Indoor CO2 Level-Based Occupancy Estimation At Low-Scale Occupant Using Statistical Learning Method

by Haolia Rahman

Submission date: 13-Mar-2022 01:25PM (UTC+0700)
Submission ID: 1782986310
File name: at_Low-Scale_Occupant_using_Statistical_Learning_Method.docx.pdf (624.86K)
Word count: 3197
Character count: 15697

Indoor CO_2 Level-Based Occupancy Estimation At Low-Scale Occupant Using Statistical Learning Method

Haolia Rahman
Dept. of Mechanical Engineering
Politeknik Negeri Jakarta
Depok, Indonesia
haolia.rahman@mesin.pnj.ac.id

Abdul Azis Abdillah
Dept. of Mechanical Engineering
Politeknik Negeri Jakarta
Depok, Indonesia
abdul.azis.a@mesin.pnj.ac.id

Asep Apriana
Dept. of Mechanical Engineering
Politeknik Negeri Jakarta
Depok, Indonesia
asep.apriana@mesin.pnj.ac.id

Devi Handaya
Dept. of Mechanical Engineering
Politeknik Negeri Jakarta
Depok, Indonesia
devi.handaya@mesin.pnj.ac.id

Idrus Assagaf
Dept. of Mechanical Engineering
Politeknik Negeri Jakarta
Depok, Indonesia
idrus.assagaf@mesin.pnj.ac.id

***Abstract*— Most of the occupancy estimations based on indoor CO_2 levels are tested on a large-scale number of occupants such the order of tens or hundreds. Logically because the pattern of the occupancy and CO_2 level is about similar at a broad range of occupants. In the present study, a small office room with an occupancy scale of 0-6 people was tested. The Statistical Learning method is used to estimate the number of occupants, including Decision Tree, Random Forest classifier, SVM, Logistic regression, *K*-Nearest Neighbor, and Neural Network. A combination of training and testing data set is applied to the methods and a comparison has been made in order to distinguish their accuracy. The result shows that the accuracy of self-estimation and cross-estimation is ranged from 86-100% and 86-94% respectively. It also found that the estimation accuracy of self-and-cross validation does not significantly increase with the increase of data set combination.**

Keywords*— *occupancy estimation, carbon dioxide, statistical learning, low-scale occupants.

I. INTRODUCTION

A real-time data of the number of occupants in a modern building is valuable information for building management to implement the building-efficient strategy. For example, they can shut down the HVAC and lighting system when the building is vacant. This strategy has proven to be effective in reducing energy consumption by 15% [1]. Various methods and devices have been introduced from the literature on the subject of occupancy estimation, among them are using a video camera, RFID tags, PIR sensors, Wi-Fi, and indoor carbon dioxide (CO_2) levels. Existing research recognizes that video camera has proven to have high accuracy in occupancy estimation reach to 93.32% [2], though this method may interfere the privacy of the occupant. Other than that, the accuracy of occupancy estimation using RFID tags and Wi-Fi-based devices is very dependent on the occupants using the device or not. While another method like a PIR sensor has a difficulty in detecting stationary occupants.

Carbon dioxide is a byproduct of human-exhaled which is an odorless, colorless, non-toxic, and non-flammable gas at room temperature. Using CO_2 sensors as one of the environmental sensors could be a relevant method to quantify the number of occupants since the magnitude of CO_2 levels is most related to the number of occupants among other environmental sensors such as temperature, humidity and sound sensor [3]. However, the debate about using indoor CO_2 levels has gained fresh prominence with many arguing such as: being sensitive to excessive openings, CO_2 removal (i.e. plants and enclosure cracks), sensor response, air mixing, and other CO_2 generation (e.g. animals and combustion). Most of the occupancy count based on indoor CO_2 levels are tested in a large room with occupancy count tens to hundreds [4,5]. At this stage, indoor CO_2 and occupancy profiles will superimpose each other resulting in a relatively small error in estimation.

Generally, CO_2-based occupancy estimation refers to physical methods and statistical methods. The physical method strongly depends on the model and its parameters [6,7]. While the statistical method depends on data sets for primary learning tests [8]. Zuraimi [9] compared the accuracy of occupancy estimation between the physical method and the statistical method and postulated that the accuracy of the statistical method is higher than the physical method.

The scope of the present study is focusing on the occupancy estimation at a small room with a low occupancy scale. We use indoor CO_2 level and ventilation rate, and the method of Statistical Learning (SL) to estimate the number of occupants. The methods are limited to Decision Tree (DT), Random Forest Classifier (RFC), Support Vector Machines (SVM), Logistic Regression (LR), *k*-Nearest Neighbor (*k*-NN), and Neural Network (NN). The aim of the present study is to compare the accuracy from the combination of several training and testing data sets.

