Data-driven Prediction of Occupant Presence and Lighting Power: A Case Study for Small Commercial Buildings

Jing Wang[1], Wangda Zuo[1], Sen Huang[2], Draguna Vrabie[2]

[1]Department of Civil, Environmental, and Architectural Engineering, University of Colorado Boulder, USA,
{jing.wang, wangda.zuo}@colorado.edu

[2]Pacific Northwest National Laboratory, USA, {huang875, draguna.vrabie}@pnnl.gov

Abstract

Commonly used deterministic methods are unable to capture the randomness in occupant behavior and its impact on electric power consumption. In this paper, we propose a new data-driven model to capture occupant behavior in a stochastic manner. Unlike existing models and prediction tools, this new model does not require occupant presence data and can learn occupants' arrival and departure time based on lighting power consumption data, which is more readily available than occupant presence data. We applied this occupant behavior model to lighting power consumption prediction and implemented the entire prediction process in Modelica. We then validated the Modelica model by comparing the predicted daily, weekly and monthly peak lighting power with measurements from two small commercial buildings. The results suggest that the prediction matches the measurement within acceptable deviations of 7%. The results also indicate that the proposed stochastic model performs better for long-term prediction of lighting power (monthly and weekly) than the short-term (daily).

Keywords: Occupant behavior modeling, occupant presence prediction, lighting power prediction, regression model, stochastic simulation

1 Introduction

The increasing penetration of renewable energy is introducing more variability within the power grid (J. Wang et al. 2018). To better balance generation and consumption, the power demand side needs to become more flexible and even more controllable. Some studies focus on estimating building load flexibility by controlling thermostatically controllable loads (TCLs) such as HVAC systems and water heaters in buildings (Wu et al. 2018; Zhao et al. 2017). Compared to TCLs, the lighting system has the advantage of shorter response time which makes it more suitable for faster demand response mechanisms (e.g., shimmy).

The stochasticity of occupant behavior and its impact on power and energy consumption presents a challenge to accurate real-time estimation of building electric loads. Traditional building energy modeling tools use static hourly schedules both for occupant presence and building equipment. This leads to discrepancies between the simulated power shape and the actual consumed power (Luo et al. 2017; Kim et al. 2017), especially for short-term prediction scenarios such as those needed for fast demand response. Limited data availability is a second challenge, as due to privacy reasons, occupant sensor data is often unavailable. These challenges must be accounted for in theoretical and model-based studies on occupant behavior and its related impacts on the power consumption and flexibility characterization of the built environment.

For commercial buildings, existing occupant presence prediction models have been developed mainly on single office rooms. Wang et al. used exponential distribution to predict the vacancy intervals of single offices (D. Wang, Federspiel, and Rubinstein 2005). Small commercial buildings have not gained enough attention concerning occupant behavior studies.

Lighting prediction models have been investigated over the past 40 years, and the research points to strong correlation between occupants' presence and the lighting status in a zone. The first published study for occupants' light switching behavior in office buildings found that switching mainly takes place when entering or vacating a space and the switch-on probability on arrival exhibits a strong correlation with minimum daylighting illuminance in the working area (Hunt 1980). Manual switch-off probability of lights is strongly correlated with the expected length of absence (Pigg, Eilers, and Reed 1996). Later, this research was expanded by the study of correlations between intermediate switch-on/-off behavior and illuminance levels (Reinhart and Voss 2003).

In this paper we propose a methodology for occupant presence and lighting power prediction based on minute-level power meter data. We apply the methodology for two small commercial buildings use cases (one bakery and one ice cream shop) and validate the prediction performance with real data collected from building sites. Here we present only the prediction of occupant presence and lighting power. In future work we will extend the methodology to other loads driven by occupant behavior.

DOI 10.3384/ECP2016973

The innovation of this work lies in: (1) The proposed method can be applied to occupant presence prediction without occupancy sensor data and it has been validated against real power meter data. (2) The method can be used for sub-hourly power demand prediction within acceptable deviations of 7%. (3) The method could be applied to other building systems and the Modelica model is extensible and scalable. The rest of the paper is organized as follows: Section 2 presents the methodology. Section 3 discusses the results. Section 4 concludes this paper with future work and limitations.

2 Methodology

Our method is based on the assumption that the usage of the lighting system and its associated power consumption is strongly determined by the presence of the occupants in the building spaces. This assumption allows us to extract occupant presence schedules from lighting power data. We then use the extracted presence data to train logistic regression models that predict people's arrival and departure times. The trained probability models are then implemented in Modelica language to reproduce building occupancy patterns. The lighting power is then predicted by multiplying the occupant presence value (0 or 1) with the observed nominal lighting power. We then extend the model to address realistic scenarios of multi-stage lighting power. To validate our model, we compare the simulation results with the lighting power data collected at two building sites and evaluate the model performance with respect to several statistical metrics.

