Identifying The Presence Of People In A Room Based On Machine Learning Techniques Using Data Of Room Control Systems

Lucia Hanfstaengl, Michael Parzinger, Markus Wirnsberger, Uli Spindler, Ulrich Wellisch
Technical University of Applied Sciences Rosenheim
Presented at the 6th expert meeting of IEA EBC Annex 71 April 8-10, 2019, Bilbao, Spain.

Abstract

When monitoring the energy performance of buildings, it may be of interest to identify the occupation periods of people in the room due to their possible impact on the energy balance. In order to be able to carry out a comprehensive energy assessment of the system and building, it is necessary to be able to classify user influence during the evaluation. This thesis investigates how the presence of people in a room can be determined cost-effectively and with little additional effort. The aim is to determine which sensors of a room control system provide sufficiently reliable data. The presence of 1-2 persons was examined on a test facility of the Technical University Rosenheim. The air-, mean radiation- and surface-temperatures, the air humidity as well as the CO_2 and VOC concentrations were measured. For the analysis, a method of supervised machine and statistical learning, random forest, is used. The smallest model error detected in predicting the presence of 1 or 2 persons from CO_2 sensor data is 1.43%. The error rates are low for all tested models if time-dynamic effects are used as predictors and the data is processed in a so-called time period form. Additionally, the ways in which this data should ideally be made available for future measurements and processed to facilitate analysis with machine and statistical learning techniques have been investigated. A further goal is to apply the models developed on measurement series in laboratory environments to real rooms and to assess the transferability of these models.

1. Introduction

1.1 Background

Calculated energy demands are used to compare buildings with each other. Depending on user behavior and the quality of construction work, actual energy consumption may vary to a greater or lesser extent from the calculated energy demand. The true energy demand of a building can therefore only be determined by a measurement of the building itself. During monitoring, user behavior has a strong influence on the result of calculations, so it is therefore of interest to know about the user behavior in the buildings when making energetic evaluations.

Multiple methods for presence detection are in discussion or under investigation. These include: movement and presence detectors, user interfaces in the room (light switches, climate control, beamer control, displays, blinds, etc.), video surveillance, or even the use of near-field communication to detect the presence of smartphones [1]. In addition, sound pressure level measurements [6] or chair sensors that detect a person sitting [7] are examined. Not all methods are suitable for obtaining information about the number of people in a room. Some of the sensors may deliver very accurate results, however data and privacy protection may create problems with user acceptance. In terms of cost efficiency, it would be advantageous to know if and which sensors of a room control system provide reliable data on the presence of people, as these systems are generally already installed in many rooms.

International research projects such as the IEA Annex 71 [2] have set themselves the goal of improving the prediction, characterization and quality assurance of the actual energy performance of buildings. Depending on the type of building and the intensity of building automation, the user is one of the biggest influencing factors on the energy demand. Therefore, there is an increasing focus on how to measure user behavior in situ[1]. Some studies aim to improve building simulation models for predicting energy demand by using more realistic user profiles[3]. Other studies focus on predicting the presence of people to improve the control of heating, ventilation and lighting systems [3][5][6][7][8]. Further studies see potential to improve facility management, e.g. in room occupancy [3]. Estimation of the presence of people in "real time" should influence the control of the building technology in such a way that energy can be saved optimally[5].

The presence of people has been examined in offices [3], open-plan offices [3][4], student apartments [5] and theatres[6].

A frequently used measurement parameter to identify the number of people present is the CO_2 concentration [1][3][5][6], because it increases with the number of people. However, the increase is also influenced by the air change, air tightness, wind speed and room volume. Also, air humidity, volatile organic compounds (VOC) and PIR sensors have been investigated with respect to presence predictions and provide results with good model accuracy[5].

