energies

MDPI

Article

Development of a Consecutive Occupancy Estimation Framework for Improving the Energy Demand Prediction Performance of Building Energy Modeling Tools

Seokho Kim [1], Yujin Song [2], Yoondong Sung [2] and Donghyun Seo [1,*]

[1] Department of Architectural Engineering of Chungbuk National University, Cheongju, Chungbuk 28644, Korea; archimpresson@gmail.com

[2] Korea Institute of Energy Research, Daejeon 34101, Korea; eugenesong@kier.re.kr (Y.S.); syd9035@kier.re.kr (Y.S.)

* Correspondence: seodh@cbnu.ac.kr; Tel.: +82-43-261-3438

Received: 23 December 2018; Accepted: 24 January 2019; Published: 29 January 2019

Abstract: To improve the energy prediction performance of a building energy model, the occupancy status information is very important. This is more important in real buildings, rather than under construction buildings, because actual building occupancy can significantly influence its energy consumption. In this study, a machine learning based framework for a consecutive occupancy estimation is proposed by utilizing internet of things data, such as indoor temperature and luminance, CO_2 density, electricity consumption of lighting, HVAC (heating, ventilation, and air conditioning), electric appliances, etc. Three machine learning based occupancy estimation algorithms (decision tree, support vector machine, artificial neural networks) are selected and evaluated in terms of the performance of estimating the occupancy status for each season. The selection process of the input variables that have crucial impact on the algorithms' performance are described in detail. Finally, an occupancy estimation framework that can repeat model training and estimation consecutively in a situation when time-series data are continuously provided over the entire measurement period is suggested. In addition, the performance of the framework is evaluated to identify how it improves the energy prediction performance of the building energy model compared to conventional energy modeling practices. The suggested framework is distinguished from similar previous studies in two ways: (1) The proposed framework reveals that input variables for the occupancy estimation model can be occasionally changed by an occupant response to certain times and seasons, and (2) the framework incorporates time-series indirect occupancy sensing data and classification algorithms to consecutively provide occupancy information for the energy modeling effort.

Keywords: occupancy status indirect occupancy sensing; R-script; decision tree; support vector machine; BCVTB

1. Introduction

1.1. Background of Study

Detailed building energy modeling tools that can predict heating and cooling energy consumption based on dynamic analysis are being widely used in academia and the industry. However, these dynamic analysis principles are only applied on an interpretation of the thermal behavior of a building construction, with respect to the ambient indoor and outdoor environment. In contrast to the high precision of dynamic analysis, relatively simplified input data are used for occupant behavior, light

and electrical equipment, HVAC (heating, ventilation, and air conditioning) control, and deployments of shades and windows, etc. [1]. Because stochastic features of these elements are unable to model precisely in building energy simulation tools [2], it has been understood that the best way to model them is to use typical schedules. Using a typical schedule is a valid and effective method if the purpose of the building energy simulation is to estimate the typical energy consumption of a target building. However, if one wants to estimate the actual energy consumption of a real building for the detailed operation of building systems or utilization in the model predictive control (MPC), it is necessary to use dynamic schedules for the input variables. In this case, the accuracy of the input occupancy information can have a profound influence on the accuracy of the building energy estimation.

1.2. Necessity and Purpose of Study

The estimation resolution of occupancy information can be divided into four levels: The estimation of occupancy status (Level-1, "Occupancy"), the estimation of the number of occupants (Level-2, "Count"), the estimation of the identity of an occupant (Level-3, "Identity"), and the estimation of the behavior of an occupant (Level-4, "Activity") [3]. Although Internet of Things (IoT) technology is widely applied in modern building environments, many recent studies have focused on the estimation of occupancy status (Level-1) and the number of occupants (Level-2) due to difficulties associated with measuring irregular occupant behavior and privacy issues. Various IT devices, including passive infrared (PIR) sensors, Radio Frequency Identification (RFID), Wi-Fi, and Beacon, as well as video cameras (which can directly measure the occupancy information), are utilized in occupancy estimation, but none of them have successfully obtained absolute reliability and utilization. This is because of the various problems associated with occupant privacy and the technical limitation to extracting occupancy status from the measured data. Excessive installation and operation costs are additional obstacles.

Garg et al. [4], Nagy et al. [5], and Gunay et al. [6] conducted studies on occupancy estimation by using a PIR sensor. They tried to provide criteria to adjust the frequency of occupancy estimation depending on the time delay for receiving occupant information. Benezeth et al. [7] and Zou et al. [8] used video cameras and image processing technology and reported an estimation accuracy of 97% and 95.3%, respectively. Amin et al. [9] and Benezeth et al. [7] highlighted the possibility that errors might be caused in image processing due to the clothing color and posture of an occupant and by the backgrounds of the target images. In addition, Shen et al. [10] reported that there might be a limitation of these sensors in some measurement situations due to confined view angles and signal range.

To overcome the disadvantages associated with the direct measurement of occupancy information, there have been various attempts to utilize indirect sensing, which is also referred to as implicit occupancy sensing, ambient sensing, and soft sensing in other research papers. The collected data is known as occupancy-related data. Among them, the most representative method is an occupancy estimation method that utilizes a CO_2 sensor because it is related to the density of occupants within a space. However, CO_2 sensor data has its own disadvantages because it is subject to the influence of HVAC operation, the air tightness of spaces, and the wind speed of the ambient environment. Therefore, it might be necessary to consider various additional factors in some cases, thus making it more complicated to perform an accurate estimation. For this reason, an increasing number of studies are being conducted using various types of indirect sensor data, such as the noise level, energy consumption, and network usage quantity.

Chen et al. [11] and Candanero et al. [12] conducted occupancy studies using measured indoor environment data and a set of various algorithms, including the extreme learning machine (ELM), artificial neural network (ANN), support vector machine (SVM), K-nearest neighbor (KNN), linear discriminant analysis (LDA), classification and regression tree (CART), gradient boosting machine (GBM), and random forest (RF). They reported the estimation accuracy of occupancy status (Level-1) at 93.5% and 99.3%, respectively. Yang et al. [13], Dong et al. [14], Khan et al. [15], Lam et al. [16], and Han et al. [17] performed measurements using a combination of direct and indirect sensors and then

estimated the occupancy status using several of the aforementioned algorithms. Most of them reported an occupancy status estimation of approximately 95%, and occupant quantity of approximately 80%.

To enhance the accuracy of occupancy estimation, several studies have used additional information. For instance, Yang et al. [18] estimated occupancy status by using indoor environment data, occupancy sensor data, and entrance door opening and closing data. They reported an accuracy of 96.0% for private rooms and an accuracy of 97.3% in an open office space. Dodier et al. [19] estimated occupancy status using the belief network based on outgoing phone call data and occupancy sensor data, which were collected from private office rooms for two days. Zhao et al. [20] estimated occupancy status using entrance door opening and closing data, Wi-Fi and Bluetooth connection data, keyboard-mouse usage data, and room sensor data. All these data were collected from private office spaces over a two week period. With the Bayesian network algorithm, they reported an occupancy status estimation accuracy of 96.7%. Wang et al. [21,22] used Wi-Fi probe data, indoor environmental measurement data, and a camera for ground truth from an open office space. To estimate the number of occupants, a M-FRNN (Markov based feedback recurrent neural network) algorithm was suggested, and then compared with many other algorithms, such as ANN, kNN, and SVM. They found around an 80% people counting accuracy with two occupants' tolerance from the office space for 60 occupants. Recently, Zou et al. [23,24] showed a very accurate Level 4 (activity) estimation result of 97.6% and Level 2 (count) estimation result of 92.5% with the Wi-Fi signal and a deep learning algorithm, DeepHare. This high accuracy of prediction is noticeable, though they implemented within controlled conditions of five typical activities. Aora et al. [25] estimated the number of occupants using the decision tree algorithm with indoor environmental data, energy consumption data, door opening and closing data, direct occupancy sensor data, and time-related data collected from office spaces for 16 days; they reported a much lower accuracy of 65%. Hailemariam et al. [26] collected data about indoor environments, energy consumption, and direct occupancy sensor from office spaces installed within a unit cubicle for 7 days to estimate the occupancy status using the decision tree algorithm, and reported the highest accuracy rate of 98.4%. Milenkovic and Amft [27] collected consumption and occupancy sensor data from private office rooms and open office space for 5 and 7 days, respectively, and estimated the occupancy status and the number of occupants using the layered hidden Markov model. They reported an accuracy rate of 87.0% and 78.0%, respectively.

Table 1 shows a summary of the data collection methods, types of gathered data, spatial and temporal resolutions, gathering periods, estimation algorithms, and the estimation accuracies of previous studies.

The literature review on occupancy estimation methods can be summarized as follows:

- An overwhelming majority of studies showed a sufficiently high accuracy rate of approximately 95% in terms of occupancy status (Level 1), meaning that those occupancy estimation methods using indirect sensors showed an insignificant difference in prediction performance compared to those using direct occupancy sensors;
- most of the studies collected data for a short period of less than one month and focused on the estimation accuracy of occupancy status and the number of occupants. However, there were not enough studies that utilized long-term measurement data to evaluate whether it could be possible to maintain accuracy in the event of seasonal changes. For instance, the correlation of indirect sensor data, such as energy consumption and window opening and closing data, with the accuracy of occupancy estimation can be changed in consideration of seasonal variations; and
- lastly, there were not enough studies that analyzed how the accuracy of occupancy estimation could change the energy consumption in the context of time series variations of occupancy and its related variables. It is believed that such studies can provide a significant impact on more accurate building energy estimation and more precise building system control.

Table 1. Summary of occupancy estimation studies in terms of input data gathering methods, time/spatial resolutions, classification algorithms, data gathering methods for ground truth, and estimation accuracies.

