HMMs Recursive Parameters Estimation for Semi-Bounded Data Modeling: Application to Occupancy Estimation in Smart Buildings

Fatemeh Rezapoor Nikroo, Manar Amayri and Nizar Bouguila
CIISE, Concordia University, Montreal, Canada

Keywords: Hidden Markov Models, Recursive Estimation, Semi-Bounded Data, Smart Buildings, Occupancy Estimation.

Abstract: Optimizing energy consumption is one of the key factors in smart buildings developments. It is crucial to estimate the number of occupants and detect their presence when it comes to energy saving in smart buildings. In this paper, we propose a Hidden Markov Models (HMM)-based approach to estimate and detect the occupancy status in smart buildings. In order to dynamically estimate the occupancy level, we develop a recursive estimation algorithm. The developed models are evaluated using two different real data sets.

1 INTRODUCTION

The daily increase in energy consumption has led to global warming. Global energy demand has continuously increased while building sector has had a major effect in this rapid growth in energy consumption (Kim et al., 2022). In this context, smart buildings promise automated systems to control energy consumption. Environmental control systems have been proved as a crucial factor in smart buildings. Another decisive factor in energy consumption concerns the occupants themselves (Gaetani et al., 2016). Automatic occupancy detection and estimation approaches allow building energy systems to manage the energy consumed. While 35% of USA's energy consumption is attributed to the heating, cooling and ventilation (HVAC) systems (Ali and Bouguila, 2022; Erickson et al., 2014), occupants' behavior has also a major influence on building energy consumption (Soltanaghaei and Whitehouse, 2016). Studies have shown that machine learning algorithms are crucial for smart buildings applications (Alawadi et al., 2020). For instance, machine learning models have been used to measure HVAC actuation levels (Ebadat et al., 2013). Furthermore, a supervised learning model has been developed in (Amayri et al., 2016) to estimate the number of occupants based on sensorial data (e.g. motion detection, power consumption, CO2 concentration sensors, microphone, or door/window positions).

Machine learning approaches can be grouped into 3 main categories: 1) generative models such as mixture models and HMMs, 2) Discriminative models such as support vector machine (SVM) and 3) Heuristic-based models which combine the 2 previous families with heuristic information. In this paper we propose HMM-based occupancy models. A first crucial factor when deploying a HMM model is the choice of probability density function (Nguyen et al., 2019). Thus, we investigate 3 distributions dedicated for semi-bounded data (i.e. positive vectors) which are detailed in section 2. The second important factor is to estimate the unknown parameters which is generally done using maximum likelihood estimation (MLE) within the expectation maximization (EM) framework. Handling real-time data of smart buildings requires continuous processing which is challenging. Therefore, one of the motivations of this paper is to propose a novel architecture to cope with real-time data. Online learning techniques provide solutions addressing real-time occupancy estimation to build models that can be continuously updated (Amayri et al., 2020). We introduce a recursive algorithm with linear time complexity to detect and/or estimate the number of occupants.

2 BACKGROUND

2.1 Hidden Markov Models

HMMs are powerful statistical models. The idea of HMM comes from a limit of Markov model in modeling problems which output is the probabilistic function of the states (Epaillard and Bouguila, 2016). This

Nikroo, F., Amayri, M. and Bouguila, N.
HMMs Recursive Parameters Estimation for Semi-Bounded Data Modeling: Application to Occupancy Estimation in Smart Buildings.
DOI: 10.5220/0011715200003491
In *Proceedings of the 12th International Conference on Smart Cities and Green ICT Systems (SMARTGREENS 2023)*, pages 81-88
ISBN: 978-989-758-651-4; ISSN: 2184-4968

function called emission probability is described in details in the following sections. Indeed, we assume more general distributions dedicated to positive vectors to model the output observable data from HMM to have more flexible models. It is noteworthy to mention the denomination of Hidden in HMM refers to the states not the parameters of the model. Markov chain is referred to a time-varying random phenomenon meeting the Markov property which indicates that the conditional probability of the forthcoming state is just based on the current state and not on historical information, which can be mathematically formulated as:

$$p(X_{t+1}|X_t, X_{t-1}, ..., X_1) = p(X_{t+1}|X_t) \quad (1)$$

HMM elements are completely defined as follow, however, it is mainly represented by three parameters $\lambda = (\pi, A, B)$ (Vaičiulytė and Sakalauskas, 2020).
1) The number of states N, each state is defined by S_i such that $S = 1, 2, ..., N$.
2) Vector $\pi = \pi_1, \pi_2, ..., \pi_N$ indicate the probability of being in each state.
3) The number of possible observations M for each state as $V = V_1, V_2, ..., V_M$. In case of our work that observations come from the distributions that we will define later, they are continuous thus, M is infinite.
4) A is the state transition probability matrix such that $a_{i,j}$, $1 \le i, j \le N$ is the probability of moving to state j at time $t+1$ while the model was in state i at time t. The constraints for transition matrix should be met:

$$a_{i,j} >= 0, \quad \sum_{j=1}^{N} a_{i,j} = 1, \quad 1 \le i, j \le N$$

5) B is the matrix showing the observation probability where $b_j(k)$ is the probability of observing V_k in state S_j.

