Hourly occupant density prediction in commercial buildings for urban energy simulation

Yue Wu1,2, Yanxia Li1,2, Chao Wang1,2, Xing Shi1,2,*

1Key Laboratory of Urban and Architectural Heritage Conservation, Ministry of Education, Nanjing, China, 210096.
2School of Architecture, Southeast University, Nanjing, China, 210096

Corresponding author e-mail: shixing_seu@163.com

Abstract. Building occupant density is a key factor influencing urban energy consumption. However, it is difficult to predict and thus often simplified to be an assumed and fixed number in urban energy simulation. In this study, the hourly occupant densities of ten representative commercial buildings in Nanjing, China were measured. The pattern of the hourly occupant density was analyzed and the key parameters defining the pattern were identified. To expand the measured hourly occupant density pattern to thousands of commercial buildings in Nanjing, five predictors, namely function, accessibility, population, business diversity, business density, were proposed. Big data technique was used to obtain the value of the five predictors for more than 3000 commercial buildings. A regression analysis was conducted to establish a model linking the five predictors with the parameters defining the hourly occupant density pattern. The methodology developed provides an effective means to predict the hourly occupant density of buildings and thus substantially improves the accuracy and reliability of urban energy consumption.

1. INTRODUCTION

The rapidly growing energy use has already raised concerns over resources shortage and environmental degradation. According to the International Energy Agency, urban energy consumption accounts for two-thirds of the world's total energy consumption. It mainly includes building, transportation, and industry sectors, of which about 60% is used for building energy consumption (Shem Heiple and David J. Sailor 2008). With buildings being a key contributor, considerable effort has been made to model the energy consumption in urban scale, which has significant implications in urban energy optimization and management.

Urban energy consumption simulation, proposed as a bottom-up urban simulation method, refers to the synthesis and computer modeling of buildings' geometric and non-geometric information. It uses the similar heat conduction principle with single building to reliably predict the dynamic urban energy. Among the various parameters, the hourly occupant density is a significant one since it has a crucial
effect on the accuracy. On one hand, personal heat dissipation is an important part of building heat source, directly affecting the air-conditioning load and therefore the building energy consumption. On the other hand, it is the setting basis of using timetable of indoor lighting system and other equipment (Wang et al. 1982) However, there are many factors influencing occupant information, like function, weather, age, transportation accessibility, etc. It is hard to predict this parameter especially in district/urban scale, due to the randomness in time and space. In the previous studies, it is generally set according to the fixed value from the specifications. However, this method has been proven to be limited and unreasonable, resulting in discrepancy of simulation (Ahmed el al. 2017, Yuvraj el al. 2010).

This paper proposed an hourly occupant density model suitable for urban energy consumption simulation based on the perspective of big data, taking the commercial buildings of Nanjing, China as an example. Firstly, 3125 commercial buildings were identified based on the POI data of Nanjing and then divided into 36 categories according to their functions, transportation accessibility, and population level, three probabilistic factors may influence occupant density of commercial buildings. After that, one in each type was chosen randomly to conduct a survey of people counting (one weekday and one day on weekends) and ten of them have been finished up to now. By analyzing the measured data, the hourly occupant patterns were simplified to three connected lines and could be defined with 4 characteristic parameters. We then proposed another two probabilistic factors: business diversity and business density. In the end, the influences of these five factors on the characteristic parameters were statistically analyzed, and the regression models between those parameters and its main influencing factors were preliminarily established respectively. The following figure shows an outline of this study.

2. COMMERCIAL BUILDINGS IDENTIFICATION BASED ON POI DATA

“POI” is the abbreviation of “Point of interests”, which represents all the geographical entities that can be abstracted into points, especially facilities closely related to people's daily life, such as shopping malls, stations and schools (Zhao and Wang 2011). Each piece of data contains attributes including its name, category, longitude and latitude, and its administrative region. It is usually applied to navigation in map software. A total of 664,755 POI points of Nanjing were obtained from the Gaode map using the FME software. In addition, a total of 261,599 polygon data of the building plan were obtained from Baidu map by self-programming method, and each contains three attributes: longitude, latitude and height.

