Energy and Buildings 69 (2014) 213–223

Contents lists available at ScienceDirect

Energy and Buildings

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

Dynamic modeling of presence of occupants using inhomogeneous Markov chains

Philip Delff Andersen[a,*], Anne Iversen[b], Henrik Madsen[c], Carsten Rode[d]

[a] Informatics and Mathematical Modelling, Richard Pedersens Plads, Technical University of Denmark, Building 321, DK-2800 Kgs. Lyngby, Denmark
[b] Danish Building Research Institute, Dr. Neergaards Vej 15, DK-2970 Hørsholm, Denmark
[c] Informatics and Mathematical Modelling, Richard Pedersens Plads, Technical University of Denmark, Building 321, DK-2800 Kgs. Lyngby, Denmark
[d] Brovej, Bygning 118, Technical University of Denmark, Building 321, DK-2800 Kgs. Lyngby, Denmark

ARTICLE INFO

Article history:
Received 27 November 2012
Accepted 1 October 2013

ABSTRACT

Occupancy modeling is a necessary step towards reliable simulation of energy consumption in buildings. This paper outlines a method for fitting recordings of presence of occupants and simulation of single-person to multiple-persons office environments. The method includes modeling of dependence on time of day, and by use of a filter of the observations it is able to capture per-employee sequence dynamics. Simulations using this method are compared with simulations using homogeneous Markov chains and show far better ability to reproduce key properties of the data.

The method is based on inhomogeneous Markov chains with where the transition probabilities are estimated using generalized linear models with polynomials, B-splines, and a filter of passed observations as inputs. For treating the dispersion of the data series, a hierarchical model structure is used where one model is for low presence rate, and another is for high presence rate.

1. Introduction

Occupants interact with the indoor environment through heat and carbon dioxide emission, switching lights on/off, opening windows, etc. Occupancy profiles are therefore a necessary input to building simulation models that include indoor environment variables, ventilation loads, electric power consumption, etc. The most common way of considering occupancy in simulation tools is by using, and if necessary repeating, one static occupancy profile [1,2]. Typically, the used profile is constant for weekdays and weekends, respectively. However, occupants do not arrive in buildings or leave buildings at fixed times. A study by [3] reported that on average the offices were occupied 46% of the time, which is supported by the study by [4] where it was found that only half of the work day was spent at the work station. Therefore building systems controlled by occupant presence have shown great energy saving potential in office buildings.

Recently, occupants presence models have been developed by [5,4,6–8]. These models include behavior of occupants based on empirical data. Refs. [6] and [9] both developed occupant presence models as a first order Markov chain. Wang's data fits well with the exponential distribution when observing individual offices and vacant intervals. However the exponential model was not validated for occupied intervals. Ref. [4] considered occupant presence as an inhomogeneous Markov chain interrupted by occasional periods of long absence. By using a profile of probability of presence as input to a Markov chain they were able to reproduce intermediate periods of presence and absence distributed exponentially with a time-dependent coefficient as well as fluctuations of arrivals, departures and typical breaks. They defined a parameter called the "parameter of mobility". This parameter indicates how much people move in an out of the zone, by correlating the tendency of coming to work with the tendency of leaving.

Ref. [5] looked at occupancy based on more detailed prior knowledge about time consumption on different tasks in a working day. As input to their model they included information on the intermediate activities of the occupants such as 'receiving unexpected visitor', 'walking to printer', and 'having lunch'. They were able to simulate occupancy patterns using a probabilistic method for different intermediate activities. The model by [5] is a step towards a more behavioral approach to simulating occupancy.

The focus of the study presented in this paper is to develop a model for presence of occupants for simulation of single person presence sequences in an office environment. The study seeks answers to the following questions:

- How can the dependence of the tendency of being present on the time of day be modeled?

* Corresponding author. Tel: +45 45253402; fax: +45 45882673.
E-mail addresses: pdel@imm.dtu.dk, philip@delff.dk (P.D. Andersen), aiv@sbi.aau.dk (A. Iversen).

0378-7788/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.enbuild.2013.10.001

Nomenclature

AIC	Akaike information criterion
BIC	Bayes information criterion
HOR	high occupancy rate
LOR	low occupancy rate
rmse	root mean square error
τ	a time stamp in continuous time
t	a time stamp in continuous time
T	maximum of t, i.e. $t \in [0, T]$
n	a time stamp in discrete time
N	maximum of n, i.e. $n \in \{0, a, \ldots, N\}$
$\{X_n\}$	a random process in discrete time
X_n	the state of the random process $\{X_n\}$ at time n
x_n	the observation of the random process $\{X_n\}$ at time n
$\boldsymbol{X}^{(i)}$	the ith sequence of observations
$p_n \in [0, 1]$	the unconditioned probability of $X_n = 1$
$\boldsymbol{A}$	a matrix
$\boldsymbol{A}^T$	$\boldsymbol{A}$ transposed
M	number of states in a Markov chain
I	the characteristic function
Q	the number of sequences of observations
$\log : \mathbb{R} \to \mathbb{R}$	the natural logarithm
$\mathbb{N}$	the set of natural numbers, $\{1, b, \ldots\}$
$i, j, k \in \mathbb{Z}$	integers
μ_i	the mean of the ith sequence of observations
α	the intercept in the linear domain of the generalized linear model
β_i	the weighting of the ith basis spline in the linear domain of the generalized linear model
ρ_i	the weighting of the ith power of time of day in the linear domain of the generalized linear model
γ	the weighting of the exponential smoothing in the linear domain of the generalized linear model
λ	the parameter for the exponential smoothing filter
Λ_n	the value of the exponential smoothing filter at time step, n
θ	a parameter vector

- Can model dependence based on past behavior improve predictions?

The aim is to present a framework for modeling occupancy in an office environment and then apply it to fitting data that is believed to be representative. The focus will be on modeling what can be considered "typical" occupant presence sequences, where "typical" is to be judged from data. Sequences of very little presence are expected to be frequent because of vacation, sickness, etc. Sequences of significantly more presence than "typical" will be omitted. The focus is on single-person simulation, and correlation structures in data will not be modeled.

The outcome is techniques for an occupancy simulation model that can be used in building simulation programs when simulating demand responsive systems such as lighting or ventilation systems.

2. Methods

In this section, the data collection method and the mathematical framework to be used in the analysis will be described.

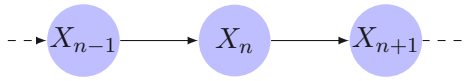

Fig. 1. Illustration of dependence in a Markov chain. The Markov condition says that the distribution of future states (here X_{n+1} and on) conditioned on the present state (X_n) and all past states (up to X_{n-1}) is the same as the distribution of future states that are only conditioned on the present state. Therefore, in the graph, X_{n-1} and X_{n+1} are only connected through X_n.

2.1. Data collection

Occupancy patterns have been measured in an office building in San Francisco, CA. Data comes from ballast status records in the control system and have been registered every 2 min. If an occupant is present at the workspace, the lamp is switched on, and the ballast status is on. Once the workspace is unoccupied the lights drop to preliminary power and are turned off after a delay of 20 min. The occupants cannot override anything manually. The data collected have been corrected for the delay by setting the last 20 min of intervals of "presence" to "absence". However, absences shorter than 20 min have not been encountered because of the delay in the equipment.