II. RELATED WORK

Methods based on SL to estimate the number of occupants using indoor CO_2 level has been numerously explored by numbers of researchers. Hailemariam [10] has attempted to evaluate the impact of the DT method using multiple sensors of such as CO_2 level, current, light, motion, and sound which the accuracy of occupancy detection was shown to have 80.02-98.44%. Hailemariam reveals that the DT method can improve occupancy detection systems based on motion sensors alone. Not only for the purpose of occupancy detection, DT method is also applicable for occupancy estimations [11, 12, 13].

The method of RF is essentially a collection of Decision Trees. Candanedo conducted a series of trials in occupancy estimation which he mixed several methods of SL and many combinations of sensors [14]. He reveals that the accuracy of the estimation using RF shows quantities of 78.76% and

64.21% at first and second testing respectively. However, Kallio [13] measures that RF has lower accuracy than DT in occupancy estimation using a one-year data set.

Support Vector Regression is a well-known method that has advantages in optimizing pattern recognition systems with good generalization capabilities [15]. This method has been widely applied to predicting and estimating the number of occupants based on indoor CO_2 level [16] which fairly average error estimation. Meanwhile, Chen [17] has combined the Inhomogeneous Hidden Markov Model with Multinomial Logistic Regression and shows that the combination is more effective than Hidden Markov Model solely.

III. METHOD

A. Testbed

The testbed is a single zone of an office room with an area and height of 34 m^2 and 2.6 m respectively. It is equipped with sensors, and ventilation systems such as return and supply ducting, and centrifugal fans as illustrated in Fig. 1. The airflow passing through the ducting is basically measured using a velocity meter (hot wire) and converted into flow rate based on the logarithmic Chebyshev method [18]. The airflow rate enables to be adjusted via a controllable fan which is embedded in both supply and return ducting. The indoor CO_2 level is measured using two CO_2 sensors located in the return ducting and in the middle of the room, while the outdoor CO_2 concentration is measured via CO_2 sensor located in the supply ducting. The type of CO_2 sensor is a Kimo probe connected to the transmitter C310-HO series with the accuracy ±3% of reading or ±50 ppm and the velocity sensor is Kanomax 6501 series with the accuracy ±3% of reading or ±0.01 kPa. The ventilation scheme used for the present measurement was proportional to the number of occupants, which increase and decrease of the airflow rate according to occupant entering and leaving the room. A laser beam is installed in between the frame of the door to record the ground truth of the occupants as well as input control of ventilation rate. The interval of collecting data was set in one minute which stores in the DAQ system.

B. Statistical model

Decision tree is one of the supervised machine learning algorithms, which can be used for classification and regression problems. The tree structure is broken down into smaller parts called a node. The node may have more than two branches depending on the attribute test conditions and the selected attribute. The attribute selection measures include entropy (S) and information gain (S, A) as formulated below:

$$\text{Entropy}(S) = \sum_{i=1}^{c} - P_i \log_2 P_i, \quad (1)$$

where entropy is the amount of information that acquired to describe the sample, c is the number of partitions S, and P_i is the proportion of category i elements over the total number of recorded sample.

$$\text{Gain}(S, A) = (S) - \sum_{v \in \, values\,(A)} \frac{|S_v|}{|S|} (S_v), \quad (2)$$

Where A is the variable which is tested, v is the possible value for the variable A, and S_v is the number of samples for the value v. The $|S_v|$ and $|S|$ is the cardinality of S_v and S, respectively. Entropy (Sv) is the entropy for a sample that has a value of v.

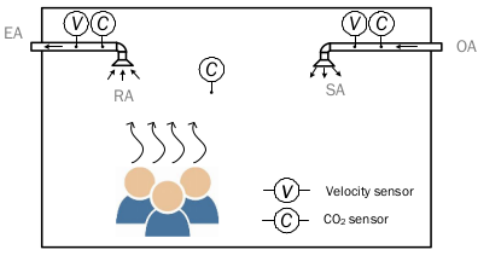

Fig. 1. The schematic diagram of sensors in the test-chamber

Theoretically, the RF method will improve the accuracy of the DT method since the RF method is a combination of each tree from a selected DT model. Determination of the classification by RF is taken based on the voting results of the formed tree.

