Developing a Hidden Markov Model for occupancy prediction in high-density higher education buildings

Journal of Building Engineering 73 (2023) 106795

Contents lists available at ScienceDirect

Journal of Building Engineering

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

Developing a Hidden Markov model for occupancy prediction in high-density higher education buildings

Bashar Alfalah[a,b], Mehdi Shahrestani[a,*], Li Shao[a]

[a] School of the Built Environment, University of Reading, UK
[b] College of Architectural and Planning, Imam Abdulrahman Bin Faisal University, Saudi Arabia

ARTICLE INFO

Keywords:
Occupancy
Prediction
Hidden Markov model
Higher education institution
Ground truth data
Sensor

ABSTRACT

The number of occupants in a building as well as the occupancy patterns and profiles have a significant impact on the building energy consumption associated with heating, cooling, ventilation, and lighting systems. A range of approaches has been used in previous studies for building occupancy prediction. However, there were limitations associated with the methods used for data collection. For instance, the prediction of occupancy based on the concentration of CO_2 level adds an additional level of uncertainty into the estimation of the building occupancy level. To avoid such uncertainties, high-resolution passive infrared video camera sensors were used to capture occupancy data for an extended period of one year. The collected occupancy data have then been used to develop a Hidden Markov Model (HMM), to predict the occupancy levels in a high-density higher education case study building. The model is trained under different scenarios to find the most appropriate set of attributes associated with occupancy data that can be used to develop the data driven HMM. The occupancy attributes considered include day, week, month, and term. Moreover, the collected dataset was optimally split into 70% for training and 30% for validation using cross-validation, which yields high prediction accuracy. The results of the prediction model under different scenarios were evaluated using root mean square error and Kullback Leibler (KL) divergence.

1. Introduction

The buildings are designed to provide a comfortable environment for their occupants. The number and presence of building occupants have a different impact on energy consumption during different hours of workdays and weekends. The common belief is that occupants consume more energy during working hours. However, studies conducted on commercial buildings showed that more than 50% of wasted energy was during unoccupied hours [1–3]. The last decade has shown an increase in interest in building occupancy modelling. Information regarding the occupancy numbers at different times of the day provides opportunities to use spaces more efficiently and operate buildings more energy efficient.

Machine learning models such as Hidden Markov Model (HMM) [4–17] and Artificial Neural Network (ANN) have been extensively developed in previous studies to predict occupancy numbers and/or energy consumption in various types of buildings. In a study conducted by Candanedo et al. [4], HMM was used to predict the occupancy schedule in a residential building. The model was developed using occupancy status data collected from several sensors, including temperature, humidity, carbon dioxide, light sensors, and occupancy status. Occupancy data were collected for a period of one month, excluding days when the building was unoccupied.

[*] Corresponding author.
E-mail address: m.shahrestani@reading.ac.uk (M. Shahrestani).

https://doi.org/10.1016/j.jobe.2023.106795
Received 3 December 2022; Received in revised form 22 April 2023; Accepted 7 May 2023
Available online 9 May 2023

The model was able to predict the average occupancy profiles for seven days ahead. In another study conducted by Ruy and Moon [7], HMM was developed to predict the occupancy numbers in Building Integrated Control Testbed (BICT) at Dankook University. Several data were collected for seven days, including occupancy profile, carbon dioxide, temperature, light usage, and appliances. The developed model was able to predict the occupancy number for one week. Lui et al. [9] developed HMM to predict the occupancy status of presence or absence in a single office. The occupancy data was collected using passive infrared (PIR) motion detectors for 4 h. The result shows high accuracy in predicting the occupancy status for 2 h. Another study by Lam et al. [11] developed HMM to predict the occupancy numbers in an open space office. In order to detect the presence of occupants, several sensors were used, including temperature, carbon dioxide, humidity, and motion sensors. The data were collected for 55 days in 1-min intervals. The result of the model predicted one week of occupancy numbers with an average accuracy of 68%.

Other studies have developed Artificial Neural Network (ANN) models for building occupancy prediction [16–29]. Zuraimi et al. [18] developed ANN model to predict the number of occupants in a lecture theatre. Occupancy data were collected for 80 working days using a camera and carbon dioxide sensors. The photos captured by the camera were used to count the number of occupants in the study area manually. The developed model predicted the occupancy number for five working days ahead. The result of the prediction model was evaluated by applying root mean square error (RMSE), which showed high predictive performance. In another study, Ekwevugbe et al. [19] developed ANN model to predict the occupancy numbers. Several sensors such as PIR, carbon dioxide, temperature, sound detection, and camera were used to detect six occupants in an open-plan office for a period of one month. The model predicts the occupancy numbers during occupied periods of ten days ahead. In a study conducted by Ashouri et al. [21], Wi-Fi signals were used to collect the occupancy data in an office building to train the ANN model. The occupancy data were collected for two days aiming to predict the occupancy numbers for one day ahead. The model was able to predict the number of occupants during working days with high accuracy of 90%. Alam et al. [24] also developed ANN model to predict the occupancy numbers. The occupancy data collected from a single office using carbon dioxide sensors were used for training the model to predict the number of occupants for one day ahead.

Most of the previous studies focused on specific types of buildings, such as office and residential buildings, in developing occupancy prediction models. However, educational buildings with high occupancy rates have not been sufficiently studied. This is mainly sure to the paucity of occupancy data in these buildings and also the difficulty in collecting accurate occupancy data in such buildings. Because of these challenges in terms of occupancy data, previous studies decided to limit their scope to an area in the building such as a classroom/lecture room, selected offices, or floors with a limited number of occupants during working hours of working days. Using environmental data such as temperature, humidity, and carbon dioxide concentration, as well as the status of artificial lighting systems for estimation of occupancy, has imposed additional uncertainty on the data used for building occupancy prediction in previous studies. The novelty and contribution of this study is to address the areas of improvement and limitations of previous studies focused on the prediction of building occupancy. This research focuses on a high-density building that has not been adequately explored in the literature, as previous studies have mainly focused on residential and office buildings. Therefore, this research is uniquely designed to focus on library buildings as a high-density building type in higher education sector with stochastic behaviour in terms of occupancy patterns and profiles. Models developed in this study predict occupancy levels for the entire building, rather than a specific area, which makes the prediction model more applicable. These prediction models were developed with sufficient occupancy data collected using infrared video camera sensors with a high level of accuracy (98%) for an extended period of 12 months at 5 min intervals. Accuracy of the occupancy sensors used in this study overcomes the uncertainties associated with the robustness of data collection in the previous studies. This study developed a Hidden Markov Model using real-world occupancy data with high resolution and evaluated the claims from previous studies that removing data outliers will enhance the accuracy of prediction models. Overall, this research fills gaps in the literature by providing new insights into building occupancy modelling and presenting more reliable and accurate data for future research in this area.

2. Materials and methods

This section reviews the techniques used in previous studies to collect occupancy data in Section 2.1 and defines the approaches adopted in this study Section 2.2.1. The characteristics of the data collected and used in the modelling process are described in Section 2.2.2, which is followed by the details of the approach taken for cross-validation to find the optimal data split to be used for training and validation of the model in Section 2.2.3. Finally, Section 2.3.4 introduces the indicators applied to assess the performance of the proposed occupancy prediction model.

2.1. Methods of data collection in the previous studies

Previous studies have used different approaches in collecting occupancy data during various periods from several types of buildings to develop data-driven models. The accuracy of the data collected determines how reliable and precise the model output is in reducing the gap between actual data and predicted results of the model.

However, some of these approaches used in the previous studies for collecting occupancy data add more uncertainties to the proposed models. Some examples include collecting occupancy data using CCTV cameras and counting the number of occupants manually, which is very time-consuming takes it difficult to capture the data in short intervals to provide high-resolution occupancy data and also adds human errors in manual head counting involved in this approach [18]. Other technologies were used in the past studies, such as using the concentration of carbon dioxide, temperature, and humidity that to some extent, can be correlated to the occupancy. Using such information that is loosely related to occupancy adds significant concerns over the robustness of data-driven models developed to predict occupancy based [19,25,30,31].