Data from 86 workspaces were collected, out of which 29 were unoccupied or occupied by interns. Only data from the 57 workspaces that have been occupied by full-time staff for the entire measurement period is used.

The model fitting is based on full days in September and December 2009 and January 2010; 16 days in total. No data points are missing.

2.1.1. Description of models

All models in the present work are in discrete time. Let $t \in [0, T]$ be a continuous time scale. Choose a natural number, N, and let $\tau := T/N$. Then $t_n = n\tau$, $n \in \{0, 1, \ldots, T/\tau\}$ is a discretization of t with sample period τ. The sample period is equal to the measuring period, 2 min in this work.

The notation X_n is introduced as shorthand for the state of the discrete-time random process $\{X_t\}$ at time t_n. In other words, X_n refers to $\{X_t\}$ at time $t = n\tau$, in this case $n \cdot 2$ min.

2.1.1.1. Markov chains. A Markov chain is a time series that meets the Markov condition stating that conditioned on the present state, the future is independent of the past [10]. Let Ω represent the set of possible states of X. Then, in discrete time, $\{X_n\}$ is a Markov chain if

$$\forall k \in \mathbb{N} : n + k < N, \quad \forall s \in \Omega : \\ \mathbb{P}\left(X_{n+k} = s \,\middle|\, X_0, X_1, \ldots, X_n\right) = \mathbb{P}\left(X_{n+k} = s \,\middle|\, X_n\right) \tag{1}$$

This is illustrated in Fig. 1.

A Markov chain with M states is completely characterized at time n by the probabilities of transitions to all states:

$$\mathbb{P}X_{n+1} = j \mid X_n = i, \quad i, j \in \{1, \ldots, M\} \tag{2}$$

This means that the transition probabilities contain the distributions of the transitions from the states in the Markov chain. Hence, for each state they sum to one:

$$\forall i \in \{1, \ldots, M\}, \forall n \in \{1, \ldots N\} : \\ \sum_{j=1}^{M} \mathbb{P}(X_{n+1} = j \mid X_n = i) = 1 \tag{3}$$

Often the transition probabilities are represented in a probability transition matrix.

Because of the constraint in Eq. (3), at each time step the transition probabilities have $M - 1$ degrees of freedom for each state,

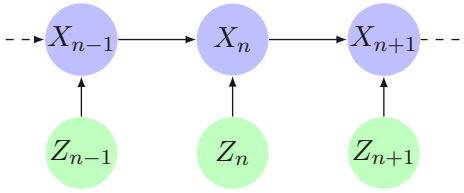

Fig. 2. Illustration of dependence in a Markov chain, $\{X_n\}$ with a covariate, $\{Z_n\}$. The input process is a deterministic process which is assumed to be known.

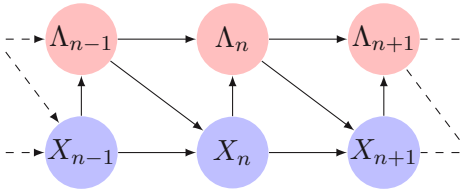

Fig. 3. A Markov chain with exponential smoothing as covariate in the transition probabilities.

corresponding to $(M-1)M$ in total for each time step. When applied to binary data, $M=2$, and hence the model has two degrees of freedom at each time step. If the transition probability matrix is constant, i.e. $\Gamma(n)=\Gamma$, $\Gamma \in \mathbb{R}^M \times \mathbb{R}^M$ the Markov chain is said to be *homogeneous*. A homogeneous Markov chain has $(M-1)M$ degrees of freedom.

2.1.1.2. Two-state Markov chains with covariates. Covariates in Markov chains with only the two states, 0 and 1, can be modeled as

$$\text{logit}\left(\mathbb{P}\left(X_{n+1}=0 \,\middle|\, X_n=0\right)\right)=Z_{1,n}\theta_1, \quad \theta_1 Z_{1,n}\in\mathbb{R}^p \tag{4a}$$

$$\text{logit}\left(\mathbb{P}\left(X_{n+1}=1 \,\middle|\, X_n=1\right)\right)=Z_{2,n}\theta_2, \quad \theta_2 Z_{2,n}\in\mathbb{R}^q \tag{4b}$$

where the logistic function denoted logit is defined as

$$\text{logit}:]0,1[\to\mathbb{R}, \quad \text{logit}(x)=\log\left(\frac{x}{1-x}\right) \tag{5}$$

and log is the natural logarithm. $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are parameter vectors while Z_1 and Z_2 are design vectors. $\mathbb{P}X_{n+1}=1 \mid X_n=0$ and $\mathbb{P}X_{n+1}=0 \mid X_n=1$ are calculated by application of Eq. (3). This formulation has the advantages that the parameters are unconstrained while the resulting probabilities span and never exceed]0, 1[. This is a *generalized linear model* [11] for binomial data, and logit is the canonical *link function* which maps from the full range of the real numbers into]0, 1[. This model has $p+q$ free parameters.

Z is a *design matrix* that can contain any observable real input. Here, functions of time will be used. One design matrix could be

$$Z^T=(1,n,n^2) \tag{6}$$

where Z^T denotes Z transposed. This would result in a second order polynomial of time to be passed through the logistic function. In (6), 1 means that an offset is included in the model (for $n=0$), and the parameter representing this offset is denoted α.

The dependence of $\{X_n\}$ on past values and on the exogenous process is illustrated in Fig. 2. Since Eqs. (4) describe transition probabilities which vary with some exogenous process, this Markov chain is *inhomogeneous*. When dependence on time of day is used, the parameter in the linear domain of the generalized linear models will be denoted ρ_i where i is the power of the time of day.

Generalized linear models are implemented in **R** and can be fitted using the `glm` function.

2.1.1.3. Natural splines. Splines are piecewise polynomial functions. In this work, B-splines with natural boundary conditions are used. These are piecewise third order polynomials with the boundary condition that the second derivatives are zero at the end-points [12]. The polynomials are between *knots* for which number and positions have to be chosen. In this work, knots are always equidistantly spread.

In **R**, the basis functions of natural splines can be calculated using the `splines` package. By using the basis functions in the design matrix, splines are fitted as input to the generalized linear model.

Where natural splines are used in the general linear models, the parameters in the linear domain are denoted β_i where i means that it relates to the ith basis function.

2.1.1.4. Exponential smoothing. Exponential smoothing is a low-pass filter. It is a weighted average, with the weights decaying exponentially with time difference. The speed of the decay is contained in the only parameter, $\lambda \in [0,1]$:

$$\Lambda_n=\lambda X_n+(1-\lambda)\Lambda_{n-1} \tag{7}$$

Since $\{\Lambda_n\}$ is a weighted average of $\{X_n\}$, it has the same range as $\{X_n\}$.

