Ph.D. programme in Urban and Architectural Management and Valuation

Ph.D. Thesis

Exploring the applications and limitations of location-based social network data in urban spatiotemporal analysis

Centre for Land Policy and Valuations (CPSV)
Department of Architectural Technology (TA)
Barcelona School of Architecture (ETSAB)
Universitat Politècnica de Catalunya (UPC)

Ph.D. candidate: Liya Yang
Supervisors: Dr. Carlos Marmolejo Duarte
Dr. Pablo Martí Ciriquián

July 2021

Abstract

Nowadays, the widespread utilization of intelligent mobile and located-embedded services extends the border of social network sites (SNS) to physical-spatial space. The appearance of SNS changes our daily life meanwhile promotes the shift of the research paradigm of all academic disciplines. The data generated from SNS is named location-based social network(LBSN) data.

As to urban studies, LBSN data provides a promising opportunity to understand the quotidian life and cities, and thus it brings a profound impact on both urban theories and empirical studies, such as the urban spatiotemporal analysis and urban sentiment analysis. However, despite the opportunities that LBSN data provides, some challenges and limitations also associate with related researches of urban studies. Who are using LBSN applications? How many degrees could social media data represent the actual situation of the physical city? What is the relation between LBSN data and urban issues?

Therefore, taking advantage of previous researches, this dissertation seeks to explore the applications and limitations of location-based social media data for urban spatiotemporal analysis based on a comprehensive summation of current works and three empirical analyses based on different sources of LBSN data in different cities. The ultimate purpose of the dissertation is expected to gain new knowledge that could help future urban studies and urban planning.

The dissertation reviews the historical evolution of LBSN data and delimitates the definition of LBSN data. Meanwhile, it tries to construct the theoretic connection between LBSN data and human spatiotemporal behaviors. A snowballing literature study is adopted in the literature review of the dissertation. It summarizes urban applications and limitations leveraging LBSN data and compares their results for increasing knowledge that is currently lacking. Three empirical studies that utilize different LBSN dataset to conduct innovative researches regarding the urban structure, functional relations, and urban

sentiments. Some popular algorithms of spatiotemporal analysis are involved in these studies.

In the Sina Weibo data case, the spatiotemporal variation of Weibo activities reflected how people occupied the urban space dynamically in Beijing, China. The Foursquare case study calculated and confirmed the functional relationship between places in Barcelona, Spain. The Twitter data project investigates the relationship between the urban environment and public sentiments. The results confirm some phenomena that were also observed by other research. Furthermore, they also drill deeper into the relationship between LBSN data and the urban space.

The result argues the irreplaceable position of LBSN data in urban studies and states the potentials of LBSN data from the perspective of scientific urban planning. This dissertation reveals the potentials and limitations of LBSN data at the level of non-government research. LBSN data as a data bridge, connect social activities with geo-space. In the future, cooperating with a keen understanding of society and other datasets, LBSN data can create more possibilities for urban daily life and urban studies.

Keywords: LBSN, urban management, social media, urban planning, Twitter, Weibo, Foursquare, human activities

Resumen

Hoy en día, la utilización generalizada de servicios móviles inteligentes y servicios integrados ubicados extiende la frontera de los sitios de redes sociales (SNS) al espacio físico-espacial. La aparición de los SNS cambia nuestra vida diaria mientras tanto promueve el cambio del paradigma de investigación de todas las disciplinas académicas. Los datos generados a partir de SNS se denominan datos de redes sociales basadas en la ubicación (LBSN).

En cuanto a los estudios urbanos, los datos de LBSN brindan una oportunidad prometedora para comprender la vida cotidiana y las ciudades y, por lo tanto, tienen un impacto profundo tanto en las teorías urbanas como en los estudios empíricos, como el análisis espacio-temporal urbano y el análisis del sentimiento urbano. Sin embargo, a pesar de las oportunidades que brindan los datos de LBSN, algunos desafíos y limitaciones también se asocian con investigaciones relacionadas de estudios urbanos. ¿Quiénes utilizan las aplicaciones LBSN? ¿En cuántos grados podrían representar los datos de las redes sociales la situación real de la ciudad física? ¿Cuál es la relación entre los datos LBSN y los problemas urbanos?

Por lo tanto, aprovechando investigaciones previas, esta disertación busca explorar las aplicaciones y limitaciones de los datos de redes sociales basados en la ubicación para el análisis espacio-temporal urbano basado en un resumen completo de trabajos actuales y tres análisis empíricos basados en diferentes fuentes de datos LBSN en diferentes ciudades. Se espera que el propósito final de la disertación sea obtener nuevos conocimientos que puedan ayudar a futuros estudios urbanos y planificación urbana.

Este trabajo revisa la evolución histórica de los datos LBSN y delimita la definición de datos LBSN. Mientras tanto, intenta construir la conexión teórica entre los datos LBSN y los comportamientos espacio-temporales humanos. Se adopta un estudio de la literatura como una bola de nieve en la revisión de la literatura de la disertación. Resume las aplicaciones urbanas y las limitaciones que aprovechan los datos de LBSN y compara sus resultados para aumentar el conocimiento que falta actualmente. Tres estudios empíricos que utilizan

diferentes conjuntos de datos de LBSN para realizar investigaciones innovadoras sobre la estructura urbana, las relaciones funcionales y los sentimientos urbanos. Algunos algoritmos populares de análisis espacio-temporal están involucrados en estos estudios.

En el caso de datos de Sina Weibo, la variación espacio-temporal de las actividades de Weibo reflejaba cómo la gente ocupaba el espacio urbano de forma dinámica en Beijing, China. El caso de estudio de Foursquare calculó y confirmó la relación funcional entre lugares en Barcelona, España. El proyecto de datos de Twitter investiga la relación entre el entorno urbano y los sentimientos públicos. Los resultados confirman algunos fenómenos que también fueron observados por otras investigaciones. Además, también profundizan en la relación entre los datos de LBSN y el espacio urbano.

Palabras Clave:
LBSN, gestión urbana, redes sociales, urbanismo, Twitter, Weibo, Foursquare, actividades humanas

Acknowledgements

It's four years and a half arduous journey. Fortunately, I have received a great deal of support and assistance from our Centre de Política de Sòl i Valoracions. The brilliant and amiable professors, their insights, have played a central role in the development of the dissertation.

I am particularly grateful to my supervisor, Professor Carlos Marmolejo, who has given me tremendous help since the master's program. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. Meanwhile, you did not hesitate to praise me when I gained some progress. These details gave me a lot of strength during the study. More importantly, your academic passion and professional training impressed me and taught me to investigate this project with a critical eye. (P.S. I still keep those critical comments for future review) Finally, the dissertation started from a naive proposal but has become a cornerstone of my future research. Besides, thank you for all the patient support and opportunities that hanseled to me. Each time I learned a lot from it.

I would also like to thank my second tutor, Professor Pablo Martín, for the valuable guidance and data sources throughout my studies. I still remember the first meeting that we met in Barcelona, which helped to choose the right processing route of the study. The Twitter dataset that I needed is the key element for completing the third case study and my dissertation.

I am sincerely thankful to Professor Josep Roca Cladera and our dear secretary Rolando Biere Arenas. Thank you for the warm support. Each time when I asked for a favor, you always responded to me quickly. I was glad to meet you at the ETSAB's cafeteria on many earlier mornings.

In addition, I would like to thank Dr. Mateusz Gyurkovich and Dr. Damian Poklewski-Koziełł. It was an inspiring and pleasure to have discussions with you. I could not have furnished the dissertation without your help.

Finally, I am deeply thankful to my close friend Hui, my cousin, and my friends who have given me a lot of mental support and love. It is impossible to complete the journey without you. Many thanks to my parents for giving me the free choice of my life. If time is a parallel space, thank the myself in this space for choosing and accomplishing the Ph.D. path.

Contents

Abstract 1
Resumen 3
Acknowledgements 5
Contents 6
Chapter I. 14
I.1. The purpose and significance of the study 14
I.2. Statement of the problem 16
I.3. Scope of the study 18
I.4. Research Objectives 21
I.5. Contributions and highlights of chapters 22
Chapter II. 23
II.1. Background: a brief evolution of Social Network Site 24
II.2. Understanding LBSN data: concepts of social media and location-based social network 28
II.3. Related theories of urban spatiotemporal analysis leveraging LBSN data 31
3.1. Understanding spatiotemporal behaviors: a perspective of time geography. 31
3.2. Understanding LBSN behaviors: uses and gratification theory 36
3.3. Understanding spatiotemporal analysis: data and models 51
Chapter III. 59
III.1. Methodology of the literature review 59
III.2. Framework of the empirical study 61
2.1. LBSN data: types, components, characteristics 61
2.2. General data mining process of urban studies leveraging LBSN data 69
2.3. Outline of case studies 73
Chapter IV. 76
IV.1. Existing literature review 76

IV.2. An overview 78
IV.3. Human mobility 85
IV.4. Urban structure 92
IV.5. Urban land uses and functions 99
IV.6. Urban sentiment / public perception analysis 106
IV.7. Urban health and well-being 112
IV.8. Event detection 117
IV.9. Summary 120
Chapter V. 122
V.1. A survey of the literature review 123
V.2. General limitations of LBSN data 129
2.1. Limitations of the data 129
2.2. Incomparability of datasets and methods 130
V.3. Demographic representativeness 131
V.4. Bias of human online behaviors 134
V.5. Representativeness of spatial human mobility 138
5.1. Precision of LBSN location 138
5.2. Bias of uneven spatial distribution 141
V.6. Representativeness of semantic analysis 143
6.1. The uncertainty of evaluation of sentiments and opinions from texts 143
6.2. Problems of LBSN expressions 144
6.3. Incomparability of data and methods 145
V.7. Summary 149
Chapter VI. 151
VI.1. Case study I 151
Analysis of the Spatial Structure of Beijing from the point view of Weibo Data 151
1.1. Introduction 151
1.2. Literature review 153
1.3. Methodology 157

1.4. Identification of Weibo sub-centers 165
1.5. Results for the identification of Weibo sub-centers 166
1.6. Discussion and conclusion 181
VI.2. Case study II 184
Identifying functional relations of urban places through Foursquare from Barcelona 184
2.1. Introduction 184
2.2. State of the art 186
2.3. Methods 189
2.4. Results 202
2.5. Discussion and conclusions 223
VI.3. Case study III 227
Quantifying the relationship between public sentiment and urban environment in Barcelona 227
3.1. Introduction 227
3.2. Literature review 229
3.3. Methods and materials 235
3.2.1 Sentiment classification 237
3.2.2 Human inspection of Spanish-English translation and sentiment classification 238
3.4. Results 246
3.5. Conclusion 265
Chapter VII. Discussions and Conclusions 267
VII.1. Summary of findings 267
VII.2. Discussions and implications 270
2.1. The indulgence of data: whether LBSN data was over-used? 270
2.2. Discussion: toward citizen science and scientific urban planning 272
VII.3. Conclusion 275
Appendix A. Completed table of regression model 276
References 279

Index of Figure

Figure 1 The terminology of Big data 19
Figure 2 the current popular social network sites 25
Figure 3 A timeline of development of SNS 27
Figure 4 Searching interests of the word "social media" 28
Figure 5 The word cloud of the definition of social media 30
Figure 6 An ancient Egypt map 32
Figure 7 A prototype of time geography 34
Figure 8 A GIS example of space-time prism of time geography 35
Figure 9 Approaches of LBSN behavior investigation 40
Figure 10 Study structure of LBSN behaviors 41
Figure 11 Ranking of use- motives among Facebook, Twitter, Instagram, and Snapchat 50
Figure 12 Increasing cost efficiencies of computation with a marked step change from mechanical to electronic processing technologies 52
Figure 13 Movement of global merchant fleets 53
Figure 14 Visualization of a four-day's sample: above: New York, below: Tokyo 54
Figure 15 The distribution of bacon tags 55
Figure 16 Specification of geo-spatiotemporal model 56
Figure 17 Construction of time object 57
Figure 18 Framework of urban computing 58
Figure 19 Procedure of literature review 60
Figure 20 Graph representation of LBSN data 64
Figure 21 Core features of LBSN data 65
Figure 22 Spatial distribution and temporal behaviors of Instagram and Foursquare users 66
Figure 23 Movement trajectory of tourist generated from geotagged photos 67
Figure 24 Plaza de España in the view of Google Maps 68
Figure 25 Framework of LBSN data analysis 69
Figure 26 Flowchart of data collection 71
Figure 27 Location of case studies 73
Figure 28 Number of reviewer of articles by year 78
Figure 29 Number of publication by searching keywords: Urban & LBSN 79
Figure 30 Number of reviewed articles by country 80
Figure 31 Data sources of reviewed articles 82

Figure 32 Breakdown of reviewed articles in different urban issues 84
Figure 33 Spatial density of Weibo check-ins 93
Figure 34 Extracting AOI from Flickr photo in NewYork; (a) locations of Flickr photos; (b) point clusters detected by DBSCAN 94
Figure 35 Graphical representation of the LDA model. 102
Figure 36 Different paths between Euston Square and Tate Modern 108
Figure 37 Daily SMAI for whole of UK over six week 113
Figure 38 Division of representativeness of LBSN data 123
Figure 39 Breakdown of reviewed articles regarding LBSN limitations 124
Figure 40 Year-distribution of reviewed papers 125
Figure 41 Frequency of LBSN data sources 125
Figure 42 Distribution of studied countries 126
Figure 43 Country-distribution of study cases 127
Figure 44 Distribution of LBSN data in each domain 128
Figure 45 Age distribution of Twitter and Facebook users from UK 133
Figure 46 Difference of check-ins between gender 136
Figure 47 Results of Foursquare check-ins against GPS traces 139
Figure 48 Result of survey of geotagged tweets in Japan 140
Figure 49 Quantitative methods of sentiment analysis from 2010 to 2020 147
Figure 50 The functional zones of Beijing according to the 2004 City Master Plan 154
Figure 51 The layout of Beijing Master Plan 2004 155
Figure 52 Schematic diagram of the monitoring range 160
Figure 53 Detail of Beijing urban tissue 162
Figure 54 Anselin Local Moran's I Result 164
Figure 55 Temporal distribution of Weibo activities 167
Figure 56 3D Weibo density panorama 168
Figure 57 3D view of Weibo density in the central area of Beijing 169
Figure 58 3D view of urban contexture in the Guomao CBD area 170
Figure 59 3D view of Weibo density in traditional city center and Haidian 171
Figure 60 Distribution of potential Weibo sub-centers in Beijing 173
Figure 61 Division of workday's potential sub-centers 177
Figure 62 Confirmed sub-centers at the weekend and overlapping potential areas 181
Figure 63 Distribution of check-ins in Barcelona Metropolitan Region 190
Figure 64 Monthly Check-ins of Foursquare in Barcelona 191
Figure 65 Outline of analzying process 192
Figure 66 Foursquare users' check-ins and stay duration 194
Figure 67 Spatial activities of tourists and residents 203
Figure 68 Major tourist spatial flows and POIs' check-ins 206

Figure 69 Major local spatial flows and POIs' check-ins 207
Figure 70 The distribution of popular POIs around Pl. Catalunya 208
Figure 71 Heat map of tourist usage-flows matrix 211
Figure 72 Heat map of tourist usage-flows matrix 212
Figure 73 Matrix of interaction values among tourist usages 213
Figure 74 Matrix of interaction values among local usages 214
Figure 75 Proxcal plot of interaction value matrix (tourist group) 215
Figure 76 Proxcal plot of interaction value matrix(local group) 216
Figure 77 Graph of paired usages with prominent interaction values (tourist group) ... 220
Figure 78 Graph of paired usages with prominent interaction values (local group)..... 221
Figure 79 Comparison between functional proximity and spatial proximity(tourist group) 222
Figure 80 Comparison between functional proximity and spatial proximity(local group) 223
Figure 81 Distribution of tweets' language 236
Figure 82 Total density of tweets of Barcelona based on AEBs 241
Figure 83 Spatial distribution of tweets in public spaces of Barcelona city 246
Figure 84 Spatial distribution of tweets in central area of Barcelona 247
Figure 85 Distribution of the three languages in public spaces of Barcelona..... 248
Figure 86 Variation of percentage of sentimental tweets 251
Figure 87 Variation of the net sentiment score 251
Figure 88 Weekly variation of negative tweets 252
Figure 89 Distribution of positive and negative sentiments 253
Figure 90 Spatial distribution of sentimental tweets and net sentiment 254
Figure 91 Spatial distribution of sociodemographic, human mobility and socioeconomic activities 258
Figure 92 Spatial distribution of built environment 261
Figure 93 Distribution of net sentiment at AEBs level used in the model..... 264

Index of Tables

Table 1 Taxonomy of gratifications on media uses 43
Table 2 List of keywords searching via Google scholar 60
Table 3 Criteria of selection of academic articles 61
Table 4 Characteristics of different LBSN data 63
Table 5 Components of LBSN data 63
Table 6 Objects of data cleaning 71

Table 7 Summary of case studies 74
Table 8 Summary of existing literature review 77
Table 9 The distribution of publication source 81
Table 10 Data sources of reviewed articles 83
Table 11 Summary of representative studies of human mobility 89
Table 12 Summary of analyzed studies of urban structure 97
Table 13 Summary of analyzed studies of urban functions/ land uses 104
Table 14 Summary of representative studies of urban sentiments 110
Table 15 Summary of representative studies of urban health and well-being 115
Table 16 Summary of analyzed studies of event detection 119
Table 17 Demographic compositions of popular social media platforms of U.S. 131
Table 18 Summary of representative studies of demographic bias 137
Table 19 Comparative summary of different data collection techniques 138
Table 20 Summary of representative spatial bias studies 142
Table 21 Typical intricate elements of semantic analysis 145
Table 22 Summary of representative studies of semantic analysis issues 148
Table 23 Parameters of monitoring circles 159
Table 24 Three different sizes of quadrat 163
Table 25 Statistical description of potential Weibo sub-centers and distance 172
Table 26 The land use type of overlapping potential sub-centers 174
Table 27 Model Summary of the four periods 176
Table 28 Model Summary of Workdays and nights of workdays 178
Table 29 The tested sub-centers of weekend based on two standards 179
Table 30 The land use type of confirmed sub-centers at the weekend 180
Table 31 Components of Foursquare data 191
Table 32 Summary of valid users 192
Table 33 Four thresholds of improvement 195
Table 34 The summary of locals and tourists 196
Table 35 New category of Foursquare POIs 197
Table 36 Spatial distribution of check-ins among municipality 204
Table 37 Number of visitors to major attractions in Barcelona 206
Table 38 Difference of urban usages between tourists and locals 209
Table 39 Prominent interaction values 217
Table 40 Thresholds of sentiment classification 237
Table 41 Result of Spanish sampling sentiment evaluation 238
Table 42 Number and Reasons of different sentiment classification 239
Table 43 Urban environment indicators 244
Table 44 Spatial distribution of Tweets in public spaces 247

Table 45 The top forty high-frequency words 249
Table 46 Results of Sentiment classification 250
Table 47 Pearson correlation between urban indicators and sentiment indicators 255
Table 48 Model summary of sentiment density 262
Table 49 Model summary of the net sentiments 265

Chapter I.