Another data collection method was using signals, for example Wi-Fi to detect the number of occupants in buildings by, for example, communicating to occupants' smartphones. This method has several drawbacks, such as not all occupants carrying their phones, phones are switching off, the limitation of detecting other devices signals such as smartwatches, tablets, laptops, and other devices that can connect to Wi-Fi routers [32–34]. Moreover, the new phones change the IP address automatically, so the data collected may be twice the actual numbers. Due to the ability of this technique to provide relatively reliable data only in a short period, there was a tendency in the previous studies that used this method to develop relatively short predictions, for example, three days ahead based on the data collected from the past nine days [21,35]. Using Wi-Fi technology implies a limitation on the extent of data collection. For example, the majority of previous studies that used this method for data collection limited the scope of their data collection to a zone or space with a limited number of occupants [5,6,9,11,13,24]. This was mainly because the expectation from participants could have been better communicated if the data is collected from a smaller space with a limited number of occupants. Moreover, previous studies mainly focused on collecting the data during working days and working hours and excluded the days with a few occupants, such as weekends, where knowing the occupancy patterns and profile during out of working hours and holidays plays a key role inefficient management of building services [4,7,18,22,30].

2.2. Adopted approach in this study

The aim of this study is to predict the occupancy levels in a case study building. Hidden Markov Model was adopted to be developed under several scenarios for predicting the occupancy levels in a higher education institutional building. The type of case study building selected has a limited focus from previous studies due to a lack of occupancy data and the complexity of following a reliable approach for data collection from a building with high occupancy density. Therefore, in this study, an extended period of occupancy data was collected using high-accuracy infrared video camera sensors, with the aims of collecting reliable occupancy data at building level and avoid the uncertainties faced in the previous studies. As the sensors used in previous studies, including PIR and $CO_{2,}$ were incapable of providing information to a building level, as they were limited to a space level. Moreover, the collected occupancy data covered 12 months in a high resolution of 5-min intervals compared to only a few weeks of data collected in previous studies [4,7,11]. Moreover, the collected and consequently the predicted data represented the occupancy in the entire building covering working days, weekends, vacations, and holidays, whereas the previous studies focused on working hours during working days [18,21]. In addition, cross-validation was deployed to determine the optimal spile of the data for training and validation of the proposed prediction model.

2.2.1. Hidden Markov model

Hidden Markov Model (HMM) was originally introduced by Baum and Egon [33] in the late 1960s. HMM is a popular statistical method that is successfully applied in modelling data, statistical pattern recognition and classification [34]. Over the last few years, HMM has proven that its mathematical structure can emerge as a significant method for predicting different data types, such as the occupancy number and energy consumption in various types of buildings. HMM was one of the most developed models in the previous studies, promising high prediction accuracy. HMM works by considering the current state that depends on the past state, where the state space is considered hidden from the observer [36]. It is modelled by the transition and emission probabilities, which are hidden from the observer. HMM takes sequential input data to get the future sequence's score, representing the model's output. Fig. 1 illustrates the structure of the connection state of HMM, showing the hidden states and the observations that are emitted from each hidden state based on the emission probability distribution [36,37]. In this study, Hidden Markov Model (HMM) was selected as it proves the ability to predict with high accuracy in many studies. Moreover, the selection was also based on the nature of the intended inputs of the occupancy data and the desired output of occupancy levels. The work with occupants that have stochastic behaviour makes it hard to predict, where the model addresses the issues by predicting the current/present states based on the previous one in a non-deterministic way. The data collected has a pattern; therefore, HMM seems an appropriate choice, unlike if the data had random characteristics when HMM would not be a suitable choice for such a condition. The development of HMM was based on using maximum likelihood algorithms, as it is a well-established method that has been proven to work well for HMMs [45]. The choice of algorithms as it has the ability to estimate the model parameters by maximising the probability of the observed data. Constructing the

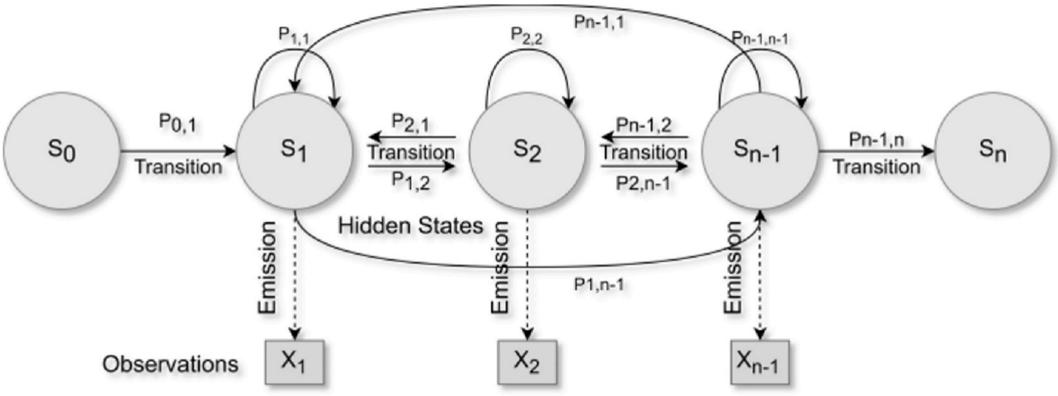


Fig. 1. The structure of HMM identifies the hidden states (Sn), the probability from states (P), and the observations (Xn).

HMM gives the ability to interact with the input parameters, which can be modified, extended, or eliminated to formulate a suitable training set to develop a data-driven HMM that can provide a high prediction accuracy.

2.2.2. Occupancy data

In this study, the occupancy data was collected from a higher educational library building at the Reading University in England. The number of occupants in the building was collected for 12 months. The occupancy data was collected using infrared video camera sensors installed at the main entrance of the building. The sensors record the occupants entering and leaving the building 24/7 at 5 min intervals. The occupancy patterns and profiles in the building were analysed in our previous study using a set of clustering analysis techniques, resulting in the identification of three occupancy patterns and profiles [38]. The three identified patterns were characterised by low, medium, and high occupancy densities. Pattern 1 is associated with medium-density, Pattern 2 is associated with high-density, and Pattern 3 is associated with low-density of occupancy numbers. Pattern 1 contains the months of January, February, and October. Pattern 2 contains four months, March, April, May, and November. Pattern 3 contains the months, June, July, August, September, and December. Fig. 2 illustrates the resolution of the occupancy data collected in three dimensions. According to Melfi et al. [39], these dimensions represent the quality of information collected from the sensor. Most previous studies represented the collected data as to whether the space/building is occupied or not without offering the options for providing the percentage of expected occupancy or an actual number of occupants in the space.

Occupancy patterns in office buildings are fairly stable over time. Such buildings typically have a higher number of occupants during office hours during working days, and their patterns of occupants can be generalised for future periods [40]. There are still days when the number of occupants changes when there are visitors, but they are limited. However, such a stable occupancy pattern is not available for buildings with high occupancy density, for example, library buildings. There is diversity in terms of activities in educational buildings that can be seen reflected in the occupancy numbers and their fluctuations on hourly, daily, weekly, monthly, and termly bases. The diversity of activities in various rooms and floors in library buildings has not been covered sufficiently in previous studies as the focus was on a space in a building. That led to an issue of scalability of the collected and developed model. Moreover, it is difficult to generalise the diversity of occupancy numbers in higher education buildings when collected for a short period of time due to the rapid changing in occupancy patterns during academic terms and vacation periods over one academic year. For this reason, occupancy data covering an extended period (more than a few months) is needed to be used in developing a prediction model for such buildings.

The occupancy data of 12 months is plotted against the day of the year, as shown in Fig. 3. Three occupancy patterns can be recognised from Fig. 3, two patterns can be associated with academic terms (Spring/Autumn Terms and Summer Terms) with a high density of occupants and another pattern associated with vacation periods with a low number of occupants that occurred between academic terms in different periods. Probability Distribution Function (PDF) was utilised in order to determine if the data collected contained outliers or, in other words, seasonal effects. The results showed that there were outliers in the data mainly related to the vacation periods. The duration of data identified related to Christmas Vacation, Easter Vacation, and Summer Vacation are highlighted by black dash lines. The identified outliers using the PDF were not data points that were outside the main body of the data, missing data, or data with zero value. The identified data has seasonal effects that cause a significant drop in the number of occupants due to the building function during vacation periods compared to the occupancy numbers in other academic terms, which is the reason for indicating such periods as an outlier. However, in this study, the identified outliers have been evaluated to determine if the inclusion/exclusion of outliers influences the performance of the prediction model.

