CO_2-based grey-box model to estimate airflow rate and room occupancy

Sebastian Wolf[1,*], *Maria* Justo Alonso[2], *Davide* Calì[1], *John* Krogstie[2] , *Hans Martin* Mathisen[2], and *Henrik* Madsen[1,2]

[1]Technical University of Denmark
[2]Nowegian University for Science and Technology

Abstract. In the existing building stock, heating, cooling and ventilation often run on fixed schedules assuming maximal occupancy. However, fitting the control of the HVAC system to the building's real demand offers large potential for energy savings over the status quo. Building occupants' presence as well as mechanically supplied and infiltrated airflow rates provide information that enables to define tailored strategies for demand-controlled ventilation. Hence, real-time estimations of these quantities are a valuable input to demand-controlled built environments. In this work, the use of stochastic differential equations (SDE) to estimate the room occupancy, infiltration air-rate and ventilation air-rate is investigated. In particular, a grey-box model based on a carbon dioxide (CO_2) mass balance equation is presented. The model combines knowledge about the physical system with statistical, data-driven parameter estimation. Furthermore, the proposed model contains uncertainty parameters. This is in contrast to purely deterministic models based on ordinary differential equations, where uncertainty is usually disregarded. The suggested model has been tested in a naturally ventilated and in a mechanically ventilated environment; the performance in these two cases has been compared. We show that the ability to address measurement errors and non-homogeneous conditions in the room air implies that the suggested SDE-based grey-box approach is suitable in the context of demand-controlled ventilation.

* Corresponding author: sewo@dtu.dk

1 Introduction

Heating, cooling and ventilation in buildings usually run on fixed schedules, in many cases even constantly throughout the day, all days. Furthermore, ventilation systems often run with a constant air flow rate that is adjusted to maximum occupancy. Hence, reducing operation hours and airflow rates to the required extent enables potential energy-savings. For this reason, reliable room occupancy estimates are needed to provide valuable information for an energy-efficient operation. HVAC control strategies would benefit from accurate presence estimates [1] based on reliable measurements. Moreover, reliable data-driven presence profiles are required for the development of building occupancy models, which serve as input for building energy simulations.

In the present work, an occupancy estimation model is presented. The model estimates room occupancy based on a carbon dioxide (CO_2) mass balance equation. Alternatively, the CO_2 level can be estimated using the room occupancy as input. In contrast to earlier studies, that use ordinary differential equations (ODE) to describe the mass balance, the presented approach employs a grey-box model based on stochastic differential equations (SDE). With this, it is possible to address and quantify the uncertainty that derives from measurement errors as well as from inadequacies in the description of the physical system. The latter may concern the assumption of a homogeneously distributed pollutant concentration in the room air, the assumption of a constant infiltration rate and other oversimplifications in the model description. The stochastic framework of the model further allows parameter estimation based on a maximum likelihood approach. The results of the model being applied in two different scenarios are presented. The first data-set was recorded in a mechanically-ventilated office room in Norway. The second dataset stems from a naturally ventilated office room in Denmark.

This publication is organised as follows. An overview of occupancy estimation models in literature is given in Section 1.1. A brief description of the two data-sets is contained in Section 1.2. Section 2 is dedicated to a description of the methods used in this work. An introduction to the employed grey-box model is followed by a description of the occupancy estimation algorithm. In Section 3, the results for both data-sets are presented; and Section 4 and 5 contain a discussion and the conclusion, respectively.

1.1 Related work

Much research effort on occupancy estimation has been done, in particular in recent years ([2-6]). The work of Yang et al [7] provides a comprehensive review. The prevailing way to capture occupants' presence in buildings are motion detectors, such as passive infrared sensors, despite the following limitations: Motion sensors

- have a limited coverage range and can only detect movements in their direct line-of-sight.
- are unable to detect the number of occupants.
- fail to detect immobile occupants, e.g. someone sitting still at a computer.
- are not able to detect the beginning of vacancies (periods of absence).

