Energy & Buildings 183 (2019) 672–683

Contents lists available at ScienceDirect

Energy & Buildings

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

A Markov-Switching model for building occupant activity estimation

Sebastian Wolf[a,*], Jan Kloppenborg Møller[a], Magnus Alexander Bitsch[a], John Krogstie[b], Henrik Madsen[a,b]

[a] Technical University of Denmark, Denmark
[b] Norwegian University of Science and Technology, Norway

ARTICLE INFO

Article history:
Received 10 July 2018
Revised 19 November 2018
Accepted 22 November 2018
Available online 1 December 2018

Keywords:
Occupant behaviour
Occupant activity modelling
Occupancy detection
Hidden Markov models

ABSTRACT

Heating and ventilation strategies in buildings can be improved significantly if information about the current presence and activity level of the occupants is taken into account. Therefore, there is a high demand for inexpensive sensor-based methods to detect the occupancy or occupant activity level. It is shown that the carbon dioxide (CO_2) level in a room is dependent on the activity level rather than only on just the number of people. Therefore, this study suggests a new model based on the use of CO_2 trajectories to estimate the occupant activity level, trained on measurements both from a school classroom and from a Danish summerhouse. A hidden Markov-switching model was employed to identify the activity level. This modelling approach is a generalization of hidden Markov models, taking autocorrelation in the observations into account. This is done by an additional autoregressive part which models the persistence of the CO_2 concentration by relating the current value to past lags. The analysis of one-step prediction residuals shows that this method inherits the dynamics of the CO_2 curves much better than an ordinary hidden Markov model, and can therefore be considered a promising candidate for occupant activity estimation. Furthermore, it is shown that the presented model can be used for simulations of activity level and of the accompanying CO_2 levels.

1. Introduction

Creating a comfortable indoor climate in terms of temperature and air quality can require a lot of energy. In most buildings, this effort is made on a fixed schedule regardless of the actual occupancy state. Therefore, there is a high energy-saving potential if heating, ventilation and lighting periods are fitted to the occupancy patterns [1,2]. For automated ventilation control, however, it is not the binary occupancy state, i.e., present or absent, that should determine the ventilation rate. It is rather the total generation of CO_2 and air pollutants that play a role. The classical approach to address the rate of this generation is to associate it with the number of people in the room. However, CO_2 generation differs from person to person and is highly dependent on the physical activity level of the occupants [3,4]. Furthermore, human respiration is not always the single source of CO_2 in buildings [5]. Therefore, we suggest to identify the state for an entire room instead; and we will define the occupants' activity level by the rate the room air quality changes by the presence, physical activity and actions of all occupants as well as by ventilation. Notice that this is not necessarily proportional to the number of occupants.

In this work, we present a method for estimation, simulation and short-term forecasting of occupant activity levels. The class of models considered in this work are hidden Markov models (HMM) and autoregressive hidden Markov models (ARHMM). The latter are also referred to as Markov-switching models. The model's single input are trajectories of room CO_2 sensors which are increasingly getting integrated in building services, and are easy and relatively cheap to install, as pointed out in [6–8]. We tested the suggested approach in two environments of very different nature. The first dataset stems from a Danish primary school. Here, occupancy is very regular as pupils and teachers are present according to weekly time tables. The other dataset was measured in a Danish summerhouse. As it is used for vacation only, the occupants' schedule is rather relaxed and naturally yields a higher volatility. In general, the methodology presented here is not restricted to a type of building or occupancy behaviour. The models were built on the open source statistical software R [9]. The remainder

* Corresponding author.
E-mail address: sewo@dtu.dk (S. Wolf).

https://doi.org/10.1016/j.enbuild.2018.11.041
0378-7788/

of this section gives a brief overview of the related work the in literature.

1.1. Related work

There has been extensive research on modelling occupancy in buildings, both for office spaces and residential environments. The purpose of the models presented in literature is, for one part, the simulation of human behaviour, e.g. for building energy simulations (e.g. [10–12]); another part is dedicated to the estimation of the current occupancy status, with applications in demand-based building control (e.g. [13,14]). Since the here presented model can be seen as a merging of those two approaches, works from both applications are reviewed.

1.1.1. Simulation models

For simulation models, the occupancy state is commonly measured directly and assumed known during the development phase. The goal is to create synthetic occupancy profiles that inherit the same variation properties as the measurements. These profiles can then be used as input for building simulation tools. A commonly used approach to occupancy simulation models are Markov chains (e.g. [10,11]). In some cases, the chain's transition probabilities are time-inhomogenous, i.e., time dependent, as in the work of Andersen et al. [12]. Another approach used in literature is survival analysis ([15,16]). Here, the duration of occupied and vacant periods are sampled from a probability distribution that is based on measurements. An agent-based approach is employed by Liao et al. [17]. In contrast to most other studies, where occupancy is considered independently for single rooms, in their work the sum of occupied zones matches the total building occupation, and occupants are assigned to transition probabilities between zones. While most proposed models are probabilistic, Mahdavi and Tahmasebi [18] present a non-stochastic model. Instead of sampling from a probability distribution, the occupancy states are determined by comparing the probability of occupancy to a fixed threshold. All aforementioned models are developed in a batch training, i.e., parameters are estimated on one training set and do not change thereafter. However, Dobbs and Hencey [19] present an online-trained Markov-chain occupancy model for predictive HVAC control. The transition probabilities adapt with recent data, whereas the influence of older observations decreases by means of an included forgetting factor.

1.1.2. Estimation models

Among the studies of occupancy estimation, some models are based on physical considerations. Under the premiss that changes in the CO_2 concentration in a room are determined, apart from interactions with the exterior, only by CO_2 generation per person and the number of persons, some studies present an estimation algorithm based on the following mass balance equation,

$$\frac{\partial C_r}{\partial t} V_r = \dot{V}_{supp}(C_{supp} - C_r) + n_{occ} \cdot \dot{m}_{prod,occ} \ , \tag{1}$$

where C_r and C_{supp} are the CO_2 concentrations of room and supply air, V_r the room volume, $\dot{V}_{supp}$ the supply air volume flow, n_{occ} the number of occupants, and $\dot{m}_{prod,occ}$ the CO_2 generation per person ([6,20,21]). Other studies choose a statistical approach. For instance, Hailemariam et al. [22] use a decision tree for occupancy detection based on several features including CO_2, electricity use, light, motion and sound. Benezeth et al. [23] present a computer vision-based model. Here, camera images are automatically interpreted to detect human figures in a room.

Melfi et al. [24] explore a method to count and localise building occupants using the building's Wi-Fi network. At each wireless access point, the number of connected devices can be obtained. Further, each mobile device can be associated to the access point to which it is connected. The work shows potential for estimating occupancy on building scale, but 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 one. Kahn et al. [25] propose a method of estimating office occupancy and number of occupants using support vector machines (SVM) and k-nearest-neighbours (kNN). The features for their model are extracted from the environmental variables temperature, humidity, passive infrared (PIR) and audio levels as well as from the contextual factors computer-activity and calendar data. Newsham et al. [26] explore the possibility of detecting occupancy in offices by using data that is already collected by the occupant's computer or by an additional inexpensive sensor mounted to the computer. They use a genetic programming approach with inputs mouse activity, keyboard activity and webcam-based motion detection. They show that this model outperforms commercial PIR sensors, and can be used to decrease timeout-periods, hence saving energy in vacant periods. The results are promising, but restricted to environments, where occupancy is directly linked to working on computers.

