamic simulation applications [18]. Although the calculations are relatively accurate, and these methods are widely applicable, they do have limitations because of the difference between the calculations performed in the simulation and the actual measurements pertaining to the building. This occurs because of the increasing amount of data and variables and because there is a drastic difference in the results, depending on the user's understanding of the variables (i.e., user bias) [19]. The energy performance gap (EPG) is a representative example of the difference between dynamic simulation results and the actual energy consumption in buildings. Table 1 indicates the differences between the energy consumption estimated by simulation and the actual consumption measured after the building was constructed.

Regarding the characteristics of the majority of engineering analysis methods used in studies and reports published outside Korea, a potential common disadvantage is that the results may vary according to the user, as well as by time and money requirements. Another limitation seen in international settings is related to the supply of low-energy buildings and sustainable buildings. Research results for the initial design concerning energy usage and the performance of key technologies used in these types of buildings are different from the results obtained after the buildings are actually built. This phenomenon is referred to as the EPG [20]. In some of these studies, the difference between the values predicted by the engineering analysis method and the values measured in the actual building differ significantly [21], suggesting that the fundamental cause of these EPGs differs, according to the design–construction–operation method, as well as occupants [22]. Moreover, the difference in the results also results from errors in the energy modeling program used in the engineering analysis method. The second cause is the construction quality of the actual building under study. The EPG often occurs as a result of not having a detailed understanding regarding the implementation of certain technologies (e.g., envelope construction, window sealing, and ventilation system) regarding construction, which is due to a lack of awareness regarding eco-friendly buildings and a dearth of skilled workers. Furthermore, EPGs also occur when a detailed performance inspection is not properly conducted during the construction process, or when the construction is completed, but the performance was not fully guaranteed. Lastly, EPGs are caused by a delay in verification techniques and performance measurements for each specific technology once the building is constructed.

Table 1. Published papers on the building energy performance.

Author	Building Use	Target/Energy Source	Evaluations
Yu et al. [16]	Domestic	Gas, Electricity	Data Mining
Wilde et al. [20]	Domestic	Gas, Electricity	Monitoring
Menezes et al. [21]	Non-Domestic	Ventilation	Post-Occupancy Evaluation (POE)
Olivia et al. [22]	Hospital, School	Indoor Comfort	POE, Monitoring
Choi et al. [23]	Non-Domestic	Ventilation	POE
Hossein et al. [24]	Non-Domestic	Electricity	POE
Salehi et al. [25]	Non-Domestic	Ventilation	Dynamic Simulation
Niu et al. [26]	Domestic	Ventilation	POE
Herrando et al. [27]	Non-Domestic	Ventilation	Dynamic Simulation
Min et al. [28]	Non-Domestic	Air Handling Unit	Facility Management Review

In summary, because the energy consumption analysis or prediction studies traditionally used in the building field have been performed by experts who are highly experienced in the field and are time- and labor-intensive, it would be difficult to rely on the analysis results without expertise. Like the aforementioned EPG results, this research found a large difference in the reliability and resulting values of the analyzed results. For this reason, analysis methods have recently been developed that identify, analyze, and diagnose phenomena using data from only the actual building, instead of results calculated by default values embedded in the program.

2.2. Data-Driven Analysis Method

In recent years, many studies employing building energy prediction and analysis methods have incorporated machine learning, which helps to address the limitations of engineering analysis. This involves the extensive use of building control and energy management systems (BEMS) and BEMS data, whose use is linked to increases in the amount of energy data derived from buildings and the number of analysis variables that must be considered. Machine learning is a technique employed for discovering models, patterns, and rules within data [9]. It can be used to extract previously unknown but useful knowledge, by recognizing complex patterns in data and building statistical models based on those findings [29]. As a result, it is utilized not only in engineering, but also in various other fields, such as medicine [30]. There are also studies that have applied machine learning to the field of building energy. Building energy forecasting methodologies based on machine learning use historical data to predict the future energy use under specific constraints. These methodologies also involve analyzing building energy systems to detect malfunctions or defects [31].

A single prediction model is run using one learning algorithm and training a monolithic model throughout the model development process [32]. Various machine learning algorithms, such as Multiple Linear Regression (MLR) [33], Artificial Neural Networks (ANNs) [34], decision trees [35], and Support Vector Regression (SVR) [36], have been introduced to building energy prediction and have provided promising prediction results during the past two decades. These algorithms can process continuous real-time data derived from buildings to predict the energy use and the performance of certain equipment systems, or to detect malfunctions. Figure 1 shows a schematic diagram of a typical single prediction model. The historical data recorded from the building are divided into training (60%), verification (20%), and testing (20%) sets, after removing outliers during preprocessing. The procedures are repeated until the final result is obtained.

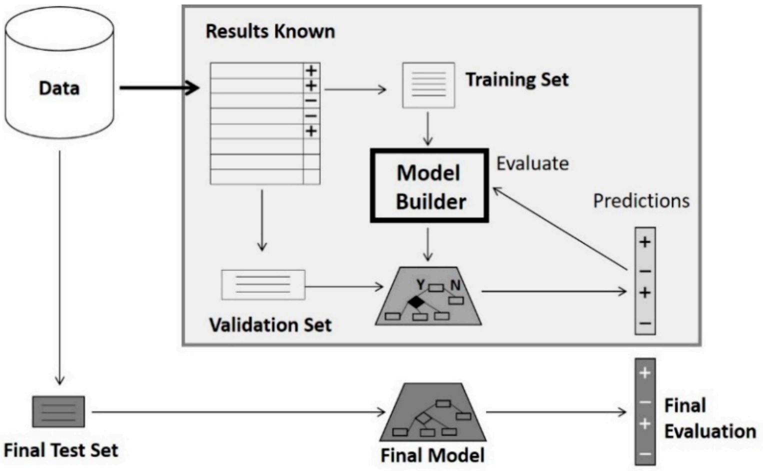


Figure 1. The process of machine learning analysis. Source: https://blog.algorithmia.com/page/50 [37].

2.2.1. Analysis of Building Energy Consumption Using Machine Learning

Although there have not been as many studies utilizing machine learning in the building energy sector as there have been in other fields of research, there has been much research on building energy consumption predictions and analytical studies that utilize machine learning. Studies by Paudel et al., [38] Yildiz et al., [39] and Rahman et al. [40] used machine learning to predict the energy consumption in buildings. Table 2 shows the target buildings and evaluation indexes for predicting building energy use. Although similarities exist with respect to utilizing actual energy usage data, predicting the usage volume, and presenting the mean absolute percentage error (MAPE) and RMSE as evaluation indicators, there is a lack of research on the analysis of the causes that induce a specific energy usage, which is one of the objectives of this study. Moreover, there has not been specific and persuasive evidence on the standard by which the variables of datasets were selected and utilized by the machine learning algorithms. The variables selected for predicting the energy use of buildings should not be based on researchers' experience; instead, they should be chosen to ensure the reliability of the results and consistency of the research process using statistical methods. Therefore, the use of machine learning for variable selection remains very important.

Table 2. Prediction of building energy consumption through machine learning.

Authors	Research Topic	Target Building	Algorithm	Evaluation Indices
Paudel et al. [38]	Prediction of building energy consumption All data vs. relevant data compared	Low-energy buildings	SVR	RMSE, R^2
Yildiz et al. [39]	Building electricity consumption prediction	Commercial buildings, educational facilities	ANN, SVR, Regression tree	MAPE, RMSE, R^2
Rahman et al. [40]	Prediction of building's fuel use Hourly data used over one year	Office buildings, supermarkets, restaurants	MLR, ANN, SVR, GP	RMSE
Moon et al. [41]	Prediction of building's electricity use	University buildings	ANN, SVR	MAPE, RMSE
Seong et al. [42]	Energy optimization model for buildings	Office buildings	ANN	MBE, CVRMSE
Support Vector Regression (SVR) Artificial Neural Network (ANN) Multi Liner Regression (MLR)		Root Mean Square Error (RMSE) Mean Absolute Percentage Error (MAPE) Coefficient of Variation of the Root Mean Square Error (CVRMSE)		

2.2.2. Analysis of Building Energy Consumption by the Clustering Method

Clustering is a common technique in unsupervised learning, which is the machine learning method of identifying a specific pattern in data without the ground level. Clustering refers to a group of methodologies that classify data with similar attributes into a number of groups [29]. Few studies have used the clustering analysis method for the analysis of building energy [43–47]. Most of them have focused on the analysis of consumption patterns of specific buildings. However, they only mention the possible causes of transmission and distribution, instead of providing details through experimental results, indicating that they only used the method with the aim of studying the methodology itself.

Table 3 outlines the results of recent studies that employed the clustering method and classifies them in terms of the algorithms used and their evaluation indices. Recent studies on clustering that are related to this study tend to use centroid-based K-means algorithms, and most studies were not studies looking for key variables that contribute to energy consumption, but were only used to analyze building energy consumption patterns with clustering algorithms.

Table 3. Building energy analysis using the clustering method.

Authors	Research Topic	Target Building	Algorithm	Evaluation Indices
Naganathan et al. [43]	Identifying the loss of energy during transmission and distribution	105 buildings	K-means	-
Ko et al. [44]	Improving the estimation accuracy of building energy consumption	Office buildings	K-means	R^2
Yang et al. [45]	Utilized K-shape clustering for analyzing building energy use patterns	Educational facilities	K-Shape SVR	RMS
Moon et al. [46]	Analysis of cooling and heating energy consumption patterns in office buildings	Office buildings	K-medoids	-
Hwang et al. [47]	Building energy demand predictions using hierarchical clustering	No information on target buildings	Hierarchical Clustering	APE, R^2

To summarize the literature review, there are studies that have used machine learning methods to predict the energy consumption in buildings, and some studies have used clustering methods to analyze energy usage patterns. However, there are no studies that have used machine learning and clustering at the same time to determine the cause of energy consumption.

3. Approach

The purpose of this study was to identify the power consumption of buildings and determine the variables that contribute to their use. This study confirms that key variables obtained using clustering in machine learning strongly correlate with the energy consumption of buildings. For this purpose, it is important to first identify building energy consumption patterns. This was done based on a total of 16 variables related to building energy usage, in addition to the electricity consumption of the target building. Then, the variables that best define the energy consumption patterns were identified. The major elements affecting consumption are not identified by simulation, or by the analyst's experience, but through a machine learning process. The structure of the research work is described below and Figure 2 shows simplify the process of this study from the number one to six.