The following flowchart (Figure 1) shows the research workflow for the results presented in this paper.

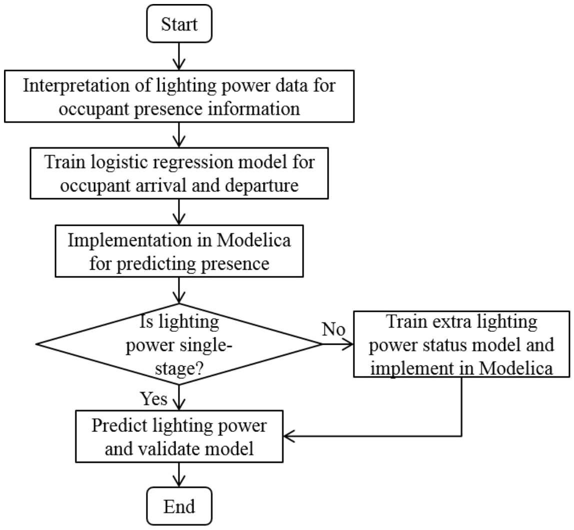
Start
Interpretation of lighting power data for occupant presence information
Train logistic regression model for occupant arrival and departure
Implementation in Modelica for predicting presence
Is lighting power single-stage?
No
Train extra lighting power status model and implement in Modelica
Yes
Predict lighting power and validate model
End

Figure 1. Research and modeling workflow.

2.1 Determine Occupant Presence

In this section, we discuss the extraction of occupant presence information from the lighting power data. As indicated in the literature review, occupant arrival time and departure time has a strong correlation with the lighting power utilization: According to Hunt's work (Hunt 1980), the action of turning on the lights depends on the minimum illuminance level on the working plane upon arrival and people tend to leave the lights on until the space is fully empty. This is consistent with our observation on the lighting power data in the two studied buildings (C2: ice cream shop and F1: bakery). As plotted in Figure 2 and Figure 3, once the lights are turned on, they will remain on for the whole day until all the people leave the space. This means that in this case the illuminance level is not a strong driver for the light utilization. In our preparation work where we used regression of lighting power based on indoor illuminance levels, prediction accuracy was relatively low. In this paper, we will assume that people in the two studied buildings are not sensitive to the illuminance levels and will turn on the lights once they enter the space and will keep the lights on while they are there. In other words, lights are not switched on to increase workplace illuminance levels, but rather to show potential customers that the store is open. Based on this assumption, we extract the occupant presence information from the lighting power data and regard it as the ground truth.

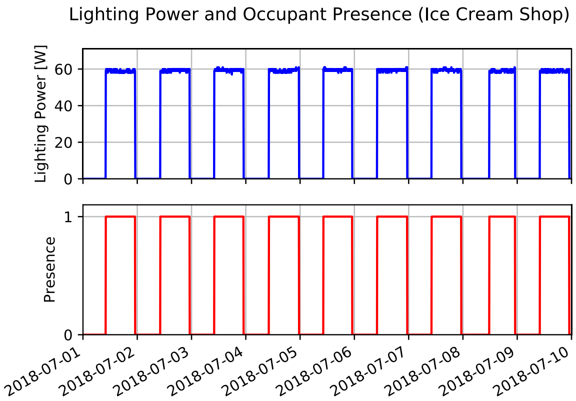

Figure 2. Lighting power and occupant presence (C2: ice cream shop).

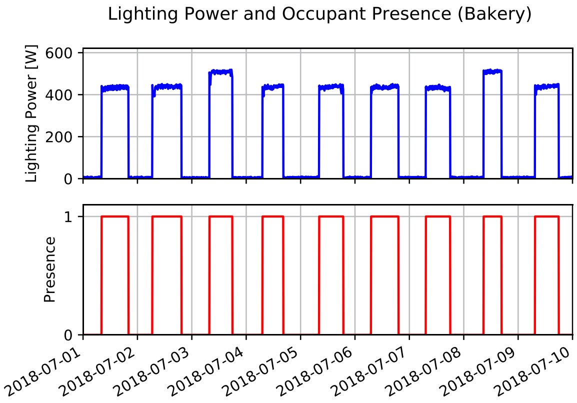

Figure 3. Lighting power and occupant presence (F1: bakery).