Another common framework for occupancy estimation are hidden Markov models (HMM). The advantage of this model class is that it provides a complete stochastic framework based on the likelihood theory, inherently providing parameter and state estimators, measures of uncertainty and forecasts. Dong et al. [27] apply three different machine learning approaches to estimate occupancy and the number of occupants: HMM, Support Vector Machines and Artificial Neural Networks. They find that HMM perform best, when comparing the predicted states to the actually observed number of occupants. Dong and Lam [28] additionally apply a Gaussian mixture model to detect the number of occupants. Liisberg et al. [29] investigate in depth the use of HMM for occupancy modelling based on smart-meter electricity observations of private apartments in Spain. They test the possibilities of classifying the occupancy states based on hourly electricity data. Candanedo et al. [30] investigate different combinations of predictor variables and time scales for the use of occupancy detection. Their model is a homogeneous HMM with a Gaussian state-dependent distribution. Occupancy is treated as a binary variable. They find that a model with first-order difference of CO_2 concentration performs best. The humidity ratio is the second best predictor in their analysis. Ai et al. [31] compare a HMM with an autoregressive hidden Markov model (ARHMM) to estimate the number of office occupants, and conclude that the ARHMM has a higher accuracy. They use data from a controlled office lab with 6 to 10 occupants. A sensor network of PIR, CO_2, temperature, relative humidity, air-velocity, global thermometer and reed switches is used for the estimation.

In contrast to the present work, Ai et al. [31] use a time-homogeneous Markov chain for the underlying state transitions. Hence, the probabilities for occupancy changes are time-independent which appears to be an overly strong assumption. Furthermore, the number of model states is given a priori, whereas in the present work, it is used as a tuning parameter, providing the model with more generality.

Most studies that make use of CO_2 trajectories focus on office buildings, where the physical activity level is relatively homogeneous and CO_2 generation mainly originates from human respiration. Much less attention has been drawn to other types of buildings with less homogeneous activities and other sources of CO_2 generation such as smoking or cooking with a gas stove. Only Calì et al. [6] and Candanedo et al. [30], who include rooms of residential houses, such as kitchen, sleeping room and laundry room in their application, and Dedesko et al. [5] who estimate presence and movements in hospital rooms, make attempts to model non-office buildings.

(a) School (b) Summerhouse

Fig. 1. Pictures of the classroom and the summerhouse.

1.1.3. Motivation

The CO_2 level in a room changes due to ventilation, infiltration, human exhalation and combustion processes. Any adjustment of windows or other ventilation, changes in occupancy or changes in the outdoor CO_2 level change the course of the CO_2 curve in the room. Furthermore, the room CO_2 level is affected by the occupants' metabolic equivalent (MET) which relates to physical activity. A proper description of these factors leads to a white-box model, i.e., a model that attempts to explicitly describe the system's internal structure, in this case including a CO_2 mass balance equation. Such a model is rather complex. Its estimation requires an extensive prior knowledge of the room parameters, and continuous measurements of the above mentioned quantities. Further, a white-box model needs to be adjusted every time it is applied to a new environment. The here presented model, on the other hand, attempts to find states in the CO_2 dynamics by a statistical black-box model which requires no prior knowledge. The model seeks to find the number of states which is most suitable for the description of the observed data, and estimates an autoregressive process for each state. As such, the states are characterised by their mean value and their correlation to preceding values in the process. Further, the model estimates time-dependent transition probabilities between the states processes. This leads to a suitable mathematical representation of the CO_2 dynamics in the room. Nevertheless, this approach comes with the cost of having no direct physical interpretation of these estimated states. An a posteriori interpretation is possible and was done for the data in this work. However, we argue that a human interpretation is not necessarily required if the model is used e.g. to aid an automated control system of the ventilation.

In the present work, we analysed CO_2 trajectories of a school classroom and of a summerhouse, representing two environments very far from each other on the spectrum of predictability. In contrast to the reviewed works, the estimated quantity in this work is not occupancy (binary or number of occupants), but the activity level as defined above, i.e., the rate of air quality change by the actions of all occupants as well as by ventilation and infiltration. Within a HMM framework we identify the number of distinguishable states needed to describe the variations in the activity level, and then estimate the state for each time point. Since the transition between states is modelled by an inhomogeneous Markov chain, the model further allows for realistic simulation of the activity states and corresponding CO_2 levels.

The present work is inspired by the work of Bitsch [32]. For a more comprehensive overview of the literature, the reader is referred to the work of Shen et al. [8]. The authors review and classify approaches of implicit occupancy detection, with a focus on approaches that make use of existing or potentially available data streams that are related to occupancy.

1.2. Data description

This section provides a brief description of the two datasets used in this work. Fig. 1 shows photographs of the respective rooms.

1.2.1. Summerhouse

One of the datasets analysed in this paper stems from a summerhouse on a Danish island. The dwelling has a floor area of 89 m^2 and is used for vacation only. It is occupied by different occupants throughout the year. The sensor used in this work is installed in a room used as living room. We analysed CO_2 data from 1st to 31st August 2015 on a five-minute scale. A two day sample of the data is shown in Fig. 2(b).

1.2.2. School

Another dataset was measured in a classroom of a Danish primary school near Copenhagen. Usually, six pupils but up to twelve pupils and two teachers occupy the room in a period from 8 a.m. to 2 p.m or 3 p.m. The low number of pupils occurs since the room is used for children that have difficulties to follow the usual classes. The classroom has a floor area of 41 m^2, is located in the first upper floor and has six operable windows, three westwards and three eastwards. Mechanical ventilation is carried out every morning before school starts, and again after school ends. During teaching hours, the ventilation would be too loud. According to the teachers, the classroom windows are opened for intermittent ventilation if the teachers feel that the air is stuffy. The sensor is placed at the north side of the room at a height of about 1.50 m. The data reaches from 18th of January to 7th of February 2017 on a five-minute scale. A two day sample of the data is shown in Fig. 2(a).

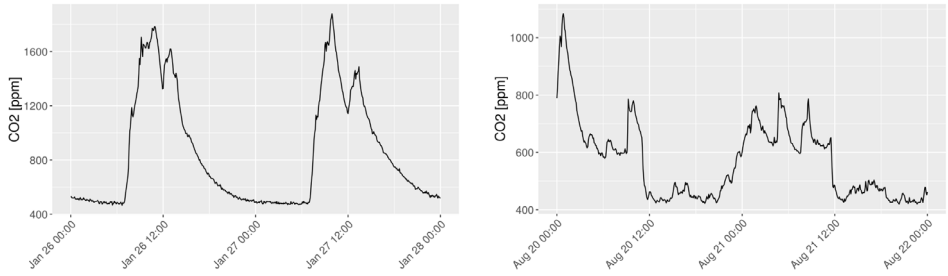
2. Methods

This section gives a brief overview of the employed methods in this work. In Section 2.1 and Section 2.2, ordinary and hidden Markov chains are defined, respectively. A brief definition of autoregressive linear time series model is given in Section 2.3, which enables us to subsequently define Markov-switching models in Section 2.4. In Section 2.5, it is described how the model parameters were estimated using maximum likelihood estimation, for the homogeneous and the inhomogeneous case.

2.1. Markov chains

For a finite set M with $|M| = m$, an m-state Markov chain is a sequence of M-valued random variables $\{X_t\}$ with

$$P(X_t|\mathbf{X}^{(t-1)}) = P(X_t|X_{t-1}), \tag{2}$$



(a) School data (b) Summerhouse data

Fig. 2. Global decoding of the 2 state homogeneous HMM. States represented by colours and by step function.

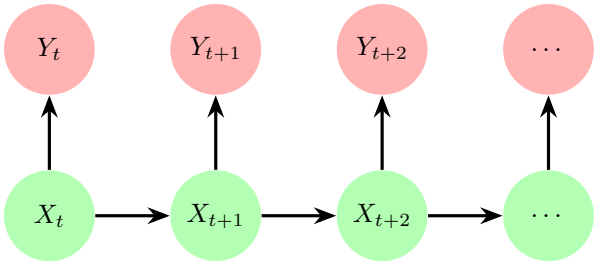


Fig. 3. HMM flow chart.

where $\mathbf{X}^{(t-1)} = X_{t-1}, X_{t-2}, \ldots, X_0$ is set of variables of $\{X_t\}$ up to time $t-1$. Eq. (2) is referred to as the Markov property. In words, the probability of the current time step, conditioned on the entire history, depends only on the previous time step. For $i, j \in M$ the conditional probability

$$P(X_t = i | X_{t-1} = j) = \gamma_{i,j}(t) \tag{3}$$

is called transition probability from state j to state i (at time t). The matrix $\Gamma_t = \{\gamma_{i,j}(t)\}$ is called transition probability matrix (TPM). If Γ does not depend on the time t, the process is called a homogeneous Markov chain. Otherwise it is called an inhomogeneous Markov chain.

