

Spatial Interpretation of Urban Hotspots based on SNS Data and Machine Learning

任, 宇杰

<https://hdl.handle.net/2324/4784382>

出版情報：九州大学, 2021, 博士（工学）, 課程博士
バージョン：
権利関係：

Spatial Interpretation of Urban Hotspots based on SNS Data and Machine Learning

Yujie Ren

2021

Contents

List of Figures.....	7
List of Tables.....	10
Acknowledgement.....	12
Abstract.....	14
Chapter 1	
General Introduction.....	17
1.1 Research background.....	17
1.1.1 Aggregation of urban elements and generation of hotspots.....	17
1.1.2 Link between urban hotspots and urban system.....	17
1.1.3 Advanced spatial analysis with Big-Data and AI.....	18
1.2 Literature review.....	19
1.2.1 Study on the concept and importance of urban hotspots.....	19
1.2.2 Study on the identification of urban hotspots.....	20
1.2.3 Study on the application of Big-Data and AI in urban analysis...20	
1.3 Research objectives.....	21
1.4 Research structure.....	22

References.....25

Chapter 2

A New Urban Hotspot Identification Method29

2.1 Origin of this new conception29

2.1.1 Limitations in previous urban hotspots identification method ...29

2.1.2 Advance of Big-Data era and SNS platform.....33

2.2 Procedure of new urban hotspot identification method35

2.2.1 Overall technological process.....35

2.2.2 SNS data collection and pre-processing.....37

2.2.3 KANN-DBSCAN, an advanced clustering algorithm42

2.2.4 Concave hull, an algorithm to determine the boundary45

2.2.5 Hierarchical clustering, an algorithm to extract function46

2.3 Empirical experiment of new identification method46

2.3.1 Urban hotspots identification results46

2.3.2 Accuracy testing of the new method.....49

2.4 Summary on new urban hotspot identification method51

References.....53

Chapter 3

Spatial-temporal Distribution of Urban Hotspot57

3.1 Necessity of clarifying urban hotspots dynamically.....57

3.2 Urban hotspots at different time periods and dates59

 3.2.1 Division of time period and date59

 3.2.2 Identification results on different time61

3.3 Emerging and changing of urban hotspots.....63

 3.3.1 Basic types of change of urban hotspots.....63

 3.3.2 Changes of urban hotspots across time period and date64

 3.3.3 Rules of the emerging and changing of urban hotspots.....66

 3.3.4 Attribute variations under the rules of urban hotspots.....68

3.4 Relationship between the rules and built environment71

 3.4.1 Model construction72

 3.4.2 Output of binary logistic regression models.....75

 3.4.3 Discussions on the urban hotspot rules and built environment ..77

3.5 Summary on the spatial-temporal distribution of urban hotspot.....78

References.....80

Chapter 4

Street-view Impression of Urban Hotspots83

4.1 Additional landscape value of urban hotspots.....	83
4.2 Extraction of street-view impression using SNS data.....	85
4.2.1 Street-view impression extraction process	85
4.2.2 SNS online-review data collection.....	87
4.2.3 Filter landscape from online-review photos	89
4.2.4 Extract evaluation from online-review text.....	91
4.2.5 Extraction results of street-view impression.....	93
4.3 Comparison on the landscape preference	96
4.3.1 Landscape preference comparison procedure.....	96
4.3.2 Construction and performance of the models	98
4.3.3 Extraction of urban landscape cognition difference	100
4.4 Summary on the street-view impression.....	103
References.....	106

Chapter 5

Simulation of Urban Hotspots.....	110
5.1 Significance of urban hotspot simulation	110
5.2 Methodology on urban hotspot simulation	111
5.2.1 Urban spatial simulation in previous works.....	111

5.2.2 Data preparation for urban hotspot simulation115

5.2.3 Algorithms for urban hotspots simulation..... 117

5.3 Urban hotspot simulation results 121

5.3.1 Simulation results from multinomial logistic regression 121

5.3.2 Simulation results from random forests 125

5.3.3 Simulation results from convolutional neural network 127

5.4 Comparison of simulation accuracy of urban hotspots..... 131

5.5 Summary on the urban hotspot simulation 134

References..... 136

Chapter 6

Conclusions 139

6.1 Main findings..... 139

6.2 Academic contributions 142

6.3 Future directions and recommendations..... 144

List of Figures

Figure 1-1 Overall technological process of this study

Figure 2-1 Application of theoretical model in identifying urban hotspots

Figure 2-2 Mainstream algorithms in identifying urban hotspots

Figure 2-3 Samples of geo-tagged SNS platform data

Figure 2-4 Workflow of acquiring data from API

Figure 2-5 Overall technological process of the new urban hotspot identification method

Figure 2-6 Time tag acquisition method of SNS data from Sina Weibo

Figure 2-7 Samples of the pre-processing SNS data

Figure 2-8 Principle of DBSCAN

Figure 2-9 Main process steps of DBSCAN algorithm

Figure 2-10 Algorithms to convert clouds to polygons

Figure 2-11 Results of urban hotspots identification with new method

Figure 2-12 Function of urban hotspots

Figure 2-13 Urban hotspots extracted from POI sample datasets

Figure 3-1 Empirical cases of variability of urban hotspots

Figure 3-2 Spatial distribution of check-in points at different periods

Figure 3-3 Number of check-in points at different periods

Figure 3-4 Proportions of check-in points at different periods

Figure 3-5 Spatial distribution of urban hotspots in downtown of Nanjing

Figure 3-6 Basic changes of urban hotspots

Figure 3-7 Spatial distribution of the basic changes of urban hotspots (Weekday ~ Off day)

Figure 3-8 Spatial distribution of the basic changes of urban hotspots (Morning & Afternoon & Night)

Figure 3-9 Six main rules for the emergence and change of urban hotspots

Figure 3-10 Attributes variations under the emerging rules of urban hotspots

Figure 3-11 Spatial distribution of the 6 rules

Figure 3-12 Data related to urban built environment

Figure 4-1 Research process of extraction of street-view impression

Figure 4-2 SNS data collection sites (location of urban hotspots)

Figure 4-3 Image semantic segmentation algorithm to judge landscape type

Figure 4-4 Cleaning and division standard of landscape elements

Figure 4-5 Comment opinion extraction algorithm

Figure 4-6 Text sentiment analysis algorithm

Figure 4-7 Natural landscape's street-view impression

Figure 4-8 Cultural landscape's street-view impression

Figure 4-9 Research process of extraction of landscape preference

Figure 4-10 Data augmentation tools

Figure 4-11 Accuracy of landscape preference prediction models

Figure 5-1 The evolution of OpenStreetMap community

Figure 5-2 The vector digital map slices collection

Figure 5-3 Sample vector and remote sensing image slices

Figure 5-4 Roadmap of urban hotspot simulation model based on CNN

Figure 5-5 Urban hotspot simulation results from multinomial logistic regression

Figure 5-6 Parameter optimization of random forest models

Figure 5-7 Variable importance estimation from random forest models

Figure 5-8 Urban hotspot simulation results from random forest

Figure 5-9 CNN image classifier based on vector digital map slices

Figure 5-10 Urban hotspot simulation results from vector CNN

Figure 5-11 CNN image classifier based on raster remote sensing slices

Figure 5-12 Urban hotspot simulation results from raster CNN

List of Tables

Table 2-1 Summary of research on the identification of urban hotspots

Table 2-2 Sample twitter data

Table 2-3 Sample Sina Weibo data

Table 2-4 Classification rules of category_name

Table 2-5 Summary of research on the identification of urban hotspots

Table 3-1 Evidence on the impact of urban hotspots on urban system

Table 3-2 Descriptive statistic on the attributes of urban hotspots

Table 3-3 Result of logistic regression models

Table 4-1 Sample SNS online-review dataset

Table 4-2 Input data of prediction models

Table 4-3 Prediction results

Table 4-4 Prediction results

Table 5-1 Factors related to urban built environment

Table 5-2 Urban built environment factors used in regression models

Table 5-3 Numerical elements related to the urban built environment

Table 5-4 Results of multinomial logistic regression

Table 5-5 Simulation accuracy comparison

Acknowledgement

Acknowledgement

Acknowledgement

Foremost, I would like to express my sincere gratitude to my advisor Prof. Shichen Zhao for the continuous support of my Ph.D. study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Takeru Sakai, Prof. Nobuo Mishima, and Prof. Divigalpitiya Prasanna, for their encouragement, insightful comments, and hard questions.

My sincere thanks also go to my fellow lab mates in *Urban Planning Lab, Kyushu University*: Dr. Chen Qi, Dr. Xu Feifan, Dr. Dang Sheng, Che Youlu, Chen Siting, Du Mengge, Ni Molin, Song Jingying, Chi jiemeng, Zhang Yichuan, Tsai Chiayu, Higuchi Go, and Ryuta Yamamoto for the discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last three years. I am also grateful to my friends in Kyushu University: Cui Yue, Zhu Yifan, Wu Wei and Tang Xun for their enlightening me the first glance of research.

Last but not the least, I would like to thank my family: my parents for giving birth to me at the first place and supporting me spiritually throughout my life; and my girlfriend Fan Tianhui for her accompany during my journey of Ph.D..

Abstract

Abstract

Abstract

With the rapid development of cities, the physical elements and population in the urban spaces show a trend of spatial aggregation. Urban hotspots are the products of this phenomenon of aggregation, which have been defined as urban regions undertaking relatively higher densities of urban built-up environment elements and population flows. Prior studies have noted that urban hotspots play important roles as urban sub-centers with different functions (transportation hubs, economic cores, production clusters, etc.) and representatives of the urban landmarks. However, the previous research on the identification methods and occurrence patterns of urban hotspots was limited by the lack of information contained in the data source, while very little was found in the literature on the landscape feature and evaluation of urban hotspots. On this basis, this study takes urban hotspots as the research objects and conducted an in-depth spatial interpretation analysis of urban hotspots with the support of new data sources (SNS data) and new methods (machine learning algorithms). Specifically, this research was undertaken to design an optimal urban hotspot identification method and evaluate the spatial-temporal distribution patterns and landscape characteristics of urban hotspots. In addition, considering the great importance of urban hotspot to urban system, its emergence and attribute characteristics are also predicted. The results of this study illustrated that:

In Chapter 2, this study proposes a new dynamic identification method of urban hotspots based on KANN-DBSCAN, concave hull algorithms, and SNS check-in datasets. Benefiting from the explosive growth of SNS data and the continuous progress of computer vision algorithms, this method can not only accurately grasp the location of urban hotspots from urban space, but also estimate the function, popularity, occurrence date, and occurrence time periods of urban hotspots. This study provides an optimal method for urban researchers to understand hotspots regions in urban systems.

In Chapter 3, this study focuses on the spatial-temporal distribution and characteristics of urban hotspots. Using the urban hotspot identification method proposed in Chapter 2, the location and detailed features of urban hotspots at

different dates (weekdays and off days), time periods (morning, afternoon, and night) are extracted. Afterward, by comparing and summarizing the variation of location and feature of urban hotspots, it has been found that the emergence and change patterns of urban hotspots generally follow six basic rules. This provides a quantitative basis for urban researchers to understand the structure and function of urban spaces dynamically and real-timely.

In Chapter 4, this study realizes the extraction of landscape evaluation and preference on urban hotspots based on online review data from SNS users (images and texts) and machine learning algorithms (NLP and CNN). Following this new approach, the subjective evaluation and preference of visitors and tourists on the landscape of urban hotspots in the case study region (Nanjing, China) are estimated. This study realizes the understanding of the deficiencies of urban hotspots in landscape construction as urban landmarks and external windows and provides method support for urban scene analysis in Big-Data era.

In Chapter 5, this study predicts the occurrence status, date, period, and function of urban hotspots and also compares the performance of different data sources (numerical data, digital map vector data, remote sensing raster map data) combined with different algorithms (regression model, random forest and convolution neural network) in urban hotspot simulation. The results illustrate that the combinations of different data sources and algorithms have different simulation accuracy in various scenarios. This study provides guidance for urban hotspot simulation under different scenarios and objectives.

In general, with the help of SNS data and machine learning algorithms, this study interprets the urban hotspot from a spatial perspective. This study proposes a new urban hotspot identification method, which improves the analysis of spatial-temporal distribution and characteristics of urban hotspots. In addition, this study also realizes the landscape feature analysis and spatial simulation of urban hotspots. This study provides a quantitative basis and method support for urban researchers to better understand the dynamics of urban space and present the application prospect of SNS data and machine learning algorithms in the field of urban research, as well.

Keywords: Urban hotspot; SNS data; Machine Learning Algorithm

Chapter 1
General Introduction

Chapter 1

General Introduction

1.1 Research background

1.1.1 Aggregation of urban elements and generation of hotspots

Elements in urban spaces have shown a trend of high concentration during rapid urbanization (Batty et al. 1989). With the improvement of the regional development level, the economic, cultural, talent, transportation, and even political elements in the urban spaces will show the pattern of shortening the distance between similar elements in spatial distribution and finally focus on specific areas (Jacobs-Crisioni et al. 2014).

Therefore, researchers defined the areas where a greater number of urban elements, events or population gathered relative to other locations as urban hotspots. In the past two decades, the concept of urban hotspots had been extended as urban sub-centers and urban landmarks (Sherman 1995; Li et al. 2018). This urges the recognition and understanding of urban hotspots as a crucial matter for urban planners and designers. Why some places are more popular than others and how these hotspots exist in urban space are the two key questions whose answer may lead to an understanding of the local population's dynamics and the reasons behind the preference for some places over others (Marti et al. 2017).

1.1.2 Link between urban hotspots and urban system

Urban hotspots with different functions are demonstrated to have significant impacts on the urban transportation system, climate environment, and real estate market, respectively. As reported in the literature, Bilkova et al. (2016) found that the aggregation of urban commercial and production factors is the main reason for the rise of regional land prices. The research by Qin and Zhen (2019) also

showed that the popularity of typical urban hotspots such as catering spaces in urban areas will greatly affect the travel patterns of urban residents. What's more, Jia's work (2014) even showed that the hierarchies of urban hotspots could affect the CO₂ emission and air pollution level. This urges the understanding of urban hotspots dynamically as a crucial matter for urban planners and designers.

In general, summarizing the temporal and spatial distribution and characteristics of urban hotspots is not only help to understand the dynamic change pattern of urban built environmental elements and people flow, but also can be used to predict the development trend of urban economy, transportation, and environmental sub-systems.

1.1.3 Advanced spatial analysis with Big-Data and AI

As the core elements of urban space and important influencing factors of urban economy, production, transportation, and environmental systems, urban hotspots have been regarded as important objects in the field of urban spatial analysis since its concept was put forward. Discussions regarding the identification methods of urban hotspots have dominated the research on urban hotspots in recent years. However, the research on urban hotspots based on previous data and methods can not accurately identify the number and scope of hotspots in cities and neglects the dynamics of urban residents and the real situation of using these hotspots.

The outbreak of multi-source social networking service (SNS) data provides new opportunities for urban research. With the advance of the information era, open-sourced SNS data (i.e., Twitter, Flickr, and Weibo) were introduced into urban studies to identify urban regional characteristics and population flow (Yan et al. 2019; Ota et al. 2017; Kitamura et al. 2019). On the other hand, the methods of urban spatial analysis have been comprehensively improved with the continuous popularization and progress in the computer field (especially machine learning algorithms) in recent years. As a complex system, the problems in urban space are not able to be estimated and predicted by a statistical model. At present, machine learning algorithm has achieved good results in the fields of urban functional area discrimination, pedestrian flow prediction, traffic volume prediction and so on (Yu et al. 2021; Cohen and Dalyot 2020; Fan et al. 2019).

Therefore, we believe that SNS data, and machine learning algorithms have great possibilities to help us carry out an in-depth spatial interpretation analysis on urban hotspots.

1.2 Literature review

On the basis of the above three research backgrounds, it has been made clearly that urban hotspots are of great importance to the city and also pointed out the good opportunity to interpret urban hotspots in combination with new data and new methods in the current era. In this chapter, this study will review the literature related to urban hotspots and find the focus and future direction of relevant research.

1.2.1 Study on the concept and importance of urban hotspots

The concept of urban hotspots originated from the field of criminology. Sherman (1995) defined the areas where a greater number of criminal or disorderly events happen relative to other locations as hotspots. At the early stage, urban hotspots are defined as the space where specific types of physical elements or events exist or occur in the cities.

In the field of Humanities (such as human geography and economy), the concept of urban hotspots is extended to "human activity gathering area". Malmberg noted that the nature of human being to be clustered in certain places, leading to national and regional specialization, and the geographic conditions, such as factor endowments, cultures, and business environment, determine business and other types of human activities (Malmberg et al. 1996). Considering that Malmberg also mentioned the clustering phenomenon whereby human activities influence and are influenced by geography will not easily disappear over time (Malmberg et al. 1996), many economists and geographers began to work on this clustering phenomenon.

Overall, although the definitions of urban hotspots in different fields are different and even controversial, urban hotspots could be recognized as the locations where relatively high levels of an attribute or activity happened in this

research and play the roles of urban centers with different functions and urban landmarks. The dominating attribute or activity determines the function of the particular urban hotspot.

1.2.2 Study on the identification of urban hotspots

Discussions regarding the identification methods of urban hotspots have dominated the research on urban hotspots in recent years. At the early stage, theoretical models represented by Porter's diamond model and central place theory are the prevailing methods. Yang (1994) predicted the spatial structure of urban hotspot areas in Beijing and achieved good results based on Porter's diamond model. Afterward, Malmberg et al. (1996) outlined another theoretical urban cluster extraction method based on local economy and knowledge.

With the advance of the information era, open-sourced geo-tagged SNS data (i.e., Twitter, Flickr, and Weibo) were introduced into urban studies to identify urban regional characteristics and population flow (Yan et al. 2019; Ota et al. 2017; Kitamura et al. 2019), which provide new approaches for the identification of urban hotspots. One dominating method in this field could be spatial clustering, and it aims to partition similar spatial data into subclasses, called spatial clusters. Liu et al. (2010) and Shi et al. (2017) introduced spatial clustering algorithms into the identification of urban hotspots. With the point of interest (POI) data from the digital map, urban spaces with relatively higher levels of attributes to neighboring locations were extracted and recognized as urban hotspots.

Regarding the formation of urban hotspots as a spatial process, using surface network methods to extract urban hotspots is another main approach of the identification of urban hotspots. Sadahiro (2001; 2003) sketched a hotspot extraction method based on surface network and topological changes, which enables shape extraction and dynamic observation of hotspots. Hu et al. (2014) further utilized this method and conducted a case study on the variations of urban hotspots using taxicab mobile data. This type of method based on kernel density estimation (KDE) could visualize the underlying pattern of spatial points and works well with mobile objects data.

1.2.3 Study on the application of Big-Data and AI in urban analysis

With the advent of the Big Data era, geo-tagged social network data provide a new perspective for urban hotspots research. Kaplan and Haenlein summarized the challenges and opportunities of utilizing social media data in urban studies. He pointed out the birth of platforms such as Wikipedia, YouTube, Facebook, Second Life, and Twitter, and the arrival of the Web 2.0 era, bringing new vitality and perspective to social life and scientific research (Kaplan and Haenlein 2010; Martí et al. 2019). After that, urban researchers around the world began to use these social network data (especially Yelp and Twitter) to analyze hot issues such as urban disasters, the pressure of urban residents' survival, and the choice of tourists' destinations (Guntuku et al. 2018; Hamstead et al. 2018; Salas-Olmedo et al. 2018). Their research proves that in the era of big data, geo-tagged data from both social networks and traditional ways can help to better understand the spatial structure of urban space.

On the other hand, the methods of urban spatial analysis have been comprehensively improved with the continuous popularization and progress in the computer field (especially machine learning algorithms) in recent years. As a complex system, the problems in urban space are not able to be estimated and predicted by statistical model. Combining the large volume of SNS data and the high performance of machine learning methods, urban spatial analysis and machine learning models match each other very well. At present, machine learning algorithm has achieved good results in the fields of urban functional area discrimination, pedestrian flow prediction, traffic volume prediction and so on (Yu et al. 2021; Cohen and Dalyot 2020; Fan et al. 2019).

1.3 Research objectives

Therefore, after summarizing the research background and literature review of urban hotspots, it is believed that there is still room for improvement in the identification methods of urban hotspots in relevant studies, and there is a lack in the research about the dynamic change patterns and landscape characteristics of urban hotspots, as well. At the same time, the emerging SNS data and machine learning algorithms, which are highly consistent with the urban spatial analysis,

are able to provide new opportunities to improve and enhance urban hotspot research. This study is done with the following research objectives:

- *Design a new SNS-based urban hotspot identification method in this research. This method is proposed to obtain the location, shape, function, and popularity of urban hotspots real-timely.*
- *Clarify how urban hotspots emerge and change. The location and attribute changing trends of urban hotspots spanning different time periods and dates will be summarized.*
- *Extract the evaluation of tourists and visitors on the landscape of urban hotspots that play the roles of urban landmarks and external windows with the help of SNS online review data and machine learning algorithms.*
- *Simulate the emergence and attributes of urban hotspots. The simulation accuracy under different data sources and algorithms will also be compared.*

1.4 Research structure

The specific contents of all chapters are detailed as follows:

In Chapter 1, the basic definition and importance of urban hotspots are introduced. At the same time, the progress and shortcomings of relative research in the field of urban hotspots are summarized as well. According to the findings of the literature review, the research objectives of this study are formally put forward, and the research scheme is preliminarily designed.

In Chapter 2, this study introduces a new urban hotspot identification method, which can extract the location, shape, function, and popularity of urban hotspots from urban space real-timely by SNS datasets, KANN-DBSCAN, and concave hull algorithms. Specifically, this study presents the detailed operation flow of this new method, including data source, data processing, and recognition results. Moreover, the accuracy, resource consumption, advantages, and disadvantages of the new method are also discussed.

In Chapter 3, this study focuses on the spatial-temporal distribution and variation characteristics of urban hotspots, which have been proved to affect urban economic, transportation, and environmental systems. The urban hotspots identification method designed in Chapter 2 is applied in a case study (the downtown of Nanjing City, China), and the spatial distribution of urban hotspots at different time periods, on different days is grasped. Afterward, this study further summarizes the location and characteristic variations of urban hotspots across different times in the case study area into rules that reflect the spatial-temporal patterns of urban hotspots.

In Chapter 4, this study turns to shed on the landscape features of urban hotspots. In addition to being identified as the sub-centers of the cities, urban hotspots also play the role of the gathering places of urban residents and attract a large number of local and foreign tourists and visitors. In order to find out the tourists' evaluation of the scenery of urban hotspots, this study collects the online word-of-mouth review data from SNS platforms and builds machine learning models to estimate the evaluation results. Moreover, the landscape preference of residents with different properties was also compared.

In Chapter 5, this study attempts to predict the emergence and attributes of urban hotspots. This study realizes the simulation on the location and various characteristics of urban hotspots using machine learning algorithms under the condition of obtaining open-sourced digital maps and satellite remote sensing image data. Moreover, this study compares the accuracy and efficiency of urban hotspots predicting approaches based on data from different sources and different algorithms to summarize the optimal scheme.

In Chapter 6, the main findings on urban hotspots' identification, spatial-temporal distribution, landscape characteristic, and simulation will be summarized. And the academic contributions, innovation, and future direction of this study will be further discussed as well.

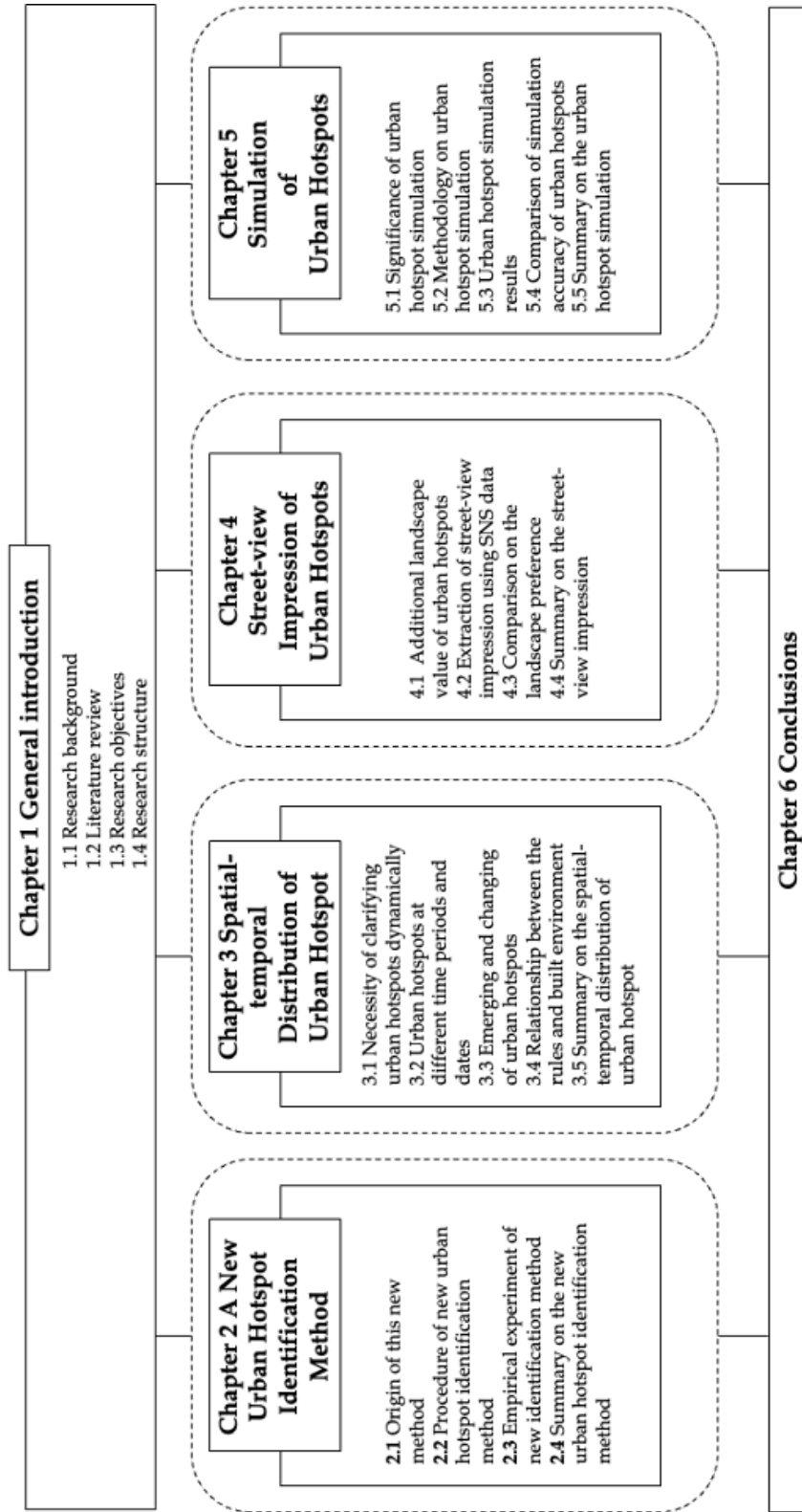


Figure 1-1 Overall technological process of this study

References

- 1) Batty, M., Longley, P., & Fotheringham, S. (1989). Urban growth and form: scaling, fractal geometry, and diffusion-limited aggregation. *Environment and planning A*, 21(11), 1447-1472.
- 2) Bilková, K., Krizan, F., & Barlík, P. (2016). Consumers preferences of shopping centers in Bratislava (Slovakia). *Human Geographies*, 10(1), 23.
- 3) Cohen, A., & Dalyot, S. (2020). Machine-learning prediction models for pedestrian traffic flow levels: Towards optimizing walking routes for blind pedestrians. *Transactions in GIS*, 24(5), 1264-1279.
- 4) Fan, Z., Liu, C., Cai, D., & Yue, S. (2019). Research on black spot identification of safety in urban traffic accidents based on machine learning method. *Safety science*, 118, 607-616.
- 5) Guntuku, S. C., Buffone, A., Jaidka, K., Eichstaedt, J. C., & Ungar, L. H. (2019, July). Understanding and measuring psychological stress using social media. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, pp. 214-225).
- 6) Hamstead, Z. A., Fisher, D., Ilieva, R. T., Wood, S. A., McPhearson, T., & Kremer, P. (2018). Geolocated social media as a rapid indicator of park visitation and equitable park access. *Computers, Environment and Urban Systems*, 72, 38-50.
- 7) Jacobs-Crisioni, C., Rietveld, P., & Koomen, E. (2014). The impact of spatial aggregation on urban development analyses. *Applied Geography*, 47, 46-56.
- 8) Jia, T., Carling, K., & Håkansson, J. (2013). Trips and their CO₂ emissions to and from a shopping center. *Journal of Transport Geography*, 33, 135-145.
- 9) Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, 53(1), 59-68.

- 10) Kitamura, T., Honma, K., & Imai, K. (2019). Tourism characteristics throughout Japan determined from geotagged Flickr photo data: Interest analysis of foreign tourists based on estimated place of residence and tag clustering. *Journal of Architecture and Planning (Transactions of AIJ)*, 84(755), 187-197.
- 11) Li, J., Long, Y., & Dang, A. (2018). Live-Work-Play Centers of Chinese cities: Identification and temporal evolution with emerging data. *Computers, Environment and Urban Systems*, 71, 58-66.
- 12) Liu, H., Wang, L., Sherman, D., Gao, Y., & Wu, Q. (2010). An object-based conceptual framework and computational method for representing and analyzing coastal morphological changes. *International Journal of Geographical Information Science*, 24(7), 1015-1041.
- 13) Malmberg, A., Sölvell, Ö., & Zander, I. (1996). Spatial clustering, local accumulation of knowledge and firm competitiveness. *Geografiska Annaler: Series B, Human Geography*, 78(2), 85-97.
- 14) Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2017). Using locative social media and urban cartographies to identify and locate successful urban plazas. *Cities*, 64, 66-78.
- 15) Ota, K., Imai, K., & Honma, K. (2017). A study on identification of regional characteristics based on temporal-spatial analysis of geotagged-tweet data. *Journal of Architecture and Planning (Transactions of AIJ)*, 82(731), 283-289.
- 16) Qin, X., Zhen, F., & Gong, Y. (2019). Combination of big and small data: Empirical study on the distribution and factors of catering space popularity in Nanjing, China. *Journal of Urban Planning and Development*, 145(1), 05018022.
- 17) Sadahiro, Y. (2001). Analysis of surface changes using primitive events. *International Journal of Geographical Information Science*, 15(6), 523-538.

- 18) Sadahiro, Y. (2003). Stability of the surface generated from distributed points of uncertain location. *International Journal of Geographical Information Science*, 17(2), 139-156.
- 19) Salas-Olmedo, M. H., Moya-Gómez, B., García-Palomares, J. C., & Gutiérrez, J. (2018). Tourists' digital footprint in cities: Comparing Big Data sources. *Tourism Management*, 66, 13-25.
- 20) Sherman, L. W. (1995). Hotspots of crime and criminal careers of places. *Crime and place*, 4, 35-52.
- 21) Shi, B., Zhao, J., & Chen, P. J. (2017). Exploring urban tourism crowding in Shanghai via crowdsourcing geospatial data. *Current Issues in Tourism*, 20(11), 1186-1209.
- 22) Yan, L., Duarte, F., Wang, D., Zheng, S., & Ratti, C. (2019). Exploring the effect of air pollution on social activity in China using geotagged social media check-in data. *Cities*, 91, 116-125.
- 23) Yang, W. (1994). The retailing and services center and network of Beijing: Then, now and long before. *Acta Geographica Sinica*, 49(1), 9-17.
- 24) Yu, Z., Jing, Y., Yang, G., & Sun, R. (2021). A new urban functional zone-based climate zoning system for urban temperature study. *Remote Sensing*, 13(2), 251.

Chapter 2

A New Urban Hotspot Identification Method

Chapter 2

A New Urban Hotspot Identification Method

2.1 Origin of this new conception

Firstly, the reason and opportunity of proposing a new urban hotspot identification method will be introduced. Generally, the improvement rooms within the existing identification methods and the rise of SNS platforms and Big-Data era are concluded as the two main factors.

2.1.1 Limitations in previous urban hotspots identification method

As discussed in chapter 1.2, studies regarding the identification methods of urban hotspots have dominated the research on urban hotspots in recent years. However, due to the different definitions of urban hotspots in various fields, the requirements for accuracy, efficiency, and content of urban hotspots identification results are also various. In summary, the identification method of urban hotspots has experienced the transformation from theoretical models to quantitative algorithms.

(1) Period I: Theoretical models

Around the 1990s, researchers in the field of urban planning and economics took the lead in discovering that various elements and people in urban space will show a trend of aggregation (Li 2020). To accurately grasp the occurrence pattern of this spatial agglomeration phenomenon, researchers have introduced the central place theory, diamond model, and other theoretical models into urban hotspot research (Berry & Garrison 1958; Porter 2011; Oliver 2001). At this stage, the main data source of urban hotspot identification is the two-dimensional map. Researchers analyze and predict the spatial distribution of urban hotspots by combining theoretical models and personal experience. As shown in Figure 2-1, a case of application of the theoretical model in identifying urban hotspots is

presented. After making statistics on the distribution of political, demographic, and economic resources in Beijing in 1980, Yang et al. (1994) used Varignon's Theorem to analyze and predict the distribution of urban business centers in Beijing in the 1990s. After testing, the prediction has achieved good results, which is an case of urban hotspot identification based on the theoretical model.