Ref No.	Resolution			Accuracy	Classification Algorithm	Ground Truth	Data Gathering					Data Collecting Period
							Virtual Sensor			Direct Occupancy Sensor	Time Information	
	Occupancy	Spatial	Temporal				Environment	Energy Usage	Contextual Information			
[11]	Level 1, 2	Room	Min.	93.5%, 74.2%	ELM, ANN, SVM, KNN, LDA, CART	Camera	Temperature, RH, CO_2, Air-Pressure					30 days
[13]	Level 1, 2	Room	Min.	98.2%, 97.8%	SNM, KNN, ANN, NB, TAN, DT	Camera Touchscreen	Temperature, RH, CO_2, Light, Sound		Door	Motion, Infrared		20 days
[15]	Level 1, 2	Room	Min.	95.5%, 78.0%	SVM, KNN	Camera, Observation	Temperature, RH, Light, Sound			PIR	Meeting schedule	10 days
[12]	Level 1	Room	Min.	99.3%	LDA, CART, RF, GBM	Camera	Temperature. RH, Humidity ratio, CO_2, Light				Time stamp, Date stamp	
[14]	Level 2	Room	Min.	65–90%	SVM, ANN, HMM	Camera	Temperature, RH, CO_2, Light, Outdoor-Temperature, Sound, DewPoint, PM2.5, CO, TVOC			Motion		44 days
[24]	Level 1	Room	Min.	98.4%	Decision tree	Camera	CO_2, Light, Sound	Current (pc)		Motion		7 days
[18]	Level 2	Room	Min.	87.6%	Radial Basis Function (RBF) neural network	Camera Touchscreen	Temperature, RH, CO_2, Light, Sound			Motion, PIR		20 days
[19]	Level 1	Room	Sec.		Belief network	Manual, Camera			Outbound phone call	Motion		2 days
[20]	Level 1	Room	Sec.	97.0%	Bayesian network	Manual			Wi-Fi, Keyboard-mouse, Bluetooth, Door	Motion, Chair sensors		2 weeks
[16]	Level 2	Floor	Min.	80.0%	SVM, ANN, HMM	Video camera	Temperature, RH, CO_2, Light, Outdoor-Temperature, Sound, DewPoint, PM2.5, CO, TVOC			Motion		44 days
[17]	Level 2	Room	Min.	80.8%	SVM, HMM, Autoregressive Hidden Markov Model	Manual recording	Temperature, RH, CO_2			PIR		3 weeks
[21] [22]	Level 2	Open office	Min.	80.0%	M-FRNN, ANN, kNN, SVM	Camera	Temperature, RH, CO_2, CO, Pressure, Airflow				Time	9 days
[23] [24]	Level 2 Level 4	Rooms	Min.	92.8% 97.6%	kNN, CARM, RF, SVM, CNN	Controlled condition						2 days
[25]	Level 2	Room	Min.	65.0%	Decision tree	Recorded videos	Temperature, RH, CO_2, Light	Power (laptop)	Door, Window	Motion	Time stamp, Date Stamp	16 days
[27]	Level 1, 2	Room	Min.	87.0%, 78.0%	Layered Hidden Markov Model	Manual recording, Ultrasound range finder		Power (plug)		PIR		5 days, 7 days

In this study, long-term measured data that is explicitly and implicitly related to occupancy information of a private office space is used and processed to compare the performance of occupancy status estimation algorithms. Three machine learning algorithms are selected for the analysis. The analysis is implemented for each season to find which measured variables have a crucial impact on the occupancy estimation. Throughout the analysis processes, the development of a framework that could continuously process the time-series measurement data to provide occupancy estimation results into a building energy model has been implemented to predict building energy consumption. Lastly, this study shows how the accuracy of the building energy model could be improved if continuous occupancy information from the framework is provided in a real building energy model.

2. Data Collection and Preprocessing

2.1. Description on Target Space and Collected Data

This study selected an occupied private office space measuring 22.51 m^2 in area as an experimental space in which to collect direct and indirect occupancy-related data. Figure 1 shows the inside view of the target space. Table 2 shows general information about the target space and energy consuming devices. The light emitting diode (LED) lighting is turned on and off with manual switches and has a dimming controller. The electric heat pump (EHP) is used as the cooling and heating equipment. During the heating season, the steam radiator is primarily used, but when central heating is not supplied, individually controllable EHP and auxiliary electric heaters are used.

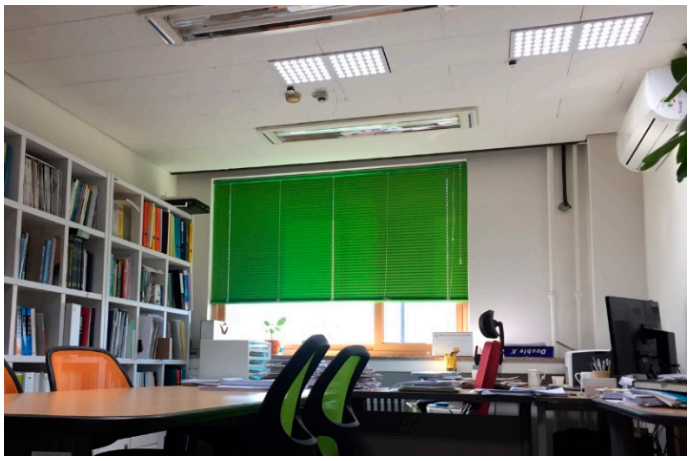

Figure 1. Inside view of the target space.

Table 2. General information on the target space and energy consuming devices.

Category	Description
Location	Cheongju-si, Republic of Korea
Room area	22.51 m^2
Room purpose	Private office
Occupant number	1 person
Lighting equipment	LED (auto dimming control)
Heating and cooling equipment	EHP, Auxiliary heater, Steam radiator
Office equipment	Desktop PC
Control	All equipment except the steam radiator is individually controlled by an occupant.

For indirect occupancy measurement, this study collected data regarding dry-bulb temperature, relative humidity, CO_2 level, illuminance, electricity consumption data for LED lights, desktop (PC), and electric heat pump (EHP), as well as occupancy data that is used as reference values to verify the

performance of occupancy estimation. Figure 2 and Table 3 offer details about sensors and stored data. The data collection activity lasted for approximately nine months, from 1 December 2016 to 30 August 2017, but energy consumption data was not measured for June and July due to a data logger malfunction. A total of 210,432 data points was collected for an actual period of seven months and used in this study. Because raw data had different collection time intervals depending on the type of data logger used, and because there were various types of missing or anomaly values, the data quality control pre-process was implemented with R language.

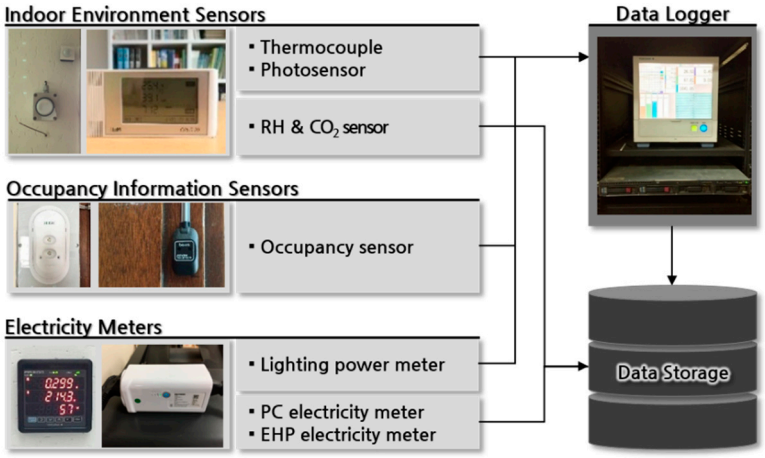

Figure 2. Status of indoor and outdoor environment, energy and occupancy sensor items, and data storage.

Table 3. Detailed information on measurement data and equipment.

Logger	Measuring Element	Sensor (Manufacture)	Resolution	Data Interval
Logger 1	Temperature Illuminance Lighting power Occupancy status	TX-FF-0.32-1P (FUKUDEN) HD2021T AA-SP (Deltaohm) PR300 (Yokogawa) PN1500 (Botem)	0.5 °C (at −25~100 °C) ±0.005 klux (at 0.02~2 klux) ±0.5 W 98.61%	1 min
Logger 2	Relative humidity CO_2 concentration	OPUS20 TCO (Lufft)	±2% RH ±50 ppm	15 min
Logger 3	EHP energy consumption PC energy consumption	Enertalk Plug (Encored Technologies)	±0.9%	1 s

2.2. Quality Control and Pre-Processing of Measurement Data

Because the collected data had different storage intervals that ranged from 1 s to 15 min depending on the logger used, they were converted into 15 min interval data and the time format was unified into a single format recommended by ISO 8601:2004 [28]. The missing values occurred at a ratio of 0.12–2.25% depending on the loggers, as shown in Table 4.

The missing data that occurred for more than 1 h were regarded as long-term missing values. These missing values were filled with average values of the same missing period of the previous weekday. In addition, the missing data that occurred for a continued period of less than 1 h were classified as short-term missing data. They were interpolated using the spline interpolation method [29]. The interpolation was implemented using Stats [30], a basic package of the R [31].

Table 4. Quality analysis of measurement data.