Motion sensors are often employed for lighting control, which benefits from the sensors' short reaction time. For lighting, the above-mentioned limitations are not major, since lighting usually does not depend on the number of people (greater than one) in the room, and a failed presence detection is easily recognised and corrected by the occupants. However, movement within a defined space is just one of many information sources that can indicate occupants' presence. Among other sources, camera images, PC usage information and environmental variables such as temperature, humidity, noise and CO_2 level can be used to gather occupants' presence patterns. To this end, there has been extensive research on model-based occupancy estimation making use of this wide range of input variables. Several studies in the literature use a statistical or machine learning approach. The authors in [6], for instance, explore the use of decision trees for occupancy detection based on features from a combination of CO_2, electric current, lighting, motion and sound sensors in a single office cubicle. They found in their study that a decision tree, trained only on features (i.e., input variables) from a motion sensor outperformed any other combination of inputs, even the inclusion of all sensors. In the light of these results, they suggest exploring alternative classification methods that are less prone to overfitting. The work of [5] presents a model based on a computer vision algorithm. In their work, camera images are automatically interpreted to detect human figures in a defined space. However, privacy concerns make the usage of video-images generally problematic. Even if the images are recorded in a low resolution and are not stored, the fact alone that cameras are installed may intrude the occupants' perceived privacy. The authors in [8] explore a method to localise and count building occupants using a building's Wi-Fi network. Each mobile device is associated to the wireless access point to which it is connected, providing an estimate of the device's location. Moreover, the number of connected devices can be obtained at each access point which provides an estimate of the number of occupants in one area. Showing potential for estimating occupancy on building scale, the work reveals that the method is not suitable for a finer spatial resolution, since access point ranges overlap, and mobile devices do not necessarily connect to the closest point.

A different framework for occupancy estimation are Hidden Markov Models (HMM). These are statistical models that allow to draw inference about the unobserved occupancy state from of one or more observed variables. The authors in [9] use of HMM for occupancy modelling

based on observations from smart electricity meters. However, in most cases, the observations contain environmental variables such as the CO_2 concentration ([10-12]). An auto-regressive Hidden Markov Model, which is an extension of the above-mentioned HMM, is applied in the work of [13]. It is shown that this extension addresses and compensates for auto-correlation in the CO_2 observations that are otherwise neglected.

Another way of describing the relation between CO_2 and occupancy is a mass balance equation ([2,14-16]). The advantage of this methodology lies in the direct physical interpretation of the model. Here, it is assumed that changes of the CO_2 level in the room air are uniquely defined by the air exchange rate and the CO_2 production of human respiration. Mathematically, this results in an ODE in the mass (or volume) of CO_2 in the air.

The work of [2] states the importance of occupancy estimation for demand-controlled ventilation. They present an occupancy estimation model based on a CO_2 mass balance equation and show that their dynamic model outperforms the steady-state method that was proposed in the ASHRAE standard at that time. In their works, all physical parameters are assumed known by measurements. Therefore, the practical use of this model is limited to situations where this information is available. The authors in [14] apply the same approach to estimate occupancy and use the estimates as external input for two different model-based controllers for ventilation control. The model presented by [15] is also based on a CO_2 mass balance. The model is tested on different residential and non-residential buildings with promising results pertaining occupancy estimation. In contrast to the aforementioned works, in [15], most of the model parameters are estimated through an optimisation algorithm. However, the CO_2 generation per person is assumed known and constant. A limitation of the deterministic ODE approach described in the works above is that it does not account for disturbances, such as a non-homogeneous air distribution, differences in the occupants' CO_2 production and other simplifications in the model as well as for measurement errors. A grey-box model, which is generally a model based on physical considerations completed by data-driven estimations for the unknown factors (parameters) and can overcome these limitations. In particular, SDE can be used instead of ODE. SDE include a noise term in the system equation, addressing disturbances in the system. An additional observation equation including a term for measurement error completes the model. The quantification of the error terms can be used to estimate the correct state value (in this case the CO_2 level). The strength of this approach lies in its ability to address and quantify the uncertainty that corresponds to simplifications in the physical model and to measurement error. An SDE model based on a CO_2 mass balance equation to estimate the infiltration rate, the ventilation rate and the CO_2 generation per person is presented by [17]. However, in their method, occupancy is considered a model input, and is hence not estimated. In the present work, a model similar to the one presented in [17], but additionally able to estimate the number of occupants, is suggested.