In the framework of Eqs. (4), the design matrix for a model using exponential smoothing and no covariates is

$$Z_n=(1,\Lambda_{n-1})^T \tag{8}$$

Fig. 3 is a graph of the information flow using exponential smoothing and no covariates. As seen from Fig. 3, the Markov condition is still respected when using the exponential smoothing as input as long as the most recent, and only the past states of $\{X_n\}$ are used in the design matrix as in Eq. (8). Notice that the exponential smoothing adds two parameters to the model, one is the exponent, λ, the other is the parameter in the linear domain of the generalized linear model. The latter is denoted γ.

Finally, both filtered states and exogenous processes can be used in the design matrix. A graph of this model is shown in Fig. 4. The design matrix can now include both a column with ones, polynomial functions of time of day, basis splines of time of day, and exponential smoothing of the observations.

2.1.2. Model performance assessment

The model estimation is based on the maximum likelihood principle. Let X_n be 1 for occupant presence, 0 for occupant absence at time n. Assume that it follows the Bernoulli distribution with parameter, p, the probability of $X_n=1$. Then the likelihood function of p given the observation, x_n, is:

$$\mathcal{L}(p;x_n)=\mathbb{P}X_n=x_n=\begin{cases}1-p, & x_n=0\\ p, & x_n=1\end{cases} \tag{9}$$

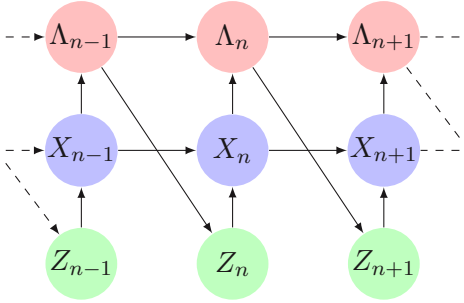

Fig. 4. A Markov chain with an exogenous process and exponential smoothing as covariate in the transition probabilities.

The joint likelihood of observations $x_1, x_2, \ldots, x_N$ is the product of the individual likelihood values:

$$\mathcal{L}(p; x^{(N)}) = \prod_{n=1}^{N} \mathcal{L}(p; x_n) \tag{10}$$

The *maximum likelihood estimate* of p refers to the value of the parameter that maximizes the likelihood function.

$$\hat{p}(x^{(N)}) = \text{argmax}_p \mathcal{L}(p; x^{(N)}) \tag{11}$$

Here, p can also be a function of other parameters, $\boldsymbol{\theta}$. Then the maximum likelihood estimate of $\boldsymbol{\theta}$ is parameters that maximizes the likelihood function.

$$\hat{\boldsymbol{\theta}}(x^{(N)}) = arg\,\max_{\boldsymbol{\theta}} \mathcal{L}(p(\boldsymbol{\theta}); x^{(N)}) \tag{12}$$

Instead of the likelihood function itself, the logarithm of the likelihood function, simply called the log-likelihood and denoted ℓ, is often used. This has the advantage that the joint log-likelihood function is a sum instead of a product:

$$\begin{aligned} \ell(p(\boldsymbol{\theta}); x^{(N)}) &= \log\left(\prod_{n=1}^{N} \mathcal{L}(p(\boldsymbol{\theta}); x_n)\right) \\ &= \sum_{n=1}^{N} \log\left(\mathcal{L}(x_n; p(\boldsymbol{\theta}))\right) \end{aligned} \tag{13}$$

Since the natural logarithm is an increasing function of all positive numbers, the log-likelihood can be maximized just as well as the likelihood itself.

In a homogeneous Markov chain, the transition probabilities can be estimated in the same way. The transition from i to j is a Bernoulli experiment that happens with probability $\mathbb{P}X_{n+1} = j \mid X_n = i$. Notice that the likelihood function of the conditional probability in (10) should only be based on the data where $X_n = i$.

For an inhomogeneous Markov chain, a parametric relation over time can be determined using parametric expressions of the transition probabilities as in (12).

2.1.2.1. The estimation routine. For a given smoothing parameter, λ, for the exponential smoothing (7), the following steps are carried out.

1 Exponential smoothing is calculated for the whole data sequence using (7).
2 Parametric expressions (basis splines or other polynomial expressions) of time of day are calculated.
3 The design matrix is formed by exponential smoothing (one column) and polynomial relations (several columns, for splines, one less than the number of knots).
4 The parameters in the general linear model are fitted using `glm` in **R**.

The log-likelihood value of this total model is used as the objective function in an optimization algorithm. For the optimization, the implementation of the Brent algorithm in `optimize` in **R** is used [13].

2.1.2.2. Information criteria. For testing models against each other, likelihood-ratio tests can be used if the models are *nested* (one model can be obtained by equaling parameters in the other to zero). Since change of positions of the spline knots leads to models that are not sub-models of each other (not nested), an *information criterion* is needed to compare the performance of different models.

The Akaike information criterion (AIC) is a popular choice of information criterion [14]. For the model, S, it is given by

$$\text{AIC}(S) = -2 \cdot \ell_S + 2 \cdot k \tag{14}$$

where ℓ_S is the log-likelihood value of the parameters of S at the maximum likelihood estimate. k is the number of parameters in the model. However, it may be an advantage to use the Bayesian information criterion (BIC) which takes the amount of data into account.

$$\text{BIC}(S) = -2 \cdot \ell_S + \log(N) \cdot k \tag{15}$$

where N is the number of data points. In this work, BIC is used for model choice.

3. Results

3.1. Data overview and preparation

Data recordings for every 2 min from 57 sensors over 16 full days were considered. The first records were from August 2009, the last from January 2010. The time stamps in the data files were in PST/PDT (Pacific Standard Time/Pacific Daylight Time). Working hours were assumed to follow local time. Therefore "time of day" is used for modeling referring to the local time zone, i.e. PST/PDT.

3.1.1. Choosing periods to model

It was investigated if some times of the day, some sensors, or even whole days should be skipped. The total number of activated sensors was inspected throughout each of the available days to ensure that none of them were holidays. The total number of occupants was plotted for all of the 16 considered days in the upper region of Fig. 5. Two days look a bit different than the rest with lower occupancy in the afternoon, but none of the days were so different that they could be considered non-working days. They are a Tuesday and a Friday and hence not one day of the week that could be different from the others. Apart from these 2 days where there is slightly lower afternoon occupancy, the days are quite similar. All days were kept for the analysis.

Narrow spikes of high occupancy, even after 8 p.m., are seen in many – if not all – of the sequences of total occupancy. This means that the status of the sensors are correlated. The spikes are unlikely to be caused by employees coming to and leaving their desk but rather by one or more persons activating several sensors. It is known that a guard walks through the building every night and this could be the cause of some of these spikes. Since these spikes are likely not to caused by usage of the workspaces, they are not of considered particularly interesting in this work.