Introduction

I.1. The purpose and significance of the study

Nowadays, shopping, working, entertainment, even dating, all aspects of daily life are cooperating with the Internet. The Internet is becoming a "natural" urban setting, in which Social Network Sites (SNS) are a pivotal part of the new environment. These Social Network Sites, such as Facebook, Twitter, YouTube, and TikTok, attract millions of users by incorporating various information and communication. Furthermore, the widespread utilization of intelligent mobile and located-embedded services extends the border of SNS to physical-spatial space. Such a data is named as location-based social network(LBSN) data. For example, users can upload content that they want to share and mark their locations at the same time. Conversely, places, like restaurants and tourist attractions, can be visualized online by comments and labels of people on various social networking sites. The appearance of SNS changes our daily life meanwhile promotes the shift of research paradigm of all academic disciplines.

Location-based social media data (LBSN) provides a promising opportunity to understand the quotidian life and cities. As to urban studies, LBSN data brings a profound impact on both urban theories and empirical studies. Regarding spatiotemporal human movements, it allows researchers to observe the city more precisely: from group movements (Kannangara, Xie, Tanin, Harwood, & Karunasekera, 2020) to urban regional mobility(Long, Han, Tu, & Shu, 2015; J. Yuan, Zheng, & Xie, 2012). Moreover, the semantic information of LBSN data can disclose the function and people's perception about places. In other words, the interaction between the urban built environment and citizens can be observed by LBSN data. "Cities have become highly interconnected techno-social systems embedded with intricate and complicated spatial and social networks"(W. Luo, Wang, Liu, & Gao, 2019). Therefore, it brings new insights to the urban analysis, such as urban mobility, social segregation, urban sentiment, etc. "We live perhaps on the eve of an epistemological revolution that will result in the end of

conventional urban theory and the creation of data-driven rather than knowledge-driven urban science" (Rickards, Gleeson, Boyle, & O'Callaghan, 2016).

In 2007, Jim Gary described the change of science paradigms (Hey, Tansley, & Tolle, 2009)– from empirical description to theoretical generalization, simulation of phenomena, and then data exploration right now. The availability of large volumes of LBSN data leads to a new paradigm of academic research for social science(Cambria, Wang, & White, 2014; Lazer et al., 2009), medical science, and environmental engineering. For example, Nagel et al. (2013)explored the relationship between the outbreak of flu and the related information on Twitter. Sakaki, Okazaki, and Matsuo (2010)investigated the interaction of event -- earthquake in Twitter and produced a spatiotemporal model that can predict the location and the center of the earthquake in Japan. Therefore, data-driven researches have becoming one of the main philosophies of academic researches. It is necessary to discuss the pros and cons of the new paradigm.

However, despite opportunities that LBSN data provides, some challenges and limitations also associate with related researches of urban studies(Martí, Serrano-Estrada, & Nolasco-Cirugeda, 2019), such as the representativeness of LBSN data(Mitchell, Frank, Harris, Dodds, & Danforth, 2013), the quality of the datasets(Steiger, De Albuquerque, & Zipf, 2015) and the incomparability of methods(Ruths & Pfeffer, 2014). For example, part of bias of data is caused by the obscure filtering process of data streams, different social media data sources, different time span of data tracing, and among others. It lacks a universal database to guarantee the quality of data of researches, if not impossible. In fact, it is difficult to evaluate the effectiveness of researches, even within the same city or region. On the other hand, the challenge lies in the development of new theories and methodologies to explain and the interconnections of spatial space and the LBSN data. There are hundreds of urban studies related to LBSN data have been generated every year. However, the theoretic integration of cities, urban geography and LBSN data are not sufficient, due to the seas of data and the diversity of urban issues.

Therefore, taking advantage of previous researches, this dissertation seeks to explore the applications and limitations of location based social media data for urban spatiotemporal analysis based on a comprehensive summation of current literatures and three empirical analyses based on different sources of LBSN data in different cities. The ultimate purpose of the dissertation is expected to gain new knowledge that knowledge that could help the future urban studies and urban planning.

I.2. Statement of the problem

Compared with traditional mobility survey data, LBSN data is generated by users spontaneously and can be traced without time lag. When people start to use social media, the data is recorded simultaneously. LBSN data can deliver the geo-information, social information and people's mobility to some degree. What's more, the recording scale of social media data continues to expand in both of time and space dimensions. Taking advantage of these, LBSN data can eliminate some biasness of traditional survey and provide abundance data for dynamic researches. For example, traffic flows are usually obtained by official traffic surveys, which fails to catch the dynamics of human movements immediately.

However, the limitations of LBSN data are also under debate. First of all, as to the data itself, the missing of social identification in LBSN data causes the disconnection of visual and physical world, and biasness of data results. How many degrees could daily activities data represent the actual situation of the physical city? LBSN data are affected by demographic factors, user's preference and social statue, and urban infrastructures. How can these bias affect the results when we try to use these data to make spatial analysis? While huge volume data are recorded every second, the part that we could investigate is merely the data with regular patterns or characters. Since those data providers (e.g. Twitter, Facebook) do not public the detail filtering algorithm of accessible data, it is hard to investigate the whether these data are sufficient to represent the characteristics of the physical world. Secondly, with the growing of researches using social media

data, the applied methods of analysis lack a systematic summarization which allows to evaluate the potential of LBSN data. For example, from 2005 -2013, nearly half of academic studies that are based on Twitter data were in the United States (Steiger, De Albuquerque, et al., 2015). If the research scope shifts to other areas where twitter is not popular, it is hard to infer that the same methods or parameters of the method are also suitable for those areas.

Furthermore, the relationship between the virtual space and the physical space is not well recognized. For example, why are people willing to use them and expose their location? How does the physical urban structure impact on these data? How do these data reflect the urban structure, as the temporal attribute of the urban structure is highlighted? Many works of literature only focus on the phenomenon and information flows of urban areas. The deeper theoretic integration of geography, social theories, and social network data is urgent to develop. Previous theories, such as time geography (Hägerstrand, 1970), action space(Horton & Reynolds, 1970), and mental map(Lynch, 1960) described such relationships from different standpoints: the spatial and temporal location of movements, influential factors that form people's action space, and the recognition process of the built environment. However, they are independent of each other. Few theoretic frameworks try to reunion them and generate a comprehensive mechanism to decompose the relationships between the human movements, social space, and the urban environment.

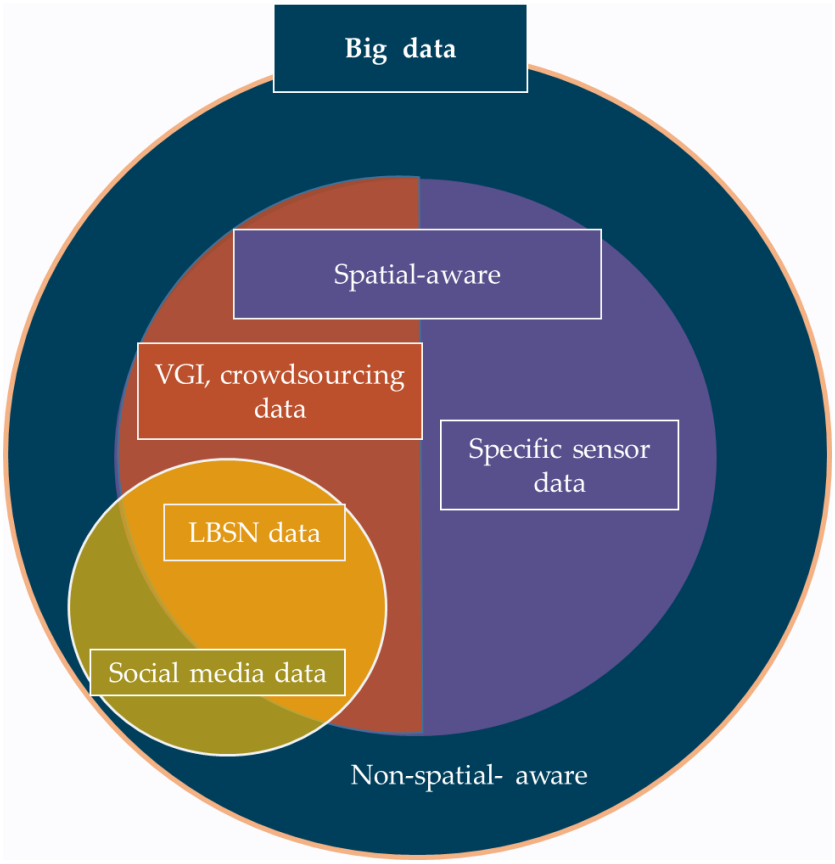
Nevertheless, the primary purpose of this study is not to develop an omni-theory of urban spatiotemporal analysis. Instead, the subject matter explored in this dissertation is foremost driven by the previous empirical studies based on LBSN data and three study cases in Barcelona and Beijing. This dissertation aims to summarize applications using LBSN data and compare their results for increasing knowledge that are currently lacking.

I.3. Scope of the study

Before delimitating the scope of the study, it is necessary to clarify several popular terms that are frequently appeared in academic articles: Big data, spatial-aware data, crowdsourcing data, volunteered geographic information(VGI), social media data, and LBSN data.

The term of "Big data" usually means "datasets are so large and complex that they become awkward to work with using standard statistical software"(Snijders, Matzat, & Reips, 2012). Furthermore, this broad definition can be described as a process of alchemy of large-scale data: "'big data' often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of dataset." (Cavanillas J M. et.al, 2014) Therefore, it is not a new creation or discovery. These data are derived from the reality and acquired by emerging technologies. "Big data" is a more precise and detailed description of human society and the environment.

Under this definition, all aforementioned data could be considered as "Big data"(**Figure 1**). According to the availability of location detection, "Big data" can be divided into spatial-aware and non-spatial-aware data.



Source: own elaboration

Figure 1 The terminology of Big data

In the field of urban studies, there are two types of spatial-aware data that are very popular: 1) specific sensor data – such as remote sensing images, SPOT images, and aerial photographs; 2) data from popular applications or urban services system – such as taxi-tracking, public transportation card, and social media data. The former one belongs to geographical data that can depict their monitored objects with a high precision. The latter one are produced by human activities spontaneously and provides more social-dimensional information of users.

With the development of technologies, more and more electronic devices are equipped with geo-location services. The nearly instant positioning becomes available and affordable for commercial organizations and common people. Many applications, such as GPS navigation, open source map, and social media sites, allows people to access their current locations and create places on the map.

Therefore, Goodchild (2007) brought up the term *volunteered geographic information* (VGI) to describe the phenomenon that citizens create geographic information using various electronic devices. However, the citizen is only one link of the generation of geospatial data that includes the prior data modeling, data integration, and representation. Therefore, Heipke (2010) chose "*crowdsourcing* to describe data acquisition by large and diverse groups of people, who in many cases are not trained surveyors and who do not have special computer knowledge, using web technology". In essence, the two concepts refer to the same type of data.

Social media data is generated from social network sites, which contain contents related to users' social life, such as opinions, pictures, and social networks. Some social media data belong to non-spatial-aware, such as YouTube, Reedit[1], and WhatsApp[2]. Combined with the geo-location services, location-based social networks (LBSN) enhance the connectivity between the Internet space and the physical space. Because of the great commercial potentials, many social media companies provide free and commercial API to visit these data, such as Twitter and Google Maps.

By "urban spatiotemporal analysis using LBSN data" we refer to the urban analysis that is based on the *spatiotemporal signal* (Steiger, De Albuquerque, et al., 2015) of LBSN data which consists of a geo-location and a timestamp. These spatiotemporal signals are generated when users of social network sites post information on platforms. Therefore, analyzing these signals is useful to understand the citizen's movements, spatial structure, land uses, and patterns of urban mobility. For example, Marmolejo and Cerda Troncoso (2017) identified the urban subcenters using citizen trip-chains in Barcelona. In fact, the human-centric perspective has become the new frontier of urban studies. In addition, combining with the semantic and graphical information, LBSN data enriches social details of locations, such as events in a location, perceptions of a place, and usages of places. However, the semantic analysis of LBSN data will be restricted to the urban sentiment analysis due to the complexity of semantic analysis.

[1] https://www.reddit.com/

[2] https://www.whatsapp.com/?lang=en

The dissertation mainly investigates three influential LBSN data: Twitter, Sina Weibo, and Foursquare. Twitter and Sina Weibo are micro-blogs that allow users to publish short messages, images, and videos as well as share their positions. Twitter is a world-wide platform except a few countries, however, Sina Weibo is the biggest platform of micro-blog in China. Foursquare is a local search-and-discovery service application, which could provide practical information about living for users. Compared with Twitter or Sina Weibo, Foursquare focuses on the places rather than individuals. Therefore, these three data source could represent the universality of LBSN data while the difference of them are presented in terms of the different type and region.

Our case studies especially give attention to the urban structure, functional relationships of urban places and the urban sentiment. Utilizing these data, the quantitative part explores three major urban issues: urban structure, urban function, and public sentiment. One the one hand, these studies demonstrate advantages of LBSN data for urban analysis, which disclosure many aspect of urban life and can be proxy of human activities in terms of a dynamic view. On the other hand, cooperated with statistical data from the local government, we develop new applications of LBSN data that highlight the "invisible" relationships in cities: urban active areas, functional relationship between places and the correlation between urban environment and public sentiment.

I.4. Research Objectives

The general objective of this dissertation focus on investigating the applications and limitations of location-based social media data in urban spatiotemporal analysis, which can support for future urban studies and urban planning.

Specific objectives:

1. To study the evolution, definitions, and characteristics of LBSN data.
2. To summarize the applications of LBSN data through a systematic literature review.

3. To identify potentials and concerns of LBSN data in urban spatiotemporal analysis.
4. To develop innovative urban analysis using LBSN datasets.
5. To discuss the potential role of LBSN data in future research and urban planning.

I.5. Contributions and highlights of chapters

Chapter II introduces the historical development of social media network sites and reviews the definition of social media, location-based social network data. More importantly, it explores why people use LBSN applications, how to understand spatiotemporal behaviors, and how to analyze massive spatiotemporal behaviors.

Chapter of Methodology explains the method snowballing literature studies that adopt in the literature review of the dissertation. As each case study has distinct objectives and methods, this chapter mainly introduces the framework of the three empirical studies, the general LBSN data structure, and the mining process.

Chapter IV and V review the main applications and limitations of LBSN data in the field of urban spatiotemporal analysis during the past decade. The dissertation also conducts a systematic survey that summarizes the spatial and topic distribution of these researches.

Chapter VI consists of three empirical studies that utilize different LBSN datasets to investigate the urban structure, functional relations, and urban sentiments. The results confirm some phenomena that were also observed by other researchers, such as the spatial agglomeration of LBSN data, and the relationship between socio-demographic profiles and the Twitter sentiments. Furthermore, they also drill deeper into the relationship between LBSN data and the urban space.

Combined with previous investigations, the final chapter discusses the current status and future applications of LBSN data. Moreover, it argues the irreplaceable

position of LBSN data in urban studies and states the potentials of LBSN data from the perspective of scientific urban planning.

Chapter II.

Background and related theories of LBSN data analysis

II.1. Background: a brief evolution of Social Network Site

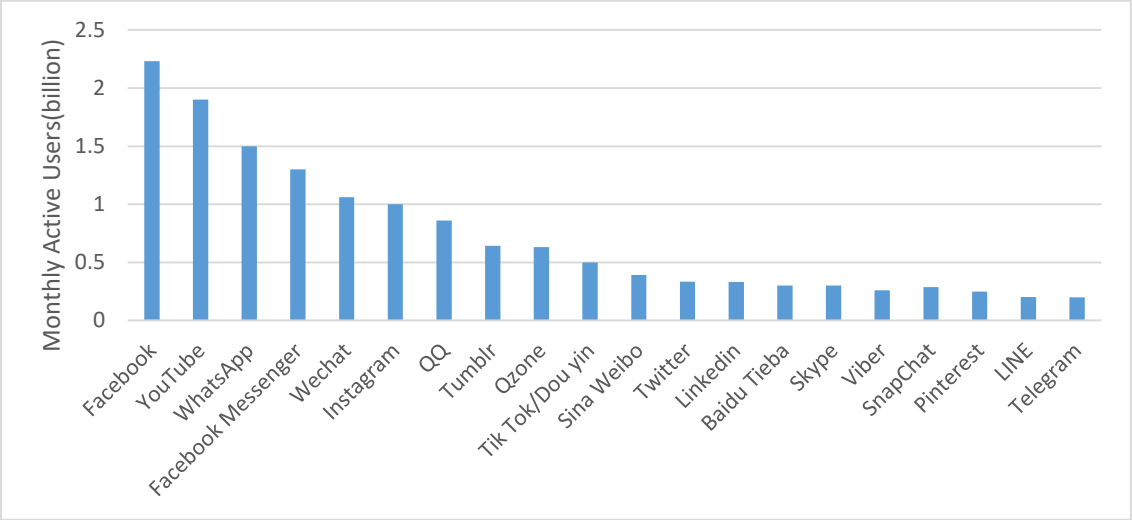
The growing availability of LBSN data is accompanying with the development of social network sites and location-based services (LBS). Social network sites provide services that allow people to accomplish social activities online, such as self-presentation and communication. Location-based services support users to access their current positions, which are based on global positioning system(GPS), wireless local-area (WLAN), or Radio-frequency identification (RFID) (Kolodziej & Hjelm, 2017).

A social network means "a social structure made up of individuals connected by one or more specific types of interdependency, such as friendship, common interests, and shared knowledge "(Y. Zheng, 2011). A social structure is a term of sociology that is used to describe the stable form or pattern that people interact with each other. In other words, social networks are social relationships.

Therefore, a social network site connects and reflects such relationships via Internet platforms. According to the definition of Boyd and Ellison (2007), social network sites is a kind of web-based services, which allows individuals to create online profiles, connect each other and share information. There is interchangeable name – "social networking site" also appears in public disclosure. The characteristics of these social network sites may differ from site to site, which are decided by their functions. For example, compared with Instagram, Facebook has stronger ties of social relationships because Facebook aims to construct the online social network among people while Instagram aims to share visual things.