2 Methods

This section introduces the methods used in this work. First, two data collections, to which the presented model was applied, are described. Further, the grey-box model is presented which describes the variation in the room's CO_2 levels dependent on the occupancy. Subsequently, it is described how the model parameters are estimated, and how the model is used to estimate the room's occupancy.

2.1 Data collection description

2.1.1 Dataset 1

CO_2 level in parts per million (ppm) and occupancy were measured during nine week days in August 2018 in a mechanically-ventilated office in Trondheim, Norway designed for six occupants. The building is built in 1962, was renovated in 2016. The room is located in the ground floor has a floor area of 40 m^2. It has three north facing windows that were closed during the entire measurement period, and a glass wall with the door to a corridor in the south. The door was closed except for short periods corresponding to entering or leaving the room. The timestep of the CO_2 and occupancy recordings is five minutes. The occupancy data was recorded manually. The CO_2 sensor was located in the center of the room at a height of 1.30 m. The measurements were taken using a Vaisala GMP222 Carbon Dioxide Probe which has resolution of 10 ppm.

2.1.2 Dataset 2

Dataset 2 stems from a naturally ventilated five persons office room near Copenhagen, Denmark. On four consecutive days in February, Wednesday to Saturday, the CO_2 level in ppm was measured by a SenseAir S8 sensor on a timestep of five minutes. The office room is located in the ground floor of a 1960's building. It has two operable windows facing west and one door to a corridor facing east. The windows were closed during the entire measurement period. The room has a floor area of about 27.5 m^2 and a height of about 3 m. During the measurements, the CO_2 sensor was located in the center of the south wall at a height of 1.60 m. The 30 cm difference in the sensor height compared to Dataset 1 derives from the fact that the shelves on which the sensors were placed were of different heights. However, we do not expect a significant influence on the results due to this difference. Occupancy was recorded only on the last day of the measurement period. The acoustic noise level in decibel was recorded during the entire period.

2.2 Grey-box model

Under the following assumptions, the variation of the CO_2 concentration in a room can be described by the mass balance Equation (1),

- The CO_2 generation rate per person is constant over time and for all occupants.
- The outdoor CO_2 concentration is constant.
- The room's CO_2 level is influenced by human exhalation (production), natural and mechanical ventilation, and by air infiltration (removal) only.
- The room air is spatially homogeneous.

$$dX_t = -\left[n_{air} \cdot (X_t - c_e) + \frac{\dot{c}_{occ}}{V} \cdot n_{occ}\right] dt \qquad (1)$$

with

$$n_{air} = n_{mec} + n_{nat} + n_{inf} \qquad (2)$$

where X_t is the room CO_2 level, c_e the outdoor CO_2 level, $\dot{c}_{occ}$ the CO_2 production per occupant, V the room volume, n_{occ} the number of occupants in the room, and n_{mec}, n_{nat} and n_{inf} are the air change rates of mechanical ventilation, natural ventilation and infiltration, respectively. In reality, however, the above-mentioned assumptions will never hold entirely, and the here-described mass balance equation can only approximate the correct CO_2 value ([2,15,18]). One way to address this uncertainty is to employ a grey-box model, which takes the physical equation as a basis and derives the unknown model and uncertainty parameters from data [19]. Hence, a noise term is introduced in the differential equation. This is done by adding a Wiener process (also Brownian motion [20]), which represents the integrated version of a Gaussian white-noise process. This results in the SDE in Equation (3). Furthermore, it is assumed that the CO_2 sensors used are not fully precise. For this reason, an additional observation equation (4) with a measurement error term is added.

$$dX_t = -[n_{air} \cdot (X_t - c_e) + \dot{c}_{occ} \cdot n_{occ}]dt + \sigma \cdot d\omega \qquad (3)$$

$$Y_{t_k} = X_t + \sigma_\varepsilon \qquad (4)$$

2.3 Parameter estimation

To estimate the model parameters, occupancy is assumed known. The parameters are estimated using the maximum likelihood estimation (MLE) approach, which is outlined in the following: The likelihood function is the joint probability of all CO_2 observations as a function of the model parameters θ. In the MLE method, the likelihood is maximised with respect to θ. The parameters which maximise the likelihood are the maximum likelihood estimates $\hat{\theta}$. In practice, usually the logarithm of the likelihood, called log-likelihood, is maximised, which leads to the same parameter estimates. The premiss of the MLE method is that, of all possible parameters, the most suitable are those which are most consistent with the observed data. In the above-described grey-box model, the joint probability of the observations can be expressed by the product of the one-step predictions $X_{i+1|i}$. It can be shown that $X_{i+1|i}$ are Gaussian distributed. Hence, they are fully characterised by their mean $\mu_{t_{i+1}|t_i}$ and variance $\sigma^2_{t_{i+1}|t_i}$. By using a Kalman filter ([21]), it can be shown that the log-likelihood is given by:

$$logL(\theta) = -\frac{1}{2}\sum log\,(2\pi) + log\left(\sigma^2_{t_i|t_{i-1}} + \sigma^2_\varepsilon\right) + \frac{\left(y_i - \mu_{t_i|t_{i-1}}\right)^2}{\sigma^2_{t_i|t_{i-1}} + \sigma^2_\varepsilon}, \qquad (5)$$

where y_i are the CO_2 observations. The log-likelihood can be maximised using numerical optimisation in R to obtain $\hat{\theta}$.

2.4 Occupancy estimation

In the second part of the model development, the CO_2 level is assumed known, and the occupancy state is assumed unknown. As model parameters, the maximum likelihood estimates $\hat{\theta}$ of Section 2.2 are used. In order to estimate the occupancy state vector n_{occ}, the likelihood function is maximised with respect to n_{occ}. This is a complex problem for two reasons. The number of unknowns equals the number of observations; hence, the dimension of the optimisation is very high. In other words, many parameters have to be estimated simultaneously. Second, n_{occ} is non-negative and integer-valued. This constraint is not respected by most numerical optimisers. Therefore, a custom optimisation routine is employed. The estimate vector n_{occ} is initialised as the zero vector. Subsequently, the vector is increased by one at that time point where the increase in likelihood is highest. The algorithm terminates when an increase in occupancy does not lead to an increase in likelihood for any time point.

3 Results

3.1 Dataset 1

Since the windows were closed in the recorded period, we assume $n_{nat} = 0$ in Equation (3). The data was divided into a training set of four consecutive days and a test set of five consecutive days. In a first step, the model parameters were estimated on the training set using the grey-box model described in Section 2.1. The parameter

estimates are shown in Table 1. The values in parentheses represent the standard error.

Table 1: Parameter estimates

	Dataset 1	**Dataset 2**
n_{inf} $[1/h]$	0.12 (-)	0.5 (0.03)
n_{mec} $[1/h]$	4.2 (0.09)	-
$\dot{c}_{occ}$ $[l/h]$	16.5 (-)	15.7 (0.51)
$c_e[ppm]$	444.3 (1.7)	405.3 (2.5)
$log\ (\sigma)$	2.48 (0.01)	1.55 (0.07)
$log\ (\sigma_\varepsilon)$	-4.78 (18.85)	0.26 (0.07)

Subsequently, the CO_2 levels were estimated for the training and test set, employing the estimated model and using occupancy as the known input. The CO_2 estimates are shown in Figure 1 and 2, respectively. The graphs include 5-minute and 1-day forecasts. The estimates of the training data (Fig. 1) are in-sample estimates, since the model parameters were estimated on the same dataset in this case. The forecasts on the test set (Fig. 2), however, are out-of-sample estimates as the set for parameter estimation (training set) and the set for CO_2 estimation (test set) are independent. Both for training set and test set, the CO_2 estimates follow the measurements well.