After data pre-processing, the POI data was reclassified and assigned. That is because different types of POI function differently in the process of identifying commercial buildings. For example, when there are two “convenience store” points and one “shopping mall” point in a building, the latter obviously plays a more important role in determining the type of the building. Considering that urban energy consumption simulation is the ultimate goal of the research and it is easily affected by floor area, those POI points were assigned according to their general area determined by field research and expert interview. The weights indicate the importance of different POI points in identification. Part of the results are shown in Table 1.

| No. | Category | Secondary category | Tertiary category | Examples | Weights |
|-----|----------|--------------------|------------------|----------|---------|

Table 1 The reclassification and assignment of POI points.
There are often many kinds of stores in a commercial building, which means that there are many points of different POI in a building plan. Accordingly, the processed POI data and plan data were combined through the spatial joint tool in ArcGIS, and the number of points were counted in each plan (each building). Then the weights in the above table were assigned to get the area occupied by each type of POI in each building. According to the results, the following four conditions were proposed to decide whether it is a commercial building:

1. \( \frac{S_{\text{commercial (including catering)}}}{S_{\text{total}}} = \frac{\sum_{i=1}^{n} S_i}{S_{\text{total}}} \geq 50\% \)

2. \( \frac{S_{\text{commercial (without catering)}}}{S_{\text{commercial (including catering)}}} = \frac{\sum_{i=1}^{n} S_i}{\sum_{i=1}^{n} S_i} \geq 60\% \)

3. Not within the residential area

4. Building area should be above 500m²

Here, (1) and (2) were used to ensure commercial buildings which mainly involve retail could be elected. (3) was used to ensure the building was not an auxiliary facility of a community. A total of 3,125 commercial buildings were identified in the end.
Two districts where commercial buildings are relatively abundant were used for testing. Comparing the commercial buildings identified by this method and field research, the accuracy rate reached 82.7% and 81% respectively, proving that the method is of strong adaptability.

3. SELECTION OF FIELD RESEARCH BUILDINGS

In order to select representative buildings for field research, 3,125 commercial buildings were divided into 36 categories by function, transportation accessibility, and population level. One in each type was chosen as a subject randomly, and 10 of them have been finished up to now. Those three are probabilistic factors influencing the occupant density determined through literature studies (Feng et al. 2009) combined with practical experience. In the following, the definition and determination of those factors are introduced in detail.

Function refers to a certain demand in the daily life that the building can serve, determining the attraction to a specific group of people in a certain period of time. In this study, the function was defined by the type of POI with the largest area ratio in each building plan. Four types were identified as shopping malls, supermarkets, electronic stores and building materials stores. We named them as F1 (F for Function), F2, F3 and F4 respectively.

Transportation accessibility refers to the difficulty of obtaining the service from a certain location to a commercial building. “Service area analysis” tool in ArcGIS was used to quantify the difficulty. Firstly, the road system was simulated in the light of the real road data of Nanjing. Then, the area each building could be access to within a certain period of time by a certain means of transportation through the road system was calculated using that tool (Yang and Song 2004). The shorter the time required, the more people tend to choose the building as a destination. So different weights were deployed to indicate the trend. Integrated service areas were calculated by the following equations finally.

\[ S_{\text{private}} = 0.5 \times S_{15 \text{ minutes}} + 0.3 \times S_{30 \text{ minutes}} + 0.2 \times S_{45 \text{ minutes}} \]  
\[ S_{\text{public}} = 0.5 \times S_{15 \text{ minutes}} + 0.3 \times S_{30 \text{ minutes}} + 0.2 \times S_{45 \text{ minutes}} \]  
\[ S_{\text{integrated}} = 0.5 \times S_{\text{private}} + 0.5 \times S_{\text{public}} \]

where Sprivate is the service area by private-driving(m2), and Spublic is the area by public transportation (including subway and bus) (m2), S15minutes, S30 minutes, S45minutes are the area accessible within 15, 30, 45 minutes respectively(m2). Sintegrated is the ultimate service area of each building(m²).

The integrated service area values of all buildings were standardized and divided into three levels. The traffic accessibility from the third level to the first level is gradually increased, and are named Tr1 (Tr for Transportation), Tr2 and Tr3 respectively.

Population level was mainly determined based on Baidu power map. When smartphone users access to Baidu products, their location information would be recorded. In accordance with this, Baidu power map counts the number of personnel in a certain area and uses different colors and brightness to reflect the spatial difference on a raster image. Each pixel of the image has a Z value which doesn't represent the actual number because it has been processed in the background. However, it reflects the relative population of a certain time in a certain area. The data used in this study was Nanjing power
map taken from Baidu Map website every hour on December 24, 2017 and December 25, 2017 during 7 am to 10 pm.

After georeferencing, raster calculation, reclassification, etc., the relative value of population in each grid was calculated by adding the "Z" of each pixel on each power map in the two days together. This value was assigned to each commercial building, which was standardized and processed into three levels. The population level from the third to the first is gradually increased, and named Po1 (P for Population), Po2 and Po3 respectively.

4. RESULTS

4.1. Field research

According to those three factors, commercial buildings were divided into 36 categories. One building of each category was randomly selected as the survey subject and ten of them has been finished until now as shown in the following table and figure.