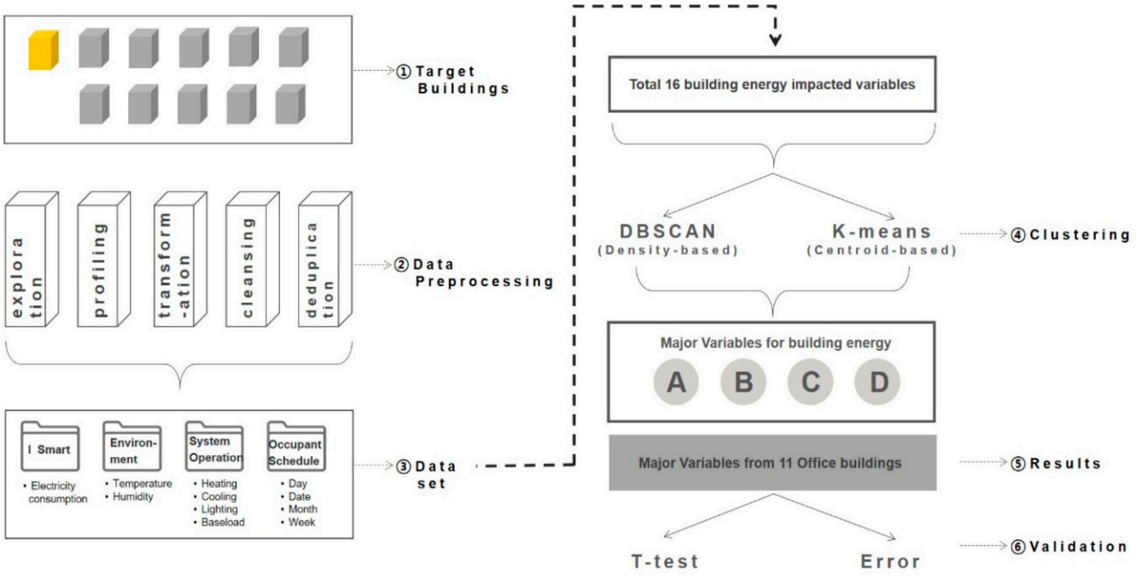


Figure 2. Schematic representation of the research process.

(1) This study targets 11 office buildings in eight different regions of South Korea. Their heating and cooling systems use electricity. The 11 buildings all have different areas, building designs, and heating and cooling systems, and are used for different purposes. In the data analysis, the physical qualities are referred to as nominal variables. In this study, only the continuous data that change in real time as a result of the actions of the occupants were used for the analysis. Four categories of variables were identified for the analysis, i.e., the amount of electric energy used by the building, external environmental data, building system data, and occupancy schedules. In total, 16 continuous variables were selected;

(2) Before analyzing the large amount of data used in machine learning, preprocessing is a necessary step that prepares data by processing missing values, removing noise, or removing outliers. In data preprocessing, cleaning is a process that fills in or removes missing values, corrects noisy data, identifies outliers, and therefore ensures the consistency of data outcomes [48];

(3) The data set in this research contains time variables (such as the year, month, day, day of the week, and hour), external environment variables (including temperature, humidity, sunlight, airflow, wind speed, and soil temperature), and energy variables that track the energy consumed during the building operating hours (in 2016, use of electricity consumption hourly data of 11 offices) [49];

(4) To determine the most important variables, unsupervised learning, which does not require a ground truth, was used. This involved using clustering to extract the most important variables that control the building energy usage and cluster similar variables together;

(5) In addition, using the *t*-test, we aimed to verify the validity of the variables, derived from machine learning, that affect the building energy consumption;

(6) For each target building, the first 16 variables related to energy consumption, and five variables derived from machine learning, were compared with the same verification logic.

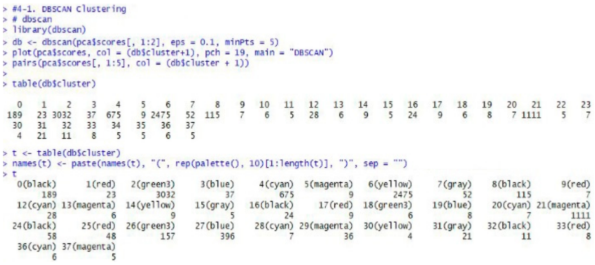
As stated in Section 3, this study aims to use energy data derived from a building and identify important variables that control most of the energy consumption. Two representative clustering methodologies used for this purpose are the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and K-means.

3.1. Clustering: DBSCAN and K-Means

Density-based spatial clustering of applications with noise, one of the most widely used density-based clustering methods, explores core objects that have a high density in a neighborhood [50]. Centroid-based clustering methods of K-means predetermine the K value and the number of clusters, and then assign each datum to one of the K clusters to minimize dispersion. Centroid-based algorithms have the advantages of clustering many data points quickly and easily, and spherical data have a relatively better reliability [28,50].

Table 4 shows a sequential list of operations performed through DBSCAN and K-means using the open-source program R 3.6.1. [51] To summarize the clustering process for extracting key variables, first, similar characteristics (density center, distance-based) in the various data were assessed to create the first cluster. Although N clusters were formed, it was impossible to determine which clusters were meaningful. Then, results were extracted only for clusters containing a large amount of data. Based on these results, the data were classified into high-energy-consumption and low-energy-consumption data clusters. Lastly, for the two clusters with large amounts of high-energy-consumption data, box plots were used to confirm which of the 16 variables related to building energy were the most important.

Table 4. Clustering process.

<table>
<tr><th>Order</th><th>R-Code and Clusters</th><th>Contents</th></tr>
<tr><td>1</td><td>
<pre>
> #4-1. DBSCAN Clustering
> # dbscan
> library(dbscan)
> db <- dbscan(pca$scores[, 1:2], eps = 0.1, minPts = 5)
> plot(pca$scores, col = (db$cluster+1), pch = 19, main = "DBSCAN")
> pairs(pca$scores[, 1:5], col = (db$cluster + 1))
>
> table(db$cluster)

   0   1   2   3   4   5   6   7   8   9   10   11   12   13   14   15   16   17   18   19   20   21   22   23
 189   23 3032   37 675   9 2475   52 115   7   6   5   28   6   9   5   24   9   6   8   7 1111   5   7
  30   31   32   33   34   35   36   37
   4   21   11   8   5   5   6   5

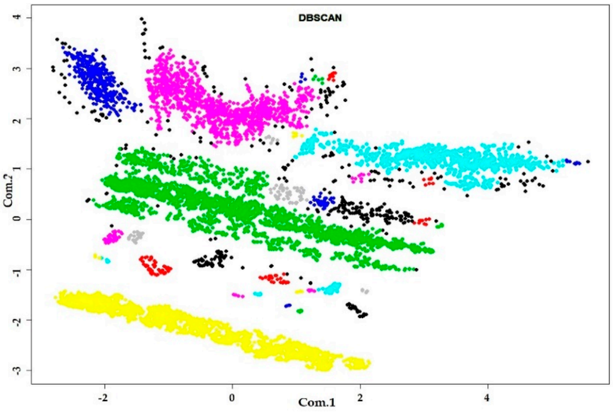
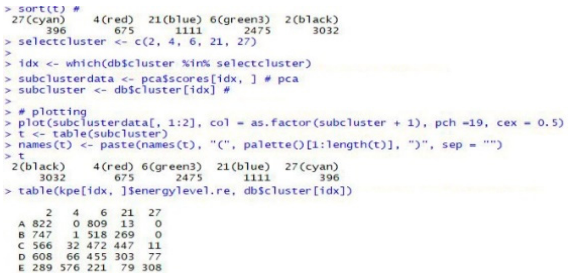
> t <- table(db$cluster)
> names(t) <- paste(names(t), "(", rep(palette(), 10)[1:length(t)], ")", sep = "")
> t
   0(black)   1(red)   2(green3)   3(blue)   4(cyan) 5(magenta) 6(yellow)   7(gray)   8(black)   9(red)
        189   23   3032   37   675   9   2475   52   115   7
   12(cyan) 13(magenta) 14(yellow)   15(gray)   16(black)   17(red) 18(green3)   19(blue)   20(cyan) 21(magenta)
         28   6   9   5   24   9   6   8   7   1111
  24(black)   25(red) 26(green3)   27(blue)   28(cyan) 29(magenta) 30(yellow)   31(gray)   32(black)   33(red)
         58   48   157   396   7   36   4   21   11   8
   36(cyan) 37(magenta)
          6   5
</pre></td><td>First cluster is formed based on similar characteristics in the data (density center). It is unable to distinguish which cluster data are significant among n clusters through seven repeated colors.</td></tr>
<tr><td>2</td><td>
</td><td>Results of re-executed DBSCAN code for n clusters sorted by the amount of data in the clusters in descending order.</td></tr>
</table>

Table 4. *Cont.*

Order	R-Code and Clusters	Contents

3	
	`> sort(t) #`
	` 27(cyan)    4(red)  21(blue) 6(green3)  2(black)`
	`      396       675      1111      2475      3032`
	`> selectcluster <- c(2, 4, 6, 21, 27)`
	`>`
	`> idx <- which(db$cluster %in% selectcluster)`
	`> subclusterdata <- pca$scores[idx, ] # pca`
	`> subcluster <- db$cluster[idx] #`
	`>`
	`> # plotting`
	`> plot(subclusterdata[, 1:2], col = as.factor(subcluster + 1), pch =19, cex = 0.5)`
	`> t <- table(subcluster)`
	`> names(t) <- paste(names(t), "(", palette()[1:length(t)], ")", sep = "")`
	`> t`
	` 2(black)    4(red) 6(green3)  21(blue)  27(cyan)`
	`     3032       675      2475      1111       396`
	`> table(kpe[idx, ]$energylevel.re, db$cluster[idx])`
	`       2   4   6  21  27`
	`  A 822   0 809  13   0`
	`  B 747   1 518 269   0`
	`  C 566  32 472 447  11`
	`  D 608  66 455 303  77`
	`  E 289 576 221  79 308`
		Extraction of results only for clusters allocating a large amount of data; re-clustering is performed based on these results for energy consumption (real-time consumption) with a relatively large amount of data.