$$b_j(k) = p(V_k^t|S_j^t), \quad 1 \le i, j \le N$$

The constraints for continuous observations are defined based on the specific probability distribution:

$$b_j(k) = \sum_{m=1}^{M} c_{jm} p(x|\theta)$$

where θ represents the parameters of the defined distribution, c_{jm} is weighting coefficient with $\sum_{m=1}^{M} c_{jm} = 1$.

$$b_j(k) >= 0, \quad 1 \le j \le N, \quad 1 \le k \le M, \quad \sum_{j=1}^{M} b_j(k) = 1$$

2.2 Inverted Dirichlet Distribution

Consider $X = (x_1, x_2, ..., x_K)$ as a vector following the inverted Dirichlet distribution $ID(\alpha)$ with parameter $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K, \alpha_{K+1})$, the probability density function has the following form (Bdiri and Bouguila, 2013):

$$p(X) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{K+1} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} (1 + \sum_{i=1}^{K} X_i)^{-\alpha_0}$$

where $\alpha_0 = \sum_{i=1}^{K+1} \alpha_i$, $\quad x_i > 0$, $i = 1...K$. Many properties of the inverted Dirichlet distribution are given in (Tiao and Cuttman, 1965; Bdiri and Bouguila, 2012).

2.3 Generalized Inverted Dirichlet Distribution

Inverted Dirichlet distribution assumes a positive correlation, therefore a generalization of it is introduced to cope with this limitation in order to have the capability of modeling wider range of real-life data (Bourouis et al., 2014a). Consider $X = (x_1, x_2, ..., x_K)$ as a vector following the Generalized Inverted Dirichlet distribution $GID(\alpha; \beta)$ with parameter $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K, \alpha_K)$ and $\beta = (\beta_1, \beta_2, ..., \beta_K)$. The probability density function has the following form (lingmnwah, 1976; Bourouis et al., 2014b):

$$p(x_1, x_2, ..., x_K) = \prod_{i=1}^{K} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \frac{x_i^{\alpha_i - 1}}{(1 + \sum_{m=1}^{i} x_m)^{\eta_i}}$$

where $\eta_i = \beta_i + \alpha_i - \beta_{i+1}$ $\quad i = 1, 2, ..., K$, $\beta_{K+1} = 0$, $\quad x_i > 0$, $i = 1...K$. By substituting $\beta_1 = \beta_2 = ... = \beta_{K-1} = 0$, inverted Dirichlet distribution with parameter $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K, \alpha_K, \beta_K)$ is obtained (lingmnwah, 1976; Bourouis et al., 2014b).

2.4 Inverted Beta-Liouville Distribution

Inverted Beta-Liouville distribution (IBL) has been proved as an efficient way of modeling positive vectors (Bouguila et al., 2022; Bourouis et al., 2021). It overcomes the limit of the Inverted Dirichlet in the aspect of positive covariance and presents less parameters as compared with generalized Inverted Dirichlet (Bouguila et al., 2022; Bourouis et al., 2021). IBL is in the family of the Beta-Liouville distribution which is a natural model for analyzing compositional data (Fang et al., 2018). Consider $r = \lambda w/(1-w)$ which w is a Beta distribution with parameters α and β as the generating variate, thus r follows an inverted Beta distribution with parameters β and λ. The generating density function is (Fang et al., 2018):

$$f(u|\theta) = \frac{1}{B(\alpha, \beta)} \frac{\lambda^\beta u^{\alpha-1}}{(\lambda + u)^{\alpha+\beta}} \quad (2)$$

where $u > 0$ and $\theta = (\alpha, \beta, \lambda)$. The Liouville distribution with positive parameters $(\alpha_1, \alpha_2, ..., \alpha_K)$ and generating density of $f(.)$ with parameter δ is defined as below (Bouguila, 2012; Fang et al., 2018):

$$p(\vec{X}|\alpha_1,...,\alpha_K,\delta) = f(u|\delta)\frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{u^{\sum_{i=1}^{K}\alpha_i - 1}}\prod_{i=1}^{K}\frac{x_i^{\alpha_i-1}}{\Gamma(\alpha_i)}$$

We have the probability density function for $\vec{X}$ as follows:

$$p(x_1,x_2,...,x_K) = \frac{\Gamma(\alpha_0)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\lambda^{\beta}(\sum_{i=1}^{K}x_i)^{\alpha-\alpha_0}}{(\lambda+\sum_{i=1}^{K}x_i)^{\alpha+\beta}}\prod_{i=1}^{K}\frac{x_i^{\alpha_i-1}}{\Gamma(\alpha_i)}$$

where $\theta = (\alpha_1,\alpha_2,...,\alpha_K,\alpha,\beta,\lambda)$, $\alpha_0 = \sum_{i=1}^{K}\alpha_i$, $x_i > 0$, $i = 1...K$. This distribution can be converted to inverted Dirichlet distribution by equalizing $\alpha = \alpha_0$ (Fang et al., 2018). Many properties of the distribution are given in (Hu et al., 2019; Koochemeshkian et al., 2020).