2.2. Hidden Markov chains

A hidden Markov chain is a probabilistic model that consists of two components: An unobserved Markov chain $\{X_t\}$ and a state-dependent process of observed values $\{Y_t\}$. We assume that Y_t only depends on the current state X_t but not on its own history $\mathbf{Y}^{(t-1)}$.

$$P(X_t | \mathbf{X}^{(t-1)}) = P(X_t | X_{t-1}) \tag{4}$$

$$P(Y_t | X_t, \mathbf{Y}^{(t-1)}, \mathbf{X}^{(t-1)}) = P(Y_t | X_t) \tag{5}$$

The distribution of $Y_t|X_t$ is called *state-dependent distribution* (SDD). In case of a Gaussian SDD, the hidden Markov chain can be expressed by

$$P(X_t = i \mid X_{t-1} = j) = (\Gamma_t)_{i,j} \tag{6}$$

$$Y_t = \mu_{X_t} + \varepsilon_{X_t,t} \tag{7}$$

where μ_{X_t} is the mean of the process in state X_t and $\varepsilon_{X_t,t}$ is Gaussian white noise, i.e. a mutually uncorrelated sequence of identically distributed normal random variables with zero mean, and with a state-dependent variance $\sigma^2_{X_t}$. The structure of a HMM is best described in a flow chart as in Fig. 3. The horizontal arrows describe the Markov chain that links the states over time. In every time step the value only depends on the previous state. The vertical arrows represent the state-dependent distribution. The current observation Y_t only depends on the current state X_t. The parameters in the model are the transition probabilities as well as the state means and variances, μ_{X_t} and $\sigma^2_{X_t}$, respectively. These parameters can be estimated using maximum likelihood estimation. Two possibilities for the estimation are given by the expectation-maximisation (EM) algorithm ([33,34]), and by direct numerical maximisation of the likelihood. In the present work, the latter approach is employed.

In the context of HMM, one is usually interested in deriving information about the unobserved states $\mathbf{X}^{(T)}$ given a sequence of observations $\mathbf{Y}^{(T)}$. Estimating the most likely state sequence given the sequence of observations is referred to as *global decoding*. It is given by

$$\arg\max_{\mathbf{x}^{(T)}} P\left(\mathbf{X}^{(T)} = \mathbf{x}^{(T)} | \mathbf{Y}^{(T)} = \mathbf{y}^{(T)}\right) \tag{8}$$

An efficient way to calculate the global decoding is called Viterbi algorithm. The core of this algorithm lies in the following observation. Let $(\hat{x}_1, \cdots, \hat{x}_T)$ denote the most likely state sequence for the observations $(y_1, \cdots, y_T)$. Then, given $\hat{x}_t = k$, the preceding state $\hat{x}_{t-1}$ can be found as that state that maximises the following expression:

$$\hat{x}_{t-1} = \arg\max_{m \in M} P(y_t | x_t = k) \cdot (\Gamma_t)_{m,k} \cdot P(x_{t-1} = m) \tag{9}$$

Hence, it is the state that maximises the product of the probability of the observation y_t being in this state, and the probability of transition to this state. For a description of the complete algorithm and a more extensive introduction to Markov chains and hidden Markov chains please refer to [33].

2.3. Autoregressive models

An autoregressive process of order p (AR(p)) can be expressed by

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t, \tag{10}$$

where c, $\phi_1, \ldots, \phi_p$ are constant parameters and ϵ_t is Gaussian white noise. The model suggests that the current value can be expressed as a linear combination of the previous p time steps up to white noise. For further reading see [35].

2.4. Markov-Switching models

A Markov-switching (also regime-switching) model is a generalization both of Markov models and AR(p) processes. It can be seen as an autoregressive model with a state-dependent mean and variance where the states follow a Markov process. An AR(1) Markov-switching process is given by

$$P(X_t = i \mid X_{t-1} = j) = (\Gamma_t)_{i,j} \tag{11}$$

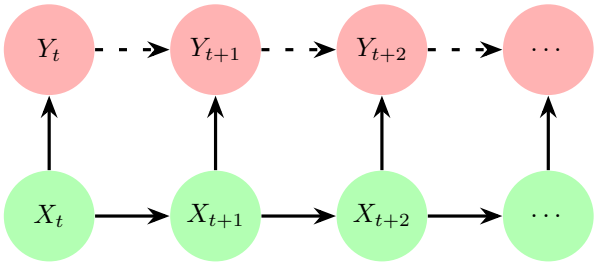


Fig. 4. Markov Switching flow chart.

$$Y_t = \phi_{X_t} Y_{t-1} + c_{X_t} + \varepsilon_{t,X_t}, \tag{12}$$

where c_{X_t} and ϕ_{X_t} are state-dependent parameters and ε_{t,X_t} is Gaussian white noise with state-dependent variance $\sigma^2_{X_t}$. The model can be interpreted as a HMM with mean values $\mu_{X_t} = \phi_{X_t} Y_{t-1} + c_{X_t}$. The structure of the Markov switching model can be seen in Fig. 4. Notice the additional dashed horizontal arrows between the observations (compare Fig. 3). They represent the dependency on the previous observation value in the model.

2.5. Parameter estimation

Let Γ be the transition probability matrix and $\mathbf{A}$ be the set of parameters associated to the state-dependent distribution. Parameter estimation is carried out by numerical maximisation of the log-likelihood function, i.e. the maximisation of the logarithm of the joint distribution of all observations with respect to the model parameters.

$$\log P(\mathbf{Y}^{(T)} |\ \Gamma, \mathbf{A}) = \log \sum_{\mathbf{X}^{(T)}} P(\mathbf{Y}^{(T)} |\ \mathbf{X}^{(T)}, \mathbf{A}) \cdot P(\mathbf{X}^{(T)} |\ \Gamma) \tag{13}$$

The numerical optimisation bears some challenges which are addressed in the following paragraphs.

2.5.1. Homogeneous Markov

The maximum likelihood estimation is carried out using the unconstrained optimiser *nlm()* in R, though the model parameter underlie certain constraints. The entries of the TPM require

$$0 \le \gamma_{i,j,t} \le 1, \tag{14}$$

$$\sum_{j=1}^{m} \gamma_{i,j,t} = 1, \quad 1 \le i \le m. \tag{15}$$

Therefore, a parameter transformation between constrained and unconstrained domain is necessary. Non-negativity is asserted by applying an increasing, non-negative function to the parameters. To account for constraint (14), we used the function expit(x)= $\exp(x)/(1+\exp(x))$ with range (0, 1). It additionally prevents numerical overflow, since for large values it converges to 1; in contrast of converging to infinity in case of the exponential function. The row sum constraint (15) can be respected by dividing each entry by its row sum. Hence, the transition probabilities can be expressed for $i \neq j$ by

$$\gamma_{i,j,t} = \frac{\operatorname{expit}(\tau_{i,j})}{1 + \sum_{k \neq i}^{m} \operatorname{expit}(\tau_{i,k})} \tag{16}$$

and for $i = j$ by

$$\gamma_{i,i,t} = \frac{1}{1 + \sum_{k \neq i}^{m} \operatorname{expit}(\tau_{i,k})}. \tag{17}$$

This results in the following transition probability matrix Γ,

$$\Gamma = \begin{pmatrix} \frac{1}{1+\sum_{k\neq 1}^{m} \operatorname{expit}(\tau_{1,k})} & \cdots & \frac{\operatorname{expit}(\tau_{1,m})}{1+\sum_{k\neq 1}^{m} \operatorname{expit}(\tau_{1,k})}) \\ \vdots & \ddots & \vdots \\ \frac{\operatorname{expit}(\tau_{m,1})}{1+\sum_{k\neq m}^{m} \operatorname{expit}(\tau_{m,k})} & \cdots & \frac{1}{1+\sum_{k\neq m}^{m} \operatorname{expit}(\tau_{m,k})} \end{pmatrix} \tag{18}$$

with $m(m-1)$ unconstrained optimisation parameters $\{\tau_{i,j}\}_{1 \le i \neq j \le m}$.