In the literature studied here, the presence of people is detected using various statistical analysis methods. Kim et al. [3] use three different machine learning methods, which are: classification and regression tree (CART),

artificial neural network (ANN) and support vector machine (SVM). Jiang et al. [4] work with the Scaled Extreme Learning Machine (FS-ELM) algorithm feature. In Pedersen et al. [5], they apply a series of decision rules (binary) to the curve of sensor data. Zuraimi et al. [6] chose three standard machine learning methods: 1) Artificial Neural Networks (ANN); 2) Prediction Error Minimization (PEM); and 3) Support Vector Machines (SVM), as these are widely used and easy to implement. Ryu et al. [8] apply a decision tree algorithm in the first step, and in the second step, a model for predicting presence is selected using a hidden markov model (HMM).

1.2 Contributions

This thesis considers the identification of the presence of people from the experimental design through to the statistical analysis, and accounts for the interaction between and the effectiveness of the methods employed. The aim is to continuously improve the test design for further investigations and to prepare it for application in the field.

Ways of using random forest to determine the presence of people and the number of persons present is being investigated specifically. The method of analysis used is a random forest method of supervised machine and statistical learning [9]. This is an ensemble method in which bootstrapping is used to create multiple, uncorrelated decision trees. This nonparametric and nonlinear method offers high flexibility, e.g. it does not require a distribution assumption or a detailed physical model. The method can deal with classification and regression tasks. In this thesis, the classification task of random forest was chosen. The method is very well suited for parallelizing calculation steps, which makes it possible to evaluate even large amounts of data quickly.

Furthermore, it is investigated which and how many sensors of a room control system are needed for the presence prediction. A range of different room sensors are included in the analysis, so that it is possible to investigate which parameters are decisive and whether the prediction accuracy increases with observations of different sensors. With the help of the so-called importance measures, the random forest offers the possibility of creating a ranking list of features (i.e. influencing variables based on the sensor values or measurement parameters) with regard to their prediction quality. In addition, on the basis of the importance measures, faulty sensors can be detected by the unimportance assigned to them.

The aim is to investigate how the data must be collected with a focus on optimal modelling (measurement interval, data storage in the database, data transfer to the analysis program).

2. Methods

2.1 Experimental design

The presence of people was estimated using a method of machine learning. The data required for the development of the models was generated in a series of experiments. The tests were carried out on a facility at the Technical University Applied Sciences Rosenheim. For more details on the test facility see Janssens[10]. The experimental facility consists of a container with three test boxes. The test boxes each have a 2/3 opaque external facade, which constitutes the only wall bordering the external air. The remaining walls border the service room of the container. The walls are thermally decoupled. Two of the test boxes (A & B) have the dimensions 2.9*2.9*2.9m³ (B*L*H). The third test box (C) is slightly bigger at 2.9*3.2*2.9m³ (B*L*H). Measurements were carried out in rooms A and C. Figure 1 shows the ground plan of the test bench.

Measuring sensors already installed in the facility were used in both test boxes. These were sensors for air temperature, radiation temperature, surface temperature, humidity, CO_2 concentration and VOC concentration. Air pressure and solar radiation were also measured. A total