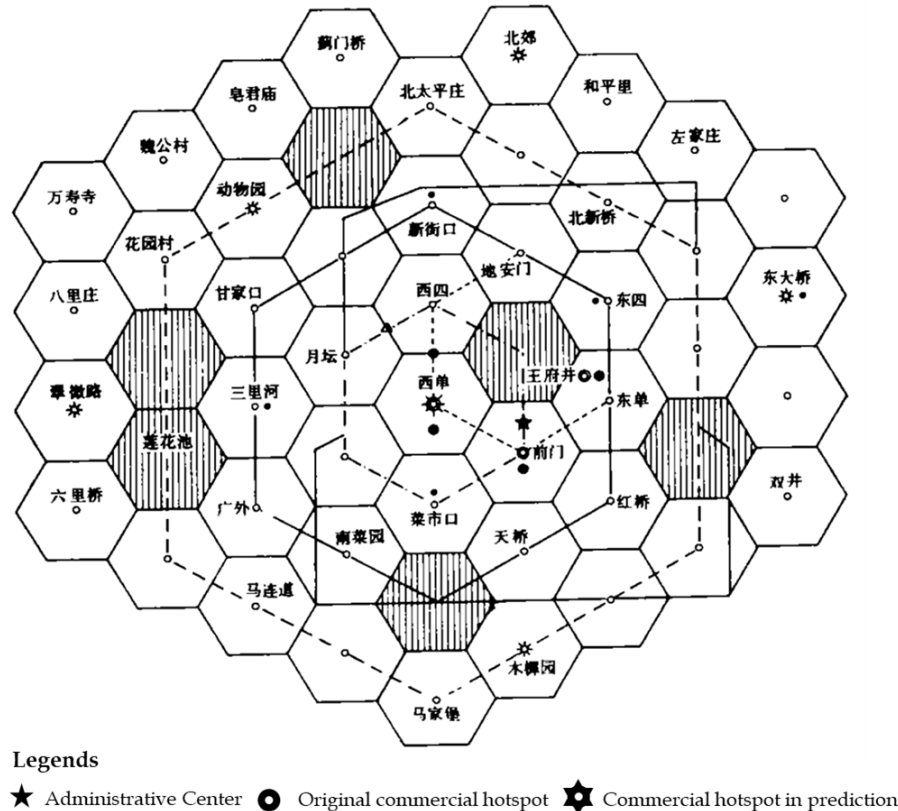


Figure 2-1 Application of theoretical model in identifying urban hotspots

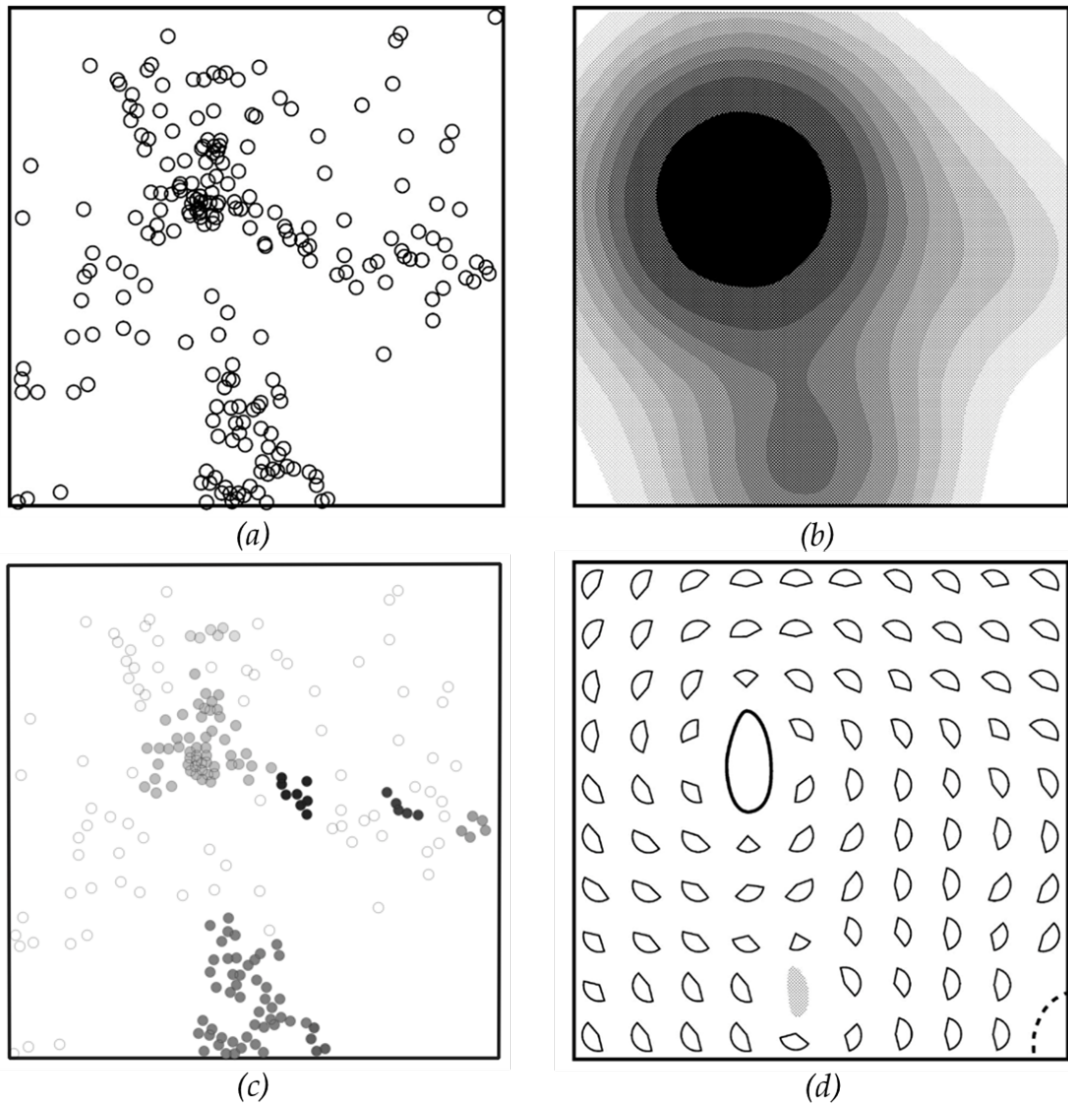
A case from 1980's Beijing based on Varignon's Theorem

Source: Yang et al. 1994

(2) Period II: Quantitative algorithms

With the advance of computer technology and the integration of geographical spatial analysis technology and urban research, utilizing quantization algorithms to calculate the distribution and feature of urban hotspots from vector datasets such as digital maps had gradually become the mainstream of urban hotspot

identification methods. Among them, kernel density, point cloud clustering, and surface network are the three most representative algorithms. Figure 2-2 shows the implementation principles of three mainstream algorithms and the results of urban hotspot identification (Figure 2-2).



Legends

- Point
- Kernel density
- Clustering results
- ▭ Slope

Figure 2-2 Mainstream algorithms in identifying urban hotspots:

(a): sample points; (b) kernel density; (c) clustering (d) surface network

These algorithms realize the identification of urban hotspots by analyzing the discrete distribution, distance, and regional variation trend of POI points from the digital map dataset respectively. After reviewing the urban hotspots identification methods at different time periods, it is found that there are still gaps within the research on the variations of urban hotspots across time, and there is still room for improvement in the identification methods of urban hotspots (Table 2-1).

Table 2-1 Summary of research on the identification of urban hotspots

Representative	Data	Methods	Distribution	Trend	Shape	Area	Popularity	Function
Yang	2D map	Theoretical model	O	×	×	×	O	×
Malmberg	2D map	Theoretical model	O	×	×	×	O	×
Liu	POI data	Clustering	×	×	O	O	×	×
Shi	POI data	Kernel density	O	O	×	×	×	O
Sadahiro	NTT data	Surface network	O	O	O	O	×	×
Hu	Taxicab	Surface network	O	O	O	O	O	×
New method			O	O	O	O	O	O

Note: Each row in the table represents one type of urban hotspot identification method. O and × respectively represent that the method has / does not have the function corresponding to the column where the symbol is located.

On the one hand, the distribution trend (i.e., location) of urban hotspots can be roughly grasped by using theoretical methods and kernel density estimation. And the specific distribution location of urban hotspots can be accurate to POI cloud-point groups with spatial clustering algorithms. However, these two methods can not accurately obtain the boundary of urban hotspots, so it is unable to estimate the actual attributes of each urban hotspot under these circumstances. On the other hand, the urban hotspot extraction method based on surface network and topological changes enable shape extraction and dynamic

observation of hotspots. However, the principle and implementation difficulty of this kind of surface-network-based algorithm is much higher than the above two kinds of quantization algorithms.

At the same time, the existed urban hotspot identification methods proposed in the research were all not able to achieve the key attribute extraction (popularity, function) and realize the dynamic analysis on urban hotspots. Therefore, this inspired me to propose an easy-to-operate urban hotspot identification method. This study designs the implementation principle and process of this new method are as simple as possible, and the location, scope, popularity, and function of urban hotspots can be obtained in real-time and accurately.

2.1.2 Advance of Big-Data era and SNS platform

As discussed in the last chapter, the existing urban hotspot identification methods do have room for improvement. Among them, this study is able to accurately extract the scope and popularity of urban hotspots by innovating and modifying the existed spatial clustering algorithm. However, to further realize the functional identification of urban hotspots, and even analyze the dynamics of urban hotspots in real-time and dynamically, it is impossible to use only the digital map POI data and public transport data that frequently used in the past.

Under these circumstances, the advent of the Big-Data era has brought new opportunities for the research of urban hotspots. The Internet has penetrated every corner of daily life. One of the most widely used applications across the age span in Big-Data era should be social network service (SNS) platforms (Wang et al. 2014), which are member-based online communities where users often begin by posting basic information about themselves -referred to as "Profiles"- and then communicate with other members in a variety of ways and on different topics. When SNS users communicate with other platform members, they will publish online review contents with spatial and temporal labels to record their daily, work, and leisure activities (Kim et al. 2021). These data, including location, time, and SNS user activity information of publication, record the spatial-temporal pattern of urban space used by urban residents (Figure 2-3). If these SNS data can be obtained in large quantities real-time, it can bring new opportunities for the identification of urban hotspots.

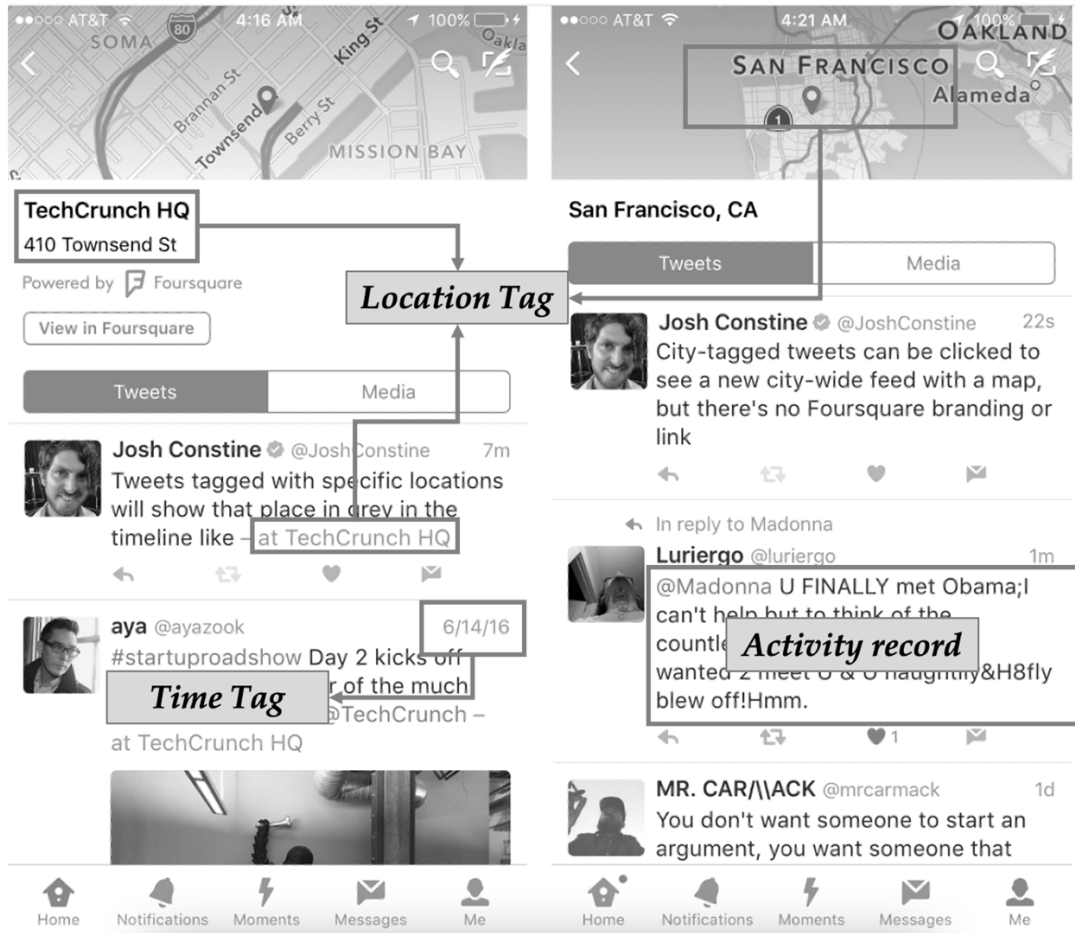


Figure 2-3 Samples of geo-tagged SNS platform data

A piece of sample data in Foursquare platform

Fortunately, in the era of Big-Data, SNS platforms have opened these data containing human activity information to researchers and developers to a certain extent through APIs. An application programming interface (i.e., API) is originally a connection between computers or between computer programs. For social network platforms, API interface is a medium to mobilize platform web services. Researchers and developers can directly access the background database of social networking platform by registering API account and organizing exclusive datasets under the premise of API rules. Figure 2-4 presented the basic principle and workflow of API.

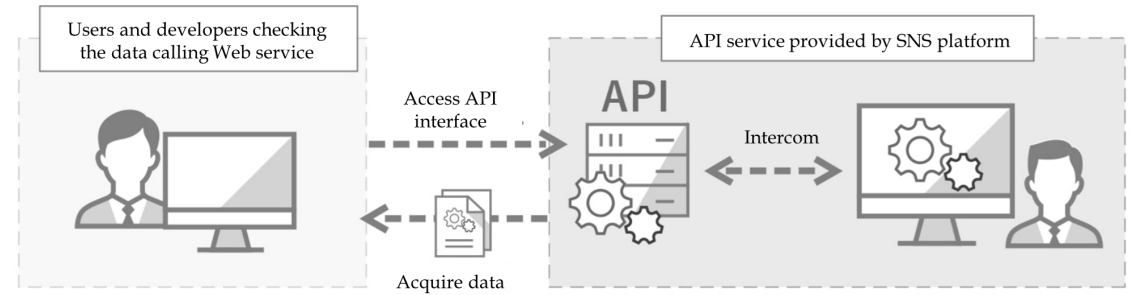


Figure 2-4 Workflow of acquiring data from API

Source: Iбата 2021

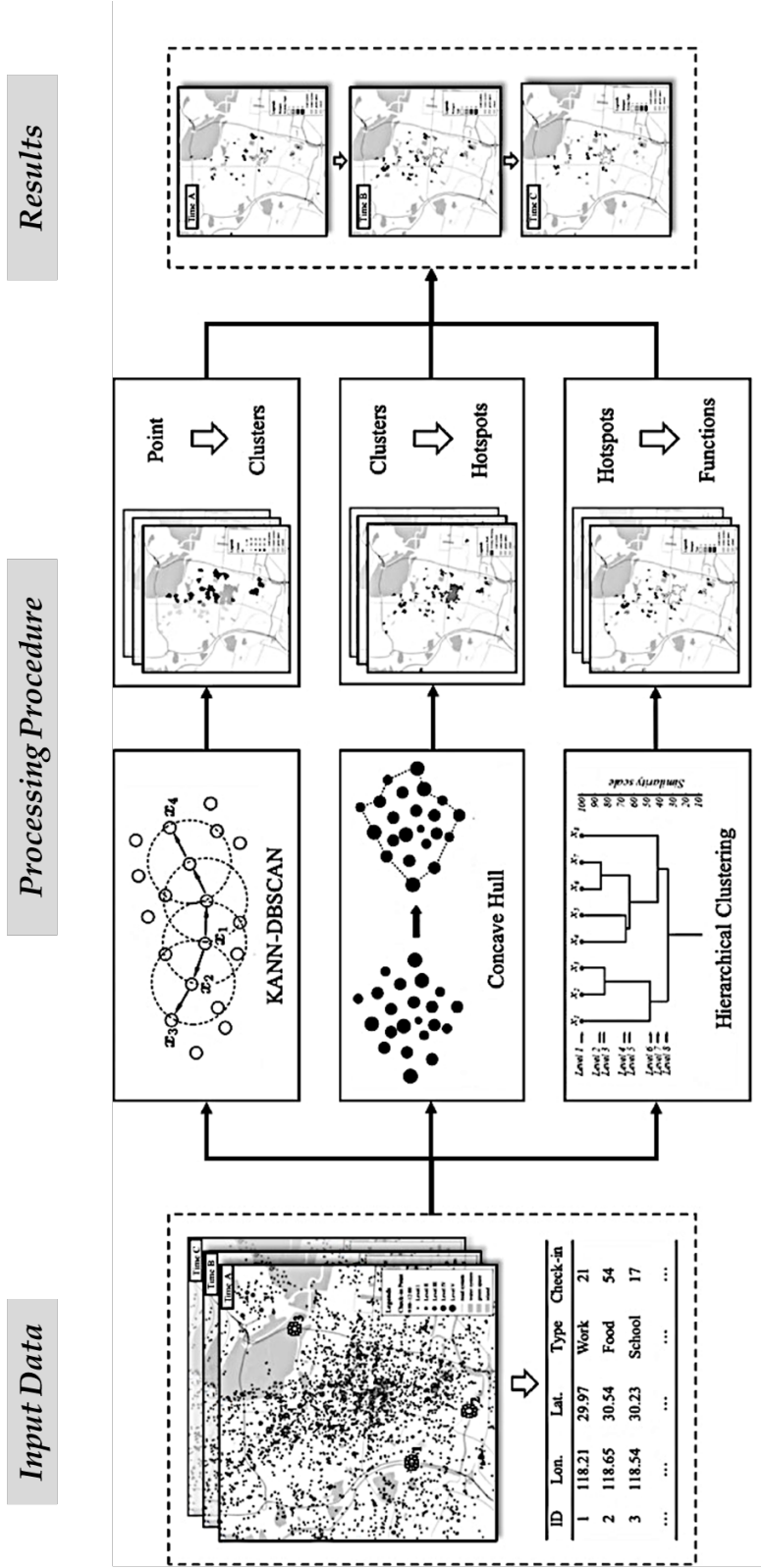
In summary, the outbreak of SNS datasets and their accessibility bring new opportunities for the identification of urban hotspots, which leads to the birth of the new concept of urban hotspots identification method proposed in this study.

2.2 Procedure of new urban hotspot identification method

By far, this study formally proposes a new urban hotspot identification method designed in this study. By improving the parameter setting method of clustering algorithm and integrating other domain algorithms, this method fills the deficiency that the traditional urban hotspot recognition algorithms cannot extract the attributes such as range, function, and popularity of urban hotspots. By replacing the metadata of the recognition algorithm with SNS check-in data from digital map and public transport data, the dynamic and real-time extraction of urban hotspots is realized. Thus, in this part, the details of the new method will be introduced step by step.

2.2.1 Overall technological process

Firstly, Figure 2-5 presents the workflow of the overall technological process of the new urban hotspot identification method:



Dynamic, real-time trend

Upgrade the parameter setting and application method of the existing clustering algorithm

Introducing SNS data into urban hotspots identification

Figure 2-5 Overall technological process of the new urban hotspot identification method

In the first step, the answer on ‘How to get and pre-process the datasets from SNS platform?’ is thoroughly discussed. The access to SNS metadata and the elimination rules of invalid data will be introduced in detail. In the second step, the main topic is the upgrade of the clustering algorithm. The main target of this step is to find out and fix the existing problem with the KANN-DBSCAN algorithm. In the third step, this study focuses on converting the clusters of SNS datasets into urban hotspots. The two computer vision algorithms are introduced to generate the boundary line of urban hotspots, and the applicability of these two algorithms is also compared. In the fourth step, a new function recognition algorithm of urban hotspots based on hierarchical clustering is added to the urban hotspot identification method proposed in this study. In general, the new method of urban hotspot identification proposed in this study includes the above four steps. In the following chapters, this study will introduce the operation details of each step in turn.

2.2.2 SNS data collection and pre-processing

As the basis of the new urban hotspot identification method, datasets from SNS platform are mainly chosen for two reasons: (1) high dynamic; (2) High information capacity. As a result, when screening the sources of SNS data, this study first needs to ensure that the data provided by the corresponding SNS platform web service meet these two standards. Table 2-2 and 2-3 respectively show the sample data obtained by the API interfaces of the two mainstream SNS platforms (Twitter and Sina Weibo). These examples will illustrate how to judge whether SNS data meets the requirements of this new method.

Table 2-2 Sample twitter data

ID	Content	Time	Type	...Like
31963	#studiolife to find #newmaterials	2014/5/1 16:57	Tweet	... 0
...
31965	safe ways to heal your #acne!	2014/5/1 16:49	Tweet	... 5

API: <https://github.com/twitter-archive/twitter-kit-android/wiki/Access-Twitter's-REST-API>

Table 2-3 Sample Sina Weibo data

ID	Title	Address	Lat.	Lon.	City	Activity type	Pic	Check-in
BFC419D	江边商店	—	99.65	33.75	975	Shopping	0	0
BFC419C	拉卜楞馍馍铺	黄河东路	99.65	33.75	975	Catering	0	0
BFC419F	积石山电焊铺	黄河东路	99.65	33.75	975	Entertainment	0	0
BFC419E	民和三川人家	黄河东路	99.65	33.75	975	Catering	0	0
BFC4199	雷家小吃店	黄河东路	99.65	33.75	975	Catering	0	0

API: <http://open.weibo.com/wiki/2/place/nearby/pois>, Record acquisition time manually

After comparison, it could be found that the SNS data records provided by Twitter API are mainly the tweet data publicly released by Twitter users, including the ID, content, publishing media, and information on its ‘retweet’ and ‘like received’ conditions. The lack of key information about the location of tweet release in these data items and the difficulty of judging the behavior and activities of tweet release make it difficult for Twitter datasets to be directly used to identify urban hotspots. In conclusion, when selecting SNS data for the identification of urban hotspots, it is necessary to ensure that the elements in the original dataset include three categories: recording the time and place of SNS comment release and the ongoing behavior in the process of comment release.

Next, the Sina Weibo dataset, which has been proved to meet the basic requirements of the SNS dataset used for urban hotspot identification, will be used as a sample (Table 2-3). Sina Weibo platform is considered as China’s answer to Twitter. Launched by Sina Corporation on 14 August 2009, it is one of the biggest social media platforms in the world, with over 445 million monthly active users in Q3 2018 (Chen et al. 2020). This study will further present how to preliminarily process these SNS data, specifically how to delete useless information and optimize its data structure. As shown in Table 2-3, each raw check-in data item contained 9 types of attributes: *poiid*, *title*, *address*, *longitude*, *latitude*, *city_code*, *category_name*, *checkin_num*, *photo_num*. After completing the

data collection, this study needs to remove the useless attributes in each piece of data and modify the values of some attributes. Specifically, this study retained the attributes of *latitude*, *longitude*, *category_name*, and *checkin_num* of each original data. Within them, the location of the occurrences of check-in activities is recorded in the attributes of *latitude* and *longitude*. The type and intensity of the occurrences of check-in activities are reflected in the attributes of *category_name* and *checkin_num*, respectively.

In addition, Sina Weibo database is updated in real time. Every time a new check-in record is generated in specific region, the record will be automatically added to the database of the corresponding spaces. The use of Sina SNS platform API is manual real-time collection. This study is able to collect the check-in data generated at time points *a* and *b* respectively and compare the difference of check-in records generated in this time interval to obtain the check-in activity record data generated in time interval from *b* to *a* (Figure 2-6). This is the way to get the time of the occurrences of check-in activities from the Sina Weibo dataset. In this study, the check-in dataset is divided into two types of weekdays (28, 29, 30 April 2014) and off days (27 April and 1, 2, 3 May 2014), and the average value of the check-in quantity generated at each check-in point in 3 time periods (morning 6 ~12 a.m., afternoon 12 ~ 18 p.m. and night 18 ~ 24 p.m.) is calculated.

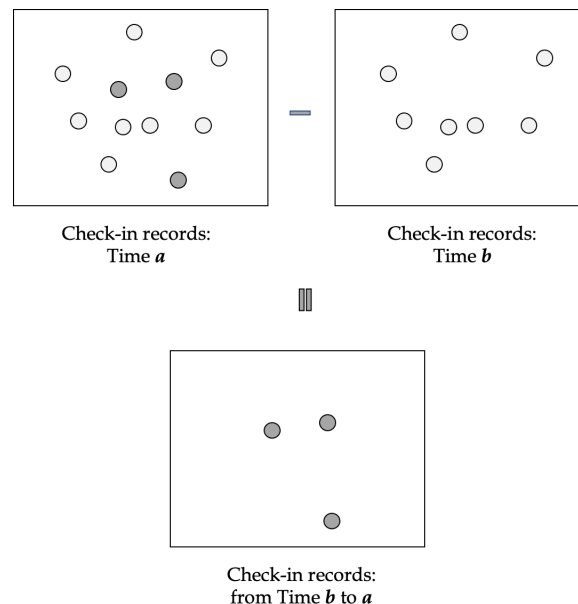


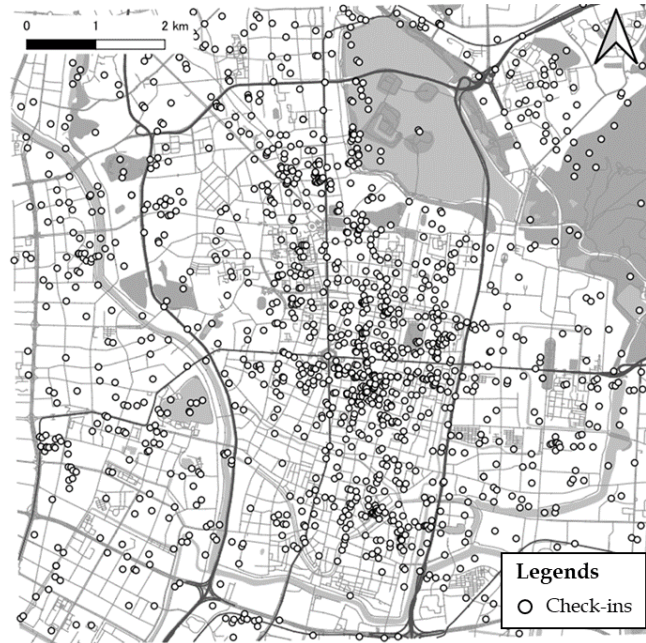
Figure 2-6 Time tag acquisition method of SNS data from Sina Weibo

So far, the Sina Weibo meta dataset has been converted to include the time and place information of check-in behavior, and further record the type and intensity of user check-in activities of Sina Weibo. The original check-in points are divided by the Weibo platform into over 200 activity types, this study merged the original activity types into 7 activity categories of *leisure, service, residence, tourism, transportation, work and others* according to land use regulations, check-in activities and Yan et al.'s work (2019). The classification rules of *category_name* are presented in Table 2-4.

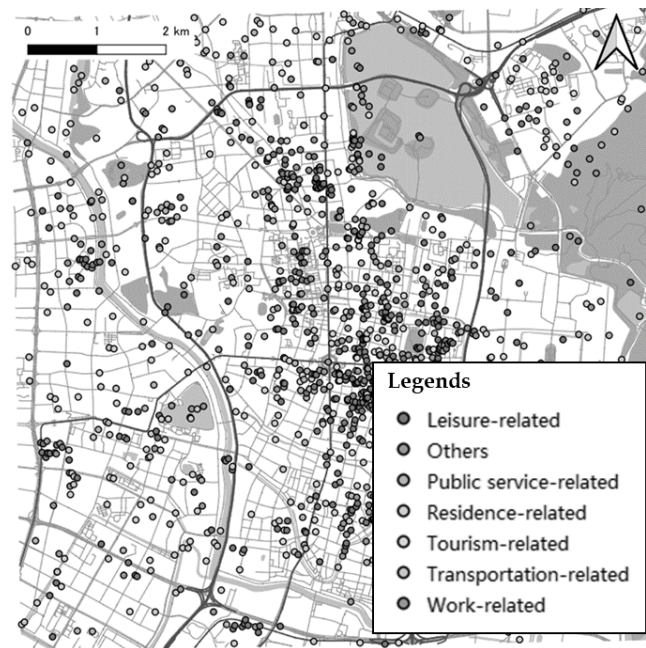
Table 2-4 Classification rules of *category_name*

Activity type	Original label (Category_name)
Leisure	Mall, Restaurant, Gym, Bar, Museum, Gallery, etc.
Service	Municipal government, Police station, etc.
Residence	Community, Apartment, Block, etc.
Tourism	Hotels, Temples, Scenic spots, Landmark buildings, etc.
Transportation	Railway station, Bus station, Subway station, Port, etc.
Work	Factory, Company, Office building, etc.
Others	Religious organizations, Construction sites, etc.

Finally, the metadata from Sina Weibo API is converted into the standard format of *time, location, activity type, and activity density*. Here, this study takes the preprocessing process and results of a set of Sina Weibo formatted original data as an example to show the process and effect of the whole procedures of data cleaning. The main sample check-in data is collected from April 27 to May 3, 2014 in the core area of Nanjing, China, which is between latitudes 32.009° and 32.090° N and longitudes 118.721° and 118.826° S. Finally, a total of 28316 check-in points was obtained. It is worth mentioning that this data cleaning method is not only applicable to data of Sina Weibo platform, but also applicable as long as the format of metadata directly or indirectly contains four basic attributes of: time, location, activity type, and activity density (Figure 2-7).



(a)



(b)

Figure 2-7 Samples of the pre-processing SNS data:

(a): raw data form; (b) cleaned data form

2.2.3 KANN-DBSCAN, an advanced clustering algorithm

Density-based spatial clustering of applications with noise (DBSCAN) is a spatial clustering algorithm which could find dense areas and expand these recursively to find dense arbitrarily shaped clusters (Li et al., 2019). The basic idea of DBSCAN algorithm is to locate regions of high density that are separated from one another by regions of low density (Figure 2-8). This algorithm is utilized to classified spatial points into sub-groups of core points, border points and noise points, and the density-connected core and border points are defined as clusters.

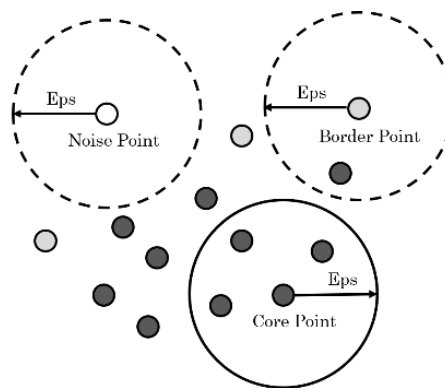


Figure 2-8 Principle of DBSCAN

(Minpts=4)

In DBSCAN algorithm, the circle of each point with itself as the center and Eps as the radius is defined as its 'neighborhood region'. The points where the number of points in its 'neighborhood region' exceeds the set MinPts are defined as the core points. If greater than 0 but less than MinPts, they are set as border points. And those without points involved in 'neighborhood region' are defined as noise points (Figure 2-9 a). Next, after eliminating the noise points, the remaining parts need to be divided into different clusters according to their connection relationship. Specifically, if two points are in each other's 'neighborhood region', they are directly-density-reachable (DDR) to each other (Figure 2-9 b). If two points have a DDR relationship with the third point at the same time, then the relationship between them is defined as density-reachable

(DR) (Figure 2-9 c). Finally, if two points have a DR relationship with the third point at the same time, then the relationship between them is defined as density-connected (DC) (Figure 2-9 d). In the DBSCAN algorithm, the points which are density-connected (DC) to each other will be classified into a cluster.

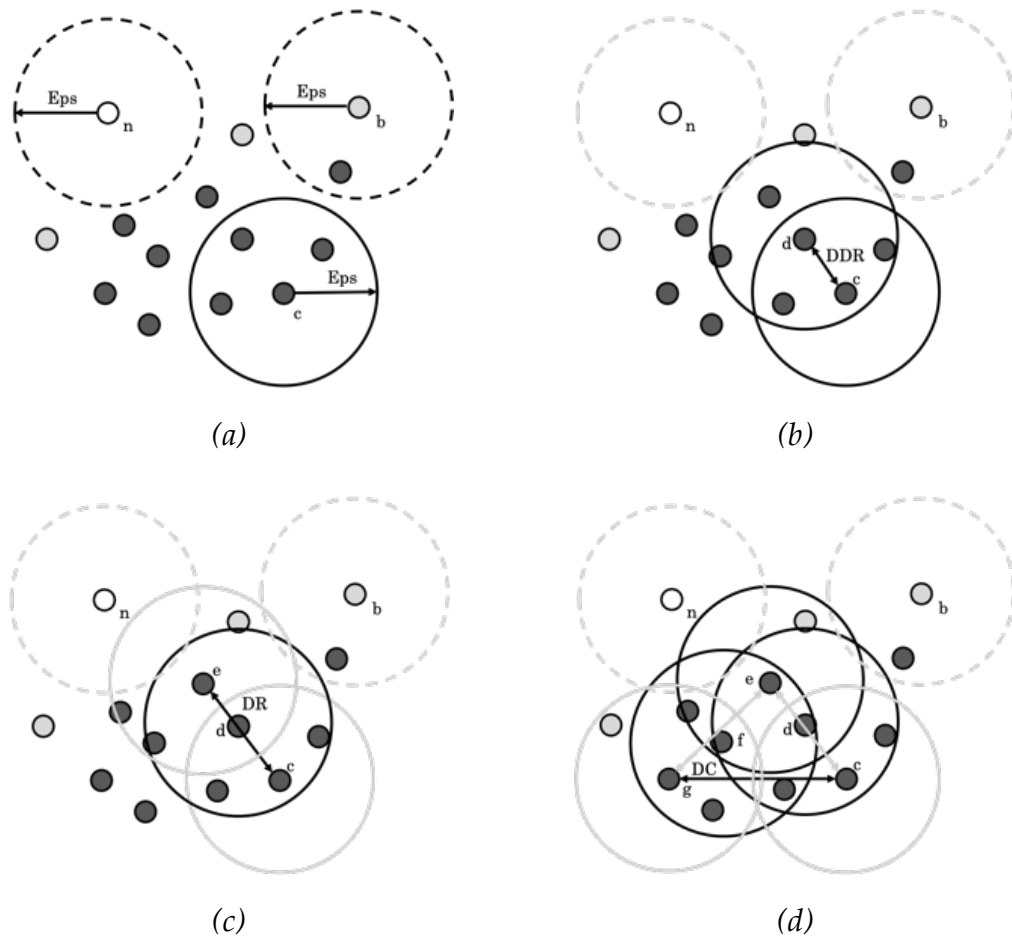


Figure 2-9 Main process steps of DBSCAN algorithm:

(a) Noise, border and core points; (b) Directly-density-reachable (DDR); (c) Density-reachable (DR); (d) Density-connected (DC)

Two main input parameters to DBSCAN are 'Eps' and 'MinPts'. 'Eps' represents radius of the 'neighborhood region' and 'MinPts' represents the minimum number of points that should be contained within that neighborhood. In this study, 'Eps' denotes the radius of the 'neighborhood region' of check-in

points contained in an urban hotspot, and 'MinPts' denotes the threshold (minimum value) of number of check-in points contained in an urban hotspot. When introducing DBSCAN algorithm for real-world case studies, how to determine the value of the core input parameters of 'Eps' and 'MinPts' greatly affects the accuracy of the calculation results. In previous studies, researchers often determined the values of 'Eps' and 'MinPts' based on personal experience or previous research results. In this study, in order to find the hotspots accurately, a new method of self-adaptive parameters determination of DBSCAN algorithm (KANN-DBSCAN) which could generate the optimal values of 'Eps' and 'MinPts' parameters automatically is selected to determine the inputs of DBSCAN.