2.5.2. Inhomogeneous Markov

One challenge of modelling an inhomogeneous Markov chain is a high number of parameters to be estimated. Estimation of transition probability matrices for every time step $t = 1, \ldots, T$ of the entire available data would result in $Tm(m-1)$ free parameters. However, several assumptions can be made to reduce this number. First, we assume that the transition probabilities do not change between days. Hence, T reduces to just the number of time steps for one day, say N. In the case of a five minute resolution, N equals 288. Furthermore, it is presumed that the transition probabilities do not change abruptly between two time steps but rather describe a continuous function over time, and can thus be interpolated. In this work, cubic periodic B-splines, i.e., piecewise polynomial functions, with four equidistant knots were used to interpolate the transition probabilities on the diagonal of the TPM. The spline basis consists of four column vectors and is denoted by $X = (X_1, \ldots, X_N)^T \in \mathbb{R}^{N \times 4}$. The spline parameters are denoted by $\beta = (\beta_1, \ldots, \beta_m) \in \mathbb{R}^{4 \times m}$. The off-diagonal parameters are estimated according to paragraph 2.5.1 with $m(m-2)$ degrees of freedom. This results in the following TPM.

$$\Gamma_t = \begin{pmatrix} \frac{\operatorname{expit}(X_t\beta_1)}{\operatorname{expit}(X_t\beta_1)+\sum_{i\neq 1}^{m} \operatorname{expit}(\tau_{1,i})} & \cdots & \frac{\operatorname{expit}(\tau_{1,m})}{\operatorname{expit}(X_t\beta_1)+\sum_{i\neq m}^{m} \operatorname{expit}(\tau_{1,i})} \\ \vdots & \ddots & \vdots \\ \frac{\operatorname{expit}(\tau_{m,1})}{\operatorname{expit}(X_t\beta_m)+\sum_{i\neq m}^{m} \operatorname{expit}(\tau_{m,i})} & \cdots & \frac{\operatorname{expit}(X_t\beta_m)}{\operatorname{expit}(X_t\beta_m)+\sum_{i\neq m}^{m} \operatorname{expit}(\tau_{m,i})} \end{pmatrix} \tag{19}$$

The number of parameters to be estimated is $4m + m(m-2)$.

2.5.3. State-dependent distribution

In case of homogeneous and inhomogeneous HMM, the parameters μ_{X_t} and $\sigma^2_{X_t}$, i.e., a number of $2m$ parameters, need to be estimated for the state-dependent distribution. In case of an AR(1) Markov-Switching model, the parameters are c_{X_t}, ϕ_{X_t} and $\sigma^2_{X_t}$, what corresponds to a number of $3m$. To account for the non-negativity of the variance we applied the exponential function

$$\sigma^2_{X_t} = \exp(\xi_{X_t}) > 0\ , \tag{20}$$

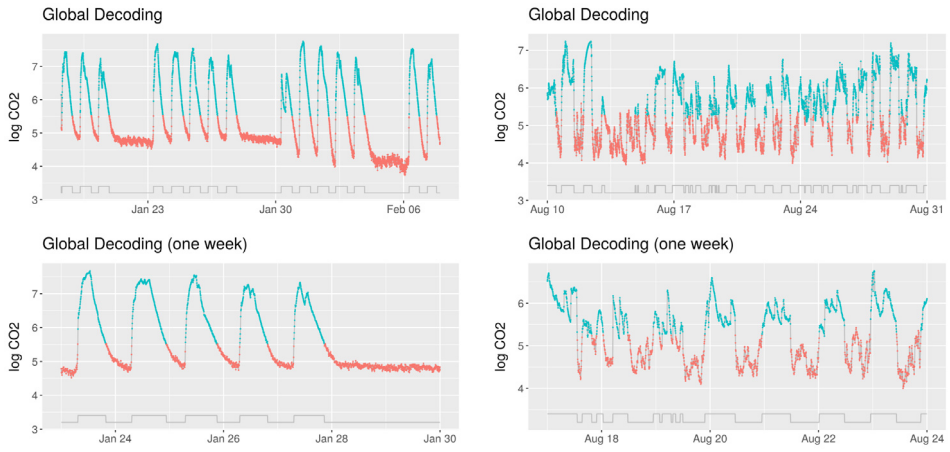
and optimised the unconstrained parameter ξ_{X_t}.

3. Results

In this section, we describe the results of univariate models within the above described framework with observations

$$Y_t = \log(C_t^{\text{in}} - C_t^{\text{out}})\ , \tag{21}$$

where C_t^{in} is the measured indoor CO_2 concentration in parts per million (ppm) at time t, and C_t^{out} represents the outdoor CO_2 concentration. The latter is here assumed to take a constant level of 350 ppm. It is chosen to be constant, since variations in the outdoor CO_2 level are expected to be negligible for the purpose of this model. The value is chosen to be slightly below the minimum level of the recorded indoor CO_2 concentration. The authors are aware that the atmospheric CO_2 level is higher than 350 ppm; it



(a) School data (b) Summerhouse data

Fig. 5. Global decoding of the 2 state homogeneous HMM. States represented by colours and by step function.

was rather a decision made for numerical reasons to set this value as low. The logarithm function is applied to the difference as a standard data transformation of skewed, non-negative distributions to behave more Gaussian, in the sense that the resulting variance will be more constant and symmetric. In the following, we refer to Y_t as CO_2 level, neglecting that we actually mean the transformation in (21).

3.1. Model selection

The selection of the most suitable model among the class of the above-described class of models was carried out by means of several methodologies. As a first step, we looked at the visualization of the global decoding. This way, one can check whether the decoding is in agreement with how human intuition would divide the data into distinct states.

Additionally, the one-step prediction (pseudo) residuals ([33]) were analyzed with respect to their desired attributes: normality and temporal independence. These attributes indicate that the model captures all systematic variation in the data, and only random noise (white noise) is left. In a next step, Akaike's information criterion (AIC) and the Bayes information criterion (BIC) were used to compare models directly with respect to their balance of goodness-of-fit and complexity. Both AIC and BIC are measures that favour a high likelihood while penalizing the number of model parameters. For a more detailed introduction see [36].

3.1.1. Initial model

The objective of model selection is to find the least complex model that describes the variation of the data well. Therefore, we begin the analysis with the simplest of the potential models. This is a 2-state model with a homogeneous Markov chain. Fig. 5 shows the global decoding for this model. The states are indicated by the colours. The same information is given by the step function below the graph. One can see that the decoding separates high and low CO_2 regions fairly well. However, the transition to the "active" state happens at a point where the CO_2 level is relative high and the occupants have assumingly already arrived. The analysis of one-step residuals is given in Fig. 6. The figure shows the residuals over time in the upper left, the residual histogram in the upper right, the plot of sample quantiles against theoretical quantiles (QQ-plot) in the lower left and the autocorrelation function of the residuals in the lower right. The figure reveals weaknesses of the model. The residuals clearly show a non-random pattern over time. The histogram is far from the desired Gaussian bell curve. The QQ-plot deviates significantly. The autocorrelation function shows high dependencies of the residuals to their past values (lags).