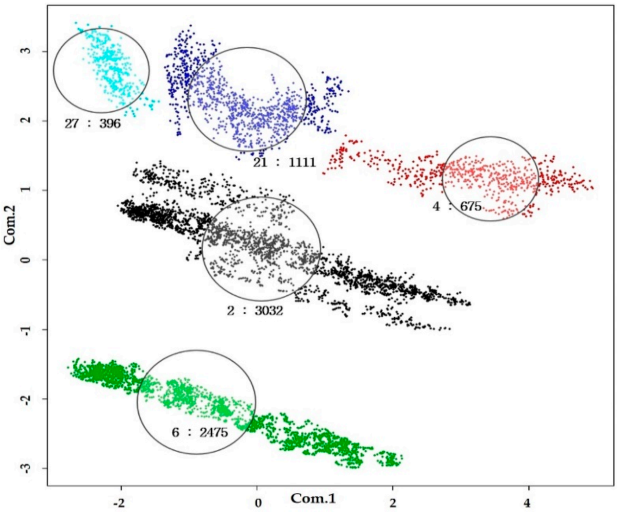
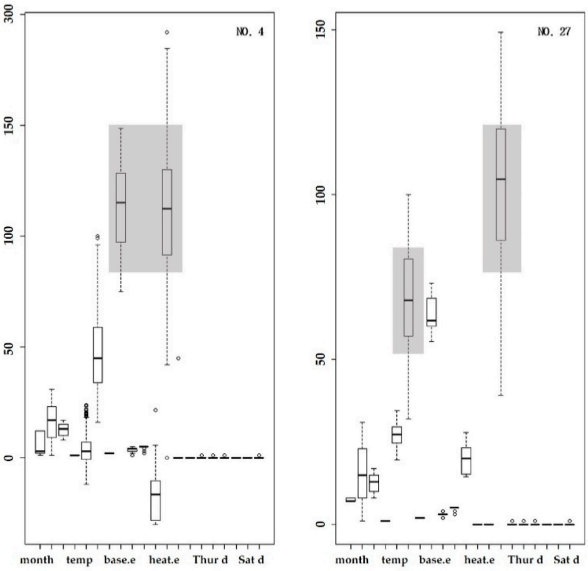
4		Formation of five clusters with large amounts of data. In particular, the red cluster (4) and the sky-blue cluster (27) are classified as high energy consumption clusters (clusters with assigned high consumption level E).
5		The important variables of Cluster 4 and Cluster 27 are displayed in a box plot. Among the 16 energy variables, in the red cluster, baseload and heating energy were determined to be important variables, while in the sky-blue cluster, cooling energy and humidity were determined to be important variables.

Figure 3 shows the results after clustering using the density-based DBSCAN code to derive the clusters for the 11 buildings, where two parameters were required for DBSCAN: epsilon (ε) and the minimum number of points required to form a cluster (minPts). Epsilon is a distance parameter that defines the radius used to search for nearby neighbors. In Figure 4, the distance-based K-means results are shown. The consumption data were categorized based on energy usage and divided into five classes (energy consumption level A to E), with A representing the lowest energy consumption and E representing the highest. The energy consumption cluster results (e.g., size and shape, density of clusters, and number) of the buildings for each region are different. This occurs because the characteristics of the 11 regions' building envelope, heating and cooling systems, and building use are different.

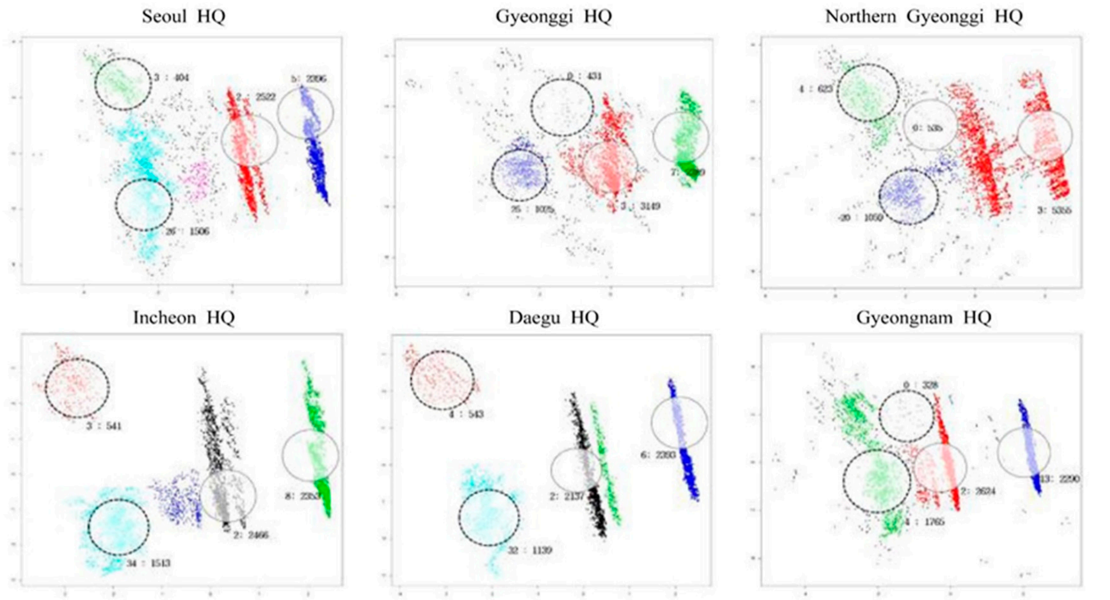


Figure 3. Result of DBSCAN clustering.

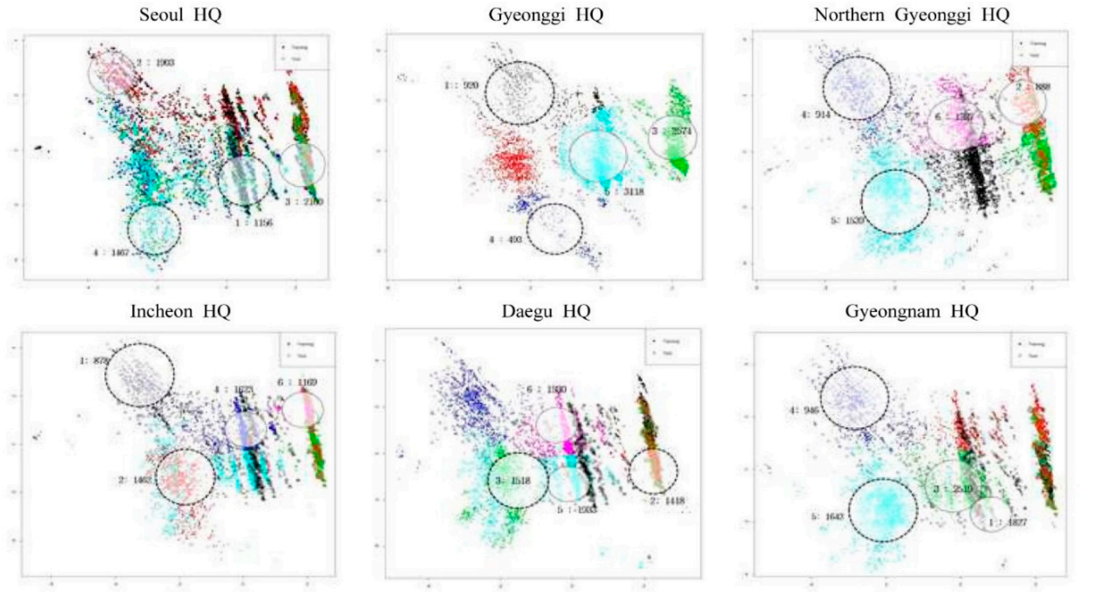


Figure 4. Result of K-means clustering.

3.2. Clustering Result

This study identified important variables affecting the energy consumption of 11 regional buildings. It classified them into high and low energy consumption categories to identify the clusters. Data obtained from a building are a mix of continuous and categorical data; hence, it is difficult to determine the unique features of the building using only the building energy data. Therefore, two different clustering algorithms were employed. A comparison was made to determine which of the two clustering algorithms correctly analyzes building energy data characteristics and energy consumption patterns. The four qualitative evaluation items are as follows [7,52]: the shape of a cluster and density of color, slope of an inverse model, sensitivity of an outlier (determined by the mean of the absolute values of the standardized variables), and grades that are accurately distributed against the amount of energy consumption when it is graded based on relative criteria. Table 4 was developed to determine which of the two clustering algorithms can accurately analyze the data features and energy consumption patterns of the building by weighting the above-mentioned four qualitative evaluation indices. The selected clustering algorithm was then used to categorize buildings only with variables contributing to the energy consumption, which were identified from the buildings; this result is expected to lead to a consistent analysis and provide meaningful results.

Table 5 shows the results of four qualitative indicators for assessing the two aforementioned clustering algorithms. The cluster density using DBSCAN was 2.53, which was 1.03 times higher than that using K-means based on converting data into the energy consumption based on the outdoor temperature. This result means that the use of DBSCAN rather than K-means can lead to meaningful results for deriving the key variables that directly affect consumption among the 16 variables that affect the building energy consumption.

Table 5. Comparison of DBSCAN and K-means.