Unlike RF, the goal of the SVM algorithm is to find the best hyperplane in N-dimensional space (a space with N-number of features) that serves as a clear separator for the input data points. The SVM algorithm determines the best hyperplane that is able to separate a two classes which have a maximum margin. The optimization problem of SVM is as follows:

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i, \quad (3)$$

subject to :

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \; \xi_i \geq 0, \; i = 1,2,\ldots,N, \quad (4)$$

where $x \in R^n$ is a training data point, $y \in \{1, -1\}$ is its label, $b \in R$ called a bias, ϕ is a mapping function, w is weights, and C is the parameter to control slack variable ξ.

In the case of classification, logistic regression works by calculating the class probability of a sample. The LR equation model is formulated as :

$$\log\left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_N X_N, \quad (5)$$

where Y=(0,1) is a binary label, $X = (X_1, .., X_n)$ are n explanatory variables selected based on the Akaike Information Criterion and $\beta = (\beta_0, \ldots, \beta_n)$ is the estimated regression coefficient.

Method of k-NN estimates the conditional distribution of Y given X by calculating the closest distance and classifying the observation data into the class with the highest probability. The k-NN look for the positive integer k observations closest to the test observation x_0 and estimates the conditional probability that it belongs to class j as follows :

$$\Pr(Y = j|X = x_0) = \frac{1}{k} \sum_{i \in N_0} I(y_i = j), \quad (6)$$

where N_0 is the closest set of k-obesrvations and $I(y_i = j)$ is an label whose value is 1 if the observation value in N_0 is not 0 and vice versa. After calculate these probabilities then the method assigns the greatest probability value of x_0 to the corresponding class.

The algorithm of NN were inspired by perceptron and neurons in the human brain. The perceptron accepts input in the form of a numeric number then processes it to produce an output. A perceptron consists of several components, namely: input (x_i) whereas $= 1,2,\ldots,m$, weights (w_i) and bias (w_o),

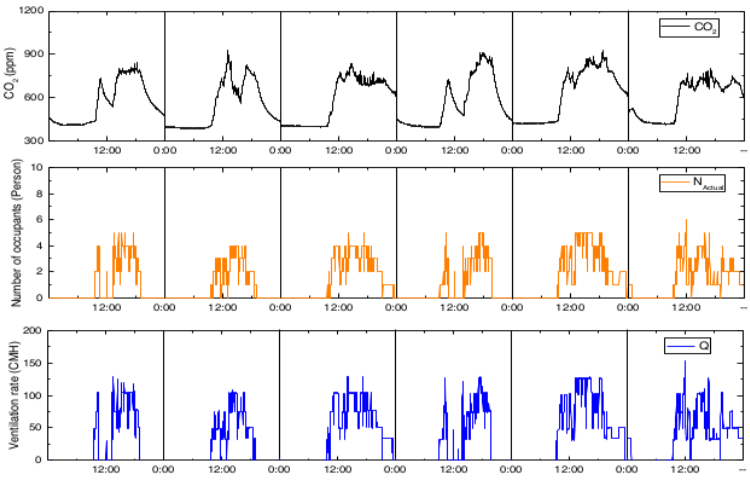

Fig. 2. Indoor CO_2 profile, actual number of occupans, and venrtilation rate profile.

activation function or non-linearity function (g), and output ($\hat{y}$) which formulated as follow:

$$\hat{y} = g(w_o + \sum_{i=1}^{m} x_i w_i). \quad (7)$$

We use Python 3.6.9 to run the algorithm and calculate the accuracy of occupancy estimation from the experimental data set.

IV. Result and Discussion

The indoor CO_2 level as a result of the occupancy (N) and ventilation rate (Q) pattern can be found in Fig. 2 which was taken for 6 days separately. The CO_2 level increase and decrease as occupant enter and leave the room with the average of occupancy 2-3 person at an occupied time. The ventilation rate profile almost fit with the occupancy profile which more less affect the CO_2 profile. The estimation accuracy of each model is summarized in Table 1 which is denoted as NDT, NRFC, NSVM, NLR, N*k*-NN, and NNN. We divided the learning data set into training (TR) and testing (TE), which combination has varied within increased and decreased by 10%. Similarly, results of the estimation accuracy are presented in TR for self-validation and TE for cross-validation. It is evident from the Table that the DT and RFC methods having perfectly accurate using self-validation followed by SVM and *k*-NN, NN, and LR respectively. The accuracy of cross-validation shows that RFC, SVM, and *k*-NN have the highest accuracy, followed by NN, DT, and LR. What is more interesting about the data in this Table is that the estimation accuracy of cross-validation at all methods does not significantly increase with the increase of TR data set (or decrease in TE data set). Moreover, the accuracy of self-validation was found almost stagnant at the increase and decrease of data set combination.