KANN-DBSCAN algorithm is a method to automatically determine the input parameters Eps and MinPts of the DBSCAN algorithm. Specifically, the candidate Eps parameters are firstly obtained with the help of K-Average nearest neighborhood algorithm. For dataset D:

$$D_{n \times n} = \{Dist(i, j) | 1 \leq i \leq n, 1 \leq j \leq n\} \quad (2-1)$$

Where, $D_{n \times n}$ denotes a real symmetric matrix ($n \times n$); n denotes the number of items contained in dataset D; $Dist(i, j)$ denotes the distance from item i to j in dataset D.

The elements of each row in the distance matrix $D_{n \times n}$ are arranged in ascending order. The distance vector D_0 denotes a composed of the elements in the first column represents the distance from the object to itself, all of which are recorded as 0. Then the k-nearest neighbor distance vector of the data point composed of the elements in column K is D_K . Further, the K-average nearest neighbor distance of vector $\overline{D_K}$ can be obtained by averaging the elements in the vector. And take it as a candidate Eps parameter:

$$D_{Eps} = \{\overline{D_K} | 1 \leq K \leq n\} \quad (2-2)$$

After getting the candidate Eps, the corresponding candidate Minpts parameter will be estimated by mathematical expectation.

$$MinPts = \frac{1}{n} \sum_{i=1}^m P_i \quad (2-3)$$

Where, P_i denotes the number of Eps domain objects of the i_{th} object, and N is the total amount of data in dataset D .

Finally, input each pair of candidate Eps and MinPts in DBSCAN algorithm and observe the output cluster quantity. When the number of generated clusters is the same for the first three consecutive times, the clustering results tend to be stable, and the number of clusters is defined as the optimal number of clusters in KANN-DBSCAN. The Eps and MinPts parameters corresponding to the optimal number of clusters are set as the optimal input parameters of DBSCAN. In addition, the source code of KANN-DBSCAN algorithm is references from Github and Li et al. (2019) (<https://github.com/liyihao17/KANN-DBSCAN>).

2.2.4 Concave hull, an algorithm to determine the boundary

After clustering all check-in points into sub-groups and removing noise points, this study forwards another step to detect the scope of each urban clusters. Drawing the boundary for a cloud of points had always been the main research field for the computer vision (CV). Among all the related CV algorithms, convex hull and concave hull are the two most popular ones:

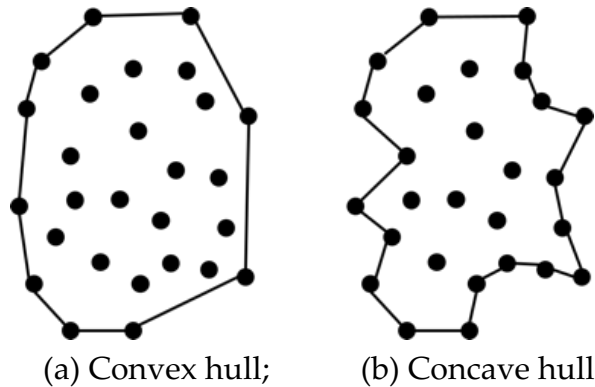


Figure 2-10 Algorithms to convert clouds to polygons

A convex hull of a set of points is the uniquely defined shape that minimizes the area that contain all the points, without having any angle that exceed 180 degrees between two neighboring edges, as seen in Figure 2-10 (a). A concave hull of a set of points could be defined as the shape which minimizes the area of

the containing shape, but allowing any angle between the edges, as seen in Figure 2-10 (b). Since the concave hull had shown a great performance in dealing with real world problems, like detecting the urban boundaries (Cao et al. 2013), in this study, the area covered by the concave hulls of check-in clusters are defined as urban hotspots.

2.2.5 Hierarchical clustering, an algorithm to extract function

Before further investigate the detailed attributes of the identified urban hotspots, this study determines the functional orientations of them according to the proportions of 7 types of contained activities (mention in Table 2-5) with hierarchical clustering (Khalil 2021). The form of basic urban hotspot dataset required for hierarchical aggregation is shown in the following table.

Table 2-5 Summary of research on the identification of urban hotspots

ID	Leisure	Service	Residence	Tourism	Transport	Work	Others
1	0.00%	13.31%	86.69%	0.00%	0.00%	0.00%	0.00%
2	74.33%	12.38%	0.00%	0.00%	0.00%	13.29%	0.00%
...

Note: Each row in the table represents the proportion of 7 types of check-in activities within the boundary line of various urban hotspots.

2.3 Empirical experiment of new identification method

In this chapter, in order to prove the effectiveness of the urban hotspot identification method proposed, this study will present the urban hotspot identification process and results under the designed method and verify its accuracy by comparing the corresponding identification results with that under previous data sources and methods.

2.3.1 Urban hotspots identification results

Using the urban hotspot identification method designed, this study extracted a total of 216 urban hotspots from the check-in record dataset in the core area of Nanjing, China from sub-datasets corresponding to different periods (morning, afternoon and night of weekdays and off days) generated in Chapter 2.2.2.

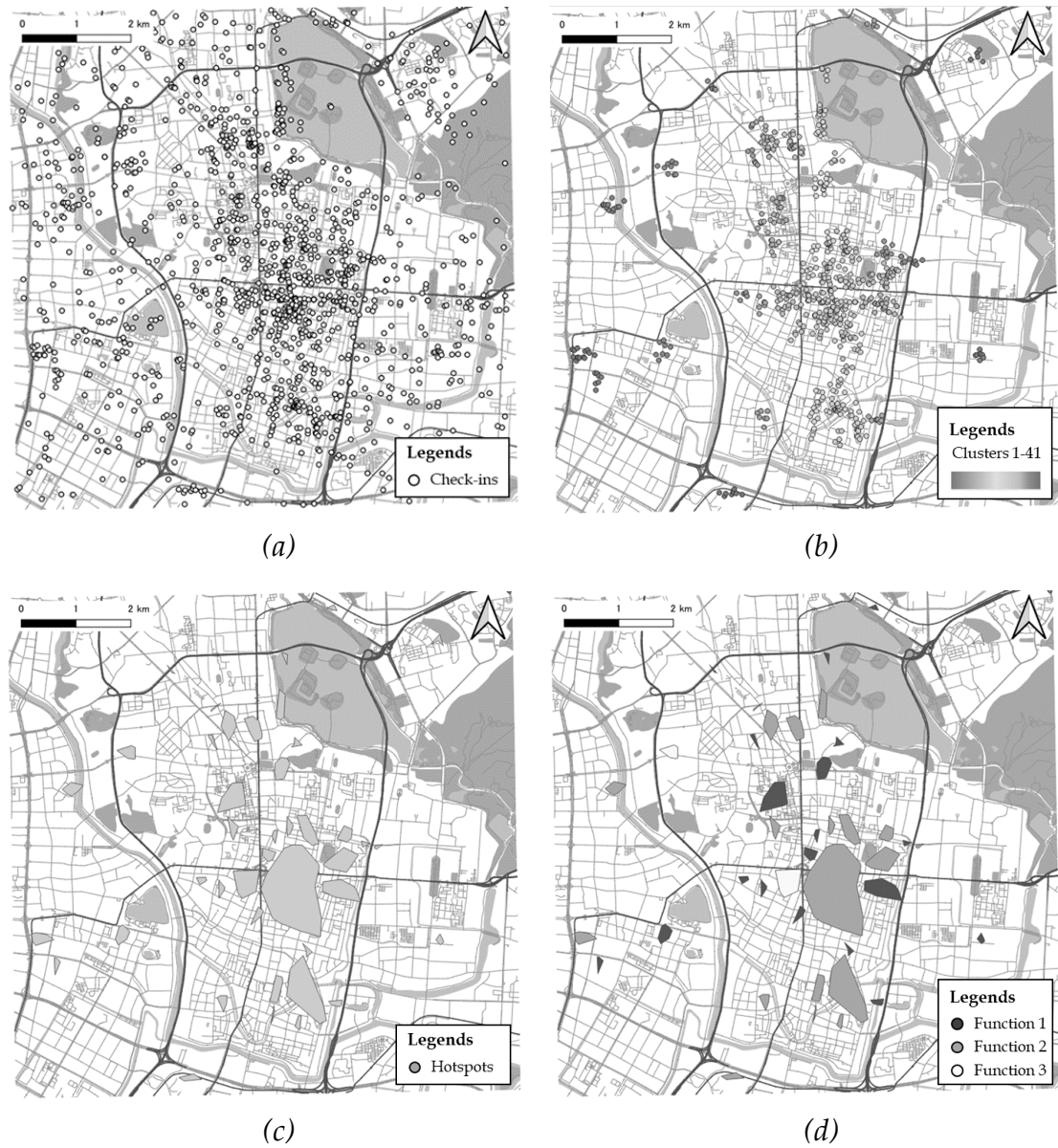


Figure 2-11 Results of urban hotspots identification with new method:

(a) Original data; (b) KANN-DBSCAN result; (c) urban hotspots' boundary identification result; (d) urban hotspots' function identification result

Since the number of check-in records contained in the check-in dataset collected in the afternoon of the off days is the largest in the six subsets, this study will take the process that the check-in points contained in the subset are gradually transformed into urban hotspots as an example to show the operation procedure of the urban hotspots identification method designed in this study (Figure 2-11). and the urban hotspot identification results of other dates and periods can be seen in Figure 3-5. Among them, Figure 2-11 (a) shows the spatial distribution of all check-in records generated in the afternoon of the off days in Nanjing core area. Figure 2-11 (b) shows the clustering results of using KANN-DBSCAN algorithm (parameters: $Minpts=15$, $Eps=0.002$) to convert the check-in point records into clusters of 41. Figure 2-11 (c) shows 41 polygons which represent the 41 urban hotspots identified from the check-in dataset of afternoon of the off days. All the polygons are the concave hulls of the 41 sub-groups of check-in points in Figure 2-11(b). And the area of each concave hull and the quantity of check-ins within the boundary of each concave hull are defined as the area and popularity of urban hotspots. Figure 2-11(d) shows the function identification results from hierarchical clustering. 3 types of urban hotspots' functional orientations are extracted. Afterwards, this study determined the functional orientations of them according to the proportions of 7 types of contained activities (mention in chapter 3.2) with hierarchical clustering. The results were presented in Figure 2-12. The output of classification showed that: (1) By observing the changes of Normalized RMS Distance during the operation of hierarchical clustering (Kabacoff 2010), it is found that the appropriate number of clusters in this study is 3.

As Robert I. Kabacoff (2010) pointed out in the book "R in Action", the optimal number of clusters could be determined by calculating the within sum of normalized rooted mean squared (RMS) distance of the objects in groups. Theoretically, as the number of clusters increases, the number of objects in each category decreases, and the sum of distances between objects will get closer and closer. Therefore, this study introduced the R-package of '*nbclust*' recommended in "R in action" to evaluate the performance of clustering results in cases of the number of clusters is set from 2 to 15 (Charrad et al. 2014). The results illustrated that within the set indices for testing the performance of clustering, 60% (9/15)

proposed 3 as the best number of clusters. Therefore, according to the majority rule, the best number of clusters in this study is 3.

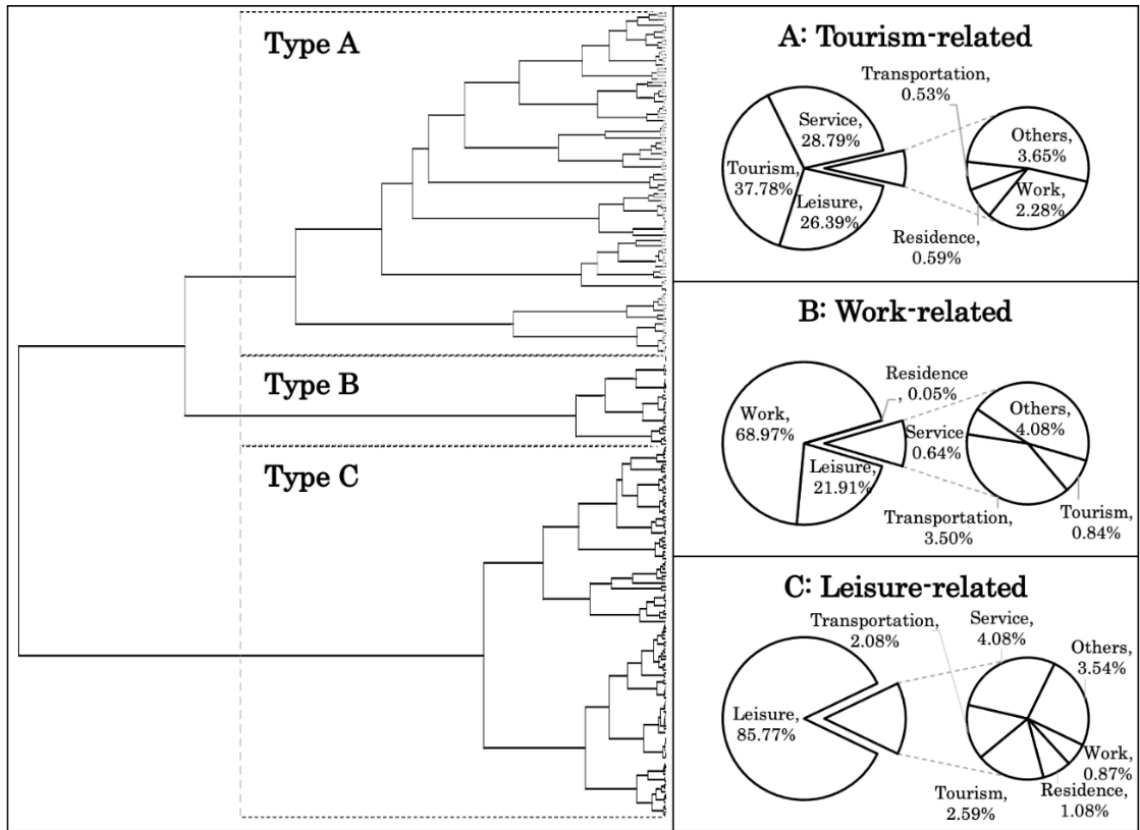


Figure 2-12 Function of urban hotspots

That is to say, the urban hotspots identified in this study can be divided into three categories according to the proportion of human activities involved. The proportion of all types of human activities contained in the three clusters were shown in the 3 pie charts on the right side of Figure 2-12. (2) According to the dominant types of check-in activities, the three types of urban hotspots are responsible for urban tourism zones (Function 1 in Figure 2-11, Tourism: 37.78%; Service: 28.79%; Leisure: 26.39%), urban leisure and entertainment centers (Function 2 in Figure 2-11, Leisure: 85.77%), and urban workspaces (Function 3 in Figure 2-11, Work: 68.79%; Leisure: 21.91%).

2.3.2 Accuracy testing of the new method

In addition, in order to verify the accuracy of urban hotspot identification based on SNS data, POI points from digital map platforms are also obtained.

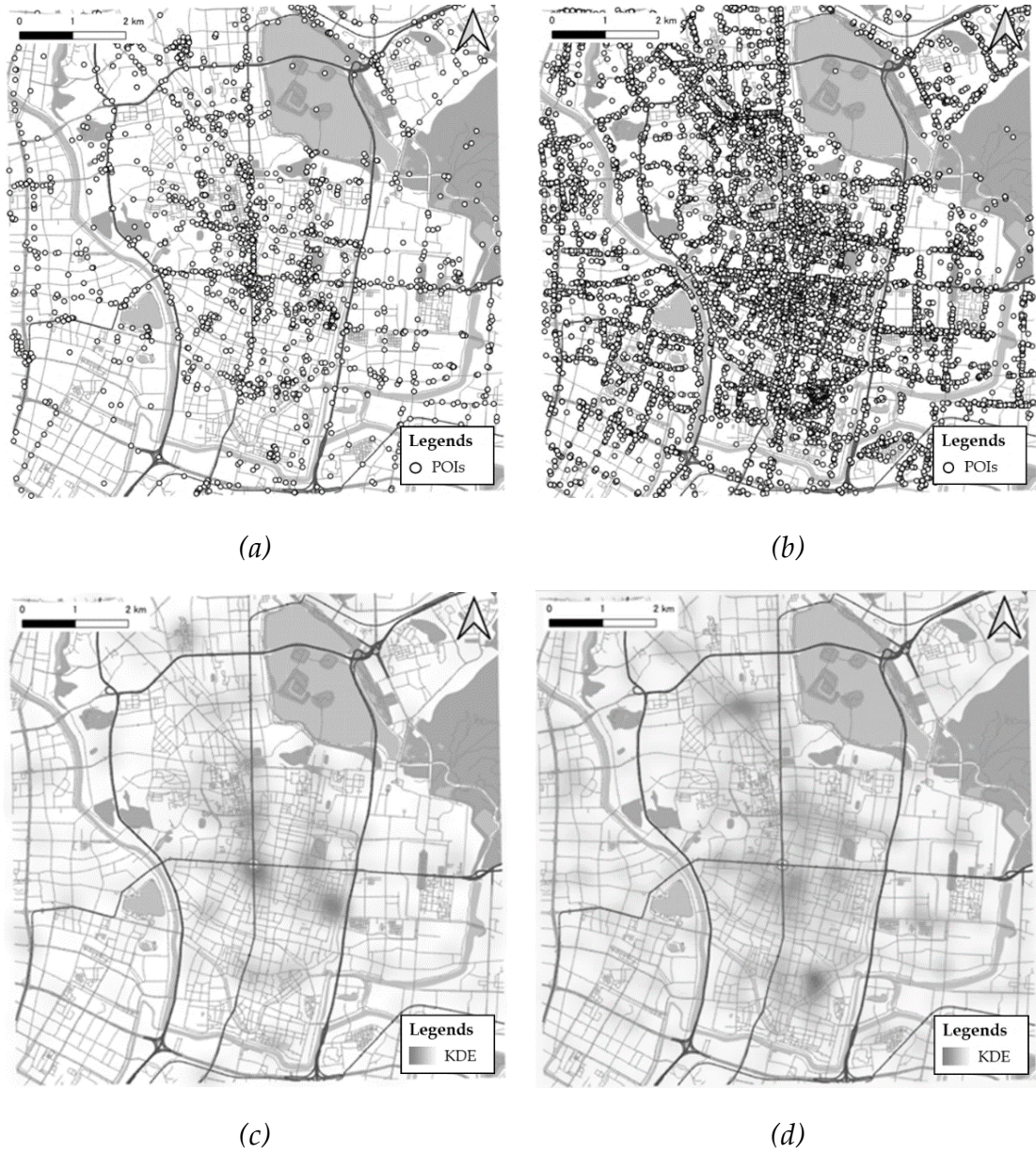


Figure 2-13 Urban hotspots extracted from POI sample datasets

(a) Openstreetmap POIs (a total of 1751); (b) Gaode map POIs ((a total of 10766);

(c) Openstreet map POIs KDE; (d) Gaode map POIs KDE

Furthermore, Figure 2-13 shows the sample POI data points and the results of the quantitative method of estimating the kernel density distribution of urban POIs by using the traditional kernel density method, and then estimating the distribution of urban hotspots. By observing the heatmap map Kernel Density Estimation (KDE) related to OpenStreetMap and Gaode digital map POI (Liu et al. 2017), this study is able to find that the POI density cores mainly distributed near the geometric center of the central area of Nanjing (Xinjiekou CBD), Northwest (Hunan Road Pedestrian Street) and Southeast (Confucius Temple scenic spot). The result is consistent with the recognition result based on KANN-DBSCAN and check-in data shown in 2-11. At the same time, this study also found that the results of Gaode map's POI heatmap with larger POI data samples are closer to the results of the new method proposed in this study. Generally speaking, the result of urban hotspots heatmap recognition based on check-in data is close to that of POI cluster identification in terms of location, pattern and structure. The recognition results are also consistent with the overall master planning of Nanjing. Therefore, the urban hotspot identification method based on KANN-DBSCAN and Sina Weibo check-in data has achieved good accuracy.

2.4 Summary on new urban hotspot identification method

In this study, a new urban hotspot identification method based on SNS data, KANN-DBSCAN algorithm and concave hull algorithm is proposed. Besides on the real-time extraction of urban hotspots, this new method can also obtain more attributes of urban hotspots. Through the above analysis and calculation, the following findings about the features and advantages of the new method are obtained in this study:

Benefiting from the time tags contained in the platform user check-in data provided by the SNS platform, the real-time extraction of urban hotspots by obtaining the check-in data records generated in different periods is realized.

Benefiting from the activity-type label contained in the platform user check-in data provided by SNS platform, the main functions of urban hotspots in urban

space are estimated by calculating the activity composition implemented by SNS platform users in each urban hotspot area.

With the help of parameter self-setting of KANN-DBSCAN algorithm, the automatic and reasonable extraction of the scope of urban hotspots is realized. This method is able to automatically adjust the values of key parameters of MinPts and Eps in clustering algorithm according to the size of dataset, so as to ensure the scientificity of urban hotspot size (because in traditional methods, these parameters need to be determined subjectively according to the experience of experts).

With the help of concave hull algorithm, it is found that the best method to delimit the boundary of urban hotspots should be concave hull algorithm. With the detected boundary, the area and popularity and function of urban hotspots could also be estimated by counting the scale, amount, and activity proportion of involving check-in points.

By comparing the urban hotspot identification results based on traditional data source (digital map POI data) and method (kernel density algorithm) with the experimental results of this study, it is found that the urban hotspot location identified by the new method is consistent with the traditional research and has high accuracy.

References

- 1) Berry, B. J., & Garrison, W. L. (1958). Recent developments of central place theory. *Papers in Regional Science*, 4(1), 107-120.
- 2) Cao, Z., Wang, S., Forestier, G., Puissant, A., & Eick, C. F. (2013, August). Analyzing the composition of cities using spatial clustering. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing* (pp. 1-8).
- 3) Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a dataset. *Journal of statistical software*, 61, 1-36.
- 4) Chen, W., Lai, K. K., & Cai, Y. (2020). Exploring public mood toward commodity markets: a comparative study of user behavior on Sina Weibo and Twitter. *Internet Research*.
- 5) Hu, Y., Miller, H. J., & Li, X. (2014). Detecting and analyzing mobility hotspots using surface networks. *Transactions in GIS*, 18(6), 911-935.
- 6) Ibata, T. (n.d.). Explains two methods of linking EC sites with API. Takashi Ibata. <https://www.ebisumart.com/blog/ec-api/>
- 7) Kabacoff, R. I. (2015). *R in action: data analysis and graphics with R*. Simon and Schuster.
- 8) Khalil, A. (2021, July 30). Hierarchical Clustering Algorithm For Machine Learning. Medium. <https://medium.com/geekculture/hierarchical-clustering-simply-explained-f86b9ed96db7>
- 9) Kim, D., Seo, D., & Kwon, Y. (2021). Novel trends in SNS customers in food and beverage patronage: An empirical study of metropolitan cities in South Korea. *Land Use Policy*, 101, 105214.
- 10) Li, W., Yan, S., Jiang, Y., Zhang, S., & Wang, C. (2019). Research on method of self-adaptive determination of DBSCAN algorithm parameters. *Comput. Eng. Appl*, 55(5), 1-7.

- 11) Li, X. (2020). Recognition of Urban Polycentric Structure Based on Spatial Aggregation Characteristics of POI Elements: A Case of Zhengzhou City. *Beijing Da Xue Xue Bao*, 56(4), 692-702.
- 12) Liu, H., Wang, L., Sherman, D., Gao, Y., & Wu, Q. (2010). An object-based conceptual framework and computational method for representing and analyzing coastal morphological changes. *International Journal of Geographical Information Science*, 24(7), 1015-1041.
- 13) Liu, X., He, J., Yao, Y., Zhang, J., Liang, H., Wang, H., & Hong, Y. (2017). Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science*, 31(8), 1675-1696.
- 14) Malmberg, A., Sölvell, Ö., & Zander, I. (1996). Spatial clustering, local accumulation of knowledge and firm competitiveness. *Geografiska Annaler: Series B, Human Geography*, 78(2), 85-97.
- 15) Oliver, P. N. (2001). Pierre Varignon and the parallelogram theorem. *The Mathematics Teacher*, 94(4), 316-319.
- 16) Porter, M. E. (2011). *Competitive advantage of nations: creating and sustaining superior performance*. simon and schuster.
- 17) Sadahiro, Y. (2001). Analysis of surface changes using primitive events. *International Journal of Geographical Information Science*, 15(6), 523-538.
- 18) Sadahiro, Y. (2003). Stability of the surface generated from distributed points of uncertain location. *International Journal of Geographical Information Science*, 17(2), 139-156.
- 19) Shi, B., Zhao, J., & Chen, P. J. (2017). Exploring urban tourism crowding in Shanghai via crowdsourcing geospatial data. *Current Issues in Tourism*, 20(11), 1186-1209.
- 20) Wang, J. L., Jackson, L. A., Gaskin, J., & Wang, H. Z. (2014). The effects of Social Networking Site (SNS) use on college students' friendship and well-being. *Computers in Human Behavior*, 37, 229-236.

- 21) Yang, W. (1994). The retailing and services center and network of Beijing: Then, now and long before. *Acta Geographica Sinica*, 49(1), 9-17.

Chapter 3

Spatial-temporal Distribution of Urban Hotspot

Chapter 3

Spatial-temporal Distribution of Urban Hotspot

3.1 Necessity of clarifying urban hotspots dynamically

In this chapter, this study will try to clarify the temporal and spatial patterns of the emergence and change of urban hotspots from the perspective of dynamics. Therefore, the necessity of clarifying urban hotspots dynamically will be explained first.

(1) Variability of urban hotspots

One the one hand, as mentioned above, urban hotspots in this study are defined as the gathering areas of similar built-up environmental elements and people flow in the city (Newling 1969). The built environmental elements in the city will continue to advance with the development progress of the city and the planning and design scheme (Limtanakool et al. 2009). Therefore, urban hotspots will show different characteristics in different periods and dates with the changes of urban built environment and human flow. This is evidenced in a recent work by Li et al. (2018).

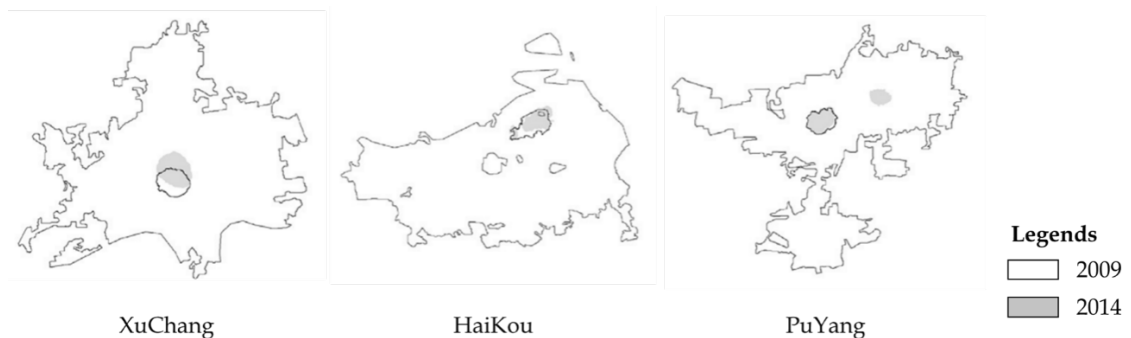


Figure 3-1 Empirical cases of variability of urban hotspots

Source: Li et al. (2018)

Li et al. (2018) sorted out the location of LWP (Live-Work-Play) centers in 35 Chinese cities in 2009 and 2014 and found that the location of urban centers will change over time, which greatly support viewpoint that urban hotspots is of variability (Figure 3-1).

(2) Impact of urban hotspots on urban system

On the other hand, the constant change of urban hotspots will also lead to the variability of the variation of their great impacts on the urban system (Jiang et al. 2012). In detail, the urban hotspots had been demonstrated to be important for our daily life and even the urban environment (McDonald 1987). Different types of urban hotspots will have a significant impact on different aspects of the urban sub-systems.

Table 3-1 Evidence on the impact of urban hotspots on urban system

Representative	Type of hotspots	Urban sub-system	Link
Bilkova (2016)	Commercial center	Real estate market	Rise of regional land prices
Qin (2019)	Catering center	Population pattern	travel patterns of residents
Jia (2014)	All types	Eco-system	CO ₂ emission and air pollution

For instance, as summarized in Table 3-1, Bilkova (2016) et al. found that the aggregation of urban commercial and production factors is the main reason for the rise of regional land prices. The research by Qin and Zhen (2019) also showed that the popularity of typical urban hotspots such as catering spaces in urban area will greatly affect the travel patterns of urban residents. What’s more, Jia’s work (2014) even shown that the hierarchies of urban hotspots could affect the CO₂ emission and air pollution level.

In general, summarizing the temporal and spatial distribution characteristics of urban hotspots is not only helpful to understand the dynamic change pattern of urban built environmental elements and people flow, but also can be used to predict the development trend of urban economy, transportation, and environmental sub-systems.

3.2 Urban hotspots at different time periods and dates

Starting from this chapter, this study will continue to explore the temporal and spatial distribution characteristics of urban hotspots with the help of a case from the central area of Nanjing, Jiangsu Province, China, which is between latitudes 32.009° and 32.090° N and longitudes 118.721° and 118.826° S.

3.2.1 Division of time period and date

Since the main propose of this study is to investigate the dynamic changes of urban hotspots, the first thing to be determined is to determine which time periods and dates the research will analyze the dynamic changes of urban hotspots. The results of this study are hoped to be universal and representative and can reflect the temporal and spatial distribution characteristics of urban hotspots in all regions to a certain extent. Therefore, this study will specifically answer the following two questions:

- *How do urban hotspots change in a day (24 hours)?*
- *What is difference between the urban hotspots on off days and weekdays?*

To do that, a total of 28316 check-in points from Sina Weibo platform was obtained from April 27 to May 3, 2014, in the core area of Nanjing. Furthermore, this study extracts the check-in activities in a specific time period by calculating the difference in the number of check-ins generated at different time points (Chapter 2.2). This study sets the time points of data collection as 6, 12, 18 and 24 o'clock and calculate the check-in data generated in the 3 time periods of morning (06:00-12:00), afternoon (12:00-18:00) and night (18:00-24:00) from April 27 to May 3, 2014. Thus, the check-in dataset is divided into two types of weekdays (28, 29, 30 April 2014) and off days (27 April and 1, 2, 3 May 2014), and the average value of the check-in quantity generated at each check-in point in 3 time periods is calculated. So far, this study has reorganized the raw data into 6 subsets that reflect the spatial distribution and intensity of check-in activities from morning to night on weekdays and off days, respectively. The segmentation results of SNS dataset are shown in Figure 3-2.