Measurement Elements	Total Number of Data Points	Number of Missing Data Points			Missing Rate		
		Total	Short-Term	Long-Term	Total	Short-Term	Long-Term
Temperature	26,304	593	522	71	2.25%	0.27%	1.99%
Relative humidity	26,304	32	22	10	0.12%	0.04%	0.08%
CO_2 concentration	26,304	32	22	10	0.12%	0.04%	0.08%
Illuminance	26,304	587	520	67	2.23%	0.25%	1.98%
Lighting power	26,304	587	520	67	2.23%	0.25%	1.98%
PC usage	26,304	172	126	46	0.48%	0.16%	0.32%
EHP usage	26,304	157	118	39	0.42%	0.13%	0.29%
Occupancy	26,304	587	520	67	2.23%	0.25%	1.98%

In addition to the collected data variables, derived variables, such as the change rates of temperature, relative humidity, and CO_2 concentration, were also considered to determine if they had a higher correlation with occupancy information than the measured values. Table 5 lists descriptions of the derived variables in reference to the studies by Zhang et al. [32] and Chen et al. [11]. In addition, by referring to the study by Candanedo and Feldheim [12], daily accumulated hours (unit: Second) and the hour of each measurement time were used as input variables. Finally, 21 input variables (seven basic variables measured by indirect sensors with 14 derived variables) were used for occupancy estimation analysis.

Table 5. Types and calculation methods of the derived variables.

Notation	Calculation	Description
FD1_	$raw(i) - raw(i-1)$	First order difference
SD_	$FD(i) - FD(i-1)$	Second order difference
FD2_	$raw(i) - raw(i-2)$	Variation of first order difference
MA1h_	$\left(\sum_{i-3}^{i} raw(i)\right)/4$	1-h moving average
CSD		Cumulative seconds of a day
Hour		Hour of the measured time

3. Development and Performance Analysis of Consecutive Occupancy Estimation Framework

Most of the previous studies on occupancy estimation have used classification algorithms and occupancy-related measurement data for less than one or two months. In this case, the correlation between the indirect sensor data with occupancy status may vary with seasonal changes, even in the same space. It is uncertain whether a trained occupancy estimation model based on short-term measured data can maintain accuracy in different seasonal periods. Therefore, a framework that can estimate occupancy information using long-term and time-series data is suggested by expanding the short-term occupancy estimation methodologies. For this work, Section 3.1 provides an outline of the selected occupancy estimation algorithms. The evaluation results for the performance of each algorithm for each season are described in Section 3.2. In Section 3.3, an occupancy estimation framework that can repeat training and estimation consecutively in a situation where time-series data are continuously provided over the entire period is suggested. In addition, the analysis results are compared with those from a short-term occupancy estimation performance. Figure 3 presents a schematic diagram of the processes conducted in Chapter 3.

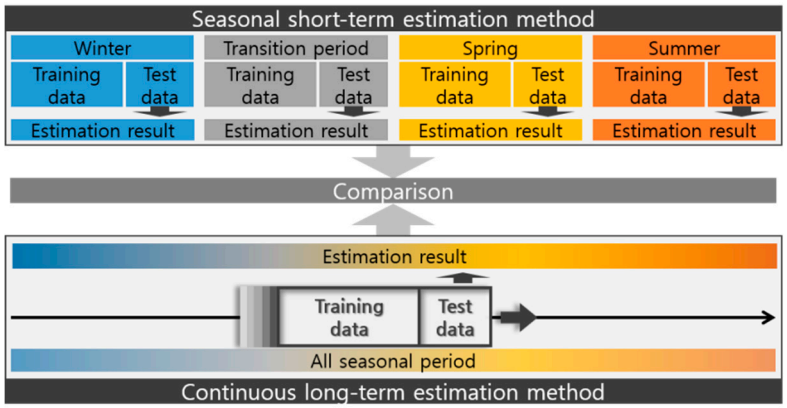

Figure 3. Schematic diagram of the comparison process between long-term and short-term data based occupancy estimation performances.

3.1. Selection of Occupancy Estimation Algorithms and Parameter Tuning

From the literature review, three classification algorithms widely used for occupancy estimation were selected: CART, SVM, and ANN. The CART classifies output values using input variables and reference values that can maximize the homogeneity of output values. The criterion that is used to determine the homogeneity of the output values before and after classification is called impurity. The smaller the impurity, the higher the homogeneity. At each classification stage, the impurity before and after classification is compared to evaluate the level of improvement in homogeneity, and then those input variables and their values that are considered the most appropriate as classification criteria are selected. By using the selected classification criteria, the data classification process is repeated at each stage. When it is completed, a classification model is created [33].

The SVM is characterized by its method of finding hyperplanes, which are boundaries that divide various types of data in multi-dimensional spaces. The SVM classifies data based on various hyperplanes to identify the most optimal hyperplane that can maximize the margin, which is the space between different types of data. In the case of a simple classification, data can be classified through a linear hyperplane. However, if data cannot be classified using a linear hyperplane, the data in a high-dimensional space is repositioned through the kernel function in order to determine the hyperplanes on the high dimensional space to classify the data. In this study, because the kernel function known as Gaussian RBF (radial basis function) is used, parameters, Gamma and C, are defined through the tuning process.

The ANN consists of an input layer where input variable values are entered, a hidden layer where the target values of output variables are adjusted by applying the weighted value of each variable, and an output layer where the final results are produced. Each layer can adjust the level of accuracy by changing the number of neurons. In addition, the hidden layer can have more than two layers [34].

In the case of the SVM and ANN, a parameter must be determined before the training of a classification model because the performance of the classification model, such as the estimation speed and accuracy, can be influenced [35]. To determine a proper parameter tuning value, a grid-search method with a cross verification was selected [36]. In the case of the SVM, by referring to the case of Hsu et al. [36], parameters were selected through the Grid-search method using 10-fold cross validation. The tuning parameters, C (cost) and γ (gamma), were determined from the ranges of 10^{-1}, 10, 10^{1}, 10^{2}, and 10^{3} and the ranges of $10^{-3}, 10^{-2}, 10^{-1}, 10, 10^{1}$, respectively. In the case of the ANN, the number of hidden layers and hidden neurons are determined by parameter tuning. In reference to the previous studies conducted using the ANN (Dong et al. [14]; Yang et al. [13]; Ekwevugbe et al. [37]; Yang et al. [18]; Jiang et al. [38]; Chen et al. [11]; Zuraimi et al. [39]; Li and Dong [40]), the parameters were determined through the Grid search by using 10-fold cross-validation on 10–50 units, with a

hidden layer at a one or two units and the hidden neuron at a 10 unit. Table 6 shows the final results of the determined parameter values that show the highest accuracy.

Table 6. Selection Results of parameter tuning of classification algorithms.

Seasonal Period	SVM		ANN	
	c	γ	Hidden Layer	Hidden Neuron
Winter	1	1	1	10
Transition period	10	10	1	10
Spring	10	10	2	30
Summer	0.1	0.1	1	10

3.2. Performance Evaluation of Seasonal Short-Term Occupancy Estimation

The data collected for seven months was divided into four seasonal periods, winter (December to March), transitional season (April), spring (May), and summer (August), in consideration of indoor and outdoor environment and electricity usage characteristics of devices. The selected final classification models with indirect measurement data of each season were analyzed in terms of performance of occupancy estimation and the selected input variables each season. The seasonal short-term occupancy estimation was conducted in the following process: Key variable selection, parameter tuning, training model, and testing the model and selecting the best model, as shown in Figure 4.

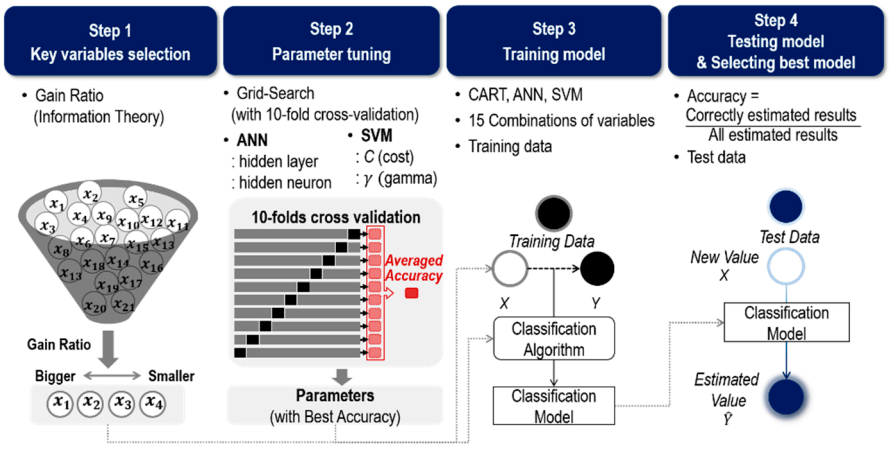

Figure 4. Diagram of seasonal short-term occupancy estimation model selection process.

3.2.1. Selection of Key Input Variables

If the number of input variables is "n", the training and verification processes of the classification models must be carried out $2^n - 1$ times to consider all possible combination models. This might require excessive calculation time depending on the number of input variables. Therefore, it is necessary to filter out those variables that have relatively less influence on occupancy estimation. For the selection method of key variables, the entropy-based information theory method, which evaluates the impurity of output values within a dataset, was introduced. The information theory method has been used in recent studies for a similar purpose (Dong et al. [14]; Zhang et al. [32]; Ekwevugbe et al. [37]; Yang et al. [18]; Arora et al. [25]; Amayri et al. [41]; Ryu and Moon [42]; Masood et al. [43]). In this study, the gain ratio, one of the information theory methods, was used. Equations (1)–(3) show the calculation methods of the gain ratio in the classification method (Han et al. [44]) used in this research:

$$\text{Entropy} = -\sum_{i=1}^{m} p_i \log_2(p_i), \quad (1)$$

$$\text{Information gain} = \text{Entropy}(D) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Entropy}(D_j), \quad (2)$$

$$\text{Gain ratio} = \frac{\text{Information gain}}{-\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2\left(\frac{|D_j|}{|D|}\right)} \quad (3)$$

where i is the number of types of output values, and j is the number of types of input variables. In addition, because the output variable in this study is the state of occupancy, i has a value of either 1 or 2. Because a total of 21 input variables are used, j is an integer between 1 and 21. p_i is the probability that the output value of data selected randomly among the dataset, D, might be the same as the value of the i-th output variable. D_j is a dataset that is classified by the j-th input variable. If the gain ratio has a greater value, it means it has a greater influence on the classification of output variables. After calculating the gain ratio values of 21 input variables by season, four variables with the highest influence were selected as the input variables for the seasonal classification models. Table 7 shows the rankings of the seasonal input variables based on their gain ratio values.