In detail, the occupancy estimation pattern of six methods obtained from 50% of TR and TE data set is shown in Fig. 3. The observation period from Fig. 3 was captured for 16 hours which mostly during the occupied time. From the chart, the occupancy estimation profiles (blue line) are superimposed with the ground truth (NGT) (red line). Most of the methods are fit well when the room is unoccupied and they respond well when indoor CO_2 levels start to build up at early occupation. Even though estimations are very difficult to catch all the NGT fluctuations (short change) but still follow the global trend of NGT.

TABLE I. The Accuracy of Six Methods.

Data TR (%)	Data TE (%)	Accuracy (%)											
		NDT		*NRFC*		*NSVM*		*NLR*		*Nk-NN*		*NNN*	
		TR	*TE*	*TR*	*TE*	*TR*	*TE*	*TR*	*TE*	*TR*	*TE*	*TR*	*TE*
10	90	100	89	100	91	93	93	88	86	93	91	89	86
20	80	100	90	100	92	95	92	87	86	94	92	88	87
30	70	100	89	100	92	94	92	87	87	94	92	87	87
40	60	100	89	100	92	93	93	86	87	93	92	88	88
50	50	100	89	100	92	93	93	87	87	93	92	88	88
60	40	100	89	100	92	93	93	86	87	93	93	88	89
70	30	100	90	100	93	93	93	87	87	93	93	88	89
80	20	100	91	100	94	93	94	87	87	93	94	87	88
90	10	100	91	100	94	93	94	87	87	93	94	91	93
Average		100	90	100	92	93	93	87	87	93	93	88	88

NDT = occupancy estimation using Decision Tree
NRFC = occupancy estimation using Random Forest classifier
NSVM = occupancy estimation using Support Vector Machines
NLR = occupancy estimation using Logistic Regression
Nk-NN = occupancy estimation using *k*-Nearest Neighbor
NNN = occupancy estimation using Neural Network

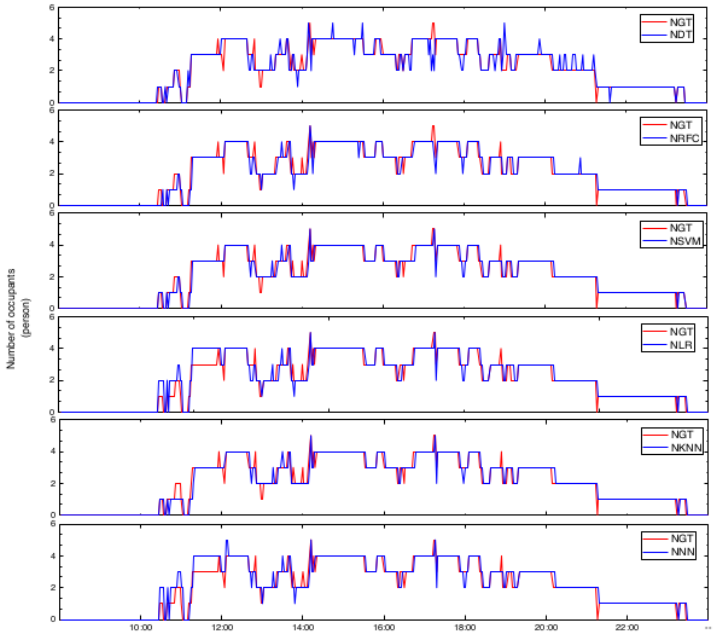

Fig. 3. The occupancy estimations using six methods of SL.

V. Conclusion

Six methods of Statistical Learning have been successfully implemented for occupancy estimations at room with low-scale occupants based on indoor CO_2 levels. Our results indicate that methods of RFC, SVM, and *k*-NN have the highest accuracy when validated using testing data set (cross-validation), followed by NN, DT, and LR. We also found from the combination of training and testing data set that the increase of training or decrease of testing data set, does not significantly increase the accuracy of cross-estimation. The empirical findings in this study provide a new understanding of how the occupancy estimation profile can follow most of the ground truth profile, but still have difficulty following a short change of real occupancy profile.

Acknowledgment

This work was funded by UP2M Politeknik Negeri Jakarta through the PDUPT scheme (contract number: B.264/PL3.B/PN.003/2021).

References

[1] CIBSE, Building Control System, Chartered Institute of Building Services Engineers, London, 2009.

[2] P. W.Tien, S. Wei, and J. Calautit, "A computer vision-based occupancy and equipment usage detection approach for reducing building energy demand," Energies, vol. 14(1), 156, 2021.

