Energy and Buildings 112 (2016) 28–39

Contents lists available at ScienceDirect

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models

Luis M. Candanedo*, Véronique Feldheim

University of Mons, Belgium

ARTICLE INFO

Article history:
Received 16 July 2015
Received in revised form
16 November 2015
Accepted 28 November 2015
Available online 2 December 2015

ABSTRACT

The accuracy of the prediction of occupancy in an office room using data from light, temperature, humidity and CO_2 sensors has been evaluated with different statistical classification models using the open source program R. Three data sets were used in this work, one for training, and two for testing the models considering the office door opened and closed during occupancy. Typically the best accuracies (ranging from 95% to 99%) are obtained from training Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART) and Random Forest (RF) models. The results show that a proper selection of features together with an appropriate classification model can have an important impact on the accuracy prediction. Information from the time stamp has been included in the models, and usually it increases the accuracy of the detection. Interestingly, using only one predictor (temperature) the LDA model was able to estimate the occupancy with accuracies of 85% and 83% in the two testing sets.

1. Introduction

The accurate determination of occupancy detection in buildings has been recently estimated to save energy in the order of 30 to 42% [1–3]. Experimental measurements reported that energy savings was 37% in [4] and between 29% and 80% [5] when occupancy data was used as an input for HVAC control algorithms. Nowadays, with the affordability of sensors increasing and becoming more ubiquitous, together with affordable computing power for automation systems it makes the determination of occupancy a very promising approach to lower energy consumption by appropriate control of HVAC and lighting systems in buildings. Other applications for occupancy detection include security and determination of building occupant behaviors. A system that could accurately detect the presence of the occupants without using a camera is very interesting due to privacy concerns.

This research has used data recorded from light, temperature, humidity and CO_2 sensors as a means to detect occupancy and a digital camera to establish ground occupancy for supervised classification model training. Combinations of these sensors can already be found in many buildings. The trained and tested models in this work are Random Forest (RF), Gradient Boosting Machines (GBM), Linear Discriminant Analysis (LDA) and Classification and Regression Trees (CART). These statistical classification models are already implemented in the open source program R. To the extent of our knowledge, the use and performance of RF, GBM and LDA models have not been reported before in the scientific literature for the occupancy detection task.

This work builds upon previous research that pointed out that better occupancy detection could be achieved by using monitoring equipment with higher resolution and accuracy. It also deals with model and feature selection by testing different feature combinations that have not been reported before. Also, for the first time, the measurements time stamp has been considered and included in the classification models.

The present work mostly deals with real time occupancy detection rather than with the development task of a new statistical learning algorithm. Because of that, this section focuses mostly on literature related to occupancy detection with sensors data.

1.1. Occupancy modeling previous work

A stochastic occupancy model of occupancy profiles based on a survey was developed by [6]. The model is useful for inputs to domestic energy models that require occupancy data in a time series and identifies differences between weekend and weekdays.

A model of occupancy was built from data of digital video cameras, passive infrared detection and CO_2 sensors [7]. The model used

* Correspondence to: Thermal Engineering and Combustion Laboratory, Rue de l' Epargne 56, 7000 Mons, Belgium. Tel.: +32 0 65 37 4471; fax: +32 0 65 37 4400.
E-mail address: Luismiguel.candanedoibarra@umons.ac.be (L.M. Candanedo).

http://dx.doi.org/10.1016/j.enbuild.2015.11.071
0378-7788/

Bayesian statistics to account for the role of previous information. The researchers reported that the model reduced the average error from 70% to 11%.

Two models for predicting occupancy were presented by [8]. The first model used multivariate Gaussian distribution to the sensed data (from a digital camera) and the second model is an Agent Based Model (ABM) that can be used for simulating mobility patterns.

The occupancy of one room was estimated with an agent-based model by [9]. The study examined the effect of adding noise to the data in order to account for people that occupy the building for short periods. This research was further advanced by proposing a graphical model of occupancy evolution in multi-zone buildings [10].

The number of occupants in a room was modeled using data for CO_2, temperature and ventilation to build a dynamic model [11]. The developed model had the returned occupancy level 88% of the time and was better than predictions obtained with Support Vector Machine and Neural network estimators. The raw data from the experiment was made available online.

Occupancy models were developed from data of a wireless sensor network that monitored occupancy [2]. The equipment employed cameras to count the number of occupants. The models were used in EnergyPlus to estimate the energy consumption. The authors calculated that it is possible to have an annual energy saving of 42% while maintaining adequate comfort standards.

1.2. Real time occupancy detection

A study of building occupancy detection using sensor belief networks was presented by [12]. For the study 3 passive infrared (PIR) sensors were used, together with a sensor that detects when a telephone handset was "off the hook". The truth occupancy was found using a video camera and tagged by human observers. The data was used to calibrate the belief network.

Prediction of user behavior when using a pattern discovery model was presented by [3]. The models were used in EnergyPlus and the estimated energy savings of 30% were reported. Among the data collected were noise levels, illumination, motion, CO_2 temperature and relative humidity. The models were based on unsupervised approach to avoid training or modeling of the environment at hand. In another paper [13] reported that an accuracy of 80% was obtained in the detection of the number of occupants using hidden Markov models. The authors also used support vector machines (SVM) and Neural Networks (NN). In [14] different data features were studied, and ordered using the concept of relative information gain (RIG). The correlation between the number of occupants was ranked as 77.65% for humidity, 73.42% for acoustics, 67.14% for CO_2 and 37.39% for temperature.

A presence sensor platform using a wireless system was developed and implemented by [15]. The system was based on PIR sensor, and a reed sensor to detect when a door is opened.

A model for real time estimation of building occupancy was presented by [16]. The model used data from a video camera and passive-infrared sensors. The occupancy detection was improved by using the extended Kalman filter, which combined the sensor data with the people movement.

An average of 73% accuracy on the occupancy number detection using Hidden Markov Models during testing periods was reported by [17].

Data was collected from a wireless sensor network for building occupancy by [1] and it was estimated that it is possible to achieve 42% annual savings with occupancy detection. The occupancy was established using seven cameras that recovered images at 1.5 fps (frames per second).

Occupancy detection from sensor measurements was presented by [18]. The research used the Latent Dirichlet Allocation unsupervised probabilistic model to analyze the occupancy logged data in a three floor building.

Real time occupancy detection using decision trees was presented by [19]. The reported accuracy was 97.9% when using data from a passive infrared motion sensor. The study included data from light, sound, CO_2, power use and motion sensors. The authors pointed out that a decrease in the accuracy (classification results) was found when adding many sensor readings and the reason is suspected to be overfitting.

An electronic device that used Radio-frequency Identification (RFID) tags and motion sensors were used to detect and predict occupancy in houses [20]. The researchers found that the system saved gas and reduced MissTime, defined as the time that the house was occupied but not warm.

A vision-based system for occupancy detection and activity analysis was presented in [21]. It used a camera and automatic image analysis to find the humans in the field of vision. The system was able to count the number of occupants in the images. The system was reported to have 97% accuracy in detection.

RFID tags were used in [22] to measure and monitor occupancy. The field test found an average zone detection accuracy of 88% for stationary occupants and 62% for mobile occupants.

A model that used a combination of temperature, humidity, CO_2, light, sound and motion was used to detect the number of occupants [23]. The model was based on a radial basis function neural network and implemented in Matlab. The reported accuracy was 87.62% for self-estimation and 64.83% when the trained model was applied to another room. A stated limitation of the study was the fact that the sensors used in the study were not calibrated.

Sensor data from CO_2, sound, relative humidity, air temperature, computer temperature and motion was used to train a neural network model [24,25]. The reported accuracy ranged from 75 to 84.5% with the model.

In [26] a wireless occupancy detection system it was presented that each node used a thermal camera and a PIR sensor to estimate occupancy. The article presents three models for the regression model: *k*-nearest neighbor, linear regression and artificial neural networks.

The electricity consumption from digital meters in 5 houses was used to train classification models to establish building occupancy [27]. Typically the accuracy was above 80% for the occupancy determination. It was found that the probability of using more power when the house is occupied is higher. However, it was found that the probability of having lower consumption even when the house is occupied is significant since the occupants might not be using any electrical devices.