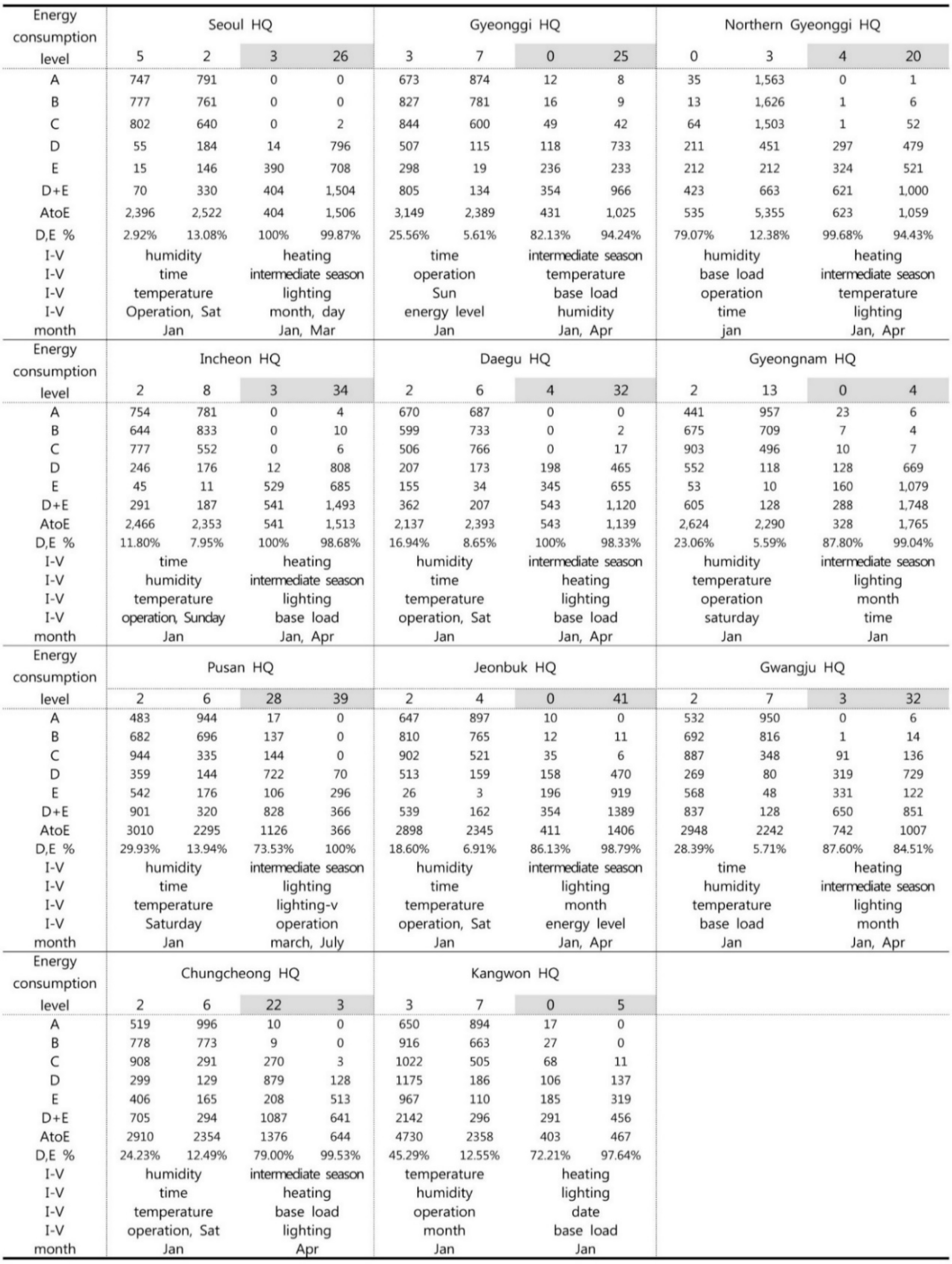
	DBSCAN				**K-Means**			
Criteria	**Level of Clustering**	**Inverse Model ***	**Outlier**	**Energy Consumption Grading**	**Level of Clustering**	**Inverse Model**	**Outlier**	**Energy Consumption Grading**
Weighted value	1.00	0.75	0.50	0.25	1.00	0.75	0.50	0.25
Seoul	3.00	2.25	1.50	0.75	1.00	2.25	1.50	0.75
Gyeonggi	3.00	2.25	1.50	0.50	2.00	2.25	1.00	0.50
Northern Gyeonggi	3.00	2.25	1.00	2.00	1.00	2.25	1.50	0.75
Incheon	3.00	1.50	1.50	0.75	2.00	2.25	1.50	0.75
Daegu	3.00	2.25	1.00	0.75	1.00	1.50	1.50	0.75
Gyeongnam	3.00	2.25	1.50	0.50	1.00	1.50	1.50	0.50
Pusan	3.00	0.75	1.50	0.50	1.00	1.50	1.50	0.50
Jeonbuk	3.00	1.50	1.00	0.50	1.00	1.50	1.50	0.75
Gwangju	3.00	2.25	1.00	0.25	1.00	1.50	1.50	0.75
Chung cheong	3.00	1.50	1.00	0.50	1.00	1.50	1.50	0.25
Kangwon	3.00	1.50	1.00	0.25	1.00	1.50	1.00	0.50
mean	3.00	1.84	1.23	0.66	1.18	1.77	1.41	0.61

* A graph showing both winter heating energy consumption (left slope) and summer cooling energy consumption (right slope) compared with outdoor air temperature, with the interim period without heating or cooling in the middle.

4. Research Results

This section discusses the result of clustering (DBSCAN) of the main variables affecting building energy consumption in 11 buildings in eight regions.

Figure 5 shows the results of classifying the energy consumption (A to E) of each building and deriving the variables that affect the high-energy-use clusters. First, the data were sorted into energy consumption levels A to E by hour, from a low to high energy consumption. Clusters with more levels D and E than A and B can be considered high-energy-consumption clusters. The important variables (abbreviated I-V) are arranged by rank among the high-energy-consumption clusters. We conducted a level-of-importance analysis for those clusters with large amounts of high-energy-consumption data to obtain the important variables, which were then ranked by order of importance (Figure 5). For example, for the Seoul headquarters, clusters 3 and 26 contain large amounts of high-energy-consumption data. Through an analysis of the cluster attributes, we found the important variables to be the heating energy, the intermediate energy, and the lighting energy, in order of importance. January and March were identified as the months with large effects on the energy consumption.



Energy consumption level	Seoul HQ				Gyeonggi HQ				Northern Gyeonggi HQ			
	5	2	3	26	3	7	0	25	0	3	4	20
A	747	791	0	0	673	874	12	8	35	1,563	0	1
B	777	761	0	0	827	781	16	9	13	1,626	1	6
C	802	640	0	2	844	600	49	42	64	1,503	1	52
D	55	184	14	796	507	115	118	733	211	451	297	479
E	15	146	390	708	298	19	236	233	212	212	324	521
D+E	70	330	404	1,504	805	134	354	966	423	663	621	1,000
AtoE	2,396	2,522	404	1,506	3,149	2,389	431	1,025	535	5,355	623	1,059
D,E %	2.92%	13.08%	100%	99.87%	25.56%	5.61%	82.13%	94.24%	79.07%	12.38%	99.68%	94.43%
I-V	humidity		heating		time		intermediate season		humidity		heating	
I-V	time		intermediate season		operation		temperature		base load		intermediate season	
I-V	temperature		lighting		Sun		base load		operation		temperature	
I-V	Operation, Sat		month, day		energy level		humidity		time		lighting	
month	Jan		Jan, Mar		Jan		Jan, Apr		jan		Jan, Apr	
Energy consumption level	Incheon HQ				Daegu HQ				Gyeongnam HQ			
	2	8	3	34	2	6	4	32	2	13	0	4
A	754	781	0	4	670	687	0	0	441	957	23	6
B	644	833	0	10	599	733	0	2	675	709	7	4
C	777	552	0	6	506	766	0	17	903	496	10	7
D	246	176	12	808	207	173	198	465	552	118	128	669
E	45	11	529	685	155	34	345	655	53	10	160	1,079
D+E	291	187	541	1,493	362	207	543	1,120	605	128	288	1,748
AtoE	2,466	2,353	541	1,513	2,137	2,393	543	1,139	2,624	2,290	328	1,765
D,E %	11.80%	7.95%	100%	98.68%	16.94%	8.65%	100%	98.33%	23.06%	5.59%	87.80%	99.04%
I-V	time		heating		humidity		intermediate season		humidity		intermediate season	
I-V	humidity		intermediate season		time		heating		temperature		lighting	
I-V	temperature		lighting		temperature		lighting		operation		month	
I-V	operation, Sunday		base load		operation, Sat		base load		saturday		time	
month	Jan		Jan, Apr		Jan		Jan, Apr		Jan		Jan	
Energy consumption level	Pusan HQ				Jeonbuk HQ				Gwangju HQ			
	2	6	28	39	2	4	0	41	2	7	3	32
A	483	944	17	0	647	897	10	0	532	950	0	6
B	682	696	137	0	810	765	12	11	692	816	1	14
C	944	335	144	0	902	521	35	6	887	348	91	136
D	359	144	722	70	513	159	158	470	269	80	319	729
E	542	176	106	296	26	3	196	919	568	48	331	122
D+E	901	320	828	366	539	162	354	1389	837	128	650	851
AtoE	3010	2295	1126	366	2898	2345	411	1406	2948	2242	742	1007
D,E %	29.93%	13.94%	73.53%	100%	18.60%	6.91%	86.13%	98.79%	28.39%	5.71%	87.60%	84.51%
I-V	humidity		intermediate season		humidity		intermediate season		time		heating	
I-V	time		lighting		time		lighting		humidity		intermediate season	
I-V	temperature		lighting-v		temperature		month		temperature		lighting	
I-V	Saturday		operation		operation, Sat		energy level		base load		month	
month	Jan		march, July		Jan		Jan, Apr		Jan		Jan, Apr	
Energy consumption level	Chungcheong HQ				Kangwon HQ							
	2	6	22	3	3	7	0	5				
A	519	996	10	0	650	894	17	0				
B	778	773	9	0	916	663	27	0				
C	908	291	270	3	1022	505	68	11				
D	299	129	879	128	1175	186	106	137				
E	406	165	208	513	967	110	185	319				
D+E	705	294	1087	641	2142	296	291	456				
AtoE	2910	2354	1376	644	4730	2358	403	467				
D,E %	24.23%	12.49%	79.00%	99.53%	45.29%	12.55%	72.21%	97.64%				
I-V	humidity		intermediate season		temperature		heating					
I-V	time		heating		humidity		lighting					
I-V	temperature		base load		operation		date					
I-V	operation, Sat		lighting		month		base load					
month	Jan		Apr		Jan		Jan					

A: the lowest energy consumption
E: the highest energy consumption
D + E: relatively high energy consumption

A to E: total energy consumption
D, E%: high energy consumption/total energy consumption
I–V: Important variables by consumption section

Figure 5. Result of the clustering of 11 buildings.

Figures 6 and 7 show the location of the buildings and the classification of the important variables derived from the high-energy clusters of the 11 regions, respectively. The temperature and humidity, which can be called environmental variables, were excluded because of duplicate derivation from the 11 regions. In the Seoul, Gyeonggi, Incheon, Daegu, and Gyeongnam buildings, lighting and heating energy were the most influential variables, followed by baseload energy and intermediate energy. It can be concluded that occupancy is closely related to building energy consumption. Moreover, the difference in energy consumption between winter heating and summer cooling was also found to be an important variable for office buildings. Lighting and baseload energy were also found to be important variables (in that order) in the six regions of North Gyeonggi, Busan, Jeonbuk, Gwangju, Chungcheong, and Kangwon. Occupancy and related lighting energy variables, which can be considered characteristics of office buildings, were found to be variables affecting high energy consumption for all 11 regions, rather than regional variables related to the building's location (i.e., latitude and topography). In addition, baseload energy, which is involved in the operation of the building's systems and the outlet load, is also considered to be related to occupancy and was found to have the greatest effect on energy consumption in office buildings. These results indicate that the most important variables for energy usage in non-residential buildings are related to lighting and baseload energy consumption, based on the building's occupancy, rather than to regional and topographical characteristics based on the building's location.