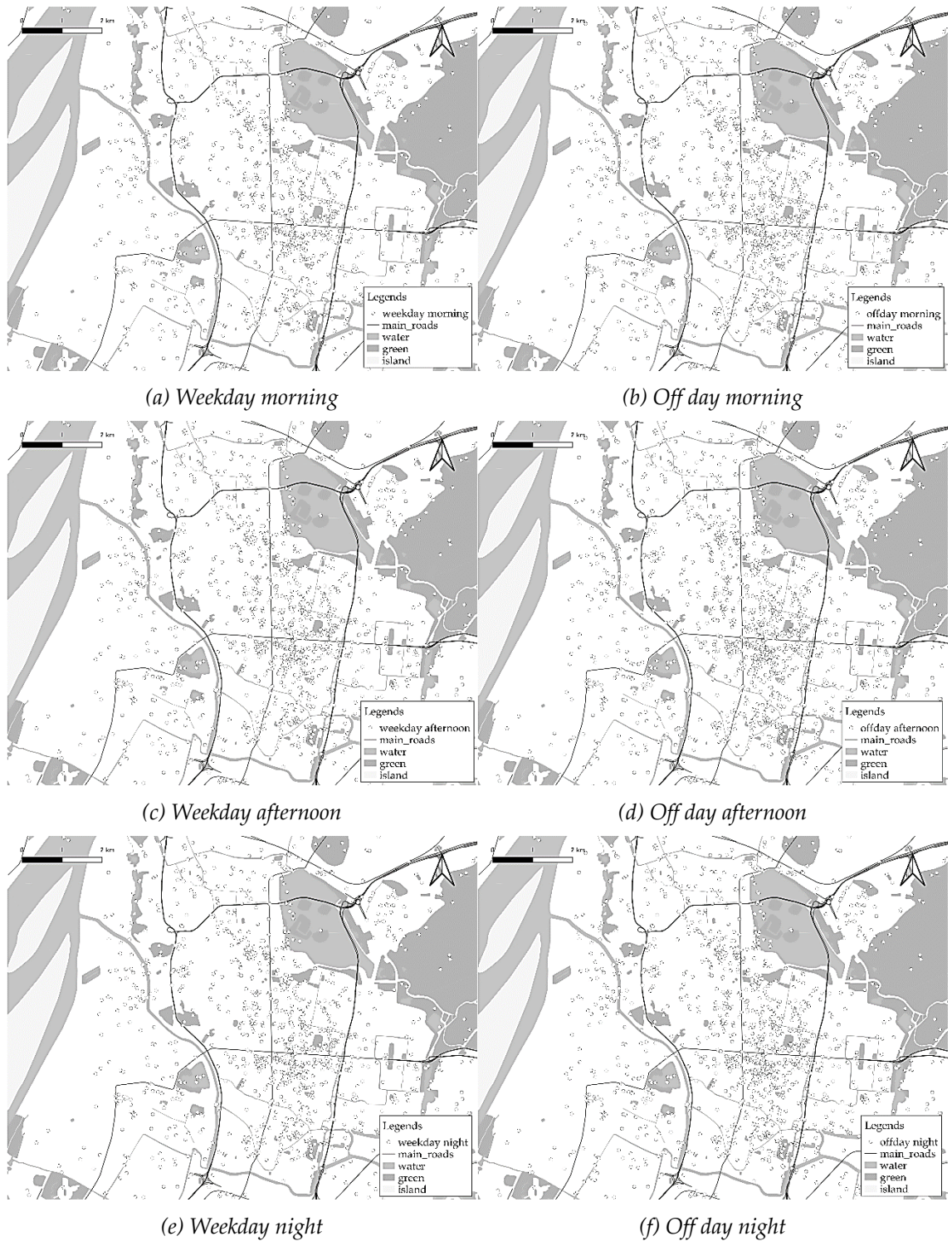


Figure 3-2 Spatial distribution of check-in points at different periods

Note: Each point may generate more than one check-in activity

In all six periods, the number of check-in points generated in the afternoon of the rest day in the study area is the largest, reaching 6299. In contrast, the number of check-in points in the morning of the working day is the lowest, only 3465. In comparison, there are two main findings. On the one hand, the number of check-in points of off days in the study area is much higher than that in working days. Each corresponding period is 43% higher on average (about 1683 times). On the other hand, it is common that the number of check-in points in the afternoon is the highest, which is much higher than that in the morning and evening. The general pattern is presented in Figure 3-3.

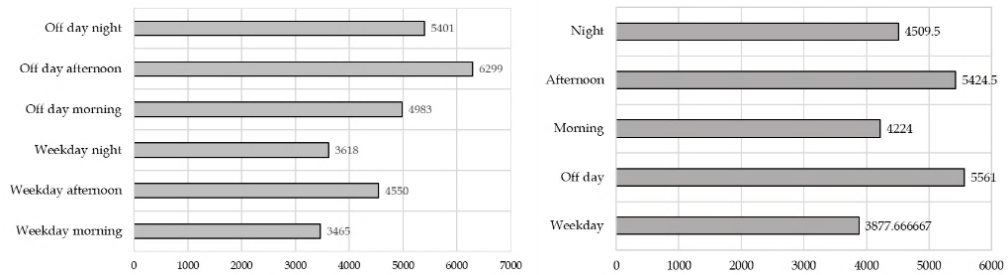


Figure 3-3 Number of check-in points at different periods

In addition, this study also investigated the proportion of check-in points including check-in activity types in 6 periods. The comparison results show that the proportion of the seven main check-in activities is basically similar in each period, and only in the proportion of leisure and entertainment activities, it shows the characteristics that working days are more than rest days.

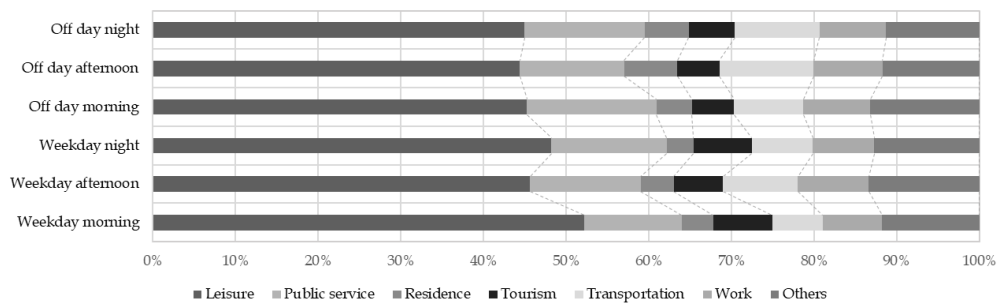


Figure 3-4 Proportions of check-in points at different periods

3.2.2 Identification results on different time



(a) Weekday morning (Eps = 0.00021, MinPts=13) (b) Off day morning (Eps = 0.00020, MinPts=14)



(c) Weekday afternoon (Eps = 0.00020, MinPts=15) (d) Off day afternoon (Eps = 0.00020, MinPts=15)



(e) Weekday night (Eps = 0.00021, MinPts=14) (f) Off day night (Eps = 0.00021, MinPts=14)

Figure 3-5 Spatial distribution of urban hotspots in downtown of Nanjing

Note: The urban hotspots with different patterns represent different functions

Using KANN- DBSCAN, the check-in points were classified into 216 sub-groups, and the input parameters were shown in Figure 3-5. The concave hulls of 216 sub-groups of check-in points denoted urban hotspots identified. And the area of each concave hull and the quantity of check-ins within the boundary of each concave hull are defined as the area and popularity of urban hotspots. The 216 urban hotspots identified in this study can be divided into three categories of tourism, work and leisure related types according to the proportion of human activities involved.

With the help of QGIS, the spatial distribution of urban hotspots extracted from 6 time periods was visualized (Figure 3-5). It can be seen that large-scale urban hotspots occurred more frequently near the center of the study area (where the two main roads intersect on the diagram) in the afternoon. The relative density of urban hotspots on off days (weekends and holidays) was greater than that on weekdays. In addition, it was also illustrated in the figure that the morphology of urban hotspots changed by time at the same space, which seemed to be accompanied by changes of functional orientations. Therefore, this study will analyze this phenomenon in detail in the following chapters.

3.3 Emerging and changing of urban hotspots

In this chapter, this study will compare the location and attribute changes of urban hotspots across different periods and dates and continue to try to summarize the emergence and change law of urban hotspots with a case study in the main urban area of Nanjing, China.

3.3.1 Basic types of change of urban hotspots

Firstly, this study will define the standard for observing the dynamic change of urban hotspots in this study - morphological change. In terms of morphology, the changes of urban hotspots mainly contained five basic types of maintain, merge, split, emerge and disappear (Figure 3-6).

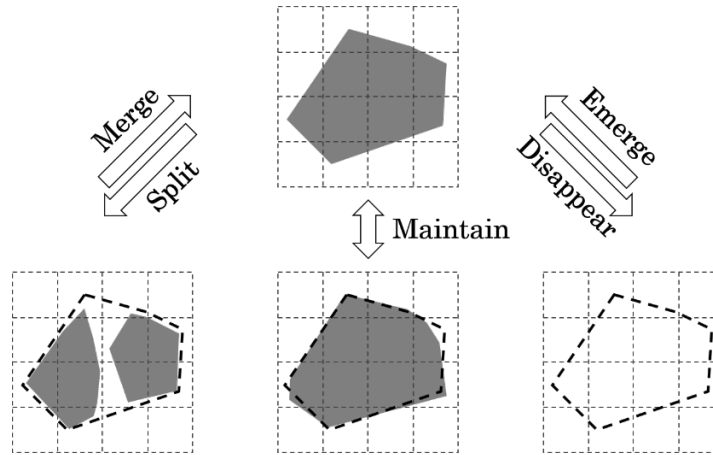


Figure 3-6 Basic changes of urban hotspots

Specifically, if the outlines of an urban hotspots have a complete coverage relationship before and after the change, or their outlines intersect, this type of change is defined as maintain; If there emerge more than 2 new urban hotspots within the original outline of an urban hotspot undergoing a period of time, this type of change is defined as split; otherwise, it is defined as merge; If an urban hotspot appears in an area where no urban hotspot originally exists, this change is defined as emerge; If an urban hotspot vanishes in its surrounding areas after a period of time, this change is defined as disappear.

3.3.2 Changes of urban hotspots across time period and date

In this chapter, this study further explored the rules of the emerging and changing of urban hotspots according to the order and types of the occurrences of five basic changes of urban hotspots across different time periods, and the spatial distribution of four types of changes except maintain was visualized (Figure 3-7 and Figure 3-8).

Specifically, in the following figures, this study presented the emergence and change of urban hotspots in five situations (weekday to off day; weekday morning to afternoon; weekday afternoon to night; off day morning to afternoon; off day afternoon to night). Specifically, each transparent polygon with black solid outline represented the original state of an urban hotspot, and each grey filled polygon represented its state after change. The text box in the upper left

corner of each figure showed the time point of change starting and ending. Furthermore, the areas covered by solid rectangular boxes, dotted solid boxes, solid circles and dotted circles represented the urban spaces where the basic changes of merge, split, emerge and disappear happened, respectively. In addition, the remaining urban hotspots that are not surrounded by any graphics experience the basic changes of 'Maintain'.

(1) Weekday ~ Off day

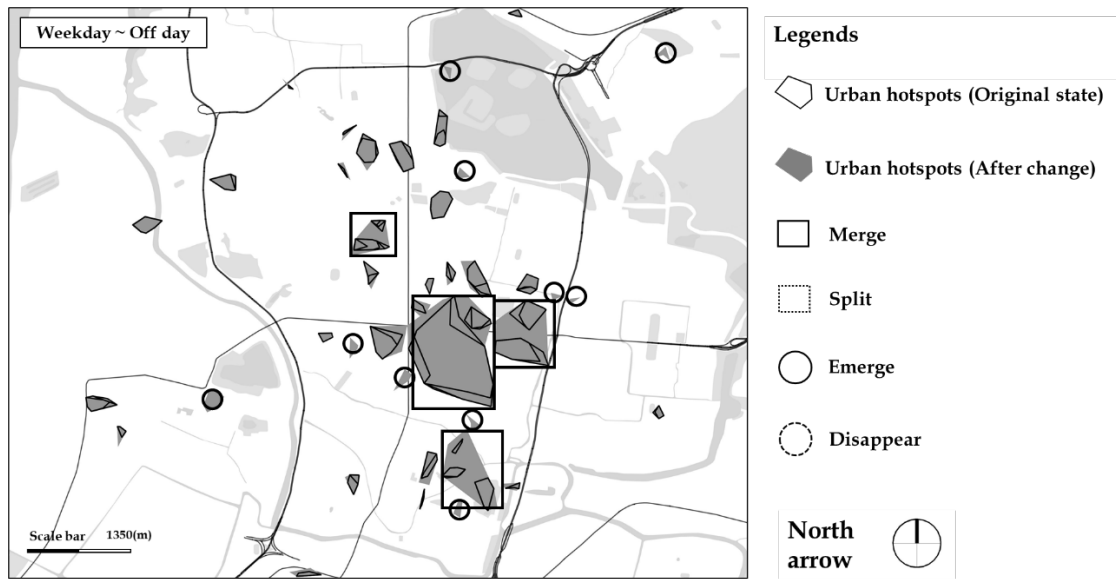


Figure 3-7 Spatial distribution of the basic changes of urban hotspots

Weekday ~ Off day

Other parts of urban hotspots change their shapes with across different time periods or only appear at specific time points. Specifically, the Figure 3-7 presented the emergence of urban hotspots on weekdays and off days. From the figure, it is found that: (1) The urban hotspots contained in the solid circles of Figure 3-7 only appears on the off days and disappear on the weekdays. (2) The urban hotspots contained in the solid boxes of Figure 3-7 merge with other hotspots around them on off days.

(2) Morning & Afternoon & Night

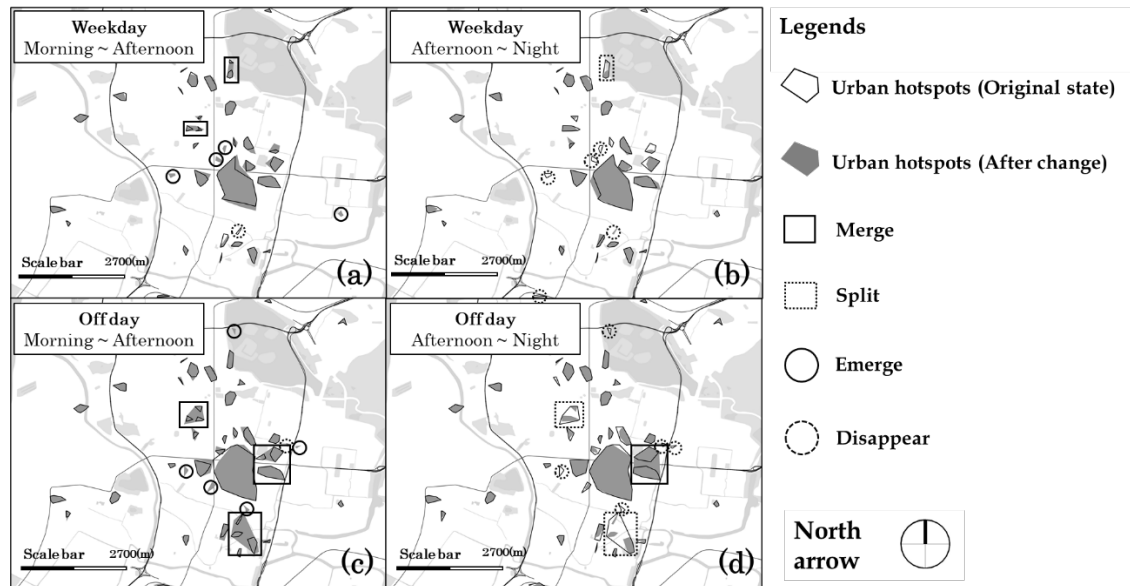


Figure 3-8 Spatial distribution of the basic changes of urban hotspots

Morning & Afternoon & Night

Moreover, the spatial distribution of the basic changes of merge, split, emerge and disappear across different periods of time illustrated in the Figure 3-8 indicated that: (1) The urban hotspots contained by both the solid boxes of Figure 3-8 (a, b) and dotted boxes of Figure 3-8 (c, d) merge with surrounding hotspots from morning to afternoon, and split back to their original states at night. This phenomenon appears on both weekdays and off days. (2) The urban hotspots contained by both the solid boxes of Figure 3-8 (a, b) and Figure 3-8 (c, d) merge for 2 times from off day morning to night. (3) The urban hotspots contained by the solid or dotted circles of Figure 3-8 (a, b, c, d) vanished and appeared intermittently on both weekdays and off days. The global spatial distribution of basic changes of urban hotspots illustrated that a part of urban hotspots didn't experience the basic changes of merge, split, emerge or disappear but maintain their original form across different time periods.

3.3.3 Rules of the emerging and changing of urban hotspots

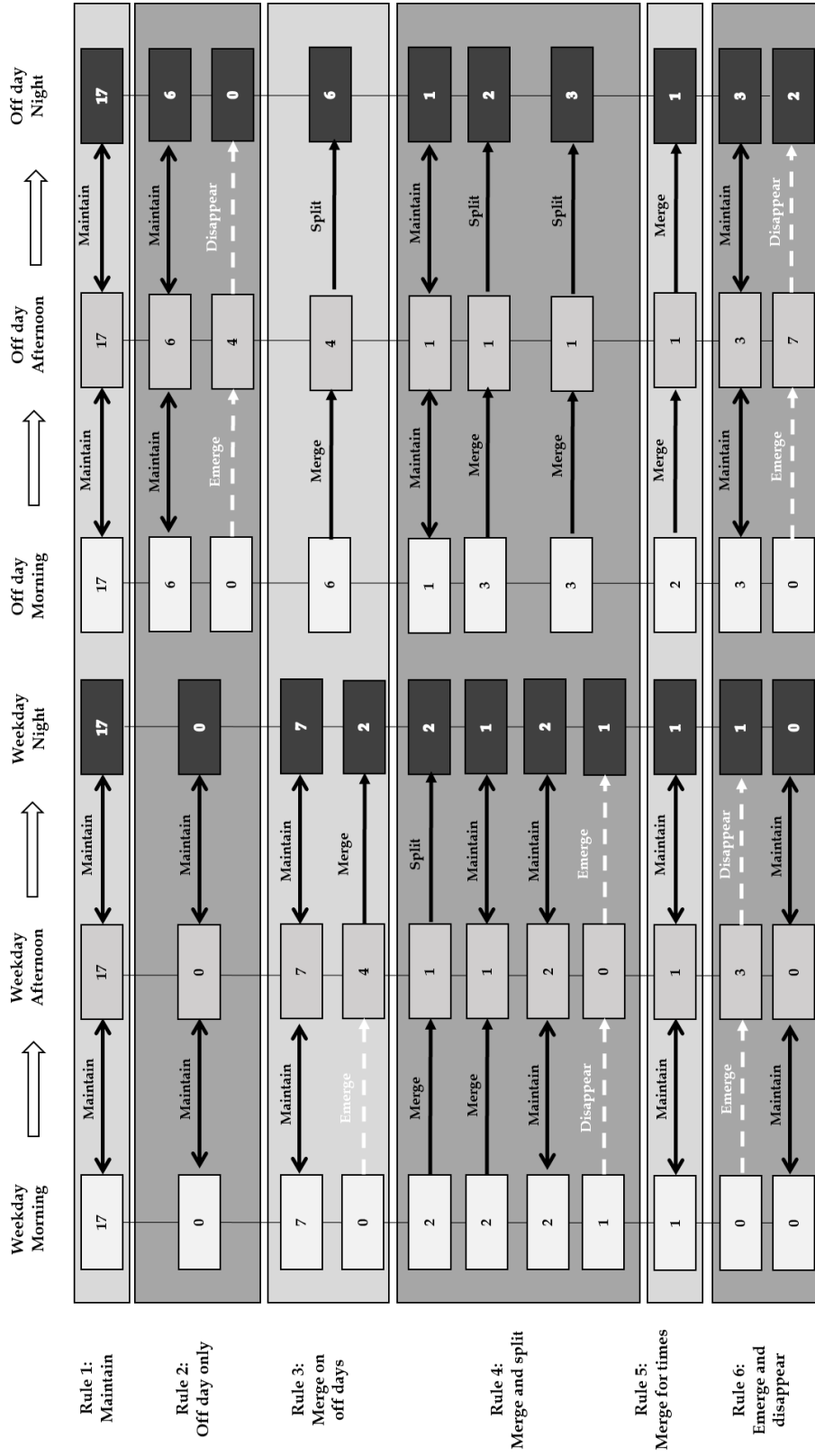


Figure 3-9 Six main rules for the emergence and change of urban hotspots

On the whole, urban hotspots were divided into 6 categories according to the order and types of 5 basic changes experienced by them, and the changes experienced by each category represents a common rule. Therefore, the emergence and change of urban hotspots could be summarized into 6 rules of ‘Maintain’, ‘Off day only’, ‘Merge on off days’, ‘Merge and split’, ‘Merge for times’ and ‘Emerge and disappear’. It should be noted that the variation process of a particular urban hotspot may follow more than one rule. Specifically, the variation processes of urban hotspots were visualized in Figure 3-9. The rows and columns where boxes were located represented the emerging and changing rules they experienced and the time periods they were in, respectively. In addition, the numbers in the boxes represent the quantity of urban hotspots, and the arrows connecting two boxes and the words above represented the types of basic changes.

3.3.4 Attribute variations under the rules of urban hotspots

In the previous chapter, this study has summarized six main emerging and changing rules of urban hotspots according to their morphological change. Here, this study focused on the attribute variations following each rule. Table 3-2 presents a descriptive statistic on the attributes of urban hotspots under six rules.

Table 3-2 Descriptive statistic on the attributes of urban hotspots

Rule	Average		Median		Std.Dev.	
	Popularity	Area	Popularity	Area	Popularity	Area
1	1146.7	37764.3	1140.7	37638.2	202.3	4223.4
2	115.4	12913.3	90.6	9267.7	57.0	8365.4
3	2241.1	186292.6	958.4	42924.6	1911.7	200369.3
4	950.0	33476.1	265.7	26716.5	871.3	24177.8
5	1159.4	59208.4	831.9	42263.9	1191.0	57736.8
6	92.5	8339.7	91.9	8280.4	41.5	4105.3
Total	1118.9	60264.3	1008.1	36129.9	1128.5	107014.9

The results showed that: (1) The standard deviations (Std.Dev.) of the area and popularity of urban hotspots in rule 1 are only 4223.4 m² and 202.3 respectively, which are far lower than the total level of 107914.9 m² and 1128.5. Among the 6 rules, they were only in the 5th and 4th places. This further indicated that the attributes of urban hotspots in rule 1 were located in a relatively small range around the average value. Compared with the hotspots under other rules, no matter what state they were in, their attributes variation little. (2) Urban hotspots under other rules, especially those urban hotspots that had undergone split and merge types of basic changes, had more discrete data distribution characteristics (relatively large standard deviation). Therefore, this study further analyzed the area, popularity and function of urban hotspots at various stages of their emergences and changes (Figure 3-10).

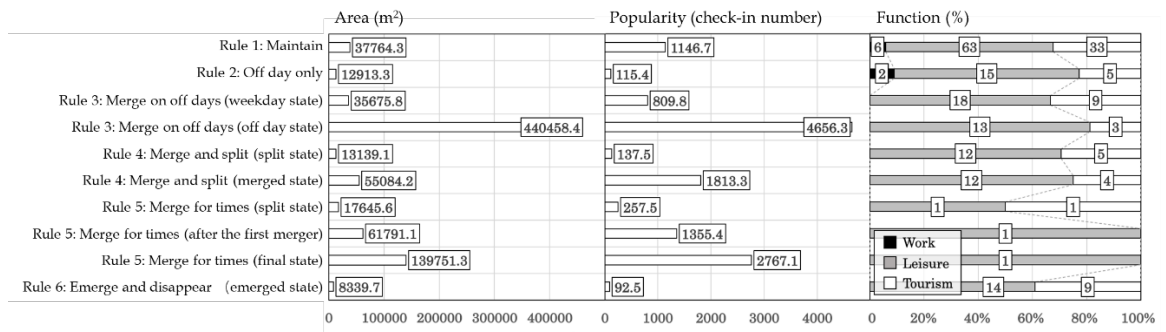


Figure 3-10 Attributes variations under the emerging rules of urban hotspots

The two bar plots on left presented the attributes of the area and popularity of urban hotspots at different states (the numbers beside the bar represent the average values area and popularity) and the stacked histogram on right showed the functional structures (the lengths of the columns represent the proportion of the number of urban hotspots with three functions; the numbers on the columns represent the specific quantity of hotspots with corresponding function).

(1) Area & Popularity

To mention first, since the variation trend of area of urban hotspots was similar to their popularity, so they were discussed together. The Figure 3-10 illustrated that: (1) The area and popularity of urban hotspots in rule 1 and rule 3 (weekday

state) were near the overall median. Specifically, the median area of all urban hotspots was 36129.9 m², while the average area of corresponding urban hotspots included in the two rules were 37764.3 m² and 35675.8 m² respectively. Similarly, the median popularity of all urban hotspots was 1008.1, while the average popularity of corresponding urban hotspots in the two rules were 1146.7 and 809.8, respectively. (2) The area and popularity of urban hotspots in rule 2, 6 (emerge state) and rule 4,5 (split state) were extremely low. Specifically, the average values of their areas were only 12913.3, 8339.7, 13139.1 and 17645.6 m², respectively, far below the median value (36380.8 m²). Similarly, their popularity values were 115.4, 92.5, 137.5 and 257.5, which were far from the median (1016.9). (3) When the urban hotspots in rule 4 and 5 had completed a merger with the surrounding ones, their area and popularity be doubled several times. Specifically, their area increased to 55084.2 and 61791.1 m², respectively, and their popularity surged to 1813.3 and 1355.4. These values were distributed in the range of 1-2 times the median. (4) When the urban hotspots in rule 3 and 5 reached the final state of their merging (i.e., the state of off day or after multiple merging), their area and popularity reached the maximum value in the study period. Specifically, their areas were 440458.4 and 139751.3 m², and the popularity reached 4656.3 and 2767.1, respectively, ranking the first and second place in the study period.

(2) Function

In addition to the area and popularity, the function of urban hotspots under different rules also changed across different time periods. The stacked histogram on right showed the functional structure of urban hotspots in different states following 6 rules. Specifically, there is no difference in the functional structure of urban hotspots covered by rule 1, 2, 6 (which experience the basic changes of 'Maintain', 'Emerge' and 'Disappear') in different stages of their changes. However, when the urban hotspots included in rules 3, 4, 5 changed from split state to merged state, the number of tourism-related hotspots decreased, while the number of leisure-related hotspots increased, which indicating that the functional orientation of tourism-related hotspots turned into leisure-related ones after merger.

Generally, the 6 rules of the emerging and changing of urban hotspots and the attribute variations accompanied by these rules can be summarized as follows:

(1) A part of urban hotspots with small area and low popularity only appears in a certain period of time. They either only emerge on off days (Rule 2) or vanishing and appearing intermittently in a day (Rule 6). Other part of them merges surrounding ones in the afternoon, but split back to their original state at night (Rule 4).

(2) Most urban hotspots with median area and popularity didn't experience morphological change (Rule 1), and the remaining small parts are either divided into their spilt states (Rule 4) or merged into large-area and high-popularity ones (Rule 5) accompany with function change from tourism-related to leisure-related.

(3) It is worth noting that if the merger took place in the process from weekdays to off days, it formed urban hotspots with largest area and highest popularity in the study period (Rule 3). Similarly, this process accompanies with function change from tourism-related to leisure-related, as well.

3.4 Relationship between the rules and built environment

In the previous chapters, this study has summarized the emerging and changing of urban hotspots into six rules and discussed the attribute variations following these rules. Here, this study focuses on the relationship between the emerging and changing rules of urban hotspots and urban built environment. Previous studies have demonstrated that urban hotspots appear in the areas where urban elements are concentrated and close to the urban center (Li et al. 2018).

This study plan to further analyze whether there is a certain link between the urban built environment and the variations of urban hotspots in this chapter. Firstly, this study visualized the spatial distribution of 6 emerging and changing rules of urban hotspots in Figure 3-11. This study divided the study area into standard square grids, each of which is 0.0075 unit of longitude and latitude wide.

The selection of scale is the result of considering the complexity of calculation and the accuracy of data.

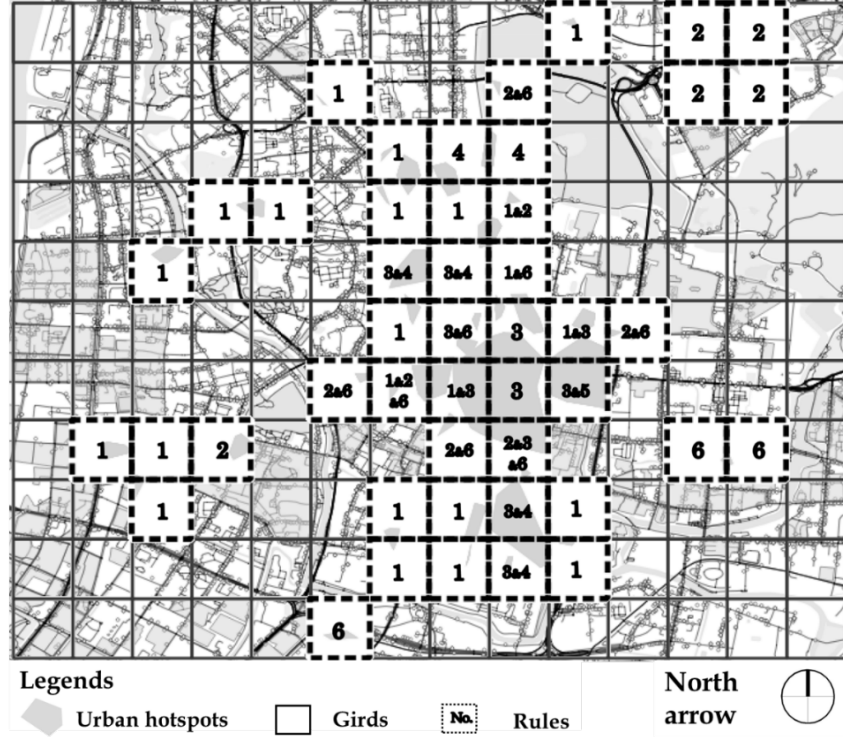


Figure 3-11 Spatial distribution of the 6 rules

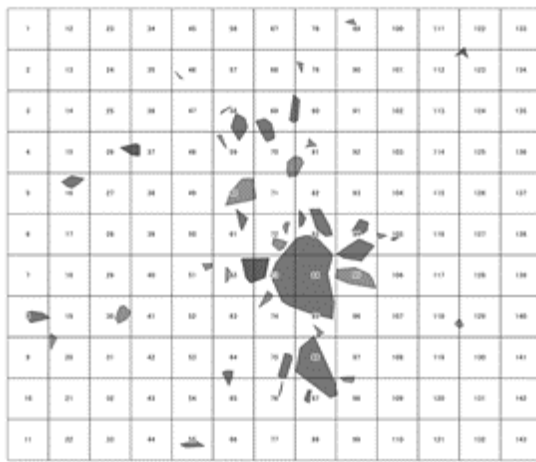
Note: The numbers on the figure represent the number of the emergence and change law of urban hotspots

This study also calculated the number of transport infrastructure (amount of bus stations, metro stations, taxicab stations per km²), road density (total length of roads per km², km/km²), building density (coverage per centage of building footprints, %) and the distance to the urban center (the distance from the geometric center of the grid to the city government, km) of each grid and set them as elements of the urban built environment factors.

3.4.1 Model construction

In order to further explore the quantitative relationship between the probability of the occurrence of six types of rules and the built environment, this

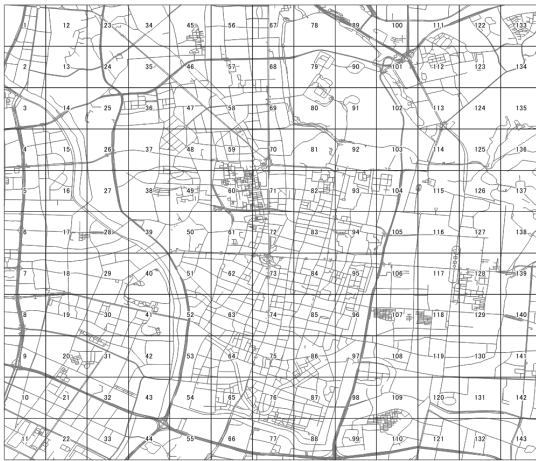
study constructed binary logistic regression models for six types of emerging and changing rules. Specifically, the observations used to quantify the regression relationship were the 0.0075 degrees grids, and the dependent and independent variables were the occurrence states of 6 emerging and changing rules of urban hotspots and 4 urban built environment factors of number of transport infrastructure (NT), road density (RD), building density (BD) and distance to the urban center (DC), respectively.



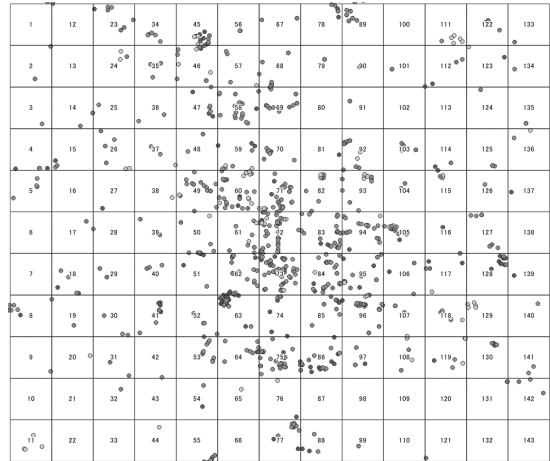
(a) Urban hotspots



(b) Building footprints



(c) Road network



(d) Transport infrastructure

Figure 3-12 Data related to urban built environment

Data source: OpenstreetMap, (<https://www.openstreetmap.org>)

In addition, the identification results of this study have indicated that urban hotspots have obvious characteristics of concentrated distribution. Therefore, to eliminate the influence of adjacent urban hotspots on the accuracy of the regression results, this study also added the number of hotspots per km² in adjacent grids (NH) as an instrumental variable.

The data related to the urban built environment used in this research is collected from OpenStreetMap (<https://www.openstreetmap.org>) platform. OpenStreetMap is a web-based mapping platform. It has been the most popular web map source for use in location-based services with specific emphasis on pedestrian navigation, tourist guide applications, and other location-based academic works (El-Ashmawy 2016). Specifically, the main data include the vector points of public infrastructure, vector lines of the road network and vector polygons of building footprints. In order to facilitate the quantitative analysis of the relationship between urban built environment and urban hotspots, this study divided the study area into standard square grids, each of which is 0.0075 unit of longitude and latitude wide. The selection of scale is the result of considering the complexity of calculation and the accuracy of data. Finally, this study calculated the 4 urban built environment factors of the number of transport infrastructure (amount of bus stations, metro stations, taxicab stations per km²), building density (coverage per centage of building footprints, %), road density (total length of roads per km², km/km²) and the distance to the city center (the distance from the geometric center of each grid to the city government, km) of all grids, which had been proved to be the main factors influencing the formation of urban hotspots in previous literatures (McMillen and Smith 2003).