Occupancy accuracy detection was reported in single occupancy rooms ranging from 92.2% to 98.2% according to the algorithm used for detection [28]. The authors reported that CO_2, door status and light variables have important contributions to the modeling results.

A review of occupancy detection systems and the evaluation of experimental results from chair sensors was presented by [29]. In this research, cushions having micro switches wired to a wireless transmitter were added to chairs of a conference room. The ground truth occupancy and the cushion sensors recorded were compared for 8 h and the error was 0% for the determination of the number of occupants.

A summary of the reported accuracies, the classification models and the sensors or parameters is presented in Table 1.

After reviewing the literature, a few key points are identified:

- The accuracy of the prediction was not reported in all the surveyed literature, making it difficult to compare performance among different approaches.

Table 1
Models, parameters and reported accuracies for occupancy detection.

Source	Classification models employed	Sensors/parameters	Accuracy for occupancy
[13]	Hidden markov models, Neural networks, Support Vector Machines (SVM)	CO_2 inside roomCO_2 outside room	NA
[18]	Latent dirichlet allocation	PIR	NA
[19]	Decision Trees (DT)	CO_2, computer current, light, PIR, sound	Ranging from 81% to 98.441% (only PIR)Only light: 81.01%Only sound: 90.78%Only CO_2: 94.68%
[23]	Radial basis function neural network	Lighting, sound, Reed sensor, CO_2, temperature, RH, PIR	*Note*: Accuracy for number of occupants63.23–66.43%
[24]	Artificial Neural Networks (MATLAB and WEKA[30])	CO_2, sound, relative humidity, air temperature, computer temperature, PIR	*Note*: Accuracy for number of occupants70.4–72.37%
[25]	Artificial Neural Networks (WEKA)	Temperature, humidity, light, Volatile Organic Compounds (VOCs), CO_2	*Note*: Accuracy for number of occupants67–69%
[26]	*K*-nearest neighbors, Linear regression, and artificial neural networks	PIR, Thermal array sensor	NA
[27]	Support Vector machine (SVM), K-nearest neighbor (KNN), Thresholding	Electric power consumption (W)	59–90%
[28]	Support Vector machine (SVM), *k*-nearest neighbor (KNN), Artificial Neural Network (ANN), naïve Bayesian (NB), tree augmented naïve Bayes network (TAN), decision tree (DT). Used WEKA.	CO_2, Reed sensor (for door), relative humidity, temperature, light, sound, PIR	88.9–98.2%For DT algorithms in two rooms:CO_2: 66.36–89.86%Light: 58.88–69.52%T: 55.26–65.32%CO_2 and T: 69.15–89.12%

- The best reported accuracy (100%) on the number of occupants employed individual chair sensors that transmitted the data with radios.
- Poor sensor calibration and low resolution seem to lower the accuracy prediction.
- Previous research has shown that using too many features may cause a decrease in the accuracy.
- The data sets from previous research are not readily available.
- Many interesting features combinations have not been evaluated, or reported in the literature.
- The performance of the other classification models has not been evaluated (LDAs, RFs, and GBM models). These models are used in the field of statistical learning when the answer is a qualitative variable [31], for example, during the diagnosis of a disease, out of 3 possible ones, digit image recognition, online fraudulent transactions detection among others.
- The date/time variable has not been employed for the classification.

The present work evaluates the performance of widely used classification models that are already available in the R programming language. Sensors with high resolution are employed. Models considering the possible shared feature combinations reported in the literature in Table 1 are compared here. The accuracy of the predictions will be evaluated and the datasets will be provided in order to allow for researchers to compare their implementations.

2. Data collection and setup

An office room with approximate dimensions of $5.85\text{m} \times 3.50\text{m} \times 3.53\text{m}$ $(W \times D \times H)$ was monitored for the following variables: temperature, humidity, light and CO_2 levels. A microcontroller was employed to acquire the data. A ZigBee radio was connected to it and was used to transmit the information to a recording station.

A digital camera was used to determine if the room was occupied or not. The camera time stamped pictures every minute and these were studied manually to label the data. See Fig. 1 for photograph of the setup.

The time stamp of the data has been exploited in this work by extracting the number of seconds from midnight for each day (NSM). The date stamp is also exploited by classifying it either a weekend (0) or a weekday (1), and this variable will be referred to Week Status (WS).

Table 2 presents the main features of the monitoring sensors:

An additional feature/variable included in data model is the humidity ratio (W). The humidity ratio in kg_w/kg_{da} was calculated using the measured temperature and relative humidity. The saturation pressure over liquid water (p_{ws} in Pa) was calculated with [32]:

$$\ln(p_{ws}) = \frac{C_1}{T} + C_2 + C_3 T + C_4 T^2 + C_5 T^3 + C_6 \ln(T)$$

where:

C_1 = 5.8002206E + 03	C_3 = −4.86402396E-02	C_5 = −1.44592093E − 08
C_2 = 1.3914993E + 00	C_4 = 4.1764768E − 05	C_6 = 6.5459673E + 00

T is the absolute temperature, $K = {}^\circ\text{C} + 273.15$.

The relative humidity, φ is defined as,

$$\varphi = \left.\frac{p_w}{p_{ws}}\right|_{t,p}$$

The humidity ratio W, is calculated using the following equation:

$$W = 0.622 \frac{p_w}{p - p_w}$$

where, p is taken as the standard atmospheric pressure, 101.325 kPa.

2.1.1. Environmental conditions

The data was recorded during winter in Mons, Belgium during the month of February. The room was heated by hot water radiators

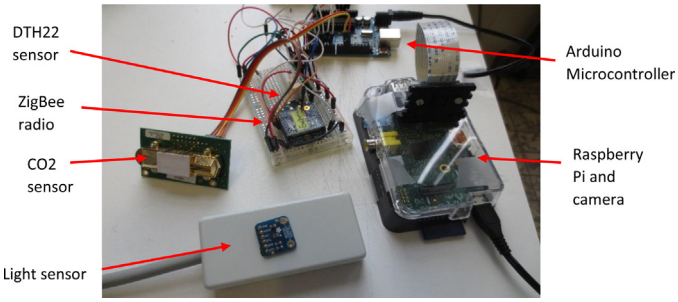

Fig. 1. Data acquisition setup showing the light, CO_2, DHT22 (temperature/humidity) sensors, a ZigBee radio and a microcontroller card and the digital camera controlled by a Raspberry Pi.

Table 2
Monitoring equipment.

Sensor	Parameter	Accuracy	Resolution	Range
DHT22	Temperature	±0.5 °C	0.1 °C	−40–80 °C
DHT22	Humidity	±3% RH	0.1%	0–100%RH
TSL2561	Light	NA	1 Lux	1–40,000
Telaire 6613	CO_2	400–1250 ± 30 ppm1250–2000 ± 5% of reading + 30 ppm	1 ppm	0–2000 ppm

that kept the room above 19 °C. In order to estimate the difference in occupancy detection accuracy given by the models, they are tested for data sets when the office door is open and closed. The readings were recorded at time intervals of 14 s or 3 to 4 times per minute, and then averaged for the corresponding minute.

The sensors were placed on a desk as shown in Fig. 2. The distance to the closest occupant was 1.1 m and to the second occupant about 2.9 m.

3. Recorded data and data sets description

Fig. 3 shows data between the morning of February 3, 2015 and the morning of February 4, 2015. The figure was created with the ggplot2 package [33]. As can be seen in the plot, just before 8:00 when the room is occupied by the first person, all the sensors show an increase in their measurements. The office lights are turned on. When the second occupant arrived just past 09:00 the slope of the humidity and CO_2 curves also increases. The Humidity and the CO_2 seems to have a morning maximum reading just after 11:00. Another interesting feature of the plot is that the CO_2 and humidity curves show a very similar pattern, especially for the humidity ratio and the CO_2. When the room is not occupied around 13:00 and 13:30, the light sensor registers a significant drop in the light measurement (lights off), also the CO_2, humidity and temperature sensors register a small drop in their readings. When the room is left vacant after 18:00 all the sensors show a downward trend, but the most dramatic drop is the light measurement. The next day around 7:30 in the morning when the room is occupied again by the first person, all the sensor readings start to increase noticeably. Again the slope for the humidity ratio and CO_2 increases when the second occupant arrives just past 9:00.