Figure 2 lists the top 20 of the most popular social network sites (SNS) based on their monthly active users in April of 2019. Monthly active users (MAU) means the number of users visit or use a SNS over one month. Therefore, 2.3 billion MAU means that more than 30% of world population used Facebook in that month. In

some area, the percentage of SNS users is even higher. For example, Wechat is the major SNS application in China which owns 1.06 billion MAU. Considering the total population of China is about 1.42 billion, it means that almost every Chinese adult install Wechat application.



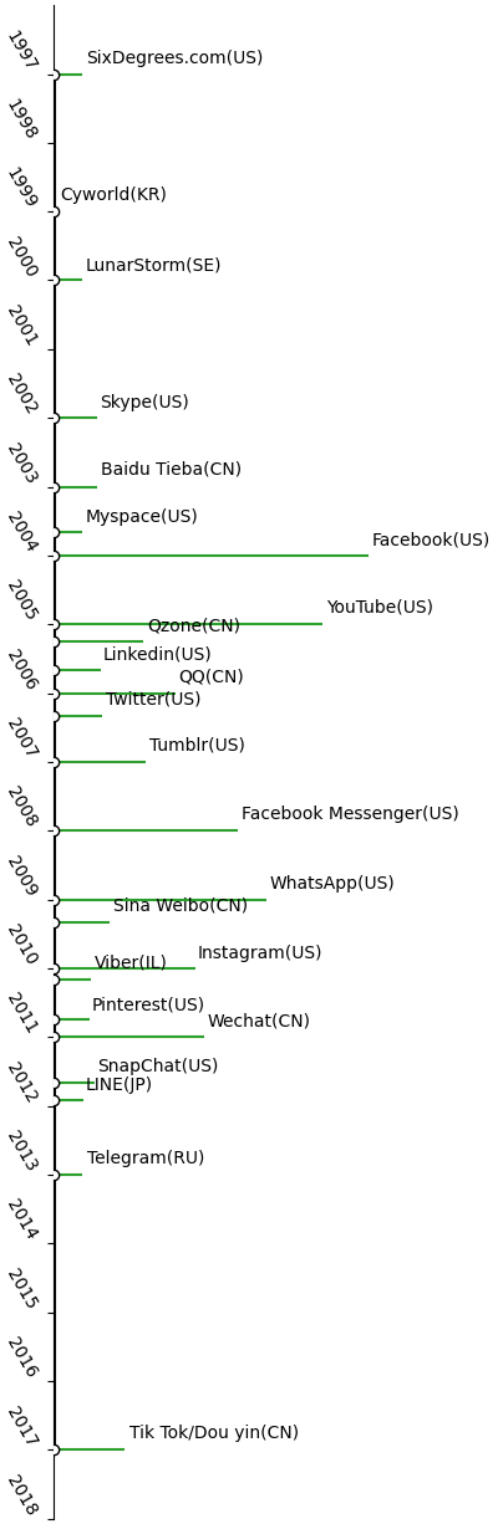
Source:https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Figure 2 the current popular social network sites

In fact, the development of SNS has been merely more than 20 years, since the first SNS – Sixdegree.com was established in 1997. It was named after the theory of six-degree separation which claims that the social connections (a friend of a friend) of any two people is less than six steps. Registered users could find interested person through the site. However, SNS has not gain enough popularity until 2003. In 2003, the first influential SNS – MySpace appeared and became the largest SNS in next five years. The number of users surpassed 100 million in 2006.

Figure 3 outlines those most influential SNS and their launched time. Since 2004, SNS has grown all over the world rapidly due to the popularization of the Internet. Facebook was launched in 2004. Two years later, Twitter, as another type of social media that focused on exchanging short messages, was created and successfully spread over the world. The majority of SNS are created from the United States; a few of them (Wechat, QQ, Qzone, Douyin/ TikTok, Sina Weibo, and Baidu Tieba) are from China; LINE and Telegram are widely used in Japan and Russia separately. At the same time, different types of SNS also developed rapidly. For example, YouTube is a video-sharing website; Whatsup belongs to an

application of personal-immediate communication. Although the popularity of social media applications is different in different countries, it is undeniable that SNS has been part of people's daily life and change the way that people consume information. The rising of video-sharing SNS, such as Instagram and TikTok, indicates that more and more people covey and receive information via the form of videos rather than texts.

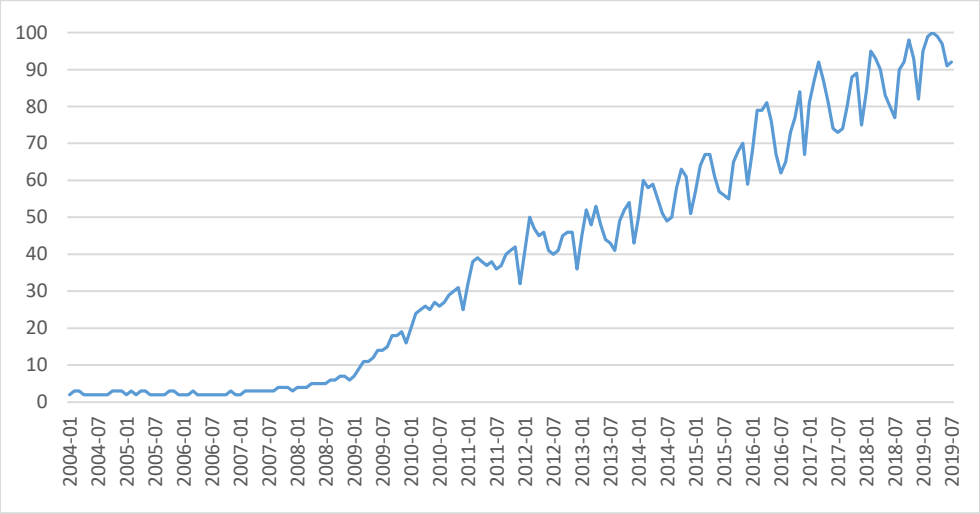


source: own-elaboration

Figure 3 A timeline of development of SNS

II.2. Understanding LBSN data: concepts of social media and location-based social network

As SNS is spreading globally, "social media" becomes a popular word. According to Google trends, the worldwide search-interest about "social media" increased significantly after 2009 (**Figure 4**). The searching interest has increased nearly 100 times. The definition of social media also varies according to different academic researcher as the following.



Source: Google Trends.

Note: the vertical axis describes the search-interest using a score range from 0 to 100.

Figure 4 Searching interests of the word "social media"

In 2016, Merriam-Webster defined social media as "forms of electronic communication through which users create online communities to share information, ideas, personal messages, and other content." In this sense, social media is a kind of communication tools which move the activity of communication from offline to online. The similar definition that can be found in academic literature:

> Social media represents "the technologies or applications that people use in developing and maintaining their social networking sites. This involves the posting of multimedia information (e.g., text, images, audio, video), location-based services (e.g., Foursquare), gaming (e.g. Farmville, Mafia Wars)"(Albarran, 2013).

However, social media actually contains meaning beyond the technological applications. "User-generated content" is a fundamental difference between traditional media and social media. The published information is no longer produced solely by media companies or the government. Any individual can register an account on SNS and publish information：

> "Social media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content" (Kaplan & Haenlein, 2010).

> "Internet-based, disentrained, and persistent channels of mass personal communication facilitating perceptions of interactions among users, deriving value primarily from user-generated content"(Carr & Hayes, 2015).

Further, social media forms an environment or a digital space for our society when the volume of users are huge enough. In this sense, social media is nothing new in human society. It is just a form of exchange of information which is based on Internet. "Humans in social relation produce them"(Fuchs, 2017, p. 37):

> Social media is "an environment in which information is "passed from one person to another along social connections, to create a distributed discussion or community"(Standage, 2013, p. 3).

> "Social media is made up of various user-driven platforms that facilitate diffusion of compelling content, dialogue creation, and communication to a broader audience. It is essentially a digital space created by the people and for the people, and provides an environment that is conducive for interactions and networking to occur at different levels (for instance, personal, professional, business, marketing, political, and societal)"(Kapoor et al., 2018).

We put these definitions into a word cloud (**Figure 5**). The top fifteen of high-frequency words of these definitions display the key characteristics of social media.

The larger size of the word appears; the higher frequency it has. It is obvious that social media has a strong attribute of the social interaction. Combined these words, social media can be summarized as an Internet-based environment for communication and personal interaction. Its essential function is to help people to communicate and interact with others. Such a social communication is beyond "true relationship" (such as friendship or colleagues). Two individuals can interact with each other due to the common interests or an activity, even a place.

Source: own-elaboration

Figure 5 The word cloud of the definition of social media

Therefore, a location-based social network (LBSN) not only inherits the social information, but also provides spatial-aware functions that indicate the location of information(Wilson, 2012). It connects the social networks with the physical world in the form of spatiotemporal signals. It allows people to know about places meanwhile adds the social dimension in places. For example, people can make check-ins and comments at a tourist attraction. In reverse, people also can access the information of the tourist attraction that are created by other users, such as its popularity, comments and characteristics. Furthermore, the dynamic social / functional relationship between places is outlined by people's activities on LBSN site, such as the connection between hotels and tourist attractions, working places and residential places. Consequently, LBSN data enable many urban scholars to study the variation of urban dynamics using spatial temporal analysis(Collins, Hasan, & Ukkusuri, 2013; D. Paul, Li, Teja, Yu, & Frost, 2017; Yaqub et al., 2020).

II.3. Related theories of urban spatiotemporal analysis leveraging LBSN data

As previously stated, the emerging of "Big data" actually turned over the research paradigm of urban spatiotemporal analysis. Essentially, it provides the opportunity to shift the starting point of research from an aggregated spatial unit to the individual people because it allows scientists to trace the nearly real-time personal movements and information exchange. Such an advantage helps us to interpret related theories and develop new knowledge. However, back to the LBSN data itself, several critical questions are waiting to clarify. Why does LBSN data can be exploited as the data source to represent the spatiotemporal movement? What the psychological mechanism that incites people to use LBSN applications? How to record and represent human spatial behaviors on maps?

Therefore, for understanding the temporal-spatial movement of human beings, this chapter carries on the retrospect of time geography theory to cope with the LBSN data. Secondly, we attempt to explore the motives that people use LBSN applications and share their locations. Finally, a perspective of modeling explains the construction of spatiotemporal analysis.

3.1. Understanding spatiotemporal behaviors: a perspective of time geography

It is a long story that people explore the nature of time and space. "Time has often, though not always, been considered along with space, either as an extension of space or in analogy with it"(Couclelis, 1999). Geography probably is the earliest empirical summary and representation of sensing time and space. Human being has to learn the surroundings for surviving and development: from rivers, mountains, weather to tribes, villages, cities. For preserving these knowledge, people draw maps and write down notes to indicate features of the space. For utilizing the environment, people make scales of maps and measure the space by distance. As Einstein said, "There is no such thing as empty space, i.e. space without field. Space-time does not claim existence on its own, but only as a structural quality of the field"(Einstein, 1920).

Note: The eastern part of a topographical and geological map of Wadi Hammamat, in the desert between the Nile and the Red Sea, drawn on papyrus around 1150 BCE. Source: Barnard (2008)

Figure 6 An ancient Egypt map

The earlier scholars of regional study conceive the space as an "independent entity" of investigation. Space is a container in which " land uses, towns, were slotted into a preexisting set of geographic pigeon holes"(Cox, 2013). Walter Christaller's central place theory and the traditional models of urban structure (e.g. Hoyt (1939) ;Harris and Ullman (1945)) could be considered as typical examples of static-space view.

Even for those spatial-quantitative geographers in the 1960s, the subject of the investigation was still the space. They focused on the spatial interaction between places, such as flows of transportation and migrations. The social relations were put under the spotlight, just as Harvey (1973) defined "space regarded…as being contained in objects in the sense that an object can be said to exist only insofar as it contains and represents within itself relationships to other objects." So far, Human movement was a tool rather than the end.

However, just as Hägerstrand asked," what about people in regional science?" Urban study should not only focus on the socio-economic sphere and the abstracted debates, but it is also vital to understand how people access different utilities and places. "Nothing truly general can be said about aggregate regularities until it has been made for clear how far they remain invariant with

organizational differences at the micro-level."(Hägerstrand, 1970) Therefore, he introduces the concept of time geography to depict spatial human behaviors.

Time geography observes the world from a four-dimensional view – matter, space, and time. He mentions three major ways to understand time and space (Hägerstrand, 1989): subjective experience, artificial measurements such as calendar, and embedded space-time which means matter itself defines space and time by its successive configurations, such as the biological life span.

In the framework of time geography, space-time follows artificial measurement, for example, a day or a week. Human activities are describable within a limited scale during a period due to the existence of various constraints.

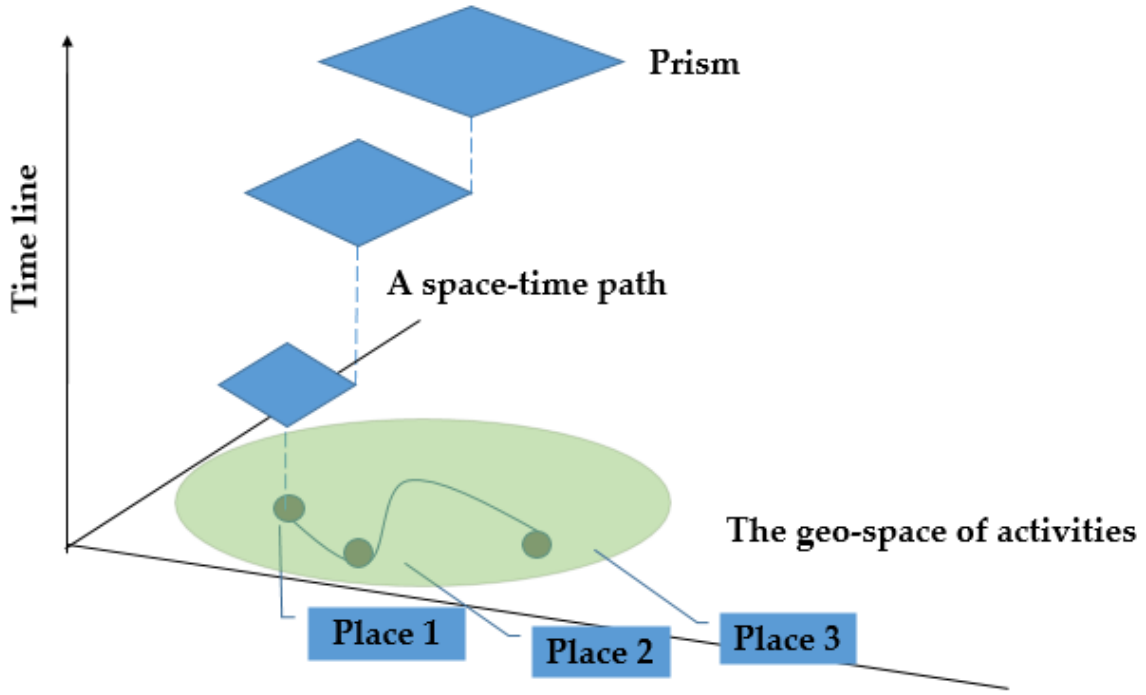
These constraints can be classified into three types:

1. capability constraints, for instance, it is impossible to appear in two places at the same time for a single person;
2. coupling constraints, such as working schedule and appointments;
3. authority constraints that limit people to access a place, such as the opening time of a school or office building. In other words, a person could only stay at the place at a particular time for a purpose.

Hence, constraints on human activities provide a feasible approach to understand urban activities. It indicates that each human movement has its purposes and spatiotemporal boundaries. "Time geography… is a discipline-transcending and still evolving perspective on everyday workings of society and the biographies of individuals." (Thrift & Pred, 1981) . Therefore, urban activities could be conceived as a network that is composed of paths of human movements and their social relations.

Time geography analyzes human activity using two basic concepts: space-time path and prism (**Figure 7**). The space-time path indicates the trace of a person during a period and also delimitates his/her potential geographical boundary of activities. The prism refers to the staying time at one place. Since the time is limited, like one day or one year, a longer duration of stay at one place means the

remaining prisms have to shrink. The project of space-time paths is the potential activity area.



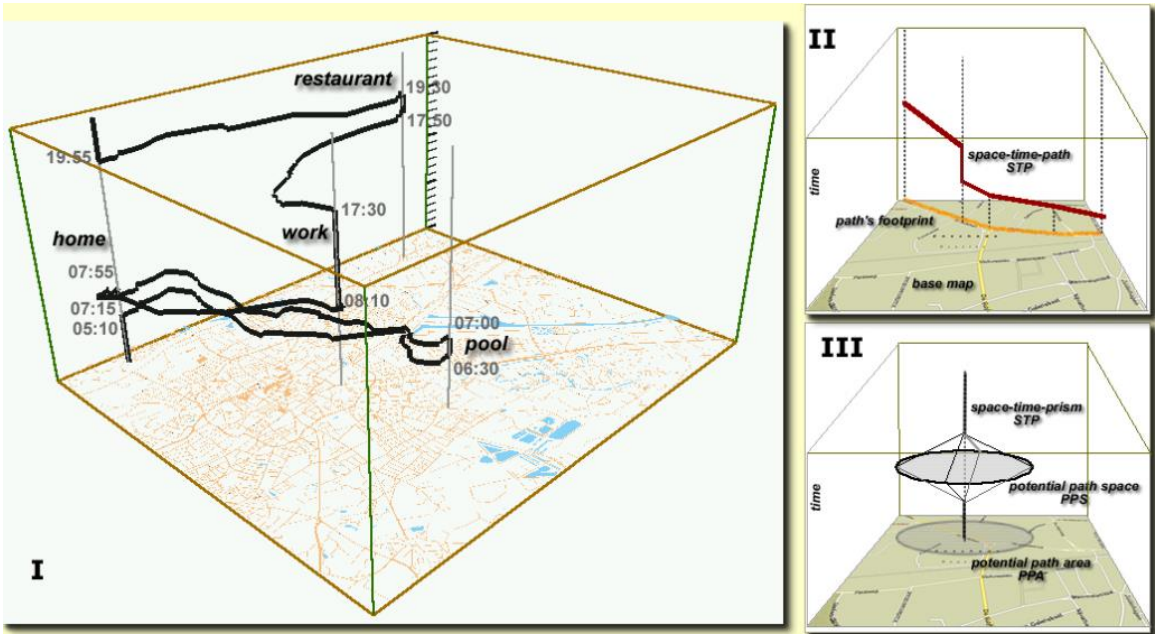
Source: own elaboration

Figure 7 A prototype of time geography

In the early phase, time geography mainly focused on empirical studies on small scales due to the limitation of the technology. For example, the PESASP program led by Bo Lenntorp designed a model for Karlstad's traffic planning in Sweden in 1976 (M. P. Kwan, 2004). It used a travel survey of 980 households to generate 5,500 items of travel records.