The lower region of Fig. 5 is a boxplot of total occupant presence in the building grouped on hour of the day. It is seen that until 6 a.m., the activity is very close to zero, except for between 5 a.m. and 6 a.m. where there is a slight activity on some of the days. From between 6 a.m. and 7 a.m. to between 10 and 11 a.m. the activity increases to around 30 simultaneous positive measurements. From between 10 and 11 a.m. to between noon and 1 p.m. the total occupation decreases to slightly more than 20 as median. This drop could be explained by a lunch break. The activity increases until between 2 and 3 p.m. after which it starts dropping. After between 3 and 4 p.m. the activity drops quickly until between 6 and 7 p.m. where the median is below 5 sensors again. Also, from this plot it is clearly seen that the many narrow peaks in occupancy after 7.30 p.m. are caused by relatively few outliers from the generally low occupancy. It is seen that the variance of the occupancy is larger in the afternoon than in the morning. Time intervals where the occupancy is small were left out, and based in Fig. 5, only model occupant presence from 6 a.m. to 7 p.m. were included in the model. Only this part of data is considered from this point.

3.1.2. Identifying outlier employees

It was then checked if data from some sensors was significantly different from the rest and should be considered outliers. It was

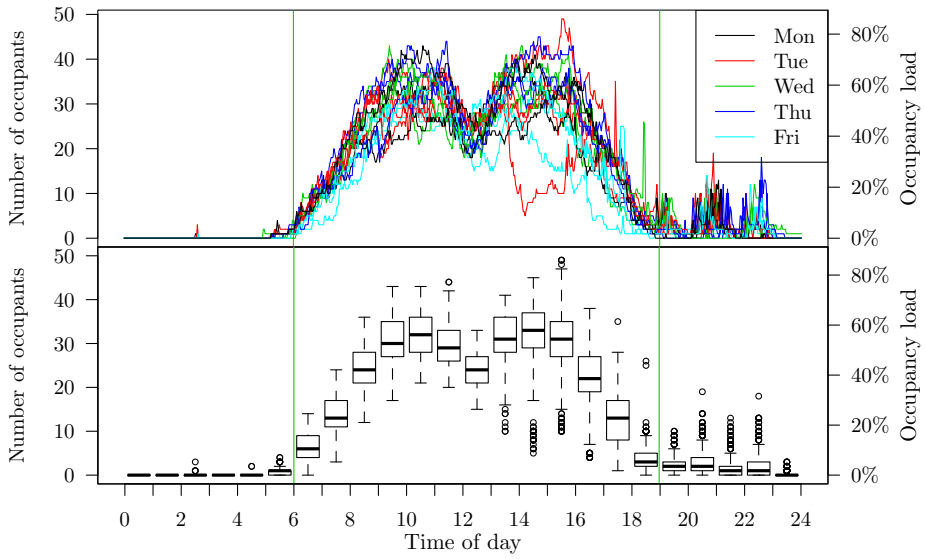

Fig. 5. Occupant presence versus time of day. The upper region shows the total number of active sensors in the office versus time of day for the 16 days considered. The lower region shows a boxplot summary of the distribution of number of present occupants, aggregated by hour of the day. The vertical green bars show the limits of the time of day kept for modeling. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

expected that single sensors would be inactive almost throughout whole days because of employees being away. A boxplot of the mean activity over each day for each sensor is shown in Fig. 6. The distribution of the daily means of the different sensors is quite different, both in medians and in variance. Many days of low occupant presence are seen, and also workspaces with generally very low occupant presence. This seems to be too many to simply disregard them as outliers and will be further investigated below. However, a few sensors have very high occupancy (6, 20, 26, 56, 57) and some of these (especially 6, 20, and 56) have low variance in occupancy. These could be located in areas that are passed by other employees throughout the day. They are considered significantly different from the rest. The vertical lines at the upper edge of the plot show the sensors that were left out.

In the data modeling description, the data considered is stripped from the outliers described here.

3.2. A hierarchic model

To determine a threshold of when to consider a sequence of measurements from one day as a working day or not, the distribution of the mean occupancy throughout a whole day of all sensors is considered. A histogram of this is seen in Fig. 7. There is a high density close to zero, and then the density is generally decreasing until mean occupant presence of a bit less than 0.2. It could be a mixture of one distribution with mode close to zero (not at work) and another with mode close to 0.6 (a work day). Based on this it is decided to make a threshold at a mean of 0.2 activity for a day-sequence. This corresponds to 2.6 h of activity. Sequences with less occupancy than 20% (from 6 a.m. to 7 p.m.) will be used to fit a *low presence rate model*, sequences with more than 20% presence are used to fit a *high presence rate model*. This is done after removal of the outliers detected in Section 3.1. The densities in Fig. 7 are

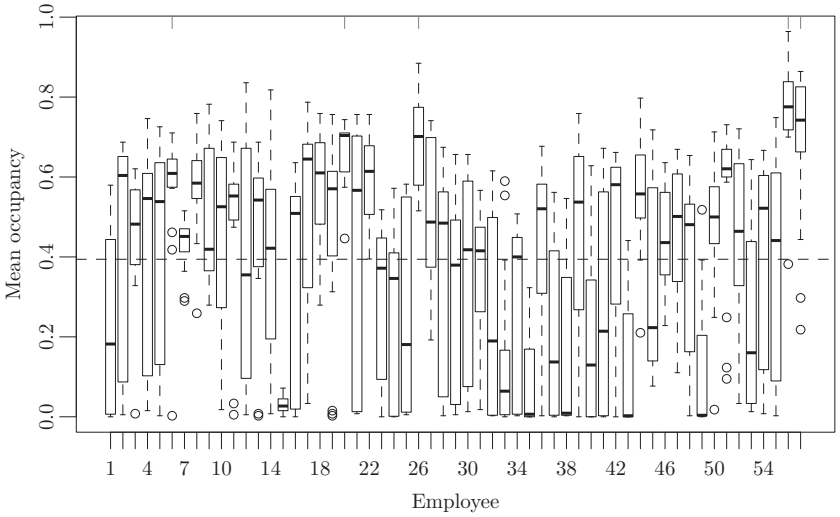

Fig. 6. Distribution of daily occupant presence for each sensor. Only 6 a.m. to 7 p.m. is considered.

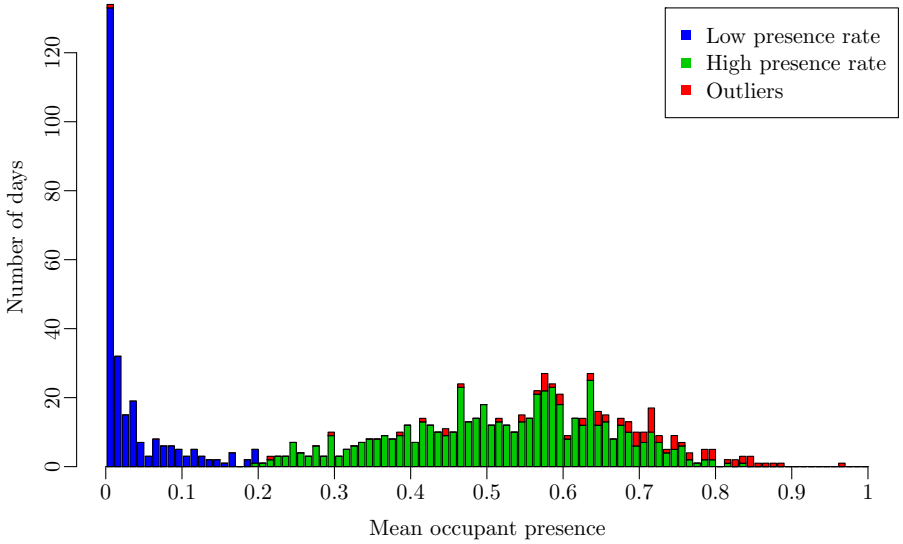

Fig. 7. Distribution of occupant presence per day for all sensors. The colors indicate what groups the different series fall into. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

colored according to this division of data. The blue color represents the sequences that fall into the "low occupancy" category, the green ones into the "high occupancy" category whereas the red ones are the outliers which are not used in the model fitting.