After collecting the occurrence and change rules of urban hotspots (occurrence or non-occurrence) and built environment data of each grid, this study plans to introduce binary logistic regression model to explore the relationship between them. The formula of this model is presented as follows:

$$p_{ij}(y_{ij} = 1|X_i) = \frac{1}{1 + e^{-(\alpha + X_i)}} \quad 3-1$$

Where, i represents the serial number of the grid; j represents the serial number of the occurrence and change rule of urban hotspots; p_{ij} denotes the probability of the occurrence of urban hotspot rule j at the grid with serial number i ; y_{ij}

denotes the actual situation of the occurrence of urban hotspot rule j at the grid with serial number i ; X_i denotes the built environment factors at grid i ; α denotes the residual item.

Next, due to the basic requirements of binary logistic regression model, this study converts the values of the corresponding variables at each grid into binary codes. Among them, if a particular type of rule occurs, the dependent variable value of the corresponding rule was encoded as 1, otherwise it was 0; if there were urban hotspots in adjacent grids, the independent variable value of NH was encoded as 1, otherwise it was 0; if there were transportation infrastructure in grid, the independent variable value of NT was encoded as 1, otherwise it was 0; if the road density, building density exceeded the average value, the independent variable values of RD and BD were encoded as 1, respectively, otherwise they were 0; if the distance from the city center below average, the independent variable values DC was encoded as 1, otherwise it was 0.

3.4.2 Output of binary logistic regression models

Inputting the re-coded variables into the models, the output results of the binary logistic regression models regarding six emerging and changing rules are obtained (Table 3-3).

Table 3-3 Result of logistic regression models

	NH	NT	RD	BD	DC
Rule1	2.31*	1.34	3.97	1.58**	0.81
Rule2	3.10	1.9	6.82	3.6	3.45*
Rule3	1.86**	6.4	2.61	2.98	1.05
Rule4	2.04**	2.16	3.05	3.14**	5.23
Rule5	0.75	1.36	2.45	2.62	3.44
Rule6	0.73*	1.04	2.09	1.93	0.42*

*Note: * represents that $p < 0.05$; ** represents that $p < 0.01$; NH, NT, RD, BD, DC represent number of hotspots in adjacent grids, number of transport infrastructure, road density, building density and the distance to the urban center respectively.*

In the above table, each number represents the odd ratio (OR) of the occurrence of urban hotspots emerging and changing rules before (coded as 1) and after (coded as 0) the value change of independent variables. Taking the relationship between NH variables and rule 1 as an example, the number 2.31 in the table indicate that the occurrence rate of rule 1 in grids with urban hotspots in adjacent grids (NH = 1) was 2.31 times of those isolated ones (NH = 0). In addition, the 6 rows in Table3 represented the models regarding the relationship between 6 rules and urban built environment factors, respectively.

The outputs of binary logistic regression models indicated that:

(1) In the five selected variables of urban built environment, road density (RD) and number of transport infrastructure (NT) had no significant impact on the emerging and changing of urban hotspots. The difference of the occurrence rate of urban hotspots emerging and changing rules between grids with different number of urban hotspots in adjacent grids (NH), building density (BD) and the distance to the city center (DC) showed a statistically significant.

(2) The occurrence rate of different rules was affected by the different built environmental factors, and the core influencing factor and mechanism was slightly different.

(3) Compared with the grids with no urban hotspots around, the grid with urban hotspots in adjacent areas was more likely to undertake the rule of 1, 3 and 4. Specifically, occurrence rates of rule 1 ,3 and 4 in grids with urban hotspots in adjacent grids (NH = 1) were 2.31, 1.86 and 2.04 times higher than those isolated ones (NH = 0), respectively.

(4) Compared with the grids with low building density (BD = 0), the grids with the building density higher than the average (BD = 1) had a higher probability of occurrence of rule 1 and 4, with OR values reach 1.58 and 3.14, respectively. That is to say, if an area has the characteristics of high building density, the possibility of the hotspots in the space maintained a stable state and merged and split in one day were 1.58 and 3.14 times of the area with low building density, respectively.

(5) Compared with the grids whose distance to the city center was greater than the mean (DC = 0), the grid whose distance to the city center was less than the

mean ($DC = 1$) has a higher probability of occurrence of rule 2, with OR value reached 3.45. That was to say, if an area was relatively close to the city center, the possibility of urban hotspots that existed in the off days was 3.45 times of that far away from the city center.

(6) The relationship between the occurrence probability of the rule 6 and the urban built environment factors is quite special. Specifically, the probability of frequent occurrence or disappearance of urban hotspots in the surrounding space was about 27% lower than that in the surrounding space without urban hotspots. And the probability of this phenomenon in the area close to the city center is only 42% of that in the far area, while the influence of these two urban factors on the incidences of other five types of rules is positive.

3.4.3 Discussions on the urban hotspot rules and built environment

Overall, the general impact of urban built environment on the emerging and changing rules of urban hotspots could be summarized as that in the urban space with more urban hotspots and transport infrastructure, higher building and road density and closer distance to the city center, the probability that urban hotspots vary following the rule of 1-5 is higher, and that of rule 6 is relatively lower. Among them, the relationship between the rules and transport infrastructure and road density didn't show a statistically significant.

Here, this study found two points that were inconsistent with conclusions of previous studies. First, the factors related to urban transportation infrastructure and road density do not show significant impact on the emerging and changing of urban hotspots, while previous research generally pointed out that urban hotspots tend to appear in areas with high accessibility (Melo et al. 2017). Second, the distance of some urban hotspots from the city center is directly proportional to the occurrence rate of their emerging and changing rules. In contrast, related studies had proposed that urban hotspots tend to appear in areas close to the city center (Leslie 2010). With the advent of the information era and the improvement of the level of transportation infrastructure construction, whether a specific urban space has high attraction, carries high-intensity human activities and evolves into an urban hotspot is not entirely determined by its accessibility and location. The attractive elements in the surrounding area of urban spaces are

likely to be another major reason for the generation of urban hotspots. This is probably the reasonable explanations for the contradictions.

A recent work of Krehl and Siedentop (2019) helps us reinforced this belief. Krehl and Siedentop (2019) studied the urban centers of Frankfurt, Cologne, Munich and Stuttgart from the macro and micro dimensions and analyzed the relationship between the spatial structure of the urban main centers and the sub centers and the built environment, accessibility and land use types in different scales. Their research results showed that the accessibility factor is closely related to the large-scale urban main centers and is demonstrated to be one of the main driving factors of urban economic development. However, from the micro perspective, the relationship between the location of urban sub centers and accessibility decline, and the main factors affecting them are the regional characteristics and economy.

3.5 Summary on the spatial-temporal distribution of urban hotspot

With the help of SNS data, KANN-DBSCAN and concave hull algorithms, this study carried out a detailed study on the spatial-temporal distribution of urban hotspots and conduct a case study in downtown area of Nanjing, China using Sina Weibo check-in data.

Generally, the spatial-temporal distribution of urban hotspots could be summarized into six emerging and changing rules, and the attribute variations accompanied by these rules can be concluded as follows: (1) A part of urban hotspots with small area and low popularity only appear in a certain period of time. They either only emerge on off days (Rule 2) or vanishing and appearing intermittently in a day (Rule 6). Other part of them merges surrounding ones in the afternoon, but split back to their original state at night (Rule 4). (2) Most urban hotspots with median area and popularity didn't experience morphological change (Rule 1), and the remaining small parts are either divided into their spilt states (Rule 4) or merged into large-area and high-popularity ones (Rule 5)

accompany with function change from tourism-related to leisure-related. (3) It is worth noting that if the merger took place in the process from weekdays to off days, it formed urban hotspots with largest area and highest popularity in the study period (Rule 3). Similarly, this process accompanies with function change from tourism-related to leisure-related, as well.

In addition, through the construction of binary logistic regression models, the quantitative relationship between the six rules and the urban built environment was combed out. This study finds that whether there are urban hotspots in the surrounding areas, the building density and the distance to the city center can have a significant impact on the emergence and change of urban hotspots. Specifically, in the urban space with more urban hotspots, higher building density and closer distance to the city center, the probability of urban hotspots appearing according to the rule of 1-5 is higher, and the probability of appearing according to the rule 6 is relatively lower.

Overall, there are two main academic contributions of this study. One of the main characteristics that distinguishes this study is the additional attention paid to the morphology and attribute variations of urban hotspots across different time periods. The related studies on the spatial and temporal distribution of urban hotspots mainly describes the location changes of urban hotspots at different time points. This study analyzes the morphological variations of urban hotspots and discusses the relationship between these variations and urban built environment. This is crucial to the understanding of urban spatial structure beyond the physical environment.

References

- 1) Bilková, K., Krizan, F., & Barlík, P. (2016). Consumers preferences of shopping centers in Bratislava (Slovakia). *Human Geographies*, 10(1), 23.
- 2) El-Ashmawy, K. L. (2016). Testing the positional accuracy of OpenStreetMap data for mapping applications. *Geodesy and Cartography*, 42(1), 25-30.
- 3) Jia, T., Carling, K., & Håkansson, J. (2013). Trips and their CO₂ emissions to and from a shopping center. *Journal of Transport Geography*, 33, 135-145.
- 4) Jiang, L., Deng, X., & Seto, K. C. (2012). Multi-level modeling of urban expansion and cultivated land conversion for urban hotspot counties in China. *Landscape and Urban Planning*, 108(2-4), 131-139.
- 5) Krehl, A., & Siedentop, S. (2019). Towards a typology of urban centers and subcenters—evidence from German city regions. *Urban Geography*, 40(1), 58-82.
- 6) Leslie, T. F. (2010). Identification and differentiation of urban centers in Phoenix through a multi-criteria kernel-density approach. *International Regional Science Review*, 33(2), 205-235.
- 7) Li, J., Long, Y., & Dang, A. (2018). Live-Work-Play Centers of Chinese cities: Identification and temporal evolution with emerging data. *Computers, Environment and Urban Systems*, 71, 58-66.
- 8) Limtanakool, N., Schwanen, T., & Dijst, M. (2009). Developments in the Dutch urban system on the basis of flows. *Regional Studies*, 43(2), 179-196.
- 9) McDonald, J. F. (1987). The identification of urban employment subcenters. *Journal of Urban Economics*, 21(2), 242-258.
- 10) McMillen, D. P., & Smith, S. C. (2003). The number of subcenters in large urban areas. *Journal of urban economics*, 53(3), 321-338.

- 11) Melo, P. C., Graham, D. J., Levinson, D., & Aarabi, S. (2017). Agglomeration, accessibility and productivity: Evidence for large metropolitan areas in the US. *Urban Studies*, 54(1), 179-195.
- 12) Newling, B. E. (1969). The spatial variation of urban population densities. *Geographical Review*, 242-252.
- 13) Qin, X., Zhen, F., & Gong, Y. (2019). Combination of big and small data: Empirical study on the distribution and factors of catering space popularity in Nanjing, China. *Journal of Urban Planning and Development*, 145(1), 05018022.

Chapter 4

Street-view Impression of Urban Hotspots

Chapter 4

Street-view Impression of Urban Hotspots

4.1 Additional landscape value of urban hotspots

As mentioned in introducing the definition of urban hotspots (Chapter 1.2.2), urban hotspot spaces not only play the role of regional business centers, transportation hubs, and industrial clusters, and affects the urban economic pattern, commuting conditions, and ecological environment, but also gather a large number of urban people flows. Imperceptibly, urban hotspots take responsibility for urban landmarks (Sun and Yu 2021). After arriving in a city, tourists and foreign visitors will give priority to visiting the hotspot areas of the city (Hartmann et al. 2020). Therefore, urban hotspots can be the windows of cities. It is also of great significance to understand the landscape characteristics of urban hotspots and the landscape evaluation of urban hotspots from tourists and foreign visitors (Latu and Bulai 2011).

(1) Urban hotspots as landmarks

By definition, urban landmarks refer to any natural or man-made structures that are recognized by the public of a particular place or city (Costonis 1971). They are as important as the orientation and planning of a city. When a person visits a city for the first time, they tend to map the city through these landmarks.

With the popularity of the Internet and the development of cross-regional transportation, urban landmarks are no longer played by the buildings with the largest regional scale and the highest height in recent decades. The most distinctive space in the city will establish a unique local symbol because of its unique natural landscape and cultural customs. These regions will naturally be known by people in the surrounding regions and even around the world with the continuous development of the Internet (Kim et al. 2020). Finally, when these areas are covered by many local and foreign people, they will be identified as

urban hotspots that play an important role in the city according to the definition of urban hotspots (Hu et al. 2019). Therefore, under the background of today's era, urban hotspots often represent the external image of the city as landmarks.

(2) Urban hotspots as tourist destination

Another reason why must study the landscape characteristics of the urban hotspot is about to be its relationship with tourism. At the beginning of tourism, when tourists travel to a destination city, the urban space related to tourism often includes (1) Transportation infrastructure for commuting activities; (2) Hotel facilities for residential activities; (3) A scenic spot for sightseeing activities (Mandić et al. 2018). With the development of modern cities and the progress of tourism, the types and quantity of urban spaces related to tourism activities show an upward trend. Some areas in the city that are not tourist attractions will also attract a large number of urban population flow (Schuckert et al. 2015).

For instance, the Manhattan block in New York and the Bund in Shanghai are the 2 typical new-style urban tourist destinations. On the one hand, from their role in cities, they are the concentration of financial and trade institutions and the economic centers of two cities (or even two countries). On the other hand, these two areas are lined with a huge number of world-famous skyscrapers and parks, both of which are adjacent to large water systems (Manhattan is surrounded by the East River, Hudson River, and Harlem River, and the Bund is close to the Huangpu River). These two economic centers are usually listed as tourist destinations while contributing a high amount of GDP (gross domestic product) to the two cities. In fact, in modern cities, with the rapid popularization of the Internet and SNS platform, a specific urban area can be quickly known by other users through online comments published by SNS users. When a place is visited a lot in the virtual space, it will have the potential to become a potential tourist destination in the real world (Wikipedia 2021).

In summary, in addition to playing the functions of urban economic center, industrial cluster, and transportation hub, urban hotspots will also attract many tourists in the form of landmarks and display the image and landscape of the city as a window to the outside world. Therefore, the research on the landscape

characteristics and street-view impression of urban hotspots has important academic and practical significance for urban planning and landscape design.

4.2 Extraction of street-view impression using SNS data

The first problem to be solved in this study is how to accurately obtain or estimate the street-view impression of urban residents and visitors on urban hotspots. Considering that the extraction of urban hotspots in this study is mainly based on the check-in data generated by SNS platform, this study will continue to propose an urban hotspot street-view extraction method relying on SNS platform data in this chapter. Similarly, this study will continue to introduce in detail the operation process and output results of the urban hotspot street-view impression analysis with the help of a case from the center of Nanjing city.

4.2.1 Street-view impression extraction process

The new approach consists of three steps: collecting SNS online review data, extracting urban scenery, and labeling subjective landscape evaluations. The detailed research procedure was presented in the research roadmap (Figure 4-1):

- *First of all, this study selected popular urban areas from Nanjing, China, which are representative hotspots with attractive human and natural urban landscapes, to collect the online review word-of-mouth data in the formats of text and picture generated within the corresponding spaces.*
- *Then, the CNN (convolutional neural network) algorithm is introduced to filter and transform the collected SNS image data into urban sceneries of natural landscapes and cultural landscapes. The classification and filtering of landscape are mainly based on the type and proportion of main landscape elements contained in the photos.*
- *Finally, the NLP (natural language processing) algorithm is introduced to extract the subjective evaluation from SNS users on the images corresponding to the landscapes of urban hotspots based on the emotional tendency contained in SNS text.*

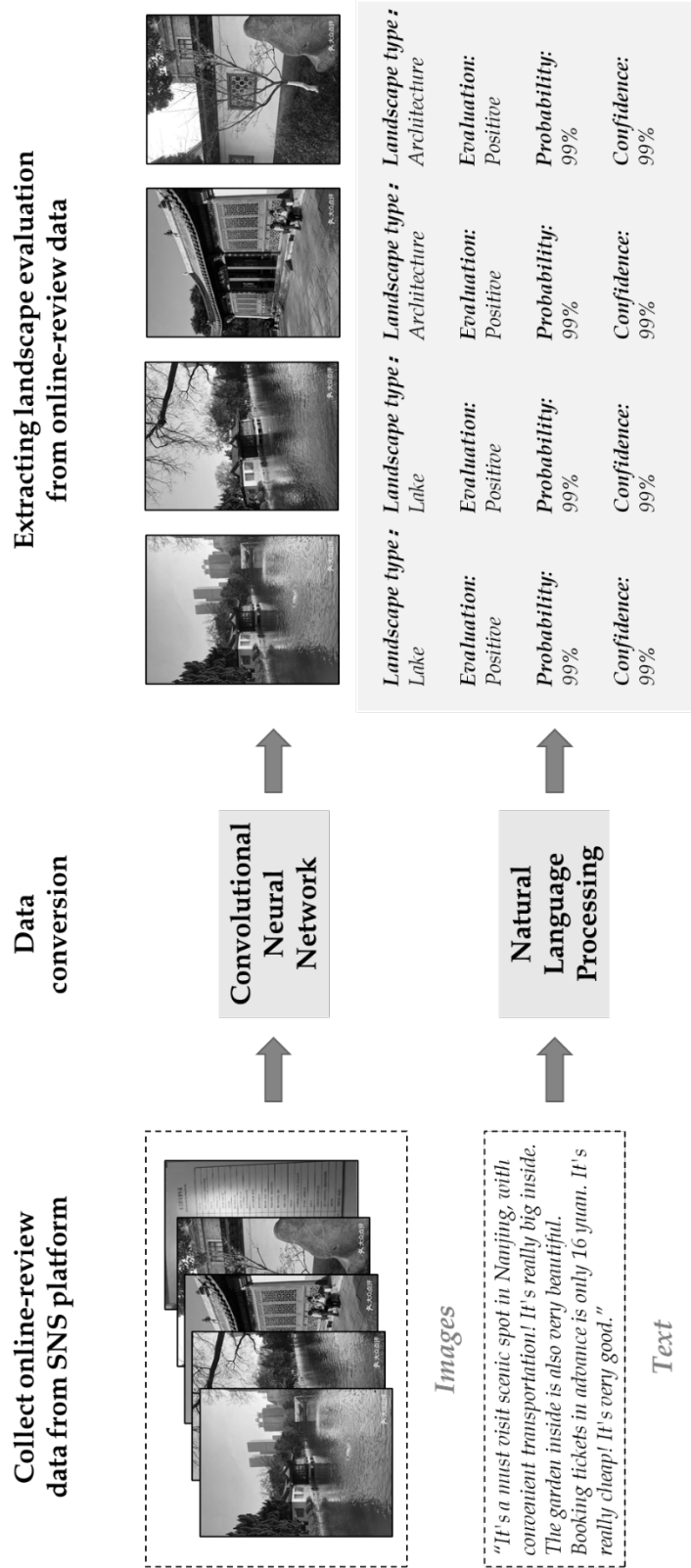


Figure 4-1 Research process of extraction of street-view impression

4.2.2 SNS online-review data collection

At the very beginning, the SNS online review data will be collected, including the images and text comments. The collection method of SNS comment data is still the API of SNS platform.

The sample SNS comment data utilized in this study was collected in 41 popular urban areas in Nanjing, China. According to the urban hotspots identification results presented in Chapters 2 and 3, these hotspot sites play the roles of urban tourist attractions, sightseeing blocks, and landmarks, which mainly consist of variety of urban cultural and natural landscapes. The spatial distribution of the experimental sites is presented in Figure 4-2.

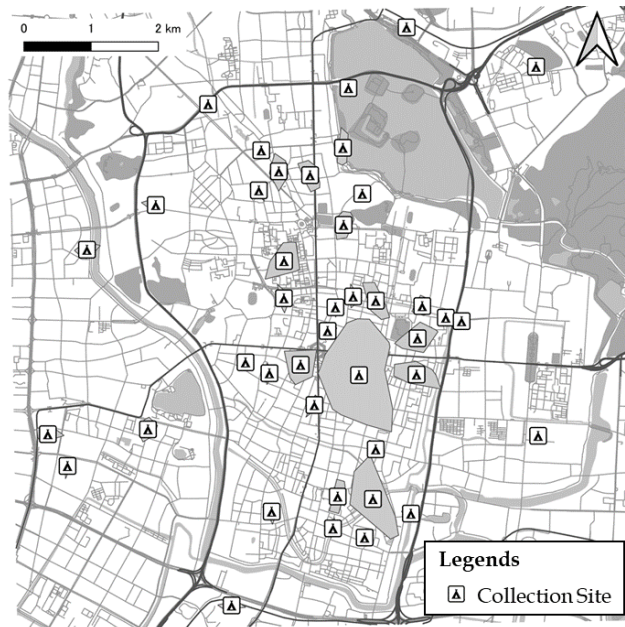






Figure 4-2 SNS data collection sites (location of urban hotspots)

Specifically, SNS data collection platforms are Dianping.com (China) and Google Map (Wikipedia 2021a). Among them, the former is mainly used to collect SNS comments from local residents and domestic tourists in Nanjing, while the latter is mainly used to collect comments from foreign tourists.

Table 4-1 Sample SNS online-review dataset

Hotspot	Text comment	Related images	Platform	Publisher
Xuanwu Lake	I went a little bit off the standard tourist track by choosing to stop in Nanjing. I don't regret the experience overall...		Google Map	Foreign tourist
Xuanwu Gate	Parque grande y antiguo. Muy transitado por turistas. Tiene barcas que se pueden alquilar para pasear por los lagos		Google Map	Foreign tourist
Xuanwu Gate	湖边的小路是亮点；此外岛上的盆景园也不错。唯一缺点就是蚊子太多了，在这么冷的天气。		Dianping	Local residents
Zijin Hill	工作之余来这里转转，公园很大，只走了一段，人还不是很多，比起西湖的话，这里更像是个公园。		Dianping	Domestic tourist
...

The significance and value of collecting online comment data from different platforms are that it can obtain the landscape evaluation of urban hotspots by people with different attributes and eliminate the impact of human attributes on the evaluation results and related errors.

The collected data is released between January 2012 and December 2019. The data collection approach is to obtain the top 10 most prepared comments, which are composed of text comments and image records reporting the corresponding areas released by the publishers with the help of a Python-based crawler tools. In addition, when collecting online word-of-mouth data, this study also checks each publisher's property information registered on Dianping.com platform and Google Map to classify the SNS users into 3 types of local residents, domestic tourists, and foreign tourists. Finally, a total of 4552 photos (local residents: 1420; domestic tourists: 2285; foreign tourists: 847) images with text comments (samples in Table 4-1).

4.2.3 Filter landscape from online-review photos

The photos obtained by the data collection method designed in chapter 4.2.2 were the travel notes pictures released by SNS platform users, which contained a large number of images with portraits and interior furniture as the main content. Therefore, this study first introduced the image semantic segmentation algorithm to filter the photos with natural and cultural urban landscape as the main object from the collected SNS photos.

What is image segmentation? It is known that an image is nothing but a collection of pixels. Image segmentation is the process of classifying each pixel in an image belonging to a certain class and hence can be thought of as a classification problem per pixel. There are two types of segmentation techniques (Li et al. 2017; Matcha 2021).

- **Semantic segmentation:** Semantic segmentation is the process of classifying each pixel belonging to a particular label. It doesn't differ across different instances of the same object. For example, if there are 2 cats in an image, semantic segmentation gives same label to all the pixels of both cats.

- Instance segmentation: Instance segmentation differs from semantic segmentation in the sense that it gives a unique label to every instance of a particular object in the image. As can be seen in the image below all 3 dogs are assigned different colors (i.e., different labels). With semantic segmentation all of them would have been assigned the same color.

In this study, the model leveraged to finish the image semantic segmentation task is PSPNet. PSPNet, or Pyramid Scene Parsing Network, is a semantic segmentation model that utilizes a pyramid parsing module that exploits global context information by different-region based context aggregation, which is one of the most well-recognized image segmentation algorithms as it won ImageNet Scene Parsing Challenge 2016 and its paper is highly cited by the computer vision community.

The PSPNet architecture takes into account the global context of the image to predict the local level predictions hence gives better performance on benchmark datasets like PASCAL VOC 2012 and cityscapes. The model was needed because FCN based pixel classifiers were not able to capture the context of the whole image (Esri 2021). The pyramid pooling module is the main part of this model as it helps the model to capture the global context in the image which helps it to classify the pixels based on the global information present in the image.

PSPNet has better performance than the mainstream Fully Convolutional Network (FCN). The model was needed because FCN based pixel classifiers were not able to capture the context of the whole image (Zhao et al. 2016). Specifically, image semantic segmentation algorithm utilized in this study was constructed through the pre-trained PSPNet built by MXNet architecture (based on ade20k dataset). This algorithm is mainly used to extract the main landscape elements (a total of 150 types of landscape elements) of SNS images (Figure 4-3 & 4-4) and then judge the landscape type (natural or cultural) described in the image. And The 150 types of landscape elements extracted from the ade20k-based model is listed at: <https://github.com/CSAILVision/sceneparsing>. After this step, the images with portraits and interior furniture (cabinets, tables, etc.) as the main visual elements were removed, and a total of 2576 landscape photos (natural landscape: 1292; cultural landscape: 1284) were retained.

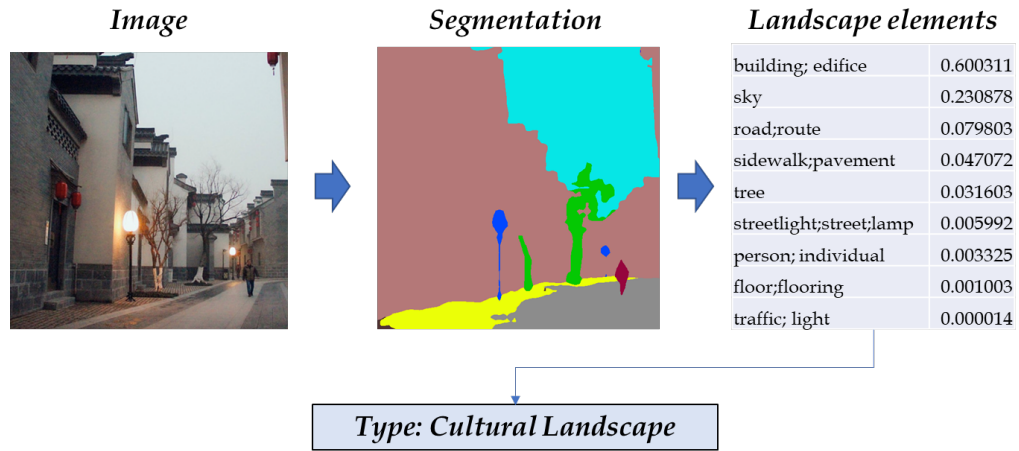


Figure 4-3 Image semantic segmentation algorithm to judge landscape type

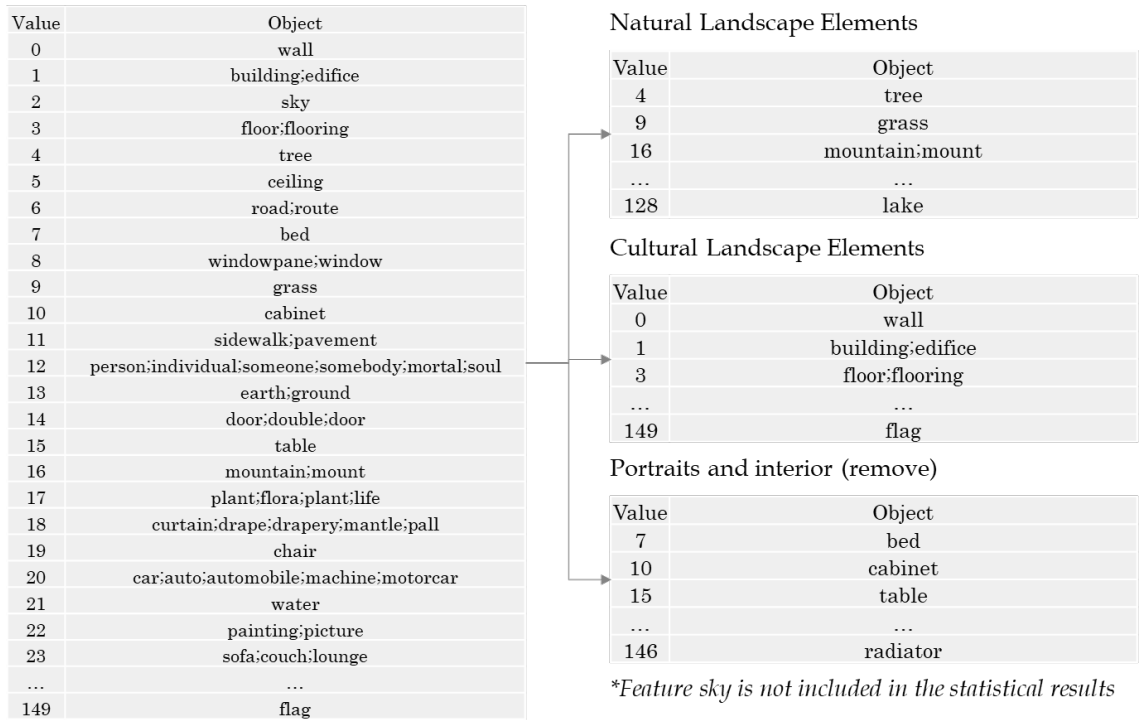


Figure 4-4 Cleaning and division standard of landscape elements

4.2.4 Extract evaluation from online-review text

(1) Comment opinion extraction

The text comments collected in this study correspond to at least one picture of the urban scene. Since this study plan to extract the subjective evaluation of the corresponding scene from the text comments accurately, this study needs to remove the content irrelevant to the urban landscape from the texts. Here, this study introduces comment opinion extraction algorithm in NLP technology. The option extraction algorithm is mainly used to extract and understand the object described by the online comment text and its evaluation.

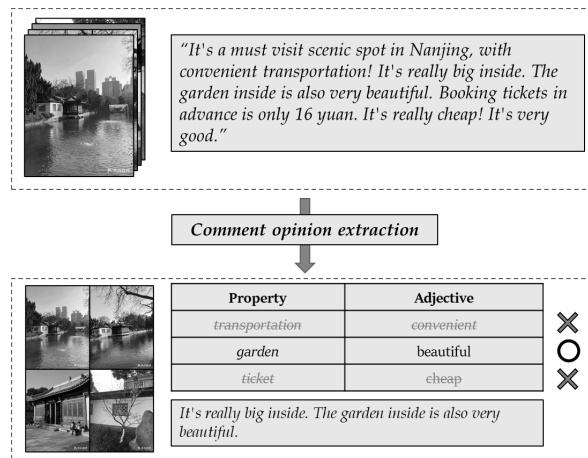


Figure 4-5 Comment opinion extraction algorithm

Using the comment option extraction algorithm, this study is able to divide a long string of words into different types of words (adjective, apposition, adverb, conjunction, noun, etc.). Specifically, the implementation of comment opinion extraction algorithm is based on Baidu Brain framework and Python programming language. This algorithm could extract the core objects and evaluation results by analyzing the attributes and dependency of the words in text paragraph in GBK format. As shown in Figure 4-5, using this algorithm, the main evaluation objects (attribute words) and evaluation results (adjective words) of each comment text are split and extracted, and the paragraphs irrelevant to landscape evaluation are eliminated.

(2) Text sentiment analysis

After filtering the collected SNS data into 2576 urban landscape photos with subjective comments corresponding to their dominating visual elements, this study plans to judge the emotional tendency contained in the comment paragraph and then infer the preference of the SNS publisher to the urban landscapes in the photos. Here, this study introduces the text sentiment analysis algorithm in NLP technology.

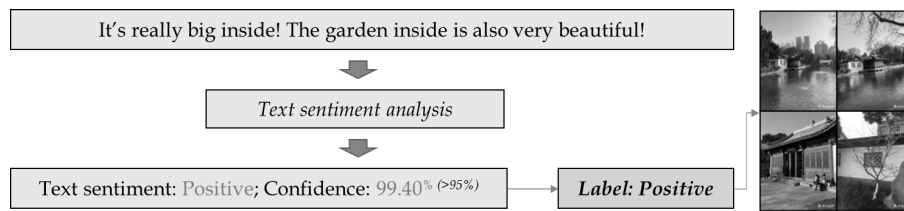


Figure 4-6 Text sentiment analysis algorithm

The implementation of the algorithm is based on Baidu Brain framework and Python programming language as well. The algorithm analyzes the synchronous and semantic aspects of a text by matching the words in the paragraph with the pre-constructed bag of words, which is a list of positive terms (like good, beautiful, useful, interesting, etc.) and negative terms (such as bad, ugly, uncomfortable, frustrated, etc.). The text with subjective opinion information is judged by the category of emotional polarity (positive, negative and neutral as labels), and the confidence level. As shown in Figure 4-6, using this algorithm, the sentiment extracted from the text is recognized as the subjective evaluation of corresponding urban scene photos by SNS publishers. The sentiment with confidence value over the threshold value of 95% is further utilized as the label for the urban landscape images. After this step, 1290 landscape photos with significant emotional labels (Positive/Negative) were retained.

4.2.5 Extraction results of street-view impression

Using the urban hotspot street-view impression extraction method proposed in chapters 4.2.1 to 4.2.4, the street-view impression results of 41 urban hotspots in the central area of Nanjing by urban residents and tourists are extracted.

The specific results are shown in Figure 4-7 and 4-8. This study presents the spatial distribution of landscape evaluation praise rate of urban hotspots with

the main landscape types of natural and cultural landscape, respectively. By analyzing the relationship between the landscape characteristics of urban hotspots and the street-view impression, this study tries to find out the problems existing in the landscape design and planning of urban hotspots.

Specifically, the main landscape type of each urban hotspot is determined by the type of landscape photos generated in the hotspot area (refer to chapter 4.2.3). When more than 50% of the main elements in the landscape photos corresponding to an urban hotspot area are natural landscape, the main landscape type of the urban hotspot is defined as natural landscape. On the contrary, it is defined as cultural landscape.