In the 1980s, Allan Pred began to introduce time geography into the macro-scale of human society, integrate with the theory of structuration to explain the relation between individual activities and social reproduction. Since social reproduction could be decomposed by various types of social institutions, such as schools and factories, the intersections of individual paths and institutions establish a kind of relationship of production (Pred, 1981). For example, a workplace may exploit eight hours of an employee, and thus his/her spatial scale

of movement is limited. In other words, the occupation of time and space is a reflection of social reproduction.



Note: I. The whole traveling path. II. The simplified space-time path. The vertical line in the path represents the stay of duration at the place. III. The space-time-prism (STP). Source: Kraak (2003)

Figure 8 A GIS example of space-time prism of time geography

With the development of Geographic Information Systems(GIS), the theory of time geography gained a repulsion through the computational visualization of the theory's entities(M.-P. Kwan, 1999; Miller, 1991, 2005). For example, Miller (1991) established a GIS-based algorithm to calculate the potential path areas. Kraak (2003) visualized one day of his traveling trajectory in Netherland**(Figure 8**). Utilizing the time-geographic view, M.-P. Kwan (1999) studied the pattern of home-work movements between different gender in Columbus, U.S. The result proved that women had a higher level of fixed constraint due to domestic responsibility. The theoretic structure of time geography is also applied in the modeling of transportation flows(Miller & Bridwell, 2009; Neutens, Schwanen, & Witlox, 2011).

After the proliferation of LBSN data and other spatial-aware data, instant spatiotemporal movements can be observed and stored by computers. It not only

proves the hypothesis of time geography, but it also can provide large-scale observations of human mobility. However, time geography has not developed further. One explanation is that current algorithms of spatiotemporal analysis are more flexible and precise than the structure of time geography. The basic research unit of time geography is a path. Conversely, LBSN data is based on a point with a timestamp and a location. Naturally, LBSN data can support the analysis of various temporal and spatial scales. Time geography, as a model, only fits for describing patterns of routine individual mobility. It is not an ideal model to study abnormal mobility and spatial aggregated mobility.

Nevertheless, the outstanding contribution of time geography is that it takes people as a meaning but also a matter. "Society is not only a set of minds and intangible roles and institutions in interaction(Hägerstrand, 1989)." The spatiotemporal human movement is the entity of these social relations and interactions. The appearance of a person at a space-time location has a purpose and is constrained by his/her socioeconomic conditions.

3.2. Understanding LBSN behaviors: uses and gratification theory

This section discusses the motives that people use social media applications for understating the online behaviors and bridging the gap between the virtual and real world. In concrete, several critical questions are helpful to deconstruct the puzzle:

6. What motives do people decide to use LBSN applications?
7. What motives do LBSN users share their locations and other personal information?
8. Why do people choose different LBSN applications?

Without a doubt, the mechanism of LBSN uses is multiplex. "No single factor is theorized to drive media use; it is the interaction among needs, individual differences, and social context that predicts use"(Lucas & Sherry, 2004). It does not exist a universal theory that is able to fully explain the acceptance of LBSN applications and motives of LBSN behaviors. However, as we stated in the definition of social media, social media is of attributes of mass media, social functions, and technological products. Therefore, LBSN uses could be investigated

through two widely used approaches: the uses and gratifications theory(Diddi & LaRose, 2006; Katz, Blumler, & Gurevitch, 1973; Whiting & Williams, 2013)and the information technology acceptance research(Frigui, Rouibah, & Marzocchi, 2013; O'cass & Fenech, 2003). The former one interprets the LBSN as a new form of mass media and electronic productions. The latter explores the acceptance and adoption of LBSN.

3.2.1 Why does uses and gratifications (U&G) theory become popular?

Before the innovation of LBSN applications, the two approaches have had a plethora of case studies. The research of information technology acceptance focus on why and how people accept new technology(F. D. Davis, 1993; Rondan-Cataluña, Arenas-Gaitán, & Ramírez-Correa, 2015; Shih, 2004). Venkatesh, Morris, Davis, and Davis (2003) did a completed literature review of users 'decision-making of information systems and developed a unified model to explain the reason that employees accept an information system. However, as the exponential increase of LBSN users, the study of information technology acceptance seemed to fall behind the development of the real world.

By contrast, uses and gratifications (U&G) theory explores the motives behind the " media consumption or non-media-based activities"(Lucas & Sherry, 2004). The analysis of audiences' motives and choices for mass media could be date back to the study of radio audience in 1935(Cantril & Allport, 1935) and uses of newspaper (Berelson, 1949). Compared with information technology acceptance, uses and gratifications (U&G) theory is more flexible and adaptable to the study of LBSN behaviors. Firstly, information technology acceptance mainly confines to work scenarios where employees consider to use the software or not, and thus it does not explain the complexity of individual' s behaviors and choices beyond workplaces. Whereas, considering the media attribute of LBSN, scholars can modify the U&G model quickly and apply it to various LBSN applications.

Secondly, with the widespread use of computers and smart devices, a user may adopt many LBSN applications in his/her devices for different motivations. For example, Instagram is for sharing photos of daily life; and LinkedIn aims to

build work connections. Essentially, these applications are created for satisfying human needs. Just as Google's co-founder Larry Page said, "the perfect search engine would understand exactly what you mean and give back exactly what you want.[3]" Such a motive-driven philosophy is also shared by most designs of LBSN applications. Therefore, U&G theory is an ideal tool for analyzing the uses and LBSN.

3.2.2 Philosophy and methodology of uses and gratification (U&G) theory

Strictly speaking, U&G theory is more close to a systematic approach rather than a rigorous theory. First of all, models are transforming with the progressing of technologies, and the variables of models are changing with particular media and social environments. It lacks a completed list or catalog of human needs and gratifications. Moreover, the absence of a completed relevant theory is able to map each need to a particular behavior. Each behavior or decision is a continuum result of this person's past, current status, and circumstances. However, it is possible to investigate the issue from a reversed direction, *i.e.* from "needs – gratifications" to "gratifications - needs". Hence, the philosophy of the U&G approach starts from the function of a media to connect with needs(Katz et al., 1973) . For example, sharing location may satisfy the need of self-representation or social interaction. Watching a video may be correlated to relax or information seeking.

Katz et al. (1973) summarized five basic assumptions of U&G theories: 1) the audience plays an active role during media communications; 2) the audience has an initiative in media choice for satisfying needs; 3) the media only meets partly human needs, and the degree of gratification of these needs varies from different media; 4) the methodological cornerstone of U&G theory is that people have sufficient self-awareness to declare their interests and motives in particular situations; 5) the gratification theory should exclude the cultural value judgment of mass communication.

[3] Source: https://sites.google.com/site/jurgensencompositionprojectweb/about/philosophy

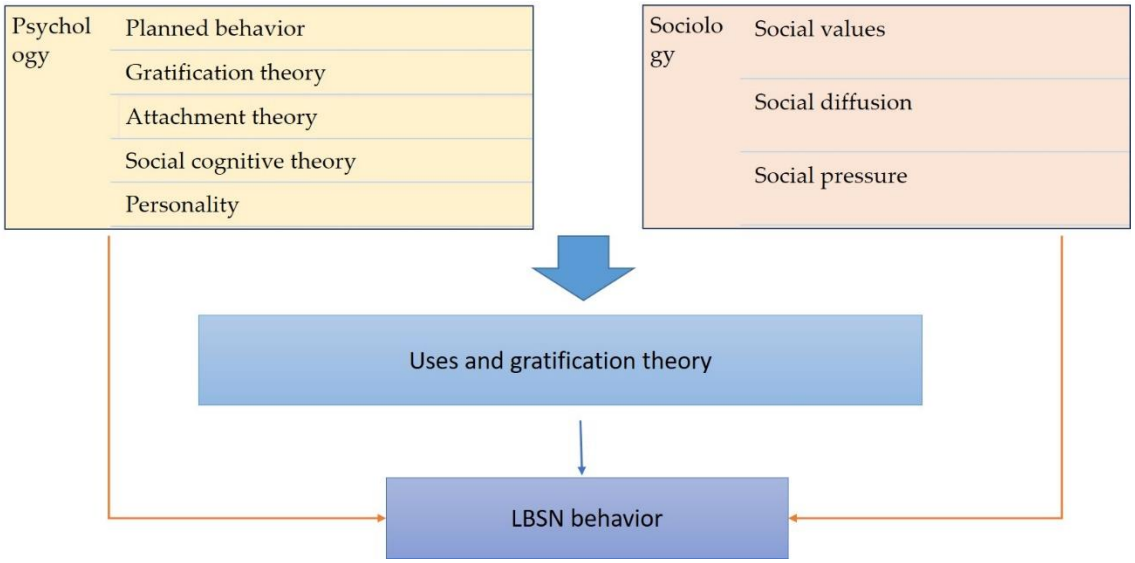
These assumptions delimitate the subject and study scale of U&G theory, which concentrates on personal choices of mass media rather than a macro-narrative of the society. Firstly, the subject of the research is the audience that consumes mass media, though early scholars primitively focused on categorization of audiences' responses(Ruggiero, 2000). The active role of audiences indicates that people can actively choose the media for satisfying needs. Secondly, the exclusion of cultural values and the user-centric perspective limits the scope of study on the relationship between user's motives and mass media.

Regarding theoretical origin, previous researches are central to several linked but distinct theories. In general, psychology and sociology contribute to the theoretic foundation of the U&G approach (**Figure 9**). Meanwhile, some scholars also use the two theories to investigate LBSN uses(Amichai-Hamburger & Vinitzky, 2010; L. Chen, Hu, Shu, & Chen, 2019; Haridakis & Hanson, 2009). For example, Hughes, Rowe, Batey, and Lee (2012) investigated the correlation between personality and social media (Facebook and Twitter) usages. They found that personality was indeed correlated with online social interactive and information seeking, though the influence was not decisive. Yoo, Choi, Choi, and Rho (2014) studied the impact of social appearance (a person's public image) and social conformity (a form of social pressure) on Twitter user's behaviors in South Korea. The result showed that the social conformity had a positive correlation with the frequency of Twitter use. Meanwhile, the social appearance would affect the trustworthiness of Twitter information.

However, neither of them could conceive of a unified or completed model to analyze all influential factors. Moreover, such a social or personality influence cannot reach a rigorous measurement because it is impossible to calculate the precise proportion of these factors in the behavior. The impact of these factors would change as input variables changed. Furthermore, the causal effect is unclear when we try to build connections between the abstract and macro concepts and the particular individual behaviors. For example, in the case of Yoo et al. (2014) , social conformity leads to the increase of consuming Twitter through users' perceived values, such as utilitarian values, hedonic values, and social capital values. However, it is debatable that to what extent these values are connected

with social conformity. It is also arguable that whether social conformity and social values still have an evident impact on the use frequency of Twitter when other variables enter the model, such as personal addiction and their professions.

In summary, it is not difficult to postulate the hypothesis of such relationships theoretically because each person has his/her social constraints, due to people's social attributes. However, the obscurity of these relationships should be recognized. It causes that models of U&G are individualistic and discrete. It could be called the "continued flaws in U&G theory".(Ruggiero, 2000)



Source: own elaboration

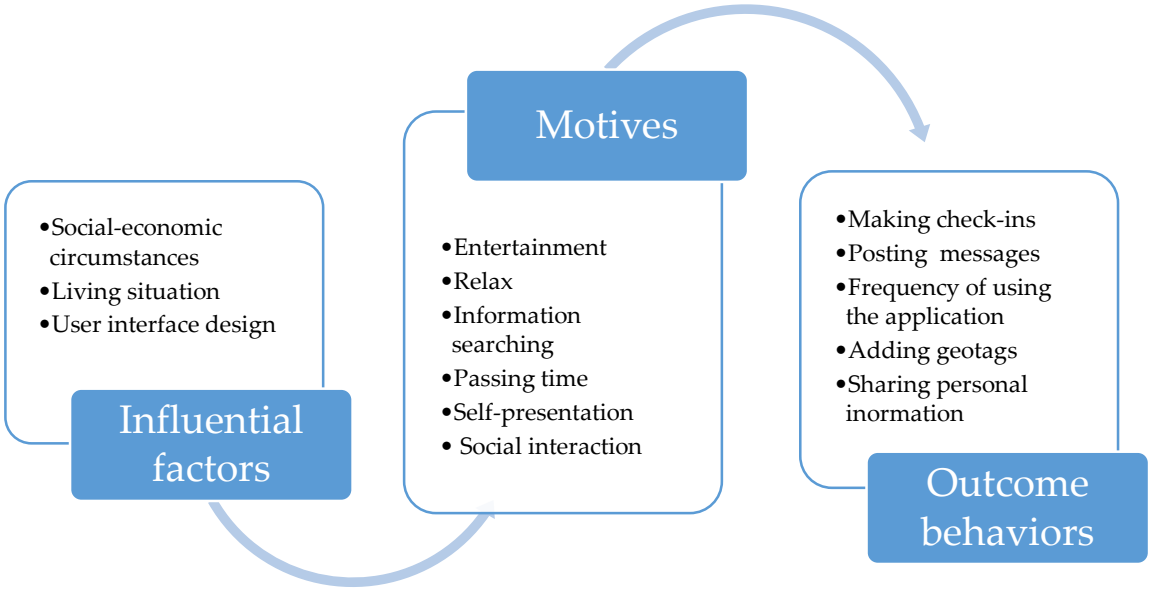
Figure 9 Approaches of LBSN behavior investigation

Despite these perceived imperfections, the U&G approach keeps its cutting-edge position in the era of LBSN. On the one hand, nowadays, people have more and more choices on social media productions for various purposes. Therefore, on a certain level, the analysis of gratifications and motivations is the most precise and direct way to understand users. On the other hand, the heuristic paradigm of U&G models leaves enough space for improvement and adaptation. For example, Kircaburun, Alhabash, Tosuntaş, and Griffiths (2020) included personality traits into their U&G model for predicting problematic social media uses (PSMU)among

college students in Turkey. The result indicated that PSMU was correlated with several demographic and personality characteristics, such as being female, introverted and less conscientious.

Thirdly, as the accumulation of these individualistic investigations, the benchmark of the U&G approach, such as the questionnaire and the typology of needs and gratification, is constantly furnishing and examining. Moreover, a plethora of worldwide studies will reveal some common conclusions of media uses. For example, the hedonic motives of using media and social media have been observed by many studies(Lindqvist, Cranshaw, Wiese, Hong, & Zimmerman, 2011; I. Pak, Teh, & Cheah, 2018; Palmgreen & Rayburn, 1979; Whiting & Williams, 2013).

Nevertheless, combining these historical reviews, we could outline the general theoretical mechanism of LBSN behaviors**(Figure 10**). The models of gratifications focus on the relationship between motives and outcome behaviors. The social-economic environment, personal living status, and software influence and form people's motives.



Source: own elaboration

Figure 10 Study structure of LBSN behaviors

For explaining the establishment of the correlation between behaviors and motives, we take the study of Palmgreen and Rayburn (1979) as an example. They created two concepts to describe the relationship: gratification sought(GS) and gratification obtained (GO). For instance, assuming the question is why people watch TV, the information seeking from TV programs can be analyzed as:

GS. I watch TV to keep up with current news.

GO. The TV program helps me to keep up with news.

Therefore, the need of information seeking is satisfied by watching the TV program. The behavior of watching TV news is correspondent to the need.

In terms of methodology, it consists of the catalog of uses and gratifications and the mathematic model. Each case may yield its own classification because they investigate different medium and users under different societies. For example, Rubin, Perse, and Barbato (1988) developed six motives for interpersonal communication: pleasure, affection, inclusion, escape, relaxation, and control. LaRose, Mastro, and Eastin (2001) formulated a U&G model to explain the Internet usages leveraging social cognitive theory, which includes ten indicators: expected activity, pleasing sensory, novel sensory, social outcome expectations, expectations of negative Internet outcomes, self-efficacy, self-regulatory, self-disparagement, self-slighting, and self-perceptions of Internet addiction. The following section is going to discuss these classifications in detail.

Concerning the mathematical model, regression analysis is the common method for testing various relationships, including classical multiple linear regression(LaRose et al., 2001), hierarchical regression(Kircaburun et al., 2020), Partial least square regression(Liu, Cheung, & Lee, 2010; Palmgreen & Rayburn, 1979; Thompson, Wang, & Daya, 2019; Yoo et al., 2014). Among them, Partial least square regression (PLS) is the most popular statistical approach for U&G model. It is a good method to correlate two matrix of variables (i.e. the dependent variables and independent variables)(Abdi, 2003); and is an alternative of the classical linear regression due to its robustness(Geladi & Kowalski, 1986). However, it only applies to small sample size.

3.2.3 Taxonomy of gratifications on LBSN uses and behaviors

For better understanding U&G models of LBSN uses, this dissertation summarizes several representative models of gratification, along with the taxonomy of gratification, respondents of surveys, and the medium, are displayed in **Table 1** to facilitate a ready comparison. We also traced some media uses in earlier epochs and placed them in the table. It revealed that human needs did not change a lot regardless of the evolution of technology.

Table 1 Taxonomy of gratifications on media uses

Authors	Respondents	Country	Medium	Taxonomy of gratifications
Palmgreen and Rayburn (1979)	A random sample of telephone interviews	US	Television	1) Relaxation 2) Learning about things 3) Communicatory utility 4) To forget 5) To pass time 6) Companionship 7) Entertainment
James, Wotring, and Forrest (1995)	Random users from bulletin board services(BBS[4]) - CompuServe and Prodigy	US	Electronic Bulletin Boards	1) Entertainment/Interest 2) Info/Education 3) Business 1) Communication medium appeal 2) Socialize
Perse and Dunn (1998)	National phone survey	US	Home computer CD-ROM	1) Learn 2) Entertain 3) Excite 4) Relax 5) Forget 6) Lonely 7) Busy 8) Habit 9) Friends
LaRose et al. (2001)	College students	US	Internet	1) expected activity 2) pleasing sensory 3) novel sensory 4) social outcome expectations

[4] BBS is similar to current online forums.