To convert the lighting power data into occupant presence information, we first cleaned the power meter data by removing obvious outliers such as values that are extremely large for lighting systems. Then, we selected the threshold for determining occupant presence (e.g., 0 for absent; 1 for present) to avoid oscillations in presence status. For instance, the threshold for the ice cream shop is 50 W; for the bakery is 350 W. Any power value above this threshold is converted to 1 and below this threshold into a 0. Because the power data has 1-minute resolutions, we will make the assumption that presence or absence of 1 minute can be neglected and we will filter out two consecutive changes of occupant presence to eliminate frequent oscillations in the resulted presence data.

The lighting power shapes shown in Figure 2 and Figure 3 indicate the different characteristics of the two buildings. For the ice cream shop, only one power value occurs every day regardless of weekday or weekend. However, for the bakery, two distinct levels are observed in the power shape. Hence, for his case, we divide the power shape into two parts namely base lighting power (Figure 4) and additional lighting power (Figure 5) and we model them separately. This two-stage lighting behavior is probably caused by zoning of the lighting system. The expression for multi-stage lighting power can be described with Eq. 1.

$$P(t) = a_0(t)P_{base} + a_1(t)P_{extr,1} + \cdots + a_{n-1}(t)P_{extr,n-1} \quad (1)$$

P is the lighting power; a_i is the binary variable that indicates the status of base or extra lighting; n is the number of stages. Both P and a_i are time dependent. Here, a_0 indicates the building occupancy and the rest of them indicates the on/off of extra lighting devices. a_i is predicted with logistic regression models introduced in Section 2.2. P_{base} and $P_{extr,i}$ are the average power value of each stage. For C2, $n = 1$; for F1, $n = 2$.

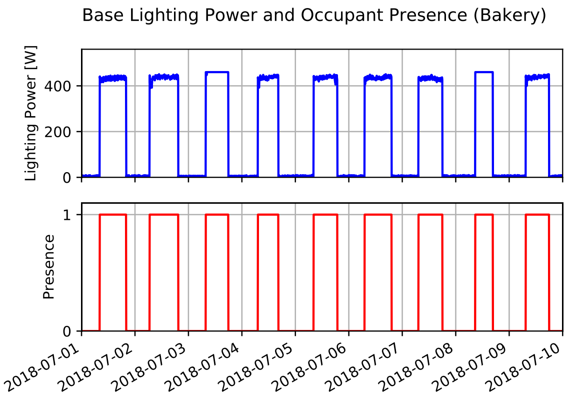

Figure 4. Base lighting power and occupant presence in F1 bakery.

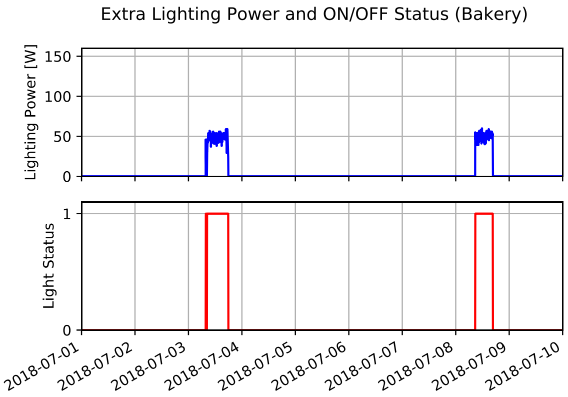

Figure 5. Extra lighting power and lighting status in F1 bakery.

2.2 Train Logistic Regression Models

The prediction of occupant presence could be viewed as a classification problem. As discussed before, the arrival and departure behavior in the two studied buildings follows the same pattern for weekdays and weekends regardless of the indoor illuminance level. Hence, the main feature for classifying occupant presence is the time of the day. We chose logistic regression as our model for the training because: (1) it is a linear classifier and is easy to train; (2) it can reach the same level of accuracy as non-linear classifiers; (3) it is easy to implement in Modelica. We divided the arrival and departure behavior into two models and trained them separately as they have opposite trends along time of the day.

To rule out the impact of seasonal change in the occupant behavior, the training and validation datasets were selected from the summer of 2018. June and July data were used for the training and August data was used for the validation. The accuracy is defined as the rate of classifying the data point into the right group. The confusion matrices for the test datasets of all the regression models are shown in Table 1. The format of the confusion matrices follows the pattern in Table 2.

Table 1. Confusion Matrices for Classification Performance.

C2 Arrival	3693	44
	31	624
C2 Departure	283	60
	4	1849

F1 Arrival	2736	132
	118	1406
F1 Departure	1797	260
	273	2062
F1 Extra On	16	0
	3	0

Table 2. Example Confusion Matrix (C2 Arrival).