Table2. The surveyed buildings overview.

| Serial Number | Function | Transportation accessibility | Population level |
|---------------|----------|-------------------------------|------------------|
| 1             | 1        | 1                             | 1                |
| 2             | 1        | 1                             | 2                |
| 3             | 1        | 3                             | 3                |
| 4             | 2        | 1                             | 2                |
| 5             | 2        | 1                             | 3                |
| 6             | 3        | 1                             | 1                |
| 7             | 3        | 1                             | 2                |
| 8             | 4        | 1                             | 3                |
| 9             | 4        | 2                             | 2                |
| 10            | 4        | 2                             | 3                |

Figure 1. The distribution of surveyed buildings and site photos

The number of people was measured by field research during the business hours and recorded hourly. First, the number of people entering and leaving each hour is subtracted to get the number of
stranded at the end of each period, and the value is divided by the effective area of the buildingl. Then, the hourly occupant density was obtained.

4.2. Qualitative analysis

The hourly occupant density patterns of the ten buildings are given in Figure 3 in light color. Some results can be seen from it.

(1) Indoor occupant density is varied evidently with time and behaves different features in weekdays and weekends. (2) It can be found that, in general, the density of shopping malls (NO.1,2,3) and supermarkets (No.4,5) perform higher than the type of electronic stores (No.6,7) and building material stores (NO.8,9,10), which proves that the density is affected by function. (3) Transportation accessibility and population level also have impacts on density. For example, No.8 and No.10 have the same F and Po; however, the average occupant density in weekday and weekend of No.8 whose Tr is the first class is higher than No.10 whose Tr is the second class. (4) Overall, the hourly occupant density pattern of each building shows the characteristics of “double-peak” and “three-stage”. The red segment indicates the period of surging increase in number of people after opening, the yellow segment indicates a marginal increase with fluctuation, and the blue segment indicates the sharp
decline 2-3 hours before the ending time. (5) It can be found that the number of people is not zero at the beginning and the end because the number of staff inside were unable to count beyond business hours. (6) Except for F, Tr and Po, two other factors that may affect the occupant density of commercial buildings can be proposed through the analysis. They are the business diversity and the business density. For example, although No. 9 mainly sells building materials, it also operates restaurants, snack bars, fast food restaurants, etc., which attract a large number of customers. Although the No.10 building is mainly engaged in the electronic products, it not only has many snack bars and fast food restaurants inside, but also is located in the traditional business district of Nanjing, which may result in its high occupant density.

Therefore, the two new parameters of each building were quantitatively calculated. The business diversity was set as the number of POI types contained in each building plan, which was named Di (Diversity). The business density is the number of commercial buildings in the radius of 500m centered on itself and named De (Density).

Moreover, the curves were simplified according to the feature of “three-stage” and “double-peak”. Each one was simplified into three connected lines: a sharp growth to the first peak, a slight growth to the second peak, and a significant decline. Besides, the number of people were set zero one hour before the opening and after the end. The simplified pattern is shown in the Figure 3 in dark color. Then, each pattern could be defined by six characteristic parameters: t1, t2, D2, t3, D3, t4 where t1 and t4 are known for a particular building. So, we just need to know whether there is a relationship between the predictors and the remaining characteristic parameters of the curve. If so, what kind of relationship it is.

4.3. Correlation analysis

In this paper, the data of weekday was taken as an example for regression analysis. SPSS was used to quantify the correlation between t2, D2, t3, D3 and F, Tr, Po, Di, De separately. Since Tr, Po, Di and De are numerical variables, Pearson correlation coefficient “r” was used to measure the correlation; whereas F is a categorical variable, so Kendall correlation coefficient “τ” was used. (τ, also known as the concordance coefficient, is used to measure the correlation of categorical data, whose meaning is similar to the Pearson)(Feng et al.2009) The correlation coefficient of each factor (On weekday) can be seen in Table 3.

| Characteristic parameters | F  | Tr | Po  | Di   | De   |
|---------------------------|----|----|-----|------|------|
| t2                        | τ  | γ  | γ   | γ    | γ    |
|                           | -0.054 | -0.346 | 0.145 | *0.475 | 0.144 |
| D2                        | τ  | γ  | γ   | γ    | γ    |
|                           | *-0.368 | 0.297 | *0.365 | *0.646 | 0.031 |
| t3                        | τ  | γ  | γ   | γ    | γ    |
|                           | -0.116 | *0.806 | *0.46 | *0.35 | *0.593 |
| D3                        | τ  | γ  | γ   | γ    | γ    |
|                           | *-0.515 | 0.103 | *0.377 | *0.585 | 0.006 |
In general, the variables are considered to be irrelevant if the absolute value of the correlation coefficient is less than 0.3 (Sun 2007). As a result, it can be seen from the table that Tr(-0.346) and Di(0.475) are suitable to be the explanatory variables to build regression model for $t_2$. Similarly, F(-0.368), Po(0.365), Di(0.646) are selected for $D_2$; Tr(0.806), P(0.46), Di(0.35) and De(0.593) for $t_3$; F(-0.515), P(0.377), Di(0.585) for $D_3$.