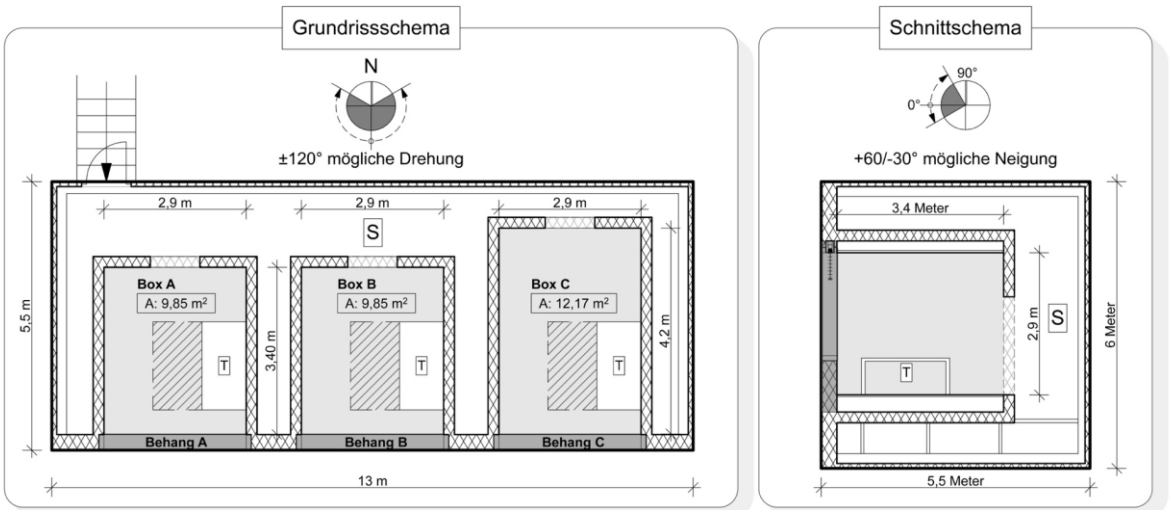

Figure 1 Layout of the test facility

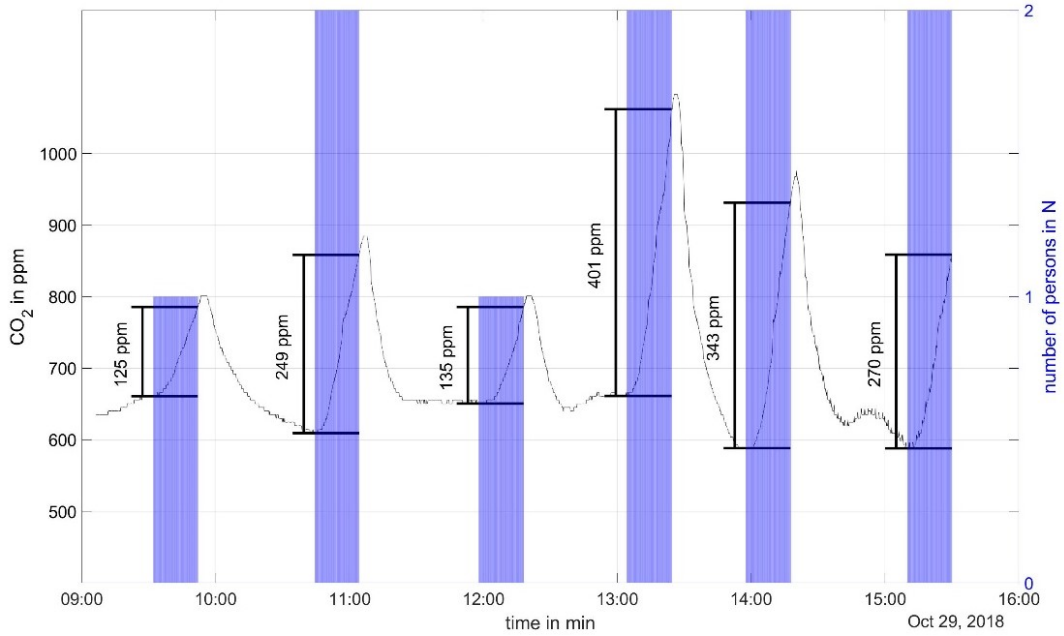

Figure 2 Exemplary CO2 concentration curve and number of persons in room C on a measurement day. Additionally, recorded is the increase of the CO2 concentration in the presence phases

of 34 sensors recorded data, most of which were air and surface temperatures.
The number of people present for the training set data was determined with a switch. In addition, a PIR sensor was installed in the entrance area of the test box. During the experiment, one or two people were in room A or C for a period of 20 minutes. After the persons had left the room, the room was ventilated by opening the test box door. Forty experiments were carried out in each room (A and C) with both one and two people present, so that 160 experiments were carried out in total.

2.2 Results from experimental study

Figure 2 shows an example of the profile of the CO_2 concentration during a measurement day. The height of the bars indicates the number of people present. The CO_2 concentration increases after entering the test box, and decreases after leaving the test box and opening the test box door. In addition, the change in the CO_2 concentration during the presence phases is displayed.