	Predicted No	*Predicted Yes*
Actual No	3693	44
Actual Yes	31	624

The accuracy of the classifier is then calculated with Eq. 2.

$$Accuracy = \frac{No.\ of\ correctly\ classified\ points}{No.\ of\ total\ data\ points} \qquad (2)$$

For building F1, the lighting power is divided into the base power and the extra power. The base part reflects occupants' arrival and departure and is regressed in dependence on time of the day. The frequency (i.e., number of total times) of extra lights on of F1 in 2018 is plotted in bars (Figure 8). From the figure, we can see that the status of the extra lighting has a correlation with day of week. Hence, the feature for this part is chosen as day of week. Also, from the figure, we see that the total frequency of extra lights on in 2018 is only 8.8%. To deal with the imbalance in the training dataset, we adopted the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al. 2002), which made the minority (extra lights on) class equal to the majority class (extra lights off) by creating synthetic samples of the minority class. The logistic regression parameters for each model are listed in Table 3. The probability

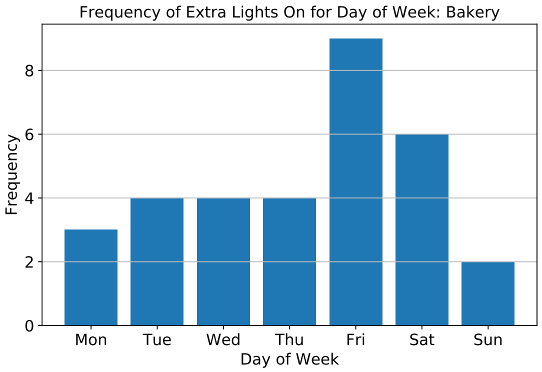

Figure 8. Extra lights on frequency for day of week in F1 bakery (2018).

Table 3. Logistic Regression Parameters.

		Accuracy	β0	β1	β2	β3	β4	β5	β6	β7
C2	Arrival	0.98	-27.1983	0.0447	N/A					
	Departure	0.97	34.6877	-0.0249						
F1	Arrival	0.94	-11.9311	0.0254	N/A					
	Departure	0.88	13.7769	-0.0125						
	Extra On	0.84	-0.8309	-0.4829	0.4967	-0.2586	0.4967	-0.1171	-0.4829	-0.4829

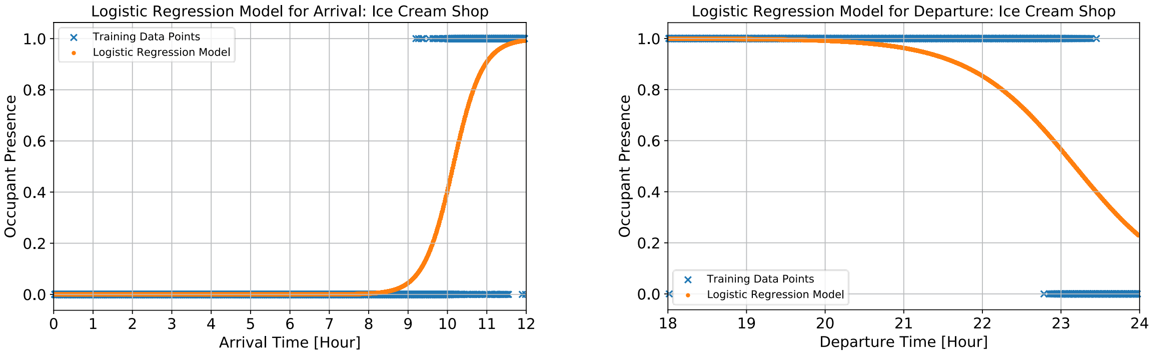

Figure 6. Logistic regression model for arrival (left) and departure (right) in C2 ice cream shop.

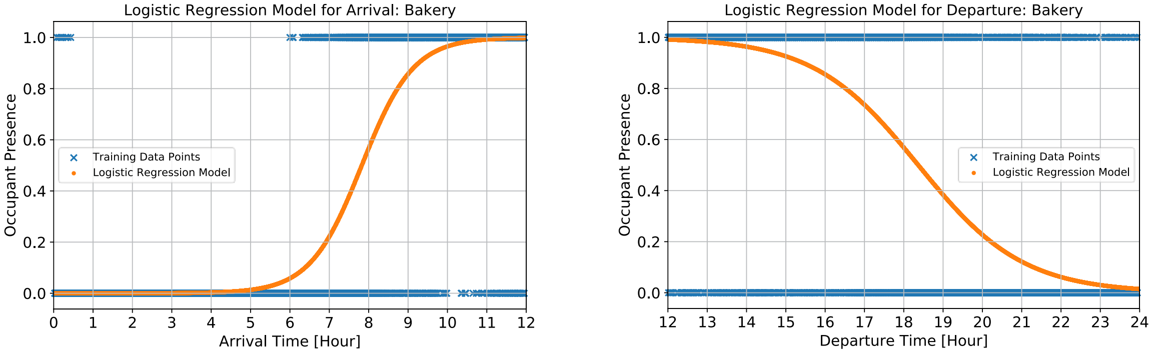

Figure 7. Logistic regression model for arrival (left) and departure (right) in F1 bakery.