3 RECURSIVE MODEL

To learn the HMM parameters, we need to estimate the distribution parameters of the HMM (Vaičiulytė and Sakalauskas, 2020). To achieve this goal we calculate the log likelihood of each probability density function then maximize it. Each probability density function is explained in a different sub-section. However, the whole process is the same. It is noteworthy that ψ used in equations below is the Digamma function which is the logarithmic derivative of the Gamma function. The probability that sequence x is observed while the system is at specific time t considering state q, which can be any in range of $q = 1,,2,...,N$, is:

$$\frac{\pi_t^{<q>}\log distribution(x_t|\alpha<q>)}{\sum_{j=1}^{N}\pi_t^{<j>}\log distribution(x_t|\alpha<j>)} \quad (3)$$

Now the summation over t indicates the probability of the model being in specific state considering specific output sequence x is observed:

$$\sum_{t=1}^{T}\frac{\pi_t^{<q>}\log distribution(x_t|\alpha<q>)}{\sum_{j=1}^{N}\pi_t^{<j>}\log distribution(x_t|\alpha<j>)} \quad (4)$$

where "*distribution*" refers to each distribution we have used in this paper. Here, we define some same variables which are used in the rest of the paper for all the 3 distributions.

$$\theta_t^q = \pi_t^q \log distribution(x_t|\vec{\theta}^{<q>}) \quad (5)$$

$$\omega_t^{<q,s>} = \omega_{t-1}^{<q,s>} + \frac{1}{t}\left(\frac{\log(x_t^s)\theta_t^q}{\sum_{j=1}^{N}\theta_t^j} - \omega_{t-1}^{<q,s>}\right) \quad (6)$$

$$\gamma_t^q = \gamma_{t-1}^q + \frac{1}{t}\left(\frac{\theta_t^q}{\sum_{j=1}^{N}\theta_t^j} - \gamma_{t-1}^q\right) \quad (7)$$

where $1 \leq q \leq N$, $1 \leq s,i \leq K$.

3.1 Inverted Dirichlet Distribution

The log likelihood function of inverted Dirichlet (ID) distribution is as follows:

$$\log ID(\vec{X}|\vec{\theta}) = \log\Gamma(\sum_{i=1}^{K+1}\alpha_i) + \sum_{i=1}^{K}(\alpha_i - 1)\log x_i - (\sum_{i=1}^{K+1}\alpha_i)\log(1+\sum_{i=1}^{K}x_i) - \sum_{i=1}^{K+1}\log\Gamma(\alpha_i)$$

To maximize the above equation, derivative of it with respect to each parameter is shown below:

$$\frac{\partial \log ID(\vec{X}|\vec{\theta})}{\partial\alpha_j} = \begin{cases}\psi(\sum_{i=1}^{K+1}\alpha_i) + \log x_j - \log(1+\sum_{i=1}^{K}x_j) - \psi(\alpha_j) & \text{if } j = 1,2,...,K \\ \psi(\sum_{i=1}^{K+1}\alpha_i) - \log(1+\sum_{i=1}^{K}x_j) - \psi(\alpha_j) & \text{if} j = K+1\end{cases}$$

Now the batch formula to derive α_i is as below:

$$\alpha_i = \psi^{-1}[\psi(\sum_{i=1}^{K+1}(\alpha_i^q)) + \frac{\frac{1}{T}\sum_{t=1}^{T}\log(x_t^s)\frac{\pi_t^{<q>}\log ID(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N}\pi_t^{<j>}\log ID(x_t|\vec{\theta}^{<j>})}}{\frac{1}{T}\sum_{t=1}^{T}\frac{\pi_t^{<q>}\log ID(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N}\pi_t^{<j>}\log ID(x_t|\vec{\theta}^{<j>})}} - \frac{1}{T}\sum_{t=1}^{T}\frac{\pi_t^{<q>}\log ID(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N}\pi_t^{<j>}\log ID(x_t|\vec{\theta}^{<j>})}] \quad i = 1,2,...K$$

$$\alpha_{K+1} = \psi^{-1}[\psi(\sum_{i=1}^{K+1}(\alpha_i^q)) - \frac{1}{T}\sum_{t=1}^{T}\frac{\pi_t^{<q>}\log ID(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N}\pi_t^{<j>}\log ID(x_t|\vec{\theta}^{<j>})}] \quad (8)$$

where $1 \leq q \leq N$, $1 \leq s \leq K$. We assume each observation occurs in specific time step t. Hence, in our recursive model, the parameters are updated as time goes until they meet the condition of $|\theta_{new} - \theta_{old}| <= \varepsilon$. The middle variables and the main parameters are:

$$\alpha_t^{<q,s>} = \psi^{-1}[\psi(\sum_{i=1}^{K+1}(\alpha_i^q)) + \frac{\omega_t^{<q>}}{\gamma_t^q} - \gamma_t^q] \quad (9)$$

$$\alpha_t^{<q,K+1>} = \psi^{-1}[\psi(\sum_{i=1}^{K+1}(\alpha_i^q)) + \frac{\omega_t^{<q>}}{\gamma_t^q} - \gamma_t^q] \quad (10)$$

where $1 \leq q \leq N$, $1 \leq s,i \leq K$.