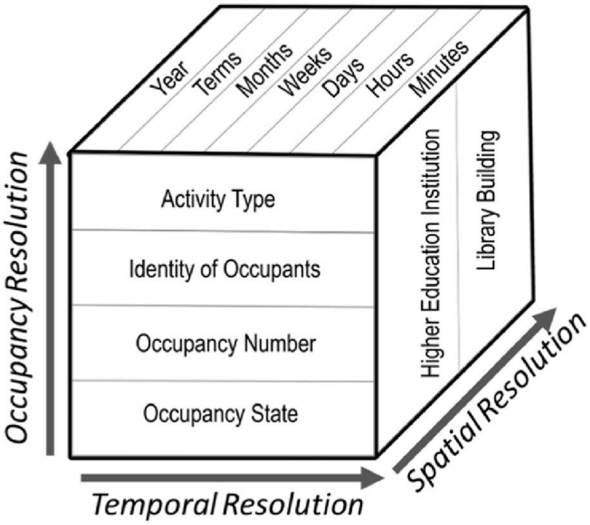


Fig. 2. The modified resolution accuracy of the collected occupancy data by the sensor shows different accuracy of temporal, spatial, and occupancy resolution. (Modified from Melfi et al. [39]).

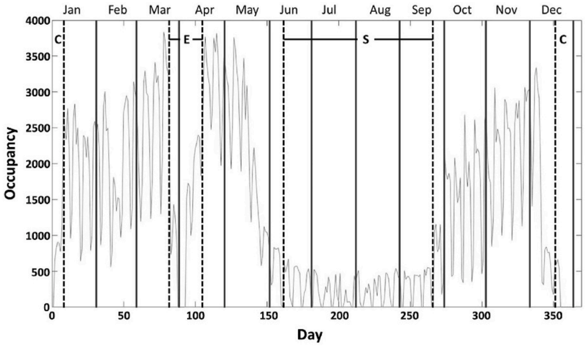


Fig. 3. PDF identified the duration of the outliers in the dataset; the black dash line with letters represents; C: Christmas Vacation, E: Easter Vacation, and S: Summer Vacation.

Various occupancy attributes were identified based on the nature of the occupancy data collected over a period of 12 months and the type of case study of the higher educational building. These attributes are the day of the week, week of the month, the month of the year and academic terms. A value has been assigned to each identified attribute, as shown in Table 1. To provide additional clarity about the weeks and academic terms attributes, any extra day of the month that did not fall into the first four weeks of the month was assigned to week five. Additionally, the attribute academic term includes three academic terms, as well as the vacations that occurred between them. The occupancy data were discretisation into bins for Hidden Markov Model using the binning technique. The technique divided the observed range (the number of occupants) into equal-density bins resulting in creating bins that represent an equivalent density with the associated values. The highest number recorded in the building was 3835 occupants, divided into 20 bins based on the binning technique (Tabel 2). Therefore, the 20 bins are uniformly covering the occupancy data in a range of 0–4000.

2.2.3. Cross validation

Cross validation (CV) is an effective method for choosing the optimal prediction model [41]. However, the method involves a long process of dividing the data to be used for training and validation and running the model for a broad range of data splits as well as comparison and evaluation of the results associated with each iteration which is very intensive in terms of time for processing [42]. In this study, CV was applied aiming to find the optimal split of the data for training the model and validating the results. The model was developed using occupancy attributes associated with 12 months of occupancy data as inputs. For this reason, the use of CV was to avoid training the model using a random splitting of the data. Generalising an academic term over the entire year was inappropriate because the occupancy patterns differ from one academic term to another. The generalisation of the random split will deteriorate the accuracy of the model and cause overfitting of the model results.

Fig. 4 shows the split of the data for training and validation by utilising CV. The various data split (12 folds) to be used for training and validation processes have been tested in predicting the occupancy levels and were evaluated using RMSE to determine the optimal iteration that gives high accuracy in occupancy prediction. In this study, cross-validation is conducted, and the outcomes show using the first 70% of data to train the model and the last 30% for validation; the details are explained in section 4.4.

2.2.4. Statistical indicators for evaluating the performance of the prediction model

Indicators in this study were used to evaluate the performance of the occupancy prediction model by measuring the similarities and differences between predicted and actual datasets. The first indicator is Kullback Leibler divergence, or simply the KL divergence, which is used to measure the differences between two datasets over the same variable (x). It is also used as a non-symmetric method to measure the difference between two datasets, p(x) and q(x) Equation (1) [43]. If the two datasets are equal, the KL divergence value will be zero. The second indicator used was Root Mean Square Error (RMSE), which is one of the well-known fitness indicators in

Table 1
The identified attributes of occupancy data and their assigned values.

Days	Value	Weeks	Value	Months	Value	Academic Terms	Value
Monday	1	Week 1	1	January	1	Christmas Vacation	1
Tuesday	2	Week 2	2	February	2	Spring Term	2
Wednesday	3	Week 3	3	March	3	Easter Vacation	3
Thursday	4	Week 4	4	April	4	Summer Term	4
Friday	5	Week 5	5	May	5	Summer Vacation	5
Saturday	6			June	6	Autumn Term	6
Sunday	7			July	7		
				August	8		
				September	9		
				October	10		
				November	11		
				December	12		

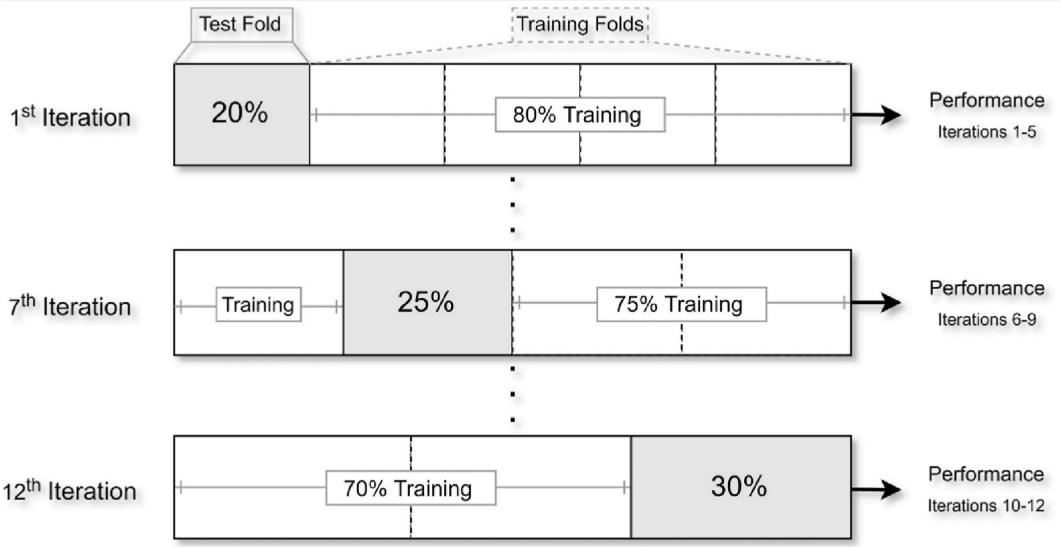


Fig. 4. 12-fold cross-validation diagram. The dataset was divided into five subsets in 20% test fold, four subsets in 25% test fold, and three subsets in 30% test fold. The performance of each iteration is measured by using the RMSE indicator.

regression models [44]. RMSE is used to calculate the variance between two datasets. Equation (2) is used to calculate the value of RMSE.

$$KL\ (P(x)\ \|\ Q(x)) = \sum P(x) Log(Q(x)\ /\ P(x)) \tag{1}$$

$$RMSE = \sqrt{1/n} \sum_{i=1} (oi - fi)^2 \tag{2}$$

3. Model development

A hidden Markov model was constructed aiming to predict the occupancy levels. The model was developed based on the inputs of several sets of data including a range of occupancy attributes as shown in Table 1, where each set represents a scenario. The attributes associated with each occupancy data point include the day of the week, week of the month, month of the year, and academic term. The model was trained using three sets of occupancy attributes in three scenarios. Scenario 1 applied the entire attributes noted in Table 3, where Scenario 2 and Scenario 3 excluded the month of the year attribute before training the model. In addition, the outliers presented in Fig. 3 were eliminated in Scenario 3 before training the model.

3.1. Scenario 1

The first scenario is to be trained using all of the four identified occupancy attributes, including the day of the week, week of the month, the month of the year, and academic terms, for the aim of evaluating the performance of the model in using various inputs. The domain of the attributes is defined as observable variables (transition), and the hidden variable represents the occupancy numbers (emission). The size of the hidden variable domain is 20, as shown in the number of bins adopted (see Table 2). The size of the observable variables domain is 2520 as there are 7 days per week, 5 weeks per month, 12 months per year and 6 terms considered for the period of 12-month data collection (7 × 5 × 12 × 6 = 2520). The matrix sizes of the transition and the emission were constructed to

Table 2
Dividing the occupancy numbers into 20 bins.