Table 7. Evaluation results of seasonal input variables' influence based on the gain ratio.

Rank	Winter		Transition_Period		Spring		Summer	
1	Light_Power (W)	0.6103	PC_Usage (Wh)	0.4330	PC_Usage (Wh)	0.3380	Light_Power (W)	0.7868
2	PC_Usage (Wh)	0.3140	EHP_Usage (Wh)	0.3988	EHP_Usage (Wh)	0.3372	EHP_Usage (Wh)	0.7283
3	ILLUM (lux)	0.2354	Light_Power (W)	0.2305	ILLUM (lux)	0.2726	ILLUM (lux)	0.2718
4	TEMP (°C)	0.1355	ILLUM (lux)	0.2267	Light_Power (W)	0.2631	CO_2 (ppm)	0.2635
5	MA1h_TEMP	0.1243	CO_2 (ppm)	0.1791	MA1h_RH	0.2487	PC_Usage (Wh)	0.2436
6	FD1_CO_2	0.1195	MA1h_CO_2	0.1455	FD2_CO_2	0.1460	FD2_ CO_2	0.2022
7	CO_2 (ppm)	0.1095	TEMP (°C)	0.1335	RH (%)	0.1364	MA1h_CO_2	0.1965
8	EHP_Usage (Wh)	0.1002	MA1h_TEMP	0.1281	FD1_CO_2	0.1188	FD1_CO_2	0.1832
9	FD2_CO_2	0.0958	FD1_CO_2	0.1186	CO_2 (ppm)	0.1182	TEMP (°C)	0.1257
10	MA1h_CO_2	0.0941	CSD	0.1121	MA1h_CO_2	0.0941	MA1h_TEMP	0.1174
11	SD1_CO_2	0.0791	FD2_CO_2	0.0965	CSD	0.0911	SD1_TEMP	0.0959
12	HOUR	0.0701	SD1_CO_2	0.0810	HOUR	0.0888	SD1_CO_2	0.0827
13	CSD	0.0682	HOUR	0.0716	SD1_CO_2	0.0852	FD2_TEMP	0.0752
14	FD1_TEMP	0.0642	SD1_TEMP	0.0715	FD2_TEMP	0.0772	SD1_RH	0.0743
15	SD1_TEMP	0.0635	FD2_TEMP	0.0688	SD1_TEMP	0.0754	FD1_TEMP	0.0694
16	FD2_TEMP	0.0619	FD1_TEMP	0.0661	FD1_TEMP	0.0717	FD1_RH	0.0632
17	SD1_RH	0.0470	FD2_RH	0.0385	TEMP (°C)	0.0651	FD2_RH	0.0626
18	FD1_RH	0.0412	FD1_RH	0.0318	MA1h_TEMP	0.0628	HOUR	0.0502
19	FD2_RH	0.0340	RH (%)	0.0187	FD2_RH	0.0370	RH (%)	0.0499
20	MA1h_RH	0.0150	MA1h_RH	0.0175	SD1_RH	0.0301	CSD	0.0485
21	RH (%)	0.0138	SD1_RH	0.0130	FD1_RH	0.0251	MA1h_RH	0.0409

The four input variables with the highest gain ratio values for each season are all measured variables, and those that are derived from the measured variables and derived from time information were not adopted. The two top ranked variables of each season are the variables that measured energy usage from LED, PC, and EHP.

The Lights_Power parameter was the greatest influential input variable during winter and summer, but its influence was reduced during the transitional period and spring. This can be interpreted as the result of changes in the behavioral patterns of the occupant. For instance, if the occupant was not present in the target space during winter and summer vacations, the lights were turned off for most of the unoccupied times. During the transitional season and spring, when the semester was in session, the occupant frequently left the target space due to lectures or short meetings, so the space was often left vacant with the lights on. Similarly, the EHP had a relatively lower influence during winter because the central heating radiator and auxiliary space heater were used together with the EHP, depending on the central heater operation time.

3.2.2. Training and Verification of Classification Models

After the training and verification processes of 15 ($2^4 - 1$) classification models for each season, the models with the highest level of accuracy were selected as a final classification model. In each

process, the first two thirds of data of each season were used for training and the remaining one third were used for verification of the trained classification models. R was used for the training of the classification models with classification algorithms. In detail, the rpart package was used for the CART (Therneau et al. [45]; Therneau et al. [46]), the e1071 package was used for the SVM (Meyer et al. [47]; Karatzoglou et al. [48]), and the neuralnet package was used for the ANN (Fritsch et al. [49]; Gunther and Fritsch [50]).

Table 8 summarizes the selected classification models with the highest overall accuracy throughout the training and verification processes for each season. Figure 5 shows the daily accuracy of the selected seasonal models per classification algorithms. In terms of the overall accuracy of each classification algorithm, all algorithms show the highest overall accuracy in summer, followed by the transitional period, winter, and spring. Regarding the seasonal accuracy of the algorithms, the CART shows the highest accuracy during winter, the transitional period, and summer. The SVM shows a relatively higher accuracy in spring compared to the other algorithms. The ANN is the most inefficient in terms of computation time and accuracy.

Table 8. Selected final classification models and their accuracies for classification algorithms by seasons.

Algorithm		Winter	Transition Period	Spring	Summer
CART	Model	Light_Power + PC_Usage + TEMP	EHP_Usage + PC_Usage + ILLUM	EHP_Usage + PC_Usage	Light_Power
	Accuracy	94.58%	97.19%	91.33%	97.18%
SVM	Model	Light_Power + PC_Usage + TEMP	EHP_Usage + PC_Usage	PC_Usage + ILLUM	Light_Power
	Accuracy	93.26%	96.25%	93.55%	97.08%
ANN	Model	ILLUM + PC_Usage + TEMP	PC_Usage + ILLUM	EHP_Usage + PC_Usage	Light_Power + PC_Usage
	Accuracy	93.03%	96.46%	90.52%	97.08%

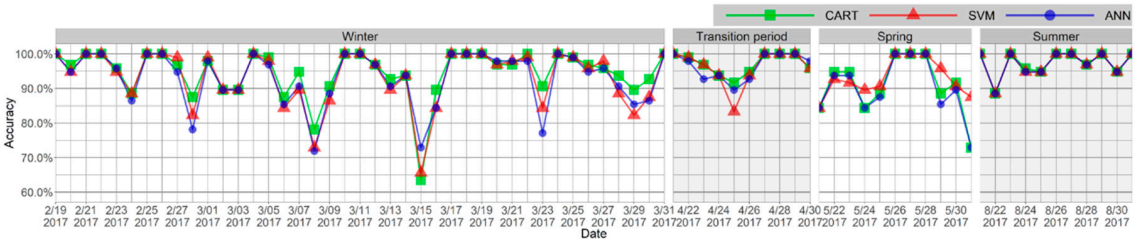

Figure 5. Seasonal changes of daily accuracy of selected final occupancy estimation models with three classification algorithms.

In terms of accuracy consistency for all seasons, the SVM model maintains a higher level of accuracy while the other models show a substantially lower accuracy in spring. The low accuracies are incurred as a result of the fact that there were some days during winter and spring when the accuracy level drops below 80%. This is because the occupant vacated the target space for a long period on the days without turning off lights and the PC, meaning the space usage characteristics on that day deviated far from the typical usage characteristics.

3.3. Framework Development for Consecutive Occupancy Estimation with Time-Series Data

If the predicted energy demand is based on actual occupancy information, the data would be more reliable and useful for advanced building energy management and many other applications. For this result, a consecutive occupancy estimation with time-series data is needed. Because the variables that have a significant impact on the occupancy estimation may vary at any moment, as seen in Section 3.2, a consecutive occupancy estimation framework is suggested for dealing with long-term time-series data. For this framework development, a moving window concept was introduced and occupancy estimation over the entire measurement period was conducted by repeating the variables selection,

training model, final model selection, and verification process. Figure 6 depicts the consecutive occupancy estimation scheme for time-series data processing.

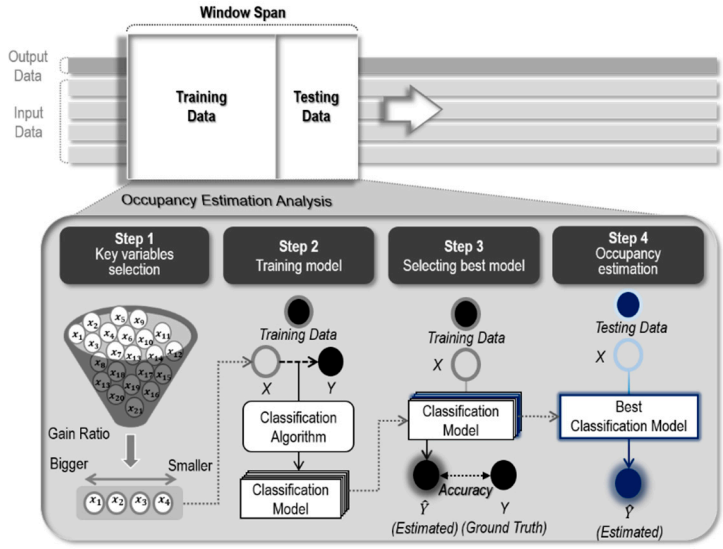

Figure 6. Diagram of the consecutive long-term occupancy estimation framework.