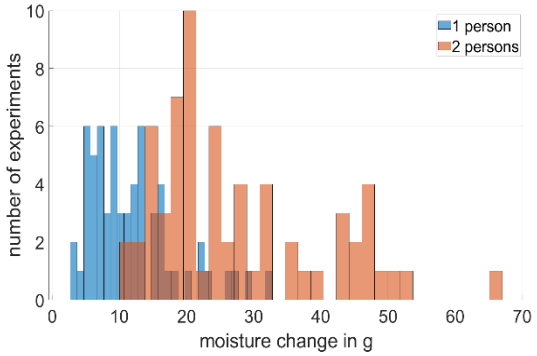

Figure 3 Humidity change in the room during the 20 minute occupancy period, from all experiments

Figure 3 shows the change in humidity due to the release of moisture from people during occupancy, of 20 minutes of all measurements. The values are subdivided into the humidity increase from one person and two persons. The increase in humidity for one person is in the range of 10g. For two people, the increase in humidity is usually greater than 10g and can reach over 50g. The two different numbers of people present form two distinct frequency distributions with little overlap.

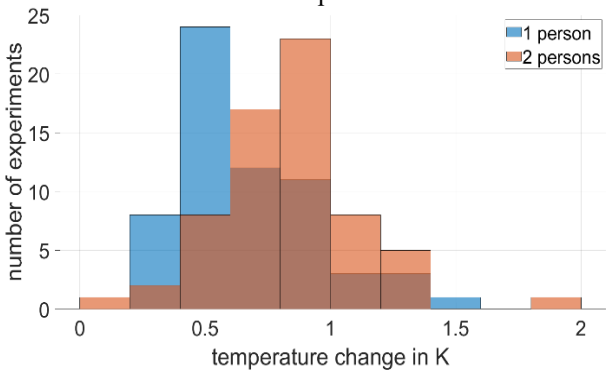

Figure 4 Temperature increase during the 20 minute occupancy period, from all experiments

Figure 4 shows the increase in temperature during occupancy, of 20 min, of all measurements. For one person, the maximum temperature difference of all tests is between 0.4 K and 0.6 K. For two people, the maximum is between 0.8 K and 1.0 K. There is a high overlap between 0.4 K and 1.0 K.

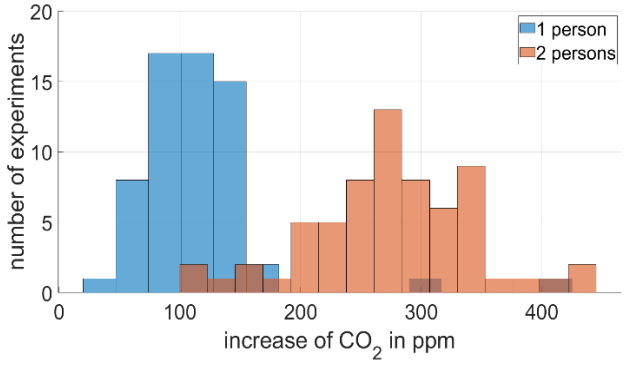

Figure 5 Increase of the CO2 concentration during the 20 minute occupancy period, from all experiments

The increase in CO_2 concentration during occupancy from all experiments is shown in Figure 5. The increase in CO_2 concentration for one person ranges from between 70ppm and 170ppm. For two people, the range extends from 100 ppm to 430ppm.

2.3 Data preprocessing

In general, monitoring-measurement data is often available in a multivariate time series structure that may not be synchronized. For the application of the random forest algorithm, a data structure centered on the event response variable must be created. With regard to realistic applications, only simultaneous or possible influencing variables (features) of the past should be assigned to the response.

Successful and efficient data analysis using methods of machine and statistical learning starts with the adequate recording of the measuring sensors.

For the analyses, a data structure is required in which the observations of all variables and the values of the potential influence variables (with as few missing values as possible) are assigned in an event-oriented manner to each point in time or time interval. A time-synchronized, multivariate time series structure of the sensor-measured values (with a low missing component) is suitable as a basis for this.

In the following, two different table structures are explained and evaluated with random forest. These are the table structures we call ‘time point form’ and ‘time period form’. Figure 6 displays those table forms. A data structure that contains one observation per point in time is the time point form, and is characterized by the fact that each row represents a specific point in time, and the columns, in addition to the response, are formed by the sensors used. In the case where there is one observation per time interval or period, a table is available in time period form. In addition to the response, there are then several sensor values at different times within the time period form for each time period. For example, if the time period of 5 minutes is specified and the sensors measure every minute, then the table in the time period form contains for each sensor the columns with the measured value after one, two, three, four and five minutes.