[3] R. Zhang, K. P. Lam, Y. Chiou, and B. Dong, "Information-theoretic environment features selection for occupancy detection in open office spaces," Build. Sim., vol. 5, pp. 179–188. 2012.

[4] Z. Sun, S. Wang, and Z. Ma, "In-situ implementation and validation of a CO_2-based adaptive demand-controlled ventilation strategy in a multi-zone office building." Build. and Env., vol. 46(1), pp. 124–133, 2011.

[5] T. Lu, X. Lü, and M. Viljanen, "A novel and dynamic demand-controlled ventilation strategy for CO_2 control and energy saving in buildings," En. and Build., vol. 43, pp. 2499–2508, 2011.

[6] H. Rahman and H. Han, "Bayesian estimation of occupancy distribution in a multi-room office building based on CO_2 concentrations." Build. Sim., vol. 11, pp. 575–583, 2018.

[7] H. Rahman and H. Han, "Real-time ventilation control based on a Bayesian estimation of occupancy," Build. Sim., vol. 14(5), pp. 1487–1497, 2021.

[8] A.G. Alam, H. Rahman, J.K. Kim, and H. Han, "Uncertainties in neural network model based on carbon dioxide concentration for occupancy estimation." J. of Mech. Sci. and Tech., vol. 31, 2573–2580, 2017.

[9] M.S. Zuraimi, A. Pantazaras A, K.A Chaturvedi, J.J Yang, K.W. Tham, S.E. Lee, "Predicting occupancy counts using physical and statistical CO_2-based modeling methodologies." Buil. and Env., vol. 123, pp. 517–528, 2017.

[10] E. Hailemariam, R. Goldstein, R, Attar, and A. Khan, "Real-time occupancy detection using Decision Trees with multiple sensor types," Sym. on Sim. for Arch. and Urban Design, 2011.

[11] M. H. Toutiaee, "Occupancy detection in room using sensor data," arXiv:2101.03616v1.

[12] M. Amayria, A. Arorab, S. Ploixa, S. Bandhyopadyayc, Q.D. Ngod, and V.R. Badarlab, "Estimating occupancy in heterogeneous sensor environment," En. and Build., vol. 129, pp. 46–58, 2016.

[13] J. Kallio, J. Tervonen, P. Rasanen, R. Makynen, J. Koivusaari, and J. Peltola, "Forecasting office indoor CO_2 concentration using machine learning with a one-year dataset", Build. and Env., vol. 187, 107409, 2021.

[14] L.M. Candanedo, and V. Feldheim, Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models," En. and Build., vol. 112, pp. 28–39, 2016.

[15] A.A. Abdillah and Suwarno, "Diagnosis of diabetes using support vector machines with radial basis function Kernels," Int. J. of Tech, vol. 7(5), pp. 849–858, 2016.

[16] Z. Han, R.X. Gao, and Z. Fan, "Occupancy and indoor environment quality sensing for smart buildings," IEEE Int. Inst. and Meas. Tech. Conf. Proc..

[17] Z. Chen, Q. Zhu, M.K. Masood, and Y.C. Soh, "Environmental sensors based occupancy estimation in buildings IHMM-MLR," vol. 13 (5), pp. 2184 – 2193, 2017.

[18] ISO 3966, Measurement of fluid flow in closed conduits—velocity area method using pitot static tubes. The Int. Org. for Stand. 2008

Indoor CO2 Level-Based Occupancy Estimation At Low-Scale Occupant Using Statistical Learning Method

ORIGINALITY REPORT

5%	1%	4%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Dewi Yanti Liliana, Diah Priharsari. "Tsunami Early Warning Detection using Bayesian Classifier", 2019 2nd International Conference of Computer and Informatics Engineering (IC2IE), 2019 Publication	1%
2	Luis Rueda, Kodjo Agbossou, Alben Cardenas, Nilson Henao, Sousso Kelouwani. "A comprehensive review of approaches to building occupancy detection", Building and Environment, 2020 Publication	1%
3	Alzahra Badi, Abdulrhman Mohammed Osama Elzwaie, Mohamed Alshareef Baayu, Faraj Mohamed Darrat et al. "Detection of Self–installed Mobile Repeaters", The 7th International Conference on Engineering & MIS 2021, 2021 Publication	1%
4	koreascience.kr Internet Source	1%

5	Kurniabudi, Deris Stiawan, Darmawijoyo, Mohd Yazid Bin Idris, Alwi M. Bamhdi, Rahmat Budiarto. "CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection", IEEE Access, 2020 Publication	1 %

Exclude quotes	On	Exclude matches	< 17 words
Exclude bibliography	On		