(1) Natural landscape

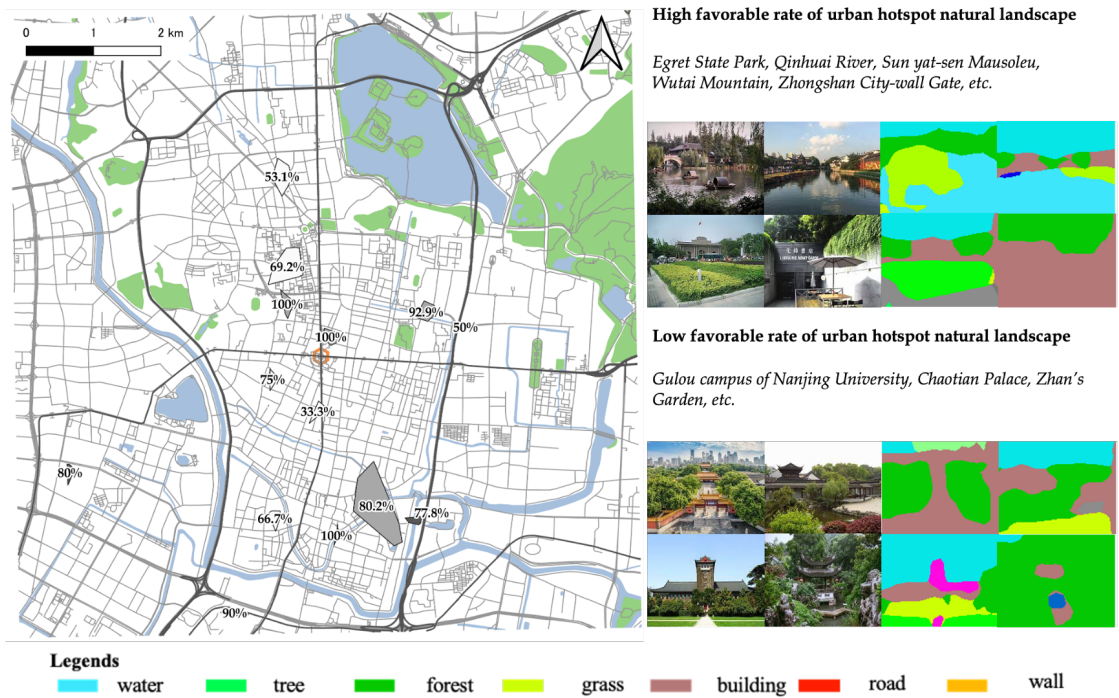


Figure 4-7 Natural landscape's street-view impression

Figure 4-7 shows the urban hotspots (five for each) with the highest and lowest favorable rates for the natural landscape in the center of Nanjing. By analyzing the landscape characteristics (landscape element types and landscape element composition) of these urban hotspots, it has been found that:

- The landscape of urban hotspots with a large area of water and exquisite flower beds as the landscape theme can generally get better subjective evaluation from visitors.
- On the contrary, urban hotspots with antique buildings and classical gardens as the core natural landscape have generally received low praise.
- In addition, the street-view impression of urban hotspots with rockery landscape as the main body is controversial.

(2) Cultural landscape

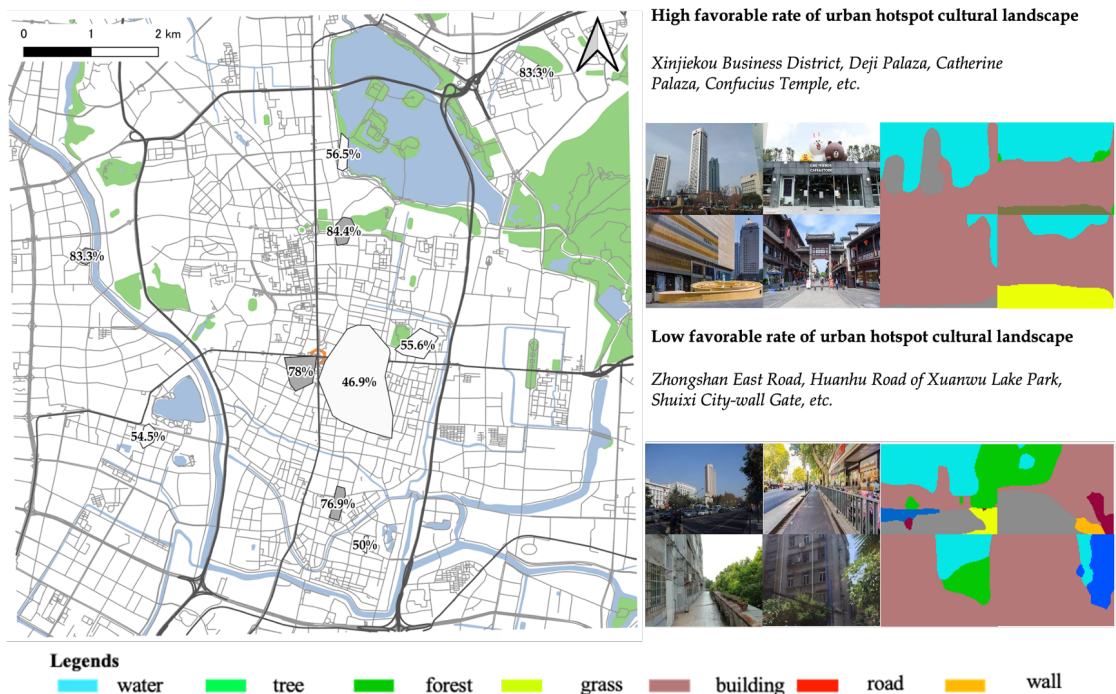


Figure 4-8 Cultural landscape's street-view impression

Figure 4-8 shows the urban hotspots (five for each) with the highest and lowest favorable rates for the cultural landscape in the center of Nanjing. By analyzing the landscape characteristics (landscape element types and landscape element composition) of these urban hotspots, it is found that:

- The street-view impression of urban hotspots with extreme height and large-scale modern style buildings as the main body of cultural landscape is better.

- *The cultural landscape composed of daily and life flavor urban streets and antique commercial streets is not liked by visitors.*

4.3 Comparison on the landscape preference

Subjective cognition on urban landscapes derives from the interactions between the physical characteristics of an urban landscape and the psychological responses of the beholder (Abello and Bernáldez 1986). It has played an important role in the process of urban planning and design (Howley et al. 2012). Several investigations have pointed out that the cognitive preference varies depending on the physical characteristics of urban scenes and the demographic attributes of observers (Wang and Zhao 2017). Therefore, exploring the relationships between demographic attributes and urban scene preference provides valuable information for landscape architects and urban planners (Kuper 2017). Specifically, the identification of similarities in cognitive preference across groups would assist the development of general guidelines for the planning and design, whereas the demonstration of group differences would help to sort out the cases or conditions of a specific group's preference (Howley 2011). However, these traditional methods inevitably have the problems of the high cost of data acquisition and difficulty of data processing. In certain extreme scenarios (such as different countries, time periods, or languages), it is even difficult to accurately obtain the landscape preferences of urban residents due to the small sample size of data (Oshi et al. 2006; Oshi et al. 2007).

4.3.1 Landscape preference comparison procedure

On this basis, this study plans to design a new framework based on machine learning algorithms (CNN and NLP algorithms) to make the extraction and analysis of urban landscape preferences more convenient and faster (Satoshi and Kotaro 2019). This study conducts an empirical experiment to introduce the construction and implementation process of this method in detail and showed the practicability and superiority of this algorithm by analyzing the experimental results (Figure 4-9).

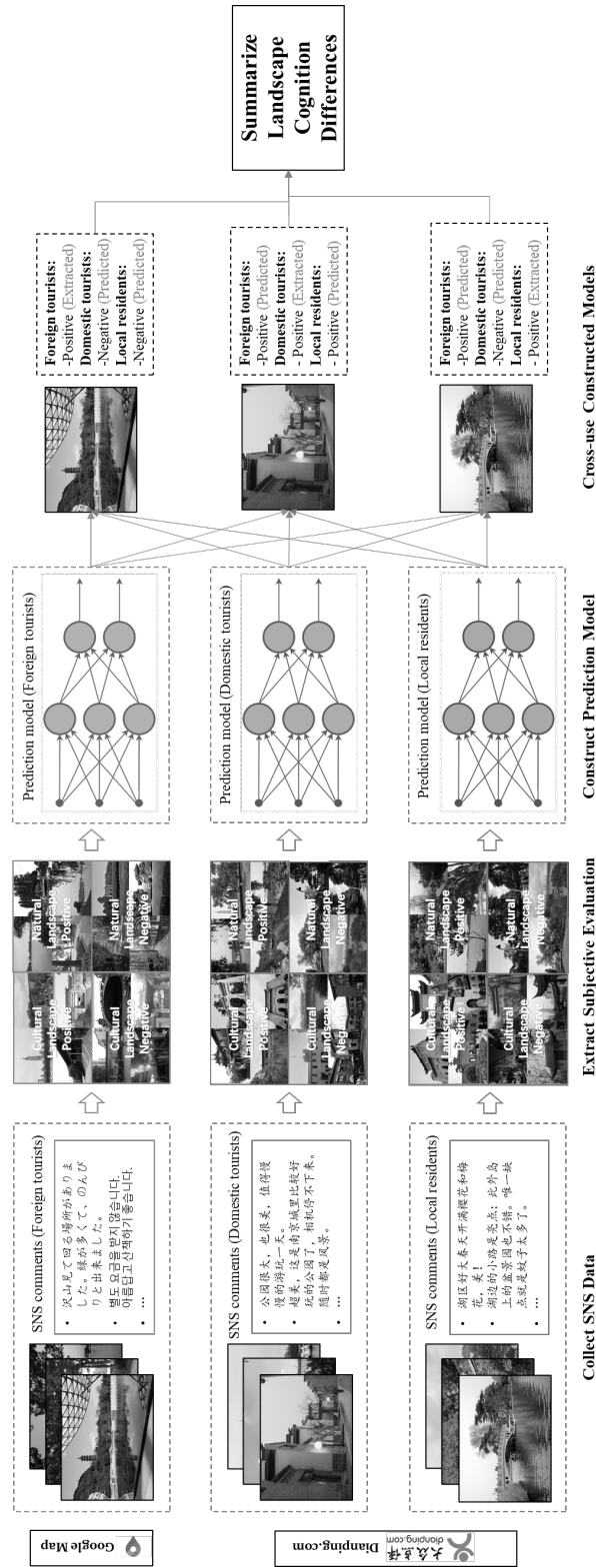


Figure 4-9 Research process of extraction of landscape preference

4.3.2 Construction and performance of the models

The 1290 filtered photos with subjective sentiment labels were classified into 12 groups according to the publisher's attributes and label types (Table 4-2). Then, this study divided these 1290 photos into training data and test data of landscape preference prediction model according to 80% to 20%. Specifically, if 80% of the total number of each group of photos is not an integer, this study rounds it down and records the results as training data, and the rest as testing data. As a result, this study input 1016 of the photos as training data into 6 CNN architectures to generate 6 prediction models which can represent the cognitive preference for the cultural and natural landscape of local residents, domestic tourists, and foreign tourists, respectively. The remaining 274 photos will be used as testing data for the model accuracy test.

Table 4-2 Input data of prediction models

Group	Attribute	Type	Label	Training (80%)	Testing (20%)	Sum
1		Natural	Positive	118	30	148
2	Local	Landscape	Negative	28	10	38
3	Residents	Cultural	Positive	140	36	176
4		Landscape	Negative	36	11	47
5		Natural	Positive	160	40	200
6	Domestic	Landscape	Negative	40	11	51
7	Tourists	Cultural	Positive	196	51	247
8		Landscape	Negative	48	14	62
9		Natural	Positive	118	30	148
10	Foreign	Landscape	Negative	28	11	39
11	Tourists	Cultural	Positive	84	23	107
12		Landscape	Negative	20	7	27
Total amount				1016	274	1290

It is worth mentioning that due to the lack of data volume corresponding to each label in the input dataset, this study utilized the method of data

augmentation (add noise, blur, expose, flip and rotate) to expand the size of the training dataset (Figure 4-10). After each original image is optimized by five data enhancement tools (cross-use and superpose-use) in Apple Core ML architecture, the overall volume of the database could meet the training data requirements of the image classifier model based on CNN network.

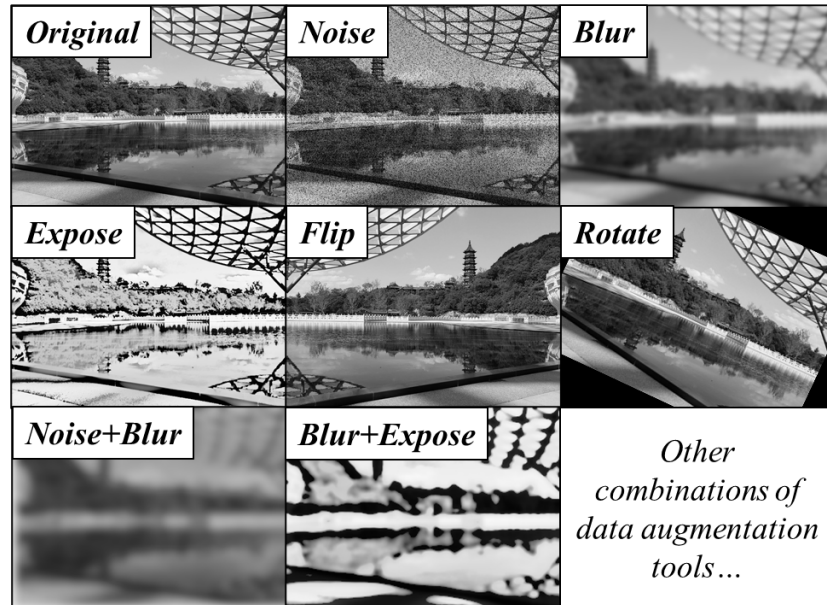


Figure 4-10 Data augmentation tools

Finally, this study trained 6 landscape preference prediction models based on CNN network. The accuracy of the 6 urban scene preference prediction models built in this study ranged from 98.0% to 100% and 84.4% to 96.8% at 25th iteration in the training dataset and the testing dataset (Figure 4-11), respectively. Specifically, the model constructed for predicting domestic tourist's preference on natural landscapes showed the highest accuracy in both training and validation, while that on cultural landscapes is the lowest.

In general, since the accuracy of landscape preference prediction models reached 99.55% and 91.07% in the process of training and validation averagely, the models can accurately predict the subjective preferences of three groups of people for different types of urban landscapes.

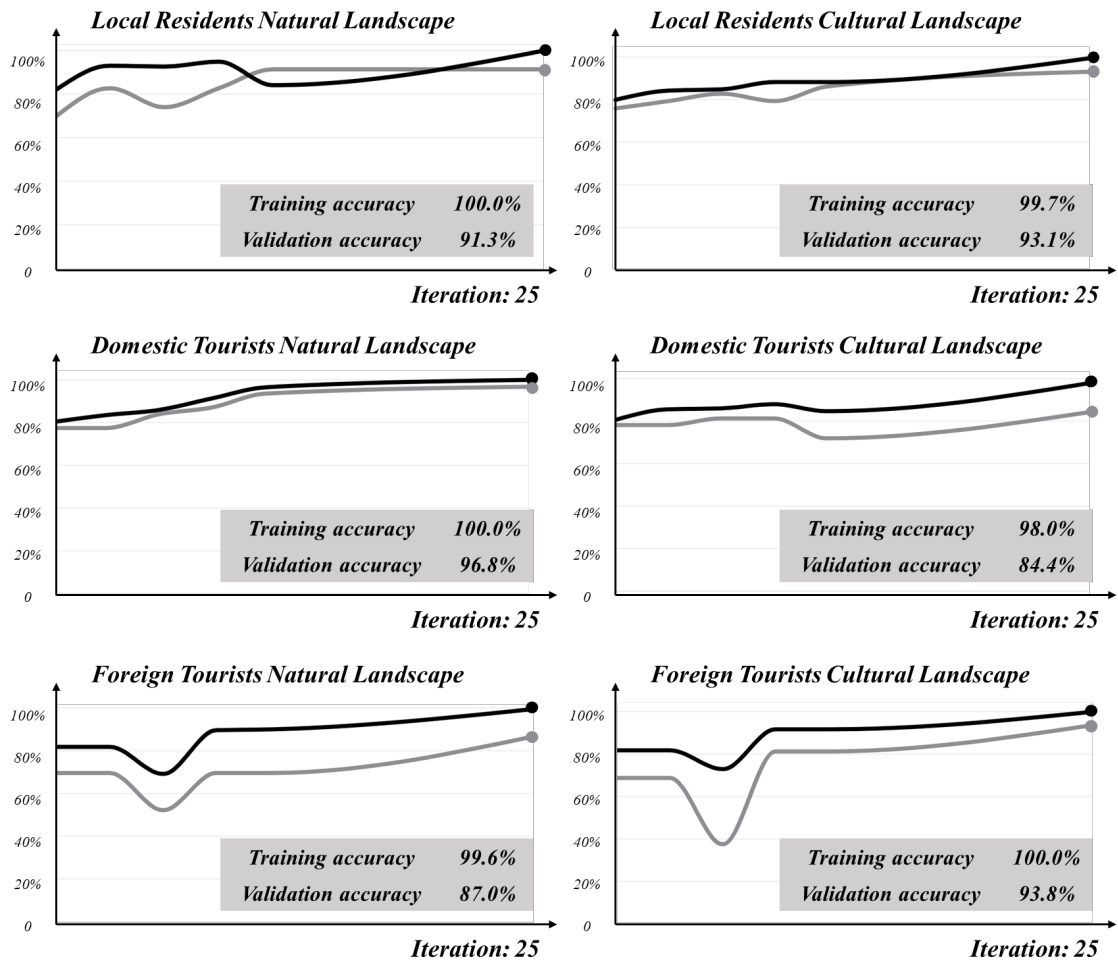


Figure 4-11 Accuracy of landscape preference prediction models

4.3.3 Extraction of urban landscape cognition difference

Next, the 6 models were cross used to estimate the cognition differences in Nanjing’s urban landscape within local residents, domestic tourists, and foreign tourists. Specifically, 1290 urban landscape photos used in this study for training the prediction models all contain subjective evaluation from SNS users with one particular attribute. This study further used the models to predict the subjective evaluation results of SNS users with the other two groups of people and record the experimental records in which the confidence parameter of the model prediction results is more than 95%.

Table 4-3 Prediction results

Label	L-P-p.	L-N-p.	L-Cons.	D-P-p.	D-N-p.	D-Cons.	F-P-p.	F-N-p.	F-Cons.
L-NL-P (148)				120	11	81.1%	123	4	83.1%
L-NL-N (38)				2	28	73.7%	5	22	57.9%
L-CL-P (176)		--		129	8	73.3%	141	6	80.1%
L-CL-N (47)				3	31	66.0%	11	28	59.6%
D-NL-P (200)	150	14	75.0%				139	13	69.5%
D-NL-N (51)	2	37	72.6%				4	31	60.8%
D-CL-P (247)	193	14	78.1%		--		187	13	75.7%
D-CL-N (62)	6	50	80.7%				5	46	74.2%
F-NL-P (148)	116	7	78.4%	104	24	70.3%			
F-NL-N (39)	8	27	69.2%	13	16	41.0%			
F-CL-P (107)	79	14	73.8%	74	13	69.2%		--	
F-CL-N (27)	4	14	51.9%	6	19	70.4%			

*Note: L, D, and F denote local residents, domestic tourists and foreign tourists, respectively; NL and CL denote natural landscape and cultural landscape, respectively; P and N denote positive and negative subjective sentiment, respectively; Cons. denotes the consistency between the sentiment label and prediction results; p. denotes that the prediction results with confidence value over 95%.

The prediction results evidently demonstrated that there are significant differences in the subjective evaluation results on Nanjing’s urban landscapes among local residents, domestic tourists, and foreign tourists at the 95% confidence level. As shown in Table 4-3, the consistency between the label and the prediction sentiment from the other two groups of SNS users only ranged from 41.0%~83.1%. And the difference between the labels and prediction results existed in 29.3% of photos. Hence, this study classifies the photos according to the evaluation results and summarizes the landscape cognitive bias by analyzing the content, elements, and composition characteristics of all types of urban landscapes.

Base on the classification and content analysis on the selected typical urban scenes (3 photos with highest confidence values were selected as samples) with cognition differences based on CNN network (Table 4-4), the main findings can be summarized into the following three points:

Table 4-4 Prediction results

		Sentiment	Photo a	Photo b	Photo c	Element a	Element b	Element c							
Urban Landscape with same cognition results	Natural Landscape	Positive													
		Negative													
	Cultural Landscape	Positive													
		Negative													
Cognition bias between residents and tourists	Natural Landscape	Resident: Positive Tourist: Negative													
		Resident: Negative Tourist: Positive													
	Cultural Landscape	Resident: Positive Tourist: Negative													
		Resident: Negative Tourist: Positive													
Cognition bias between domestic and foreign tourists	Natural Landscape	Domestic: Positive Foreign: Negative													
		Domestic: Negative Foreign: Positive													
	Cultural Landscape	Domestic: Positive Foreign: Negative													
		Domestic: Negative Foreign: Positive													
Legends															
			water		tree		forest		grass		building		road		wall

Firstly, the first row in Table 4-4 showed the consistency of the landscape preferences of the three attribute groups. In the natural landscape, the subjective evaluation results of local residents, domestic tourists, and foreign tourists on the waterfront landscape composed of a large area of water and forest and the mountain landscape with bare peaks as the core are consistent. People give a positive and negative evaluation to these two kinds of landscapes respectively. Similarly, the subjective evaluation results of the three groups of people on the cultural landscape of street view composed of low-rise modern edifices and urban lanes within dilapidated flat rooms are the same. The former is preferred by all groups of people, while the latter is vice versa.

Secondly, the second row in Table 4-4 showed the differences in landscape preferences between local residents and (domestic and foreign) tourists. Local residents and tourists have a different subjective evaluation of the natural landscape of urban parks with different characteristics. Local residents prefer the large area of forest on both sides of the artificial Road and Bridge in the park, while tourists give a higher evaluation on the park landscapes with historical pavilions and stone tablets as the core. This difference also exists in the human landscape with different styles of buildings as the main body. Local residents prefer the edifices which mixed the style of modern and “Minguo” (19th-20th century of China), while tourists prefer reconstructed antique style buildings.

Thirdly, the third row in Table 4-4 showed the differences in landscape preferences between domestic people (residents and tourists) and foreign tourists. Specifically, the unique style of statues, rockeries can make the natural landscape favored by domestic people, while foreign tourists will pay more attention to the open grasslands when evaluating natural landscapes. In the cultural landscape, buildings with characteristics of historic doors, murals, and decorations with prominent colors (i.e., golden) are the key factors to attract domestic people and foreign tourists, respectively.

4.4 Summary on the street-view impression

In general, this study combines the online review data of SNS users and machine learning algorithms to design an urban hotspot street-view impression and preference prediction model. This model can not only estimate the subjective cognitive results of urban residents, visitors, and tourists on the urban landscape but also make it possible to quantitatively compare the landscape preferences of residents with different demographic characteristics. In a case study targeting the street-view impression and preference differences of local residents, domestic tourists, and foreign tourists on the landscape of urban hotspots, this study collected the text and picture comments generated in the core area of Nanjing Dianping.com and Google map platforms to construct urban scene preference models representing the cognition of three groups of people on the natural and cultural landscape and summarized their landscape evaluation and preference by comparing the outputs from different models. The results showed that:

With the help of NLP and CNN techniques, taking the core area of Nanjing as an example, this study demonstrated a complete experiment of urban hotspot street-view impression and feature extraction. The experimental results show that: (a) For natural landscapes, the street-view of urban hotspots with a large area of water and exquisite flower beds as the landscape theme can generally get a better subjective evaluation from visitors. On the contrary, urban hotspots with antique buildings and classical gardens as the core natural landscape have generally received low praise. In addition, the landscape evaluation of urban hotspots with rockery landscape as the main body is controversial. (b) For cultural landscapes, the street-view of urban hotspots with extreme height and large-scale modern-style buildings as the main body of cultural landscape is better. The cultural landscape composed of daily and life flavor urban streets and antique commercial streets is not liked by urban residents, visitors, and tourists.

This study also constructed 6 models representing the cognitive preference on Nanjing's cultural and natural landscape of local residents, domestic tourists, and foreign tourists, respectively. The accuracy of the prediction models built in this study is averagely 99.55% and 91.07% at the 25th iteration in the training dataset and the testing dataset, respectively, which means that the constructed models can accurately predict the subjective preferences of three groups of people on different types of urban scenes. By cross using the subjective

preferences prediction models constructed for different attributes of people, cognition differences between local residents, domestic tourists, and foreign tourists are extracted. The main landscape contents, elements, and composition characteristics of urban scenes are closely related to the subjective preferences of three groups of people on urban landscapes.

In summary, the main contribution of this study is to propose a quantitative framework of landscape cognition prediction without visiting and investigation by using machine learning algorithms and SNS data. It is also worth mentioning that the approach proposed in this study can not only extract the landscape preferences of different groups of people but also be used to predict people's subjective evaluation of future urban design and planning (with CGs or rendering images).

References

- 1) Abello, R. P., & Bernáldez, F. G. (1986). Landscape preference and personality. *Landscape and urban planning*, 13, 19-28.
- 2) Costonis, J. J. (1971). The Chicago Plan: Incentive zoning and the preservation of urban landmarks. *Harv. L. Rev.*, 85, 574.
- 3) Esri. (2021). How PSPNet works? | ArcGIS Developer. ArcGIS. Retrieved October 12, 2021, from [https://developers.arcgis.com/python/guide/how-
pspnet-works/](https://developers.arcgis.com/python/guide/how-
pspnet-works/)
- 4) Hartmann, R., & Su, M. M. (2020). Tourism to Lu Gou Qiao: enduring scenic qualities of a landmark bridge and a difficult legacy of a conflict site. *Journal of Heritage Tourism*, 1-11.
- 5) Howley, P. (2011). Landscape aesthetics: Assessing the general public's preferences towards rural landscapes. *Ecological Economics*, 72, 161-169.
- 6) Howley, P., Donoghue, C. O., & Hynes, S. (2012). Exploring public preferences for traditional farming landscapes. *Landscape and urban planning*, 104(1), 66-74.
- 7) Hu, Q., Bai, G., Wang, S., & Ai, M. (2019). Extraction and monitoring approach of dynamic urban commercial area using check-in data from Weibo. *Sustainable cities and society*, 45, 508-521.
- 8) Iatu, C., & Bulai, M. (2011). New approach in evaluating tourism attractiveness in the region of Moldavia (Romania). *International Journal of Energy and Environment*, 5(2), 165-174.
- 9) Kim, D., Kang, Y., Park, Y., Kim, N., & Lee, J. (2020). Understanding tourists' urban images with geotagged photos using convolutional neural networks. *Spatial Information Research*, 28(2), 241-255.
- 10) Kuper, R. (2017). Evaluations of landscape preference, complexity, and coherence for designed digital landscape models. *Landscape and Urban Planning*, 157, 407-421.

- 11) Li, F. F. (2017). CS231n Convolutional Neural Networks for Visual Recognition. Cs231n. <https://cs231n.github.io/>
- 12) Mandić, A., Mrnjavac, Ž., & Kordić, L. (2018). Tourism infrastructure, recreational facilities and tourism development. *Tourism and hospitality management*, 24(1), 41-62.
- 13) Matcha, A. C. N. (2021, May 20). A 2021 guide to Semantic Segmentation. AI & Machine Learning Blog. <https://nanonets.com/blog/semantic-image-segmentation-2020/>
- 14) Navío-Marco, J., Ruiz-Gómez, L. M., & Sevilla-Sevilla, C. (2018). Progress in information technology and tourism management: 30 years on and 20 years after the internet-Revisiting Buhalis & Law's landmark study about eTourism. *Tourism management*, 69, 460-470.
- 15) Oishi, H., Murakawa, S., & Nishina, D. (2006). A study on the characteristics of the preference for regional landscapes based on the free descriptive answer by subjects. *Journal of Environmental Engineering(Transaction of AIJ)*, (599), 135-142.
- 16) Oishi, H., Wa, S. M., & Nishina, D. (2007). An Analysis on the Physical Characteristics for Preferable Landscapes Based on The Photographs Taken by the Subjects. *Journal of Environmental Engineering(Transaction of AIJ)*, (611), 75-82.
- 17) Satoshi, Y., & Kotaro, O. (2019). Development and verification of the impression deduction model for city landscape with deep learning. *J. Archit. Plan.(Trans. AIJ)*, 84, 1323-1331.
- 18) Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel & Tourism Marketing*, 32(5), 608-621.
- 19) Sun, S., & Yu, Y. (2021). Dimension and formation of placeness of commercial public space in city center: A case study of Deji Plaza in Nanjing. *Frontiers of Architectural Research*, 10(2), 229-239.

- 20) Wang, R., & Zhao, J. (2017). Demographic groups' differences in visual preference for vegetated landscapes in urban green space. *Sustainable cities and society*, 28, 350-357.
- 21) Wikipedia contributors. (2021, October 11). Manhattan. Wikipedia. <https://en.wikipedia.org/wiki/Manhattan>
- 22) Wikipedia contributors. (2021a, October 10). Meituan. Wikipedia. <https://en.wikipedia.org/wiki/Meituan>
- 23) Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881-2890).

Chapter 5

Simulation of Urban Hotspots

Chapter 5

Simulation of Urban Hotspots

5.1 Significance of urban hotspot simulation

As concluded several times in previous chapters, urban hotspots at different locations with different characteristics are demonstrated to be key factors in the urban system (Krehl and Siedentop 2019). If the urban hotspots could be predicted, the future development trend of urban sub-systems, especially transportation, ecological, real estate and etc., will be foreseen, which is of great importance for urban planning. However, the accuracy of urban hotspot simulation chiefly depends on the data API and data quality of the SNS platform. The users of mainstream SNS platforms in the world are not evenly distributed, and the difficulty of API data collection of SNS platforms is also limited by regional regulations and laws. For a specific region, the inability to obtain the check-in data matching the quality and quantity of urban hotspot identification is the main obstacle to regional urban hotspot identification (Chen et al. 2014; El-Ashmawy 2016).

In the previous chapters, this study has conducted in-depth research on the emergence, change pattern, and characteristics of urban hotspots, and preliminarily mastered the objective property of urban hotspots, the key factor of the city. The results show that there is a significant correlation between the temporal and spatial distribution and attribute changes of urban hotspots and the built-up environmental elements of urban space. This inspired us to collect data related to the built environment to predict the temporal and spatial characteristics and attributes of urban hotspots. Compared with the check-in and online-review comment data of the SNS platform, the data of the digital map platform is easier to obtain, and the coverage of the dataset is wider.

Taking OpenStreetMap as an example, the open-source map platform provides users with vector maps of roads, buildings, green space, water, and other elements in urban space free of charge. The spatial-temporal statistical analysis of OSM’s full history since 2008 showed that humanitarian mapping efforts added 60.5 million buildings and 4.5 million roads to the map (Herfort et al. 2021). Overall, by acquiring these open vector data, this study could sort out the built environment information covering almost any regions of the world.

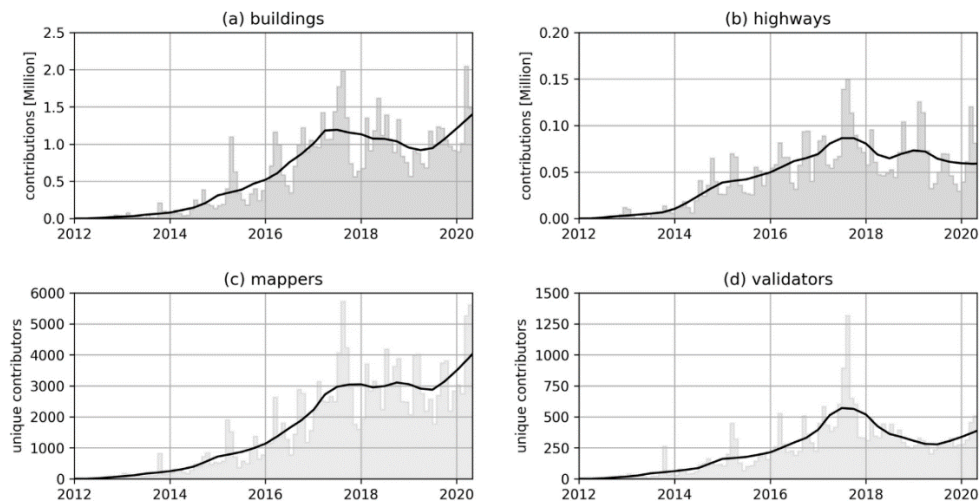


Figure 5-1 The evolution of OpenStreetMap community

Source: Herfort et al. (2021)

In this chapter, this study will introduce how to predict the temporal and spatial location and specific attributes of urban hotspots. The main purpose of this chapter is not only to verify the feasibility of simulation but also to compare the performance of different data and algorithms in predicting urban hotspots.

5.2 Methodology on urban hotspot simulation

5.2.1 Urban spatial simulation in previous works

In the definition of this study, urban hotspots are essentially a kind of special urban space. They play the role of urban sub-center and people gathering area.

Therefore, the simulation of urban hotspots is also a kind of urban spatial simulation. Therefore, before determining the method of predicting urban hotspots, this study first made a literature review on the relevant contents of urban space simulation in previous studies.

(1) Urban hotspot simulation based on regression models

Regression analysis is a method of predictive modeling technology, which studies the relationship between the dependent variable (target) and independent variable (predictor). This technique is used in forecasting, time series modeling, and finding causal relationships between variables. In the study of urban spatial analysis (Chapter 3), it has been found that the characteristics of urban hotspots are related to the built environment elements of the cell and its surrounding area. Therefore, based on the relationship between urban spatial attributes and built environment elements, many researchers have constructed regression models to predict urban land use and population distribution pattern. By fitting the functional relationship between the two, the researchers realize the simulation of urban space (Galea et al. 2005).