3.1.2. Homogeneous vs. inhomogeneous

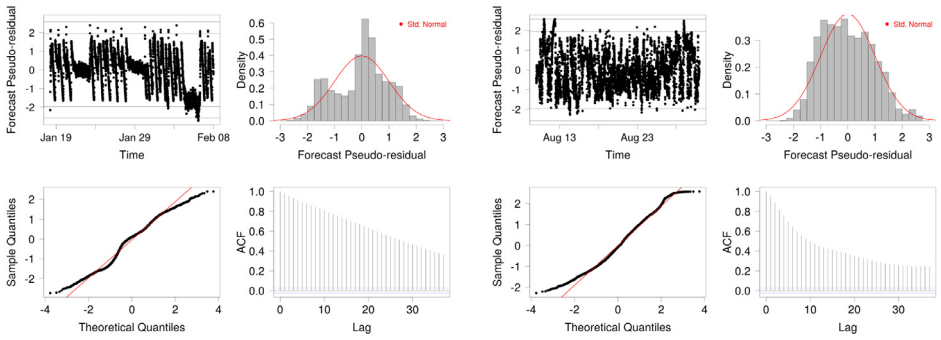
Activity patterns of building occupants are obviously time-dependent. One way to capture this time-dependency is to allow the transition probabilities to vary in time, i.e., to use an inhomogeneous Markov chain for the states. This approach comes with a higher number of parameters. It is therefore sensible to test whether this extra complexity pays off in terms of model performance. Figures of global decoding and residual analysis of a 2-state inhomogeneous model showed very similar results to the homogeneous case in Figs. 5 and 6, respectively, and are therefore omitted. A comparison of AIC and BIC for models with up to 5 states are given in Table 1. Both criteria favour the inhomogeneous model, even though its higher complexity is penalised. For this reason, we restrict the class of models considered in the further analysis to inhomogeneous models.

3.1.3. Hidden Markov vs. Markov Switching

A weakness of the above models described in Section 3.1.2 is that the one-step residuals show heavy temporal dependency. It is apparent that the autocorrelation of the CO_2 values over time is not sufficiently captured. To overcome this, we introduce an additional autoregressive (AR) parameter ϕ_{X_t} as in Eq. (12) that describes the dependence between observations over time. The resulting model structure is called Markov-switching. In this model, the current value of Y_t is normally distributed with autoregressive and state-dependent mean $\phi_{X_t} \cdot y_{t-1} + c_{X_t}$ and state-dependent variance $\sigma^2_{X_t}$. Due to the findings above, the transition probabilities are chosen to be time-dependent. Fig. 7 shows the residual analysis of a 2-state Markov-switching model. It shows a clear improvement over the models without AR parameter, described in Section 3.1.2. The residuals show less temporal dependence and are evidently normally distributed. From Table 1 it is apparent that the Markov-switching model outperforms the ordinary HMM in terms of the information criteria.

3.1.4. Number of states

One important tuning parameter in the model is the number of hidden states m. This parameter is not directly optimised by

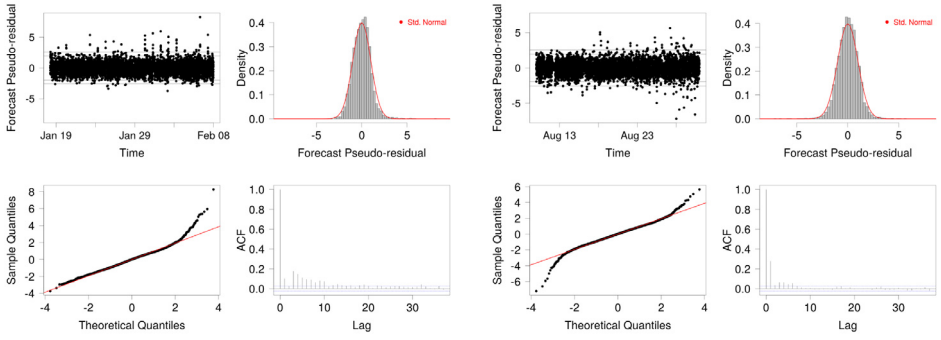


(a) School data (b) Summerhouse data

Fig. 6. Residual analysis of the 2 state homogeneous HMM.

Table 1
Comparison of AIC and BIC (school data).

	parameters	par. (tpm)	par. (sdd)	AIC	BIC
m=2 states					
Homogeneous HMM	6	$m(m-1)$	$2m$	7792	7832
Inhomogeneous HMM	12	$m(m-2)+4m$	$2m$	7627	7721
Markov-Switching	14	$m(m-2)+4m$	$3m$	-19187	-19080
m=3 states					
Homogeneous HMM	12	$m(m-1)$	$2m$	3236	3316
Inhomogeneous HMM	21	$m(m-2)+4m$	$2m$	3118	3279
Markov-Switching	24	$m(m-2)+4m$	$3m$	-20326	-20146
m=4 states					
Homogeneous HMM	20	$m(m-1)$	$2m$	-690	-556
Inhomogeneous HMM	32	$m(m-2)+4m$	$2m$	-827	-586
Markov-Switching	36	$m(m-2)+4m$	$3m$	-20770	-20503
m=5 states					
Homogeneous HMM	30	$m(m-1)$	$2m$	-2963	-2762
Inhomogeneous HMM	45	$m(m-2)+4m$	$2m$	-3101	-2766
Markov-Switching	50	$m(m-2)+4m$	$3m$	-21392	-21023



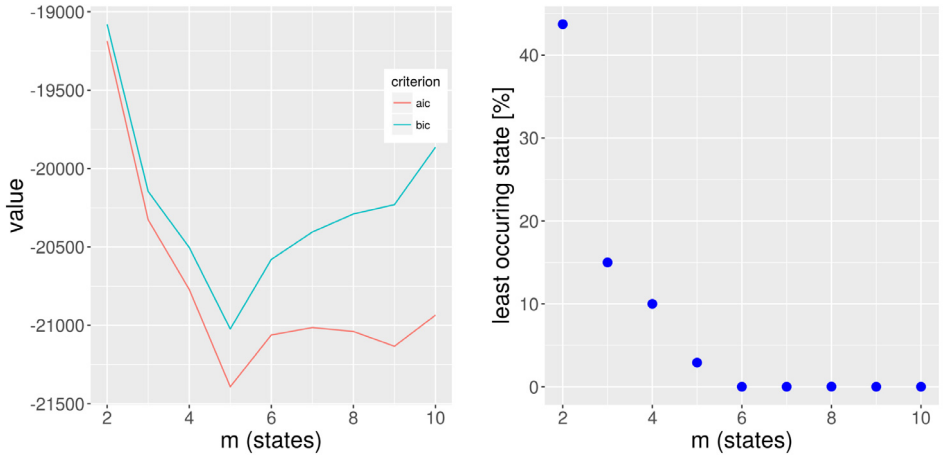
(a) School data (b) Summerhouse data

Fig. 7. Residual analysis of the 2 state Markov switching model.

the maximum likelihood estimation. However, models with different values of m can be compared based on their maximised likelihood. Hence, AIC and BIC can be used. Fig. 8(a) shows the AIC and BIC for the models with 2 to 10 states. Both criteria favour the model with $m = 5$ as their value is minimal here. Another measure for the justification of adding additional states is the actual usage of all available states in the model. If one of the states is barely applied in the global decoding, this might indicate that this state is obsolete. Fig. 8(b) shows the percentage of usage for the state that occurs least often in the decoding. For $m \geq 6$ this value is very close to 0 which indicates that there are more states in the model than useful. Furthermore, the analysis of residuals showed that the model does not improve significantly for values greater than five.