Fig. 4 gives a pairs plot showing the relationship for all the variables. The blue dots represent the occupied status, the green the non-occupied status. As can easily be seen in the temperature and humidity plot there is no clear separation boundary for the blue and green dots. The same thing happens for the combinations of temperature and CO_2, humidity and CO_2, temperature and humidity ratio, humidity and humidity ratio and CO_2 and humidity ratio. However, the pair combinations of temperature and light, humidity and light, light and CO_2 and light and humidity ratio, show a clear separation trend for the blue and green dots which indicates that these pair combinations are good candidates for training the classification models.

To have a better idea of the correlation between variables, the correlation matrix was calculated and is presented in Table 3 and its associated *p*-values (Table 4). As can be seen in the Table 4, the correlations are indeed significant for the pairs with very low associated *p*-values. In contrast, correlations are not significant for the pairs: NSM–Humidity and Week Status–NSM. Fig. 5 shows a visualization of the correlation matrix created with the Corrplot package.

3.1. Data sets

Three data sets have been used to train and test the classification models. They are summarized in Table 5. For all the data sets, the temperature, humidity, the derived humidity ratio, light, CO_2, occupancy status (0 for non-occupied, 1 for occupied) and time stamp are defined. The class distributions are also shown there.

3.2. Classification models performance

The statistical models CART, RF, GBM and LDA are used in this work. The CART models stratify the region where the predictions are done, the predictor space, into a number of simple regions [31]. Random Forests are models that make an effort to improve the accuracy of the prediction by creating many classification trees. When building the classification trees, a random sample of the predictors (2 or 3 for example) is selected and the best splits are used [31]. The prediction is based on the maximum number of votes to one classification status from each of the trees. During training, a set of observations not used to obtain the trees are used to estimate the error and are referred to as the out-of-bag (OOB) observations. The GBM models, also known as boosting, also try to improve the prediction for a decision tree by using information from previously generated trees [31]. LDA models use Baye's

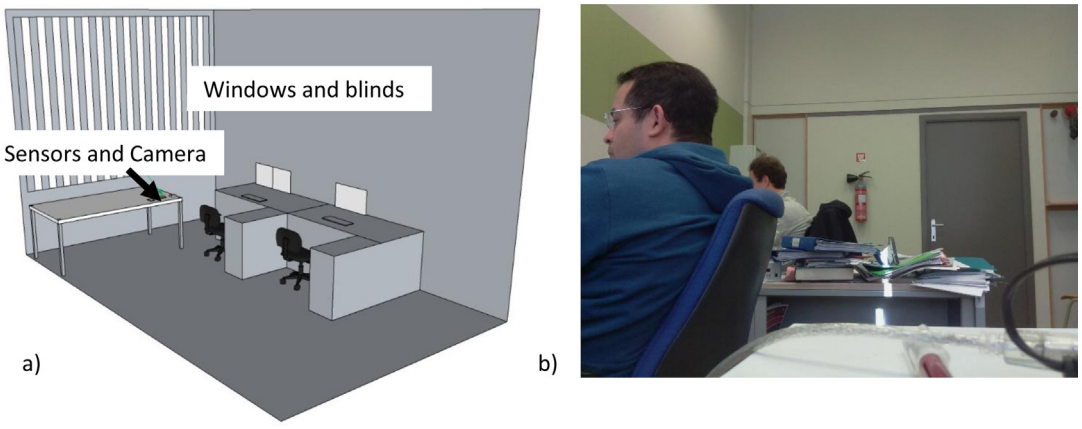

Fig. 2. (a) Room sketch showing the position of the sensors and the position of the occupants (b) Example of one of the pictures from the digital camera used to establish ground occupancy.

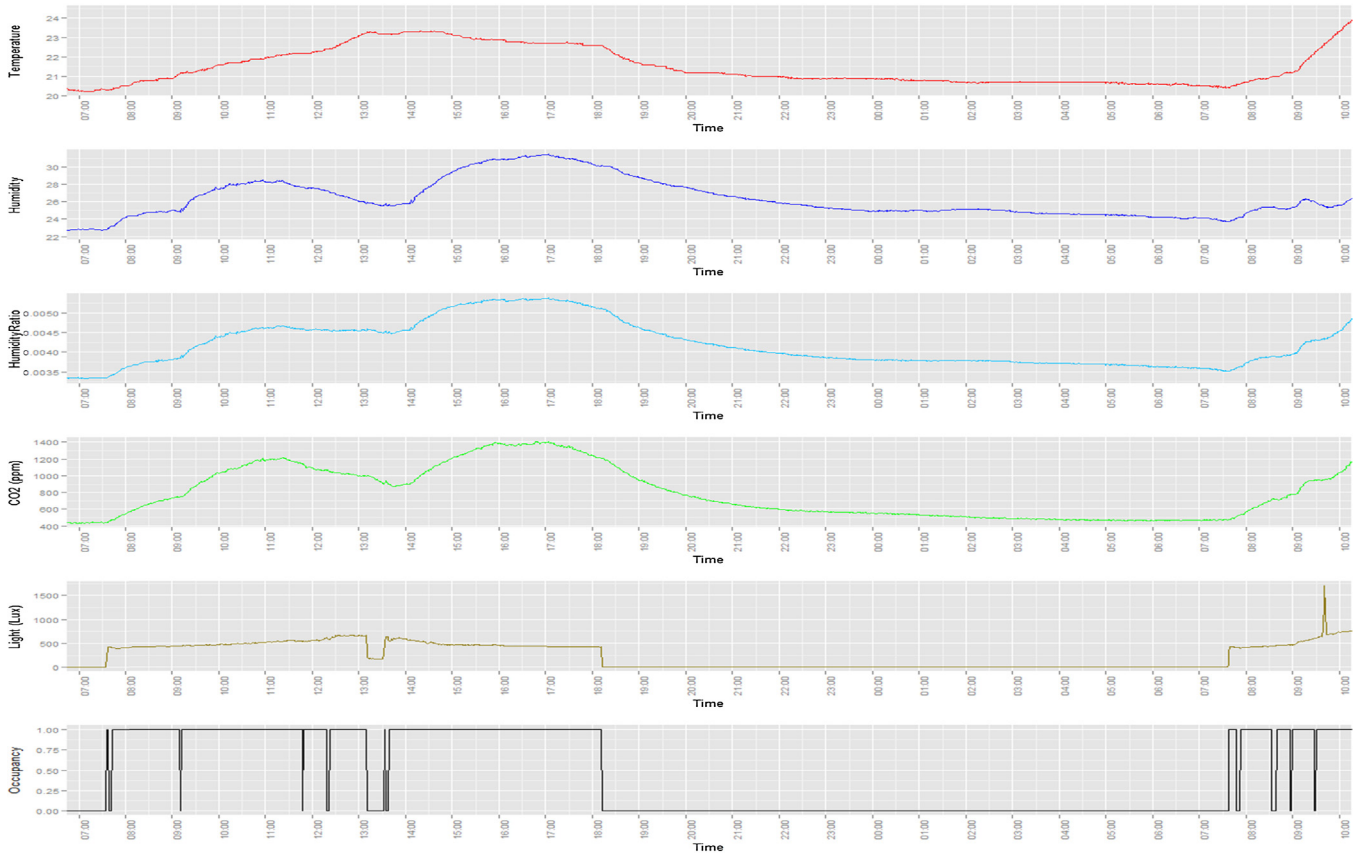

Fig. 3. Measurement profiles for 1 day (morning of February 3, 2015 to February 4, 2015).