Bins	Range	Bins	Range
1	0–199	11	2000–2199
2	200–399	12	2200–2399
3	400–599	13	2400–2599
4	600–799	14	2600–2799
5	800–999	15	2800–2999
6	1000–1199	16	3000–3199
7	1200–1399	17	3200–3399
8	1400–1599	18	3400–3599
9	1600–1799	19	3600–3799
10	1800–1999	20	3800–3999

Table 3
The occupancy attributes inputs used for developing the scenarios.

Attribute	Scenario 1	Scenario 2	Scenario 3
Day [1]	✓	✓	✓
Week [2]	✓	✓	✓
Month [3]	✓		
Term [4]	✓	✓	✓
Outlier	✓	✓	

be 20 × 20 and 20 × 2520, respectively.

3.2. Scenario 2

The second scenario used selected occupancy attributes, including the day of the week, week of the month, and academic terms, for the aim of evaluating the model when developing more efficiently using optimal input data. A discrete dataset of 20 bins was used, as shown in Table 2. However, to avoid the potential problem of over-fitting that may cause by using all attributes as in Scenario 1, the month of the year attribute was eliminated from the training dataset. The other attributes (day of the week, week of the month, and academic terms) remained in the training dataset. The size of the hidden variable domain is 20, but the size of the observable variables domain is 210 as there are 7 days per week, 5 weeks per month, and 6 terms considered for the period of 12-month data collection (7 × 5 × 6 = 210). The size of the transition and emission matrix in Scenario 2 is 20 × 20 and 20 × 210, respectively.

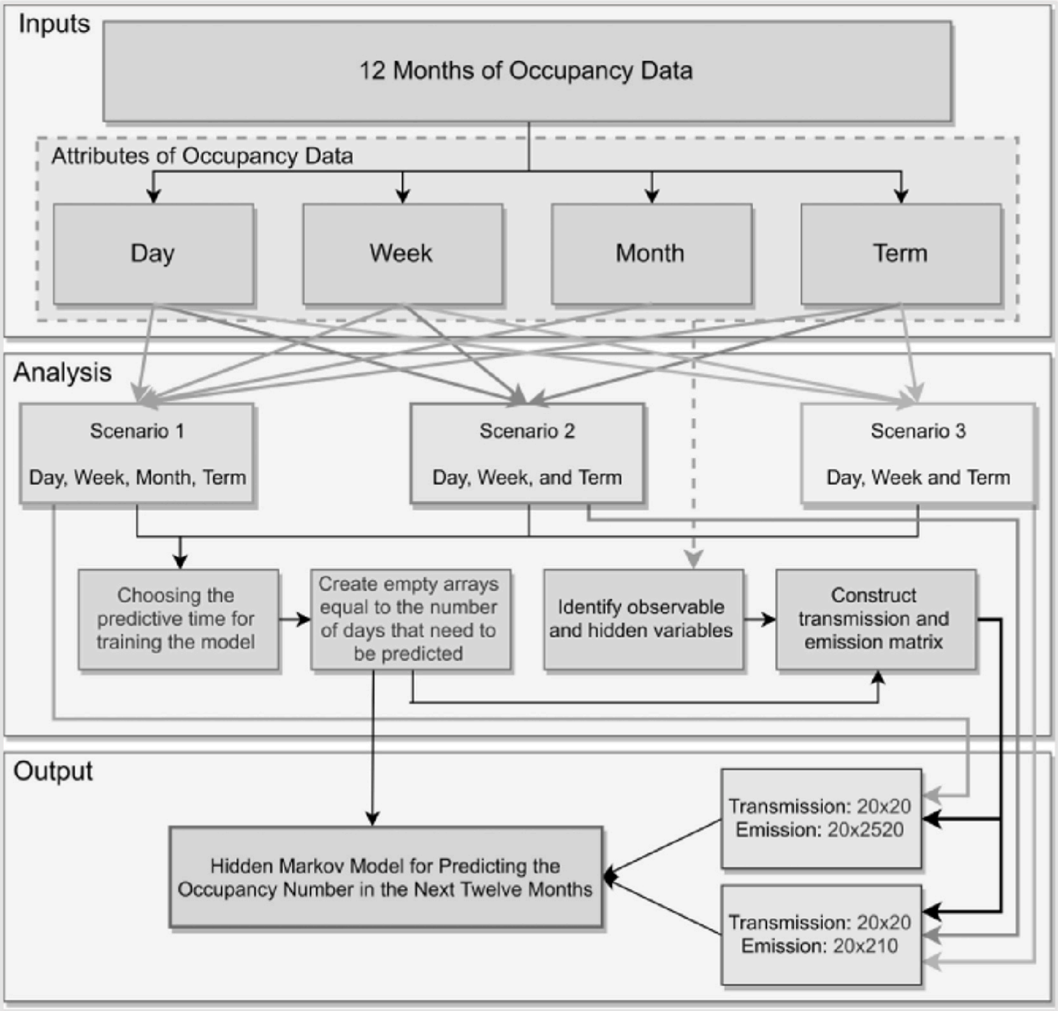
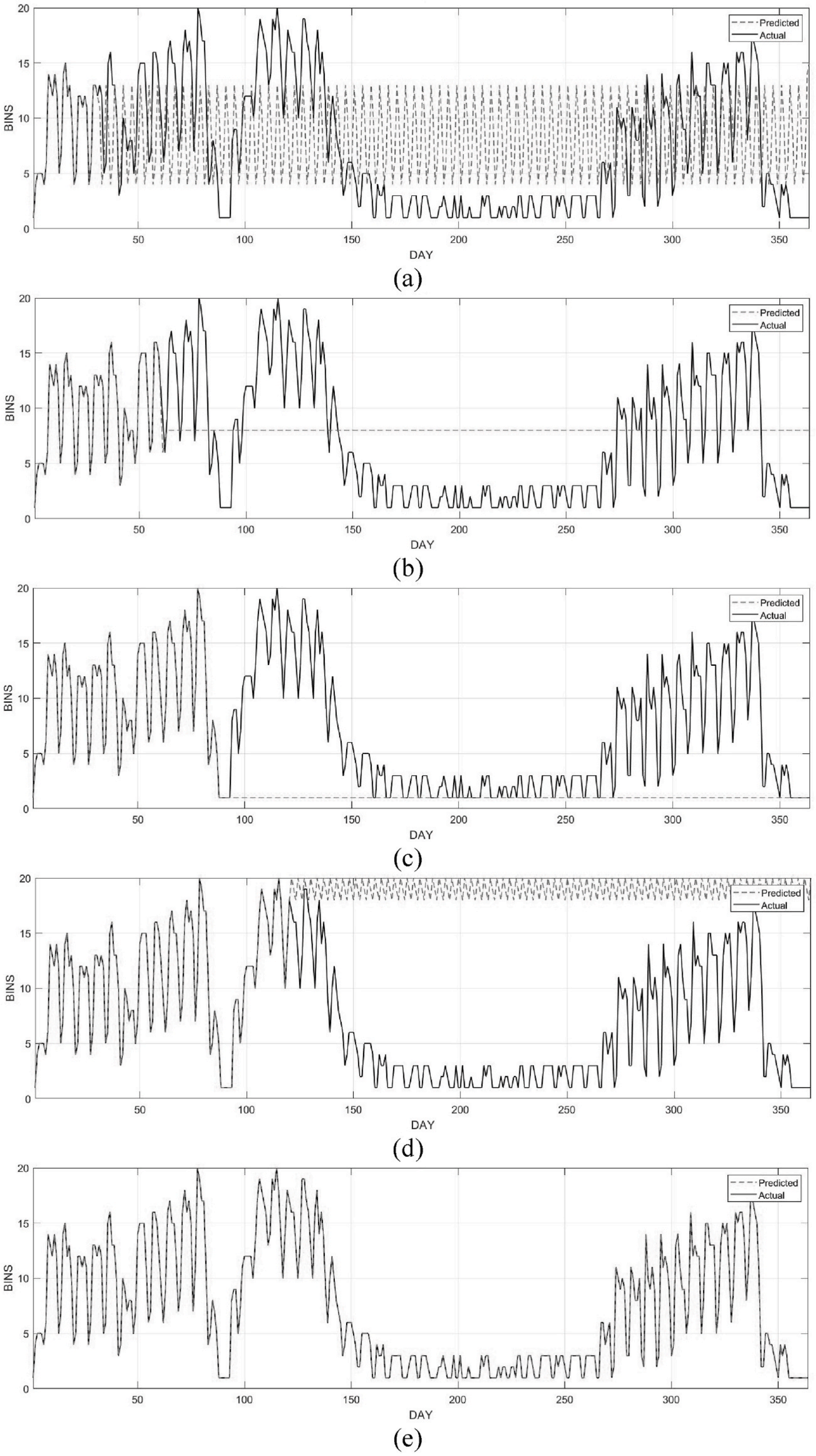