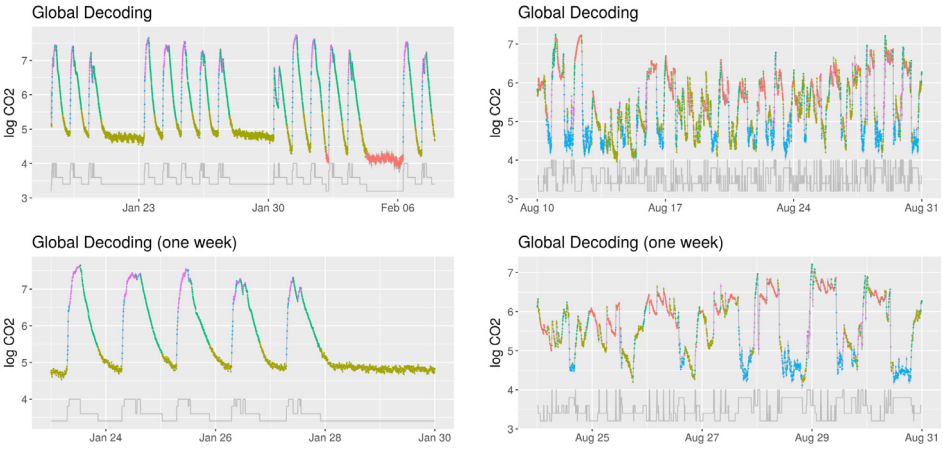
3.2. Final model

Global decoding and residual analysis of a 5-state inhomogeneous Markov-Switching model are given in Figs. 9 and 10, respectively. This model is referred to as the final model in the following. From Fig. 10, it is evident that, for both datasets, the residuals behave close to what is desired, i.e., independently and identically Gaussian distributed. The autocorrelation function shows indeed significant correlations in the first two time lags which might indicate that a moving average (MA) component in the model would describe the data even more accurate. Nevertheless, for the sake of parsimony, the final model is here seen to be accurate enough. The estimates of this model for the stationary mean μ, the autore-



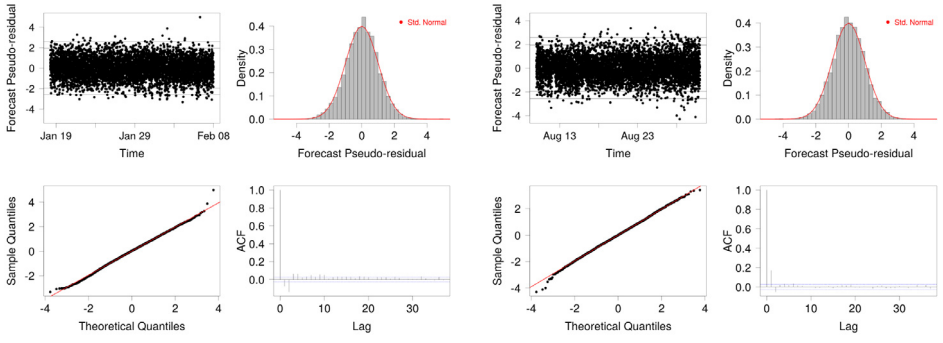
(a) Information criteria (b) Least occurring state

Fig. 8. Selection of the number of states (school data).



(a) School data (b) Summerhouse data

Fig. 9. Global decoding of the 5 state inhomogeneous Markov switching model. States represented by colours and by step function.



(a) School data (b) Summerhouse data

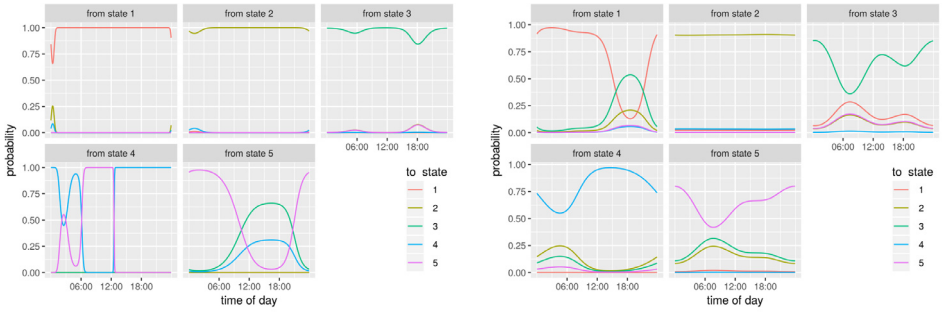
Fig. 10. Residual analysis of the 5 state inhomogeneous Markov switching model.

gressive parameter ϕ and the standard deviation σ are given in Table 2.

Fig. 11 shows the transition probabilities of the final model as a function of time of day for school (left) and summerhouse (right). In the case of the school, the probabilities are based only on data from Monday to Friday, because no transitions are expected on weekends. For the summerhouse, all seven days of a week were taken into account.

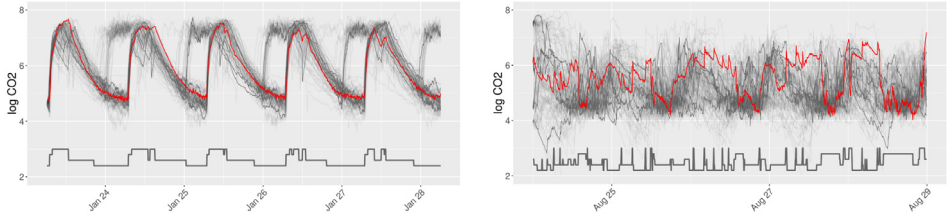
Table 2
Estimates of the final model.

	School data			Summerhouse data		
State i	μ_i	ϕ_i	σ_i	μ_i	ϕ_i	σ_i
1	4.131	0.6554	0.08173	7.569	1.0052	0.02241
2	4.754	0.9599	0.05427	6.172	0.8839	0.26615
3	4.200	0.9909	0.01862	4.512	0.8276	0.11633
4	7.290	0.8824	0.08784	7.524	0.9629	0.05027
5	7.562	0.9459	0.02370	3.983	0.9839	0.05534



(a) School data (b) Summerhouse data

Fig. 11. Transition probabilities of the final model.



(a) School data (b) Summerhouse data

Fig. 12. 100 simulations (grey) with measured values (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3. Simulation

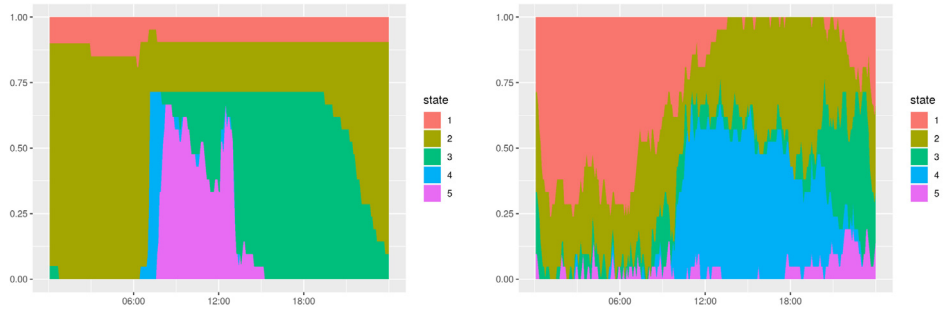
Fig. 12 shows simulations of the log CO_2 levels. The results of one hundred simulation runs of five week days are plotted. The red line indicates the measured log CO_2 values for this period. As expected, the simulations of the school data show much more similarity, to each other and to the measurements, than in the case of the summerhouse data. The simulations of the summerhouse look more chaotic. However, it is apparent that the simulated variation is of same magnitude as in the measurements, indicating a plausible range of values. The bottom part of the two graphs shows the median of the one hundred accompanying activity state simulations.

3.4. Interpretation of states

Occupancy has two rather obvious states, present and absent. Even though it is not the primary objective of this work to estimate binary occupancy, the interpretation attempt shown in Table 3 is possible from the global decoding. We can use this interpretation of activity states to evaluate the model with respect to its ability to distinguish occupied and vacant intervals. Ground truth occupancy was not recorded for either of the data sets. However, the average acoustic noise level between two time steps was measured. This can be an indicator for presence if the environment is quiet in periods of absence. The acoustic noise levels above a reference measurement level of 32 dBA were compared to the activity states which were associated to presence. These are states four and five for the school data, and states three, four and five for the summerhouse data. Fig. 14 shows this comparison for one week. The graphic shows the acoustic noise level (black) along with a comparison of the binary occurrence of acoustic noise (dashed blue) with the model's "presence estimates" (red). Here, "presence estimate" refers to the above-mentioned activity states that are assigned to presence in Table 3. By Considering the noise signal the "true" value, the discrimination indicators *true-positive rate, true-negative rate* and *accuracy* are obtained. For building control purposes, false negatives (wrongly concluded vacancy) and false positives (wrongly concluded occupancy) may have implications of very different nature, as acknowledged in [8]. False negatives are usually more crucial as they may lead to discomfort or annoyance of the occupants, whereas false positives lead to additional energy

Table 3
Interpretation of states.