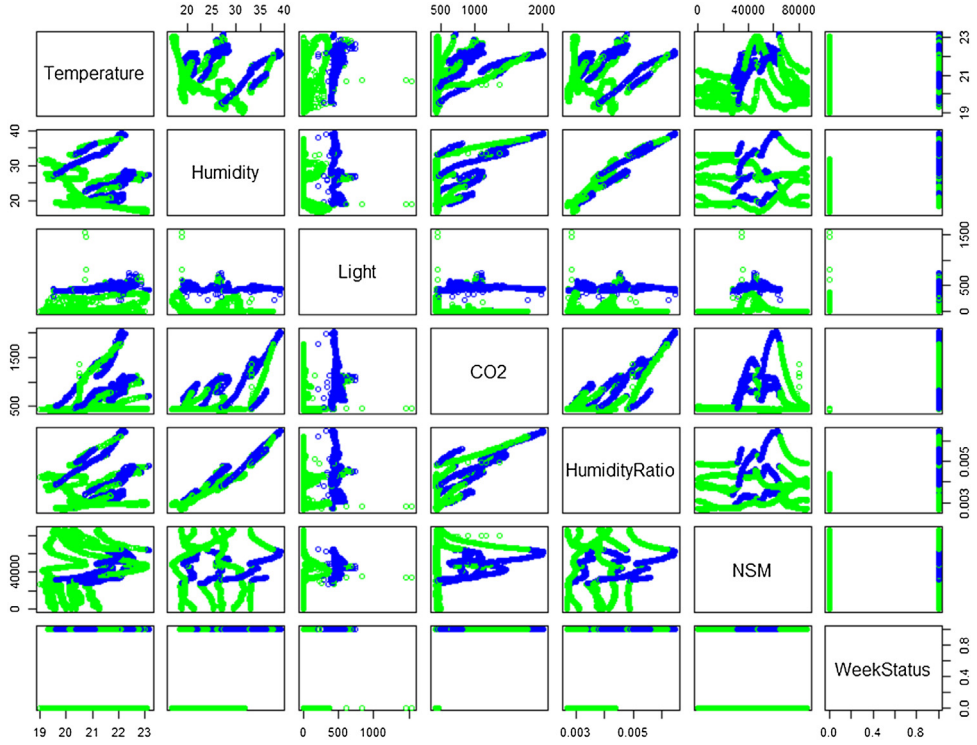

Fig. 4. Pairs plot. The blue data correspond to the occupied status and green to the not occupied status (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Table 3
Correlation matrix.

	Temp.	Humidity	Light	CO_2	H. Ratio	NSM	WS
Temp.	1.00	−0.14	0.65	0.56	0.15	0.26	0.42
Humidity	−0.14	1.00	0.04	0.44	0.96	0.02	0.11
Light	0.65	0.04	1.00	0.66	0.23	0.09	0.28
CO_2	0.56	0.44	0.66	1.00	0.63	0.21	0.39
H. Ratio	0.15	0.96	0.23	0.63	1.00	0.10	0.24
NSM	0.26	0.02	0.09	0.21	0.10	1.00	−0.01
Week status	0.42	0.11	0.28	0.39	0.24	−0.01	1.00

Table 4
Associated *p*-values of the correlations.

	Temp.	Humidity	Light	CO_2	H. Ratio	NSM	WS
Temp.		0e + 00	0e + 00	0e + 00	0e + 00	0e + 00	0e + 00
Humidity	0e + 00		6e − 04	0e + 00	0e + 00	1.25e − 1	0e + 00
Light	0e + 00	6e − 04		0e + 00	0e + 00	1.154e − 14	0e + 00
CO_2	0e + 00	0e + 00	0e + 00		0e + 00	0e + 00	0e + 00
H. Ratio	0e + 00	0e + 00	0e + 00	0e + 00		0e + 00	0e + 00
NSM	0e + 00	1.25e − 1	1.154e − 14	0e + 00	0e + 00		3.269e − 1
Week Status	0e + 00	0e + 00	0e + 00	0e + 00	0e + 00	3.269e − 1	

Table 5
Data set description.

Data set	Number of observations	Data Class Distribution (%)		Comment
		0 (non-occupied)	1 (occupied)	
Training	8143 of 7 variables	0.79	0.21	Measurements taken mostly with the door closed during occupied status
Testing 1	2665 of 7 variables	0.64	0.36	Measurements taken mostly with the door closed during occupied status
Testing 2	9752 of 7 variables	0.79	0.21	Measurements taken mostly with the door open during occupied status

theorem to estimate probabilities under the assumption that each of the variables/features follows a normal distribution. Then the model builds a classifier based on a linear combination of the features [31]. For more information about the models please refer to [31,34] and the [35] documentation.

During model development and testing, logistic regression models were tested. However, the algorithms did not converge. As reported by [31], when the classes are well-separated (see Fig. 4), the parameters estimated for the logistic regression models are unstable and the linear discriminant analysis model is better suited for this problem.

After the statistical models have been trained, they are evaluated against the training, testing 1 and testing 2 data sets using a confusion matrix to calculate the model accuracy. The training of the model was done with the Classification and REgression Training package (CARET) [35] that is available in R [36].

For RF models, the model training was done using 25 bootstrapped samples. Then the algorithm automatically selects the best number of splits according to the highest accuracy of the trained RFs. The number of trees for each RF model during training was set to 500, after it was verified that there was no significant decrease in the error prediction with more trees in the forest. For the GBM models the tuning parameters are the following: the maximum number of trees (which was set to 600, with 20 trees increments), the interaction depth (tested for 1, 2, 3 and 4), and the shrinkage (set to 0.1). The model also used 25 bootstrapped samples. The algorithm selects the best accuracy from the training combinations and selects the best combination of interaction depth and number of trees of the forest. The LDA models do not need any tuning parameter for model training. For the CART models, CARET automatically selects the best tuning parameter (complexity pruning) after using 25 bootstrapped samples.

The confusion matrix of observed and predicted values is presented in Table 6:

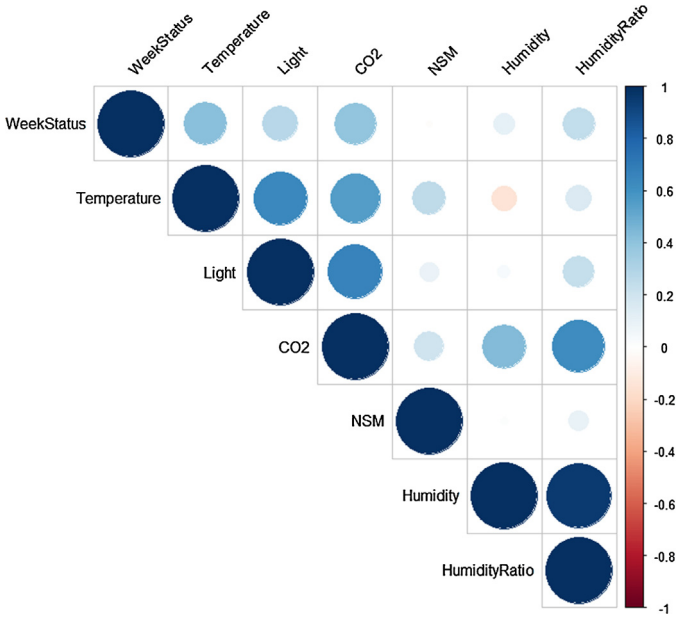

Fig. 5. Correlation Plot. The positive correlations are shown in blue and negative correlation in red. The sizes of the circles are proportional to the correlation values in Table 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Table 6
Confusion matrix.

			Reference Event 0 (Not Occupied)	Reference No Event 1 (Occupied)
Predicted	Event	0 (Not Occupied)	*A*	*B*
Predicted	No Event	1 (Occupied)	*C*	*D*

Table 7
Models performance.