Table 5-1 Factors related to urban built environment

Built Environment factors	Items
Road network	<i>Width, length, and density of national highway, main roads and sub roads</i>
Public service	<i>Stations, medical facilities, primary and secondary schools, police stations, government town halls, commercial facilities</i>
Land use	<i>Commercial land, open space, agricultural land, forest and water area</i>
Land functional zones	<i>Urban block planning (Residential area, commercial area, industrial area, etc.), Urban spatial scheme, Etc.</i>
Physical and geographical conditions	<i>Elevation, vegetation, water body, temperature, precipitation, climate zone, etc.</i>

Others	<i>Population, land price, crime rate, traffic accident rate and natural disaster rate</i>
---------------	--

Table 5-1 shows the urban factors as dependent and independent variables predicted by the regression model in previous studies. Using linear regression, logistic regression, polynomial regression, and other regression models, researchers have tried to use urban built-up environmental factors to predict urban spatial characteristics and made a quantitative evaluation of the simulation accuracy. Generally speaking, the regression model still occupies the mainstream of urban space simulation research. However, the main purpose of urban simulation research based on the regression model is to explore the relationship between urban built-up environmental factors and spatial characteristics selected as independent variables and dependent variables (Mazumdar et al. 2018). Due to the complexity and uncertainty of the urban system, the simulation based on the regression model often does not have high precision.

(2) Urban hotspot simulation based on machine learning algorithms

As an interdisciplinary and artificial intelligence subject, machine learning has been gradually introduced into the field of urban space research and urban simulation. In machine learning, after "training", the system can use professional algorithms and a large number of databases to study, learn and make simulations and suggestions. The simulation model presented to new data can be adjusted without human intervention and learned from previous iterations to produce more reliable and repeatable decisions and results.

Table 5-2 Urban built environment factors used in regression models

Representative	Simulation object	Simulation algorithm
Liu et al. (2015)	Atmospheric pollutant	BP and RBF Neural Networks
Shang et al. (2017)	Traffic flow	Support vector regression
Wu et al. (2020)	Livability of urban spaces	Decision forest
Li and Hu (2019)	Theft risk	ML-nonlinear regression

Liu et al. (2021)	Land rent price	Regression tree and LASSO
Chen et al. (2021)	Urban thermal comfort	Deep forest

Table 5-2 presented some empirical cases of urban space simulation using machine learning methods. An important reason why advanced analysis based on machine learning is becoming more and more popular is that it can provide business advantages for almost every industry. Machine learning is useful as long as there are a large number of data and simulation models that need to be adjusted regularly (Intel 2021).

Urban problems are often complex, and the essence of urban problems can not be simulated and explained by conventional linear or nonlinear function models. Therefore, the accuracy and applicability of simulation based on traditional statistical models such as multiple linear regression (multiple regression) and vector autoregression in the field of urban research are being challenged by machine learning methods. This is because in principle, the statistical model is used to find a mathematical model that can minimize the mean square error of all data, and the purpose of machine learning is to obtain a model that can be predicted repeatedly. This is also the reason why the data needs to be divided into the training set and test set in the process of machine learning modeling. Usually, the performance of the machine learning model is judged by comparing its accuracy in the test set. In other words, in essence, machine learning methods are designed to make simulations, and statistical models are established to infer the relationship between variables.

With the development of machine learning technology, another progress in urban spatial simulation is the expansion of data sources. When using the traditional statistical model for spatial simulation, the built environment elements used as the original data (independent variables) are displayed in the form of numerical values. With the development of machine learning technology, another progress in urban spatial simulation is the expansion of data sources. When using the traditional statistical model for spatial simulation, the built environment elements used as the original data (independent variables) are displayed in the form of numerical values. In the field of machine learning, computer vision technology allows us to take pictures containing the elements of

the urban built environment as the original data. Because computer vision technology is a science that studies how to make machines "see". Further, it refers to the use of cameras and computers instead of human eyes for machine vision such as target recognition, tracking and measurement, and further image processing, which is more suitable for human eyes to observe or transmit to instruments for detection. Kitajima et al. (2019) pointed that the urban built environment dataset in the form of pictures has advantages of (1) more intuitive expression; (2) Contains the advantages of richer information.

After literature review, this study has found that urban spatial simulation methods, data, scales, and objects are different. In general, good simulation results have been achieved in all aspects. Then, specific to the special, real-time and variable urban core elements such as urban hotspots, how to select and process the original data for predicting urban hotspots and how to select the appropriate simulation algorithm (regression model or machine learning model) are important problems to be solved in this research. Specifically, this study will introduce our data collection and processing for urban hotspot spatial simulation, as well as the principles and implementation methods of various spatial simulation algorithms. In this chapter, my main research purpose is to realize the emergence and attribute simulation of urban hotspots by collecting data containing urban built environment information by using the correlation between built environment and urban hotspots.

5.2.2 Data preparation for urban hotspot simulation

The simulation method will be used to extract urban hotspots and master their attribute characteristics from urban space in the scenario where the SNS platform check-in data cannot be collected adequately. Therefore, the original data used for simulation including the characteristics of the urban built environment needs to have the characteristics of (1) "low acquisition difficulty" and (2) "wide data coverage". As a result, the OpenStreetMap platform, which has been mentioned in Chapters 3.4 and 5.1, is still introduced as the data source (Figure 3-12). This study collected vector map data from the open-source OpenStreetMap platform (including information of various built environmental elements such as roads, buildings, water bodies, and green spaces), and transformed them into images

and data tables respectively. Specifically, this study divides the complete research area into grids of the same size, and then counts the distribution density and existence of various built-up environmental elements in each grid.

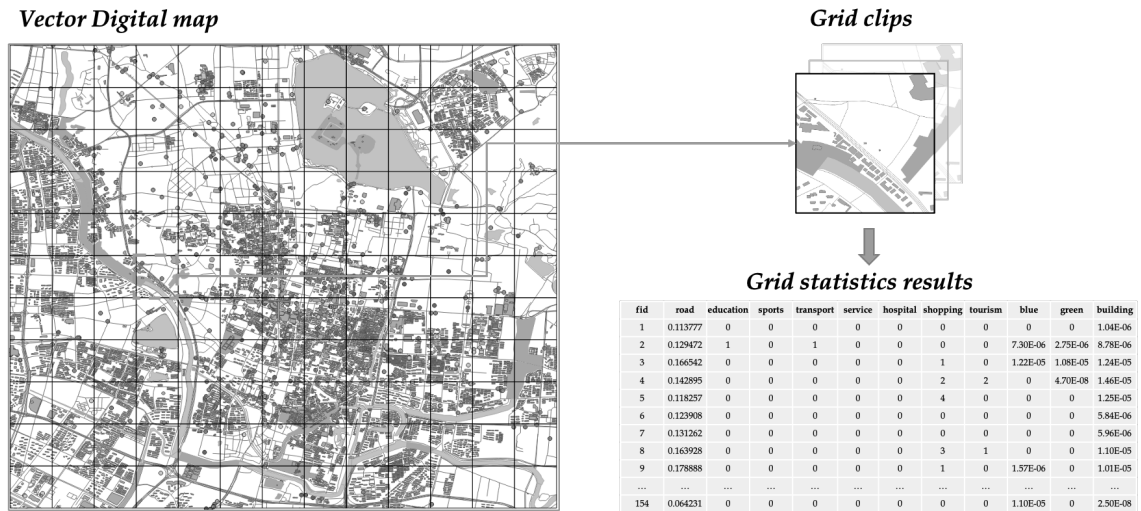


Figure 5-2 The vector digital map slices collection

Table 5-3 Numerical elements related to the urban built environment

Item	Content	Unit
MR	Total length of urban main-roads in the grid	Kilometer (km)
SR	Total length of urban sub-roads in the grid	Kilometer (km)
DC	Distance from grid center point to city center	Kilometer
PB	Number of bus stops in the grid	-
SS	Number of public service facilities related to social services in the grid	-
PS	Number of public service facilities related to product services in the grid	-
LS	Number of public service facilities related to living services in the grid	-
FH	Average floor height of buildings in the grid	meter
FA	Total area of building footprints in the grid	square meter
BA	Total building area (including all floors) in grid	square meter

WB	Total area of water bodies in the grid	square meter
GS	Total area of green spaces in the grid	square meter

As shown in Figure 5-2, various types of built environmental vector elements are set in different formats. The map contains the distribution of roads, various infrastructures, green space, and water bodies in the city. Next, this study cut the study area into grids. The statistical results of each grid and the built environmental elements on each will be used as the original data for urban hotspot simulation, the numerical elements related to the built environment that need to be counted in each grid are shown in Table 5-3.

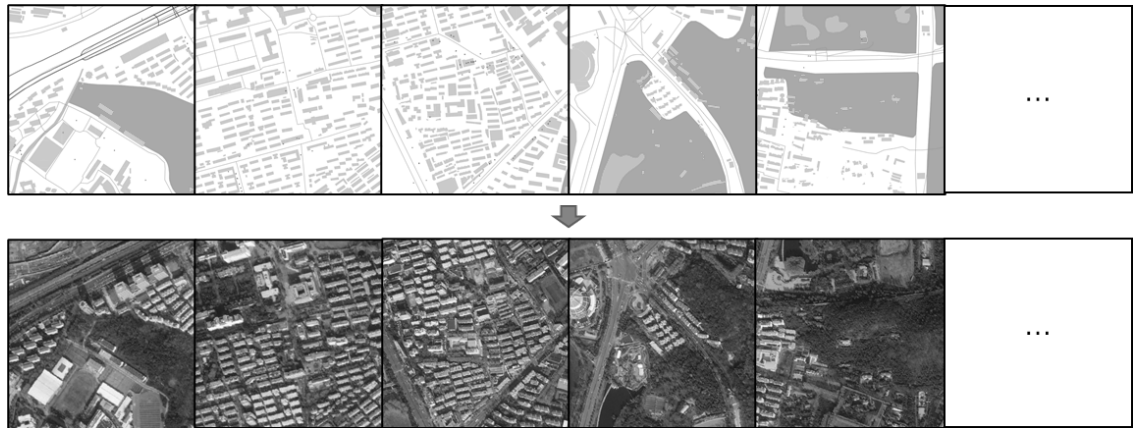


Figure 5-3 Sample vector and remote sensing image slices

In addition to vector map slices, this study also plans to collect remote sensing maps as a data source for urban hotspot recognition based on machine learning algorithms. The remote sensing data is collected from USGS EarthExplorer (<https://earthexplorer.usgs.gov>). The specific utilization approach of the processed data (numeric data, vector grid, and remote sensing slices) will be further introduced in the next chapter, and the sample vector and remote sensing image slices are presented in the upper and lower parts of Figure 5-3, respectively.

5.2.3 Algorithms for urban hotspots simulation

After sorting out the data needed to predict urban hotspots, this study will introduce how this study uses these data to predict the emergence and attributes of urban hotspots. Again, this study will not only realize the urban hotspot simulation based on the characteristics of the urban built environment but also summarize the performance of different data and algorithms in the simulation, especially the parameter selection, setting, and accuracy of related algorithms. Specifically, the simulation performance of regression models and emerging machine learning models widely used in previous studies will be compared under different scenarios, different simulation objectives, and different data sources.

(1) Multinomial logistic regression

On the one hand, the regression model will be combined with numerical data to predict the emergence and attributes of urban hotspots. In the regression model, the occurrence (occurrence or absence), and attributes (date, period, function) of urban hotspots in each grid will be set as the dependent variables of the regression model. It is worth mentioning here that after repeated consideration and combined with practical experience, it is believed that the absolute popularity and subjective landscape evaluation of urban hotspots are not able to be accurately predicted, so they are not included in the dependent variables of the regression model. The absolute popularity of urban hotspots reflects the frequency of visits by residents and tourists, which is greatly affected by the basic situation of regional population. The landscape characteristics of urban hotspots are largely determined by subjective ideas, regional urban style and design characteristics.

The urban built-up environmental elements (location, transportation, public facilities, development intensity, greening, etc.) contained in Table 5-3 will be set as the independent variables of the regression models. According to the attributes of dependent variables to be predicted, logistic regression model will be selected as regression models for predicting urban hotspots. It is worth mentioning that in Chapter 3.4, this study used a variant of logistic regression model - binary logistic regression model to explore the relationship between the change law of urban hotspots and urban built environment. In the case at that

time, this study constructed six binary logistic regression models to estimate the relationship between the occurrence of each law (represented by 0 and 1 respectively) and the built environment elements. Here, when constructing a simulation, many dependent variables in the results to be predicted are not presented in the form of binary variables. For example, the function (Entertainment, tourism and daily life) and occurrence time (weekday, off day, morning, afternoon and night) of urban hotspots are multi-classification variables. In order to deal with this problem, softmax regression model will be further introduced to undertake the multi-classification task.

Softmax regression is the general form of logistic regression. Logistic regression is used for two- dimension classification, while softmax regression is used for multi-classification. For input dataset $\{(x_1, y_1), (x_1, y_1), \dots, (x_1, y_1)\}$ with k categories, i.e. $y_i \in \{1, 2, \dots, k\}$, the a softmax regression is mainly used to estimate the probability that x_i belongs to each classification:

$$h_{\theta}(x_i) = \begin{bmatrix} p(y_i = 1|x_i; \theta) \\ p(y_i = 2|x_i; \theta) \\ \vdots \\ p(y_i = k|x_i; \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^k e^{\theta_j^T x_i}} \begin{bmatrix} e^{\theta_1^T x_i} \\ e^{\theta_2^T x_i} \\ \vdots \\ e^{\theta_k^T x_i} \end{bmatrix} \quad 5-1$$

Where, $\theta_1, \theta_2, \dots, \theta_k \in \theta$ denote the parameters to be estimated in the model, the function of $1/\sum_{j=1}^k e^{\theta_j^T x_i}$ Is to make the estimation result of probability distribution in $[0,1]$ and the sum of which is equal to 1. Using the softmax regression model (Stanford University 2021), it is estimated that the probability that x_i belongs to category j is:

$$p(y_i = j|x_i; \theta) = \frac{e^{\theta_j^T x_i}}{\sum_{l=1}^k e^{\theta_l^T x_i}} \quad 5-2$$

Finally, the highest value of the probability of belonging to X and j categories will be set as the specific category to which the simulation sample belongs.

(2) Random decision forest

In addition to statistical models, machine learning algorithms will also be introduced into the simulation of urban hotspots. Among them, for numeric data,

except for estimating the parameters of regression model and realizing the simulation with traditional econometric methods, there are also machine learning classifier models suitable for numeric data scenarios. Among them, the Random decision forest (or random forest) algorithm is mainly designed to figure the multi-class and multi-label classification problems. Random forest model operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

Generally, there are four reasons for me to choose random forest as the machine learning algorithm to predict urban hotspots with numeric data. First, the random forest algorithm will not have the common over fitting problems in statistical model and regression analysis. It can process high-dimensional (many features) data without dimensionality reduction and feature selection. Secondly, it can judge the importance and interaction of features by comparing the differences of simulation results under different spanning tree construction and pruning. Third, for unbalanced datasets, it can balance errors. If a large part of the features is lost, the accuracy can still be maintained. Fourth, in a random forest, this study doesn't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree. Specifically, in the input dataset of random forest model in this study, the occurrence, function and time of urban hotspots will be set as y respectively. In the regression model, the elements representing the regional built-up environment variables as dependent variables will be set as x (R, S. E. 2021).

(3) CNN (Convolutional neural network) image classifier

The above two methods are used to predict by counting the values of urban built-up environmental elements and the occurrence and attributes of urban hotspots in each delimited grid. So, can the "picture data" containing more information also predict urban hotspots? In order to verify this problem, this study utilizes the digital map vector slices and remote sensing map grid slices prepared in Chapter 5.2.2 and the occurrence and attribute features of urban hotspots at the grid corresponding to each slice to input into the convolutional neural network to construct the image classifier model. Specifically, the

construction principle of the classifier is similar to the picture preference extractor constructed in Chapter 4.3 (Figure 5-4).

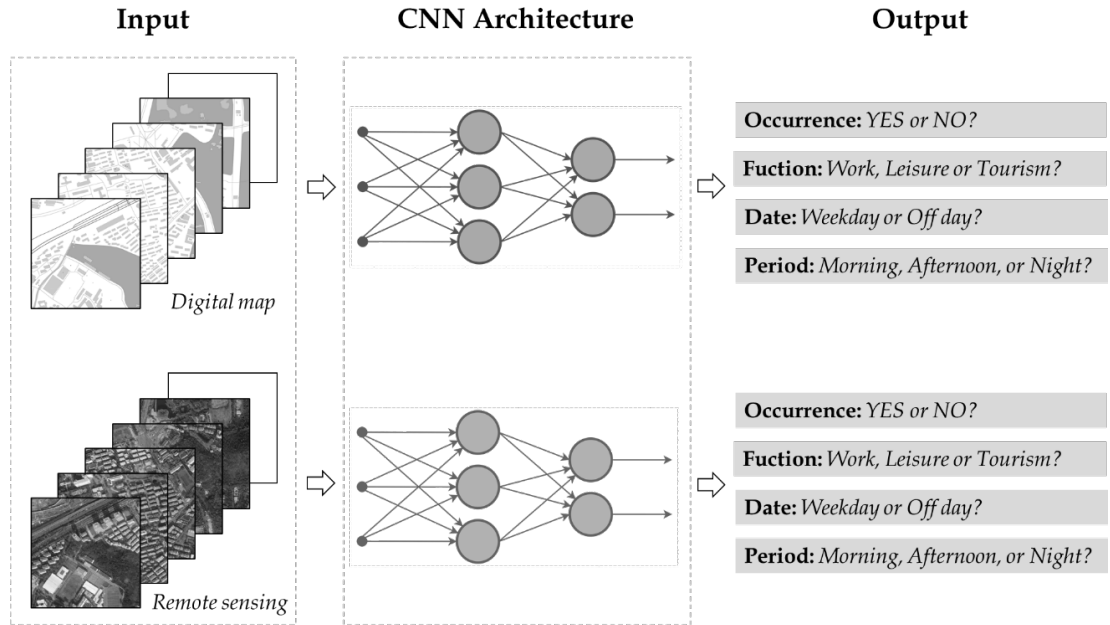


Figure 5-4 Roadmap of urban hotspot simulation model based on CNN

Specifically, in this study, four image classifier models based on convolutional neural network will be constructed with vector digital map slice and grid remote sensing image slice as input data respectively. The four classifier models will be used to predict the occurrence, function, occurrence date and occurrence period of urban hotspots respectively. Similar to chapter 4.3, the construction of image classifier model is based on Apple's CoreML framework using Xcode.

5.3 Urban hotspot simulation results

In this chapter, this study will present and compare the simulation results of urban hotspots based on different dataset and algorithms.

5.3.1 Simulation results from multinomial logistic regression

Table 5-4 Results of multinomial logistic regression

	Total	Date			Period			Function					
		Off-day	All time	12 ~ 18 p.m.	12 ~ 24 pm	All periods	Tourism	Leisure	Leisure & Tourism	Work	Work & Tourism	Work & Leisure	All functions
MR coef.	0.0971	0.1596	2.5687	0.6720	1.0490	6.1492	2.5493	(1.5338)	2.1331	0.1078	2.3745	(0.8185)	2.6800
MR std.	0.4237	1.1177	1.1219	1.2447	1.2559	1.2590	0.2890	0.2740	0.3093	0.3002	0.3166	0.2784	0.3161
SR coef.	1.7173	(2.8679)	1.2505	1.3754	2.7379	1.2919	(0.2314)	0.1542	4.5084	3.8798	(1.4998)	(0.0043)	0.1508
SR std.	1.7218	0.4383	0.4556	1.0891	1.0937	1.0973	0.3450	0.3465	0.3879	0.3494	0.3276	0.3450	0.3465
DC coef.	(0.3213)	3.4774	1.0562	3.8101	1.9168	2.7426	(0.8065)	(1.4091)	0.2666	0.7023	1.9494	6.3244	(1.0397)
DC std.	3.4054	3.4013	3.4354	1.6934	1.7277	1.7341	0.6095	0.5957	0.6121	0.6435	0.6322	0.6917	0.5990
PB coef.	4.5406	1.2546	(1.5080)	3.1027	0.1996	1.4257	1.9746	0.3938	0.5852	0.7737	(1.5339)	1.7088	4.0361
PB std.	0.1435	0.8966	0.8981	0.6733	0.6823	0.6912	0.8502	0.8540	0.8558	0.8592	0.8323	0.8724	0.8910
SS coef.	2.0090	1.3092	0.4276	1.1545	1.8199	(0.1091)	0.6521	0.0411	4.0511	0.9356	0.8012	(1.1985)	1.7832
SS std.	1.4037	2.4525	2.4666	1.9926	2.0173	2.0251	0.7446	0.7450	0.7832	0.7692	0.7540	0.7291	0.7627
PS coef.	(0.2909)	3.1333	2.1750	0.9679	(0.6799)	1.7457	1.4520	1.6311	0.2392	2.5938	0.3716	(1.9886)	0.7477
PS std.	1.6589	1.5676	1.5842	1.5152	1.5310	1.5359	0.4685	0.4845	0.4708	0.4842	0.4728	0.4427	0.4761
LS coef.	2.3249	2.9003	4.0094	2.9636	3.6761	(2.2219)	1.5838	(0.6562)	1.0505	0.9899	0.1375	2.6739	2.7712
LS std.	2.3495	2.5935	2.6170	1.9420	1.9682	1.9723	0.3919	0.3855	0.4019	0.4178	0.3935	0.4266	0.4199
FH coef.	2.8266	2.4684	1.1596	2.6690	4.6067	4.2865	0.3889	1.9093	0.7112	1.0632	1.2625	1.6286	2.5028
FH std.	1.2124	1.9839	1.9960	1.5430	1.5630	1.5893	2.5057	2.5244	2.5125	2.5256	2.5204	2.5269	2.5310
FA coef.	2.5016	3.0861	2.1160	0.0536	0.6157	1.6750	0.4264	(0.0410)	0.8444	(0.2506)	(2.2481)	0.9607	0.1540
FA std.	3.8814	3.7254	3.7642	2.2624	2.3000	2.3160	1.5193	1.5189	1.5274	1.5566	1.4932	1.5318	1.5209
BA coef.	0.7392	2.8261	1.2551	0.0143	(0.5350)	5.6935	1.2067	5.7715	0.9192	0.4785	3.5937	(0.4025)	4.5615
BA std.	1.3335	0.2701	0.2834	0.9701	0.9729	1.0000	2.5730	2.6296	2.5818	2.5757	2.6148	2.5678	2.6191
WB coef.	4.3736	1.4817	(1.3253)	2.2656	2.2085	2.3163	3.5706	3.2501	2.3344	1.2627	(1.7768)	(1.6470)	0.7222
WB std.	2.0012	3.1991	3.2191	0.5022	0.5343	0.5428	0.8048	0.8366	0.8270	0.8368	0.7841	0.7834	0.8121
CS coef.	2.1297	(2.4398)	0.5508	3.1139	1.0104	(0.0428)	(1.2924)	3.7144	1.1055	(0.0613)	0.4939	(0.7277)	(1.1022)
CS std.	2.6781	3.5824	3.6092	0.6741	0.7102	0.7204	0.9700	1.0065	0.9806	1.0059	0.9758	0.9606	0.9589

Firstly, this study presents the results of using regression model to predict urban hotspots step by step. This study constructed four logistic regression models to predict the emergence and attribute characteristics of cities. Table 5-4 shows the output results of multinomial logistic regression. The multinomial logistic regression model is estimated with R language with the package of 'nnet'.

Specifically, each column in Table 5-4 represents the possibility difference between the occurrence and nonoccurrence of urban hotspots in a specific scenario. The 2nd column represents the relationship between the overall existence of urban hotspots and urban hotspots. Columns 3 ~ 4 represent the relationship between the date of urban hotspots (rest day or working day?) and the built environment of the city. Columns 5 ~ 7 represent the relationship between the time period of urban hotspots (morning, noon or evening? Or across multiple periods?) and the urban built environment. Columns 8 ~ 14 represent the relationship between the functions of urban hotspots (work, leisure or tourism? Or two or more functions at the same time?) and the urban built environment. Each line of coef. represents the impact of the change of built environment attribute value of a class of cities on the possibility of urban hotspots. Each row of std. represents the corresponding standard deviation. In order to better illustrate the interpretation method of each number in table 5-4, this study takes the two estimation results corresponding to variable of MR (i.e., the length of main road) in the third column "off day" as an example. First, the numbers in this column are from a multiple logistic regression model used to estimate the date of urban hotspots. The number 0.1596 in the first row indicates that for each unit (1km) increase in the length of the main road in the grid, the probability of urban hotspots in the grid appearing on the rest day will be increased by 15.96% compared with that not appearing., and the standard deviation of the simulation result at all grids is averagely 1.1177. In addition, it is worth mentioning that any urban hotspot simulation model constructed in this chapter can predict the occurrence probability of urban hotspots at the corresponding grid under different scenarios.

For example, the simulation model of the occurrence date of urban hotspots corresponding to the third to fourth columns of table 5-4 listed in this study can output the probability of urban hotspots at the specified grid on rest days,

working days or not, and the sum of these three probability figures must be 100%. Among them, the largest possibility simulation value will be set as the simulation result. The simulation results of urban hotspots based on multinomial logistic regression are shown in Figure 5-5.



Figure 5-5 Urban hotspot simulation results from multinomial logistic regression

5.3.2 Simulation results from random forests

Next, this study continues to use this set of digital data to build four random forest models to predict the occurrence, date, period and functional attributes of urban hotspots. It is worth mentioning that the two main factors affecting the random forest model are the number of variables selected by the node branches of the decision tree and the number of decision trees in the random forest model. Therefore, it is necessary to further determine the optimal parameter value when building the model. The method of adding variables one by one is used to determine the number of variables selected by the node branches of the decision tree. Finally, this study constructed four stochastic forest classification models with the best performance, as shown in the figure below.

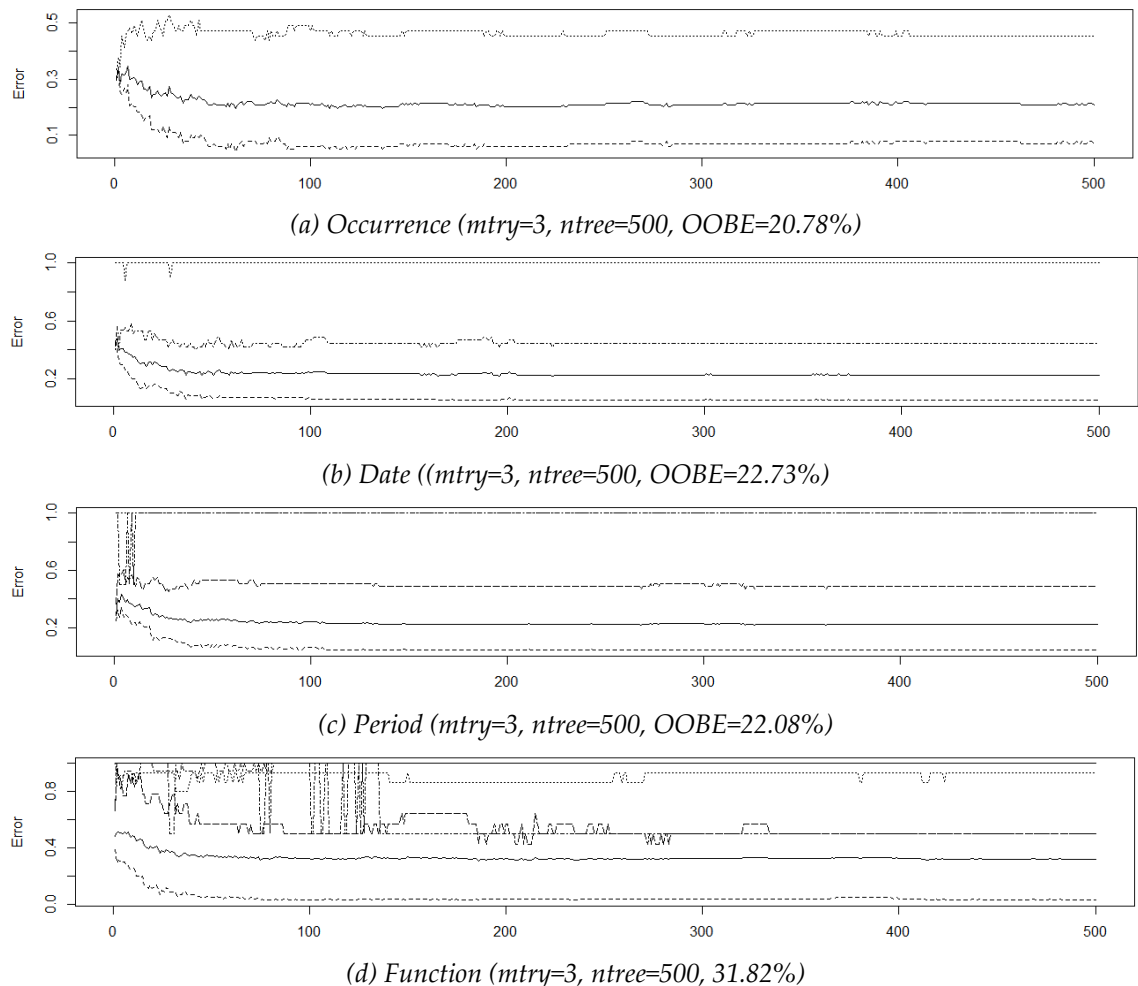


Figure 5-6 Parameter optimization of random forest models

The tool for constructing random forest model in this study is the random forest software package of R language. Figure 5-6 shows the parameter optimization results of the random forest model. It can be seen from the above figure that when the number of variables selected by the node branches of the decision tree is 3, the average misjudgment rate of the model is the lowest. When the number of decision trees is greater than 400, the model error tends to be stable, so it is finally determined that the number of decision trees in the model is about 500 to achieve the optimal model. Finally, the out of bag error (OOBE) of the four random forest models are 20.78, 22.73, 22.08 and 31.82 respectively.

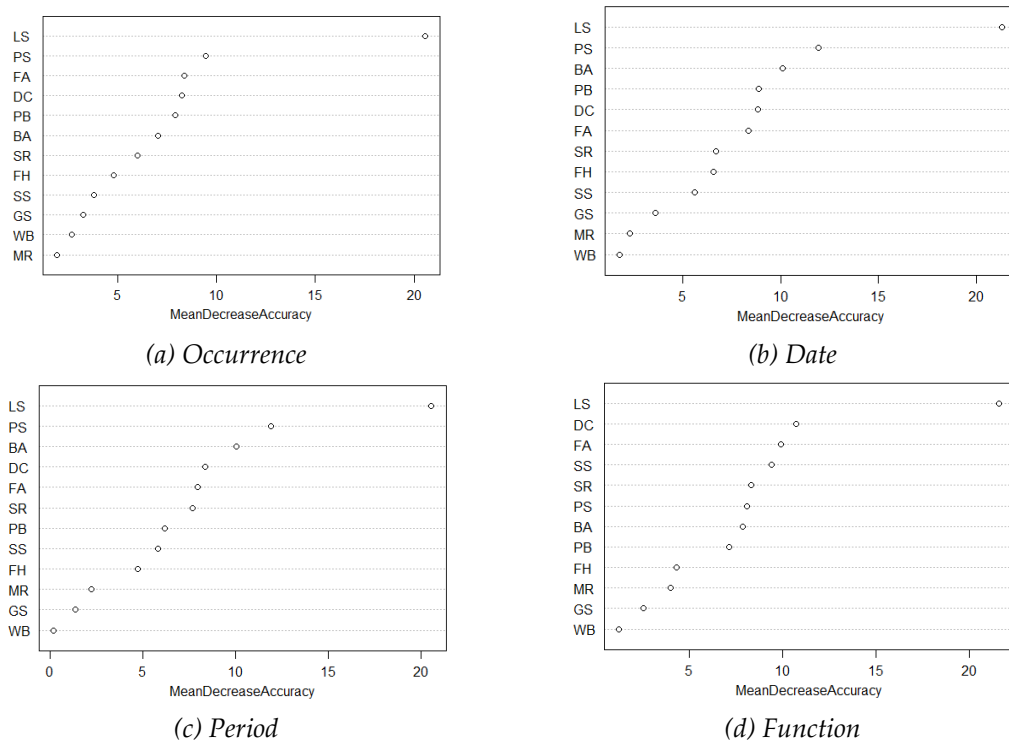


Figure 5-7 Variable importance estimation from random forest models

In addition, the random forest model additionally outputs the impact of various urban construction link elements on the accuracy of urban hotspot simulation (as shown in Figure 5-7). From the results, the number of announcement infrastructure related to life in the grid is very important for the emergence, date, period and function simulation of urban hotspots, which can increase the OOBE of the four models by 20% on average. The building bottom

area, total building area, distance from the city center and other factors can also have a significant impact on the accuracy of each model. Finally, the simulation results of urban hotspots using random forest are as follows (Figure 5-8):



Figure 5-8 Urban hotspot simulation results from random forest

5.3.3 Simulation results from convolutional neural network

Finally, using the two types of image data in Figure 5-3 and the method in Figure 5-4, this study builds four image classifiers based on convolutional neural network for each type of data. The specific method of building the model is the same as that of building the urban hotspot landscape evaluation and simulation model in Chapter 4.3.2, so it will not be described in detail.