To sum up, the beginning of this study assumed that "only key data can be used to effectively analyze building electricity consumption". This study focused on operational buildings, i.e., not buildings that are yet to be built or recently completed buildings. In the case of existing office buildings, this study aimed to determine which of the continuous energy data derived from the buildings had more influence on the electricity consumption. In general, the U-value of building materials was treated as a categorical variable because it is a set value that does not vary over time. Based on these results, we concluded that the electricity of a building can be effectively managed if only four or five major energy variables derived from the building are controlled or used as key points of operation [53,54].

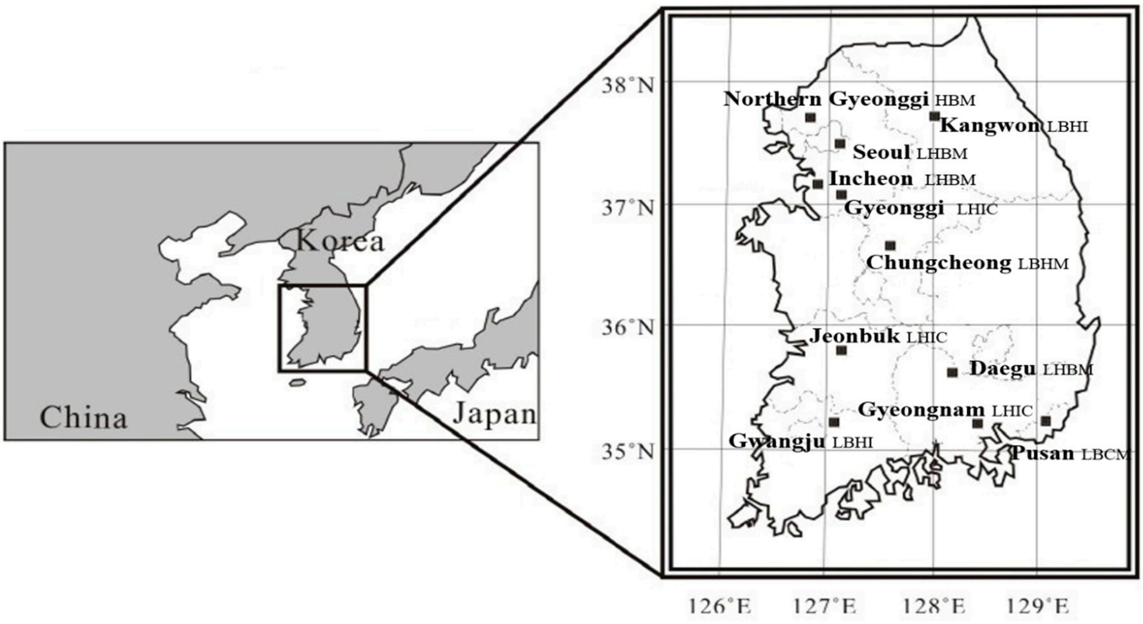


Figure 6. Office location of 11 regions. (L: Lighting energy; H: Heating energy; B: Baseload energy; C: Cooling energy; I: Intermediate energy; M: Month-time).

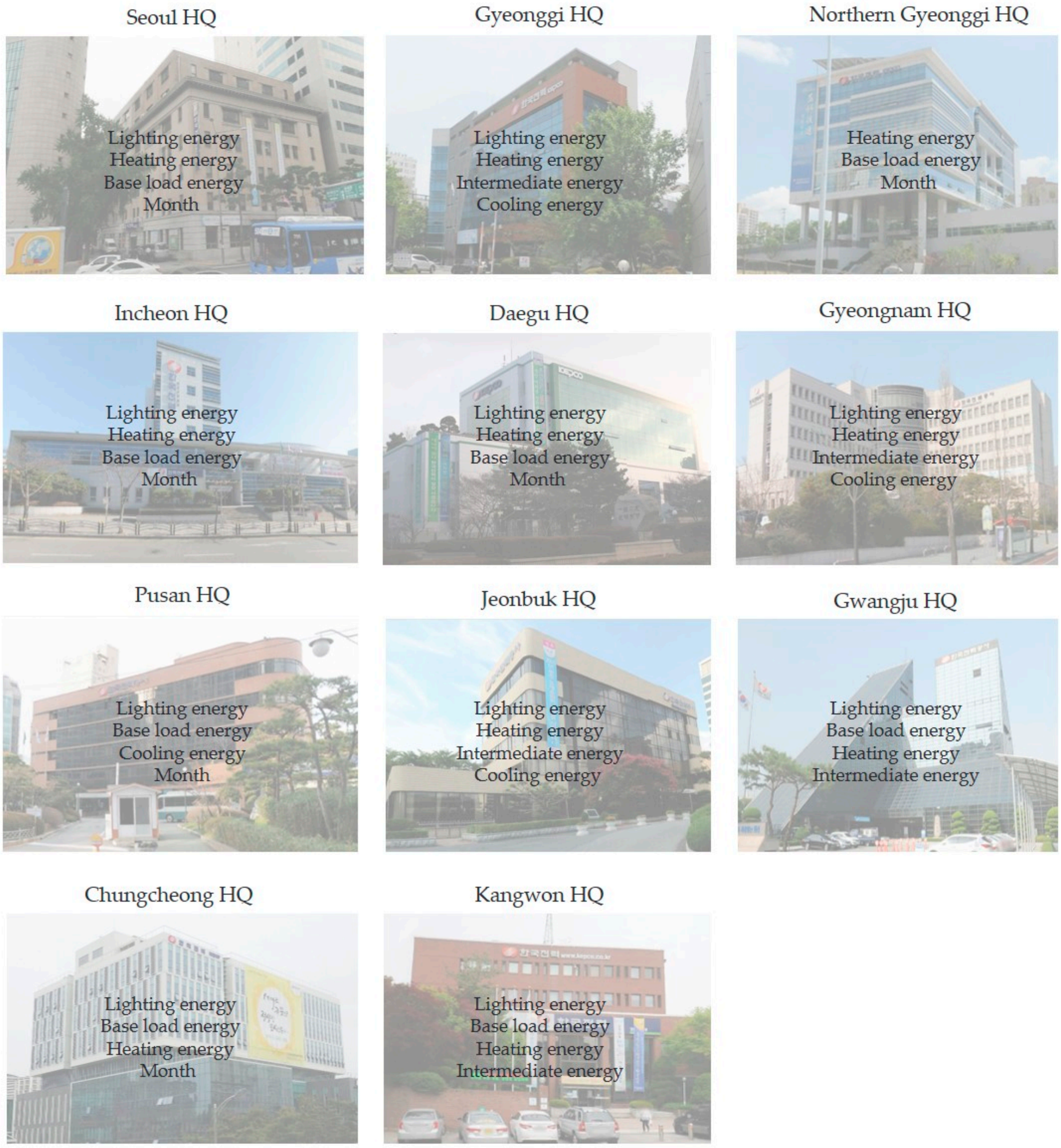


Figure 7. Results of the clustering of 11 regions and important variables.

5. Validation

5.1. Validation of Variables for High and Low Energy Consumption with a *t*-Test

Although a clustering algorithm was used to categorize the building energy data into high- and low-energy-consumption clusters, the *t*-test was used as a method to determine whether the difference between the averages of these two clusters was significant. The important variables of each cluster derived by the boxplot can also be compared using the averages of the absolute values. This comparison, however, does not reflect the range and variance of the actual values. For this reason, the *t*-test was performed to analyze whether the difference between the averages of the two different clusters, which considers the variations between them, was significant.

$$t = \frac{\overline{X_1} - \overline{X_2}}{SE_{\overline{X}1-\overline{X}2}} \; SE_{\overline{X}1-\overline{X}2} = \sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}} \tag{1}$$

The *t*-test explored whether the difference between the averages of the two different clusters (high and low energy consumption) from the building under analysis was significant. Table 6 shows the *t*-test determined six explanatory variables that led to high or low energy consumption of the building under analysis: month, temp (temperature), humi (humidity), lighting.e (energy), base.e (baseload energy), and heat.e (energy). The mean and variance of the day of the week and date variables in the two clusters were almost the same, so they were considered unnecessary for an analysis of the energy consumption.

Table 6. Result of the *t*-test of the 11 regions.

Region	Seoul	Gyeonggi	N-Gy	Incheon	Daegu	G-Nam	Pusan	Jeonbuk	Gw-Ju	Ch-Ch	Ka-W
month	0.000	0.001	0.000	0.000	0.000	0.620	0.000	0.000	0.000	0.000	0.000
date	0.803	0.616	0.728	0.219	0.024	0.531	0.450	0.484	0.288	0.177	0.739
hour	0.992	0.563	0.980	0.021	0.015	0.551	0.428	0.008	0.573	0.258	0.161
temp	0.000	0.098	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
humi	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
base.e	0.000	0.006	0.000	0.000	0.000	0.544	0.000	0.170	0.000	0.000	0.000
lit.e	0.000	0.000	0.280	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
heat.e	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.000	0.000	0.000	0.000
inter.e	0.318	0.000	0.000	0.500	0.500	0.000	0.318	0.000	0.000	0.318	0.000
cool.e	0.500	0.000	0.016	0.500	0.500	0.000	0.000	0.000	0.500	0.500	0.000
Mon	0.056	0.359	0.492	0.673	0.487	0.012	0.801	0.069	0.219	0.011	0.162
Tue	0.563	0.310	0.967	0.825	0.212	0.015	0.801	0.768	0.395	0.145	0.892
Wed	0.473	0.071	0.532	0.129	0.005	0.045	0.326	0.195	0.781	0.209	0.508
Thu	0.551	0.278	0.005	0.454	0.005	0.023	0.903	0.036	0.477	0.818	0.226
Sat	0.500	0.500	0.318	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Sun	0.500	0.318	0.157	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500

5.2. Regression Results

A regression analysis was performed using only the 16 major variables from each building, derived from the analysis process, and five or six variables selected by machine learning, to verify that the observed values, i.e., the Y variable, were well-described. Two validations were carried out for this purpose.

The coefficient of determination (R^2) of the 16 original variables was 0.8356 in the Figure 8. whereas that of the important variables was 0.8261 from the Figure 9. This suggests that the important variables alone are able to explain most of the observed variance, as the performance of the regression equation built using only the important variables is practically identical to that obtained using all the original variables.

