Stochastic-based Occupant-Centric Building Archetype Modelling Using Plug Loads
Sanam Dabirian¹, Negarsadat Rahimi¹, Ursula Ecker¹,²
¹Concordia University, Montreal, Canada
²Canada Excellence Research Chair, Montreal, Canada

Abstract
Occupant-related parameters significantly impact the urban building energy simulation uncertainty. In most existing urban-scale building energy models, fixed default occupant-related schedules are typically used, which might not necessarily capture the variation associated with occupancy. Without a more complex occupancy model within urban building energy modeling (UBEM), it is impossible to achieve a reliable energy demand estimation. For a more robust output from UBEM, occupant-related schedules should reflect the variability and diversity of the occupant behavior. This paper proposes a framework to extract representative occupant-related profiles from time-series data and model their performance considering the stochastic nature of occupant behavior.

Highlights
- Extract representative occupant-related schedules from time-series electricity consumption data
- Improve the archetype modeling by stochastic occupant-related schedules
- Reduce UBEM uncertainty

Introduction
As part of Green Economy action plans, many countries aim to become zero emission by 2050. As an intermediary goal, Canada aims to achieve a 37.5% reduction in greenhouse gas emissions by 2030 compared to 1990 (International Energy Agency 2021). The built environment has a high potential for reducing energy demand during this period. The use of urban building energy modeling (UBEM) can be a useful tool to design emission reduction scenarios and for assisting the management of building energy on the urban scale. UBEM uncertainty is significantly influenced by occupant-related parameters such as presence and their interaction with energy systems (Hong et al. 2017)(Hong et al. 2020). In most existing urban-scale building energy models, fixed default occupant-related schedules are usually used, which may not necessarily reflect occupant behavior variations. The main reason is the lack of data available to model dynamic occupancy schedules, which leads to differences between energy simulation results and the actual data. UBEM cannot accurately estimate energy demand and predict peak loads without a more complex occupancy model. UBEM requires occupant-related schedules that take into account the variability and diversity of occupant behavior in order to produce a more robust output from building energy simulation.

Archetype modeling is a common method to classify the building stock on the city level, evaluate city building energy performance, and support energy policies, building codes, new technologies, and retrofitting strategies and programs (Delmastro, Mutani, and Corgnati 2016)(Cerezo Davila, Reinhart, and Bemis 2016)(Sokol, Cerezo Davila, and Reinhart 2017)(Nouvel et al. 2017). Archetypes are the simplified definition of building characteristics, including geometry, energy systems, and occupant-related parameters. The archetype-based approach is classified into deterministic and stochastic (Lim and Zhai 2017).

In archetyping, occupant-related schedules are assumed to follow a deterministic pattern that significantly impacts energy use. Considering occupants' critical roles in a building's energy use and management, stochastic occupant-centric archetypes can better simulate district demand by incorporating the variability and stochasticity of their schedules. A more realistic district load curve can be obtained if stochastic occupant-related profiles are correctly modeled. Previous research on stochastic occupant-related schedules can only be used in building energy simulation of specific buildings, such as office and residential, but not for all building types in mixed-use districts (Dabirian et al. 2021). Furthermore, most studies have taken into account only the stochastic variations of the physics characteristics of the building involved in energy modeling, not the behavior of the occupants.

Thus, this paper outlines a framework to extract the representative occupant-related profiles from time-series data for mixed-use neighborhoods and model their performance considering the stochastic nature of occupant behavior. Also, it could be demonstrated how the stochastic-based occupant-related archetypes improve the urban building energy modeling workflow to predict demand. Using time-series data as the basis of the model makes it possible to achieve a more accurate representation of the energy demand. This dynamic model could provide relatively accurate simulation results and pave the way to identify appropriate energy management strategies.

Besides, applying stochastic-based schedules can include the variability of the occupant behavior in an urban model where similar archetypes dominate within a neighborhood. Overall, the proposed framework...
integrates flexible and reliable occupant-centric archetypes and energy demand analysis, including forecasting the impacts of the variability of occupant behavior to establish an informed basis for energy-efficient strategies and demand-side energy management.