function is expressed in Eq. 3, where p represents the probability of occupant present or extra lights on; e is the natural log base; β is the regression intercept and coefficients; m refers to the number of logistic regression independent variables. The accuracy of all the models are above 84%. Table 4 lists the probability of extra lights on for day of week in building F1.

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}} \tag{3}$$

Table 4. Probability of Extra Lights On for Day of Week from Logistic Regression.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Probability	0.21	0.42	0.25	0.42	0.28	0.21	0.21

Figure 6 and Figure 7 visualize the training data points and the logistic regression models for arrival and departure in C2 and F1. Based on our observations, occupants will arrive before 12 pm and leave after 12 pm. Hence, the arrival models are trained with data points before 12 pm and the departure models with points after 12 pm. For the ice cream shop departure model, people tend to leave very late. To increase the prediction accuracy, we used data after 6 pm to train this model.

2.3 Implement in Modelica

The implementation of the presence model and the extra lighting status model is adapted from Buildings.Occupants.Office.Lighting.Hunt1979Light in Modelica Buildings library (Wetter et al. 2014). The model is implemented as a stochastic simulation model. Every two minutes, a binary variable generator will randomly generate a binary number. The probability of this number being 1 equals the calculated probability of the occupant being present at that time of day based on the logistic regression model. Similarly, in the extra light status model, the probability of the random number being 1 equals the probability of the extra light being on at the simulated day of week.

Figure 9 depicts the layout of the two-stage lighting power prediction model for F1. The presence models generate binary signals which will be multiplied with the nominal power of each stage. The nominal powers are the calculated mean values of the lighting power in each stage. The sum of the lighting power of all stages are then compared with the actual lighting power data to validate the performance of the stochastic simulation models. An assumption is made in this model that the extra light will only be on when both of the following conditions are satisfied: (1) The extra light should be on for that day of week; (2) There are occupants in the building. The simulation was run for the whole month of August 2018 and the time step was set as 10 minutes. The actual time step was picked by Dymola to be 2 minutes due to the stochastic events.

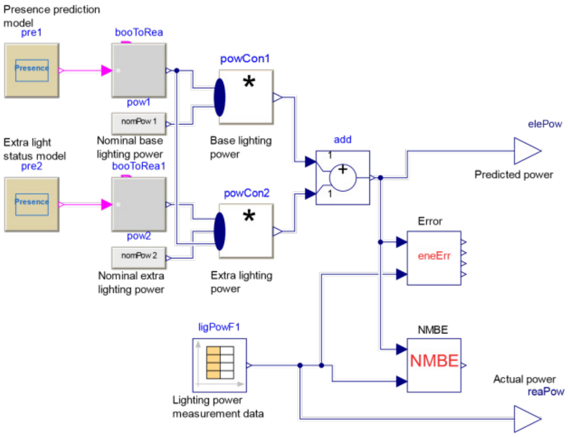

Figure 9. Modelica layout of the two-stage lighting power prediction model.

3 Results and Discussions

We evaluate both the occupant presence prediction performance and the lighting power prediction performance in this section. The presence models are evaluated with the root mean squared error (RMSE) and the coefficient of variation of RMSE (CVRMSE) of the probability distribution model. The lighting power prediction performance is evaluated with the relative error of the peak power and normalized mean bias error (NMBE). The error in the lighting power prediction is dependent on the presence prediction error as well as the error of nominal power estimation.

ASHRAE Guideline 14-2002 has requirements for whole building energy calibration (ASHRAE 2002). The smaller the time scale, the more tolerant the criteria. For example, the criteria for monthly NMBE is 5%, monthly CVRMSE is 15%, and the criteria for hourly NMBE is 10%, hourly CVRMSE is 30%. Though only the lighting system is calibrated in our work, the principle for different time scales should apply.