The model of the presence of one employee becomes a *hierarchic model*, see Fig. 8. With a certain probability, P_{HPR}, the employee is modeled with a model describing occupant presence patterns with a mean presence higher than 0.2, whereas another model with mean presence lower than 0.2 will be used with probability $1 - P_{\mathrm{HPR}}$. The model of particular interest in the present paper is the model describing presence – the "high presence rate model". Some

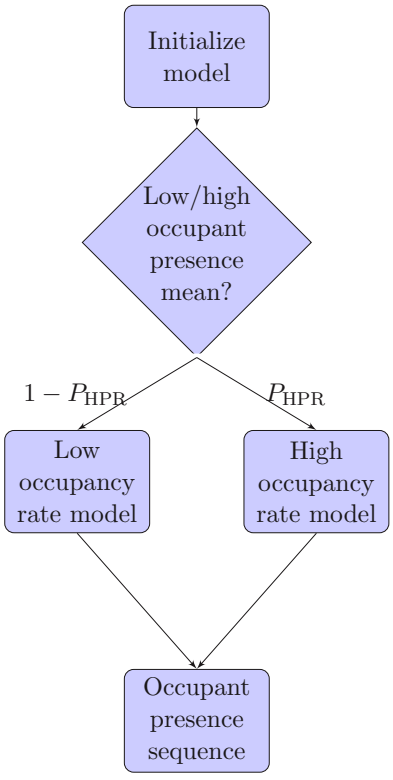

Fig. 8. The hierarchic structure of the model. With probability $\hat{P}_{\mathrm{HPR}}$ an occupant presence sequence is generated with the high presence rate model.

key properties of the partitions of the data are shown in Table 1. The procedure of estimating this model is outlined. For the low-presence sequences, the same is procedure has been carried out and the results will be given.

The probability, P_{HPR} was estimated as

$$\hat{P}_{\mathrm{HPR}} = \frac{1}{N_s}\sum_{s=1}^{N_s} I(\mu_s > 0.2) \approx 0.686 \tag{16}$$

where μ_s is the mean presence in the sequence, s.

$$\mu_s = \frac{1}{N}\sum_{n=1}^{N} X_t^{(s)} \tag{17}$$

3.3. Initial state

Since inhomogeneous processes do not have steady state properties, it was decided to base the initial conditions on the expected occupancy presence at the start of the simulations (6 a.m.). The expected presence was estimated for the HPR and the LPR independently as the mean presence of the occupants in the data sequences for the HPR and the LPR group, respectively. A Bernoulli experiment was then carried out to start each simulation in either "absence" or "presence" for each simulation. The mean value of this Bernoulli experiment was either $\hat{p}_{0,\mathrm{LPR}} = 0$ or $\hat{p}_{0,\mathrm{HPR}} \approx 0.045$.

3.4. High occupancy rate model

Two different events must be described, namely the transition from absent (0) to present (1) and from present to absent. Different models will be applied, their performances assessed, and the best one will be picked.

For every 2-min interval, the conditional probability of a transition to 1, given that 0 is observed was estimated. This is an estimate for a time of day, n. These local estimates are shown as points in Fig. 9. Also fits of generalized linear models with splines of 11 knots (10 basis splines) and different exponential smoothing levels are shown. The range of the exponential smoothing is the range that the model can take for this data. Because of the ten observations of absence which will always follow a sequence of presence, this interval is [0, 0.102]. The tendency to start working is small at 6 a.m. and it only slowly increases the first hour. Then, from 7 a.m. to 9 p.m.,

Table 1
Overview of the partitioning of the occupant presence sequences.

Group	Number of sequences	Mean	Variance of mean of sequences	Min. mean of sequences	Max. mean of sequences
HPR	571	0.521	0.018	$2.05 \cdot 10^{-1}$	0.836
LPR	260	0.031	0.002	0	0.197
Outliers	80	0.672	0.024	$3 \cdot 10^{-3}$	0.964
Total	911	0.394	0.068	0	0.964

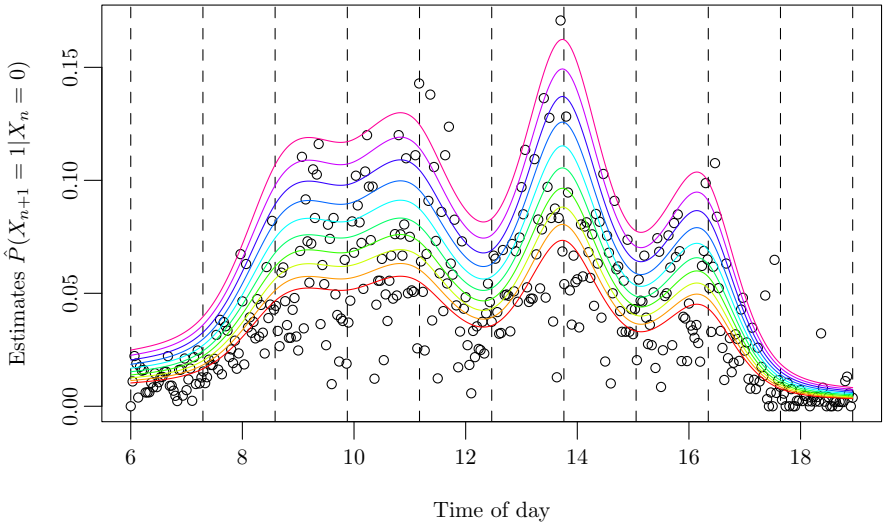

Fig. 9. Two-minute estimates of the probability of occupancy for an employee at the next time step given that he or she is absent. The smoothing level ranges from 0 to 0.102 with increasing probability of starting working.

this tendency grows rapidly. The growth is then slower but persists until around 10.30 a.m. where it starts decaying from about 7%. A "valley" is then seen over lunch time at around twelve. The global maximum is seen just before 2 p.m. after which it drops for a small valley before a local peak at 4 p.m. From there, it drops again and approaches zero at 7 p.m.

Similar estimates of local conditioned probabilities have been made for transitions from presence to presence. These are shown in Fig. 10 together with generalized linear models based on 8 knots (7 basis splines) and exponential smoothing at different levels. The smoothing levels here range from 0.186 corresponding to the lowest possible level given that the process is in "presence" to 1 – corresponding to having been in present in all history. The main tendencies are that given the smoothing level, the tendency to remain at one's work desk is quite constant except for during lunchtime and after around 3 p.m. where it drops quickly.