Time Point Form

Y	S
1	0,409
1	0,553
1	0,901
1	0,258
0	0,183
0	0,030
0	0,498
0	0,016
1	0,287
1	0,436
1	0,435
1	0,451
0	0,858
0	0,213
0	0,655
0	0,749

Time Period Form

Y	S_1	S_2	S_3	S_4
1	0,409	0,553	0,901	0,258
0	0,183	0,030	0,498	0,016
1	0,287	0,436	0,435	0,451
0	0,858	0,213	0,655	0,749

Figure 6 Example of a transformation from time point form to time period form.

2.4 Results from time form

After data preparation, the tables containing the sensors that measure every 10 seconds will be considered. This means that there are two tables in the same form with the same columns, one with the data from room A and the other with the data from room C. Models developed in room C are validated with the data from room A and vice versa. If this procedure is carried out, the following confusion matrices are obtained on the validation data, for the comparison of the estimated values with the actual occupancy values (observed values):

Time Point Form

Room A train – Room C test

		Predicted			
		0	1	2	Err
Observed	0	15655	414	1335	0.100
	1	3320	36	727	0.991
	2	3208	252	410	0.894
				Total error rate	36.25%

Time Point Form

Room C train – Room A test

		Predicted			
		0	1	2	Err
Observed	0	2911	175	10341	0.783
	1	469	45	2750	0.986
	2	572	45	2766	0.184
				Total error rate	71.50%

Figure 7 Illustration of two confusion matrices. The upper table shows the results of the model from the data of room A and the verification by the data of room C. In the lower table, the learning and validation samples are reversed.

These matrices represent contingency tables of actual presence values and occupancy values predicted by the random forest. In an initial modelling, only features belonging to the time of observation of the response were used. The results are shown in Figure 7. The table reads as follows: if 0 persons were present, this is recognized correctly 15655 times, one person is incorrectly estimated 414 times and two people are incorrectly estimated 1335 times. This results in an error rate of 0.1. The total error rate of the upper table at Figure 7 is 36.25% and of the

lower table 71.5%. Both cases show very high error rates and both models are classified as unusable. In order to improve the model quality for a second round of modelling, the differences between the sensor values at the observation time and at past measurement times are therefore included as possible features. By adding these features, which are formed over a time-lag with historical values, one progresses from a time-static model to a time-dynamic model. The difference between the time before one minute, before two minutes and every minute up to five minutes is used for each sensor. The resulting random forest models result in the following confusion matrices:

Time Point Form + differences

Room A train – Room C test

		Predicted			
		0	1	2	Err
Observed	0	16918	208	100	0.018
	1	820	2810	263	0.278
	2	237	745	2796	0.260
				Total error rate	9.53%
				Total error rate: Present yes/no	4.7%

Time Point Form + differences

Room C train – Room A test

		Predicted			
		0	1	2	Err
Observed	0	12446	600	322	0.069
	1	168	1677	1208	0.451
	2	85	85	3213	0.050
				Total error rate	12.46%
				Total error rate: Present yes/no	6.9%

Figure 8 Confusion matrices after adding temporal measurement differences. The upper table shows the results of the model from the data of room A and the verification by the data of room C. In the lower table, the learning and validation samples are swapped again.

Figure 8 shows that the overall error rate is significantly better than in the previous models without features based on temporal measurement differences, cf. Figure 7. The prediction accuracy of the model for the presence of one or two people for room A is rather low. The model of room C can hardly distinguish between one and two people in the presence of one person. However, these miscalculations can be explained by the fact that the presence of a person is checked with an accuracy of 10 seconds and the sensors require a certain reaction time until they can detect a change of a measuring parameter. Nevertheless, the overall error rates of the models are very low, especially since a lack of presence is largely detected. In addition to the previous total error rate, the table in figure 8 also contains the "Total error rate: Present yes/no". To calculate these values, a third model is formed with temporal measurement differences and binary responses, "present yes or no", and the total error rate is calculated. This changes the question of "How many people are present? " to "is there at least one person present?" (0=no, 1=yes). This procedure leads to a lower error rate, but also to a less detailed statement.