**Background**

**Occupant-related data-driven modelling**

The data-driven modeling analyzes the data to extract useful information and patterns using data mining techniques, machine learning models, and statistical methods (Fan et al. 2016). Occupant-related data-driven models are developed using robust data collected from buildings from different sources. The occupant-related patterns could be used to calculate the internal gains and probability of the occupancy presence in the buildings. Several data-driven occupancy models have been proposed to simulate the building performance at the building scale (Causone et al. 2015). Fu et al. (Fu et al. 2021) explored the chilled water consumption data as an occupancy proxy to model the building cooling energy consumption. (Happle et al. 2020) created occupancy presence schedules in 13 different U.S. cities based on location-based data such as Google Maps or Facebook. Several research works investigated the electrical meter data to extract energy use patterns (Miller and Meggers 2017). However, few studies have used time-series electricity consumption data to extract occupant-related parameters for urban energy modeling.

**Clustering**

Clustering is a data mining approach to group a dataset into N clusters based on pattern similarity. Clustering time-series data facilitates extracting valuable information from a massive and complex dataset when using supervised classification proves to be computationally difficult (Aghabozorgi et al. 2015). K-means clustering is the most common unsupervised distance-based algorithm in data-driven modeling of occupancy and archetype (Ali et al. 2019). In the K-means method, each cluster is represented by the mean of each cluster called the centroid. To initialize the optimal number of clusters, different indices such as Elbow point, Davies-Bouldin, Dunn, and Silhouette are used (Fan et al. 2021).

**Stochastic modelling in UBEM**

Stochastic occupant-related schedules provide a more accurate representation of occupant behavior variability within a building or district (Romero 2020). The majority of stochastic occupancy models are based on statistical methods (probabilistic/stochastic) such as Markov Chain processes (Wang et al. 2011). However, in some studies Monte Carlo (Sokol et al. 2017), (Li et al. 2019), and Markov chain Monte Carlo (MCMC) method (Wang and Ding 2015) have been used. (Sokol et al. 2017) proposed a model to provide the UBEM input parameters. Where a uniform distribution was fitted to occupant density, plug load, lighting power density, thermostat setpoints, and the DHW flow rate. (Schiefelbein et al. 2015), in the TEASER tool, predicted the annual occupancy, appliance, and lighting profiles using the Richardson lighting use model (Markov Chain method) to simulate district buildings' thermal demand (Richardson et al. 2008).

In this study, MCMC method is used to develop stochastic model. MCMC is applied to study the approximate distribution and obtain the fitting curve. MCMC algorithm consists of two modules: Markov-Chain and Monte Carlo techniques. In MCMC model, the data is generated by sampling from a probability distribution (PD). In this method, the random values are repeatedly drawn from the assumed parameters to converge to the real PD. The sample values are selected randomly, while the probability of the current state only depends on the previous state.

**Methodology**

The methodology steps are shown in Figure 1. It elucidates the step-by-step extraction of the representative patterns of the occupant-related schedules (e.g., plug loads) from electricity consumption profiles. The process allows the users to input various time-series data and obtain representative profiles for building performance simulations. Data-driven modeling of plug loads and the developing a stochastic model are the steps involved in the methodology for generating the schedules for plug loads for various types of buildings in various climate zones.

![Figure 1. Methodology overview](https://doi.org/10.26868/25222708.2023.1381)

In data-driven modeling, data preprocessing is the fundamental step to understanding and cleaning the dataset. The raw data collected from different sources may contain inconsistencies, errors, and noises. Thus, data pre-processing provides uniformity and improves the data quality used as input for subsequent steps in the methodology. In this study, the data preprocessing phase includes 1) cleaning and imputing the missing data; 2) dimension reduction; 3) data scaling; 4) feature creation/selection; 5) merging the data from different sources; and 6) partitioning the data into different groups. To extract the patterns and representative profiles k-means clustering is utilized. The objective of the
clustering algorithm is to minimize the sum of the Euclidean distances between each data point and their respective centroids. The outputs of clustering are the values of the cluster centroids representing the occupant-related characteristics and categorization of the profiles based on proximity/similarity to the centroids for each building type within a climate zone.

In addition to extracting the representative plug load profiles, the clustering output is also used to calculate the transition matrix that is required for the MCMC model development. The transition matrix (TM) of the plug load schedules is described as below in which \( \cdot \cdot \cdot , x_i-3, x_i-2, x_i-1 \) are the Markov chain states.

\[
P(x_i \cdot \cdot \cdot , x_i-3, x_i-2, x_i-1) = P(x_i|x_{i-1})
\]

(1)

Where \( k \) is the number of clusters, and the TM size would be \( k^2 \). In this paper, TM elements indicate the probability of the transition from one cluster to another.