The decision on a model structure is based on BIC. BIC values for the different models applied are plotted in Fig. 11. A large gain is seen in going from using homogeneity or a 1st order polynomial to at least a third order polynomial or splines. The increase in BIC between these models could be because of a suboptimal positioning of the knots. The exponential smoothing improves all the models implemented measured on BIC. The best model is found to be based on a spline with 11 knots and the exponential smoothing. This gives 13 parameters in total. Table 2 shows the parameter estimates in the chosen generalized linear model of the probability of occupancy at time $n+1$ conditioned that an employee is idle at time n.

Using likelihood-ratio tests, it was checked that all parameters in this model are significant (p-values shown in Table 2). The exponential smoothing parameter is 0.205. The glm parameter estimate related to the exponential smoothing is 8.4. Since there will never be a switch back from 0 to 1 after less than 10 zeros, the exponential smoothing level cannot exceed $1 \cdot (1 - 0.205)^{10} \approx 0.1$.

The same analysis has been carried out for modeling the probability of occupancy at time $n+1$ given that the employee is occupant at time n. The improved model found here is a generalized linear model with an intercept and 5 basis spline functions. Exponential smoothing did not improve this model significantly. The resulting parameter estimates in the generalized linear model are shown in Table 3. It is seen that some p-values (for the likelihood-ratio tests in which the parameters are zero) are large here, meaning that at least one spline basis function is insignificant. This can occur because the knot placements are not optimized but determined to be equidistant, and the number of knots is decided from BIC.

Table 4 shows an overview of the aggregated performance (all transitions – from absence as well as presence) of the best

Table 2
Parameter estimates in the HOR model of transitions from occupant absence to presence, their confidence intervals, and p-values for the test of the hypothesis that the individual parameters are zero.

Term	Estimate	2.5%	97.5%	Pr (>Chi)
α	−4.56	−4.83	−4.31	
β_1	1.88	1.58	2.19	0.000
β_2	1.50	1.08	1.92	0.000
β_3	2.03	1.65	2.41	0.000
β_4	0.71	0.34	1.09	0.000
β_5	2.73	2.35	3.11	0.000
β_6	0.55	0.13	0.97	0.010
β_7	2.24	1.84	2.64	0.000
β_8	−0.78	−1.19	−0.38	0.000
β_9	−0.77	−1.44	−0.09	0.027
β_{10}	−1.26	−1.72	−0.83	0.000
γ	7.67	6.40	8.93	0.000
λ	0.19			

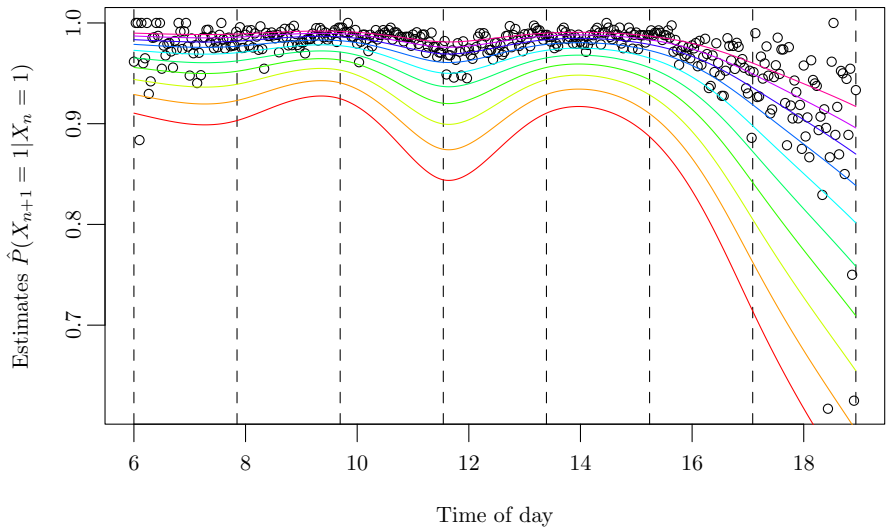

Fig. 10. Two-minute estimates of the probability of presence for an employee at the next time step given that he or she is present. The smoothing level ranges from 0.186 to 1, with increasing probability of staying work.

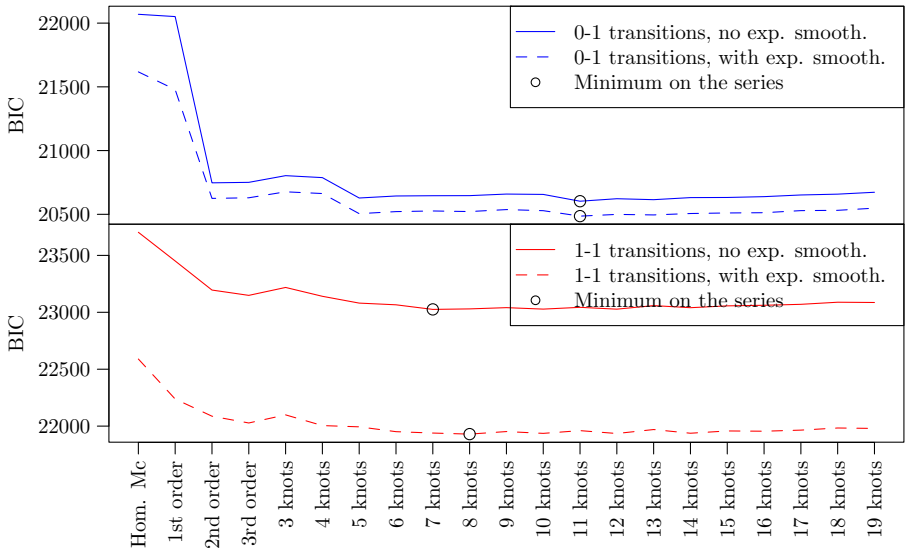

Fig. 11. BIC for different models applied to the transitions in the high presence rate part of data. 0 represents absence, 1 presence in legends.

(measured on BIC) of the different types of models that were applied on the high occupancy rate data. It is seen that the inhomogeneous models outperform the homogeneous ones measured on bias and rmse, and that the exponential smoothing further increases the performance.

Table 3
Parameter estimates in the HOR model of transitions from occupant presence to occupant presence, their confidence intervals, and p-values for the test of the hypothesis that the individual parameters are zero.

Term	Estimate	2.5%	97.5%	Pr(>Chi)
α	1.76	1.38	2.17	
β_1	0.66	0.25	1.04	0.002
β_2	−1.19	−1.68	−0.72	0.000
β_3	0.32	−0.12	0.74	0.157
β_4	−0.08	−0.55	0.37	0.729
β_5	−1.46	−1.78	−1.14	0.000
β_6	−2.22	−3.14	−1.34	0.000
β_7	−2.02	−2.34	−1.69	0.000
γ	2.82	2.68	2.97	0.000
λ	0.20			

3.5. Low occupancy rate model

The same procedure as for the high occupancy rate model has been carried through to find a low occupancy rate model. In this case, the exponential smoothing was not found significant to include in the generalized linear model. For the model of the probability of presence at time $n+1$ given that the occupant is idle at time n, a generalized linear model based on a spline and a total of six parameters was found to perform best. The parameter estimates are listed in Table 5.