Model	Parameters	Accuracy training (%)	Accuracy (%)	
			Testing 1	Testing 2
RF	*T*, *φ*, Light, CO_2, *W*, NS, WS	100.00	95.53	98.06
GBM	*T*, *φ*, Light, CO_2, *W*, NS, WS	100.00	95.76	96.10
CART	*T*, *φ*, Light, CO_2, *W*, NS, WS	99.39	94.52	96.52
LDA	*T*, *φ*, Light, CO_2, *W*, NS, WS	98.85	97.90	99.33
RF	*T*, *φ*, Light, CO_2, *W*	100.00	95.05	97.16
GBM	*T*, *φ*, Light, CO_2, *W*	99.98	93.06	95.14
CART	*T*, *φ*, Light, CO_2, *W*	99.30	95.57	96.47
LDA	*T*, *φ*, Light, CO_2, *W*	98.78	97.90	98.76
RF	*T*, φ, CO_2, W, NS, WS	100.00	94.63	64.86
GBM	*T*, *φ*, CO_2, *W*, NS, WS	100.00	91.86	51.14
CART	*T*, *φ*, CO_2, *W*, NS, WS	96.55	94.71	69.76
LDA	*T*, *φ*, CO_2, *W*, NS, WS	93.12	84.88	72.32
RF	*T*, *φ*, CO_2, *W*	99.98	68.63	32.68
GBM	*T*, *φ*, CO_2, *W*	99.56	69.53	38.81
CART	*T*, *φ*, CO_2, *W*	93.05	84.65	78.96
LDA	*T*, *φ*, CO_2, *W*	91.91	85.33	73.77
RF	*T*, *φ*, Light, *W*, NS, WS	100.00	95.65	97.21
GBM	*T*, *φ*, Light, *W*, NS, WS	100.00	95.98	97.51
CART	*T*, *φ*, Light, *W*, NS, WS	99.28	96.25	99.00
LDA	*T*, *φ*, Light, *W*, NS, WS	98.77	97.90	98.96
RF	*T*, *φ*, Light, *W*	99.95	94.00	96.73
GBM	*T*, *φ*, Light, *W*	99.96	92.42	95.27
CART	*T*, *φ*, Light, *W*	99.23	93.70	96.29
LDA	*T*, *φ*, Light, *W*	98.55	97.90	98.24
RF	*T*, *φ*, *W*, NS, WS	100.00	94.86	91.66
GBM	*T*, *φ*, *W*, NS, WS	99.98	95.50	92.49
CART	*T*, *φ*, *W*, NS, WS	88.97	90.81	89.36
LDA	*T*, *φ*, *W*, NS, WS	85.78	85.93	86.70
RF	*T*, *φ*, *W*	99.36	75.68	62.59
GBM	*T*, *φ*, *W*	97.85	76.10	63.75
CART	*T*, *φ*, *W*	88.28	84.02	86.30
LDA	*T*, *φ*, *W*	85.46	85.44	85.36
RF	*T*, Light, NS, WS	100.00	95.50	97.28
GBM	*T*, Light, NS, WS	99.88	96.36	98.81
CART	*T*, Light, NS, WS	99.28	96.25	99.00
LDA	*T*, Light, NS, WS	98.75	97.90	99.31
RF	*T*, Light	99.86	93.10	95.93
GBM	*T*, Light	99.45	95.50	98.33
CART	*T*, Light	99.08	85.10	95.43
LDA	*T*, Light	96.56	97.90	98.62
RF	*φ*, Light, NS, WS	100.00	96.92	98.31
GBM	*φ*, Light, NS, WS	99.99	96.59	98.63
CART	*φ*, Light, NS, WS	98.86	97.86	99.31
LDA	*φ*, Light, NS, WS	98.12	97.86	97.86
RF	*φ*, Light	99.96	92.35	94.36
GBM	*φ*, Light	99.75	94.52	98.71
CART	*φ*, Light	98.86	97.86	99.31
LDA	*φ*, Light	96.78	97.86	97.79
RF	Light, CO_2, NS, WS	100.00	96.06	97.42
GBM	Light, CO_2, NS, WS	100.00	96.06	98.57
CART	Light, CO_2, NS, WS	98.89	97.82	99.31
LDA	Light, CO_2, NS, WS	98.33	97.86	98.05
RF	Light, CO_2	99.95	92.61	97.41
GBM	Light, CO_2	99.26	93.70	98.34
CART	Light, CO_2	98.89	97.82	99.31
LDA	Light, CO_2	97.53	97.86	97.86
RF	CO_2, *T*, NS, WS	100.00	95.65	57.85
GBM	CO_2, *T*, NS, WS	99.98	96.14	77.71
CART	CO_2, *T*, NS, WS	96.55	94.71	69.76
LDA	CO_2, *T*, NS, WS	91.01	88.37	80.42
RF	CO_2, *T*	99.88	75.31	36.93
GBM	CO_2, *T*	97.76	72.98	46.46
CART	CO_2, *T*	93.05	84.65	78.96
LDA	CO_2, *T*	90.26	87.62	80.40
RF	Light, *W*, NS, WS	100.00	96.47	98.58
GBM	Light, *W*, NS, WS	99.99	96.74	99.03
CART	Light, *W*, NS, WS	99.37	96.89	98.89
LDA	Light, *W*, NS, WS	97.79	97.86	97.55
RF	Light, *W*	99.98	95.76	97.68
GBM	Light, *W*	99.52	96.10	98.98
CART	Light, *W*	98.94	96.66	98.90
LDA	Light, *W*	96.81	97.86	97.39
RF	*T*, *φ*, NS, WS	100.00	94.67	91.91
GBM	*T*, *φ*, NS, WS	99.95	95.61	92.85

Table 7 (*Continued*)

Model	Parameters	Accuracy training (%)	Accuracy (%)	
			Testing 1	Testing 2
CART	T, φ, NS, WS	88.97	90.81	89.36
LDA	T, φ, NS, WS	85.45	85.93	88.50
RF	T, φ	99.36	75.65	67.30
GBM	T, φ	97.85	77.49	65.34
CART	T, φ	88.24	84.02	86.30
LDA	T, φ	83.67	85.67	84.70
RF	T, NS, WS	99.88	95.08	92.52
GBM	T, NS, WS	99.75	95.91	92.57
CART	T, NS, WS	88.97	90.81	89.36
LDA	T, NS, WS	86.09	86.23	87.18
RF	T	87.87	70.73	85.33
GBM	T	87.44	64.69	85.63
CART	T	85.88	66.53	86.51
LDA	T	83.38	85.33	83.64
RF	Light, NS, WS	99.98	97.00	97.81
GBM	Light, NS, WS	99.74	95.95	98.47
CART	Light, NS, WS	98.93	97.22	99.26
LDA	Light, NS, WS	97.53	97.86	97.84
RF	Light	99.20	95.68	97.91
GBM	Light	98.77	97.86	99.32
CART	Light	98.78	97.86	99.31
LDA	Light	96.38	97.86	97.70
RF	CO_2, NS, WS	100.00	96.36	65.20
GBM	CO_2, NS, WS	99.99	96.21	74.51
CART	CO_2, NS, WS	96.55	94.71	69.76
LDA	CO_2, NS, WS	88.90	87.02	78.36
RF	CO_2	97.53	78.76	64.21
GBM	CO_2	92.15	87.13	74.62
CART	CO_2	92.18	87.13	74.62
LDA	CO_2	88.38	86.19	79.93

The accuracy of the model prediction is calculated as:

$$\text{Accuracy} = (A + D) / (A + B + C + D)$$

As can be seen in the equation, the accuracy is calculated by the sum of true positives (A) and true negatives (D), divided by the total number of predictions. Different models have been trained considering different predictor combinations and the results are summarized in Table 7.

3.3. Discussion

High accuracies for all the model predictions are typically found when all the measured parameters are taken into account for model training. Interestingly, the Random Forest model for T, φ, CO_2 and W shows very poor performance for the two test sets. In general all the Random Forest models with the different parameter combinations seem to have a much larger accuracy in the training set than in the test sets. Fig. 6 shows the out-of-bag and classification errors. As can be seen the error does not appear to decrease after 200 trees.