State i	School data	Summerhouse data
State 1	long absence	Sleep
State 2	absence (night/weekend)	Absence
State 3	absence (break/end of day)	Presence
State 4	presence (after arrival)	High volatility
State 5	presence	Presence



(a) School data (b) Summerhouse data

Fig. 13. Time spent in each state during the course of day.

Table 4
Classification.

	School data	Summerhouse data
True-positive rate	99.4%	77.0%
True-negative rate	96.5%	74.6%
Accuracy	96.5%	75.5%

consumption. Table 4 shows the obtained true-positive rate, true-negative rate and accuracy. The close agreement of these two indicators suggest that, in this case, both are good estimators for the occupancy state.

In the summerhouse data, the interpretation of states is more difficult for several reasons. In the recorded period (August), a high degree of natural ventilation is expected through doors and windows. The occupants spend their vacation at this house and may therefore have no regular schedule. From the data we can see that, at night, high CO_2 levels coincide with silence, suggesting that the occupants sleep in the summerhouse. During the day, the values are volatile. An explanation for this can be that the occupants spend the day both inside and outside house, or that windows are kept open. A state interpretation can be found in Table 3.

Fig. 13 shows the fraction of time spent in each state as a function of time of day. For the summerhouse it is apparent that state 1, labelled as 'sleep', occurs predominantly from around midnight to 9 a.m. From 10 a.m. to 4 p.m. state 4 of 'high volatility' is the most dominant. State 3, 'presence', occurs mainly in the evening hours before midnight. States 2 and 5 do not show a strong correlation to the time of day. A comparison of noise and active presence (not sleeping) leads to the results in Table 4. As expected, the assessment criteria show lower values as they do for the school data.

3.5. Sensitivity to sensor location

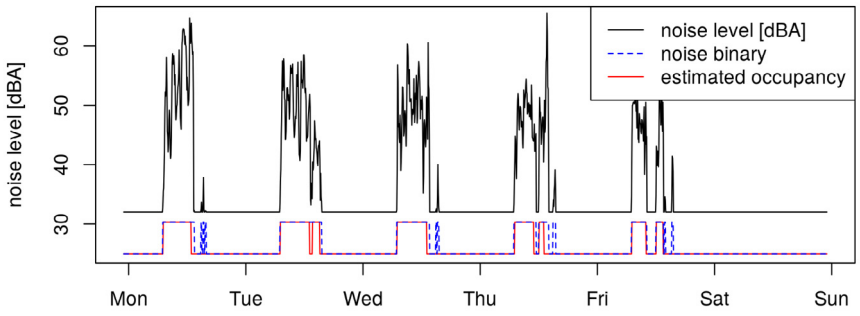
This section is dedicated to the analysis of how sensitive the state estimations are to changes in the CO_2 measurements due to different sensor locations. For this analysis, the CO_2 level was measured by two sensors in a mechanically ventilated office room with a floor area of 42 square meters. The sensors were located in opposing locations in the room. During the measurements, the room was occupied by three persons. Fig. 15 shows the state estimation of a 2-state and a 3-state Markov Switching model. The ventilation system was running from 6:00 to 20:00 on workdays. These periods are indicated in the figure by the grey bars. In the bottom of the graph, the matching of the states is indicated in green. Overall, the state estimations of the two sensor locations match 77.3% of the time in case of the 2-state model compared to an average matching of 50% by pure chance. For the 3-state model the states match 60.9% of the time compared to 33.3% pure chance probability, and for the 4-state model the matching is 53.3% compared to 25% at pure chance. The results show that the state estimates of different measurement locations coincides less for higher number of states. Though, the difference to the matching probability at pure chance is surprisingly constant with 27.3, 27.6 and 28.3 percentage points for the 2-state, 3-state and 4-state model, respectively. From the graph it is apparent that the matching is more accurate in periods of relatively few state changes.

4. Discussion

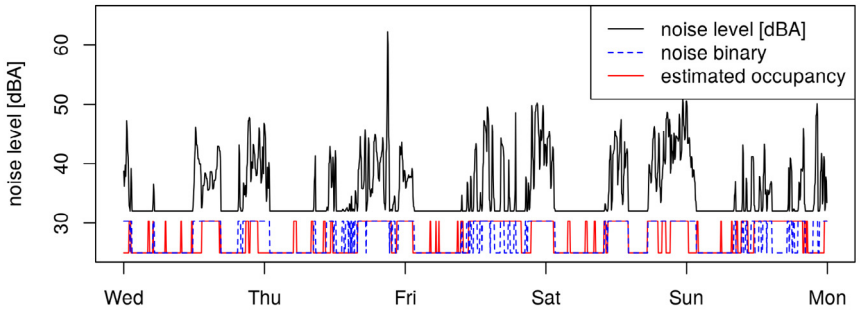
The two chosen datasets show naturally very different occupant activity patterns, and with this different CO_2 patterns. Due to the scheduled routine in the school, the CO_2 curve is much smoother, and shows a clearer periodicity, than in the summerhouse, in which the daily routine is much more arbitrary. To this end, the two environments represent two extremes of occupant activity. It is therefore encouraging that the model performs equally well in terms of one-step predictions for both sets. We follow that the presented methodology can be applied independently from the type of building and building usage.

Choosing an inhomogeneous Markov chain for the underlying state changes, i.e., allowing the transition probabilities to depend on time, showed an improvement in terms of the model performance. This agrees with the intuition that occupants' activities depend on the time of day. Furthermore, the inclusion of an autoregressive parameter improved the model significantly compared to the common HMM. The need for this parameter is fairly intuitive since the CO_2 level is clearly autocorrelated. From the autocorrelation function of the final model (Fig. 10) one could argue that an additional moving average parameter would improve the model fit. This was, however, neglected in order to keep the complexity relatively low. Furthermore, addressing the data's periodicity by including a seasonal autoregressive term of e.g. 24 hours can be a sensible modelling choice. We did not consider this in the present work, since seasonality is already taken into account by the periodic transition probabilities, and additional parameters might needlessly increase the model's complexity.

One has to keep in mind that the model learns unsupervised. It finds patterns in the data without the information of an outcome variable. This leads to states that might not be human interpretable. However, one can argue that a human interpretation may be not necessary if the model is used as input for a control algorithm. Indoor CO_2 level proved itself a good predictor for building occupant activity. It needs to be investigated whether the inclusion of other input variables can improve the presented model without making it overly complex. Additional inputs can be e.g. temperature, relative humidity or the noise level in the room.

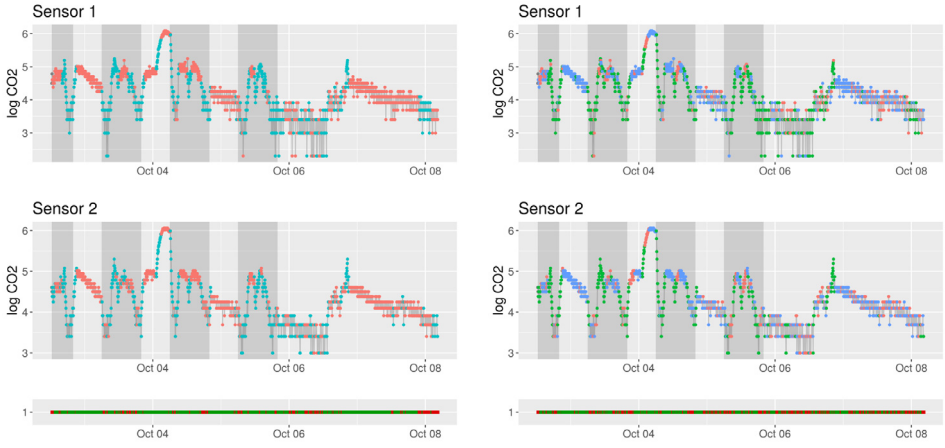


(a) School data



(b) Summerhouse data

Fig. 14. Comparison of states and noise level.