3.2 GID Distribution

The log likelihood function GID distribution is:

$$\log GID(\vec{X}|\vec{\theta}) = \sum_{i=1}^{K} \log(\alpha_i+\beta_i) + (\alpha_i - 1)\log x_i - \log\Gamma(\alpha_i) - \log\Gamma(\beta_i) - \eta \log(1+\sum_{m=1}^{i} x_m)$$

To maximize the above equation, derivative of it with respect to each parameter is shown below:

$$\frac{\partial \log GID(\vec{X}|\vec{\theta})}{\partial \alpha_j} = \psi(\alpha_j+\beta_j) + \log(x_j) - \psi(\alpha_j) - \log(1+\sum_{m=1}^{j} x_m)$$

$$\frac{\partial \log GID(\vec{X}|\vec{\theta})}{\partial \beta_j} = \psi(\alpha_j+\beta_j) - \psi(\beta_j) - \log(1+\sum_{m=1}^{j} x_m) \quad (11)$$

Now the batch formula to derive α_i and β_i are:

$$\alpha_i = \psi^{-1}[\psi(\alpha_i+\beta_i) + \frac{\frac{1}{T}\sum_{t=1}^{T} \log(x_t^s) \frac{\pi_t^{<q>} \log GID(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N} \pi_t^{<j>} \log GID(x_t|\vec{\theta}^{<j>})}}{\frac{1}{T}\sum_{t=1}^{T} \frac{\pi_t^{<q>} \log GID(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N} \pi_t^{<j>} \log GID(x_t|\vec{\theta}^{<j>})}} - \frac{1}{T}\sum_{t=1}^{T} \frac{\pi_t^{<q>} \log GID(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N} \pi_t^{<j>} \log GID(x_t|\vec{\theta}^{<j>})}] \quad (12)$$

$$\beta_i = \psi^{-1}[\psi(\alpha_i+\beta_i) - \frac{1}{T}\sum_{t=1}^{T} \frac{\pi_t^{<q>} \log GID(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N} \pi_t^{<j>} \log GID(x_t|\vec{\theta}^{<j>})}] \quad (13)$$

$$\alpha_t^{<q,s>} = \psi^{-1}[\psi(\alpha+\beta) + \frac{\omega_t^{<q>}}{\gamma_t^q} - \gamma_t^q] \quad (14)$$

$$\beta_t^{<q,s>} = \psi^{-1}[\psi(\alpha+\beta) - \gamma_t^q], \quad 1 \le q \le N \quad (15)$$

3.3 IBL Distribution

The log likelihood function IBL is as follows:

$$\log IBL(\vec{X}|\vec{\theta}) = \log\Gamma(\sum_{i=1}^{K} \alpha_i) + \log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta) + \beta\log\lambda + (\alpha - \sum_{i=1}^{K}\alpha_i)\log(\sum_{i=1}^{K} x_i) - (\alpha+\beta)\log(\lambda+\sum_{i=1}^{K} x_i) + \sum_{i=1}^{K}(\alpha_i - 1)\log x_i - \sum_{i=1}^{K}\log\Gamma(\alpha_i)$$

To maximize the above equation:

$$\frac{\partial \log IBL(\vec{X}|\vec{\theta})}{\partial \alpha_j} = \psi(\sum_{i=1}^{K}\alpha_i) + \log x_j - \psi(\alpha_j) - \log(\sum_{i=1}^{K} x_i)$$

$$\frac{\partial \log IBL(\vec{X}|\vec{\theta})}{\partial \alpha} = \psi(\alpha+\beta) - \psi(\alpha) + \log(\sum_{i=1}^{K} x_i) - \log(\lambda+\sum_{i=1}^{K} x_i)$$

$$\frac{\partial \log IBL(\vec{X}|\vec{\theta})}{\partial \beta} = \psi(\alpha+\beta) - \psi(\beta) + \log\lambda - \log(\lambda+\sum_{i=1}^{K} x_i) \quad (16)$$

$$\frac{\partial \log IBL(\vec{X}|\vec{\theta})}{\partial \lambda} = \frac{\beta}{\lambda} - \frac{\alpha+\beta}{\lambda+\sum_{i=1}^{K} x_i} \quad (17)$$

The batch formula to derive α_i, α, β, λ are as below:

$$\alpha_i = \psi^{-1}[\psi(\sum_{i=1}^{K}(\alpha_i^q)) + \frac{\frac{1}{T}\sum_{t=1}^{T} \log(x_t^s) \frac{\pi_t^{<q>} \log IBL(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N} \pi_t^{<j>} \log IBL(x_t|\vec{\theta}^{<j>})}}{\frac{1}{T}\sum_{t=1}^{T} \frac{\pi_t^{<q>} \log IBL(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N} \pi_t^{<j>} \log IBL(x_t|\vec{\theta}^{<j>})}} - \frac{1}{T}\sum_{t=1}^{T} \frac{\pi_t^{<q>} \log IBL(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N} \pi_t^{<j>} \log IBL(x_t|\vec{\theta}^{<j>})}] \quad (18)$$

$$\alpha = \psi^{-1}[\psi(\alpha+\beta) + \frac{1}{T}\sum_{t=1}^{T} \frac{\pi_t^{<q>} \log IBL(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N} \pi_t^{<j>} \log IBL(x_t|\vec{\theta}^{<j>})}]$$

$$\beta = \psi^{-1}[\psi(\alpha+\beta) + \log\lambda - \frac{1}{T}\sum_{t=1}^{T} \frac{\pi_t^{<q>} \log IBL(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N} \pi_t^{<j>} \log IBL(x_t|\vec{\theta}^{<j>})}] \quad (19)$$