The random forest method offers, with the help of the importance function (see [12]), the possibility of displaying the features sorted by their importance regarding the prediction power. A distinction is made for categorical response between the mean decrease in accuracy and the mean decrease in gini. For an exact definition and calculation of these Importance measures, see L. Breiman[9] or G. James et al.[11], In the following we focus only on the mean decrease in accuracy. The most important features for the prediction of presence in random forest were the absolute humidity, the difference between the instantaneous CO_2 concentration and the concentration 3 minutes ago, the instantaneous CO_2 concentration and the air pressure.

This ranking list can be used, for example, if you only want to restrict yourself to the most important sensors for prediction in the random forest and if you want to carry out a variable selection for modelling. A reduction of the used feature number usually leads to model stabilization and can reduce a possible overfitting of the models. For future measurement arrangements one could then also limit oneself to the most important sensors. Another application of these models is the detection of faulty sensors by including in the modelling, in addition to the sensor values, a (random) feature that consists only of pseudo-random numbers. This random feature is then assigned an "importance" in the random forest algorithm and all sensors whose "importance" is not greater than the importance of the random feature can be removed from the model building. This can also be an indication of defective sensors.

2.5 Results from period form

In the time form, the presence is estimated at a certain point in time. In comparison, the time period form estimates whether one or two persons are present in a certain time window (here 20 min). The data is transformed as shown in Figure 6 and referred to as the time period form.

In the time-form, the second model with measurement differences has produced significantly better results, so measurement differences are also taken into account when modelling with the time-form. If one forms random forest with these time period tables, one obtains the following results with a first model, cf. Figure 9.

The overall error rates are below 5%, which is a very satisfactory result.

The application of the importance function shows that the changes in the CO2 value at different points in time are most important in modelling with the time period form. If only the features based on the CO2 sensor are used, total error rates of less than 5% result. The model with measurement differences in the period form, developed on the learning sample from room A, has a total error rate of

approx. 4.5% and the model from the learning sample from room C of 1.43%.

Time Period Form + differences					
Room A train – Room C test					
		Predicted			
		0	1	2	Err
Observed	0	111	1	0	0.009
	1	1	32	0	0.030
	2	0	4	28	0.125
				Total error rate	3.39%
				Total error rate: Present yes/no	≈3%

Time Period Form + differences					
Room C train – Room A test					
		Predicted			
		0	1	2	Err
Observed	0	85	0	0	0.000
	1	0	24	3	0.111
	2	0	0	28	0.000
				Total error rate	2.14%
				Total error rate: Present yes/no	≈0%

Figure 9 Confusion matrices in the period form with temporal measurement differences.

3. Summary

This thesis has studied the possibility of estimating the presence of people using the machine learning method random forest and data from indoor climate sensors. First, a test was carried out on a test bench and the data was then evaluated descriptively and graphically. Taken on its own, the graphic evaluation of the measurements shows tendencies of the values for the presence of one and two people, but no clear statements can be made. During the statistical evaluation, two different table structures were presented as data bases for the models. It turned out that the time point form is better suited for practical application. Only in the time point form can one infer a certain point in time after the analysis with random forest. This is important if you want to determine the exact time point of occupation.

Data processing for statistical modelling is very time-consuming. Different start times from three different data loggers and different measurement intervals from 10 seconds up to 5 minutes result in the data tables having to be converted, shifted and rows removed to enable evaluation. In part, so much information was lost that modelling was no longer possible with sufficient model quality. A successful and efficient data analysis with methods of machine and statistical learning starts with well formulated objectives with respect to recording the data.