Inputs
12 Months of Occupancy Data
Attributes of Occupancy Data
Day
Week
Month
Term
Analysis
Scenario 1
Day, Week, Month, Term
Scenario 2
Day, Week, and Term
Scenario 3
Day, Week and Term
Choosing the predictive time for training the model
Create empty arrays equal to the number of days that need to be predicted
Identify observable and hidden variables
Construct transmission and emission matrix
Output
Hidden Markov Model for Predicting the Occupancy Number in the Next Twelve Months
Transmission: 20x20
Emission: 20x2520
Transmission: 20x20
Emission: 20x210

Fig. 5. The framework of developing HMM in the three different scenarios.



(a)

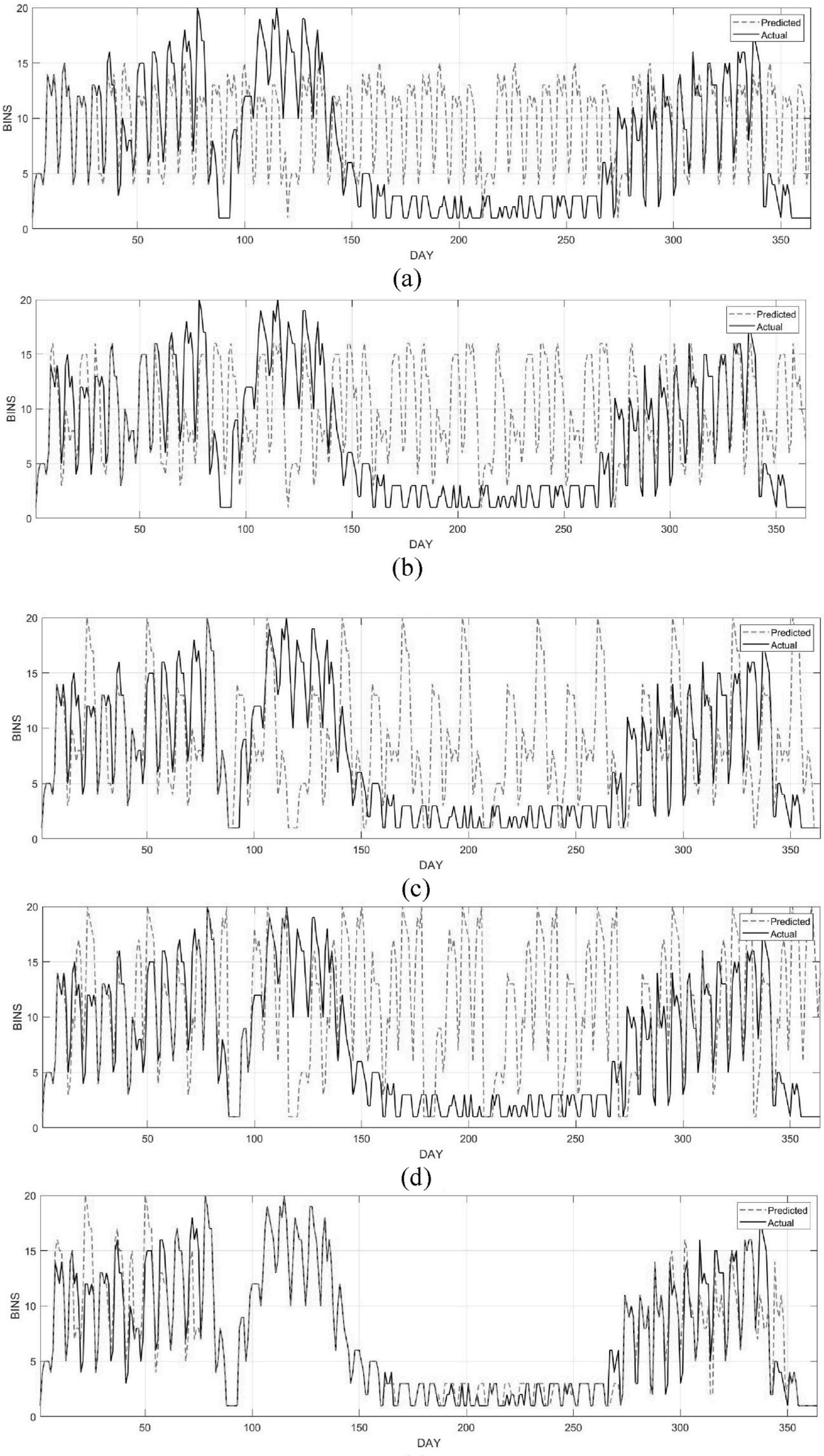
(b)

(c)

(d)

(e)

Fig. 6. The predicted results of Scenario 1 and actual data plotted against the days of the year, where the model was constructed using all attributes (Day of the week, week of the month, month of the year, and terms) for prediction using different months as input (a) one month, (b) two months, (c) three months, (d) four months, and (e) twelve months of occupancy data for training the model to predict 12 months.



(a)

(b)

(c)

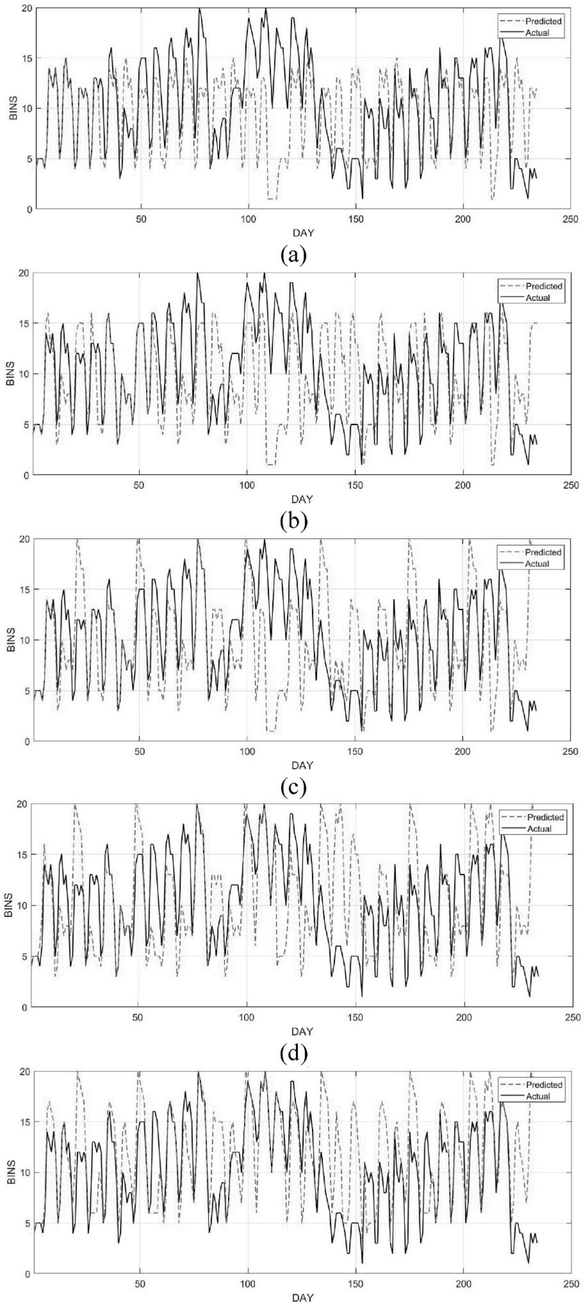
(d)

(e)

Fig. 7. The predicted result of Scenario 2 and actual data plotted against the days of the year where the model was constructed using three (Day of the week, week of the month, and terms) attributes, without the month of year attribute for prediction using different months as input (a) one month, (b) two months, (c) three months, (d) four months, and (e) twelve months of occupancy data for training the model to predict 12 months.