$$\lambda = \frac{\beta}{\alpha}\frac{1}{T}\sum_{t=1}^{T} \frac{\pi_t^{<q>} \log IBL(x_t|\vec{\theta}^{<q>})}{\sum_{j=1}^{N} \pi_t^{<j>} \log IBL(x_t|\vec{\theta}^{<j>})} \quad (20)$$

where $1 \le q \le N$, $1 \le s, i \le K$. Based on the middle variables defined earlier, above explanation of recursive logic and the batch formula of IBL, the main parameters are described as below:

$$\alpha_t^{<q,s>} = \psi^{-1}[\psi(\sum_{i=1}^{K+1}(\alpha_i^q)) + \frac{\omega_t^{<q>}}{\gamma_t^q} - \gamma_t^q] \quad (21)$$

$$\alpha_t^{<q,s>} = \psi^{-1}[\psi(\alpha+\beta)+\gamma_t^q] \tag{22}$$

$$\beta_t^{<q,s>} = \psi^{-1}[\psi(\alpha+\beta)+\log\gamma-\gamma_t^q] \tag{23}$$

$$\lambda_t^{<q,s>} = \frac{\beta}{\alpha}\gamma \tag{24}$$

where $1 \leq q \leq N$, $1 \leq s \leq K$. We present the recursive proof of the model equations:

$$\begin{aligned}\omega_t^{<q,s>} &= \omega_{t-1}^{<q,s>} + \frac{1}{t}\left(\frac{\log(x_t^s)\theta_t^q}{\sum_{j=1}^N \theta_i^j} - \omega_{t-1}^{<q,s>}\right) \\ &= \frac{1}{t-1}\sum_{i=1}^{t-1}\frac{\log(x_t^s)\theta_i^q}{\sum_{j=1}^N \theta_i^j} + \frac{1}{t}\left(\frac{\log(x_t^s)\theta_t^q}{\sum_{j=1}^N \theta_t^j}\right. \\ &\quad \left. - \frac{1}{t-1}\sum_{i=1}^{t-1}\frac{\log(x_t^s)\theta_i^q}{\sum_{j=1}^N \theta_i^j}\right) = \frac{1}{t}\sum_{i=1}^{t}\frac{\log(x_i^s)\theta_i^q}{\sum_{j=1}^N \theta_i^j}\end{aligned}$$

$$\begin{aligned}\gamma_t^q &= \gamma_{t-1}^q + \frac{1}{t}\left(\frac{\theta_t^q}{\sum_{j=1}^N(\theta_t^j)} - \gamma_{t-1}^q\right) \\ &= \frac{1}{t-1}\sum_{i=1}^{t-1}\frac{\theta_i^q}{\sum_{j=1}^N \theta_i^j} + \frac{1}{t}\left(\frac{\theta_t^q}{\sum_{j=1}^N \theta_t^j} - \frac{1}{t-1}\sum_{i=1}^{t-1}\frac{\theta_i^q}{\sum_{j=1}^N \theta_i^j}\right) \\ &= \frac{1}{t}\sum_{i=1}^{t}\frac{\theta_i^q}{\sum_{j=1}^N \theta_i^j}\end{aligned}$$

We present the algorithm for all of the models in Algorithm 1 which is based on the EM framework. The variables $\vec{\mu}_1=(\alpha_1,\alpha_2,...,\alpha_K,\alpha_{K+1})$, $\vec{\mu}_2=(\alpha_1,\alpha_2,...,\alpha_K,\alpha_K,\beta_1,\beta_2,...,\beta_K)$, $\vec{\mu}_3=(\alpha_1,\alpha_2,...,\alpha_K,\alpha,\beta,\lambda)$ are representative of ID-HMM, GID-HMM and IBL-HMM parameters, respectively.

Algorithm 1: Recursive expectation maximization algorithm for ID-HMM, GID-HMM and IBL-HMM parameter estimation.

Output: $\vec{\mu}_1$, $\vec{\mu}_2$ and $\vec{\mu}_3$ with respect to conditions, $1 \leq t \leq T, 1 \leq s \leq K, 1 \leq q \leq N$
Initialization: $\vec{\mu}_1$,$\vec{\mu}_2$,$\vec{\mu}_3$, initial-probability of each state and transition-probability between states
Input: each row of dataset; x_t, $1 \leq t \leq T$
While $|\vec{\mu}_{1_t} - \vec{\mu}_{1_{t-1}}| < \varepsilon$, $|\vec{\mu}_{2_t} - \vec{\mu}_{2_{t-1}}| < \varepsilon$, $|\vec{\mu}_{3_t} - \vec{\mu}_{3_{t-1}}| < \varepsilon$
E-step
Calculate values of $\theta_t^q, \omega_t^{<q,s>}, \gamma_t^q$;
M-step:
Update values of $\vec{\mu}_1$, $\vec{\mu}_2$, $\vec{\mu}_3$

The algorithm starts with random initialization of the parameters of both HMM and distributions we consider for data. Then, the algorithm goes through the loop for E-step and M-step until the termination criterion is met which is based on monitoring the difference between the previous value and update one of each parameter after each loop. This difference value is shown as ε and set to 0.6. As we have mentioned before, in each loop the values are updated based on the new rows of data feed to the model. We assume that each row of dataset occurs in a specific time, so that each time the new data is obtained, parameters are updated.