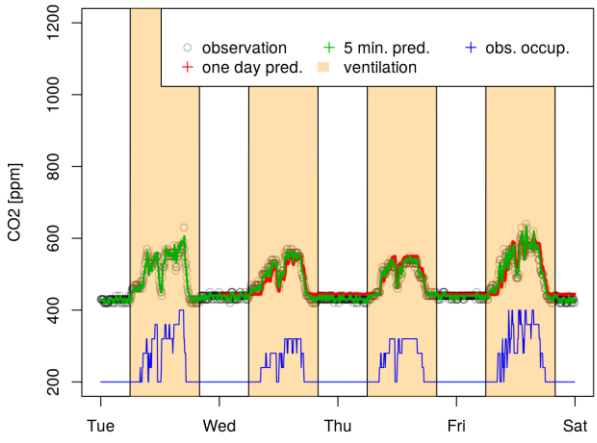

Fig 1. Dataset 1 - Training data

Finally, the occupancy was estimated using the estimated model and applying the algorithm described in Section 2.3. This was done on the training set and on the test set. The estimates are shown in Figure 3 and 4, respectively. The estimated occupancy was then compared to the measured occupancy. For binary occupancy, Table 2 shows the discrimination results, i.e. the true-positive rate (TPR), the true-negative rate (TNR) and the accuracy (ACC). Table 3 states the root mean square error of the occupancy estimation, i.e., the difference of estimated and recorded occupancy.

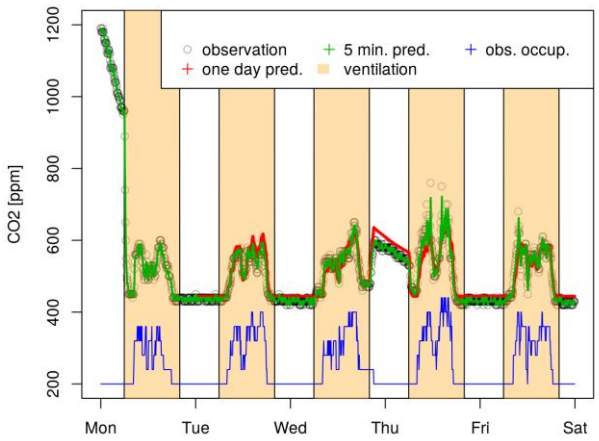

Fig 2. Dataset 1 - Test data

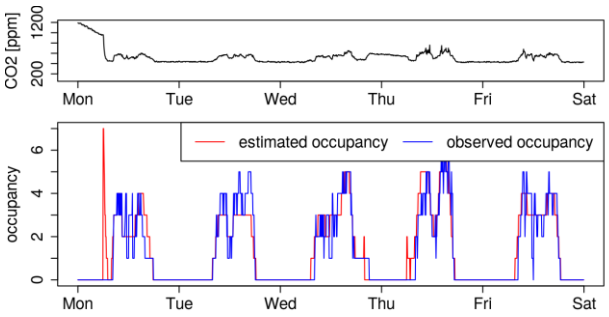

Fig 3. Dataset 1 - Training data

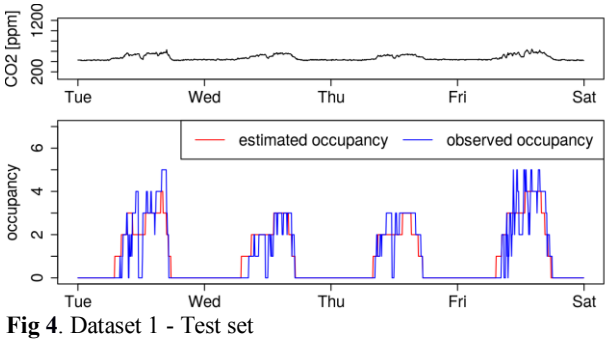

Fig 4. Dataset 1 - Test set

3.2 Dataset 2

In the case of dataset 2, the windows were closed and there is no mechanical ventilation. Therefore, both n_{nat} and n_{mec} are assumed zero. The training data consists of one day, on which occupancy was recorded manually. Due to the small sample size, the test set coincides with the entire dataset, which consists of four consecutive days. As for dataset 1, first, the model parameters were estimated on the training data using the grey-box model described in Section 2.1. The parameter estimates can be

found in Table 1. Subsequently, the CO_2 levels were estimated for the training set as shown in Figure 5. Finally, the occupancy was estimated on the test set, for which Table 2 shows discrimination results. The root mean square error could not be obtained as the ground truth occupancy was not captured during the test set period. The only validation reference are acoustic noise levels. However, these do not reveal any information about the number of occupants. The occupancy estimates of the model are shown in Figure 6.