Although this framework is based mainly on the short-term occupancy estimation process conducted in Section 3.2, the key variable selection process of the classification algorithm and final model selection process are different from the short-term data-based process. First, it is difficult to repeat the grid-search process each time if it has to be processed every time-step, as it requires too much calculation time. Therefore, the results listed in Table 6 were utilized in this framework. Second, the best model was selected by using the training error data, which was derived from the verification result with training data because obtaining verification data in an actual situation was not available.

The training period of the window span was selected as 20 days, which is the same period as the training period of the seasonal occupancy estimation. From the seven months of data, the data measured from 1 December 2016 to 20 December 2016 were selected for initial training, and the evaluation of the occupancy estimation performance was conducted on the data collected from 21 December 2016 to 31 August 2017 (for 193 days in total). The ANN algorithm was excluded because of its relatively lower accuracy and excessively long calculation time.

3.3.1. Selection of Verification Period and Window Moving Interval

The window span is combined with a training period and verification period. After the verification period is completed at every time step, the window moves to the next time step, determines a new best model again with the data in the new window span, and conducts the occupancy estimation continuously. For the moving interval period, 15 min was selected for the next estimation because the measured data set had a 15 min time step. However, a 1-day interval, which means the occupancy estimation data is updated once a day, was also considered to test the potential of minimizing training frequency as much as possible. The occupancy estimation performance with the two moving intervals are compared in this section.

Table 9 summarizes overall accuracy and standard deviation when the occupancy estimation was conducted with the moving interval of 15-min and 1-day for the entire periods. Figure 7 shows daily accuracy variations of the same estimation results. In the 15-min moving interval case, the CART and SVM shows an overall accuracy of 95.59% and 95.44%, respectively, which are 1–2% higher than those of the 1-day moving interval (93.84% and 94.55%). Standard deviations of the 15 min case were estimated as 5.24% (CART) and 5.60% (SVM), which are lower than those of the 1-day case of

11.85% (CART) and 6.90% (SVM). Particularly for the case of the CART algorithm in the 1-day case, the performance was significantly reduced when exceptional events (right after long vacation on May and August in 2017) occurred.

Table 9. Overall accuracy and standard deviation by classification algorithms and window moving intervals.

Algorithm	Moving	Overall Accuracy	Standard Deviation
CART	15 min	95.59%	5.24%
	1 day	93.84%	11.85%
SVM	15 min	95.44%	5.60%
	1 day	94.55%	6.90%

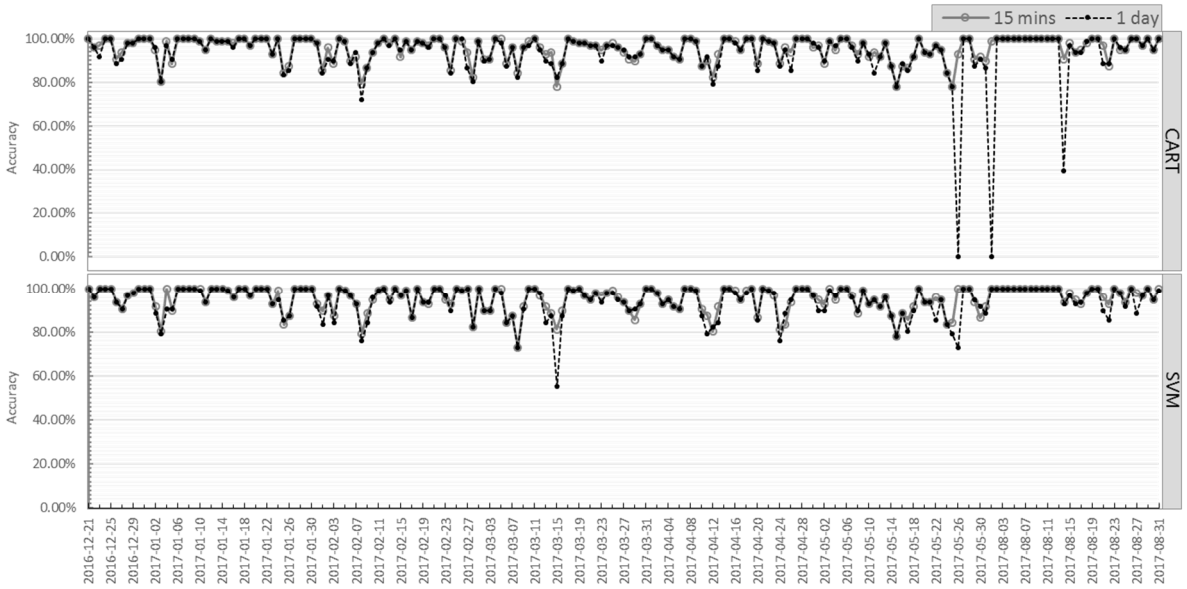

Figure 7. Comparison of daily accuracy of occupancy estimation by classification algorithms and window moving intervals.

A more detailed analysis was conducted to investigate how the above results were achieved. Two days (8th February and 1st August) with a daily accuracy of less than 80% were selected. Figure 8 shows the selected final classification models, as well as the accuracies of the estimation every 15 min. The legend lists the selected final classification models. If the occupancy estimation result is accurate, the value in each time step is one, and if the results are not accurate, the value is zero. On 8th February, the 15-min case made two incorrect estimations around 1 PM, but it produced more accurate estimations than the 1-day case after that time for the rest of the day. The figure also shows that in the consecutive process of such estimations, the 15 min interval estimation case continuously changes the final estimation model among "ILLUM + Light_Power + PC_Usage", "ILLUM + Light_Power + MA1h_ TEMP + PC_Usage", and "ILLUM + PC_Usage + TEMP" with respect to situation change.

The 15-min case shows a higher daily accuracy than that of the 1-day case by considering recent situations for the consecutive training process. This means that variations of the impact of input variables were considered in the selection of a new model. This is more apparent in the results obtained on 1 August 2017. The EHP_Usage+Light_Power + PC_Usage model was used from 00:00 to 11:30 for the 15-min case and the 15-min case continued to produce accurate results from 11:30 to 23:45 by changing the input variables of the selected final model. Meanwhile, the 1-day case kept producing inaccurate results.

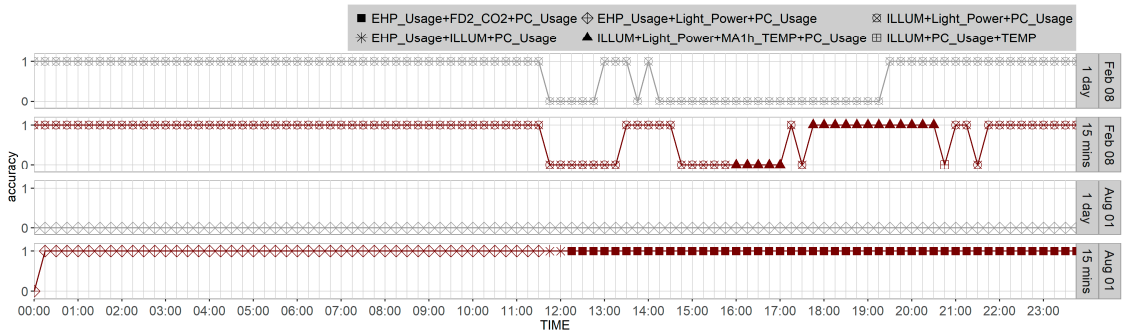

Figure 8. Comparison of estimation results and final selected classification models between 1-day and 15-min window moving interval cases with the CART algorithm.

As a result, the 15-min case was selected with a window span of 20 days and 15 min (a training period of 20 days and a verification period of 15 min). The selection, training, and verification of the occupancy estimation model process was carried out a total of 18,528 times by moving the window span at an interval of 15 min for 193 days.

3.3.2. Performance Comparison between Seasonal Short-Term and Consecutive Long-Term Occupancy Estimations

Table 10 compares the accuracy results of the consecutive long-term estimation with those of the seasonal short-term occupancy estimation for each season and overall experiment period in the last column. Figure 9 displays the same results with daily accuracy. The solid line in the figure represents the daily accuracy of the consecutive long-term estimation, while the dotted line indicates the daily accuracy of the seasonal short-term occupancy estimation.

Table 10. Comparison of seasonal and overall accuracy between short-term and long-term occupancy estimation methods.

Method	Algorithm	Winter	Transition Period	Spring	Summer	All Period
Seasonal short-term estimation	CART	94.58%	97.19%	91.33%	97.18%	94.85%
	SVM	93.26%	96.25%	93.55%	97.08%	94.28%
Continuous long-term estimation	CART	95.62%	95.76%	92.84%	98.05%	95.59%
	SVM	95.41%	94.90%	93.25%	98.29%	95.44%

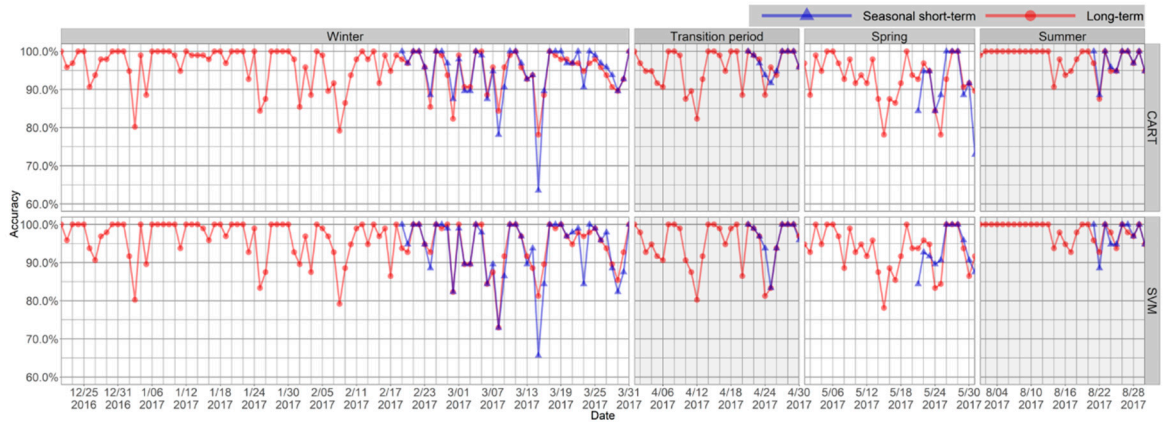

Figure 9. Comparison of daily accuracy over the entire period between short-term and long-term occupancy estimation methods.