Table 4
Performance measures for the best of each type of the models applied on the high occupancy rate part of data. The inhomogeneous Markov chain using exponential smoothing has both the better rmse and bias.

HOR Model	k	rmse	Bias	log Lik
Hom. MCs	4	0.146	$3.96 \cdot 10^{-11}$	−22875
Inh. MCs	18	0.145	$-1.73 \cdot 10^{-14}$	−21711
Inh. MCs, e.s.	23	0.144	$9.92 \cdot 10^{-15}$	−21087

Table 5
Parameter estimates in the LOR model of the probability of presence at $n+1$ given absence at n.

Term	Estimate	2.5%	97.5%	*Pr* (>Chi)
α	−6.69	−7.31	−6.13	
β_1	2.03	1.47	2.63	0.000
β_2	1.33	0.61	2.11	0.000
β_3	3.18	2.69	3.68	0.000
β_4	2.58	1.25	4.02	0.000
β_5	−1.93	−2.59	−1.32	0.000

Table 6
Parameter estimates in the LOR model of the probability of presence at $n+1$ given presence at n.

Term	Estimate	2.5%	97.5%	*Pr* (>Chi)
α	−2.88	−4.49	−1.26	
ρ_1	0.66	0.39	0.93	0.000
ρ_2	−0.03	−0.04	−0.02	0.000
γ	0.32	0.13	0.51	0.001
λ	0.59			

For the modeling of the probabilities of occupancy at time $n+1$ given occupancy at time n, the chosen generalized linear model is based on a second order polynomial and no exponential smoothing. The parameter estimates are listed in Table 6.

Table 7 lists aggregated performance measures of models on the low occupancy rate part of data. Again the inhomogeneous Markov chains perform better when measured on bias and rmse and the performance is further improved by adding exponential smoothing. However, the latter has little effect on the low occupancy rate data. In this model, exponential smoothing is only used on the transitions from presence, see Tables 5 and 6.

4. Simulations

The estimation was based on data from a 16-day period. The estimated models were then used to simulate a new 16-day period. These are simulations of the full system as sketched in Fig. 8 for as many occupant day sequences as available in data after omitting outliers. This corresponds to monitoring 52 employees for 16 days, resulting in 832 sequences in total. As in Fig. 8 each sequence is simulated with the high occupancy rate model with probability $\hat{P}_{\mathrm{HPR}}$ (see Eq. 16), and with the low occupancy rate model with probability $1-\hat{P}_{\mathrm{HPR}}$. This gave 565 sequences simulated with the high occupancy rate model and 267 simulated with the low occupancy rate model. Once the choice between LOR and HOR has been made, the initial value of the sequence is determined by a Bernoulli experiment. In the LOR model, the mean value of the Bernoulli experiment is the average of the LOR group at 6 a.m., and for the HOR model the mean value of the Bernoulli experiment is the average of the HOR group at 6 a.m.

The upper plot in Fig. 12 shows the sequences of total occupancy versus time of day for the simulated data using the model chosen in Section 3. This is to be compared with the plots in Fig. 5. The simulations all start with low occupancy (due to initial conditions), they have a peak before lunch, and one after. At 7 p.m. the occupant presence has dropped close to zero. This general tendency captures the tendency seen in the data very well. However, the data seems to vary slightly more, especially after the lunch break, mainly because of the two days discussed in Section 3.1.

The lower plot in Fig. 12 shows the mean of total occupant presence over the day and an estimated confidence interval for the total simulated occupant presence. The statistics are shown for the data series, the homogeneous Markov chain simulations (both for LOR and HOR), and the inhomogeneous MCs with and without exponential smoothing. Whereas the Markov chain due to the homogeneity does not capture the dependence of time, the two inhomogeneous models both have this ability. It is seen that the exponential smoothing does not have a big influence on the mean occupancy over the day. This is expected as exponential smoothing is a filter that influences the dynamics at per-employee level. Hence exponential smoothing is not important for mean value considerations for large systems. From the confidence intervals, it is again seen that in the afternoon, the variance in total occupancy is larger for the data than for any of the models.

The distribution of the simulated occupancy for employees throughout single days is shown in Fig. 13. This should be compared with Fig. 7. It is seen that the fitted LOR model tends to give fewer days of almost no occupancy and fewer days with occupancy over 0.1. The HOR model seems to fit the distribution in the data nicely. However, the tails of the distribution are slightly longer than what is seen in the data.

Table 7
Performance measures for the applied models on the low occupancy rate part of data. The inhomogeneous Markov chain without exponential smoothing has both the smallest rmse and the smallest bias.

LOR model	k	rmse	bias	log Lik
Hom. MCs	2	0.112	$-4.10\cdot10^{-11}$	−5875
Inh. MCs	9	0.111	$-1.70\cdot10^{-5}$	−5734
Inh. MCs, e.s.	11	0.111	$-3.43\cdot10^{-4}$	−5710

5. Discussion

A central assumption in this work is that the ballast status records are representative for each individual present. The validity of this assumption will depend on the office environment. But for the application of evaluating consumption which is controlled using passive infrared sensors this assumption is less important, since the data reflect activation of such sensors (apart from the delay on the turning off).

The outlined method clearly demonstrates its ability to model the variation in occupants' transitions between present and absent. However, splines and polynomials are only examples of how this can be done. Kernel smoothing provides other methods which could also be used.

Describing the variation of the transitions over the day is one problem, describing the per sequence dynamics is another. Using exponential smoothing of the occupant presence sequences significantly improved the prediction ability of the model, especially for the high occupancy rate data. Only this one method for describing the variation was tried, and others may do just as well or better. This result however shows that there is a need for modeling the per sequence dynamics if reliable single-occupant sequences are wanted.

The idea of using exponential smoothing comes from reading the work of [15] where exponential smoothing of observations is used as input in the model of the transition probabilities in a hidden Markov chain. Such a model was also tried on this data, but the state dependent distributions turned out to be Bernoulli distributions with practically certain success. This means that the Markov chain is practically not hidden. Hence the idea of directly observing the Markov chain. However, this limits the model framework to only two states. It is possible that for other data sets, more (hidden) states would give a better fit. Such states could be interpreted as "meeting", "short break", "gone home", etc. These would then tend to lead to absence of different lengths.