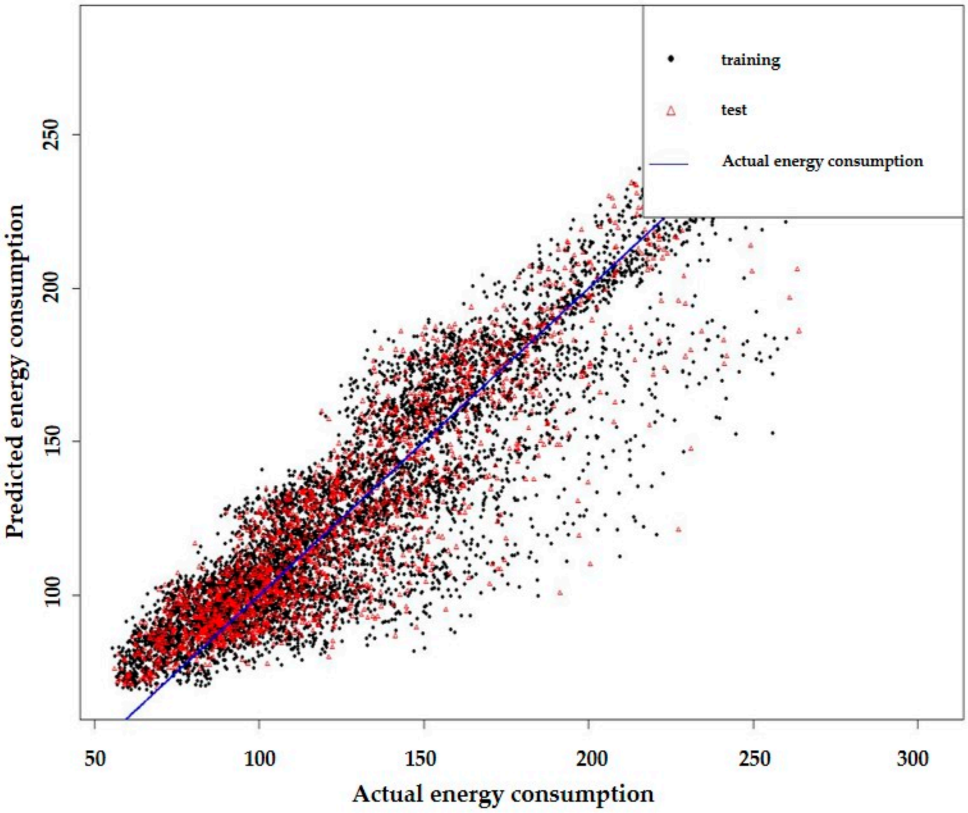


Figure 8. Results of a regression analysis of the 16 original variables.

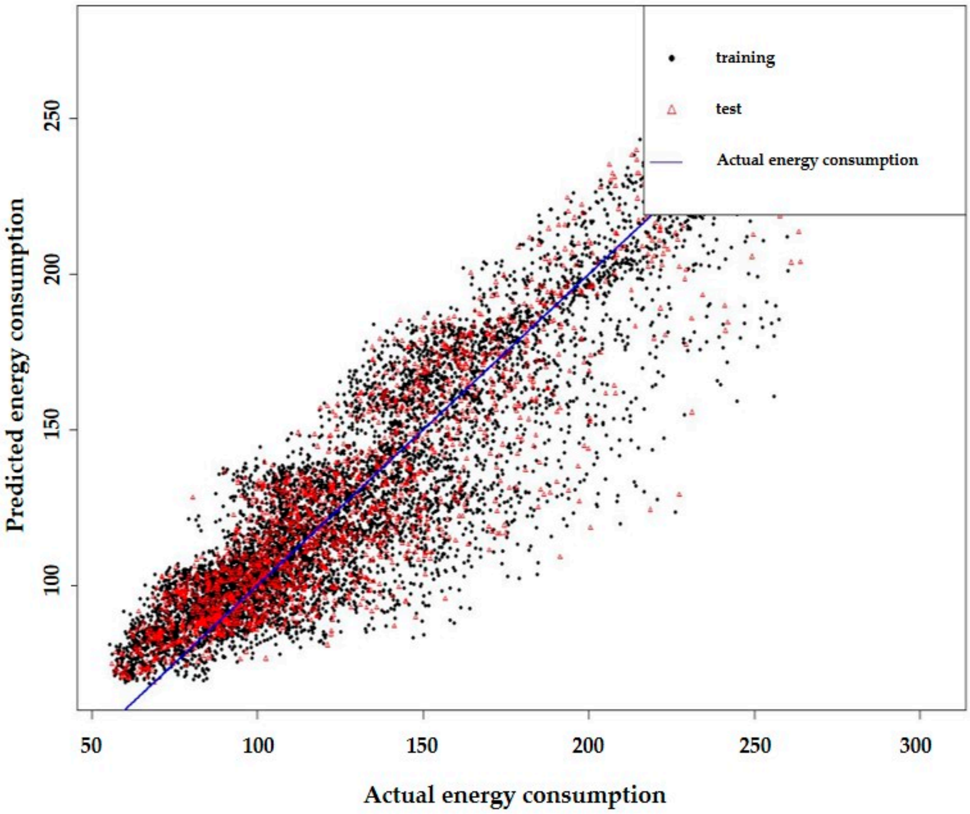


Figure 9. Results of a regression analysis of the major variables.

5.3. Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE)

The MSE takes the square root of the average of the squares of the errors from an observation and compares it to deviation between the observed values. As a measure that generalizes standard deviations, MSE is used to validate the amount of difference between the actual value and estimated value using a regression equation. Table 7 shows the results were obtained from the two regression analyses performed for the 16 original variables and important variables based on the computation of their MSE and MAPE. The MSE and MAPE results of these two regression analyses show that even though the MSE and MAPE of the important variables were only slightly higher than those of the 16 original variables, and Table 8 shows the regression analysis error and coefficient of determination of the key variables were as significant as those of the 16 original variables, which were larger in number.

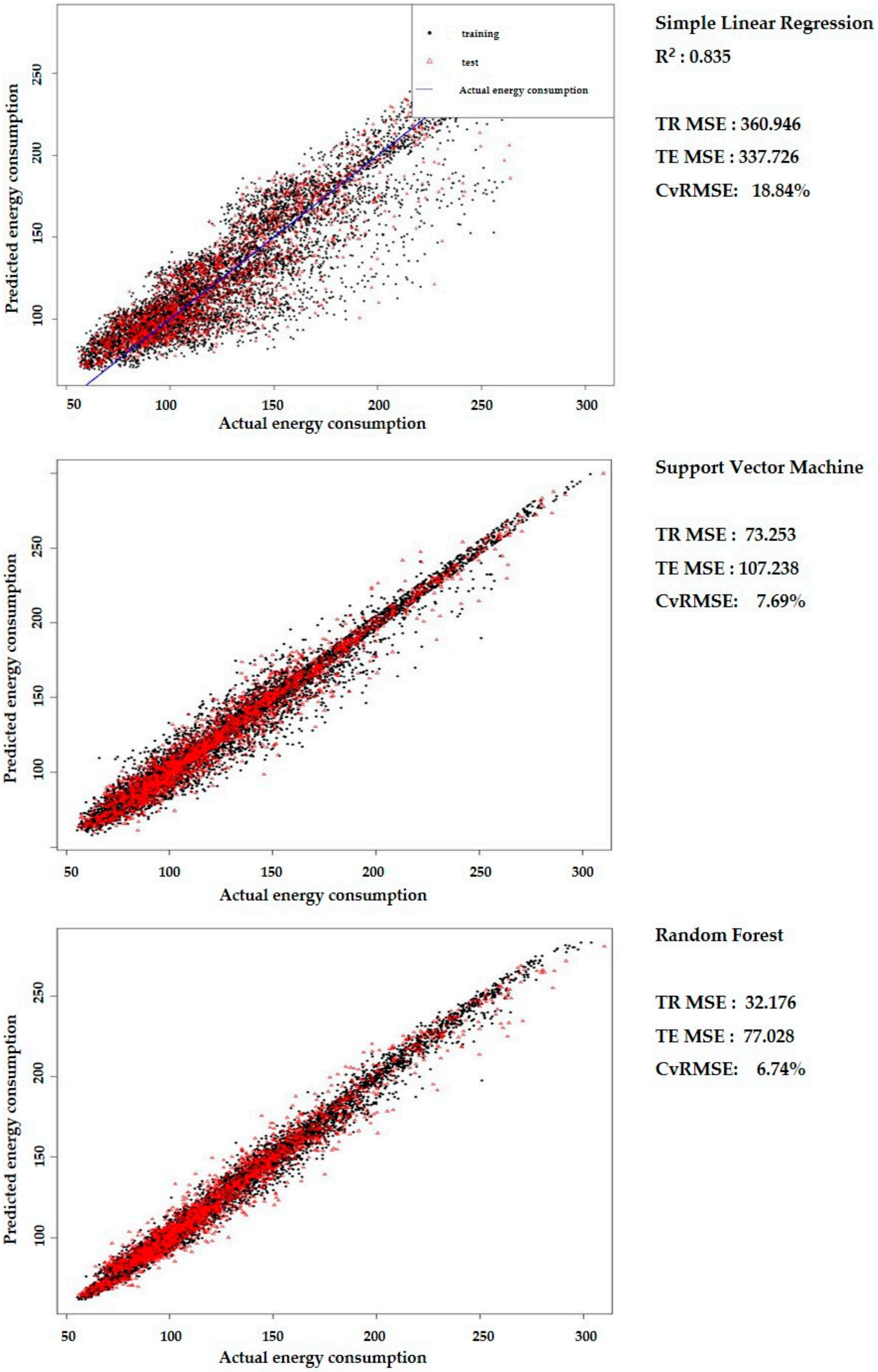
Table 7. Comparison of Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) associated with the regression analyses.

Evaluation Index	Original 16 Variables	Major Variables	Differences
MSE Training set	360.946	381.8843	21.938
MSE Test set	337.726	365.9697	28.243
MAPE Training set	11.65511	12.06266	0.40
MAPE Test set	11.65511	11.77404	0.12

Table 8. Verification of error rates with three algorithms.

Algorithms	Original 16 Variables	Major Variables	Differences
Regression R^2	0.835	0.826	0.009
Regression CvRMSE	13.84	14.60	0.8
SVM CvRMSE	7.69	9.03	1.34
Random Forest CvRMSE	6.74	7.20	0.46

Figures 10 and 11 show comparisons of energy consumption predictions, where one predicted energy consumption uses the initial 16 variables and the other prediction only uses the major variables derived by machine learning with the testbed as a subject. Apart from the R^2 value, which is the coefficient of determination, and the MSE, which is an indicator of the error rate of prediction, the results using the major variables showed error rates ranging from 7.2% to 14.6% compared with the predictions using 16 variables. The error rate using the major variables was lower than the ASHRAE Guideline 14 [55] criteria of 30%. Moreover, the random forest method showed the lowest error rates among the three different methodologies (simple linear regression, support vector machine, and random forest).