The CART and SVM algorithms produced an overall accuracy for the entire period of 95.59% and 95.44%, respectively, which is insignificantly different from the overall accuracy of the seasonal occupancy estimation of 94.85% (CART) and 94.28%(SVM). In each seasonal accuracy comparison, the differences between the short-term and long-term estimation methods are below ±2.15%, and the long-term occupancy estimation shows a higher level of accuracy than the seasonal short-term occupancy estimation except for the transitional period. Figure 9 shows the daily accuracy of the long-term occupancy estimation methods against those of the seasonal estimation methods for the two algorithms.

4. Performance Evaluation of Building Energy Consumption with the Occupancy Estimation Data

Goyal et al. [51] and Zhang et al. [52] found that occupancy information based simple feedback control of a variable air volume (VAV) box could achieve comparable saving against complex MPC control [51] and energy saving from lighting and VAV control could bring up to 23% savings [52]. The two studies used hypothetical occupancy information or direct measurement data while this research uses indirectly estimated data, which could generate realistic occupancy information easily.

A simple building energy model, but actual occupancy information-based simulation environment, was developed to evaluate the impact of actual occupancy information on building energy use. To create a more realistic energy simulation, the occupant-related input variables, such as lighting control, electricity device operation, and HVAC operation, were modeled based on occupant activities. Obtaining that actual information was not available in this research scope, so a simple assumption was applied because the target building is a unit of private space: If there is an occupant, the light, electrical devices, and HVAC work.

4.1. Establishment of Simulation Environment

4.1.1. BCVTB, R-script, and EnergyPlus Models with Occupancy Data

To provide the actual occupancy estimation results from Chapter 3.3 with the EnergyPlus energy model of the target space, the Building Control Virtual Test Bed (BCVTB) and R-script were combined. Figure 10 shows an established BCVTB simulation environment. The delivery of occupancy information was carried out at every simulation time step, and the delivered occupancy information was used in the Number of People Schedule, Lights Schedule, Electric Equipment Schedule, and Air Loop HVAC Availability Schedule in EnergyPlus.

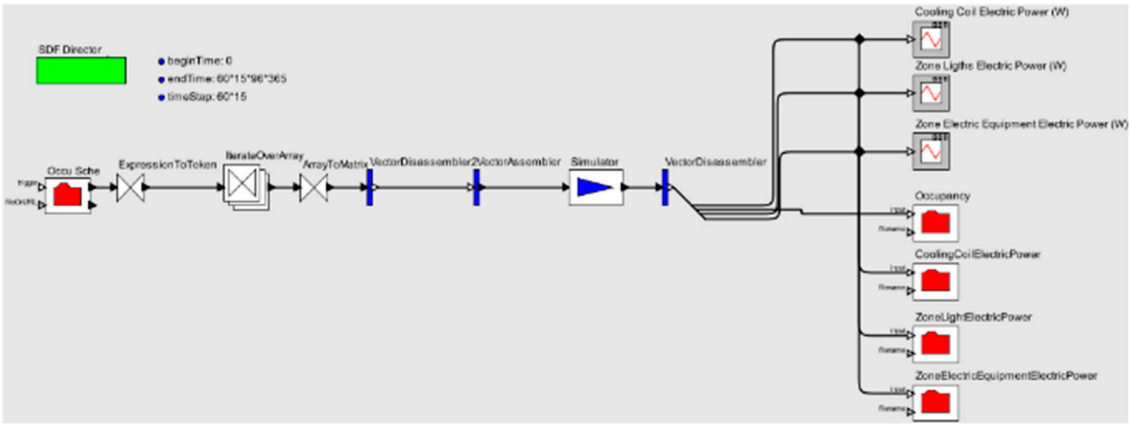

Figure 10. Developed actual occupancy based energy simulation environment on BCVTB.

4.1.2. Actual Meteorological Data for epw Input

The meteorological data used in the simulation is created using hourly observation data provided by the Korea Meteorological Administration (KMA). The downloaded weather data was converted to a

DOE weather data file (.bin) based on DOE-2.1A Reference Manual [53]. Then, the Weather Converter Tool provided by EnergyPlus was utilized to convert a DOE weather data file (.bin) into an EnergyPlus weather data (epw) file.

4.1.3. EnergyPlus Energy Model for the Target Space

The building energy model for the target space was developed using EnergyPlus version 8.1.0 based on the general building and equipment information listed in Table 11. Three electric devices, including dimmable LED lights, PC, and EHP (Electric Heat Pump) were modeled. Electricity consumption data of each device was recorded for the calibration and evaluation of the energy model. Detailed inputs of LED dimming characteristics and part load performance characteristics of EHP were not considered for modeling simplicity.

Table 11. Summary of building energy model input variables.

Variables	Input Value	
Space Info.	4.0 × 6.0 × 2.7 (m)	
Window Info.	2.0 × 1.5 (m), facing south-west	
U-Value	Wall	0.509 W/m^2·K
	Window	3.159 W/m^2·K
Interior Shade Status	100% closed blinds with 30% area retracted	
Cooling Equipment	EHP	3500 W (efficiency: 5.4 W/W)
	EHP	4300 W (efficiency: 2.9 W/W)
Heating Equipment	Steam Radiator (Central)	
	Resistive Heater (Auxiliary)	
Lighting	LED	40 W × 3 EA, Dimming
Plug Load	PC	111 W

4.2. Comparison of Energy Consumption Estimation Results

The building energy simulations with the estimated occupancy schedule measured against the reference schedule were implemented and compared to measure how the actual occupancy schedule improved the performance of the energy model. The energy consumption estimation using the actual occupancy schedule was carried out with the CART algorithm. The algorithm is considered easy to apply to an actual building application because of its relatively shorter calculation time. In case of the reference schedule, a small-scale office building schedule provided by the National Renewable Energy Laboratory (NREL) [54] because it is a prototypical schedule and is considered as conventional practice when actual information is not available for small offices. Figure 11 shows a comparison result of electricity consumption with those two types of schedule against measured energy consumption of each device for 7 days during summer.

In Figure 11, the square line indicates the measured electricity consumption data, while the circle line indicates the results of the simulation with the reference schedule, and the triangle line indicates the results of the simulation with the estimated occupancy schedule. As a matter of fact, the reference schedule case shows energy consumption based on working hours for weekdays regardless of occupancy status. However, the estimated occupancy case shows good agreement with the measured data of Elec. Equipment (PC) and this case works well, even on the weekend (20 August 2017). In the case of Lights and EHP energy consumption, the estimated occupancy case fails to predict peak energy consumption due to dimming and part-load operation features of the LED and EHP, respectively. However, the electricity consumption profiles follow very closely with the actual occupancy status.

To quantify the prediction performance of the estimated occupancy case, the root mean squared error (RMSE) and mean bias error (MBE) for each season were calculated as seen in Table 12. In terms of MBE, the estimated occupancy case performs worse than the reference schedule case due to underestimation problems. However, as the low MBEs in the reference schedule case were achieved through a summation effect of ultimately low and high energy consumption prediction, the

underestimation problem in the estimated occupancy case is not worse than the reference schedule case. In terms of RMSE, the estimated occupancy case shows a 3–88% improvement with respect to the season and electric devices. On an annual basis, the case shows a 17–33% improvement with respect to the devices. Figure 12 presents a scatter plot of the estimated electricity consumption based on the two schedules compared against the measured electricity consumption. Obviously, the estimated occupancy case (circle dot) improves energy prediction dramatically in terms of actual energy consumption per time step.

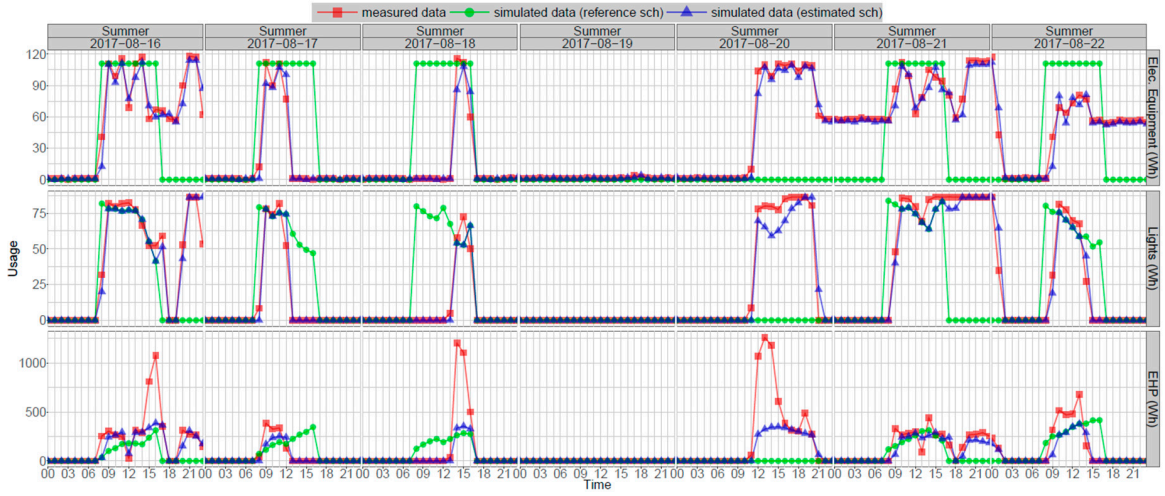

Figure 11. Comparison results of the energy simulation using the reference and estimated schedules against the measured electricity consumption for 16 August 2017~22 August 2017.