This method could be directly used to simulate occupant presence profiles in a building simulation program. However, more

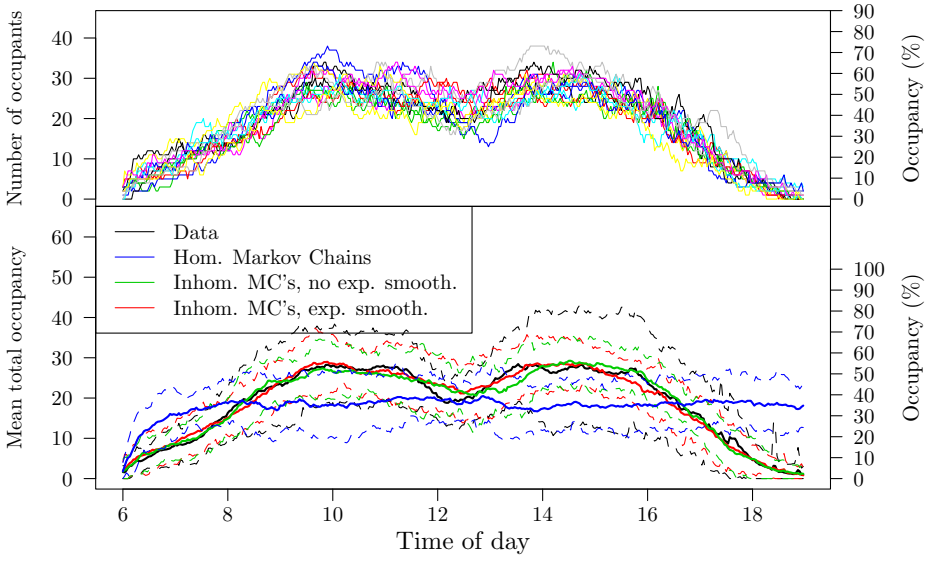

Fig. 12. Simulation of total occupant presence for a 16-day period. All employees on all days are independently simulated using the model structure as in Fig. 8. The upper plot shows the sequences simulated using the models chosen in Section 3, and the lower plot shows mean values and confidence intervals over the day for simulations using the best of different types of models.

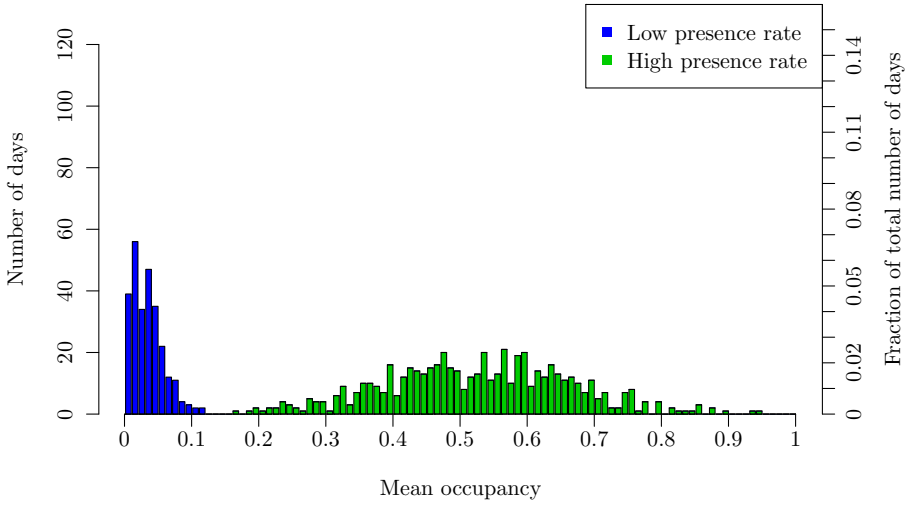

Fig. 13. Histogram of fraction of time on a day that occupants are present in simulations using the model chosen in Section 3.

data must be analyzed in order to provide good standard values for different kind of office environments and other uses.

In general the choice of model depends on the data set. Before more general conclusions can be drawn on the subject, different data sets from different sources must be analyzed.

Only independent single-occupant profiles were fitted and simulated. Correlations between occupants were not studied, neither were day-to-day correlations for single occupants. Depending on the application these correlations may be important.

6. Conclusions

Occupant presence patterns for employees in an office environment have been modeled based on data collected from electrical ballasts triggered by passive infrared sensors. After compensation for a delay in switching off the ballasts and removal of outliers, data was divided into "low occupancy rate" and "high occupancy rate" patterns which were fitted independently and the probability of activation of the two resulting models was estimated.

By use of generalized linear models based on natural splines and exponential smoothing of observations, the daily patterns were fitted. By use of the fitted models, new occupant presence patterns were simulated, and they demonstrated similar mean occupancy over the day, and the distribution of the occupancy per day had the same two-peak property as the data. The mean occupancy per versus time-of-day fit using homogeneous Markov chains did not capture the two-peaks tendency with a drop around lunch time and the drop in the afternoon.

While using exponential of the observations as a covariate in the Markov chains did not seem to have any large effect on the dependency of the time of day, it significantly improved the one-step predictions. This is thought to reflect an improved model of the dynamics of the sequences.

The outlined method can be used for generating reliable occupant presence sequences and can be included in building simulation tools. Some objectives for further studies of the subject were given, and they include modeling of modeling of data from different environments, and modeling of correlation structures between occupants and/or between days.

Acknowledgments

The authors would like to thank Francis Rubinstein and his research group at Lawrence Berkeley National Laboratory for giving us access to the measured data.

References

[1] F. Haldi, Towards a unified model of occupants' behaviour and comfort for building energy simulation, Ph.D. Thesis, 2010.
[2] P. Hoes, J.L.M. Hensen, M.G.L.C. Loomas, B. de Vries, D. Bourgeois, User behaviour in whole building simulation, Energy and Buildings 41 (2009) 295–302.
[3] D. Manicca, B. Rutledge, M.S. Rea, W. Morrow, Occupant use of manual lighting controls in private offices, Journal of the Illuminating Engineering Society 28 (1999) 42–56.
[4] J. Page, D. Robinson, N. Morel, J.-L. Scartezzini, A generalised stochastic model for the simulation of occupant presence, Energy and Buildings 40 (2008) 83–98.
[5] V. Tabak, B. de Vries, Methods for the prediction of intermediate activities by office occupants, Building and Environment 45 (2010) 1355–1372.
[6] D. Wang, C.C. Federspiel, F. Rubinstein, Modeling occupancy in single person offices, Energy and Buildings 37 (2005) 121–126.
[7] D. Bourgeois, Detailed occupancy prediction, occupancy-sensing control and advanced behavioural modelling within whole-building energy simulation, Ph.D. Thesis, 2005.
[8] C. Reinhart, Lightswitch-2002: a model for manual and automated control of electric lighting and blinds, Solar Energy 77 (2004) 15–22.
[9] I. Richardson, M. Thomson, D. Infield, A high-resolution domestic building occupancy model for energy demand simulations, Energy and Buildings 40 (2008) 1560–1566.
[10] G. Grimmett, D. Stirzaker, Probability and Random Processes, 3rd ed., Oxford University Press, 2005.
[11] H. Madsen, P. Thyregod, Introduction to General and Generalized Linear Models, 1st ed., Chapman & Hall/CRC, 2011.
[12] L. Eldén, L. Wittmeyer-Koch, H.B. Nielsen, Introduction to Numerical Computation - Analysis and MATLAB® illustrations, The Authers and Studenterlitteratur, Lund, 2004.
[13] R. Brent, Algorithms for Minimization without Derivatives, Dover, 2002.
[14] L. Wasserman, All of Statistics – A Concise Course in Statistical Inference, Springer Science + Bussiness Media, Inc, 2003.
[15] W. Zucchini, D. Raubenhaimer, I.L. McDonald, Modeling time series of animal behavior by means of latent-state model with feedback, Biometrics 64 (2008) 807–815.