4.4. Regression modelling

Linear regression is a commonly used algorithm for numeric prediction. The general linear model for categorical variables was used to build the regression model between $t_2$, $D_2$, $t_3$, $D_3$ and their main influencing factors. In order to represent the four types of functions, only three dummy variables $F_1$, $F_2$ and $F_3$ were introduced. For shopping malls, $F_1=1$, $F_2=0$, $F_3=0$; for supermarkets, $F_1=0$, $F_2=1$, $F_3=0$; for electronic stores, $F_1=0$, $F_2=0$, $F_3=1$; for building material stores, $F_1=0$, $F_2=0$, $F_3=0$. In accordance with above definition, the regression model of the four characteristic parameters based on the data of the ten buildings surveyed could be expressed as below:

\[ t_2 = 12.422 - 2.404Tr + 3.633Di \]  
\[ D_2 = -0.068 + 0.028F_1 + 0.025F_2 + 0.011F_3 + 0.027Po + 0.267Di \]  
\[ t_3 = 14.285 + 3.127Tr - 0.778Po + 4.722Di + 5.84De \]  
\[ D_3 = -0.082 + 0.003F_1 + 0.056F_2 + 0.128F_3 - 0.008Po + 0.35Di \]

Take $D_2$ as an example, according to the meaning of the dummy variable, it can be seen from the regression result of $D_2$ that the first peak value of occupant density in the shopping mall ($F_1$) is 0.003 people/m$^2$, 0.017 people/m$^2$, and 0.028 people/m$^2$ more than the supermarkets ($F_2$), the building material stores ($F_3$) and electronic stores ($F_4$) respectively. In addition, $D_2$ is increasing with the increase of the population level and the business diversity. Similarly, $t_2$, $D_3$, and $t_3$ can be explained. However, it can be found that the regression result of $D_3$ is inconsistent with the qualitative analysis, as a result of the limited amount of data.

5. CONCLUSIONS AND FUTURE WORK

In this paper, 3125 commercial buildings in Nanjing were identified based on POI data and building plan data acquired by big data technology. Then, those buildings were categorized into 36 types and ten buildings of different types were selected to be surveyed. Through summarizing the measured data, the hourly occupant density pattern was simplified into three-segment lines of “sharp growth”, “slight growth” and “significant decline” and defined with six characteristic parameters $t_1$, $t_2$, $D_2$, $t_3$, $D_3$, and $t_4$. Through correlation analysis, we found that $D_2$ and $D_3$ are mainly affected by $F$ and $Po$; $t_2$ and $t_3$ are mainly influenced by $Tr$ and $Di$, while $t_3$ also by $Po$ and $De$. The regression model between $t_2$, $D_2$, $t_3$, $D_3$ and their main factors were built, whereas the model of $D_3$ is not consistent with the analysis before as a result of limited amount of data. This method can not only be used as input to an urban simulation model to improve its accuracy and a supplement for the database of urban buildings, but also make sense for the site selection of commercial buildings. Going forward we will increase the amount of measured data. Quantitative testing of the adaptability of the model is also required.
References

[1] Agarwal Y., Balaji B., Gupta R., Lyles J., Wei M., and Weng, T 2010 Occupancy-driven energy management for smart building automation *ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building* 1-6 ACM

[2] Ahmed K., Akhondzada A., Kurnitski J., and Olesen, B 2017 Occupancy schedules for energy simulation in new pren16798-1 and iso/fdis 17772-1 standards *Sustainable Cities & Society* 35

[3] Heiple S., and Sailor D. J. 2008 Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles *Energy & Buildings* 40(8) 1426-1436.

[4] Lei F., Changchun T, Shouxiang L. and Deming K The statistical analysis and modeling of commercial buildings *Journal of Fire Science* 18(3) 130-137

[5] Wang C., Yan D., and Jiang Y. 2011 A novel approach for building occupancy simulation. *Building Simulation* 4(2) 149-167

[6] Yiming S. 2007 Analysis of correlations between variables using spss software *Journal of Xinjiang Education College* 23(2) 120-123

[7] Yujun Y, and Xiaodong S. 2004. The comparison of accessibility methods based on gis *Journal of construction science and engineering* 21(4) 27-32

[8] Zhao W., and Qingquan L. (2011). Extracting hierarchical landmarks from urban poi data. *Journal of Remote Sensing.*

Acknowledgments

This paper is financially supported by the Ministry of Science and Technology of China (project number: 2016YFC0700102).