4 EXPERIMENTAL RESULTS

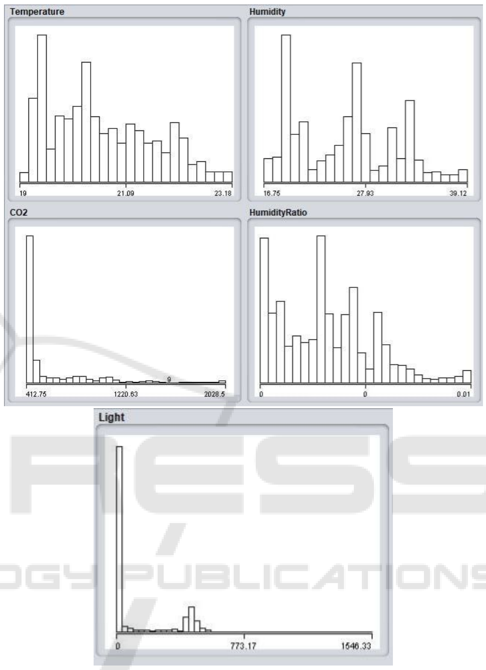

Figure 1: Visualization of occupancy detection data according to different features.

In this section we present the validation of our proposed models using 2 different real datasets for both binary and multiclass classification of occupants in smart buildings. Data that we have investigated, is collected during time. Thus, based on the logic of its application to estimate occupants in a room, each row of data shows the conditions of the room with the number of people in specific time, so we were able to train the model as each row of the data comes, to update the parameters. In that case we have successfully overcame the problem of batch learning for intensive data. The main motivation to evaluate our model on occupancy datasets is the fact that in smart buildings the goal is to automate the systems related to HVAC which offers less energy wasting along with better comfort of residents for facilities management strategic decisions. In smart buildings the sensors collecting environmental factors like the amount of CO_2

Table 1: Estimated parameters for ID-HMM, GID-HMM and IBL-HMM for occupancy detection dataset.

Model	Number of States	Parameters
ID-HMM	2	$\alpha = \begin{bmatrix} 0.2389 & 0.2414 & 0.4244 & 12.47 & 0.2266 & 13.03 \\ 0.2346 & 0.2369 & 1.196 & 12.42 & 0.2227 & 13.34 \end{bmatrix}$
GID-HMM	2	$\alpha = \begin{bmatrix} 8.723 & 8.706 & 8.708 & 8.876 & 8.633 \\ 8.786 & 8.603 & 8.610 & 8.778 & 8.586 \end{bmatrix}$ $\beta = \begin{bmatrix} 8.281 & 8.149 & 8.164 & 9.564 & 7.607 \\ 9.329 & 9.339 & 9.329 & 9.330 & 9.328 \end{bmatrix}$
IBL-HMM	2	$\vec{\alpha} = \begin{bmatrix} 0.9105 & 0.5656 & 0.0679 & 0.0322 & 0.5916 \\ 0.3517 & 0.3470 & 0.1467 & 0.1037 & 0.3807 \end{bmatrix}$ $\alpha = \begin{bmatrix} 0.7223 & 2.478 \end{bmatrix}$ $\beta = \begin{bmatrix} 0.6733 & 0.5181 \end{bmatrix}$ $\lambda = \begin{bmatrix} 2.89 & 2.03 \end{bmatrix}$

concentration, temperature, relative humidity, etc are integrated in automation settings. Thus, we have the collected information of sensors in hand allowing us to estimate the number of occupants. This is the idea in HMM, by mapping the collected data as our observation and the number of occupants as the hidden states. In all of the experiments, we define the initial probability of each states (matrix π explained earlier) as $\frac{1}{n}$ assuming n is the total number of states. Transition matrix along with each distributions' parameters are assigned randomly considering their limit according to the distribution definition. The termination criterion of the algorithm to avoid endless recursion is set to $\varepsilon = 0.6$. To analyze our datasets collected over time, we assume each record occurred in new time step. Therefore, in our recursive model, the parameters are updated as new record of data feed into the model.

Binary Classification Dataset (Occupancy Detection Dataset): The first data used to evaluate our models is related to occupancy detection. This dataset is obtained from Machine Learning Repository of University of California Irvine (UCI). Fig. 1 describes the distribution of the features. Dataset consists of 5 different features as below which are indicated in a time series: 1) Temperature, in Celsius, 2) Relative Humidity, 3) Light, in Lux, 4) CO2, in ppm, 5) Humidity Ratio, Derived quantity from temperature and relative humidity, in kgwater-vapor/kg-air.

This dataset has two different parts for training and testing. We have used the training data to train the model to estimate its parameters, then using the test data to evaluate the models. Table 1 shows the number of hidden states for the occupancy detection dataset along with the values of the model parameters. The values computed through the EM algorithm

Table 2: Accuracy, F-score, precision and recall in percent for Inverted Dirichlet HMM, Generalized Inverted Dirichlet HMM and Inverted Beta-Liouville HMM applied for occupancy detection dataset.

Model	Accuracy	F-score	precision	recall
ID-HMM	86.81	85.65	82.86	85.65
GID-HMM	86.90	87.55	89.22	86.90
IBL-HMM	84.00	79.85	86.58	84.00

are explained earlier. Model parameters, based on the distribution considered for emission probability in HMM, follow different dimensions. Table 2 indicates the evaluation results of our model based on the 3 distributions we discussed which are computed on the testing dataset.