3.3. Scenario 3

The third scenario is meant to evaluate the claim in the body of the literature that removing outliers will enhance prediction accuracy. Therefore, Scenario 3 used the same attributes selected in Scenario 2, where the month of the year attribute was eliminated and excluded the data outliers. The domain sizes of the transition and emission matrix are 20 × 20 and 20 × 210, respectively. However, to avoid generalising the outliers that were identified using PDF and illustrated in Fig. 3, the identified outliers were excluded from the dataset to train the model in this scenario. The framework of the proposed model under three different scenarios illustrates the inputs and outputs of the model shown in Fig. 5.



(a)

(b)

(c)

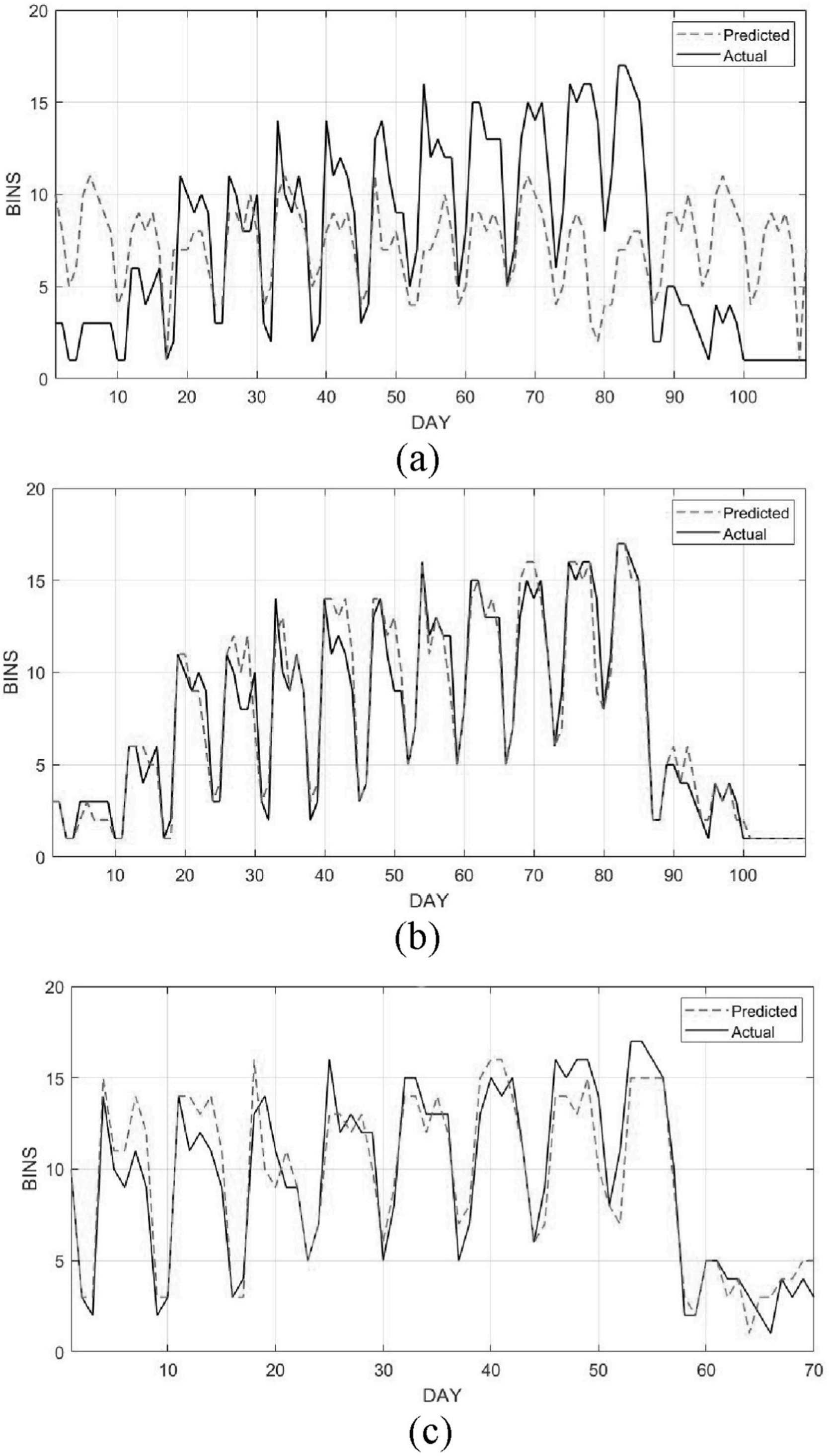
(d)

(e)

Fig. 8. The predicted results of Scenario 3 and actual data plotted against the days of the year where the model was constructed using three (Day of the week, week of the month, and terms) attributes, without the month of year attribute. It was also removing the outliers before developing the model for prediction using different months as input (a) one month, (b) two months, (c) three months, (d) four months, and (e) twelve months of occupancy data for training the model to predict 12 months.

4. Results and discussion

This section presents and discusses the results of the proposed HMM under three scenarios to determine the number of occupants in the case study building. The model used inputs of several occupancy attributes and 12 months of occupancy data for training. Cross-



(a)

(b)

(c)

Fig. 9. The usage of 70% of the data for training results of the three scenarios. Plot (a) represents Scenario 1, where four attributes are used for training the model [day, week, month, and term], (b) is Scenario 2, which was developed using three attributes [day, week, and term], and last plot (c) is Scenario 3, was developed using three attributes [day, week, and term] after removing the outlier data.

validation was used to determine the optimal split of the dataset for training and validation. The split datasets obtained from cross-validation were used to train the three scenarios. The prediction results for the three scenarios were evaluated by several statistical indicators. The idea behind training the three scenarios for 12 months was to understand the effect of different permutations of occupancy attributes and determine the one to be used for training the model to provide accurate occupancy perdition.

4.1. Scenario 1

In Scenario 1, the input data used in training the model includes day, week, month, and academic term (Table 1). Fig. 6 shows the results of the prediction using the occupancy data of 1, 2, 3, 4, and 12 months. Fig. 6a illustrates the result of using one month of occupancy data to predict the occupancy levels for 12 months. The results show that the model trained under this scenario was able to predict accurately only for the first month (the month that had been used for training the model). This pattern is observed in Fig. 6 b, c, and d, where the trained model is only able to predict the occupancy levels accurately for the period that the associated occupancy data for that period is used for training the model. The prediction of the occupancy for the rest of the month is either straight line as in Fig. 6 b and c or sinusoidal as in Fig. 6d. On the other hand, in Fig. 6e, the model trained with 12 months to predict the same period, which explains why the occupancy prediction results perfectly match the actual occupancy data. The error in the prediction model is zero, which in modelling terms, is referred to as an over-fitting problem. However, the zero-error prediction is desired in this case to ensure that the model can predict accurately and that there is no error in the algorithm used to develop the model.

4.2. Scenario 2

In Scenario 2, the model constructed with occupancy data and their associated occupancy attributes includes day, week, and academic terms that are presented in Table 3 and using the occupancy data of 1, 2, 3, 4, and 12 months as inputs for training the model. The result of the model is presented in five plots, as shown in Fig. 7. In this scenario, the month of year attribute is excluded from the data used for training the model. This is mainly due to avoid overtraining we observed for the results presented for Scenario 1, plot (e). From the result in Fig. 7, vulnerabilities in the model are caused by outliers. The model generalised the outliers in several days, such as the days 120, 211, and 274, as shown in plot (a). A sharp drop in the occupancy level occurred due to the 'reading break' on day 42, becoming a pattern in the predicted data. As a result, the same patterns are repeated periodically in Fig. 7a, b, c, and d. Fig. 7e represents the result of the model trained using 12 months of occupancy data. This plot shows the drop in the occupancy levels disappeared in the first nine months, which is providing better prediction compared to the other plots. Yet, the drops of occupancy that happened in the other plots occurred later in the prediction on months 10 and 11. Generally, Fig. 7a, b, c, and d show that the model could not differentiate between the academic terms and vacations and the predicted model handles the Summer Vacation that occurred between the days 150 and 270 as an academic term.