The MCMC algorithm is applied on each hour of data of the plug loads. The fitting function is trained with the parameters, including load and time. Firstly, different PDs are tested on the hourly data (e.g., all the records of 8:00 AM) to find the appropriate distribution that fits the data (Figure 8). In order to determine how well sample data, fit the PD, various goodness-of-fit tests and metrics such as Kolmogorov-Smirnov, Akaike information criterion (AIC), and Bayesian information criterion (BIC) are employed. The hourly plug loads are sampled for each hour from their PDs to create a daily profile in the next step.

Random sampling is implemented using pseudo-random generators available in Python. The initial value of the parameters and the number of samples are determined by the user based on the data observation. The next hour value is extracted from the related PD considering the Markov Chain transition probabilities. It means the extracted value is selected considering the previous value drawn from which cluster. The stochastic values are calculated by multiplying the TM by each sample drawn from the PDs. Then, the procedure will repeat to generate other hours of data to complete the daily plug load profile.

In order to evaluate the performance of the developed model, the Root Mean Square Error (RMSE) and R-squared (from equations (2 and 3) are used as sensitivity analysis methods. To implement the sensitivity analysis, the generated stochastic-based plug load profiles for five selective consecutive weekdays has been considered.

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}
\]

(2)

\[
R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}
\]

(3)

The RMSE indicates how accurately the generated stochastic-based plug load profiles represented the actual plug load data. \( R^2 \) is a statistical measure that indicates the proportion of the variance in the dependent variable (plug load) that can be explained by the independent variable (generated profiles).

**Results and discussion**

In this study, the methodology is applied to office buildings in climate zone 5A of the Building Data Genome Project 2.

**Data pre-processing**

To develop the data-driven occupant-centric model, an open data source, Building Data Genome Project 2 (BDGP2) (Miller et al. 2020), has been used. The data was collected from 3053 energy meters (electricity, chilled water, hot water, gas, water) from 19 different locations (1,636 non-residential buildings) over Europe and North America. The dataset mainly belongs to institutional campuses, including various building types such as offices, laboratories, classrooms, lodging, and library.

The dataset is for the years 2016 and 2017 with hourly time resolution. The metadata includes the gross floor area, construction year, building coordinates, and building primary use and sub-primary use types. The existing data has been cleaned, and the measurement units for the different energy meters have been standardized. The data processing is implemented in Python using different libraries such as Pandas, Numpy, Matplotlib, Sklearn, and Seaborn.

In order to dimension reduction, two features were calculated: the specific electricity power usage (SEPU) and the electricity usage average. The electricity usage average is calculated by dividing the total annual electricity consumption by 8760. SEPU is obtained by dividing the electricity usage average by the total floor area.

The electricity consumption metered in the BDGP2 dataset is assumed to be representative and acts as a proxy for occupancy. Thus, this study uses electricity consumption to extract the plug load. Since there is no feature to illustrate the electricity consumption components (e.g., plug loads, lighting, and HVAC usage), two conditions have been considered to assume that the electricity consumption values could be regarded as plug loads. By checking the metadata, the building has been selected when it has metered electricity, hot water, and chilled water. It means that the hot and chilled water is used to provide heating and cooling, respectively; thus, electricity mainly represents plug loads. Also, the SEPU is in the range of the ASHRAE receptacle power density.
(e.g., for office buildings, 10 W/m²). The buildings selected have a SEPU between 0 to 20 W/m².

After reducing the building data related to those buildings which do not fulfill the mentioned conditions, the missing hourly values are filled using the mean imputation technique by the average of two previous values. In order to eliminate the impact of the gross floor area of the building on the analysis process, the plug load values are normalized. In the next step, a few features, such as time components, weekday/weekend, and seasons are created to prepare the raw data into a suitable format for occupancy modeling. Moreover, visualization provides a straightforward understanding of pattern recognition. The data is visualized in daily, weekly, monthly, and seasonally format. Graphs are used to describe and explore the data. The result of this step is weekday/weekend daily profiles of electricity consumption for various sites (in different climate zones) and each building types.