3.1 Occupant Presence Prediction

RMSE represents the standard deviation of the errors and CVRMSE is the ratio of the standard deviation to the mean of the dependent variable. They both describe how concentrated the data is around the line of its best fit. Large errors are especially noticed in these metrics. The equations for calculating the two metrics are listed below. $x_{o,i}$ is the original value of the predicted variable, $x_{f,i}$ is the forecasted value, N is the number of total data points. Table 5 lists the RMSE and CVRMSE of the occupant and extra lighting status prediction models. The CVRMSE for the occupant presence models are below 25%. The CVRMSE for extra lighting prediction is 125%. This is caused by the imbalance of the training data. The probability of the extra lights being on is much lower than the probability of them

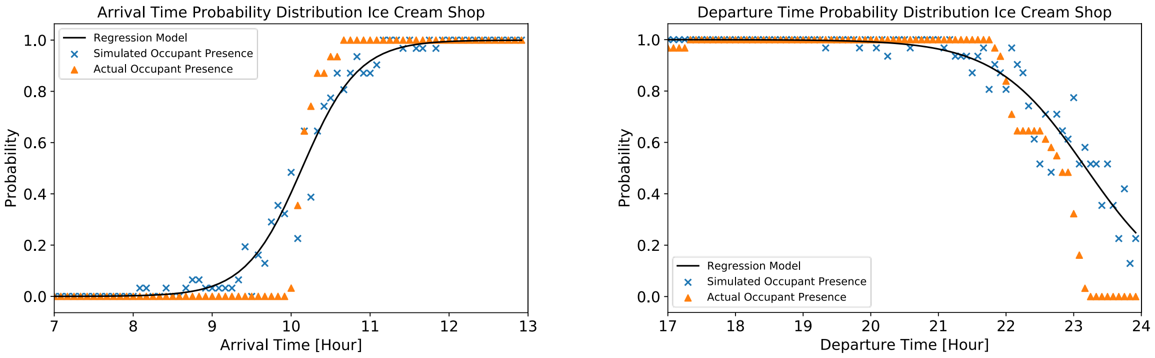

Figure 10. Arrival and departure time probability distribution (C2: ice cream shop).

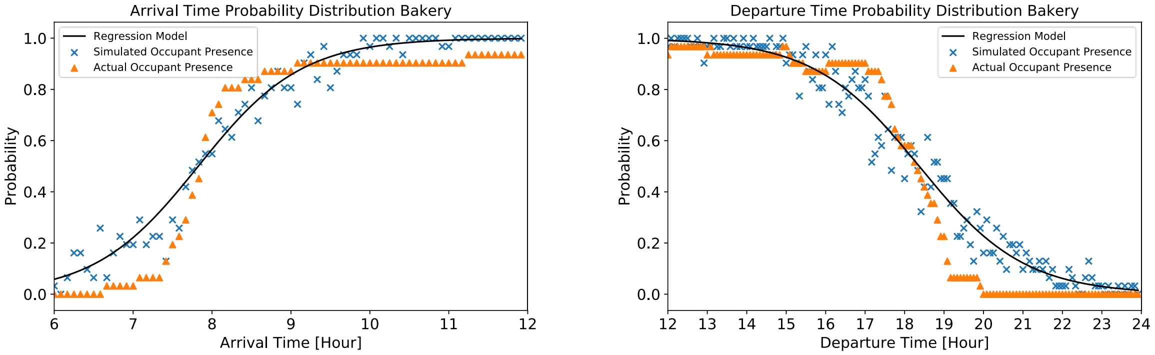

Figure 11. Arrival and departure time probability distribution (F1: bakery).

being off. Hence, the mean value $\overline{x_o}$ is very small and small errors could cause a large CVRMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(x_{f,i} - x_{o,i}\right)^2}{N}} \tag{4}$$

$$CVRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_{f,i} - x_{o,i}\right)^2}}{\overline{x_o}} \tag{5}$$

Table 5. RMSE and CVRMSE of Occupant Presence and Lighting Status Prediction Results.

	C2	*F1*	
	Occupant Presence	*Occupant Presence*	*Extra Lights*
RMSE	0.108	0.101	0.153
CVRMSE	20.9%	25.0%	125%

Figure 10 and Figure 11 plot the regression model, simulated probability distribution and the actual probability distribution of arrival and departure in the two buildings. From the figure, we see that the simulated probability distribution aligns with the regression model very well. The actual probability distribution deviates from the regression model especially during the transitional periods in the middle (e.g., 9 to 11 for C2 arrival, 17 to 21 for F1 departure). This could have been caused by the inappropriate selection of the training data. The high accuracy of the classifiers shown in Table 3 is partially because more data points are located outside the transitional period. The classifier can distinguish those points easier. Another reason could be that only one feature is used to predict occupant presence. This could have limited the shape of the logistic regression model to further fit the actual curve. More features should be explored in the future.