Simple Linear Regression

R^2 : 0.835

TR MSE : 360.946

TE MSE : 337.726

CvRMSE: 18.84%

Support Vector Machine

TR MSE : 73.253

TE MSE : 107.238

CvRMSE: 7.69%

Random Forest

TR MSE : 32.176

TE MSE : 77.028

CvRMSE: 6.74%

Figure 10. Results of the simple linear regression, support vector machine, and random forest analyses using the original variables.

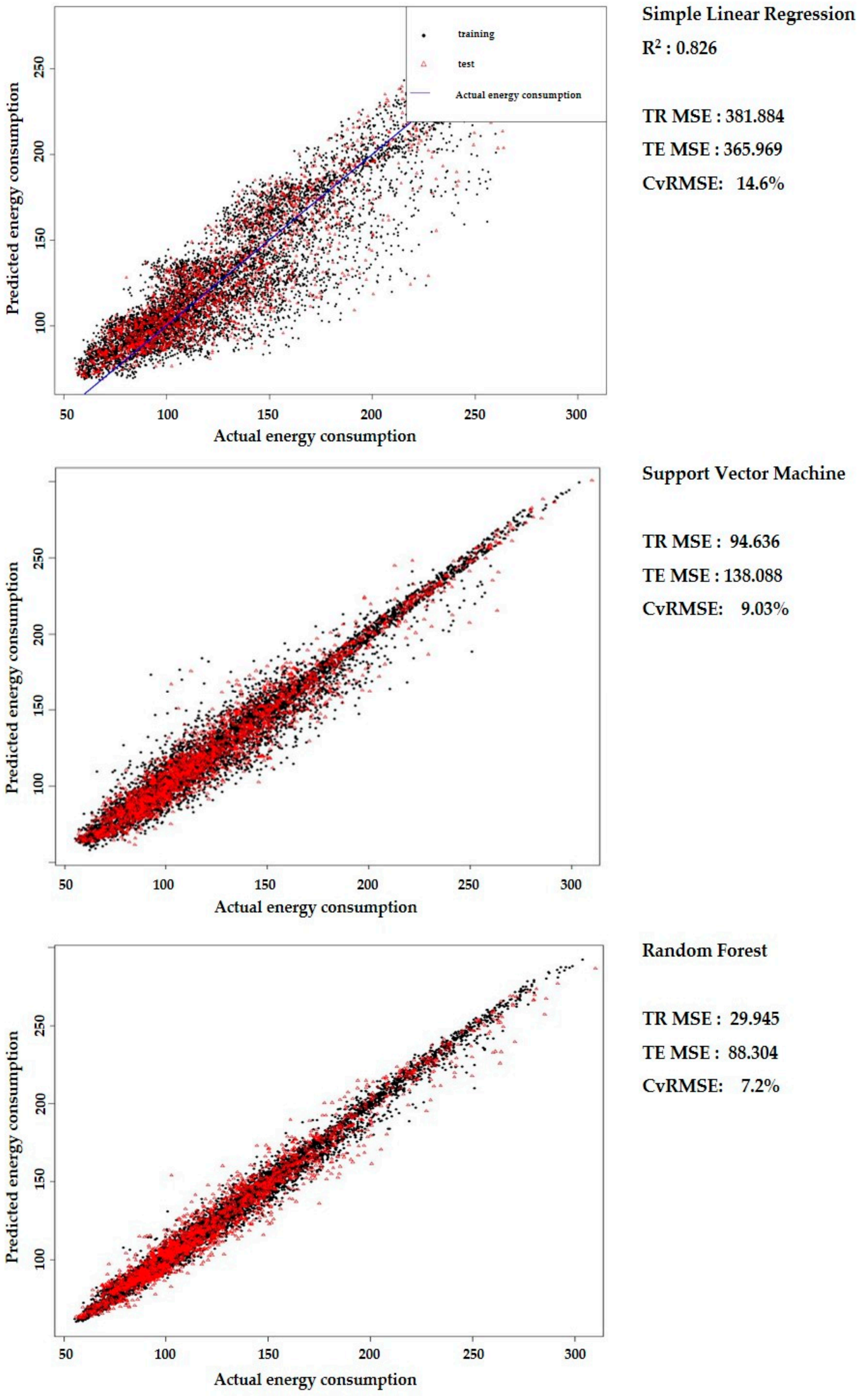


Figure 11. Results of the simple linear regression, support vector machine, and random forest analyses using only the major variables.

6. Conclusions

6.1. Research Conclusions

The purposes of this study were to examine the factors contributing to electric energy use, which accounts for more than 65% of the total building energy consumption, using clustering in machine learning based on data measured from the actual building, and then to find which variables were identified and which had a strong correlation with energy consumption.

Machine learning of the previous studies only predicted the energy usage of buildings in the building sector or confirmed the patterns of usage through clustering methods. However, this study found the key variables that affect the consumption of buildings with a high energy consumption through two different characteristics of the clustering methodology.

The results suggest that the energy consumption predicted by the major variables alone is reliably accurate and that the method can be expected to reduce energy consumption more when these major variables are controlled, or their efficiency is improved. The results can be summarized as follows:

1. The important variables for building energy consumption were derived based on machine learning clustering and were clustered into high- and low-energy-consumption clusters;
2. Based on the clustering, the energy consumption of 11 regional buildings was analyzed according to changes in the outdoor air temperature, which can reveal the building energy features;
3. *T*-tests were performed on the results of the buildings categorized into similar clusters to determine the explanatory variables that led to a high or low energy consumption;
4. Lastly, the important variables identified from this methodology were validated.

   - Comparison of R^2 values
   - Validation of the two regression equations for the 16 original variables and important variables by obtaining the MSE and MAPE.

   With respect to Conclusion 3, the important variables that had a decisive effect on the energy consumption of a single building under analysis and the 11 regional buildings (12 buildings in total) were found to be two environmental variables (temperature and humidity), lighting energy, heating energy, and a time variable (month);
5. This study determined the key variables affecting the electrical electricity consumption of buildings, especially non-residential buildings. Except for the external environment (geographic location, temperature, and humidity), the studied building's electricity consumption was found to be as important as its physical characteristics, such as an increased cooling energy, lighting energy, and baseload, due to the working conditions of the occupants. Internal heat gains varied according to occupancy time and density.

Buildings may appear similar, but vary in electricity consumption and patterns, depending on their use, size, and occupants. Therefore, the results of this study cannot necessarily be applied to all buildings. Since this study had a specific application (office building, commercial building), it can be used as an example of variables affecting the electricity consumption of a non-residential building. However, the main influencing factor may be different for a residential building or buildings with other uses (e.g., where the working hours of the occupants are not common).

6.2. Significance and Application

The significance of this study relates to three aspects. First, the analysis of building energy use by existing engineering methods was used to derive absolute values by load factors using a simulation program and to predict reductions by subloads (cooling, heating, and ventilation). Although it may be possible to predict the detailed reduction volume by energy loads or performance, the data were not based on actual data, and the reliability of the results has thus been questioned. However, predicting building energy use using major variables derived from machine learning (as in this study) utilizes the actual total energy volume of the target building, which helps to overcome the problem of reliability of the resulting value. This might occur because of differences in the actual consumption or performance versus the simulated amounts. Second, the prediction of building energy consumption applied with machine learning leads to the same results, despite the lack of experience or intuition by the researcher. Consistent results can be provided with no direct influence from an expert.

It is possible to derive major variables influencing the building energy consumption (high consumption/low consumption), identify which energy source is most responsible for consumption, and view how they influence consumption.

The major variables causing building energy consumption can be used to identify the status of energy consumption of the building and as an indicator of post hoc maintenance.

The machine-learning methodology can be widely applied to various buildings with different uses or located in different climates, and buildings can be classified by major variables influencing building energy consumption.

This method can be used for selecting the variables affecting building energy consumption and can be applied without constraints to non-residential buildings. Furthermore, if there is a collectible data set, such as real-time information regarding a building service, variables that have an impact on energy consumption can be identified without constraints.

From an economic perspective, the fourth significance is as follows. Utilizing this study's methodology in a practical work setting, it would be possible to present monitoring points for building maintenance for aging buildings. Selecting and using energy measurement sensors also make it possible to find the most significant measurement points because the data from actual equipment systems can be used to obtain clustering results (for correlation analysis and the extraction of major variables).

Author Contributions: Conceptualization, S.C. and G.-S.K.; methodology, S.C.; validation, S.C.; writing—original draft preparation, S.C.; writing—review and editing, S.C., and J.L.; visualization, S.C., and J.B.; supervision, S.-B.L.

Funding: This research was sponsored by the Korean Institute of Energy Technology Evaluation and Planning (KETEP) of the Republic of Korea (No. 2018201060010A).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Climate Change, Sustainable Development Goals. Available online: https://www.un.org/sustainabledevelopment/climate-change-2/ (accessed on 5 April 2019).
2. Choi, M.S.; Choi, D.Y. *Building Energy Consumption Sampling Survey*; Korea Energy Economics Institute: Ulsan, Korea, 2014.
3. Fayaz, M.; Kim, D. Energy Consumption Optimization and User Comfort Management in Residential Buildings Using a Bat Algorithm and Fuzzy Logic. *Energies* **2018**, *11*, 161. [CrossRef]
4. Selin, R. *The Outlook for Energy: A View to 2040*; ExxonMobil: Irving, TX, USA, 2013.
5. Sieminski, A. *International Energy Outlook*; Energy Information Administration: Washington, DC, USA, 2014.
6. KEA. *Statistics on Energy Use and Greenhouse Gas Emissions in the Industrial Sector*; KEA: Ulsan, Korea, 2017.
7. Woo, H.J.; Choi, K.W.; Kim, H.S.; Auh, J.S.; Cho, S.Y.; Baek, J.; Kim, G.S.; Leigh, S.B. A Study on Classifying Building Energy Consumption Pattern Using Actual Building Energy Data. *J. Arch. Inst. Korea Plan. Des.* **2016**, *32*, 143–151. [CrossRef]
8. Cho, S.Y.; Leigh, S.B. Comparing methodology of building energy analysis—Comparative analysis from steady-state simulation to data-driven analysis. *KIEAE J.* **2017**, *17*, 77–86. [CrossRef]
9. Moon, H.J.; Yoon, Y.R. *A Case Study on the Use of Machine Learning Technique for Building Energy Analysis*; Korea Institute of Architectural Sustainable Environment and Building Systems: Seoul, Korea, 2017.
10. Seo, W.J.; Ahn, K.U.; Park, C.S. *Utilizing Machine Learning Technology in Building Energy Diagnosis and Facility Control*; Korea Institute of Architectural Sustainable Environment and Building Systems: Seoul, Korea, 2017.