				5) expectations of negative Internet outcomes 6) self-efficacy 7) self-regulatory 8) self-disparagement 9) self-slighting 10) self-perceptions of Internet addiction
Lucas and Sherry (2004)	College students	Not mentioned	Video game	1) Challenge 2) Arousal 3) Fantasy 4) Diversion 5) Competition 6) Social interaction
Ko, Cho, and Roberts (2005)	College students	US, KR	Website	1) Information 2) Convenience 3) Entertainment 4) Social interaction
Diddi and LaRose (2006)	undergraduate students	US	News website	5) Surveillance 6) Escapism 7) Pass time 8) Entertainment 9) Habit strength 10) News quiz
Ames and Naaman (2007)	Random sampling	Not mentioned	Flickr (photo-tag)	1) Retrieval, directory 2) Search 3) Context for self 4) Memory 5) Contribution, attention 6) Ad hoc photo pooling 7) Content descriptors 8) Social signaling
Peters, Amato, and Hollenbeck (2007)	College students	US	Wireless advertising messages	1) Process 2) Socialization 3) Content
Haridakis and Hanson (2009)	College students	US	YouTube	1) Convenient entertainment 2) Interpersonal connection 3) Convenient information seeking 4) Escape: escape from some situations 5) Co-viewing 6) Social interaction
Liu et al. (2010)	A random sampling of Twitter users	CN(HK)	Twitter	1) Content: information sharing; self-documentation

				2) Process: online environment; passing time; self-expression 3) Social: social interaction 4) Technology: convenience of the medium 5) Overall satisfaction
G. M. Chen (2011)	A snowball sampling of Twitter users	US	Twitter	The need of contacting with others
Smock, Ellison, Lampe, and Wohn (2011)	College students	US	Facebook	1) Relaxing entertainment 2) Expressive information sharing 3) Escapism 4) Cool new trend 5) Companionship 6) Professional advancement 7) Social interaction 8) To meet new people
Lindqvist et al. (2011)	Random sampling	CA, US	Foursquare	1) Badges and fun 2) Social connection 3) Place discovery 4) Keeping track of places 5) Game with yourself
Whiting and Williams (2013)	In-depth interviews	Not mentioed	Commentary paper	1) Social interaction 2) Information seeking 3) Pass time 4) Entertainment 5) Relaxation 6) Expression opinions 7) Communicatory utility 1) Convenience utility 2) Information sharing 3) Surveillance/knowledge about others
Lien and Cao (2014)	Random sampling from an online survey network[5]	CN	WeChat	1) Entertainment 2) Sociality 3) Information
F. Wang, Wang, and Philip (2014)	Random sampling	EU	Gowalla (geotag)	1) Social 2) Individual
Sheldon and Bryant (2016)	College students	US	Instagram	1) Surveillance 2) Documentation 3) Coolness 4) Creativity

[5] https://www.wjx.cn/

Tasse, Liu, Sciuto, and Hong (2017)	Random sampling	US	Twitter (geotag)	1) Coolness 2) Keeping track of places 3) Place discovery 4) Social connection 5) Automatic geotag
Alhabash and Ma (2017)	College students	US	Facebook, Twitter, Instagram, Snapchat[6]	1) Information sharing 2) Self-documentation 3) Social interaction 4) Entertainment 5) Passing time 6) Self-expression 7) Medium appeal 8) Convenience
Blight, Ruppel, and Schoenbauer (2017)	College students	US	Instagram, Twitter	1) Entertainment motive 2) Information sharing 3) Escape 4) Trend 5) Companionship 6) Social interaction 7) Time pass
Khan (2017)	College students	US	YouTube	1) Information Seeking 2) Giving Information 3) Self-Status Seeking 4) Social Interaction 5) Relaxing Entertainment
Gwena, Chinyamurindi, and Marange (2018)	College students	International students (SA)	Facebook	1) Meet new people 2) Search for information 3) Share media 4) Maintain relationships 5) Connect 6) Entertainment 7) Discuss
Gan (2018)	College students	CN	Sina Weibo , WeChat	1) Hedonic 2) Affection 3) Information 4) Social
Hossain, Kim, and Jahan (2019)	Random sampling	Not mentioned	Facebook	1) Enjoyment 2) Passing time 3) Information seeking 4) Self-presentation 5) Social-interaction
Rauschnabel, Sheldon, and Herzfeldt (2019)	American Amazon Mechanical Turk (MTurk) users	US	Hashtags of social media	1) Amusing 2) Organizing 3) Designing 4) Conforming 5) Trendgaging 6) Bonding 7) Inspiring 8) Reaching 9) Summarizing

[6] https://www.snapchat.com/l/en-gb/

				10) Endorsing
Thompson et al. (2019)	A snowball sampling of Facebook users	Not mentioned	Facebook	1) Status seeking 2) Socializing 3) Entertainment 4) Pass time 5) Information sharing
Kircaburun et al. (2020)	College students	TR	WhatsApp, Instagram, YouTube, Facebook, Snapchat, Google+, and Twitter	1) Maintaining existing relationships 2) Meet new people and socializing 3) Make, express, or present more popular oneself 4) pass time 5) a task management tool Entertainment 6) Informational and education
Omar and Dequan (2020)	A snowball sampling of TikTok users	CN: 87.5%; other countries: 12/5%	TikTok	1) Social interaction 2) Archiving 3) Self-expression 4) Peeking 5) Escapism

Source: own elaboration. Note: 1. Country is the location of respondents rather than authors. 2. The taxonomy of gratifications is based on the model result of researches.

Although each paper yields different taxonomies of gratifications, they reported some common themes such as social interaction, entertainment, pass time, information seeking, self-expression, escapism, etc. It has to emphasis that these gratifications in the table are according to their final results. Therefore, Statistically, they are significant to the social media uses.

(1) Entertainment/ hedonic /amusing /enjoyment/ relaxation

Entertainment is the most common motive of social media use. It refers to obtaining entertainment and relaxation from social media. Korgaonkar and Wolin (1999) classified the relaxation into the theme of entertainment, though some scholars also tended to distinguish the relaxing motive from entertainment motive(Palmgreen & Rayburn, 1979; Whiting & Williams, 2013). Social media is widely used every day for entertainment drives regardless of different backgrounds and cultures(Manasijević, Živković, Arsić, & Milošević, 2016). Entertainment and convenience are the two most influential variables for predicting the intensity of social media uses(Alhabash & Ma, 2017). It is worth

noting that check-in behavior on Foursquare also belong to entertainment purpose due to the game-like design(Lindqvist et al., 2011).

(2) Social interaction/socializing

We tend to group social motives, such as meeting new people, maintaining relationships, bonding, making friends into the term of social interaction. As social media applications move people's social relations to the virtual world, social media naturally becomes one of the medium for satisfying the need of maintaining and developing relationships(Seidman, 2013). Even the online games also has the function of social interaction(Lucas & Sherry, 2004).

(3) Pass time

This theme firstly appeared in Palmgreen and Rayburn (1979)'s U&G study for television. It is defined as using a tool to relieve boredom and occupy the time(Papacharissi & Rubin, 2000). The difference between passing time and entertainment is whether the medium is considered as an active amusement provider or a tool that make time pass.

6) (4) Self-documentation/ Archiving

Self-documentation is similar to writing a diary online. Users utilize social media applications to record their daily life(Alhabash & Ma, 2017; Sinn & Syn, 2014) and keep track of something(Liu et al., 2010). Different from self-expression, self-documentation does not aim to convey information to the public, though the boundary is ambiguous to some degree.

(5) Self-expression/ self –presentation

Self-expression/self-presentation is defined as expressing someone's feelings, thoughts and opinions. LBSN activities can satisfy this motive by posting messages, commenting, sharing comments, fulfilling personal profiles(Omar & Dequan, 2020; Seidman, 2013; Whiting & Williams, 2013). Utz, Tanis, and Vermeulen (2012) mentioned the need of popularity also drove some users to

disclose themselves intentionally. Nevertheless, such a need could also be conceived as a need of self-presentation.

(6) Information seeking

Information seeking includes information on daily life, such as sales, events, and news. Self-education is also a type of information seeking. In the current, information seeking is more evident in YouTube uses(Haridakis & Hanson, 2009; Khan, 2017).

(7) Information sharing

The theme includes two layers: providing original information, sharing useful or personal interested information from other sources(Liu et al., 2010). This motive actually grows with the development of social media due to its interactive nature. Users are the media and the audience at the same time(Nov, Naaman, & Ye, 2010).

(8) Surveillance/knowledge about others

According to the investigation of Whiting and Williams (2013), many individuals reported that they wanted to know other's updates would keep up with others. Several behaviors are related to the motive, such as "see other people's status," "to like," and "spy on people"(Sheldon & Bryant, 2016).

(9) Convenience / medium appeal

Convenience and medium appeal can be conceived as technology gratification(Liu et al., 2010). Convenience means that the application is easy to use and access anytime and anywhere(Whiting & Williams, 2013). Medium appeal is the attractiveness of the medium that can let people continue using it.

(10) Escapism

Escapism is a motive to utilize social media to forget or avoid of some situations of the real world(Blight et al., 2017; Palmgreen & Rayburn, 1979). It does

not play a significant role in the whole LBSN uses. However, it is the most consistent factor of predicting news consumption behavior(Diddi & LaRose, 2006).

In summary, firstly, the LBSN behaviors may be associated with multiple human needs, verse versa. For example, watching others' updates possibly meets the need of entertainment, social interaction, or passing time. Conversely, self-expression may be accomplished by posting photos, leaving comments, or using hashtags. Therefore, it is impossible to build connections between each behavior and need precisely.

Secondly, regarding the gratifications across platforms, entertainment takes the dominant position(Khan, 2017).Alhabash and Ma (2017) investigated four popular social media applications and examined the difference of motives among them**(Figure 11**). In general, entertainment and convenience are the most significant gratifications. Social interaction was not as important as people usually assumed. However, participants of their study were college students, and hence needs might be different from other groups of people.

	FACEBOOK	TWITTER	INSTAGRAM	SNAPCHAT
1	Convenience	Entertainment	Entertainment	Entertainment
2	Entertainment	Convenience	Convenience	Convenience
3	Passing Time	Medium Appeal	Medium Appeal	Medium Appeal
4	Medium Appeal	Passing Time	Passing Time	Passing Time
5	Information Sharing	Self-Expression	Self-Expression	Self-Expression
6	Self-Expression	Information Sharing	Self-Documentation	Self-Documentation
7	Social Interaction	Social Interaction	Social Interaction	Social Interaction
8	Self-Documentation	Self-Documentation	Information Sharing	Information Sharing

Source: Alhabash and Ma (2017)

Figure 11 Ranking of use- motives among Facebook, Twitter, Instagram, and Snapchat

Thirdly, as most investigations focus on college students (see **Table 1**), it may lack enough general representativeness of all groups of people. Nevertheless,

some studies (Hossain et al., 2019; Lien & Cao, 2014; Omar & Dequan, 2020; Thompson et al., 2019) utilized a random sampling to conduct the analysis , which obtained similar results of motives.

The importance of U&G approach lies in that it provides a practical perspective to investigate why people use LBSN applications. It focuses on the direct relationship between LBSN uses and human needs.

With respect to studying spatiotemporal behaviors through LBSN data, these investigations prove that LBSN data is capable of depicting reality to some degree. Human activities in the physical world reflect on the LBSN data. For example, the self-documentation indicates that some users would like to disclose themselves online. Therefore, these data are not isolated from the real world.

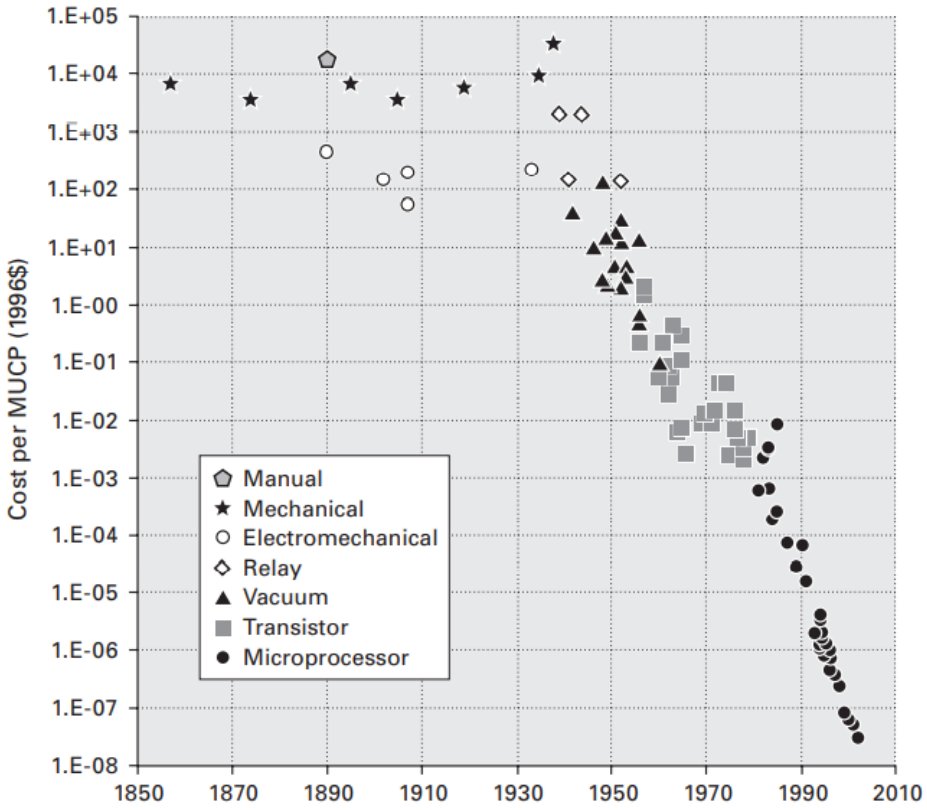
Secondly, the motive of geotagging behavior explains why people record their geo-positions. It seems that no difference with other LBSN functions: self-expression(Ames & Naaman, 2007), social interaction(Lindqvist et al., 2011; Tasse et al., 2017), entertainment(Lindqvist et al., 2011), self-documentation(Ames & Naaman, 2007). In general, "people usually consciously geotag, though a significant portion geotags unintentionally"(Tasse et al., 2017).It is not a rich data source to detect all aspects of daily life. It is not a rich data source to detect all aspects of daily life. Scholars should aware of possible bias and design the proper scope of the study.

3.3. Understanding spatiotemporal analysis: data and models

3.3.1 Data march

Critical studies of spatial information technologies have been largely outdated because mushrooming of technologies that continue improving their precision and coverage toward the quotidian life. In fact, the theorists "are contending with the technological present"(Leszczynski, 2015). **Figure 12** displays the increase of efficiency of computation in terms of cost and materials. The decrease in costs means that more people could afford the price of electronic devices. Meanwhile,

the volume of computer processors becomes smaller, and thus portable devices of high capacity become available.



Source: Kitchin and Dodge (2011). Note: MUCP: million units of computation.

Figure 12 Increasing cost efficiencies of computation with a marked step change from mechanical to electronic processing technologies

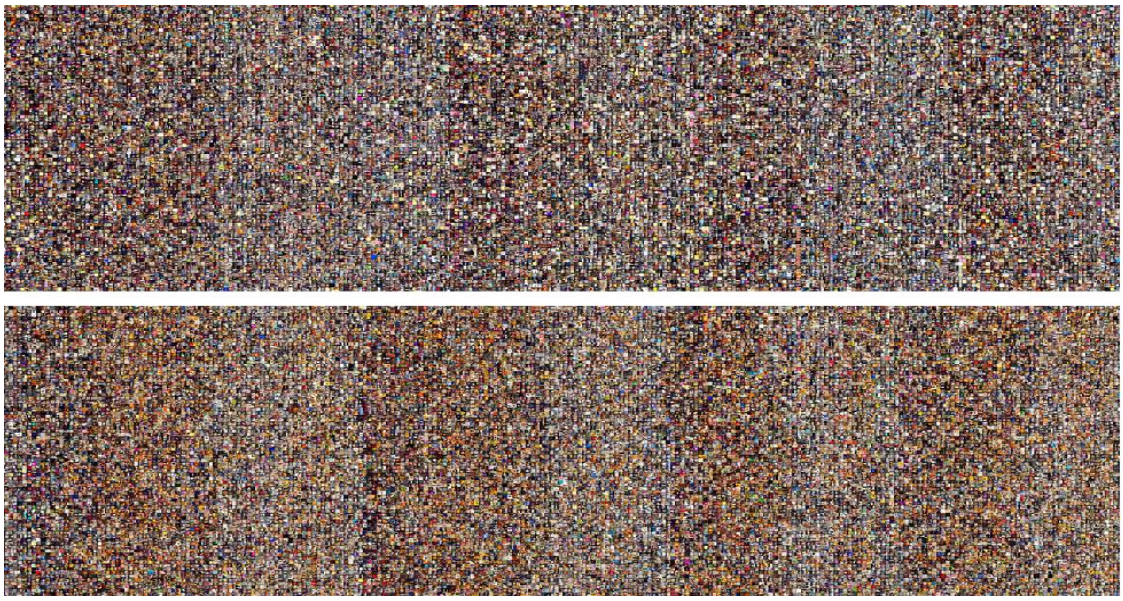
Therefore, the large –scale spatiotemporal behaviors becomes accessible for the first time in human history. An object's movement is represented as a group of three simple attributes: (longitude, latitude, timestamp). It also ends the debate about the ontology of spatiotemporal behavior(Raper, 2005), i.e. what is the best representation of it. For example, **Figure 13** displays one frame of the movement of the global merchant fleets in 2012. The original dataset contains ships' geographical locations and speeds. It shows that the enormous georeferenced data have already reached high precision and could track spatiotemporal behaviors within a long period.

Source: https://www.shipmap.org/ , captured date: 03/06/2012

Figure 13 Movement of global merchant fleets

Moreover, they are not reductive symbols of movements. On contrary, the digitalization of places and human movements connect particular social attributes with the corresponding places. For example, people can know about a restaurant's style, comments, popularity, and average price through Google Maps or Tripadvisor[7]. These data can also entail different spatiotemporal patterns of places and how people use these places. Hochman and Schwartz (2012) aggregated 550,000 photos from Instagram to show the spatiotemporal "visual rhythm ": the pattern of photo brightness varied along the time of the day(**Figure 14**).