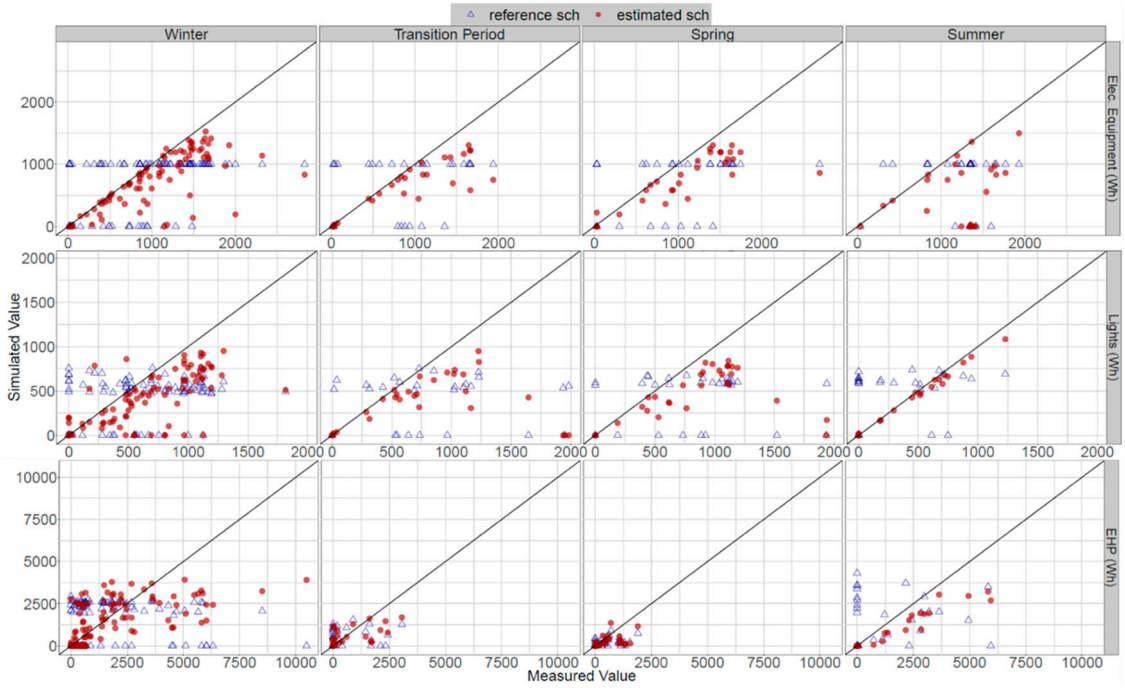

Figure 12. Scatter plot of electricity consumption with the reference and estimated schedules with respect to season and device.

Table 12. Margin of errors in energy consumption estimation results using the reference and Assumed occupancy schedules

System	Schedule	Winter		Transition Period		Spring		Summer		All Period	
		RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE
Elec. Equipment (kWh)	reference sch	0.64	−0.16	0.68	−0.19	0.71	−0.32	0.76	−0.49	0.68	−0.24
	estimated sch	0.45	−0.25	0.39	−0.25	0.48	−0.31	0.98	−0.80	0.56	−0.35
Lights (kWh)	reference sch	0.45	−0.19	0.73	−0.41	0.68	−0.40	0.42	0.15	0.54	−0.20
	estimated sch	0.36	−0.22	0.71	−0.41	0.58	−0.37	0.05	−0.03	0.45	−0.24
EHP (kWh)	reference sch	2.48	−0.20	1.02	−0.06	0.61	−0.31	2.25	0.22	2.06	−0.13
	estimated sch	1.73	−0.38	0.72	−0.14	0.56	−0.29	1.03	−0.61	1.37	−0.36

5. Results, Summary and Discussion

The major outcomes and contributions of this study are as follows.

1. This study identified the advantages and disadvantages of direct and indirect sensing methods to estimate the occupancy status through analysis of the previous research literature and actual comparison of estimation performance with some classification algorithms. As a result, this study found that utilizing indirect measurement data could yield approximately 95% of occupancy estimation accuracy, which is comparable with direct measurement data.

2. Most previous studies used short-term measured data for a period of a few months to develop occupancy estimation models. However, Tables 7 and 8 indicate the necessity of developing a new occupancy estimation model through the reselection of input variables each season and situation, because the influence of input variables on estimation accuracy may vary depending on the indoor and outdoor environment and operation of electric devices. It was observed that the performance of the selected final classification models for each season had more explanatory power by changing the decisive input variables for the models.

3. Therefore, to improve the prediction performance of energy models, the estimated occupancy information should be utilized rather than conventional typical occupancy schedules, and occupancy estimation must be continuously performed at every time step to feed the estimated results into an energy model. To enable this simulation, a framework for consecutive occupancy estimation with time-series data for the entire simulation run period was developed and suggested. The moving window concept was introduced in the framework to repeat occupancy model selection, training, and estimation at every time step in a continuous way.

4. The conducted occupancy estimation using nine months of long-term data through the suggested framework showed that the overall accuracy over the entire experimental period was calculated at 95.59% (CART) and 95.44% (SVM), while the average season accuracy of the occupancy estimation was calculated at 94.85% (CART) and 94.28% (SVM). Even the seasonal accuracy between the short-term and long-term estimation was similar to each other, showing less than a ±2.5% accuracy difference.

5. The energy model simulation study revealed that the estimated occupancy case improved the energy consumption prediction performance by 17–33% in RMSE relative to the reference schedule case. The scatterplot and line graphs showed more improvement than the number by illustrating that the estimated occupancy case model can reflect actual building energy profiles.

6. Conclusions and Future Work

To improve the energy consumption estimation accuracy of a building energy model for a certain building, the occupancy status has a significant influence and must be considered. In this study, a framework for a consecutive occupancy estimation was developed and its estimation performance was evaluated. The suggested framework is distinguished from previous similar studies by revealing that input variables for occupancy estimation model could be occasionally changed by an occupant in reaction to certain times and seasons. Finally, the estimated occupancy-data-based building energy

simulation validated an improvement in the performance of energy consumption estimation by showing that it is very close to actual energy usage profiles. It is expected that this research result and suggested framework can be utilized in various fields, such as building energy modeling, building systems control, building energy management, and facility management.

This study used a private office as a target space for occupancy and energy estimation. It is necessary to expand the proposed framework to an open office space, which is more common in office buildings. As seasonal variations required a new occupancy estimation model, the change in target space will also require a performance test of the suggested framework.

Regarding energy predicting performance with the estimated occupancy schedule, two hurdles were found. The first hurdle is improving occupancy estimation resolution from occupancy status to more detailed information, such as the number of occupants and activity of occupants. However, gaining more detailed occupant information seems very difficult in terms of feasibility and privacy issues. The second hurdle is connecting the occupant information with the energy model input. The current assumption of directly connecting occupancy status with every energy consuming device is not always true, even in open-space offices. Overcoming these two problems would be very crucial for future work to achieve better performance in building energy prediction.

Author Contributions: In this research activity, D.S. proposed core concept and methodology and edited the original draft; S.K. implemented data preprocessing, theory review, development and analysis of software, and an original draft preparation; Y.S. (Yujin Song) and Y.S. (Yoondong Sung) supervised overall analysis and validation process.

Acknowledgments: This work was supported by the Korea Institute of Energy Technology Evaluation and Planning(KETEP) grant funded by the Korea government(MOTIE) (20158530050160, "Joint Advanced Microgrid Analysis, Design, and Implementation at Military Installations in Korea").

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

References

1. Delzendeh, E.; Wu, S.; Lee, A.; Zhou, Y. The impact of occupants' behaviours on building energy analysis: A research review. *Renew. Sustain. Energy Rev.* **2017**, *80*, 1061–1071. [CrossRef]
2. Paone, A.; Bacher, J.-P. The Impact of Building Occupant Behavior on Energy Efficiency and Methods to Influence It: A Review of the State of the Art. *Energies* **2018**, *11*, 953. [CrossRef]
3. Melfi, R.; Rosenblum, B.; Nordman, B.; Christensen, K. Measuring building occupancy using existing network infrastructure. In Proceedings of the 2011 International Green Computing Conference and Work-shops (IGCC), Orlando, FL, USA, 25–28 July 2011.
4. Garg, V.; Bansal, N.K. Smart occupancy sensors to reduce energy consumption. *Energy Build.* **2000**, *32*, 81–87. [CrossRef]
5. Nagy, Z.; Yong, F.Y.; Frei, M.; Schlueter, A. Occupant centered lighting control for comfort and energy efficient building operation. *Energy Build.* **2015**, *94*, 100–108. [CrossRef]
6. Gunay, H.B.; Fuller, A.; O'Brien, W.; Beausoleil-Morrison, I. Detecting occupants' presence in office spaces: A case study. In Proceedings of the eSim 2016, Hamilton, OH, USA, 3–4 May 2016. [CrossRef]
7. Benezeth, Y.; Laurent, H.; Emile, B.; Rosenberger, C. Towards a sensor for detecting human presence and characterizing activity. *Energy Build.* **2011**, *43*, 305–314. [CrossRef]
8. Zou, J.; Zhao, Q.; Yang, W.; Wang, F. Occupancy detection in the office by analyzing surveillance videos and its application to building energy conservation. *Energy Build.* **2017**, *152*, 385–398. [CrossRef]
9. Amin, I.J.; Taylor, A.J.; Junejo, F.; Al-Habaibeh, A.; Parkin, R.M. Automated people-counting by using low-resolution infrared and visual cameras. *Measurement* **2008**, *41*, 589–599. [CrossRef]
10. Shen, W.; Newsham, G.; Gunay, B. Leveraging existing occupancy-related data for optimal control of commercial office buildings: A review. *Adv. Eng. Inform.* **2017**, *33*, 230–242. [CrossRef]
11. Chen, Z.; Masood, M.K.; Soh, Y.C. A fusion framework for occupancy estimation in office buildings based on environmental sensor data. *Energy Build.* **2016**, *133*, 790–798. [CrossRef]
12. Candanedo, L.M.; Feldheim, V. Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models. *Energy Build.* **2016**, *112*, 28–39. [CrossRef]