11. Wang, H.; Chiang, P.C.; Cai, Y.; Li, C.; Wang, X.; Chen, T.L.; Wei, S.; Huang, Q. Application of Wall and Insulation Materials on Green Building: A Review. *Sustainability* **2018**, *10*, 3331. [CrossRef]
12. Korea Energy Economics Institute. *Year Book of Energy Statistics 2018*; Korea Energy Economics Institute: Ulsan, Korea, 2019.
13. Crawley, D.B.; Hand, J.W.; Kummert, M.; Griffith, B.T. Contrasting the capabilities of building energy performance simulation programs. *Build. Environ.* **2008**, *43*, 661–673. [CrossRef]
14. Kim, Y.H.; Park, W.J.; Yang, S.H.; Kim, S.J. *Building Energy Consumption Prediction and Evaluation System*; The Society of Air-Conditioning and Refrigerating Engineers of Korea: Seoul, Korea, 2018.
15. Beak, Y.R. Thermal energy analysis program of building. *J. Mech. Sci. Technol.* **2002**, *42*, 20–21.
16. Yu, Y.; Woradechjumroen, D.; Yu, D. A review of fault detection and diagnosis methodologies on air-handling units. *Energy Build.* **2014**, *82*, 550–562. [CrossRef]
17. Fan, C.; Xiao, F.; Yan, C.; Liu, C.; Li, Z.; Wang, J. A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. *Appl. Energy* **2019**, *235*, 1551–1560. [CrossRef]
18. Harish, V.; Kumar, A. A review on modeling and simulation of building energy systems. *Renew. Sustain. Energy Rev.* **2016**, *56*, 1272–1292. [CrossRef]
19. Harish, V.S.K.V.; Kumar, A. Techniques used to construct an energy model for attaining energy efficiency in building: A review. In Proceedings of the International Conference on Control, Instrumentation, Energy and Communication (CIEC), Calcutta, India, 31 January–2 February 2014; pp. 366–370.
20. Wilde, P.D. The gap between predicted and measured energy performance of buildings. *Autom. Constr.* **2014**, *41*, 40–49. [CrossRef]
21. Menezes, A.C.; Cripps, A.; Bouchlaghem, D. Predicted vs. actual energy performance of non-domestic buildings using post occupancy evaluation data to reduce the performance gap. *Appl. Energy* **2012**, *97*, 355–364. [CrossRef]
22. Olivia, G.S.; Christopher, T.A. In-use monitoring of buildings: An overview and classification of evaluation methods. *Energy Build.* **2015**, *86*, 176–189. [CrossRef]
23. Choi, J.H.; Loftness, V.; Aziz, A. Post-occupancy evaluation of 20 office buildings as basis for future IEQ standards and guidelines. *Energy Build.* **2012**, *46*, 167–175. [CrossRef]
24. Agha-Hossein, M.; El-Jouzi, S.; Elmualim, A.; Ellis, J.; Williams, M. Post-occupancy studies of an office environment: Energy performance and occupants' satisfaction. *Build. Environ.* **2013**, *69*, 121–130. [CrossRef]
25. Salehi, M.M.; Cavka, B.T.; Frisque, A.; Whitehead, D.; Bushe, W.K. A case study: The energy performance gap of the Center for Interactive Research on Sustainability at the University of British Columbia. *J. Build. Eng.* **2015**, *4*, 127–139. [CrossRef]
26. Niu, S.; Pan, W.; Zhao, Y. A virtual reality integrated design approach to improving occupancy information integrity for closing the building energy performance gap. *Sustain. Cities Soc.* **2016**, *27*, 275–286. [CrossRef]
27. Herrando, M.; Cambra, D.; Navarro, M.; De La Cruz, L.; Millán, G.; Zabalza, I.; Bribian, I.Z. Energy Performance Certification of Faculty Buildings in Spain: The gap between estimated and real energy consumption. *Energy Convers. Manag.* **2016**, *125*, 141–153. [CrossRef]
28. Min, Z.; Morgenstern, P.; Halburd, L.M. Facilities management added value in closing the energy performance gap. *Int. J. Sustain. Built Environ.* **2016**, *5*, 197–209. [CrossRef]
29. Witten, I.H.; Frank, E.; Hall, M.A. *Data Mining: Practical Machine Learning Tools and Techniques*; Elsevier: Amsterdam, The Netherlands, 2013; pp. 45–52.
30. Lee, J.H.; Shin, J.; Realff, M.J. Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Comput. Chem. Eng.* **2018**, *114*, 111–121. [CrossRef]
31. Molina-Solana, M.; Ros, M.; Ruiz, M.D.; Gómez-Romero, J.; Martin-Bautista, M. Data science for building energy management: A review. *Renew. Sustain. Energy Rev.* **2017**, *70*, 598–609. [CrossRef]
32. Wang, Z.; Wang, Y.; Srinivasan, R.S. A novel ensemble learning approach to support building energy use prediction. *Energy Build.* **2018**, *159*, 109–122. [CrossRef]
33. Catalina, T.; Virgone, J.; Blanco, E. Development and validation of regression models to predict monthly heating demand for residential buildings. *Energy Build.* **2008**, *40*, 1825–1832. [CrossRef]
34. Ekici, B.B.; Aksoy, T.U. Prediction of building energy consumption by using artificial neural networks. *Adv. Eng. Softw.* **2009**, *40*, 356–362. [CrossRef]

35. Yu, Z.; Haghighat, F.; Fung, B.C.M.; Yoshino, H. A decision tree method for building energy demand modeling. *Energy Build.* **2010**, *42*, 1637–1646. [CrossRef]
36. Dong, B.; Cao, C.; Lee, S.E. Applying support vector machines to predict building energy consumption in tropical region. *Energy Build.* **2005**, *37*, 545–553. [CrossRef]
37. Classification: Train, Validatio, Test Split. Available online: https://blog.algorithmia.com/page/50 (accessed on 29 September 2019).
38. Paudel, S.; Elmitri, M.; Couturier, S.; Nguyen, P.H.; BrunoLacarrière, R.; Le Corre, O. A relevant data selection method for energy consumption prediction of low energy building based on support vector machine. *Energy Build.* **2016**, *138*, 240–256. [CrossRef]
39. Yildiz, B.; Bilbao, J.; Sproul, A. A review and analysis of regression and machine learning models on commercial building electricity load forecasting. *Renew. Sustain. Energy Rev.* **2017**, *73*, 1104–1122. [CrossRef]
40. Rahman, A.; Smith, A.D. Predicting fuel consumption for commercial buildings with machine learning algorithms. *Energy Build.* **2017**, *152*, 341–358. [CrossRef]
41. Moon, J.; Jun, S.; Park, J.; Choi, Y.H.; Hwang, E. An Electric Load Forecasting Scheme for University Campus Buildings Using Artificial Neural Network and Support Vector Regression. *KIPS Trans. Comput. Commun. Syst.* **2016**, *5*, 293–302. [CrossRef]
42. Seong, N.C.; Kim, J.H.; Choi, W.; Yoon, S.C.; Nassif, N. Development of Optimization Algorithms for Building Energy Model Using Artificial Neural Networks. *J. Korean Soc. Living Environ. Syst.* **2017**, *24*, 29–36. [CrossRef]
43. Naganathan, H.; Chong, W.O.; Chen, X. Building energy modeling (BEM) using clustering algorithms and semi-supervised machine learning approaches. *Autom. Constr.* **2016**, *72*, 187–194. [CrossRef]
44. Ko, J.H.; Kong, D.S.; Huh, J.H. Baseline building energy modeling of cluster inverse model by using daily energy consumption in office buildings. *Energy Build.* **2017**, *140*, 317–323. [CrossRef]
45. Yang, J.; Ning, C.; Deb, C.; Zhang, F.; Cheong, D.; Lee, S.E.; Sekhar, C.; Tham, K.W. k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energy Build.* **2017**, *146*, 27–37. [CrossRef]
46. Moon, H.J.; Jung, S.K.; Ruy, S.H. *Building Cooling and Heating Energy Consumption Pattern Analysis Based on Building Energy Management System Data Using Machine Learning Techniques*; The Society of Air-Conditioning and Refrigerating Engineers of Korea: Seoul, Korea, 2015.
47. Hwang, H.M.; Lee, S.H.; Park, J.B.; Park, Y.G.; Son, S.Y. Load Forecasting using Hierarchical Clustering Method for Building. *Trans. Korean Inst. Electr. Eng.* **2015**, *64*, 41–47. [CrossRef]
48. Shmueli, G.; Petel, N.R.; Bruce, P.C. *Data Mining for Business Intelligence*; Wiley: New York, NY, USA, 2010; pp. 91–97.
49. Cho, S.Y.; Leigh, S.B. A Study of the Possibility of Building Energy Saving through the Building Data: A Case Study of Macro to Micro Building Energy Analysis. *Korean J. Air Cond. Refrig. Eng.* **2017**, *29*, 580–591.
50. Flach, P. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*, 1st ed.; Intelligent Systems Laboratory, University of Bristol: Bristol, UK, 2012; pp. 295–328.
51. R-3.6.1 for Window. Available online: https://cran.r-project.org/bin/windows/base/ (accessed on 16 August 2019).
52. Cho, K.H.; Oh, J.H.; Kim, S.S.; Lee, B.H.; Yeo, M.S. *An Analysis of Energy Consumption Patterns in University Buildings Using Inverse Modeling*; Architectural Institute of Korea: Seoul, Korea, 2017.
53. Tronchin, L.; Manfren, M.; James, P.A.B. Linking design and operation performance analysis through model calibration: Parametric assessment on a Passive House building. *Energy* **2018**, *165*, 26–40. [CrossRef]
54. Attanasio, A.; Piscitelli, M.S.; Chiusano, S.; Capozzoli, A.; Cerquitelli, T. Towards an Automated, Fast and Interpretable Estimation Model of Heating Energy Demand: A Data-Driven Approach Exploiting Building Energy Certificates. *Energies* **2019**, *12*, 1273. [CrossRef]
55. ASHRAE. *ASHRAE's Guideline 14, Measurement of Energy and Demand Savings*; ASHRAE: Atlanta, GA, USA, 2002.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).