[7] https://www.tripadvisor.com/

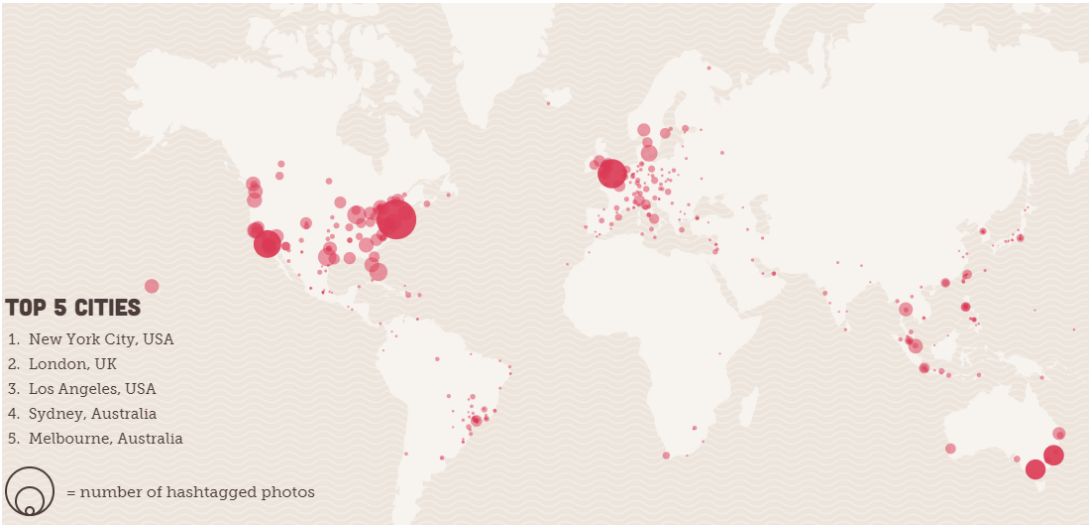


Source: Hochman and Schwartz (2012).

Figure 14 Visualization of a four-day's sample: above: New York, below: Tokyo

Meanwhile, they "more fundamentally mediate the everyday practices of urban life, subtly shaping senses of place as particular interpretations of events and locations are foregrounded or side-lined"(Graham, Zook, & Boulton, 2013). For example,

Figure 15 displays the distribution of bacon-tags on Instagram ,which provides an interesting perspective to sense the world and understand the food preferences of different places.



Source: https://cewe-photoworld.com/instagram-food-capitals/

Figure 15 The distribution of bacon tags

3.3.2 Geo-spatiotemporal models

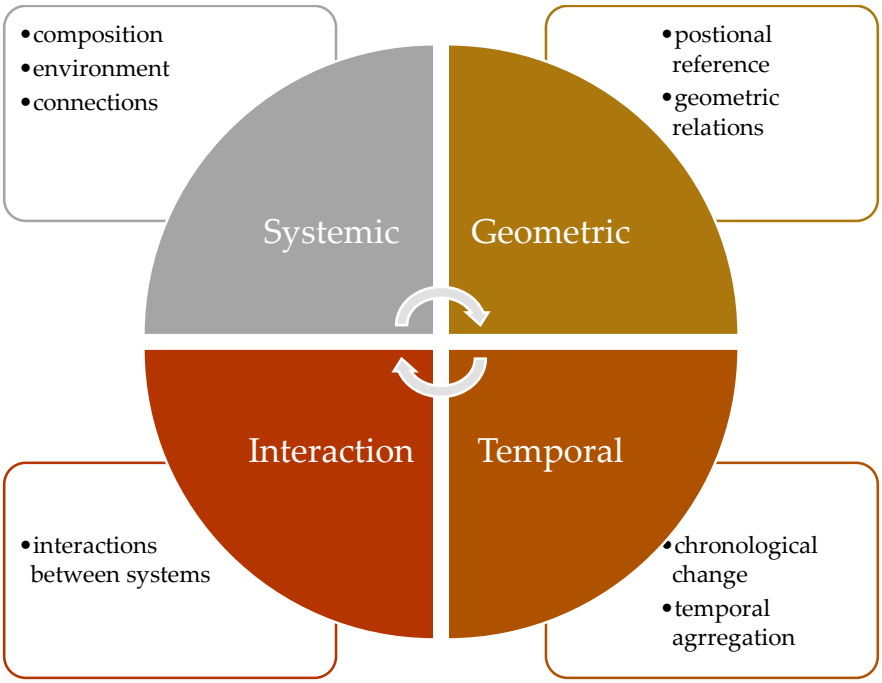
Why model? Hestenes (1997) defined a physical model as "a representation of structure in a physical system and/or its properties". A model is not only for predictions, but it is also designed to explicitly explain and suggest a system(Epstein, 2008). In essence, the model of "Big data" is partial to statistics rather than mathematics. The result is a summarization of data and inevitably contains some degrees of distortion against the real world. However, combined with the huge volume of data, models can help us to catch the key parts of things and provide a replicable approach to manipulate data. They also demonstrate great potentials to measure the complex society and raise new questions(Tonidandel, King, & Cortina, 2018).

As to geo-spatiotemporal analysis, models aim to depict the characteristics of spatiotemporal behaviors(Béjar et al., 2016; F. Luo, Cao, Mulligan, & Li, 2016), establish connections with other types of data (*e.g.* demographic and epistemic data), and finally, to disclose some properties of the world(W. Luo et al., 2019; Samani, Karimi, & Alesheikh, 2020).

Although it is difficult to summarize a general model of spatial-temporal analysis, Hestenes (1997) actually provides an excellent structure of model for spatiotemporal models (**Figure 16**). It includes four types of structure: systemic, geometric, temporal, and interaction. The systemic structure is the outline of the model, which includes composition (internal variables), environment (external variables) and connections (all interactions among variables). The geometric and temporal structure is easier to be transplanted as the spatial and temporal relations. The interaction structure indicates the relationships among variables and systems, which can be represented by functions or diagrams.

For example, a model of tourist's spatiotemporal behaviors in a destination consists of the scale of movement, speed, and transportation methods. The

environmental variables include weather, landscapes, type of road, travel purpose, etc. The temporal structure is the timeline of the movement trace. The spatial structure is represented by a series of geo-coordinates of the trace. The interaction structure could be a regression model that depicts the relationship between the speed of the tourist's movement and the external variables.



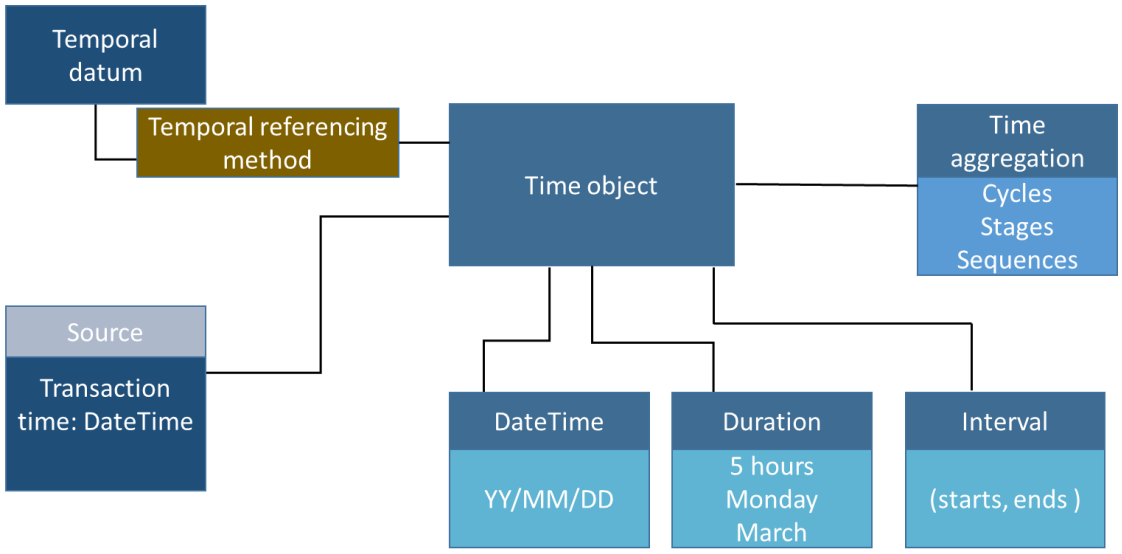
Source: reproduced and adapted from Hestenes (1997)

Figure 16 Specification of geo-spatiotemporal model

According to different goals of studies, the model could only contain part of the four structures. For example, the visualization of spatiotemporal behaviors (Hochman & Schwartz, 2012; Yaqub et al., 2020) does not need to consider the interaction structure.

Besides analytic models, the significance of models lies in the construction of the database. It can help us to collect data, define the attribute and the object of the

database, and manage the data. For example, **Figure 17** explains a conceptual model of building a time object.



Source: adapted from Koncz and Adams (2002)

Figure 17 Construction of time object

In computer language, an object is an "abstract data type with the addition of polymorphism and inheritance[8]." In simple words, an object has data and functions. The time object is created to record date time and deal with various temporal structures, such as a timeline or an interval. It starts from the temporal datum that is the Universal Time Coordinated (UTC) and the Gregorian calendar. It could be considered as the "internal" time that is the temporal reference.

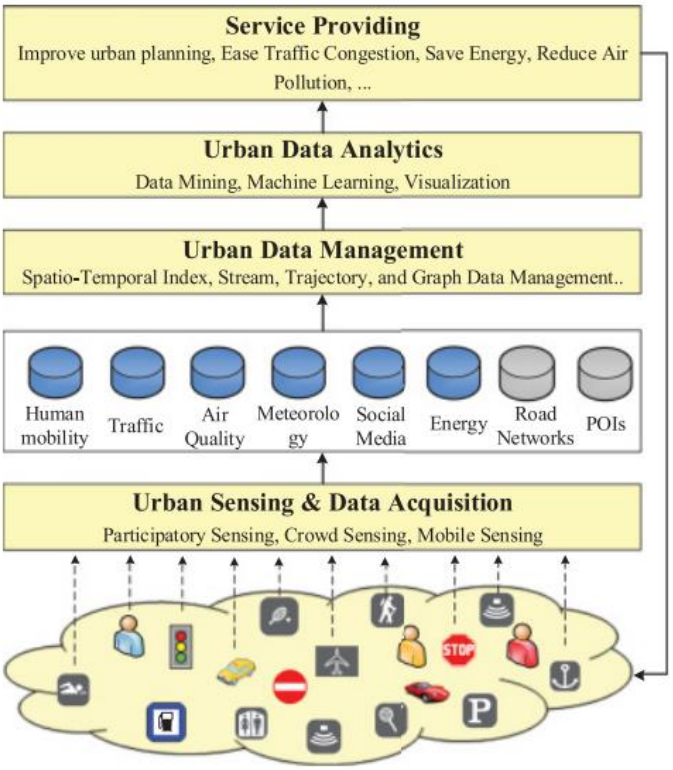
On the other hand, the time object can also receive temporal data from the outside. It collects data when activity happens. The occurrence time is the transaction time that constructs the data source.

The time object can be represented as the DateTime, time interval, and duration. The DateTime follows the regular calendar format, such as "year – month – day" or "day –month -year". The interval means the length of them between two timestamps. The duration is a time span, such as five hours, Monday,

[8]https://en.wikipedia.org/wiki/Object_(computer_science)#:~:text=An%20object%20is%20an%20abstract,found%20in%20the%20real%20world.

or March. Time object can also be aggregated into time circles, stages, and sequences.

Synthesizing data, data management models, and analytic models, spatiotemporal analysis has expanded to many urban study fields. Y. Zheng, Capra, Wolfson, and Yang (2014) proposed a new concept – "urban computing" to describe the new interdisciplinary field(**Figure 18**) :"Urban computing is a process of acquisition, integration, and analysis of big and heterogeneous data generated by diverse sources in urban spaces, such as sensors, devices, vehicles, buildings, and humans, to tackle the major issues that cities face."



Source: Y. Zheng et al. (2014)

Figure 18 Framework of urban computing

Spatiotemporal analysis leveraging LBSN data is included as a branch of the urban computing analysis. It indicates that LBSN data could cooperate with other types of data for completing urban researches.

Chapter III.

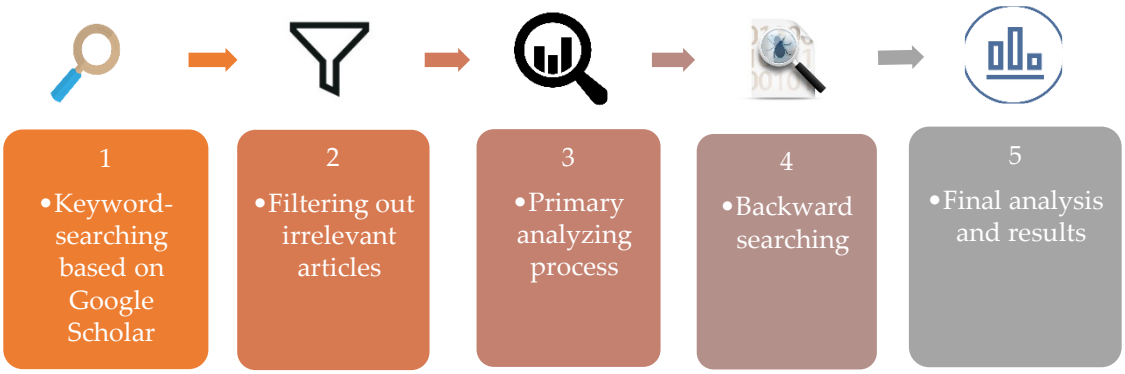
Methodology

This section introduces the general framework that describes the purposes and research logic. Firstly, the literature review plans to make a systematic survey of LBSN-related urban researches, focusing on the major limitations and applications. The section of applications summarizes popular applications of urban issues that utilize LBSN data as their data sources. The summary of limitations of LBSN data is a significant complementary to existing applications. Otherwise, the credibility of the usefulness of LBSN data is still in the dark. The review of limitations discusses major bias of LBSN data factors that affect the representativeness of LBSN data.

Secondly, the empirical study includes three independent cases from Beijing and Barcelona. We follow the "data-driven" paradigm to construct them. The starting point of the case study is the collected data, and thus each case study has a distinct data source and methodology. We will present them in chapters of empirical studies separately. Therefore, this chapter focuses on the macro-framework of case studies that consists of the general processing of LBSN data, characteristics of LBSN data, and an overview of case studies.

III.1.Methodology of the literature review

Following the guideline of snowballing literature studies(Wohlin, 2014), both literature reviews follow the same procedure of reviewing process (**Figure 19**): two rounds of collection of related articles and a quantitative analysis. Utilizing Google Scholar, a non-systematic keywords search is conducted for retrieving articles. These keywords are based on previous experiences of case studies, such as "urban ", "LBSN"," urban mobility", and "Twitter" (**Table 2**).



Source: own elaboration

Figure 19 Procedure of literature review

Table 2 List of keywords searching via Google scholar

Lists of keywords searching
LBSN +Urban; LBSN +urban structure; LBSN +public perception; LBSN +urban sentiment; LBSN + urban mobility; LBSN + urban studies; LBSN + urban planning; Twitter +public perception; Twitter +spatial structure; Weibo+ urban mobility; etc.

Source: own elaboration

The second step is to select relevant articles which require articles should utilize LBSN data and devote to analyze urban issues. For example, social diffusion(Kossinets & Watts, 2009) and computer recommendation system(Narayanan & Cherukuri, 2016) are not in the scope of our study. **Table 3** lists criteria that we select articles. Since urban spatiotemporal analysis is multidisciplinary, we do not set limits on the type and ranking of journals. However, the quality and novelty of the research is considered. Therefore, we set the third criterion to guarantee the quality of paper, and we also manually inspect

the article. The primary analyzing process will delimit the analyzing structure of literature review and primitive results.

Table 3 Criteria of selection of academic articles

1	Researches are directly related to urban spatiotemporal analysis: urban mobility, urban social and economic studies, urban health and well-being, public perception, public space, etc.
2	Researches utilize LBSN data to conduct analysis: Twitter, Foursquare, SinaWeibo, Instagram, etc.
3	Researches should be academic studies that have been qualified by journals or academic institutions.

Source: own elaboration

After that, in order to enrich the data pools of the analysis, we append articles that are from references of reviewed papers, which is so-called backward searching. Combined with the result of the two rounds of article searching, a statistical description will be produced for illustrating the general situation of the urban spatiotemporal analysis. Moreover, selected representative researches will be further introduced and summarized in each part of applications. As comparatively fewer papers specifically investigate the bias and limitations of LBSN data, the limitation part attempts to integrate the mentioned or partly investigated bias into several discussions of the representativeness of LBSN data.

III.2.Framework of the empirical study

2.1. LBSN data: types, components, characteristics

In the current, LBSN services can be classified into three types according to different focuses: human-centric, place-centric, and trajectory – centric. Both human-centric and place-centric LBSN enable users to upload information with a

geotag. However, human-centric LBSN applications, such as Twitter or Instagram, concentrate on the social and individual information rather than information of the place. Users publish their posts or comments at first and then share their geo-tags selectively. Moreover, the geotag can be the exact position where the user is posting or the position that is related to the content. For example, an Instagram user can publish pictures and add the location according to their intentions. The added geo-location can be the place where they are located or where the picture was taken. In this sense, geo-information serves the content and media.

By contrast, place-centric LBSN applications, such as Foursquare or TripAdvisor, pay attention to places. Users can create, search, and make "check-ins" and comments on venues where they are interested. These applications can provide useful information about places, such as restaurants, hotels, and buildings. Meanwhile, users are encouraged to make check-ins and give suggestions to places, such as the price, services, and their quality. The number of check-ins and comments can indicate the popularity of the place. Therefore, the place is the dominant element of these applications. Trajectory-centric LBSN services focus on tracking users' movement. For example, many sports applications help users to record their running routes and time, which have a higher accuracy of positioning and less social functions.

Table 4 measures the three types of LBSN data from seven aspects. Trajectory-centric data usually have a higher privacy level than others, and thus the availability of the data is the lowest. Therefore, most of the academic researches focus on human-centric and place-centric LBSN(Candelieri & Archetti, 2015; Cho, Myers, & Leskovec, 2011; F. Luo et al., 2016). Concerning the timeliness, both of them can support the instant location of the message. Human-centric LBSN data delivers richer social information than place-centric. However, the place-centric LBSN data has better accuracy about places. Nowadays, many applications are connecting interactively, and thus the human-centric and place-centric data can cooperate better than before. For instance, Foursquare support users to share their check-ins to Twitter at the same time. Therefore, it provides opportunities to combine Twitter data and Foursquare data in a precise way.

Table 4 Characteristics of different LBSN data

Characteristics	Types of LBSN data		
	Human-centric	**Place- centric**	**Trajectory-centric**
Focus	People	Place	Movement
Sociality	High	Medium	Low
Timeliness	High	High	Low
Semantic/ image information	Rich	Medium	Low
Precision of positioning	Low	High	High
Privacy	Low	Medium	High
Data availability	High	Medium	Low

Source: own elaboration

This dissertation also only pays attention to the human-centric and place-centric LBSN data. In general, the collected LBSN data consist of five primary attributes (**Table 5**): the unique identification, the timestamp, contents, geo-information and the social relationship. In a database, they are so-called "metadata". Each object/ observation contains different values of these attributes. Correspondingly, each object is represented as a point on the map (**Figure 20**).