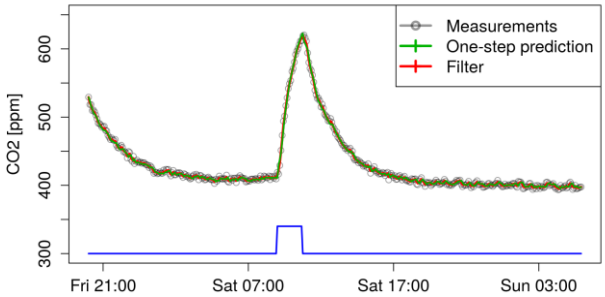

Fig 5. Dataset 2 - Training data

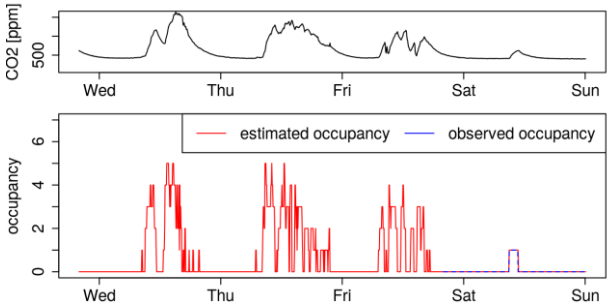

Fig 6. Dataset 2 - Test set

3.3 Comparison

Comparing the results of dataset 1 and 2, it has to be kept in mind that the occupancy was captured differently for the two test sets. For dataset 1, the reference occupancy stems from manual recordings, whereas for dataset 2, the reference derives from acoustic noise recordings in the room. As expected, for both datasets, the occupancy estimates are more accurate for the training data than for the test data. Nevertheless, the decrease in accuracy in the respective test sets is marginal, suggesting that the models are not overfitted. The training estimates of dataset 1 are less accurate than the training estimates of dataset 2. However, for results on the test sets, the opposite is true. This can likely be ascribed to the small sample size and simple structure of the training of dataset 2, compared to the test set. Overall, the test results are satisfying for both datasets.

Table 2: Discrimination Results

	TPR	**TNR**	**ACC**
Data 1 Training Set	0.83	0.99	0.92
Data 1 Test Set	0.90	0.98	0.94
Data 2 Training Set	*0.95*	1.00	0.99
Data 2 Test Set	0.81	0.90	0.88

Table 3: Root mean square error

	All data	**When occupied**
Data 1 Training Set	0.66	0.87
Data 1 Test Set	0.77	0.94

4 Discussion

For the test set of dataset 2, the root mean square was calculated once for all data points and once just for occupied periods. The latter is more meaningful, since the correct estimation of absence outside working hours is not challenging. On occupied periods, the root mean square error is 0.94 persons. However, this error rarely concerns a misclassification of the binary occupancy which can be seen from the model's high accuracy for binary occupancy. Instead, errors occur more often for high numbers of occupants.

The presented model assumes an equal CO_2 level in supply and outdoor air. However, analysis of dataset 1 showed that this assumption does not hold. Hence, model is oversimplified and could be improved by introducing different parameters for supply air and outdoor air CO_2 level.

In dataset 1, the parameter n_{inf} was estimated on a separate training dataset. The reason for this are the following conditions of the original training data: During the ventilated periods, the air exchange was dominated by the mechanical ventilation, whereas outside ventilated periods the indoor CO_2 concentration was at the level of the outdoor CO_2, since the office was unoccupied. This made it difficult to identify the infiltration rate. Hence, an additional period with turned-off mechanical ventilation, no occupants present and a high initial CO_2 level was used for the estimation of the infiltration rate.

The CO_2 production per person was not estimated but assumed with a value of 16.5 liter CO_2 per hour in dataset 1. As in the case of the infiltration rate, the reason is that the model was not fully activated by the training data. Occupancy and ventilation coincide for the major part of the data. Therefore, the ventilation rate and the CO_2 production per person were not clearly identifiable from the data. In dataset 2, the CO_2 production was estimated from the data. The estimate of 15.7 liter CO_2 per hour lies in a realistic range ([22]).

Overall, the results of parameter and occupancy estimation showed satisfying results. Since, as pointed out by [15,23,24], CO_2 sensors are increasingly getting integrated in buildings services, and are easy and relatively cheap to install, the here presented method can be considered as a candidate for the development of future demand-controlled HVAC systems.
The here presented model uses the CO_2 level exclusively as an indicator for occupant presence. It should be noted that the indoor quality also depends on factors that can be unrelated to occupancy, such as volatile organic compounds (VOC), which are not only produced by occupants but also by certain materials. Hence, a ventilation control that takes the here presented model as input for occupant detection, should additionally take pollutants not related to occupancy into account.