The models are good in predicting whether a room is occupied or not, but the results of predicting the exact number of people are not yet satisfying. The smallest model error detected in predicting the presence of one or two people from the data of CO_2 sensors comes from the model with the time period form and measurement differences, and is 1.43%. In all models tested, the error rates are low if time-dynamic effects are used as predictors and the data is processed in a so-called time period form. With the time period form, one receives better results, however it is rather unsuitable in practice. In this experiment, there were only fixed periods of occupancy of 20 minutes, but in practice one stays in a room for different lengths of time. For this reason, the time point form should be used for future studies.

With the help of an importance measurement, the sensors can be evaluated with regard to their prediction weight. The importance measurement can be used to reduce measuring systems to important sensors and thus also detect faulty sensors.

4. Conclusion

For future measurement series and with regard to field measurements, a number of requirements for the test design can be clarified by the present work.

The test facility should be upgraded in such a way that more general conditions can be set. This is for example a constant air exchange rate in the test rooms. With this measure the point in time when people leave a room could be better predicted.

Another important precondition is, that the measuring system should be exchanged and be set up in a way that all data have a uniform time stamp. This should also be observed for field measurements.

In future experiments, more information about the people that enter the room could be obtain. Among these could be documented: gender, their place of stay in the room, activity level, physical condition and state of mind, in order to integrate these effects into the statistical models.

The influence of spatial geometry on the statistical model could be investigated by measuring in all three test rooms, so that test room A and B can be compared with room C.

In addition, the number of people should be varied more strongly, and larger numbers of people and different time intervals for their occupation periods should be included in the tests.

The importance measure should be used in the future for feature selection and model optimation to reduce overfitting effects.

5. Acknowledgements

The study was carried out as part of a research project supported by the Federal Ministry for Economic Affairs and Energy.

Supported by:

Federal Ministry for Economic Affairs and Energy

on the basis of a decision by the German Bundestag

6. References

[1] Mahdavi, A.: Occupant behavior modeling for building performance simulation: Current state and future challenges. Energy and Buildings 107 (2015), p.264–278

[2] IEA EBC Annex 71 Building Energy Performance Assessment Based on In-situ Measurements. Available online: http://www.iea-ebc.org/projects/project?AnnexID=71

[3] Seokho Kim, Yujin Song, Yoondong Sung and Donghyun Seo: Development of a Consecutive Occupancy Estimation Framework for Improving the Energy Demand Development of a Consecutive Occupancy Estimation Framework for Improving the Energy Demand Prediction Performance of Building Energy Modeling Tools. Energies 2019, 12, 433

[4] Jiang, C., Masood, M. K., Soh, Y. C. u. Li, H.: Indoor occupancy estimation from carbon dioxide concentration. Energy and Buildings 131 (2016), p.132–141

[5] Pedersen, T. H., Nielsen, K. U. u. Petersen, S.: Method for room occupancy detection based on trajectory of indoor climate sensor data. Building and Environment 115 (2017), p.147–156

[6] Zuraimi, M. S., Pantazaras, A., Chaturvedi, K. A., Yang, J. J., Tham, K. W. u. Lee, S. E.: Predicting occupancy counts using physical and statistical Co2-based modeling methodologies. Building and Environment 123 (2017), p.517–528

[7] Labeodan, T., Zeiler, W., Boxem, G. u. Zhao, Y.: Occupancy measurement in commercial office buildings for demand-driven control applications—A survey and detection system evaluation. Energy and Buildings 93 (2015), p.303–314

[8] Ryu, S. H. u. Moon, H. J.: Development of an occupancy prediction model using indoor environmental data based on machine learning techniques. Building and Environment 107 (2016), p.1–9

[9] Breiman, L.: Machine Learning 45 (2001) 1, p. 5–32

[10] Janssens A.:Inventory of full scale test facilities for evaluation of building energy performances. International Energy Agency, EBC Annex 58, Reliable building energy performance characterisation based on full scale dynamic measurements. Leuven: KULeuven; (2016). Available online: http://www.iea-ebc.org/projects/project?AnnexID=58

[11] James G., Witten D., Hastie T., Tibshirani R., An Introduction to Statistical Learning. MIT Press, 2010.

[12] Liaw A., Wiener M., Classification and Regression by randomForest (2002), R News, Volume 3, Nummer 3, p. 18-22, Available online: https://CRAN.R-project.org/doc/Rnews/