Fig. 7 shows the relative variable importance for two RF models. This relative variable importance is calculated as the total decrease in node impurities from splitting on the variable measured with the Gini index, averaged over all the trees [34,37]. The relative importance of the variables is obtained with a function call in R. It is interesting to see the marked decrease in accuracy in the tests sets for the RF model without light compared to the RF model with all the variables (see Table 7). It is also interesting to see that the accuracies in the testing sets of the RF model with light and CO_2 are comparable to the RF model with all the parameters. Interestingly, the RF model for light and W is slightly better than with light and CO_2. A probable reason for this is that the W and CO_2 variables

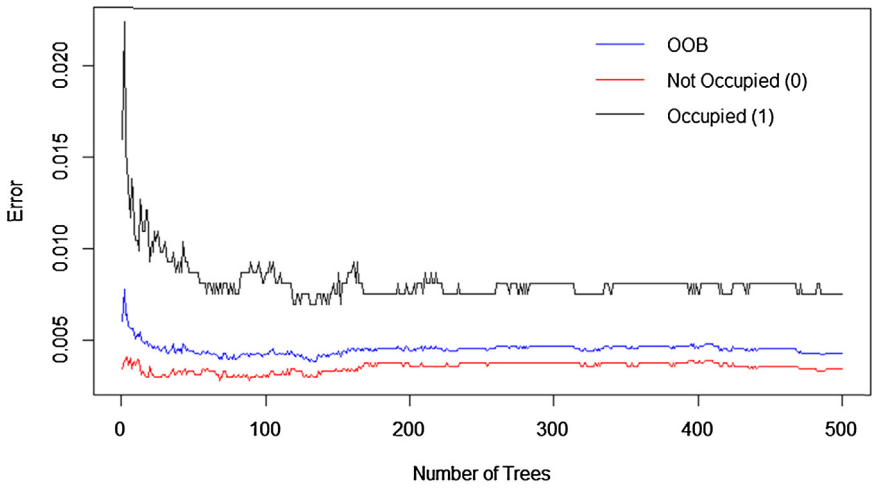

Fig. 6. Out of Bag and Classification error in the RF model with T, φ, Light, CO_2, W, NS, WS.

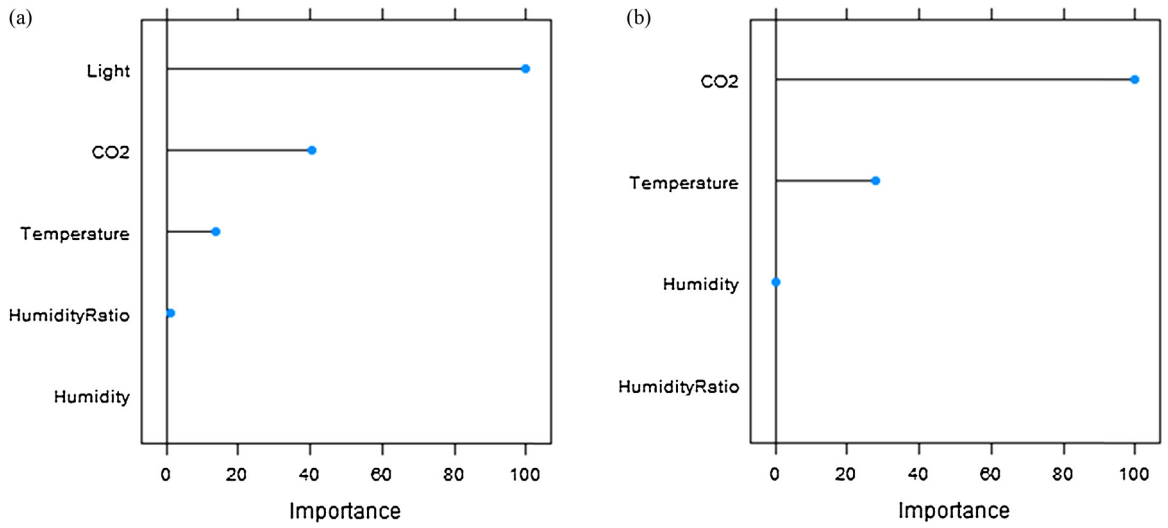

Fig. 7. a) Relative variable importance in the RF model with the parameters (T, φ, Light, CO_2, W); (b) relative variable importance in the RF model without light (T, φ, CO_2, W).

are highly correlated and therefore can act as surrogates of each other.

The maximum accuracies for the first test set were found with the LDA models and were all 97.9% when temperature and light are used in the data set for training. For the second test set the best model accuracy was 99.33% corresponding to the LDA model of T, φ, Light, CO_2, W, NS, WS.

The LDA models seem to provide a consistent accuracy prediction for all the cases in the training and test sets. This is especially true for the LDA models that use the pairs of humidity and light, light and CO_2, temperature and light, and light and humidity ratio. This was expected since it was seen in Fig. 4 that there is clear separation between the occupied and non-occupied points for those feature pairs.

When considering the effect of the NS and WS, we can see in Table 7 that in general including these two parameters increases the accuracy of the predictions regardless of the model employed (RF, GBM, CART and LDA). However, there are substantial increase in accuracies by including the NS and WS variables for the models that consider the following variables combinations (T, φ, Light, CO_2, W), (T, φ, W) and (T). For example, the accuracies for the RF model for T are between 70.73 and 85.33% and they improve to 95.08 and 92.52% by including the NS and WS parameters. By contrast, the most significant loss of model performance by including the NS and WS variables were found for the GBM, CART and LDA models for CO_2 in the testing set 2.

The GBM models have accuracies that are very close to the RF models. The GBM model for Humidity and Light shows a slight increase in accuracy for both testing sets when compared to the RF models (see Table 7). Fig. 8 shows the accuracy of the GBM model with all the parameters (T, φ, Light, CO_2, W, NS, WS) during training. As it can be seen, the best accuracies does not change significantly after about 400 trees for the max tree depths of 3 and 4.

The CART models typically are highly accurate in their prediction and their accuracies are usually very close to those of the LDA models which have given the best accuracies in the test sets. The

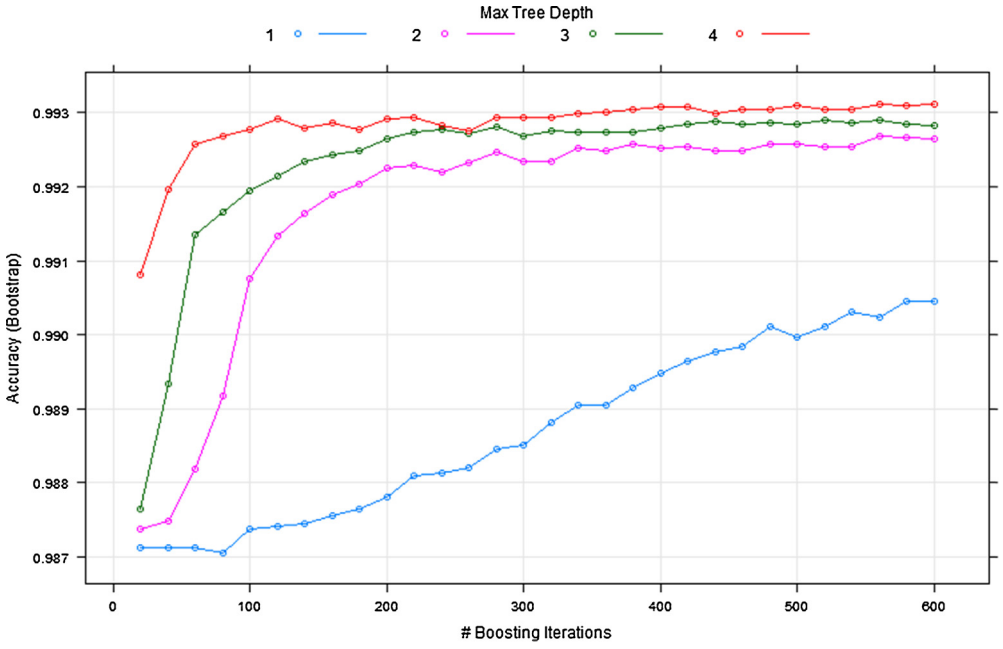

Fig. 8. Accuracy during model training for the GBM model with T, φ, Light, CO_2, W, NS, WS considering tree depth and number of trees (boosting iterations).