13. Yang, Z.; Li, N.; Becerik-Gerber, B.; Orosz, M. A multi-sensor based occupancy estimation model for supporting demand driven HVAC operations. In Proceedings of the 2012 Symposium on Simulation for Architecture and Urban Design. Society for Computer Simulation International, Orlando, FL, USA, 26–30 March 2012.
14. Dong, B.; Andrews, B.; Lam, K.P.; Höynck, M.; Zhang, R.; Chiou, Y.S.; Benitez, D. An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network. *Energy Build.* **2010**, *42*, 1038–1046. [CrossRef]
15. Khan, A.; Nicholson, J.; Mellor, S.; Jackson, D.; Ladha, K.; Ladha, C.; Oliver, P.; Plötz, T. Occupancy monitoring using environmental & context sensors and a hierarchical analysis framework. In Proceedings of the BuildSys@ SenSys, Memphis, TN, USA, 5–6 November 2014; pp. 90–99. [CrossRef]
16. Lam, K.P.; Höynck, M.; Dong, B.; Andrews, B.; Chiou, Y.S.; Zhang, R.; Benitez, D.; Choi, J. Occupancy detection through an extensive environmental sensor network in an open-plan office building. In Proceedings of the 11th International IBPSA Conference, Glasgow, Scotland, 27–30 July 2009; pp. 1452–1459.
17. Han, Z.; Gao, R.X.; Fan, Z. Occupancy and indoor environment quality sensing for smart buildings. In Proceedings of the 2012 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Graz, Austria, 13–16 May 2012; pp. 882–887.
18. Yang, Z.; Li, N.; Becerik-Gerber, B.; Orosz, M. A systematic approach to occupancy modeling in ambient sensor-rich buildings. *Simulation* **2014**, *90*, 960–977. [CrossRef]
19. Dodier, R.H.; Henze, G.P.; Tiller, D.K.; Guo, X. Building occupancy detection through sensor belief networks. *Energy Build.* **2006**, *38*, 1033–1043. [CrossRef]
20. Zhao, Y.; Zeiler, W.; Boxem, G.; Labeodan, T. Virtual occupancy sensors for real-time occupancy information in buildings. *Building Environ.* **2015**, *93*, 9–20. [CrossRef]
21. Wang, W.; Chen, J.; Hong, T.; Zhu, N. Occupancy prediction through Markov based feedback recurrent neural network (M-FRNN) algorithm with WiFi probe technology. *Building Environ.* **2018**, *138*, 160–170. [CrossRef]
22. Wang, W.; Chen, J.; Hong, T. Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings. *Autom. Construct.* **2018**, *94*, 233–243. [CrossRef]
23. Zou, H.; Zhou, Y.; Yang, J.; Spanos, C. Towards occupant activity driven smart buildings via WiFi-enabled IoT devices and deep learning. *Energy Build.* **2018**, *177*, 12–22. [CrossRef]
24. Zou, H.; Zhou, Y.; Yang, J.; Spanos, C. Device-free occupancy detection and crowd counting in smart buildings with WiFi-enabled IoT. *Energy Build.* **2018**, *174*, 309–322. [CrossRef]
25. Arora, A.; Amayri, M.; Badarla, V.; Ploix, S.; Bandyopadhyay, S. Occupancy estimation using non-intrusive sensors in energy efficient buildings. In Proceedings of the 14th Conference of IBPSA, Hyderabad, India, 7–9 December 2015; pp. 1441–1448.
26. Hailemariam, E.; Goldstein, R.; Attar, R.; Khan, A. Real-time occupancy detection using decision trees with multiple sensor types. In Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design, Boston, MA, USA, 3–7 April 2011; pp. 141–148.
27. Milenkovic, M.; Amft, O. Recognizing energy-related activities using sensors commonly installed in office buildings. *Procedia Comput. Sci.* **2013**, *19*, 669–677. [CrossRef]
28. ISO 8601 Data Elements and Interchange Formats—Information Interchange—Representation of Dates and Times. 2004. Available online: https://www.iso.org/obp/ui/#iso:std:iso:8601:ed-3:v1:en (accessed on 7 November 2017).
29. McKinley, S.; Levine, M. *Cubic Spline Interpolation*; College of the Redwoods: Eureka, CA, USA, 1998; Volume 45, pp. 1049–1060.
30. Ihaka, R.; Gentleman, R. R: A language for data analysis and graphics. *J. Comput Graphical Stat.* **1996**, *5*, 299–314.
31. The R Core Team. R: A Language and Environment for Statistical Computing. Available online: https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf (accessed on 5 October 2017).
32. Zhang, R.; Lam, K.P.; Chiou, Y.S.; Dong, B. Information-theoretic environment features selection for occupancy detection in open office spaces. *Build. Simul.* **2012**, *5*, 179–188. [CrossRef]
33. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 1984.
34. Michell, T.M. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997; pp. 81–127.

35. Staelin, C. Parameter Selection for Support Vector Machines. Hewlett-Packard Company. Available online: http://www.hpl.hp.com/techreports/2002/HPL-2002-354R1.pdf (accessed on 15 October 2017).
36. Hsu, C.W.; Chang, C.C.; Lin, C.J. A Practical Guide to Support Vector Classification. Available online: http://www.csie.ntu.edu.tw/~{}cjlin/papers/guide/guide.pdf (accessed on 29 October 2017).
37. Ekwevugbe, T.; Brown, N.; Pakka, V. Real-Time Building Occupancy Sensing for Supporting Demand Driven HVAC Operations. In Proceedings of the 13th International Conference for Enhanced Building Operations, Montreal, QC, Canada, 8–10 October 2013.
38. Jiang, C.; Masood, M.K.; Soh, Y.C.; Li, H. Indoor occupancy estimation from carbon dioxide concentration. *Energy Build.* **2016**, *131*, 132–141. [CrossRef]
39. Zuraimi, M.S.; Pantazaras, A.; Chaturvedi, K.A.; Yang, J.J.; Tham, K.W.; Lee, S.E. Predicting occupancy counts using physical and statistical CO_2-based modeling methodologies. *Build. Environ.* **2017**, *123*, 517–528. [CrossRef]
40. Li, Z.; Dong, B. A new modeling approach for short-term prediction of occupancy in residential buildings. *Build. Environ.* **2017**, *121*, 277–290. [CrossRef]
41. Amayri, M.; Arora, A.; Ploix, S.; Bandhyopadyay, S.; Ngo, Q.D.; Badarla, V.R. Estimating occupancy in heterogeneous sensor environment. *Energy Build.* **2016**, *129*, 46–58. [CrossRef]
42. Ryu, S.H.; Moon, H.J. Development of an occupancy prediction model using indoor environmental data based on machine learning techniques. *Build. Environ.* **2016**, *107*, 1–9. [CrossRef]
43. Masood, M.K.; Soh, Y.C.; Jiang, C. Occupancy estimation from environmental parameters using wrapper and hybrid feature selection. *Appl. Soft Comput.* **2017**, *60*, 482–494. [CrossRef]
44. Han, J.; Pei, J.; Kamber, M. *Data Mining: Concepts and Techniques*; Elsevier: Amsterdam, The Netherlands, 2011; pp. 336–341.
45. Therneau, T.; Atkinson, B.; Ripley, B. CRAN-Package Rpart. Available online: https://cran.r-project.org/web/packages/rpart/rpart.pdf (accessed on 8 October 2017).
46. Therneau, T.M.; Atkinson, E.J. An Introduction to Recursive Partitioning Using the RPART Routines. Available online: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf (accessed on 8 October 2017).
47. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F.; Chang, C.; Lin, C. CRAN-Package e1071. Available online: https://cran.r-project.org/web/packages/e1071/e1071.pdf (accessed on 8 October 2017).
48. Karatzoglou, A.; Meyer, D.; Hornik, K. Support Vector Machines. Available online: https://www.jstatsoft.org/article/view/v015i09/v15i09.pdf (accessed on 21 October 2017).
49. Fritsch, S.; Guenther, F.; Suling, M.; Mueller, S.M. Training of Neural Networks. Available online: https://cran.r-project.org/web/packages/neuralnet/index.html (accessed on 25 Sepember 2017).
50. Günther, F.; Fritsch, S. Neuralnet: Training of Neural Networks. Available online: https://journal.r-project.org/archive/2010/RJ-2010-006/RJ-2010-006.pdf (accessed on 8 October 2017).
51. Goyal, S.; Ingley, H.A.; Barooah, P. Occupancy-based zone-climate control for energy-efficient buildings: Complexity vs. performance. *Appl. Energy* **2013**, *106*, 209–221. [CrossRef]
52. Zhang, J.; Lutes, R.G.; Liu, G.; Brambley, M.R. Energy Savings for Occupancy Based Control (OBC) of Variable Air-Volume (VAV) Systems. 2013. Available online: https://www.pnnl.gov/main/publications/external/technical_reports/PNNL-22072.pdf (accessed on 14 January 2019).
53. DOE-2.1A Reference Manual. VIII. Weather Data. DOE. 1980. Available online: http://doe2.com/download/DOE-21E/DOE-2ReferenceManualVersion2.1A.pdf (accessed on 29 October 2017).
54. U.S. Department of Energy Commercial Reference Building Models of the National Building Stock. Available online: https://www.nrel.gov/docs/fy11osti/46861.pdf (accessed on 28 October 2017).