4.3. Scenario 3

In Scenario 3, the inputs data used in training the model include day, week, and academic term. Fig. 8 illustrates the outcomes of HMM for occupancy prediction after removing the outliers from the training dataset. The removed outliers (shown in Fig. 3) were identified by utilising the probability density function. Fig. 8a indicates that the model is able to identify the weekly variations throughout the prediction period. However, the model predicted a sharp drop on days 109 and 213, where the actual occupancy indicates a sharp rise. This error similarly occurred in Fig. 8 b, c, and d. The effect of such a generalisation became more prominent when the model is trained using 12 months, as shown in Fig. 6e. Although the major outliers had been removed, the sharp variations before and after a vacation can be observed as outliers. However, the result of the model trained using 12 months of data, Fig. 8e, was able to avoid the sharp drop in the prediction of occupancy levels for day 213.

4.4. Split the dataset

Cross validation (CV) was used to determine the optimal partitioning of the data to predict the occupancy accurately. According to Section 2.2.3, a range of permutations are considered to find the most appropriate split of actual occupancy data to be used for training and evaluation of the model while avoiding the overtraining phenomenon. The result of cross ventilation shows that 70% of the data for training and 30% of the data for validating is able to offer the highest accuracy scores of RMSE compared to other split arrangements that presented in Fig. 4. Fig. 9 shows the result of developing the models using 70% of the data for training and 30% for validation. It should be noted that Fig. 9 only includes the prediction of the occupancy for the months that are not used for training the model. Under scenario 1, Fig. 9a shows both an unreasonable fluctuation and no accuracy in the prediction of the actual occupancy. On the other hand, under scenario 2, Fig. 9b shows a good prediction result as it performed much better compared to scenario 1. Finally, under Scenario 3, the results of prediction model are presented in Fig. 9c. This scenario was developed to evaluate the claim made in previous studies that removing outliers of the data could enhance the accuracy of the model. In scenario 3, the outliers in the dataset identified by the PDF (Fig. 3) were excluded from the data used for training the model. The result of occupancy prediction in Scenario 3 (Fig. 9c) was not the most accurate one, as Scenario 2 (Fig. 9b) shows higher accuracy in prediction of occupancy. Furthermore, the trend of occupancy level in Scenario 3 (Fig. 9c) differs from that of Scenario 2 (Fig. 9b), indicating that removing outliers, including occupancy data associated with periods with low occupancy; for example, holidays, and summer vacations will not accurately reflect the reality. Hence, the developed models based on such set of inputs would not be able to predict the building occupancy accurately, particularly in high-density buildings with stochastic occupancy behaviours. Therefore, based on the results presented (Fig. 9), the claim made in previous studies regarding removing the outliers of input data to improve the accuracy of the model was found to be inaccurate. The accuracy of the proposed model under three scenarios are evaluated in the following section 4.5.

4.5. *Performance evaluation of models in prediction of occupancy*

The performance of the occupancy prediction models under three scenarios are evaluated using. Several comparison methods KL divergences and RMSE. This section first discusses the outcomes of the model when trained using 100% of the actual occupancy data, and then the results of models developed using 70% of occupancy data for training and 30% of the data for evaluation.

The performance of the HMMs developed under three scenarios using 100% of the actual occupancy data for training and validation is presented in Table 4. The outcomes of KL divergence and RMSE analyses for the model developed under Scenario 1 shows that the model is able to predict the actual occupancy of the building accurately. Although this is an ideal result (Fig. 6e), it is mainly due to the fact that the model is over-trained and predicts the results of the exact situation that it is trained for. However, the outcomes of the model developed under Scenario 2 shows a promising prediction (Fig. 7e) while avoiding the over-training phenomenon observed in Scenario 1. This has been achieved by eliminating the month of the year attribute associated with occupancy data from the process of training the model. On other hand, Scenario 3 has fewer inputs than the other two scenarios where the months of the year attribute has been eliminated, as well as the outliers (shown in Fig. 3) before training the model. Higher KL divergence and RMSE (Table 4) highlight that the model developed under this scenario is less able to predict the actual occupancy compared to those developed under the first two scenarios. The results of both KL Divergence and RMSE are in good agreement in evaluating all of the three scenarios that used different inputs. Such agreement strengthens the overall validity of the findings and shows that the results are robust and not influenced by any of the statistical indicators used.

In the next stage, the actual occupancy data is divided into 70%/30% for training and evaluation of the model, respectively. Table 5 shows the percentage and the number of the data points used in training and validating of the developed model under each scenario, as well as the RMSE results that indicate the accuracy of the model to predict the actual occupancy of the case study building. The data points in the table refer to the days of the year, which is a total of 365 days. In Scenarios 1 and 2, there are 255 data points/days, which is 70% of the entire data points over 365 days, while Scenario 3 used 164 data points for training the model. The small number of data points used in Scenario 3 is due to the fact that outliers are removed from training the model Fig. 9 c. The RMSE of the models developed under each scenario (Table 5) shows that the model developed under Scenario 2 is able to predict the actual occupancy of the building better than the models developed under the other two scenarios. Among the three models, the least effective model was the one developed under Scenario 1.

5. Conclusion

A hidden Markov model was developed under three different scenarios. The scenarios were developed using measured data collected from a cases study building with high occupancy density using infrared video camera sensors. Several sets of attributes associated with the occupancy data were examined to develop a data driven HMM. These occupancy attributes include the day of the week, the week of the month, the month of the year, and academic terms that are outlined in Table 1. HMMs were developed under three different scenarios (Table 3) where each scenario contains certain attributes of actual occupancy data for training the model. In Scenario 1, the HMM was developed and trained using all the identified attributes of the actual occupancy data. In Scenario 2, one of the attributes, the month of the year, was eliminated to avoid over-training the HMM. Finally, in Scenario 3 the month of the year attribute and the outliers identified in the occupancy data were removed from the dataset for training the HMM. The cross-validation analyses conducted for all scenarios showed that using 70%/30% of occupancy data for training/validation of the model can result in the most accurate prediction of occupancy in the case study building. Comparing the performance of the developed HMMs under each scenario revealed that the model developed under Scenario 2 was able to provide a reasonably accurate prediction for the actual occupancy of the case study building. This has been achieved by eliminating the occupancy attribute responsible for over-training of the model under Scenario 1. However, the zero-error in Scenario 1 that used all the occupancy attributes (Table 3) was a desire to ensure the ability of the model to predict accurately. In addition, Scenario 3 was developed to evaluate the claim made in previous studies that removing data outliers could enhance the accuracy of the model. According to the evaluation of the prediction model under Scenario 3, the claim proves to be inaccurate.

The ability of the model to predict the occupancy levels with high accuracy allows it to be applied to various applications, which can contribute to better management and more efficient operation of building systems. In addition, in applications such as building safety operations, it is necessary to know the number of occupants to design or reconsider fire safety measures, for example, building evacuation. These applications illustrate the importance of collecting occupancy data for high-density higher education buildings that have different characteristics than other buildings, such as office buildings. Office buildings have a stable rate of occupants over time, known occupants or employees, and fixed schedule. However, in library buildings, occupancy patterns change based on different factors, including academic terms and vacations, occupants' presence with no restrictions or a schedule. Therefore, collecting occupancy data for an extended period was needed in this study. Although long-term data collection from the entire case study building with a high occupancy density is one of the unique strengths of this research, this study focuses on only one case study building. The next step of this study is to investigate how the proposed model can predict in using a shorter period of occupancy data inputs, such as a month of an academic term, to predict the occupancy numbers for the remaining period of that academic term. Also, to evaluate the generalizability of the developed model by testing it on a new dataset and different settings from the one used in training. This will benefit the facility managers who have a short period of occupancy data in the building.

Declaration of competing interest

Authors would like to confirm no conflict of interest in this study.

Table 4
Prediction accuracy of different scenarios in developing HMM model using 100% data training.

Indicators	Results		
	Scenario 1	Scenario 2	Scenario 3
KL divergence	0.00	0.00	0.026
RMSE	0.00	420	953

Table 5
Specifying the data points and the percentage used for training and validating the three scenarios.

SCENARIO	TRAINING		VALIDATING		Accuracy - RMSE
	Data Points	%	Data Points	%	
Scenario 1	255	70%	109	30%	998
Scenario 2	255	70%	109	30%	260
Scenario 3	164	70%	70	30%	319

Data availability

Data will be made available on request.

References

[1] O.T. Masoso, L.J. Grobler, The dark side of occupants' behaviour on building energy use, Energy Build. 42 (2) (2010) 173–177.

[2] K. Anderson, K. Song, S. Lee, H. Lee, M. Park, Energy consumption in households while unoccupied: evidence from dormitories, Energy Build. 87 (2015) 335–341.