Multiclass Classification Dataset (Occupancy Estimation Dataset): This dataset is used for occupancy estimation in smart building as the goal was to estimate the number of occupants from 0 to 4. Thus, the number of hidden states in this sample will be 5. This dataset is obtained from an experiment which testbed was an office in Grenoble Institute of Technology in France (Nasfi et al., 2020). This data was obtained from 30 sensors of motions, power consumption, acoustic pressure, and door position (Amayri et al., 2020). In order to investigate the model we split the dataset into training and testing data. We use the training data to estimate the parameters of our model then apply the trained model on testing data for classification to estimate the number of occupants. Table 3 represents the values of parameters for each model learned with occupancy estimation dataset, while the evaluation metrics are shown in Table 4. The visualisation of the data is presented in Fig. 2.

Table 3: Estimated parameters for ID-HMM, GID-HMM and IBL-HMM for occupancy estimation dataset.

Model	Number of States	Parameters
ID-HMM	5	$\alpha = \begin{bmatrix} 0.5860 & 0.0859 & 0.9971 & 0.5545 & 0.2836 \\ 0.5054 & 0.9837 & 0.2110 & 0.7645 & 0.9573 \\ 0.3307 & 0.1928 & 0.3457 & 0.2385 & 0.7199 \\ 0.4826 & 0.3005 & 0.1633 & 0.8344 & 0.9434 \\ 0.5469 & 0.4305 & 0.3090 & 0.0707 & 0.6270 \end{bmatrix}$
GID-HMM	5	$\alpha = \begin{bmatrix} 7.361 & 6.149 & 4.917 & 4.447 \\ 2.146 & 2.493 & 2.591 & 2.592 \\ 4.498 & 5.912 & 6.943 & 6.747 \\ 1.057 & 1.062 & 1.369 & 1.134 \\ 8.013 & 6.171 & 1.003 & 5.850 \end{bmatrix}$ $\beta = \begin{bmatrix} 5.294 & 1.663 & 1.329 & 1.200 \\ 2.720 & 3.288 & 3.456 & 3.457 \\ 1.365 & 1.426 & 1.543 & 1.490 \\ 2.365 & 4.393 & 3.264 & 2.781 \\ 1.211 & 9.174 & 4.215 & 2.368 \end{bmatrix}$
IBL-HMM	5	$\vec{\alpha} = \begin{bmatrix} 15.25 & 40.16 & 39.95 & 39.33 \\ 0.1666 & 0.4702 & 0.7526 & 0.6555 \\ 0.3187 & 0.4525 & 0.718 & 0.6868 \\ 0.1185 & 0.1837 & 0.3916 & 0.2882 \\ 0.1212 & 0.1379 & 0.3014 & 0.2345 \end{bmatrix}$ $\alpha = \begin{bmatrix} 0.6309 & 15.27 & 13.84 & 18.46 & 8.989 \end{bmatrix}$ $\beta = \begin{bmatrix} 0.1706 & 2.971 & 73.36 & 33.32 & 38.40 \end{bmatrix}$ $\lambda = \begin{bmatrix} 6.469 & 2.001 & 0.3127 & 2.343 & 1.790 \end{bmatrix}$

Table 4: Accuracy, F-score, precision and recall in percent for Inverted Dirichlet HMM, Generalized Inverted Dirichlet HMM and Inverted Beta-Liouville HMM applied for occupancy detection dataset.

Model	Accuracy	F-score	precision	recall
ID-HMM	80.59	79.25	79.36	84.51
GID-HMM	78.24	85.27	82.87	88.02
IBL-HMM	74.53	86.86	87.93	89.94

5 CONCLUSION

In this paper, we introduce a novel approach for occupancy estimation in smart buildings through HMMs. Considering the inverted Dirichlet, GID and IBL distributions as underlying distributions describing the observation data. The models have been successfully evaluated on real-data. The goal was to reach an accurate prediction of number of occupants in a room which plays a key role in smart buildings to reduce energy consumption. One of the main motivations to apply these distributions is their flexibility. In addition, the developed models are based on a recursive approach, thus the time complexity of the algorithm

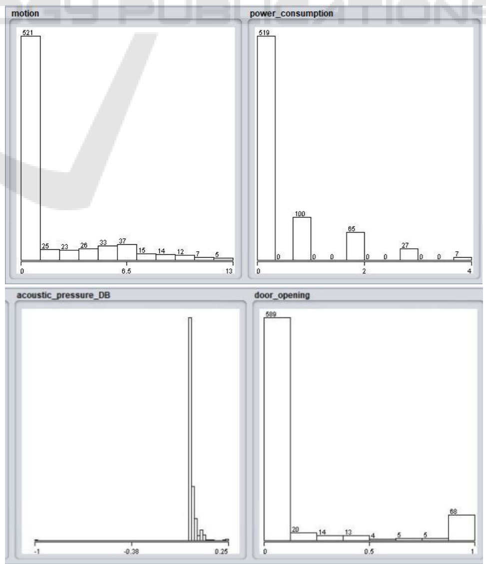

Figure 2: Visualization of occupancy estimation data according to different features.

reduces to linear time as compared with batch processing which in turn causes a substantial decrease in memory overload and computational resources. Future aspirations could be devoted to improve the initialization methods. Additionally, due to the importance of prepossessing data in machine learning models, to further enhance the model accuracy, feature selection and data quality assessment methods could be investigated.