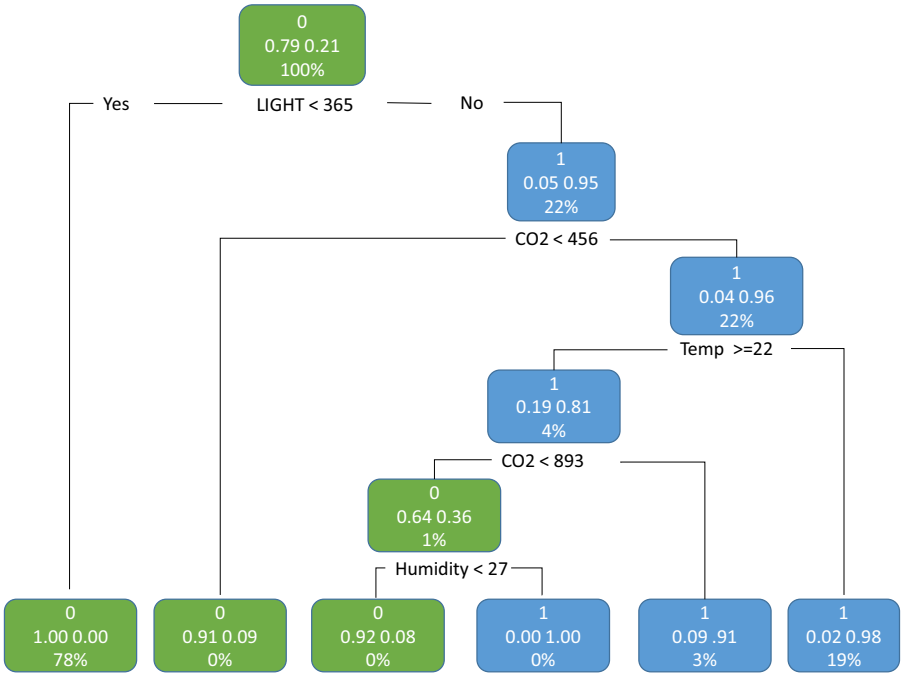

Fig. 9. Occupancy CART model for temperature, humidity, light, CO_2 and humidity ratio.

CART models are very interesting models because they are very easy to interpret. Figs. 9 and 10, show the CART models trained for temperature, humidity, light, CO_2 and humidity ratio and the CART model for light and temperature respectively. The plots were created with the rattle package [38]. The top node for both CART models is given by the light (see Fig. 11 for the detail) and it can be interpreted as the most important factor in the determination of the occupancy status. The color in the rectangles corresponds to the larger probability according to the classification status (green for 0, blue for 1) in the region to which it belongs. The percentage (rounded) of the data that fall in each node is presented.

The bottom rectangles (leaves) show the highest node purity (highest probabilities for each status). The classification task goes from top to bottom. The model predicts the status according to the test in each node that will guide the direction of the selected sub-branch of the tree, until arriving at the leaf node and assigns the corresponding prediction.

Fig. 11a shows a detail from Fig. 10: 0.79 is the probability of the evaluated data (100%) to be on the left (positive test) and there is a 0.21 probability corresponding to the right (negative test) with respect to the Boolean test done at the node (light measurement less than 365 lux). If we analyze Fig. 11b, we see that the

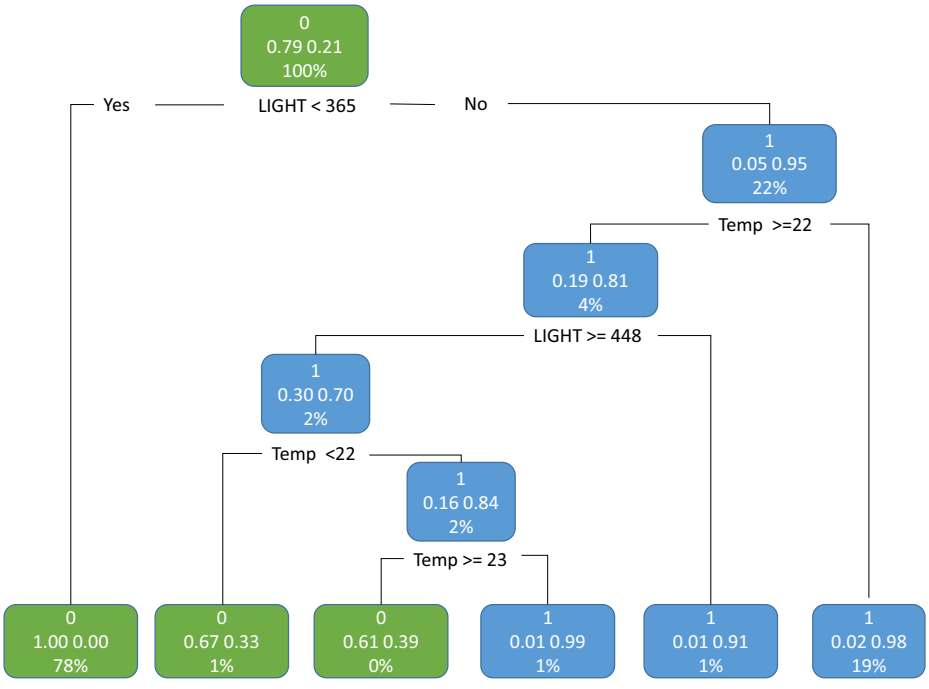

Fig. 10. Occupancy CART model for light and temperature.

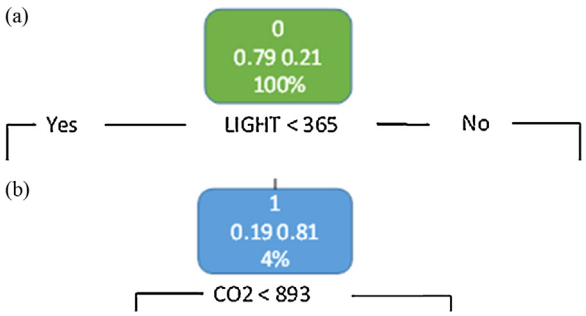

Fig. 11. (a) Detail of the Fig. 10. Fig. 11. (b): second detail of Fig. 10.

larger probability (0.81) corresponds to an occupied status (blue box—classification status of 1) and that there is a 0.19 probability of the concerned data (4% of the total amount) corresponding to CO_2 measurements inferior to 893 ppm. For this node, there is a 0.81 probability of the concerned data to have a CO_2 level higher than 893 ppm.

3.4. Comparison with results reported in the literature

Since the experimental data presented in Table 1 is not available, the testing conditions and metering equipment are not the same, nor the model implementations; it is not possible to fully compare the performance of the Decision trees models with the presented CART models. However, in this work the CART model for light have accuracies between 97 and 99%, which is above the reported value of 81% by [19] and substantially higher than the 59–69% reported in [28]. For CO_2, the CART model puts the accuracy between 75% and 87% compared to 95% in [19] and 66% and 90% in [28]. For only temperature, the CART model has accuracies of 67–87% which are higher than the 55–65% reported in [28]. Finally the CART model for CO_2 and temperature has accuracies of 79–85% which is in the range of 69–89% reported in [28].

3.5. Model implementation

CART models are very appealing since they could be easily implemented in a microcontroller. All the models can be incorporated easily in a microprocessor. Another analytic option is to remotely process the data and only transmit the control signal for the HVAC system. Regarding the sensor selection, the light sensor appears to be very important in the classification task, the one employed (TSL2651), currently can be acquired for $5.95 [39]. The CO_2 sensor can be very useful for demand control ventilation applications.

4. Conclusion

This work has shown that it is possible to obtain high accuracies in the determination of occupancy with RF, CART and LDA models. The trained Random Forest model for T, φ, CO_2 and W showed very poor performance in the testing sets (68.63 and 32.68% respectively). This is very likely due to the presence of highly correlated variables in the model (See Table 3 Correlation matrix). However, high accuracies (around 97%) were found when using only two predictors (temperature and light, light and CO_2 and light and humidity, light and humidity ratio) with the LDA models. In the initial exploratory data analysis section, it can be appreciated that these combinations show a good separation gap between the occupancy statuses.

This research has also shown that in general it is a good practice to include information related to the time of day and week status when building the classification models. The increase in accuracy was shown to be as high as 32%. Combining the time information (NS, WS) with temperature proved to yield accuracies between 86% and 96% from 65% and 86% without it.

Since the location of the sensors can affect the measurement readings, the models would have to be trained each time they are relocated.

This research has also found that using a light sensor can have higher accuracy (97–99%) for CART models compared to 81% reported in by [19] and substantially higher than the 59–69% reported in [28]. Also the CART model for temperature had accuracies of 67–87% compared to the 55–65% reported in [28]. The suspected reasons for these differences are higher sensor's accuracies, resolution and/or sensor location.