[3] K. Song, Y. Jang, M. Park, H.S. Lee, J. Ahn, Energy efficiency of end-user groups for personalised HVAC control in multi-zone buildings, Energy 206 (2020), 118116.

[4] 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.

[5] B. Ai, Z. Fan, R.X. Gao, Occupancy estimation for smart buildings by an auto-regressive hidden Markov model, in: 2014 American Control Conference, IEEE, 2014, June, pp. 2234–2239.

[6] A.A. Adamopoulou, A.M. Tryferidis, D.K. Tzovaras, A context-aware method for building occupancy prediction, Energy Build. 110 (2016) 229–244.

[7] S.H. Ryu, H.J. Moon, Development of an occupancy prediction model using indoor environmental data based on machine learning techniques, Build. Environ. vol. 107 (2016) 1–9.

[8] T. Zia, D. Bruckner, A. Zaidi, A hidden Markov model based procedure for identifying household electric loads, in: IECON 2011-37th Annual Conference of the IEEE Industrial Electronics Society, IEEE, 2011, November, pp. 3218–3223.

[9] P. Liu, S.K. Nguang, A. Partridge, Occupancy inference using pyroelectric infrared sensors through hidden Markov models, IEEE Sensor. J. 16 (4) (2015) 1062–1068.

[10] P. Hu, Z. Zhou, Q. Liu, F. Li, The HMM-based modeling for the energy level prediction in wireless sensor networks, in: 2007 2nd IEEE Conference on Industrial Electronics and Applications, IEEE, 2007, May, pp. 2253–2258.

[11] K.P. Lam, M. Höynck, B. Dong, B. Andrews, Y.S. Chiou, R. Zhang, D. Benitez, J. Choi, Occupancy detection through an extensive environmental sensor network in an open-plan office building, IBPSA Build. Simulat. 145 (2009) 1452–1459.

[12] J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, K. Whitehouse, The smart thermostat: using occupancy sensors to save energy in homes, in: Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems, 2010, November, pp. 211–224.

[13] J. Chaney, E.H. Owens, A.D. Peacock, An evidence based approach to determining residential occupancy and its role in demand response management, Energy Build. 125 (2016) 254–266.

[14] J. Virote, R. Neves-Silva, Stochastic models for building energy prediction based on occupant behavior assessment, Energy Build. 53 (2012) 183–193.

[15] B. Huchuk, S. Sanner, W. O'Brien, Comparison of machine learning models for occupancy prediction in residential buildings using connected thermostat data, Build. Environ. 160 (2019), 106177.

[16] E. Mocanu, P.H. Nguyen, M. Gibescu, W.L. Kling, Comparison of machine learning methods for estimating energy consumption in buildings, in: 2014 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), IEEE, 2014, July, pp. 1–6.

[17] I. Ullah, R. Ahmad, D. Kim, A prediction mechanism of energy consumption in residential buildings using hidden markov model, Energies 11 (2) (2018) 358.

[18] M.S. Zuraimi, A. Pantazaras, K.A. Chaturvedi, J.J. Yang, K.W. Tham, S.E. Lee, Predicting occupancy counts using physical and statistical Co2-based modeling methodologies, Build. Environ. 123 (2017) 517–528.

[19] T. Ekwevugbe, N. Brown, V. Pakka, D. Fan, Real-time building occupancy sensing using neural-network based sensor network, in: 2013 7th IEEE International Conference on Digital Ecosystems and Technologies (DEST), IEEE, 2013, July, pp. 114–119.

[20] K. Li, X. Xie, W. Xue, X. Dai, X. Chen, X. Yang, A hybrid teaching-learning artificial neural network for building electrical energy consumption prediction, Energy Build. 174 (2018) 323–334.

[21] A. Ashouri, G.R. Newsham, Z. Shi, H.B. Gunay, Day-ahead prediction of building occupancy using wi-fi signals, in: 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), IEEE, 2019, August, pp. 1237–1242.

[22] R.F. Mustapa, N.Y. Dahlan, A.I.M. Yassin, A.H.M. Nordin, Quantification of energy savings from an awareness program using NARX-ANN in an educational building, Energy Build. 215 (2020), 109899.

[23] T.G. Özbalta, A. Sezer, Y. Yıldız, Models for prediction of daily mean indoor temperature and relative humidity: education building in Izmir, Turkey, Indoor Built Environ. 21 (6) (2012) 772–781.

[24] A.G. Alam, H. Rahman, J.K. Kim, H. Han, Uncertainties in neural network model based on carbon dioxide concentration for occupancy estimation, J. Mech. Sci. Technol. 31 (5) (2017) 2573–2580.

[25] S.A. Sharif, A. Hammad, Developing surrogate ANN for selecting near-optimal building energy renovation methods considering energy consumption, LCC and LCA, J. Build. Eng. 25 (2019), 100790.

[26] W. Wang, J. Chen, T. Hong, Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings, Autom. ConStruct. 94 (2018) 233–243.

[27] X. Xu, W. Wang, T. Hong, J. Chen, Incorporating machine learning with building network analysis to predict multi-building energy use, Energy Build. 186 (2019) 80–97.
[28] M.K. Kim, Y.S. Kim, J. Srebric, Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: artificial neural network vs. linear regression, Sustain. Cities Soc. 62 (2020), 102385.
[29] S.S. Kwok, R.K. Yuen, E.W. Lee, An intelligent approach to assessing the effect of building occupancy on building cooling load prediction, Build. Environ. 46 (8) (2011) 1681–1690.
[30] 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 (7) (2010) 1038–1046.
[31] B.W. Hobson, D. Lowcay, H.B. Gunay, A. Ashouri, G.R. Newsham, Opportunistic occupancy-count estimation using sensor fusion: A case study, Build. Environ. vol. 159 (2019), 106154.
[32] V.L. Erickson, M.A. Carreira-Perpinán, A.E. Cerpa, Occupancy modeling and prediction for building energy management, ACM Trans. Sens. Netw. 10 (3) (2014) 1–28.
[33] E. Longo, A.E. Redondi, M. Cesana, Accurate occupancy estimation with Wi-Fi and bluetooth/BLE packet capture, Comput. Network. 163 (2019), 106876.
[34] S. Depatla, A. Muralidharan, Y. Mostofi, Occupancy estimation using only Wi-Fi power measurements, IEEE J. Sel. Area. Commun. 33 (7) (2015) 1381–1393.
[35] Y. Wang, L. Shao, Understanding occupancy and user behaviour through Wi-Fi-based indoor positioning, Build. Res. Inf. 46 (7) (2018) 725–737.
[36] L.E. Baum, J.A. Eagon, An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology, Bull. Am. Math. Soc. 73 (3) (1967) 360–363.
[37] M. Momenzadeh, M. Sehhati, H. Rabbani, A novel feature selection method for microarray data classification based on hidden Markov model, J. Biomed. Inf. 95 (2019), 103213.
[38] B. Alfalah, M. Shahrestani, L. Shao, Identifying occupancy patterns and profiles in higher education institution buildings with high occupancy density–A case study, Intell. Build. Int. (2022) 1–17.
[39] R. Melfi, B. Rosenblum, B. Nordman, K. Christensen, Measuring building occupancy using existing network infrastructure, in: 2011 International Green Computing Conference and Workshops, IEEE, 2011, July, pp. 1–8.
[40] ASHRAE ASHRAE Standard Standard 90.1, Energy Standard for Buildings except Low Rise Residential Buildings, American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2004.
[41] M.B. Hooten, N.T. Hobbs, A guide to Bayesian model selection for ecologists, Ecol. Monogr. 85 (1) (2015) 3–28.
[42] A. Gelman, J. Hwang, A. Vehtari, Understanding predictive information criteria for Bayesian models, Stat. Comput. 24 (6) (2014) 997–1016.
[43] O.M. Abusaid, F.M. Salem, Kullback-leibler divergence minimisation for competitive learning of self-organising maps, in: 2017 International Conference on Engineering and Technology (ICET), IEEE, 2017, August, pp. 1–6.
[44] C. Reyes, T. Hilaire, S. Paul, C.F. Mecklenbräuker, Evaluation of the root mean square error performance of the PAST-consensus algorithm, in: 2010 International ITG Workshop on Smart Antennas (WSA), IEEE, 2010, February, pp. 156–160.
[45] C.M. Bishop, N.M. Nasrabadi, in: Pattern Recognition and Machine Learning, vol. 4, springer, New York, 2006, p. 738, 4.