5 Conclusion

A model which describes the variation in room CO_2 level and can estimate room occupancy was presented. It can be used to develop demand-controlled HVAC strategies that take occupancy as input. For the first time, a grey-box model based on stochastic differential equations was employed to estimate room occupancy. The model was tested in one mechanically-ventilated and one naturally-ventilated environment. In both scenarios, it showed promising performance. In the light of the results, the model could be enhanced, e.g. by introducing additional parameters that describe the physical system more accurately, as long as overfitting is avoided.
The influence of number and location of the CO_2 sensors on the model performance is an open task for future work. Since the model achieves relatively high accuracies at this stage, a multivariate model that takes input from several sensors seems to the authors an unnecessary increase of complexity. To optimise the location of the single sensor, on the other hand, might result in an improvement of the model.
Moreover, larger training sets are required to fully activate the physical system and produce more robust parameter estimates. Furthermore, more ground truth occupancy information is needed to fully validate the model.

References

1. F. Oldewurtel, D. Sturzenegger, M. Morari, Appl, Energy **101**, 521–532 (2013).
2. S. Wang, J. Burnett, H. Chong, Indoor and Built Envir. **8** (1999) 377–391.
3. K. P. Lam, M. Höynck, B. Dong, B. Andrews, Y.-S. Chiou, R. Zhang, D. Benitez, J. Choi, Proc. of Building Sim. 2009.
4. Chenda Liao, P. Barooah, Proc. of American Control Conference, 2010, pp. 3130–3135.
5. Y. Benezeth, H. Laurant, B. Emilie, C. Rosenberger, Energy and Buildings **43** (2011) 305–314.
6. E. Hailemariam, R. Goldstein, R. Attar, A. Khan, Proc. SimAUD 2011.
7. J. Yang, M. Santamouris, S. E. Lee, Energy and Buildings **121** (2016) 344–349.
8. R. Melfi, B. Rosenblum, B. Nordman, K. Christensen, Proceedings of IGCC, (2011) 1-8.
9. J. Liisberg, J. Møller, H. Bloem, J. Cipriano, G. Mor, H. Madsen, Sustainable Cities and Society **27** (2016) 83–98.
10. B. Dong, B. Andrews, K. P. Lam, M. Höynck, R. Zhang, Y.-S. Chiou, D. Benitez, E. and B. **42** (2010) 1038–1046.
11. B. Dong, K. P. Lam, JBPS **4** (2011) 359–369.
12. L. M. Candanedo, V. Feldheim, D. Deramaix, Energy and Buildings **148** (2017) 327–341.
13. B. Ai, Z. Fan, R. X. Goa, American Control Conference (2014).
14. M. Gruber, A. Trüschel, J.-O. Dahlenbäck, Energy and Buildings **84** (2014) 548–556.
15. D. Calì, P. Matthes, K. Huchtemann, R. Streblow, D. Müller,, Building and Environment **86** (2015) 39–49.
16. F. Wang, Q. Feng, Z. Chen, Q. Zhao, Z. Cheng, Z. Jianhong, Y. Zhang, M. Jinbo, Y. Li, H. Reeve, Energy and Buildings **145** (2017) 155–162.
17. M. Macarulla, M. Casals, N. Forcada, M. Gangolells, A. Giretti, Measurement **124** (2018) 539–548.
18. Z. Sun, S. Wang, Z. Ma, Building and Environment **46** (1) (2011)
19. N. R. Kristensen, H. Madsen, S. B. Jørgensen, Proc. 13th IFAC Symposium (SYSID-2003)
20. B. Øksendal, Stochastic Differential Equations, 5th Edition, Springer, New York, 1998.
21. H. Madsen, Time Series Analysis, Chapman and Hall, 2008.
22. A. Persily, L. de Jonge, Indoor Air **27** (2017) 868–879.
23. W. Shen, G. Newsham, B. Gunay, Advanced Engineering Informatics **33** (2017) 230–242.
24. B. Gunay, A. Fuller, W. O'Brien, I. Beausoleil-Morrison, Proceedings of eSim 2016.