Table 1. Calculated features of considered buildings

<table>
<thead>
<tr>
<th>Building ID</th>
<th>Floor Area (m²)</th>
<th>Annual Energy Consumption (kWh/m²)</th>
<th>Electricity Usage Average (kWh/m²)</th>
<th>SEPU (W/m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peacock office Norm</td>
<td>3043.1</td>
<td>110.15</td>
<td>0.0126</td>
<td>12.57</td>
</tr>
<tr>
<td>Peacock office Julia</td>
<td>3913.9</td>
<td>61.11</td>
<td>0.007</td>
<td>6.98</td>
</tr>
<tr>
<td>Peacock office Annie</td>
<td>6021.8</td>
<td>114.07</td>
<td>0.013</td>
<td>13.02</td>
</tr>
<tr>
<td>Peacock office Elton</td>
<td>9901.9</td>
<td>148.55</td>
<td>0.017</td>
<td>16.06</td>
</tr>
<tr>
<td>Peacock office Edie</td>
<td>9723.9</td>
<td>79.15</td>
<td>0.009</td>
<td>9.04</td>
</tr>
<tr>
<td>Wolf office Joan</td>
<td>1355</td>
<td>101.22</td>
<td>0.0116</td>
<td>11.56</td>
</tr>
<tr>
<td>Wolf office Durham</td>
<td>5036</td>
<td>78.24</td>
<td>0.0089</td>
<td>8.93</td>
</tr>
<tr>
<td>Wolf office Rochelle</td>
<td>3115</td>
<td>45.75</td>
<td>0.0052</td>
<td>5.22</td>
</tr>
<tr>
<td>Wolf office Bobbie</td>
<td>4601</td>
<td>51.19</td>
<td>0.0058</td>
<td>5.84</td>
</tr>
<tr>
<td>Wolf office Fanuel</td>
<td>2005</td>
<td>60.87</td>
<td>0.0076</td>
<td>7.63</td>
</tr>
<tr>
<td>Wolf office Hindee</td>
<td>1129</td>
<td>68.98</td>
<td>0.0079</td>
<td>7.87</td>
</tr>
</tbody>
</table>

Although the framework was applied to most of the building types and climate zones in BDGP2, the office building in climate zone 5A has been selected to illustrate the applicability of the proposed framework among the available building types and climate zones in BDGP2. Among all 21 office buildings in climate zone 5A, 11 buildings fulfilled the model assumptions. Table 1 shows the selected buildings of the case study.

After pre-processing of the data based on the approach mentioned in Data pre-processing section, the daily profiles of the buildings in Table 1 were visualized in Figure 2. The graphs contain all of the daily profiles, including those for weekdays and weekends. Figure 10 shows a wide variety of patterns and load quantities in different buildings. It means that occupant behaviour and schedules significantly affect electricity consumption profiles (pattern and quantity) even in buildings with similar types in the same climate zone. However, there is a prominent pattern in most weekday profiles which has a rise in the morning and a decrement in the late afternoon.

Data-driven model

The daily time-series weekday and weekend profiles of all office buildings in climate zone 5A are merged in a data frame. K-means clustering, as the most common unsupervised distance-based algorithm, is considered in the workflow to explore distinct profiles.

To implement k-means clustering, the optimal number of clusters (k) as the main hyperparameter is defined based on the Silhouette score method. Figure 3 indicates an example of the Silhouette score applied to the weekday data of the case study. The optimal number of clusters is five, based on the graph. In addition to using the Silhouette score method, profile visualization can help the expert to determine the k to cover the most patterns. In the next step, the profiles are grouped into k clusters. The result of the clustering is shown in a data frame containing useful information such as the daily profiles, timestamp, name of the building, and assigned cluster. After preparing the data frame, post-processing is implemented to visualize the clusters and centroids.
Figure 3. Silhouette score analysis of weekday plug load profiles of case study

Figure 4 and Figure 5 are the visualizations of the clustered weekday and weekend plug load profiles, respectively. They show the representative plug load profiles of the case study. In these figures, the dashed curves indicate the clusters’ centroids. Color-coded clusters can be seen in the background as faded profiles. According to the Silhouette score analysis, four clusters are the optimal number of clusters. However, visualizing the profiles shows that having five clusters can better group the profiles.

Figure 6. Distribution of weekend plug load profiles in each cluster in different buildings

Figure 7. Distribution of weekday plug load profiles in each cluster in different buildings

In the next step, the TM is calculated using the clustering result described in the methodology section.

**MCMC model**

In order to apply the MCMC method to the data, TM must be calculated first. As mentioned previously, the TM elements represent the probability of transitioning from one cluster to another. To calculate the TM of the weekday profiles, the clustered weekday profiles were utilized. As the considered k for the weekday’s clustering is 5, a 5*5 matrix is calculated as follows.
In this study, electricity usage profiles of several buildings (with similar types) were analyzed, and they vary widely from a very high value to a very low value. In addition, since Monte Carlo sampling is performed from such a wide range, the Markov chain method maintains closer and more realistic ranges between two consecutive hourly values generated. However, it is still a stochastic model, and some variations will remain. The generated data would constantly shape the ever-changing occupant-related schedules and match them to govern building systems to actual human patterns. As a result of the proposed model, medium and long-term plug load predictions can be made. During each iteration, the model provides different values that can be applied to similar archetypes in different districts.