In addition, the accuracy of vector images and remote sensing images that can be obtained in this study is limited. After image slices are divided into different groups of training datasets, there is still a lack of training data. To this end, this study has used the data augmentation tools described in Figure 4-17 to optimize the volume of the original data. Finally, the production and simulation results of CNN image classifier model are as follows:

(1) Vector digital map slices

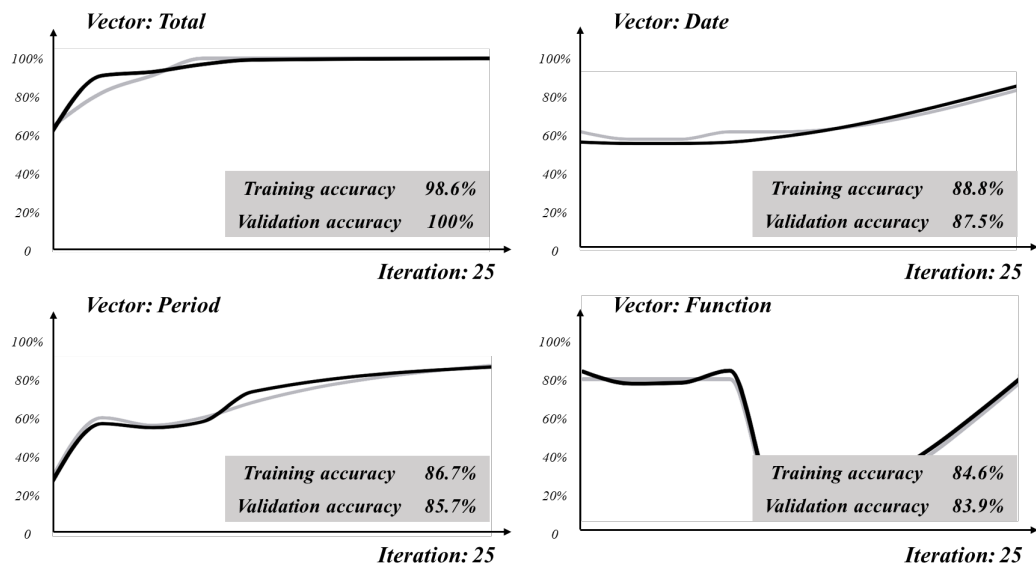


Figure 5-9 CNN image classifier based on vector digital map slices

This study trained four image classification models based on convolutional neural network. As shown in the Figure 5-9, the average accuracy of the four urban hotspot simulation models based on vector digital map slices established in this study in 25 iterations of training dataset and test dataset is more than 89.7% and 89.3%.

Therefore, these four models can accurately predict the appearance and attributes of urban hotspots under different circumstances. The simulation results from the four above models are presented as follows (Figure 5-10):

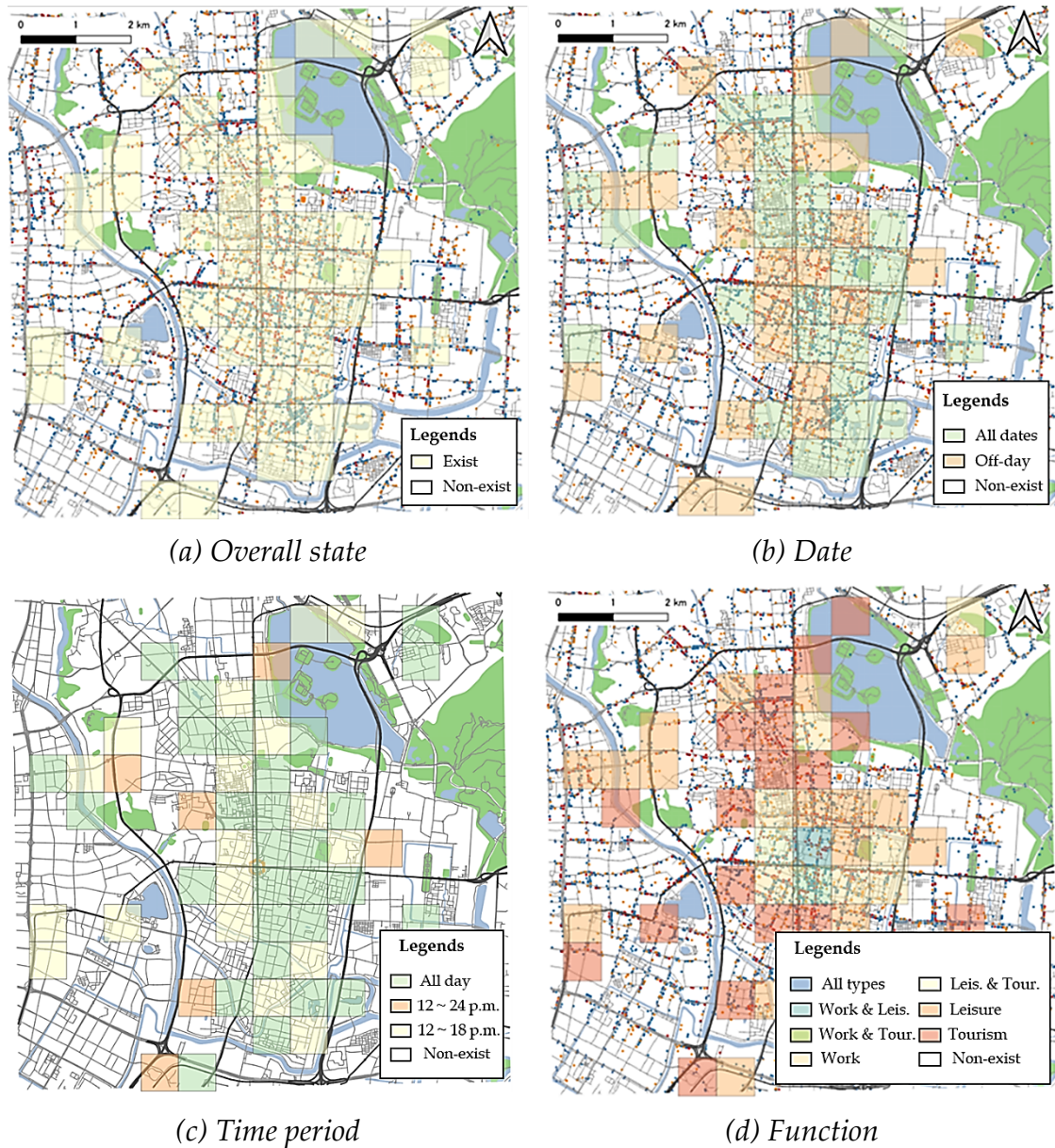


Figure 5-10 Urban hotspot simulation results from vector CNN

(2) Raster remote sensing slices

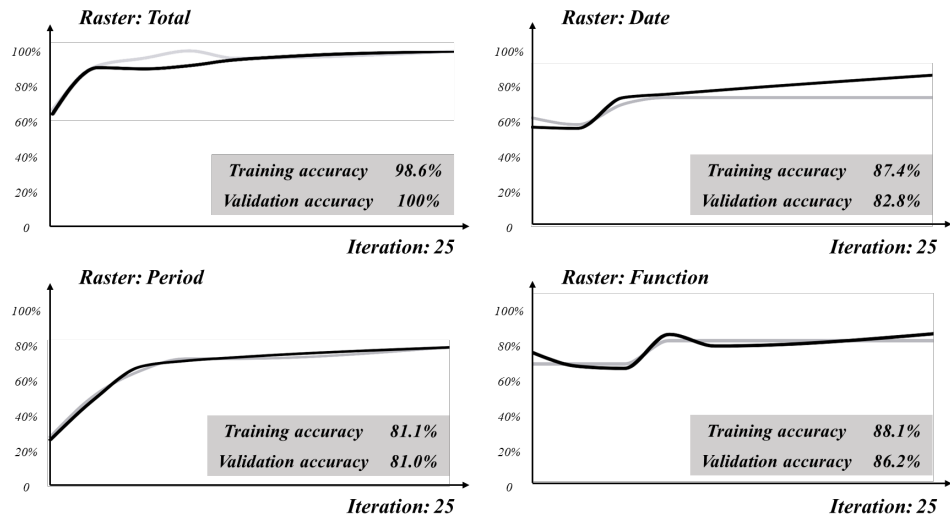
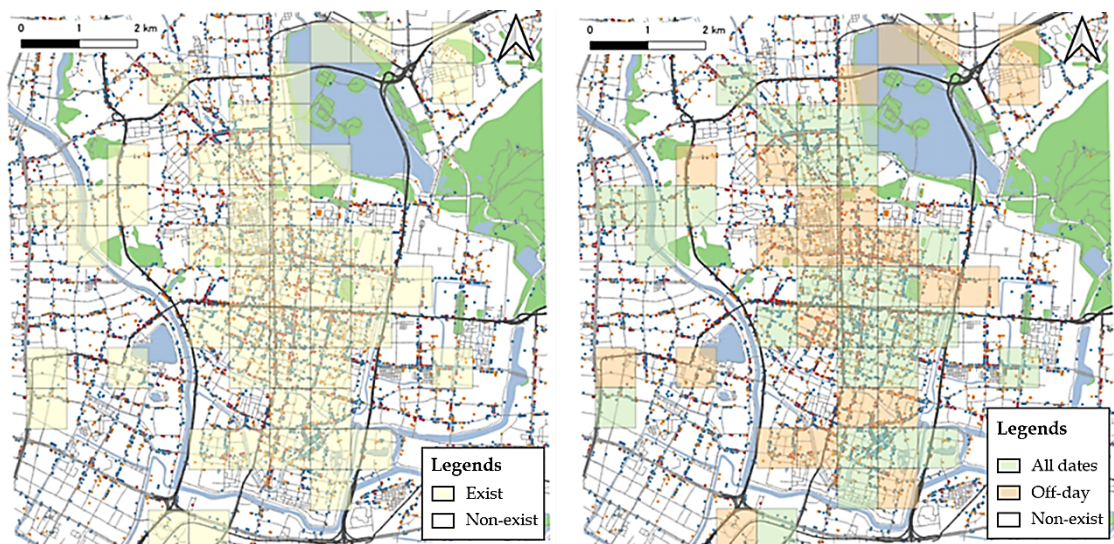


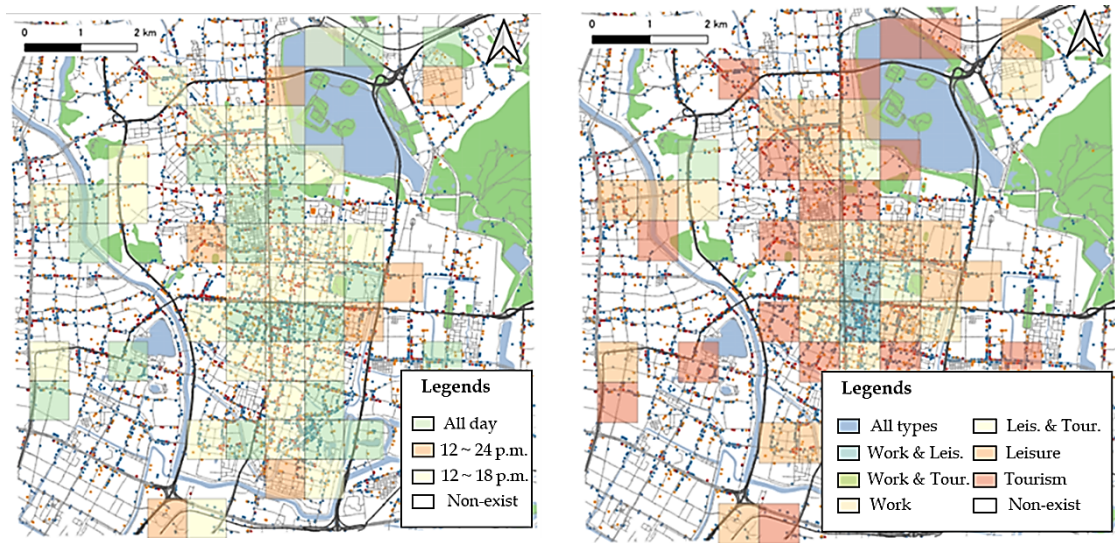
Figure 5-11 CNN image classifier based on raster remote sensing slices

Similarly, the average accuracy of the four urban hotspot simulation models based on raster remote sensing slices established in this study in 25 iterations of training dataset and test dataset is more than 88.8% and 87.5% (Figure 5-11). Therefore, these four models can also accurately predict the appearance and attributes of urban hotspots under different circumstances. The simulation results from the four above models are presented as follows (Figure 5-12):



(a) Overall state

(b) Date



(c) Time period

(d) Function

Figure 5-12 Urban hotspot simulation results from raster CNN

5.4 Comparison of simulation accuracy of urban hotspots

In this chapter, this study will horizontally compare the simulation results of four urban hotspot simulation schemes proposed in this study. Specifically, this study will count the simulation results of four models on the occurrence, date, time period and function of urban hotspots at each grid unit and calculate in detail the proportion of the correct grid in the total grid (record it as the simulation accuracy). In addition, it should be noted that there are 143 grids in the main research area of this study, of which 90 grids are identified as non-existent urban hotspots. The four urban hotspot identification schemes designed in this study can 100% predict the fact that there are no urban hotspots in these 90 grids. Therefore, whether different data and models can accurately predict the occurrence, date, period and function of urban hotspots in the remaining 53 grids under different scenarios determines the simulation accuracy of urban hotspots. As a result, this study also made additional statistics on the simulation performance of the urban hotspot simulation model in these 53 grids and recorded it as "relative accuracy".

Table 5-5 Prediction accuracy comparison

Accuracy	Multinomial logistic regression		Random forest		CNN classifier (vector)		CNN classifier (raster)		Average accuracy	
State										
Overall	138 / 143	96.5%	137 / 143	95.8%	141 / 143	98.6%	141 / 143	98.6%	557 / 572	97.3%
Relative	48 / 53	90.6%	47 / 53	88.7%	51 / 53	96.2%	51 / 53	96.2%	197 / 212	92.9%
Date										
Overall	122 / 143	85.3%	129 / 143	90.2%	127 / 143	88.8%	125 / 143	87.4%	503 / 572	87.9%
Relative	32 / 53	60.4%	39 / 53	73.6%	37 / 53	69.8%	35 / 53	66.0%	143 / 212	67.5%
Period										
Overall	107 / 143	74.8%	115 / 143	80.4%	124 / 143	86.7%	116 / 143	81.1%	462 / 572	80.8%
Relative	17 / 53	32.1%	25 / 53	47.2%	34 / 53	64.2%	26 / 53	49.1%	102 / 212	48.1%
Function										
Overall	121 / 143	84.6%	126 / 143	88.1%	121 / 143	84.6%	126 / 143	88.1%	494 / 572	86.4%
Relative	31 / 53	58.5%	36 / 53	67.9%	31 / 53	58.5%	36 / 53	67.9%	134 / 212	63.2%
General accuracy	60.4% / 85.3%		69.3% / 88.6%		72.2% / 89.7%		69.8% / 88.8%		67.9% / 88.1%	

Table 5-5 shows the comparison results of simulation accuracy of four urban hotspot simulation schemes under different scenarios. Firstly, on the whole, the overall simulation accuracy of urban hotspots of the four types of models is more than 85%. The accuracy of traditional multiple logistic regression model is relatively the lowest, only 85.3%. The simulation accuracy of convolution neural network image classifier based on vector map is the highest, with an average of 89.7%. The simulation accuracy of random forest model and urban hotspot simulation model based on grid remote sensing image is close, which are 88.6% and 88.8% respectively, with a difference of only 0.2%. Similarly, for 53 grids with urban hotspots, the performance of the four types of models is exactly the same as that in the overall mode. The order of simulation accuracy from high to low is CNN_Vector (72.2%)、CNN_Raster (69.8%), Random Forest (69.3%) and Multinomial Logistic Regression (60.4%). Next, let's take a look at the simulation performance of four types of urban hotspot simulation models in different scenarios.

Occurrence: In the task of predicting whether there will be urban hotspots in 143 grids in the study area, the four simulation methods have achieved very high simulation accuracy, and the simulation accuracy is above 95%. In contrast, the performance of the two convolutional neural network models based on image data is better in predicting the occurrence of urban hotspots, reaching 98.6%.

Date: The overall effect of the simulation of the occurrence date of urban hotspots is lower than that of the simulation of the occurrence status (occurrence or non-occurrence) of urban hotspots. The global (143 grid) simulation accuracy of the four simulation methods is distributed between 85% ~ 90%. For 53 grids with urban hotspots, the accuracy of the four algorithms is above 60%. For 53 grids with urban hotspots, the accuracy of the four algorithms is above 60%. Among the four methods, the stochastic forest model has the best performance in predicting the occurrence date of urban hotspots, and the overall and relative accuracy are the highest among the four methods, reaching 90.2% and 73.6% respectively. The simulation accuracy of traditional regression model is the lowest, and the overall and relative accuracy are only 85.3% and 60.4% respectively.

Period: The simulation of the emergence period of urban hotspots seems to be the most difficult. The overall simulation accuracy is only distributed around 75% ~ 80%. In 53 grids with urban hotspots, the simulation accuracy of the other three models is only less than 50%, except that the convolution neural network model based on vector graph slice has achieved 64.2%. The traditional multiple logistic regression model only showed a very low accuracy of 32.1%.

Function: The overall accuracy of the functional simulation of urban hotspots is better than that of time period simulation, and the effect is close to that of date simulation. The overall accuracy is also distributed between 85% ~ 90%. The difference is that the simulation accuracy of random forest model and convolution neural network model based on remote sensing map is the highest, and the overall and relative accuracy are 88.1% and 67.9% respectively. The vector CNN model, which performed best in urban hotspot function simulation before, is not suitable for predicting urban hotspot function. The overall and relative accuracy of the corresponding simulation results are at the bottom, only 84.6% and 58.5%.

5.5 Summary on the urban hotspot simulation

In general, in this chapter, by summarizing and combining the data sources and algorithms commonly used in urban spatial simulation, this study tries to predict the urban hotspot, the core element of the city. Specifically, this study predicts the occurrence status, date, period, and function of urban hotspots. The results of urban hotspot simulation are as follows:

The data sources that are suitable for urban hotspot simulation could be numeric data, vector maps, and raster remote sensing images. This study obtained the vector digital map and remote sensing image map of the study area from OpenStreetMap and USGS explore respectively. This study cuts the case study area of this study, the central area of Nanjing, into equal size grids according to the method in Chapter 3.4.

The algorithms that are suitable for urban hotspot simulation includes multiple logistic regression model and machine learning random forest classification models. Considering that the urban hotspot attributes to be predicted in this study can be transformed into classified data, the urban hotspot simulation in this study can be transformed into a classification problem of urban spatial scenes. To be specific, for the collected data in numeric form, this study utilizes the correlation between urban built environment elements and urban hotspots to build statistical multiple logistic regression model and machine learning random forest classification models to predict them. For the collected image data, this study built two convolution neural network models for vector and grid images respectively.

In general, for the simulation of different detail attributes, the combination of different data and algorithms shows their expertise. Taking pictures as raw data can achieve better accuracy than digital data. The machine learning algorithm also has better performance than the traditional statistical model in the simulation of urban hotspots. Moreover, the combination of different data sources and algorithms can show their expertise in specific simulation scenarios. The random forest model is more suitable for predicting the date of urban hotspots, while the convolution neural network based on vector and grid map slices shows better performance in predicting the time period and function of urban hotspots respectively. The four methods can predict whether there will be urban hotspots with specific functions on a specific date and at a specific time period in an urban area with an accuracy of more than 90%. On the contrary, it is difficult for existing data and methods to accurately predict when urban hotspots appear.

References

- 1) Krehl, A., & Siedentop, S. (2019). Towards a typology of urban centers and subcenters—evidence from German city regions. *Urban Geography*, 40(1), 58-82.
- 2) El-Ashmawy, K. L. (2016). Testing the positional accuracy of OpenStreetMap data for mapping applications. *Geodesy and Cartography*, 42(1), 25-30.
- 3) Chen, M., Mao, S., Zhang, Y., & Leung, V. C. (2014). Big data: related technologies, challenges and future prospects.
- 4) Herfort, B., Lautenbach, S., de Albuquerque, J. P., Anderson, J., & Zipf, A. (2021). The evolution of humanitarian mapping within the OpenStreetMap community. *Scientific reports*, 11(1), 1-15.
- 5) Myers, R. H. (1990). *Classical and modern regression with applications* (Vol. 2, p. 488). Belmont, CA: Duxbury press.
- 6) Galea, S., Ahern, J., Rudenstine, S., Wallace, Z., & Vlahov, D. (2005). Urban built environment and depression: a multilevel analysis. *Journal of Epidemiology & Community Health*, 59(10), 822-827.
- 7) Mazumdar, S., Learnihan, V., Cochrane, T., & Davey, R. (2018). The built environment and social capital: A systematic review. *Environment and Behavior*, 50(2), 119-158.
- 8) Ewing, R., & Dumbaugh, E. (2009). The built environment and traffic safety: a review of empirical evidence. *Journal of Planning Literature*, 23(4), 347-367.
- 9) Liu, J. (2015). *Temporal-spatial Variation as well as Evaluation and Prediction Models of Air pollutants in Beijing*. University of Science and Technology Beijing, Beijing.
- 10) Shang, Q. (2017). *Research on Methods for Traffic State Identification and Prediction Based on Machine Learning*. Jilin University, Jilin.

- 11) Wu, J. (2020). The Fourth Paradigm: A Research for the Predictive Model of Livability Based on Machine Learning for Smart City in The Netherlands. *Landscape Architecture* 27(5), 11-29.
- 12) Li, Y. (2020). Research on risk analysis of urban theft crime based on multi-source data. People's Public Security University of China, Beijing.
- 13) Liu, X., Huang, J., Zhao, G. (2021). A Machine Learning Approach for Community-scale Commercial Rents Mapping. *China Land Science* 2021(3), 49-57.
- 14) Chen. (2020). Prediction of indoor thermal comfort level of high-speed railway station based on deep forest. *Journal of Computer Applications*, 41(1), 258-264.
- 15) Intel. (2021.). The Future of Machine Learning, Data and Predictive Analytics. Retrieved October 14, 2021, from <https://www.intel.com/content/www/us/en/analytics/machine-learning/machine-learning-data-and-predictive-analytics.html>
- 16) Kitajima. S., Nobuhiro, R., Hidetora, T., Akinori M. (2019). A Study on the Prediction Model for the Development of Vacant Homes Using Deep Learning, *Journal of the City Planning Institute of Japan*, 54(3), 1468-1474,
- 17) Stanford University. (2021). Unsupervised Feature Learning and Deep Learning Tutorial. UFLDL Tutorial. <http://deeplearning.stanford.edu/tutorial/supervised/SoftmaxRegression/>
- 18) R, S. E. (2021, June 24). Random Forest | Introduction to Random Forest Algorithm. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>

Chapter 6
Conclusions

Chapter 6

Conclusions

6.1 Main findings

As the product of the accumulation of various spatial elements in urban space, urban hotspots have always played important roles as urban sub-centers with different functions and representatives of urban external image. On the one hand, the spatial-temporal distribution of urban hotspots will affect the chain reaction of urban transportation, economy, and environmental system. On the other hand, the landscape characteristics of urban hotspots determine the external image of the city and even the whole region to a certain extent. However, the previous research on the identification method and spatial-temporal patterns of urban hotspots is limited by the lack of information of the original data, and there is still room for improvement, while the relevant content of landscape feature extraction of urban hotspots is rarely mentioned.

On this basis, this study takes urban hotspots as the research objects, propose a new urban hotspot identification method with the support of new data sources (SNS dataset) and new methods (machine learning algorithms), and carry out in-depth research on its spatial-temporal distribution characteristics and landscape characteristics. In addition, considering the importance of urban hotspots to urban space, its emergence and attribute characteristics are predicted. Specifically, the main scientific findings of this study are as follows:

Chapter 2: With the leverage of KANN-DBSCAN, Concave hull algorithms and SNS data, this study introduces a new urban hotspot identification method. The features and advantages of the new method are obtained in this study are concluded as: (1) Benefiting from the time tags contained in the platform user check-in data provided by the SNS platform, the real-time extraction of urban hotspots by obtaining the check-in data records generated in different periods is

realized. (2) Benefiting from the activity-type label contained in the platform user check-in data provided by SNS platform, the main functions of urban hotspots in urban space are estimated by calculating the activity composition implemented by SNS platform users in each urban hotspot area. (3) With the help of parameter self-setting of KANN-DBSCAN algorithm, the automatic and reasonable extraction of the scope of urban hotspots is realized. This method is able to automatically adjust the values of key parameters of MinPts and Eps in clustering algorithm according to the size of dataset, so as to ensure the scientificity of urban hotspot size (because in traditional methods, these parameters need to be determined subjectively according to the experience of experts). (4) With the help of concave hull algorithm, it is found that the best method to delimit the boundary of urban hotspots should be concave hull algorithm. With the detected boundary, the area and popularity and function of urban hotspots could also be estimated by counting the scale, amount, and activity proportion of involving check-in points. (5) By comparing the urban hotspot identification results based on traditional data source (digital map POI data) and method (kernel density algorithm) with the experimental results of this study, it is found that the urban hotspot location identified by the new method is consistent with the traditional research and has high accuracy.

Chapter 3: This study continues to study the temporal and spatial pattern of the emergence and change of urban hotspots in the empirical research area and summarize them into 6 basic rules. (1) A part of urban hotspots with small areas and low popularity only appears in a certain period of time. They either only emerge on off days (Rule 2) or vanish and appear intermittently in a day (Rule 6). Other parts of them merge surrounding ones in the afternoon but split back to their original state at night (Rule 4). (2) Most urban hotspots with median area and popularity didn't experience morphological change (Rule 1), and the remaining small parts are either divided into their spilled states (Rule 4) or merged into large-area and high-popularity ones (Rule 5) accompany by function change from tourism-related to leisure-related. (3) It is worth noting that if the merger took place in the process from weekdays to off days, it formed urban hotspots with the largest area and highest popularity in the study period (Rule 3). Similarly, this process accompanies by function change from tourism-related

to leisure-related, as well. Moreover, through the construction of binary logistic regression models, the quantitative relationship between the occurrence of the above six rules and the urban built environment was combed out. It is found that whether there are urban hotspots in the surrounding areas, the building density and the distance to the city center can have a significant impact on the emergence and change of urban hotspots. Specifically, in the urban space with more urban hotspots, higher building density and closer distance to the city center, the probability of urban hotspots appearing according to the rule of 1-5 is higher, and the probability of appearing according to rule 6 is relatively lower.

Chapter 4: Considering the additional functions of urban hotspots as urban landmarks and landscape representatives, this study also investigates the landscape characteristics and street-view impression of urban hotspots. Generally, this study combines the online review data of social network users and machine learning algorithms to design an urban street-view impression and preference prediction model. This model can not only estimate the subjective cognitive results of urban residents on the urban street-view landscape but also make it possible to quantitatively compare the landscape preferences of residents with different demographic characteristics. In a case study targeting the street-view impression and landscape preference differences of local residents, domestic tourists, and foreign tourists on the urban hotspots' landscape, this study collected the text and picture comments generated in the core area of Nanjing Dianping.com and Google map platforms to construct urban scene preference models representing the cognition of three groups of people on the natural and cultural landscape and summarized their landscape evaluation and preference by comparing the outputs from different models. The results show that: (1) For natural landscapes, the landscape of urban hotspots with a large area of water and exquisite flower beds as the landscape theme can generally get a better subjective evaluation from visitors. On the contrary, urban hotspots with antique buildings and classical gardens as the core natural landscape have generally received low praise. In addition, the landscape evaluation of urban hotspots with rockery landscape as the main body is controversial. (2) For cultural landscapes, the landscape evaluation of urban hotspots with extreme height and large-scale modern-style buildings as the main body of cultural

landscape is better. The cultural landscape composed of daily and life flavor urban streets and antique commercial streets is not liked by urban residents, visitors, and tourists. (3) This study also constructed 6 models representing the cognitive preference on Nanjing's cultural and natural landscape of local residents, domestic tourists, and foreign tourists, respectively. By cross using the subjective preferences prediction models constructed for different attributes of people, cognition differences between local residents, domestic tourists, and foreign tourists are extracted. The main landscape contents, elements, and composition characteristics of urban scenes are closely related to the subjective preferences of three groups of people on urban landscapes.

Chapter 5: In this chapter, by summarizing and combing the data sources and algorithms used in urban spatial simulation, this study tries to simulate the urban hotspot, the core element of the city. Specifically, this study simulates the occurrence status, time, date, and function of urban hotspots. This study prepared numeric data, vector digital map data, and raster remote sensing image data to predict urban hotspots' occurrence and attributes. In general, for the simulation of different detail attributes, the combinations of different data and algorithms show their expertise. Taking images as raw data can achieve better accuracy than numeric data. The machine learning algorithms also have better performance than the traditional statistical models in the simulation of urban hotspots. Moreover, the combination of different data sources and algorithms can show their expertise in specific simulation scenarios. The random forest model is more accurate in predicting the date of urban hotspots, while the convolution neural network based on vector and raster map slices shows better performance in predicting the time period and function of urban hotspots respectively. The four methods can predict whether there will be urban hotspots with specific functions on a specific date and at a specific time period in an urban area with an accuracy of more than 90%. On the contrary, it is difficult for existing data and methods to accurately predict when urban hotspots appear.

6.2 Academic contributions

From an academic point of view, the main contributions of this study are as follows:

(1) A new dynamic identification method of urban hotspots

Benefiting from the explosive growth of SNS data and the continuous progress of computer vision algorithms, this study proposes an urban hotspot recognition method based on SNS check-in point data and concave hull algorithm. This method can not only accurately judge the location of urban hotspots in urban space, but also estimate the function, heat, occurrence date, and occurrence period of urban hotspots. This study takes researchers and urban planners one more step toward understanding the generation of urban hotspots

(2) Real-time emerging and changing rules of urban hotspots

Using the urban hotspot identification method proposed, this study realized the real-time extraction of urban hotspots. By comparing the spatial distribution and attribute characteristics of urban hotspots in different periods, this study summarizes the emergence and change patterns of urban hotspots into six laws. This provides a quantitative basis for us to understand the structure and function of urban space dynamically and in real-time.

(3) Extraction of landscape evaluation and preference

This study combines the online review data of social network users (images and texts) and machine learning algorithms (NLP and CNN) to design an urban scene evaluation and preference simulation model. Benefiting from the huge volume of SNS data in urban hotspots, this method can extract and estimate the evaluation and preference of visitors and tourists on the landscape of urban hotspots. This method provides method support for planning and designing the landscape of urban hotspots as urban landscape landmarks and external windows.

(4) Guidance for the simulation of urban hotspots

This study also compares the performance of different data sources (numeric data, digital map vector data, remote sensing map grid data) combined with different algorithms (traditional statistical model, random forest algorithm and

convolution neural network algorithm) in urban hotspot simulation. According to the expertise of different data and algorithm combinations, it provides guidance for urban hotspot simulation under different scenarios and objectives.

(5) Verification of application prospect of SNS data and Machine learning

In addition to an in-depth understanding of urban hotspots, the whole process of this study is also an empirical application case of SNS data and machine learning algorithms in urban spatial analysis. On the one hand, the results of spatial-temporal distribution characteristic of urban hotspots based on SNS check-in data prove that spatial data recording human activity intelligence contained in SNS data are helpful for urban researchers to dynamically understand urban spatial structure. On the other hand, the research results of landscape evaluation, preference extraction of urban hotspots based on CNN and NLP algorithms, and urban hotspot simulation based on random forest model respectively verify the good prospects of machine learning algorithms in urban landscape analysis and urban spatial simulation.

6.3 Future directions and recommendations

In the process of sorting out and implementing this research, some deficiencies were also found. Specifically, it is mainly reflected in the following aspects:

(1) Expanding the time span of check-in data acquisition

The emergence and change patterns of urban hotspots summarized in this study mainly reflect the changing rules of urban hotspots across off days and weekdays. Within one single day, only the analysis on the changes of urban hotspots across the morning (6-12 o'clock), afternoon (12-18 o'clock) and night (18-24 o'clock) were realized. In fact, if the temporal and spatial characteristics of urban hotspots in different seasons and different weather conditions can be obtained, as well as the hourly change characteristics of urban hotspots within 24 hours a day, this study might be able to further analyze and summarize the objective emerging and changing rules of urban hotspots in more detail. Unfortunately, the difficulty of obtaining SNS data is increasing. In another word,

the free and real-time access to check-in data from the SNS platform API has gradually been replaced by commercial purchases. At the time when the check-in data was obtained in this study, the API access frequency limit and the upper limit of data downloads ultimately determined the date time span (weekday and off day) and accuracy (each 6 hours) of the check-in data collected in this study. In the future, on the premise of reasonably and legally obtaining higher-precision SNS check-in data, the accuracy of urban hotspot emerging and changing rules will be hugely improved.

(2) Improving the accuracy of open-sourced digital and remote sensing maps

In this study, the grid unit of image data collection is set as 0.0075 degree (latitude and longitude) considering the basic accuracy of vector digital map data from OpenStreetMap platform. The grid unit is about to be 700 meters on the map. Although the OpenStreetMap platform has the largest amount of data and the highest data update frequency in the world, when setting the variable length of the basic unit of data collection to less than 0.0075 degrees, the number of urban elements contained in some grids will appear as 0. Correspondingly, in the subsequent simulation research of urban hotspots, this is only able to set the basic unit of numeric figure data collection of urban built environment and the size of remote sensing image slices to 0.0075 degrees, as well. If the accuracy of digital map data can be overcome, the conclusions of this research may be polished, and it will also become possible to compare the effect of urban hotspot simulation under different original data accuracy (grid unit).

(3) Strengthening the application of SNS data and machine learning algorithm in the field of urban planning

In this research, I have been engaged in interdisciplinary urban planning research. I mainly take an important element in urban space -urban hotspots, as the object, and try to use the check-in, text, and image online review data generated by the SNS platform to realize the identification, feature analysis, and spatial prediction of urban hotspots with the help of machine learning. This is a comprehensive study integrating the fields of information, computer science, and architecture. In the process of implementing this research, I learned how to obtain check-in, text, and image data from the developer API of the SNS platform,

and also tried to construct computer vision and natural language processing models. These new data and algorithms help me design new methods for urban hotspot identification, feature analysis and prediction, and realize the in-depth spatial interpretation. It is believed that I need to strengthen my thinking on its practical significance of introducing them into the research of urban planning. After the knowledge of these two disciplines is intertwined, how to better improve and serve the future urban planning is the key problem I need to solve in the future.