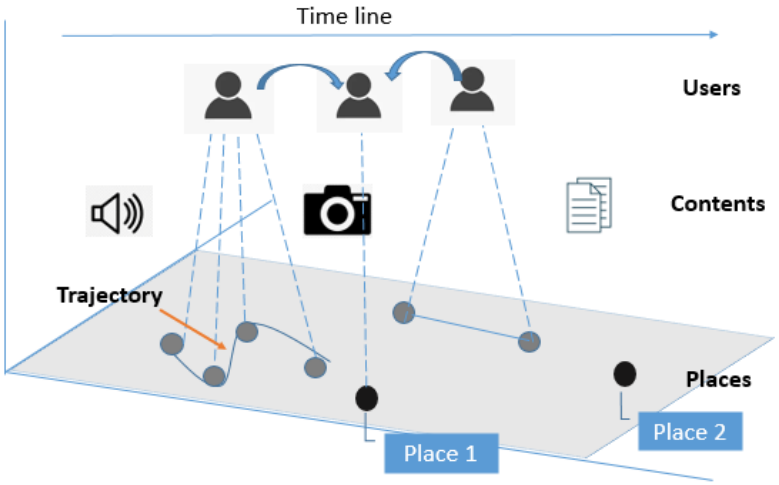
Table 5 Components of LBSN data

Attributes	Description
Identification	The unique identification of a user or a venue on the platform
Timestamp	The created time of the data
Semantic/ image contents	Texts, posted pictures or videos that are usually stored by its uniform resource locator(URL)
Geo-information	The geo-coordinates of the post, postcode and country
Social relationship	Following people and followers on the platform

Source: own elaboration

The unique identification of a user or a place ensures that they can be monitored. The timestamp provides the precise time that the observation was created on the platform. The content-attributes contain information about places, users, and the meanings behind users' behaviors, such as motivation, sentiments, opinions. Location is a significant feature for both users and places. The location of an object indicates the static position in which it is. In general, the location can be described by the coordinate system (e.g. latitude and longitude) and geo-semantic symbols (e.g. country, street, postcode).

In the case of LBSN data, the location of an observation can be revealed by multiple geo-information: geo-coordinates, postcode, country, city, users' profile, temporal habits, and languages. The social relationship can be simply described as the following people and followers of users. People can form a tie of a relationship without the necessity of having met with each other. However, it plays a crucial role in measuring the influence of the user, as a media, on the Internet. Such power sometimes can generate a great impact on individuals' life and society politically and economically. What's more, social relationships are correlated with people´s travel behaviors: 10% - 30% of human movements can be explained by social relationships (Cho et al., 2011).

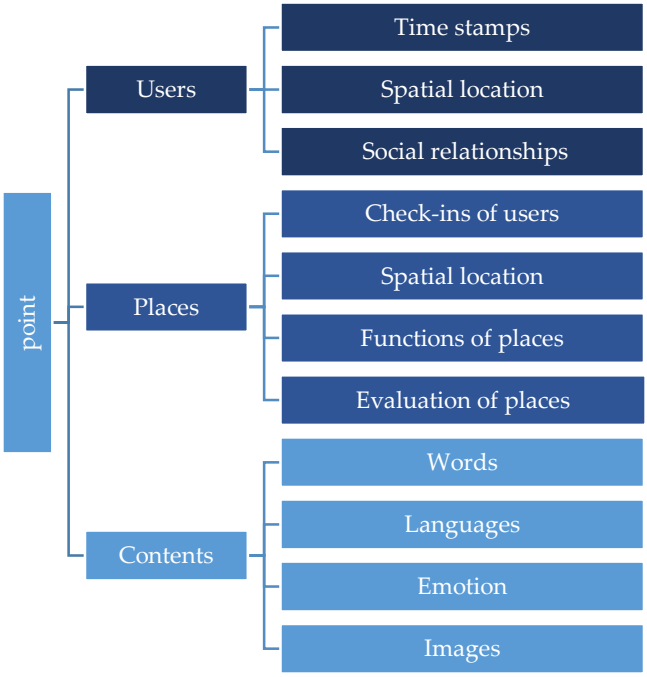


Source: own elaboration

Figure 20 Graph representation of LBSN data

All these attributes can generate valuable information for addressing different urban issues. According to different goals, these points can form into paths, clusters, and graphs. For example, we can extract the historical path of a user based on the chronological order for detecting the urban mobility. The distribution of a type of places, such as restaurants or hotels, can be aggregated using a spatial unit. Users' followers could form a graph to understand users' social networks.

Furthermore, based on these information, we can extract core features of LBSN data in terms of three perspectives: users, places and contents (**Figure 21**). With regard to users, the investigation of social relationships has the high risk of privacy issue (J. Li et al., 2020), and thus the spatiotemporal behaviors are the main scope of study.

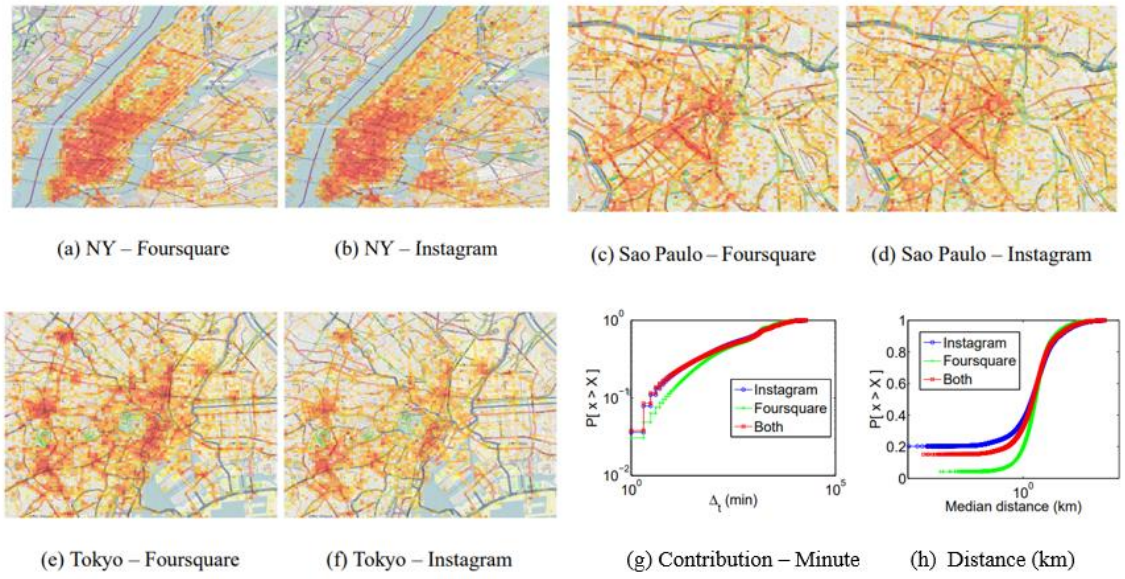


Source: own elaboration

Figure 21 Core features of LBSN data

Based on timestamps, the temporal behaviors indicate users' behaviors on the LBSN platform, such as the posting frequency and patterns. For example, Silva, Vaz de Melo, Almeida, Salles, and Loureiro (2013)analyzed user's behaviors on Instagram and Foursquare in New York, Sao Paulo, and Tokyo(**Figure 22**)in terms

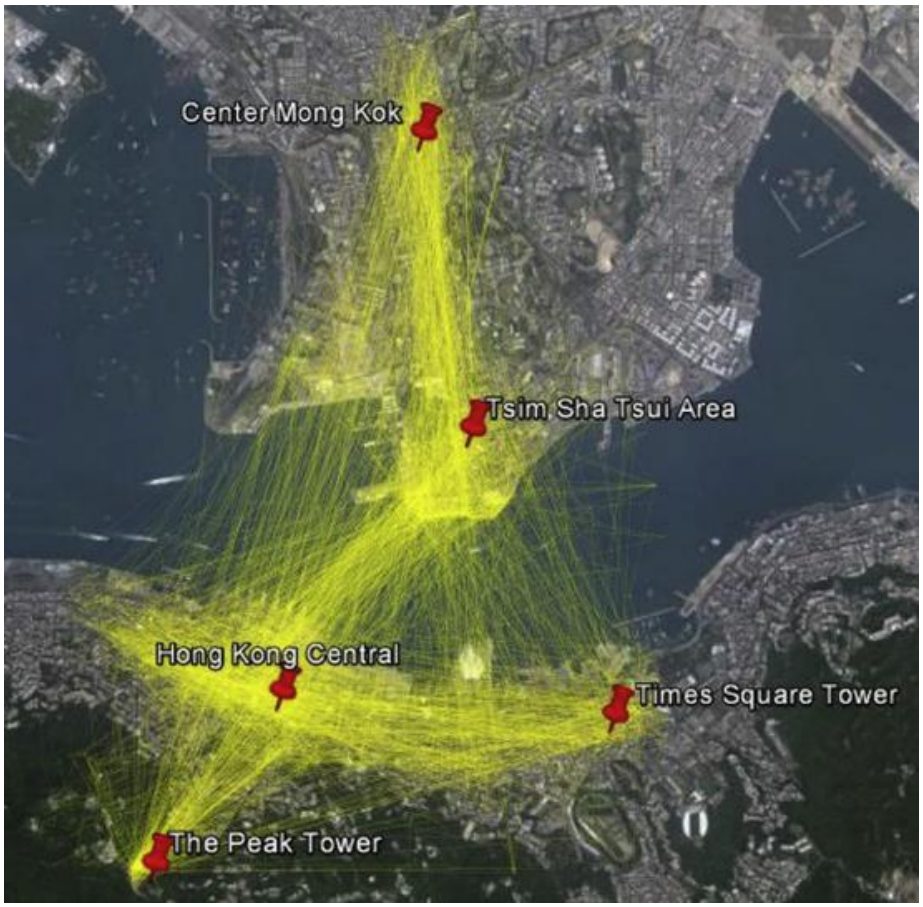
of the posting frequency and distance of two consecutive post**s** (**Figure 22** (g)(h)). They concluded that users of Instagram tended to share more contents at the same place than Foursquare users.



Source: Silva et al. (2013)

Figure 22 Spatial distribution and temporal behaviors of Instagram and Foursquare users

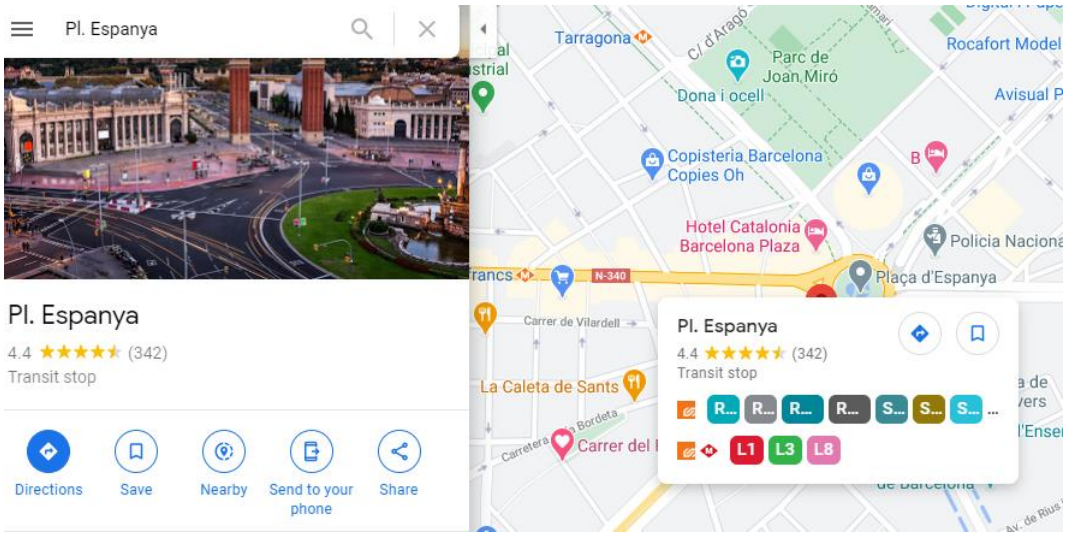
The spatial movements of users can be extracted from timestamps and the corresponding locations, which can disclose part of the urban mobility and urban spatial structure. For example, Vu, Li, Law, and Ye (2015) combined GPS data and Flickr to show tourists' main routes in Hongkong(**Figure 23**).



Source: Vu et al. (2015)

Figure 23 Movement trajectory of tourist generated from geotagged photos

Regarding to places, the temporal effect is presented by users and activities on the place, such the check-ins of users in a place during one period. The function of a place is also defined by human activities on it rather than a name of a place. For example, Plaza de España is not only one large plaza in Barcelona, Spain; but also a central transport hub and commercial area(**Figure 24**). The government updates transport information in Plaza de España. Those owners of commercial activities, such as hotels and shopping centers also can add their information to the corresponding positions on the map.



Source: Google Maps

Figure 24 Plaza de España in the view of Google Maps

The spatial locations of places form the spatial distribution of places -- the density and the trend of distribution that can reflect the land uses of a city (Andrade, Alves, & Bento, 2020; Rizwan, Wan, & Gwiazdzinski, 2020). The evaluation of places means comments, rating scores and opinions of a place. For example, nowadays, people are used to reviewing the comments and the average price of restaurants for helping them to make choices. As **Figure 24** shows, Google Maps users leave 342 rating scores to the place and some of them upload photos about the square.

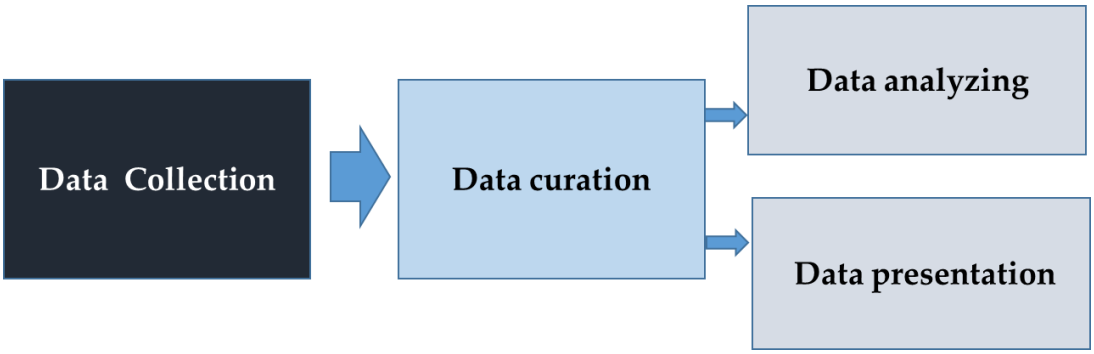
Contents includes semantic information and image information, which provides abundance social information for places and users. Therefore, LBSN data has an irreplaceable value for urban analysis. Both semantic and image analysis drives specific disciplinary for mining useful information, such as the nature language processing and distinguishing images. The features that we list are the most common used features in urban analysis. Languages can broadly identify the group of study and the location. The frequency of words is able to predict the social event and the public sentiment. The emotion that are expressed in words and sentences, such as happy, fear and sad, can be extracted and analyzed. Therefore, it could partly reveal the public perception and sentiment toward the city. Images can deliver the visual information about places, which can be a proxy

indicator of the urban land cover(Alqurashi, Kumar, & Sinha, 2016), land uses(X. Li, Zhang, & Li, 2017), and the natural environment(Leighton, Hugo, Roulin, & Amar, 2016).

Briefly, LBSN data are generated by users spontaneously, which can be retrieved through APIs that are offered by LBSN applications. A single point of LBSN data actually consists of multiple features: a user, a location or an activity. Therefore, LBSN data enable researchers to analyze and model the complex movement and relationship of people and places in different spatial granularities, such as streets, districts and cites.

2.2. General data mining process of urban studies leveraging LBSN data

The general process of LBSN data analysis includes data collection, curation, analysis, and data presentation(**Figure 25**). Data collection and curation usually cost a large portion of time during the period of research. Data collection consists of two components: base map data and the LBSN data. Base map usually includes the geographical range of studied area, administrative delimitation, and other attributes that may apply in the research.



Source: own elaboration

Figure 25 Framework of LBSN data analysis

LBSN data are mainly collected from web clawer and API. Application programming interface (API) (Park, Kim, & Ok, 2018; Plunz et al., 2019; Rizwan et al., 2020) and the web crawler (Manovich, 2016) are the most common methods to

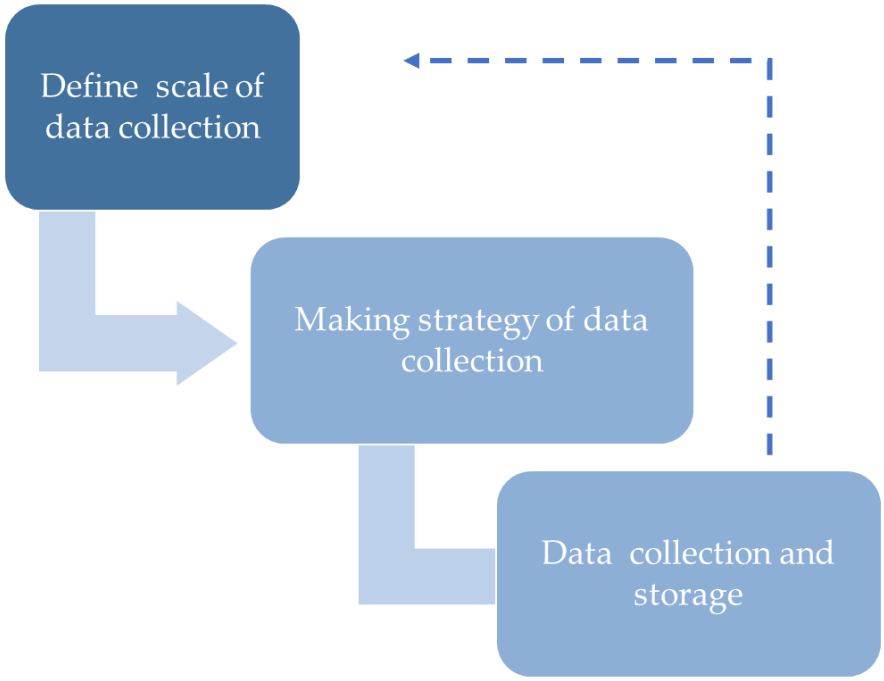
retrieve data, which can feedback nearly real-time data flows. A web crawler is a type of Internet robot that can browse web pages automatically. It is similar to a person who browser web pages, however, with a much faster speed. It is capable to find and download information on web pages multitudinously. Therefore, many LBSN applications adopt various mechanisms to protect their information from web crawler, due to the requirement of data security and commercial interests. Usually, it requires excellent programming skills to crawl data from LBSN sites. Thus, there are many commercial software earn profit from crawling data.