Table 6 compares the probability of extra lights on in F1 calculated from the simulated results and the actual data. From the table, we see that the simulated and actual results deviate on Tuesday and Wednesday. For other days, the simulation results reproduced the actual probability well.

Table 6. Comparison of Simulated and Actual Probability of Extra Lights On for Day of Week.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Simulated	0	0.29	0.29	0.14	0.29	0.29	0.14
Actual	0	0	0	0.14	0.29	0.29	0.14

3.2 Lighting Power Prediction

To evaluate the lighting power prediction performance of the models, peak power prediction relative error and NMBE are calculated on a monthly, weekly and daily basis. In this way, the lighting power prediction performance is evaluated for different time scales. As the models in this paper are mainly designed for shorter-time demand response scenarios, annual energy consumption is out of scope. Table 7 summarizes the peak power prediction accuracy. For C2, the errors are all below 2.36%. For F1, which is two-stage prediction, the errors are larger, but all stay below 6.9%. Hence, the multi-stage method performs well in predicting peak power.

Table 7. Peak Power Prediction Accuracy.

	Monthly Peak Power	*Weekly Peak Power*	*Daily Peak Power*
C2	2.36%	2.36%~2.36% (avg: 2.36%)	0.73%~2.36% (avg: 1.99%)
F1	6.90%	2.15%~6.90% (avg: 5.34%)	1.05%~6.90% (avg: 2.42%)

To further evaluate the fitness of the power curve to the real power curve, the NMBE metric is adopted, which describes the average bias in the model. NMBE is determined with Eq. 6. By definition, it is the sum of error over the sum of the actual values. This metric evaluates the fitness of the model over the whole simulation horizon.

$$NMBE = \frac{\sum_{i=1}^{N}\left(x_{f,i} - x_{o,i}\right)}{N \times \overline{x_o}} \quad (6)$$

Table 8 summarizes the daily, weekly and monthly NMBE of the lighting power. The lighting power obtained by multiplying the ground truth occupancy data with nominal power is set as the baseline for better comparison. From the table, the two-stage prediction generally has larger errors than the single-stage model. For the single-stage lighting power (C2), the monthly, weekly and daily NMBE are all within 5%, which indicates a high accuracy for power demand predictions. For the two-stage lighting power (F1), the monthly and weekly average errors are within 10%, which is still acceptable. However, we see a big deviation in the daily NMBE, and this leads to a high average value for daily NMBE. This high deviation could have been caused by an uncommon data record on Aug. 19 (see Figure 12) when the lights are only on for a short time period but the model simulated it just as usual.

Table 8. NMBE of Lighting Power Prediction.

		Baseline	*Model*
Monthly NMBE	*C2*	0.061%	3.92%
	F1	-0.55%	8.28%
Weekly NMBE	*C2*	-0.27%~0.44% (avg: 0.060%)	-0.25%~9.84% (avg: 4.07%)
	F1	-2.84%~1.30% (avg: -0.68)	0.33%~20.4% (avg: 7.92%)
Daily NMBE	*C2*	-0.56%~0.72% (avg: 0.057%)	-2.59%~23.72% (avg: 4.03%)
	F1	-12.9%~50.9% (avg: 0.39%)	-21.6%~807% (avg: 44.1%)

Additionally, as the models are simulated in a stochastic manner and the occupant presence was determined every 2 minutes, we see an obvious oscillation in lighting power in Figure 12. This feature of the model leads to that the longer the simulation time, the closer the expectation of the simulation results will be to the actual data. This explains why the model shows a better performance concerning monthly NMBE. However, short-term accuracy of the model still needs some improvement.

4 Conclusion

This paper proposed a methodology for occupant presence learning and reproducing based on lighting power metering data. The method was validated against real data. The results show that the proposed multi-stage lighting power prediction method can predict daily peak power with 2.42% relative error. The monthly and weekly NMBE of lighting power are on average below 8.28%.

Through the training and validation process of this work, we found that logistic regression models are sensitive to the quality of the training data. Ideally, the dataset should be more focused on the transitional region (i.e., where the value turns from 0 to 1 or vice versa) of the model and the two classes should be well balanced. Further, increasing the number of independent features should help improve the fitness of the probability model. The stochastic simulation results show that stochastic models can be very accurate for