REFERENCES

Alawadi, S., Mera, D., Fernandez-Delgado, M., Alkhabbas, F., Olsson, C. M., and Davidsson, P. (2020). A comparison of machine learning algorithms for forecasting indoor temperature in smart buildings. *Energy Systems*, pages 1–17.

Ali, S. and Bouguila, N. (2022). Towards scalable deployment of hidden markov models in occupancy estimation: A novel methodology applied to the study case of occupancy detection. *Energy and Buildings*, 254:111594.

Amayri, M., Arora, A., Ploix, S., Bandhyopadyay, S., Ngo, Q.-D., and Badarla, V. R. (2016). Estimating occupancy in heterogeneous sensor environment. *Energy and Buildings*, 129:46–58.

Amayri, M., Ploix, S., Bouguila, N., and Wurtz, F. (2020). Database quality assessment for interactive learning: Application to occupancy estimation. *Energy and Buildings*, 209:109578.

Bdiri, T. and Bouguila, N. (2012). Positive vectors clustering using inverted dirichlet finite mixture models. *Expert Systems with Applications*, 39(2):1869–1882.

Bdiri, T. and Bouguila, N. (2013). Bayesian learning of inverted dirichlet mixtures for svm kernels generation. *Neural Computing and Applications*, 23(5):1443–1458.

Bouguila, N. (2012). Infinite liouville mixture models with application to text and texture categorization. *Pattern Recognition Letters*, 33(2):103–110.

Bouguila, N., Fan, W., and Amayri, M. (2022). *Hidden Markov Models and Applications*. Springer.

Bourouis, S., Alroobaea, R., Rubaiee, S., Andejany, M., M.Almansour, F., and Bouguila, N. (2021). Markov chain monte carlo-based bayesian inference for learning finite and infinite inverted beta-liouville mixture models. *IEEE Access*, 9:71170–71183.

Bourouis, S., Mashrgy, M. A., and Bouguila, N. (2014a). Bayesian learning of finite generalized inverted dirichlet mixtures: Application to object classification and forgery detection. *Expert Syst. Appl.*, 41(5):2329–2336.

Bourouis, S., Mashrgy, M. A., and Bouguila, N. (2014b). Bayesian learning of finite generalized inverted dirichlet mixtures: Application to object classification and forgery detection. *Expert Systems with Applications*, 41(5):2329–2336.

Ebadat, A., Bottegal, G., Varagnolo, D., Wahlberg, B., and Johansson, K. H. (2013). Estimation of building occupancy levels through environmental signals deconvolution. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8.

Epaillard, E. and Bouguila, N. (2016). Proportional data modeling with hidden markov models based on generalized dirichlet and beta-liouville mixtures applied to anomaly detection in public areas. *Pattern Recognit.*, 55:125–136.

Erickson, V. L., Carreira-Perpi, M. A., and Cerpa, A. E. (2014). Occupancy modeling and prediction for building energy management. *ACM Transactions on Sensor Networks (TOSN)*, 10(3):1–28.

Fang, K. T., Kotz, S., and Wangng, K. (2018). *Symmetric Multivariate and Related Distributions*. Chapman and Hall/CRC.

Gaetani, I., Hoes, P.-J., and Hensen, J. L. (2016). Occupant behavior in building energy simulation: Towards a fit-for-purpose modeling strategy. *Energy and Buildings*, 121:188–204.

Hu, C., Fan, W., Du, J.-X., and Bouguila, N. (2019). A novel statistical approach for clustering positive data based on finite inverted beta-liouville mixture models. *Neurocomputing*, 333:110–123.

Kim, D., Yoon, Y., Lee, J., Mago, P., Lee, K., and Cho, H. (2022). Design and implementation of smart buildings: A review of current research trend. *Energies*, 15:4278.

Koochemeshkian, P., Zamzami, N., and Bouguila, N. (2020). Distribution-based regression models for semi-bounded data analysis. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4073–4080. IEEE.

lingmnwah, G. (1976). On the generalised inverted dirichlet distribution. *Demonstratio Mathematica*, 9(3):119–130.

Nasfi, R., Amayri, M., and Bouguila, N. (2020). A novel approach for modeling positive vectors with inverted dirichlet-based hidden markov models. *Knowledge-Based Systems*, 192:105335.

Nguyen, H., Rahmanpour, M., Manouchehri, N., Maanicshah, K., Amayri, M., and Bouguila, N. (2019). A statistical approach for unsupervised occupancy detection and estimation in smart buildings. In *2019 IEEE International Smart Cities Conference (ISC2)*, pages 414–419. IEEE.

Soltanaghaei, E. and Whitehouse, K. (2016). Walksense: Classifying home occupancy states using walkway sensing. BuildSys '16, page 167–176, New York, NY, USA. Association for Computing Machinery.

Tiao, G. G. and Cuttman, I. (1965). The inverted dirichlet distribution with applications. *Journal of the American Statistical Association*, 60(311):793–805.

Vaičiulytė, J. and Sakalauskas, L. (2020). Recursive parameter estimation algorithm of the dirichlet hidden markov model. *Journal of Statistical Computation and Simulation*, 90(2):306–323.