The MCMC analysis is applied to the plug load values for each hour, which contain the values of all buildings at that time. The developed model at this stage tests all potential PDs on the hourly values. Figure 8 presents a try of testing different PDs on one hour plug load data to find the best-fitted PD. To determine the best-fitted PD, different goodness of fit tests was conducted in addition to visualizing the PDs on the histogram.

In this study, Kolmogorov-Smirnov (P-value), Anderson-Darling, AIC, and BIC tests were used to determine the best-fitted PD for each hour. Figure 9 illustrates the best-fitted PDs on the hourly plug load profiles of the weekdays in case study buildings. Following the determination of the PDs, the sampling procedure was implemented using the pseudo-random package available in Python. The pseudo-random sampling creates the opportunity to generate a new set of data in each iteration which could be applied to simulation models for a building or the archetypes in a district. Figure 10 shows the stochastic-based plug load profiles for five selective consecutive weekdays generated by the proposed model. The dotted curve shows the average of the actual data of a similar time period.

The sensitivity analysis illustrates that the RMSE of the generated stochastic-based profiles are between 1.15 to 1.45 and the $R^2$ is between 0.7 to 0.82. The range of 1.15 to 1.45 of RMSE suggests that the generated profiles had some level of error in predicting the actual plug load values, but the error was within a reasonable range. Also, the range of $R^2$ between 0.7 to 0.82 suggests that the generated profiles were able to explain approximately
70% to 82% of the variance in the actual plug load data. This indicates a moderate to strong level of correlation between the generated profiles and the actual plug load values. However, a potential weakness of this method is that the dataset used to create the simulations was based on a year of data, so there was a wide range between the minimums and maxima. Consequently, there is a risk of prominent peaks and valleys forming between two hours in the plug load profiles. As a result, selecting appropriate sub-data (e.g., seasonal data) for the Monte Carlo method input from the datasets becomes very important. For creating stochastic profiles, the datasets used have to be selected based on seasonality, monthly variations, or similar grouping.

The application of the proposed methodology in the UEBM is to model the stochastic plug load profiles for different buildings of similar types instead of applying the fixed profiles. Having analysed a cluster of buildings with similar use types and a variety of plug loads, the developed model can be applied to the archetype in a particular climate zone. Using a transfer learning model, the learned knowledge can be applied to similar archetypes in the same climate zone. In this case, plug load schedules for similar archetypes are stochastically modelled to prevent unrealistic energy demand peaks and oversizing of the district's electrical or thermal distribution systems. Furthermore, appropriate sizing of energy systems reduces energy consumption and associated greenhouse gas emissions. Extracting the representative plug load profiles for different archetypes in various climate zones can be also used to improve the national building codes and standards.

Conclusion

Urban building energy simulation uncertainty is significantly influenced by occupant-related parameters. By assigning fixed and standard occupant-related schedules to UEBMs, a large difference is observed between the simulated and measured data. In other words, the high variation of the profiles in a similar building type of a climate zone and another climate zone illustrates that simplifying the occupant-related profiles in the UEBM reduces the accuracy and reliability of the energy demand simulation.

This study presents a stochastic-based model to provide the occupant-centric archetypes for the UEBM. The proposed framework is applied to the plug load data. The applicability of the model is tested on the office buildings in climate zone 5A from BDGP2. Using the proposed workflow, derived plug load profiles of each building type in a specific climate zone can be used to implement the energy simulation in building, archetype, and district scale. The result indicates a reasonable correlation between the generated plug load profiles and actual profile values.

The simulation outcome improves the quality of the energy efficiency measures and energy savings by enriching building energy simulations with more realistic occupant-related schedules. In addition, since the occupant-related data is not widely collected due to the absence of sub-metering devices and privacy regulations, the methodology offers a route to extract the plug load data from electricity consumption data. However, all electrical loads associated with cooling and heating are assumed to be plug loads. In case the model is extended to more buildings, in which building electricity use usually includes cooling and heating as well as car charging, it will not fully comply with this assumption. The workflow could also be used as an add-on feature to different UEBM platforms to provide the representative occupant-related profiles. This framework has limitations that will be addressed in future studies. However, the framework is generic and applicable to derive other profiles. Finally, this methodology will be validated with different buildings occupant-related data and use cases.

Acknowledgement

This research was undertaken, in part, thanks to funding from the Canada Excellence Research Chairs Program with grant number CERC-2018-00005.

References