(a) 2 state model

(b) 3 state model

Fig. 15. Comparison of states for different sensor locations.

In order to capture substantial changes of occupancy and occupant activity pattern, caused by e.g. change of occupants or the use of building, the model should be able to adapt to these changes by re-estimating its parameters with new observations. It can be therefore considered to develop an adaptive version of the presented model in future research.

5. Conclusion

A model for occupant activity estimation, simulation and short-term forecasting in buildings was presented. The model's only inputs are CO_2 level and time of day. The applied methodology is a hidden Markov framework. The model's performance showed a significant improvement when using an inhomogeneous Markov chain and including the autocorrelation between CO_2 observations, which resulted in a Markov-switching model. The model performed best for a state number of five. It was applied to two different datasets, a classroom and a Danish summerhouse. Transition probabilities and state-dependent parameters were estimated. A global decoding and 5-day simulations were carried out. Evaluation was carried out by means of one-step prediction residual analysis as well as by AIC and BIC. The results were promising in both cases.

This paper suggests to link the room CO_2 levels to occupant activity level; this link gives a good description of the dynamics and variability of the room CO_2, which in turn can be used for bet-

ter predictions and simulations. The here presented model can be used for estimation of the occupant activity level, for predictive control of the indoor air as well as input for building simulations. Further development of, for instance, a model with additional input variables or an on-line parameter estimation is left to future research.

Acknowledgements

This work has partly been carried out as a part of the CITIES, FME ZEN, and SCA projects. The authors are thankful to Innovation Fund Denmark, EU Interreg V programme, and the Norwegian Research Council and ZEN partners for the support.

References

[1] F. Oldewurtel, D. Sturzenegger, M. Morari, Importance of occupancy information for building climate control, Appl. Energy 101 (2013) 521–532.
[2] F. Oldewurtel, A. Parisio, C.N. Jones, D. Gyalistras, M. Gwerder, V. Strauch, B. Lehmann, M. Morari, Importance of occupancy information for building climate control, Energy Build. 45 (2012) 15–27.
[3] A. Persily, L. de Jonge, Carbon dioxide generation rates for building occupants, Indoor Air 27 (2017) 868–879.
[4] M.W. Qi, X.F. Li, L.B. Weschler, J. Sundell, co_2 Generation rate in chinese people, Indoor Air 24 (2014) 559–566.
[5] S. Dedesko, B. Stephens, J.A. Gilbert, A. Siegel Jeffrey, Methods to assess human occupancy and occupant activity in hospital patient rooms, Build Environ. 90 (2015) 136–145.
[6] D. Calì, P. Matthes, K. Huchtemann, R. Streblow, D. Müller, Co2 based occupancy detection algorithm: experimental analysis and validation for office and residential buildings, Build Environ. 86 (2015) 39–49.
[7] B. Gunay, A. Fuller, W. O'Brien, I. Beausoleil-Morrison, Detecting occupants' presence in office spaces: A case study, in: Proceedings of eSim 2016, Hamilton, ON, Canada, 2016.
[8] W. Shen, G. Newsham, B. Gunay, Leveraging existing occupancy-related data for optimal control of commercial office buildings: a review, Adv. Eng. Inf. 33 (2017) 230–242.
[9] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017.
[10] J. Page, D. Robinson, N. Morel, J.-L. Scartezzini, A generalized stochastic model for the simulation of occupant presence, Energy Build. 40 (2008) 83–98.
[11] I. Richardson, M. Thomson, D. Infield, A high-resolution domestic building occupancy model for energy demand simulations, Energy Build. 40 (2008) 1560–1566.
[12] P.D. Andersen, A. Iversen, H. Madsen, C. Rode, Dynamic modeling of presence of occupants using inhomogeneous markov chains, Energy Build 69 (2014) 213–223.
[13] S. Wang, X. Jin, Co2-based occupancy detection for on-line outdoor air flow control, Indoor Built. Environ. 7 (1998) 165–181.
[14] S. Wang, J. Burnett, H. Chong, Experimental validation of co2-based occupancy detection for demand-controlled ventilation., Indoor Built. Environ. 8 (1999) 377–391.
[15] D. Wang, C. Federspiel Clifford, F. Rubenstein, Modeling occupancy in single person offices, Energy Build. 37 (2005) 121–126.
[16] W.-K. Chang, T. Hong, Statistical analysis and modeling of occupancy patterns in open-plan offices using measured lighting-switch data, Build. Simul. 6 (2013) 23–32.
[17] C. Liao, Y. Lin, P. Barooah, Agent-based and graphical modelling of building occupancy, J. Build. Perform. Simul. 5 (2012) 5–25.
[18] A. Mahdavi, F. Tahmasebi, Predicting people's presence in buildings: an empirically based model performance analysis, Energy Build. 86 (2015) 349–355.
[19] J.R. Dobbs, B.M. Hencey, Model predictive hvac control with online occupancy model, Energy Build. 82 (2014) 675–684.
[20] M. Gruber, A. Trüschel, J.-O. Dahlenbäck, Co2 sensors for occupancy estimations: potential in building automation applications, Energy Build. 84 (2014) 548–556.
[21] F. Wang, Q. Feng, Z. Chen, Q. Zhao, Z. Cheng, Z. Jianhong, Y. Zhang, M. Jinbo, Y. Li, H. Reeve, Predictive control of indoor environment using occupant number detected by video data and co2 concentration, Energy Build. 145 (2017) 155–162.
[22] 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, in: SimAUD '11, Society for Computer Simulation International, San Diego, CA, USA, 2011, pp. 141–148.
[23] Y. Benezeth, H. Laurant, B. Emilie, C. Rosenberger, Towards a sensor for detecting human presence and characterizing activity, Energy Build. 43 (2011) 305–314.
[24] R. Melfi, B. Rosenblum, B. Nordman, K. Christensen, Measuring building occupancy using existing network infrastructure, in: Proceedings of Green Computing Conference and Workshops (IGCC), 2011, 2011, pp. 1–8.
[25] A. Khan, J. Nicholson, S. Mellor, D. Jackson, K. Ladha, Occupancy monitoring using environmental & context sensors and a hierarchical analysis framework, in: Proceedings of ACM International Conference on Embedded Systems For Energy-Efficient Buildings, Memphis, USA 2014, 2014, pp. 90–99.
[26] G.R. Newsham, H. Xue, C. Arsenault, J.J. Valdes, G.J. Burns, E. Scarlett, S.G. Kruithof, W. Shen, Testing the accuracy of low-cost data streams for determining single-person office occupancy and their use for energy reduction of building services, Energy Build. 135 (2017) 137–147.
[27] 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 (2010) 1038–1046.
[28] B. Dong, K.P. Lam, Building energy and comfort management through occupant behaviour pattern detection based on a large-scale environmental sensor network, J. Build. Perform. Simul. 4 (2011) 359–369.
[29] J. Liisberg, J. Møller, H. Bloem, J. Cipriano, G. Mor, H. Madsen, Hidden markov models for indirect classification of occupant behaviour, Sustain. Citie. Soc. 27 (2016) 83–98.
[30] L.M. Candanedo, V. Feldheim, D. Deramaix, A methodology based on hidden markov models for occupancy detection and a case study in a low energy residential building, Energy Build. 148 (2017) 327–341.
[31] B. Ai, Z. Fan, R.X. Goa, Occupancy estimation for smart buildings by an auto-regressive hidden markov model, in: American Control Conference 2014, 2014. Portland, Oregan, USA.
[32] M.A. Bitsch, Statistical Learning for Energy Informatics, Technical University of Denmark, DTU, Lyngby, Denmark, 2016 Master's thesis.
[33] W. Zucchini, I. McDonald, R. Langrock, Hidden Markov Models for Time Series - Second edition, CRC Press, Boca Raton, Florida, 2016.
[34] J.D. Hamilton, Time Series Analysis, Princeton University Press, Princeton, New Jersey, 1994.
[35] H. Madsen, Time Series Analysis, Chapman and Hall, New York, 2008.
[36] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning; Data Mining, Inference and Prediction, Springer, New York, 2001.