Although not treated in this work, the data suggests that for the feature pairs with good separation, the unsupervised classification model *k*-means, could identify the occupancy status effectively. Future work could also focus in the prediction of the number of occupants. Another option to improve the accuracy of the occupancy detection could be to use stochastic modeling [7].

To ensure reproducibility of the results by the research community and an eventual improved upon accuracy detection or model comparison, the data sets, together with the data processing scripts will be provided in the following repository:

⟨https://github.com/LuisM78/Occupancy-detection-data⟩

Acknowledgments

This work has received funding from the European Union's Seventh Program for research, technological development and demonstration under grant agreement no. 285173—NEED4B "New Energy Efficient Demonstration for Buildings".

References

[1] V.L. Erickson, M.Á. Carreira-Perpiñán, A.E. Cerpa, OBSERVE: Occupancy-based system for efficient reduction of HVAC energy, in: Proceedings of the 10th International Conference on, IEEE, Information Processing in Sensor Networks (IPSN), Chicago, IL, 2011, pp. 258–269.
[2] V.L. Erickson, M.Á. Carreira-Perpiñán, A.E. Cerpa, Occupancy modeling and prediction for building energy management, ACM Trans. Sensor Netw. (TOSN) 10 (3) (2014) 42.
[3] Dong B., Andrews B., (2009). Sensor-based occupancy behavioral pattern recognition for energy and comfort management in intelligent buildings. Proceedings of Building Simulation.
[4] J. Brooks, S. Goyal, R. Subramany, Y. Lin, T. Middelkoop, L. Arpan, L. Carloni, P. Barooah, An experimental investigation of occupancy-based energy-efficient control of commercial building indoor climate, in: Proceeding of the IEEE 53rd Annual Conference on, IEEE, Decision and Control (CDC), Los Angeles, CA, 2014, pp. 5680–5685.
[5] J. Brooks, S. Kumar, S. Goyal, R. Subramany, P. Barooah, Energy-efficient control of under-actuated HVAC zones in commercial buildings, Energy Build. 93 (2015) 160–168.
[6] I. Richardson, M. Thomson, D. Infield, A high-resolution domestic building occupancy model for energy demand simulations, Energy Build. 40 (8) (2008) 1560–1566.
[7] S. Meyn, A. Surana, Y. Lin, S.M. Oggianu, S. Narayanan, T.A. Frewen, A sensor-utility-network method for estimation of occupancy in buildings, in: Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on, IEEE, Shanghai, P.R. China, 2009, pp. 1494–1500.
[8] V.L. Erickson, Y. Lin, A. Kamthe, R. Brahme, A. Surana, A.E. Cerpa, M.D. Sohn, S. Narayanan, Energy efficient building environment control strategies using real-time occupancy measurements, in: Proceedings of the first ACM workshop on embedded sensing systems for energy-efficiency in buildings, ACM, Berkeley, California, 2009, pp. 19–24.
[9] C. Liao, P. Barooah, An integrated approach to occupancy modeling and estimation in commercial buildings, in: American Control Conference (ACC), IEEE, Baltimore, MD, 2010, pp. 3130–3135.
[10] C. Liao, Y. Lin, P. Barooah, Agent-based and graphical modelling of building occupancy, J. Build. Perform. Simulat. 5 (1) (2011) 5–25.
[11] A. Ebadat, G. Bottegal, D. Varagnolo, B. Wahlberg, K.H. Johansson, Estimation of building occupancy levels through environmental signals deconvolution, in: Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings, ACM, Rome, Italy, 2013, pp. 1–8.

[12] R.H. Dodier, G.P. Henze, D.K. Tiller, X. Guo, Building occupancy detection through sensor belief networks, Energy Build. 38 (9) (2006) 1033–1043.
[13] K.P. Lam, M. Höynck, B. Dong, B. Andrews, Y.-S. Chiou, R. Zhang, D. Benitez, J. Choi, Occupancy detection through an extensive environmental sensor network in an open-plan office building, IBPSA Build. Simulat. 145 (2009) 1452–1459.
[14] K.P. Lam, M. Höynck, R. Zhang, B. Andrews, Y.-S. Chiou, B. Dong, D. Benitez, Information-theoretic environmental features selection for occupancy detection in open offices, in: P.A. Strachan, N.J. Kelly, M. Kummert (Eds.), Proceedings of the Eleventh International IBPSA Conference, Citeseer, 2009, pp. 1460–1467.
[15] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, T. Weng, Occupancy-driven energy management for smart building automation, in: Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building, ACM, Zurich, Switzerland, 2010, pp. 1–6.
[16] R. Tomastik, S. Narayanan, A. Banaszuk, S. Meyn, Model-based Real-time Estimation of Building Occupancy During Emergency Egress Pedestrian and Evacuation Dynamics 2008, Springer, Berlin, Heidelberg, 2010, pp. 215–224.
[17] B. Dong, B. Andrews, K.P. Lam, M. Höynck, R. Zhang, Y.-S. Chiou, D. Benitez, An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network, Energy Build. 42 (7) (2010) 1038–1046.
[18] F. Castanedo, D. López-de-Ipina, H.K. Aghajan, R.P. Kleihorst, Building an occupancy model from sensor networks in office environments, ICDSC 3 (2011) 1–6.
[19] E. Hailemariam, R. Goldstein, R. Attar, A. Khan, Real-time occupancy detection using decision trees with multiple sensor types, in: Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design, Society for Computer Simulation International, San Diego, CA, 2011, pp. 141–148.
[20] J. Scott, A. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, N. Villar, PreHeat: controlling home heating using occupancy prediction, in: Proceedings of the 13th International Conference on Ubiquitous Computing, ACM, Beijing, China, 2011, pp. 281–290.
[21] Y. Benezeth, H. Laurent, B. Emile, C. Rosenberger, Towards a sensor for detecting human presence and characterizing activity, Energy Build. 43 (2) (2011) 305–314.
[22] N. Li, G. Calis, B. Becerik-Gerber, Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations, Automat. Construct. 24 (2012) 89–99.
[23] Z. Yang, N. Li, B. Becerik-Gerber, M. Orosz, 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, San Diego, CA, USA, 2012, pp. 49–56.
[24] Ekwevugbe T., Brown N., Pakka V. (2013). Real-time building occupancy sensing for supporting demand driven HVAC operations. 13th International Conference for Enhanced Building Operations, Montreal, Quebec.
[25] T. Ekwevugbe, N. Brown, V. Pakka, D. Fan, Real-time building occupancy sensing using neural-network based sensor network, in: 7th IEEE International Conference on IEEE, Digital Ecosystems and Technologies (DEST), Menlo Park, California, 2013, pp. 114–119.
[26] A. Beltran, V.L. Erickson, A.E. Cerpa, Thermosense: occupancy thermal based sensing for hvac control, in: Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings, ACM, Rome, Italy, 2013, pp. 11:11–11:18.
[27] W. Kleiminger, C. Beckel, T. Staake, S. Santini, Occupancy detection from electricity consumption data, in: Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings, ACM, Rome, Italy, 2013, pp. 1–8.
[28] Z. Yang, N. Li, B. Becerik-Gerber, M. Orosz, A systematic approach to occupancy modeling in ambient sensor-rich buildings, Simulation 90 (8) (2014) 960–977.
[29] T. Labeodan, W. Zeiler, G. Boxem, Y. Zhao, Occupancy measurement in commercial office buildings for demand-driven control applications—A survey and detection system evaluation, Energy Build. 93 (2015) 303–314.
[30] H. Mark, F. Eibe, H. Geoffrey, P. Bernhard, R. Peter, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. 11 (1) (2009).
[31] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, Springer, New York, 2013.
[32] ASHRAE, (2009). ASHRAE Handbook Fundamentals, Atlanta, GA.
[33] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer, New York, 2009.
[34] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Second Edition ed, Springer, New York, 2009.
[35] Kuhn M. (2015). Caret: Classification and Regression Training.
[36] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014.
[37] A. Liaw, M. Wiener, Classification and Regression by randomForest, R News 2 (3) (2002) 18–22.
[38] W. Graham, Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Springer, New York, 2011.
[39] Adafruit. (2015). ⟨http://www.adafruit.com/⟩.