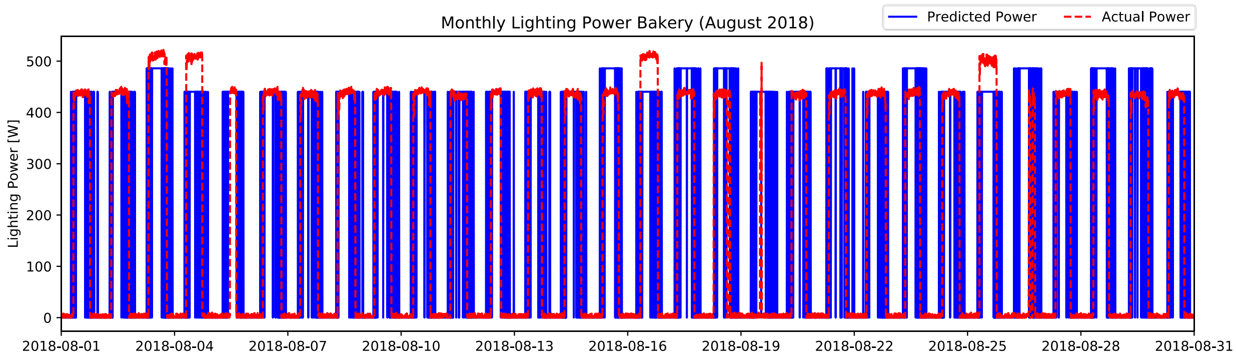

Figure 12. Monthly predicted and actual lighting power in F1 bakery.

long-term predictions. However, they cannot predict uncommon events, and this can lead to large short-term prediction errors.

This work has the limitation of not having the ground truth data for occupant presence. The presence generated from lighting power can be delayed when people arrived and did not turn the lights on. This can be cross validated with other appliance usage data in the future. In the best-case scenario, occupant surveys should be conducted to know their preferences and habits, and occupant sensors should be installed.

Acknowledgements

This research was supported by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Building Technologies Office, under Contract No. DE-AC05-76RL01830. This research was also supported by the National Science Foundation under Awards No. IIS-1802017 and IIS1633363. BIGDATA: Collaborative Research: IA: Big Data Analytics for Optimized Planning of Smart, Sustainable, and Connected Communities.

References

ASHRAE. 2002. *ASHRAE Guideline 14-2002: Measurement Of Energy And Demand Savings*.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. “SMOTE: Synthetic Minority over-Sampling Technique.” *Journal of Artificial Intelligence Research* 16: 321–57.

Hunt, D R G. 1980. “Predicting Artificial Lighting Use-a Method Based upon Observed Patterns of Behaviour.” *Lighting Research & Technology* 12 (1): 7–14.

Kim, Yang-Seon, Mohammad Heidarinejad, Matthew Dahlhausen, and Jelena Srebric. 2017. “Building Energy Model Calibration with Schedules Derived from Electricity Use Data.” *Applied Energy* 190: 997–1007.

Luo, Xuan, Khee Poh Lam, Yixing Chen, and Tianzhen Hong. 2017. “Performance Evaluation of an Agent-Based Occupancy Simulation Model.” *Building and Environment* 115: 42–53.

Pigg, S., Mark Eilers, and John Reed. 1996. “Behavioral Aspects of Lighting and Occupancy Sensors in Private Offices: A Case Study of a University Office Building.” *ACEEE 1996 Summer Study on Energy Efficiency in Buildings*, no. 8: 161–70.

Reinhart, C F, and K Voss. 2003. “Monitoring Manual Control of Electric Lighting and Blinds.” *Lighting Research & Technology* 35 (3): 243–58. https://doi.org/10.1191/1365782803li064oa.

Wang, Danni, Clifford C. Federspiel, and Francis Rubinstein. 2005. “Modeling Occupancy in Single Person Offices.” *Energy and Buildings* 37 (2): 121–26. https://doi.org/10.1016/j.enbuild.2004.06.015.

Wang, Jing, Wangda Zuo, Landolf Rhode-Barbarigos, Xing Lu, Jianhui Wang, and Yanling Lin. 2018. “Literature Review on Modeling and Simulation of Energy Infrastructures from a Resilience Perspective.” *Reliability Engineering & System Safety*.

Wetter, Michael, Wangda Zuo, Thierry S Nouidui, and Xiufeng Pang. 2014. “Modelica Buildings Library.” *Journal of Building Performance Simulation* 7 (4): 253–70.

Wu, D, H Hao, T Fu, and K Kalsi. 2018. “Regional Assessment of Virtual Battery Potential from Building Loads.” In *2018 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*, 1–5. https://doi.org/10.1109/TDC.2018.8440225.

Zhao, L, W Zhang, H Hao, and K Kalsi. 2017. “A Geometric Approach to Aggregate Flexibility Modeling of Thermostatically Controlled Loads.” *IEEE Transactions on Power Systems* 32 (6): 4721–31. https://doi.org/10.1109/TPWRS.2017.2674699.