Instead, API is easier to acquire data with limitations. API is a set of routines, protocols, and tools for building software applications. It allows users to access the database with a credential token or permission. Many popular applications, such as Twitter, Foursquare, Google place, provide various APIs for the public. The availability of data partly depends on the price because high capacity computers are the necessity for receiving data and commercial APIs are expensive. All APIs have restrictions on accessing times and duration. For example, the free Twitter API provided 1% of all tweets at a given time according to its official documentation in 2013. The commercial API – Twitter Firehose, could provide data that were double than the free one (Morstatter, Pfeffer, Liu, & Carley, 2013). The newest Twitter API[9] has a clear rule of restriction on the allocation of the number of Twitter IDs that a user can visit during 30 days. Nevertheless, it permits researchers to redistribute unlimited Twitter IDs for academic purposes.

In general, the process of collecting data from API follows similar steps(**Figure** 26): 1) defining the search scope; 2) setting the technical strategy of data collection; 3) retrieving raw data; 4) storing and cleaning data. The search area and strategy of data collection usually depend on the data provider, and thus it is possible to make several tests for validating the design of data collection. Moreover, compared with the traditional dataset, the process of cleaning, reformatting, and storing LBSN data usually involves more sophisticated algorithms and programming parts. Empty and repeated entries, and inconsistent formatting are

[9] https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

common errors that occur in the process of data collection. Therefore, this process is a tedious but important step for further analysis.



Source: own elaboration

Figure 26 Flowchart of data collection

After the data collection, the curation of data refers to clean, select and organize the dataset. Data cleaning is a process to eliminate errors of the dataset, such as duplicate entries, null entries, fake users (Mukhina, Visheratin, & Nasonov, 2020), etc. **Table 6** lists some common problems that exist in the source- dataset. The temporal abnormality means that the extreme fluctuation of users' records during a period or within a spatial range. For example, the data lost will lead the temporal activity line drop sharply. The problem of coordinates format appears when researchers try to cooperate the LBSN dataset with other spatial layers.

Table 6 Objects of data cleaning

Type of data cleaning	Problems
Data observations	duplicate entries, null entries
Temporal	abnormal activities
Spatial	coordinates system matching
User	robots, spams, inactive users, types of users
Semantic	misspelling, stop words, unrecognized words

Source: own elaboration

Spamming detection aims to clear the fake users and messages to avoid data distortion. Many scholars developed various methods to separate the normal activity and the spamming activity(B. Wang, Zubiaga, Liakata, & Procter, 2015; X. Zheng, Zeng, Chen, Yu, & Rong, 2015). Characteristics of user behaviors and content attributes are two main aspects to detect spamming activities. For example, the abusive temporal activities probably are spamming activities (Thomas, Grier, Song, & Paxson, 2011). Content features of spams, such as the length of contents, keywords, and topic are also useful in spamming detection(Blanzieri & Bryl, 2008). Except for spam, the process of user curation would also remove some inactive users or users with particular demographic backgrounds. It largely depends on the scope of the study. Semantic analysis requires an additional text cleaning for further calculation, such as unrecognized letters and stop words. Since the dataset usually contains thousands entries, the process is intractable.

Data analysis is a process to explain the properties of the data and the specific phenomenon that the data reflect. The visualization of data is to display characteristics of data and to understand the results of the analysis. In the current, many software supports large-volume data analysis and visualization, such as R[10], Python[11], Stata[12], and ArcGIS[13]. They can provide various quantitative approaches for researchers, such as regression models, spatial analysis, network analysis, and natural language toolkit. Many special commercial analysis tools are more convenient to use, such as Tableau[14] and Power BI[15].

[10] https://www.r-project.org/
[11] https://www.python.org/
[12] https://www.stata.com/
[13] https://www.arcgis.com/index.html
[14] https://www.tableau.com/
[15] https://powerbi.microsoft.com/en-us/

2.3. Outline of case studies

For providing empirical observations, the quantitative part will analyze three popular LBSN data (**Table 7**): Twitter, Sina Weibo, and Foursquare. Twitter and Foursquare data are located in Barcelona, Spain; Sina Weibo is in Beijing, China(**Figure 27**). The detailed descriptions of dataset are elucidated in each case studies.

Source: own elaboration

Figure 27 Location of case studies

The objects of research are determined by two factors: characteristics of data and the innovation of theme. The characteristics of data refer to the accessible volume, the location of dataset, and information attributes of the dataset. The innovativeness means that the study could provide some new insights into the whole academic field.

Table 7 Summary of case studies

LBSN data	Place of case studies	Focus	Temporal coverage	Volume of data	Main techniques	Publication
Sina Weibo	Beijing, China	Urban spatial structure	2016/04/11 – 2016/04/17	52,543	• Temporal similarity of activities (cosine similarity); • Classic sub-center identification(linear regression)	Yang, L. & Marmolejo, C.(2020). Analysis of the spatial structure of Beijing from the point view of Weibo Data. ACE: Architecture, City and Environment, 15(43), 9302. DOI: http://dx.doi.org/10.5821/ace.15.43.9302
Foursquare	Barcelona, Spain	Functional relationships of places	2012/04/03-2013/09/06	79,798	• Network analysis; • Interaction model	Yang, L., & Duarte, C. M. (2019). Identifying tourist-functional relations of urban places through Foursquare from Barcelona. *GeoJournal*, 1-18.
Twitter	Barcelona, Spain	Public sentiments	2016/09-2019/04	1,100,244	• Sentiment analysis(text); • Multiple linear regression	Quantifying the relationship between public sentiment and urban environment in Barcelona. (Under review)

Source: own elaboration

The first study aims to analyze spatiotemporal urban structure of Beijing metropolitan area in terms of social media activities. It utilized Weibo (the Chinese equivalent of Twitter) density to describe the citizens' activities in Beijing in one week and adopt the classic exponential model to identify Weibo sub-centers. It detects sub-centers in different time periods and compares them with the actual urban land uses.

The second project is the analysis of functional relationship between places in urban space. this study focuses on quantification of the tourist-functional relations among Places of Interest (POIs) using Foursquare data from Barcelona. This represented an effort to highlight the important functional closeness between different types of POIs whose significance is not usually obvious from their spatial relationships.

The third case study shifts the scope to the investigation of public sentiments. Based on the unit of statistical area in Barcelona city, this research uses Twitter sentiment to represent the public sentiment and develops a regression model for understanding the interrelationship from three layers: sociodemographic, built-environment, and human mobility and socioeconomic activities.

Chapter IV.

Literature Review: urban studies leveraging LBSN data

This section aims to investigate the state of current urban studies leveraging LBSN spatiotemporal analysis. It is necessary to emphasize that the principal focus is applications of urban spatiotemporal analysis and used methods. Firstly, the application is defined as the primary research field of the paper, such as urban structure and urban mobility. The research field should be limited to urban issues, otherwise, the paper will be excluded. For example, many computing researches also collect LBSN data and investigate the human mobility while their purposes are to improve algorithms(Jin et al., 2014) or build computer applications (Bagci & Karagoz, 2016). Secondly, the review of methods focuses on the summary of utilized approaches for these application rather than the evaluation of these approaches.

IV.1. Existing literature review

There are several existing surveys of applications of LBSN data make efforts to cover a specific domain (e.g. the quality of LBSN data, disaster management, LBSN data, urban computing). At earlier stage, the application of LBSN data was summarized briefly by Y. Zheng et al. (2014). In specific domains, Horita, Degrossi, de Assis, Zipf, and de Albuquerque (2013) and Haworth and Bruce (2015) reviewed the utilization of LBSN data for disaster management.

The first broad review of LBSN applications is provided by Roick and Heuser (2013). They produced a general review about the research on LBSN data, which introduced applications via three aspects: social aspects, information extraction and GIScience theories. Steiger, De Albuquerque, et al. (2015) systematically reviewed spatiotemporal analyses regarding Twitter data. According to their data pool, computer science and information science account for 76% of academic disciplines. Silva et al. (2019) summarized five aspects of urban computing with LBSN data: social and economic, city semantics(*i.e.* city dynamics), city problems, urban mobility, urban health and events detection.

Except for applications, Niu and Silva (2020) studied key authors, most cited papers , and data sources in the field of urban mobility between 2013 to 2019. The most influential researches are *Citizens as Sensors: The World of Volunteered Geography* (Goodchild, 2007) and *Understanding Individual Human Mobility Patterns* (Gonzalez, Hidalgo, & Barabasi, 2008) The former brought up the concept of VGI (Volunteered Geography Information) , which provided a theoretic cornerstone for the geospatial studies leveraging LBSN data. The latter is the first article that proved that human mobility was not random using mobile records, which supported the potentials of LBSN data for studying urban mobility.

Table 8 Summary of existing literature review

Publication	Number of reviewed paper	Domains
Horita et al. (2013)	21	Disaster management
Haworth and Bruce (2015)	Not mentioned	Disaster management
Y. Zheng et al. (2014)	Not mentioned	Urban computing
Roick and Heuser (2013)	Not mentioned	Transactions in GIS
Steiger, De Albuquerque, et al. (2015)	92	Spatiotemporal analysis
Silva et al. (2019)	65	Urban computing
Niu and Silva (2020)	226	Urban activities

Source: Own-elaboration. Note: the number is based on their description in the article, rather than their references.

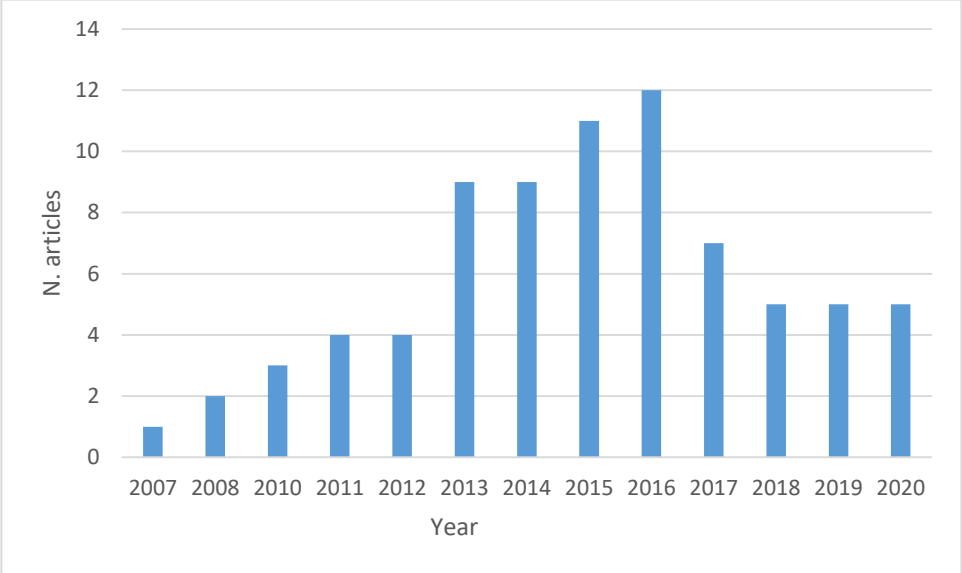
In summary, existing literature reviews failed to provide a completed retrospect of urban-centric LBSN applications. One reason is that multi-disciplinary urban study is evolving more complex and fast, as materials of studies usually leaps over several academic fields – architecture, geography, computing, psychology, etc. Secondly, numerous studies and new quantitative methods have been generated around the world every year, and thus it is a challenging task to generate a comprehensive review that envelop all sub-fields and researches. More importantly, these reviews only focus on the hundreds of case studies while the theoretical summarization of urban studies is remaining to be done.

Therefore, this section is expected to combine these existing reviews and new efforts to provide a broad and comprehensive review of urban spatiotemporal studies. The first part provides an overview of these articles that includes major applications, journals, and data sources. The second part introduces six crucial

directions of the current urban researches cooperated with LBSN data. Summary and discussions are in the final.

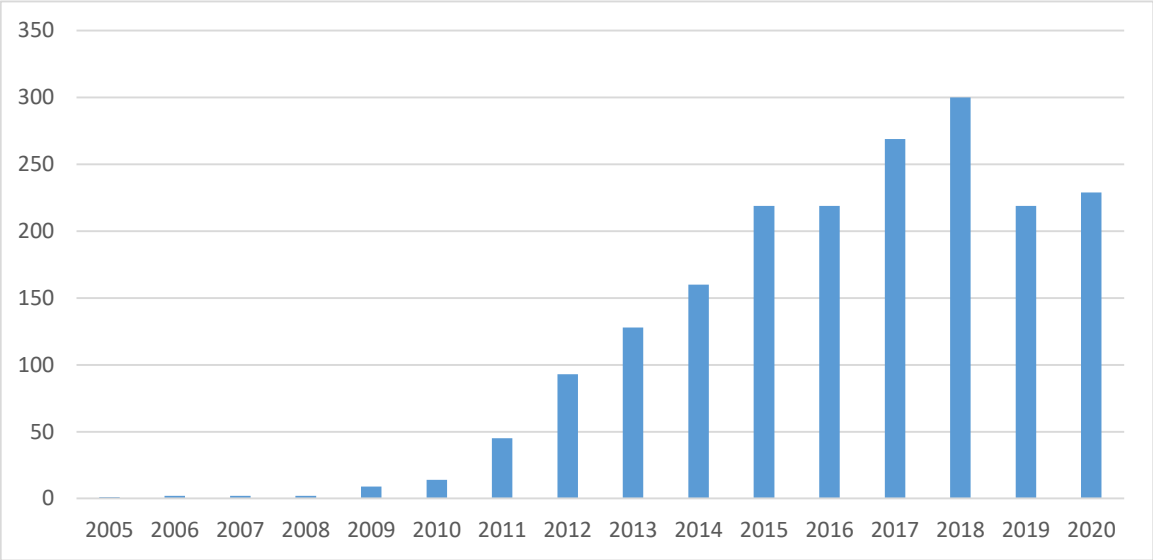
IV.2.An overview

The total number of the reviewed articles is 77, of which 55 representative pieces of research are investigated in detail. **Figure 28** displays the temporal distribution of these articles by year. The number of publications is growing after 2007, however, the peak value appeared in 2013-2016. It is different from the whole tendency of all references (**Figure 29**) because many papers with keywords "LBSN+Urban" belong to the investigation of computing algorithms. Moreover, some articles were excluded due to the lack of enough quality and novelty. According to the result of Niu and Silva (2020) , the 20 most cited references are published between 2003 to 2015. It indicates that the initial and the most innovative studies have been largely finished during that period.



Source: own elaboration

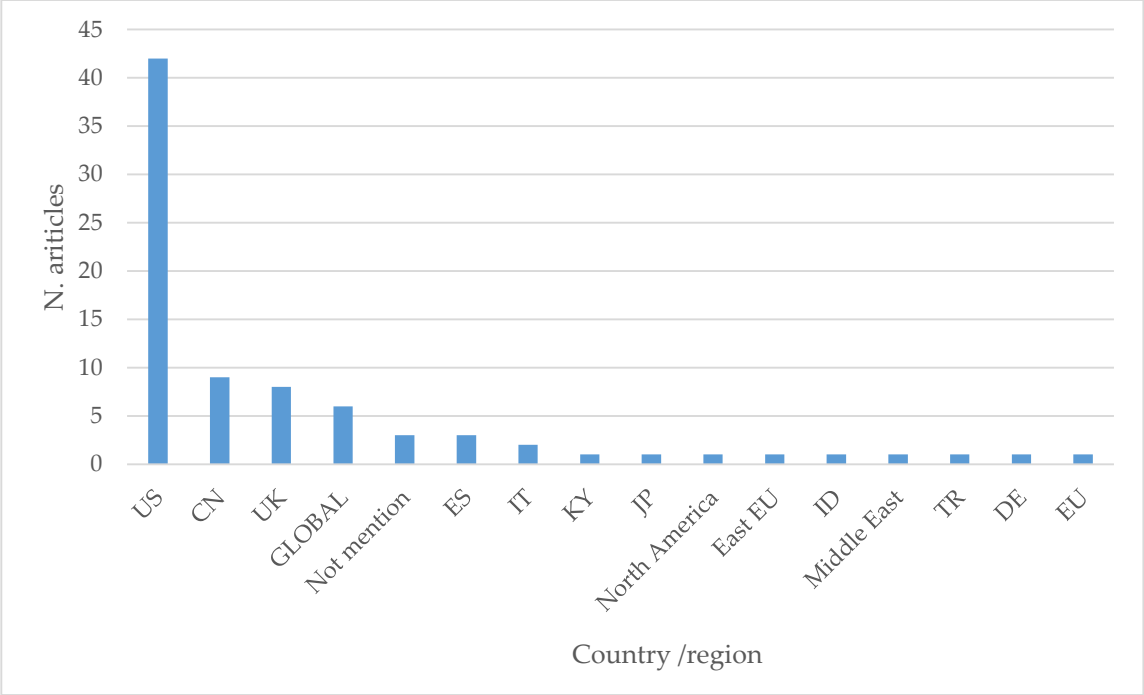
Figure 28 Number of reviewer of articles by year



Source: own elaboration

Figure 29 Number of publication by searching keywords: Urban & LBSN

Figure 30 describes the spatial distribution of these articles. About half of papers are from the United States, which is similar to the previous study (Steiger, De Albuquerque, et al., 2015). The followings are China and the United Kingdom. Three articles do not mention their geographic range due to the consideration of privacy. The number of research largely depends on the availability of data. Therefore, regions/ cities with more open data are more easily under the spotlight.



Source: own elaboration. Note: KY: Kenya, ID: Indonesia, TR: Turkey, DE: Germany.

Figure 30 Number of reviewed articles by country

The source of publication **Table 9** indicates which academic disciplines utilize LBSN data to investigate urban issues. In general, computer science and geoscience contribute more than 70% of them because most of the applications involve computing calculation and modeling. ACM, AAAI, WWW, and IEEE are the top level of international computing conferences. In the field of urban studies, *Computers, Environment and Urban Systems* is at the leading position of the number of articles. Meanwhile, it is worth to noticing that LBSN data has been widely applied in different disciplines of research, from psychology, behavior study to transportation planning. In other words, LBSN data also expand the research scope of urban issues.

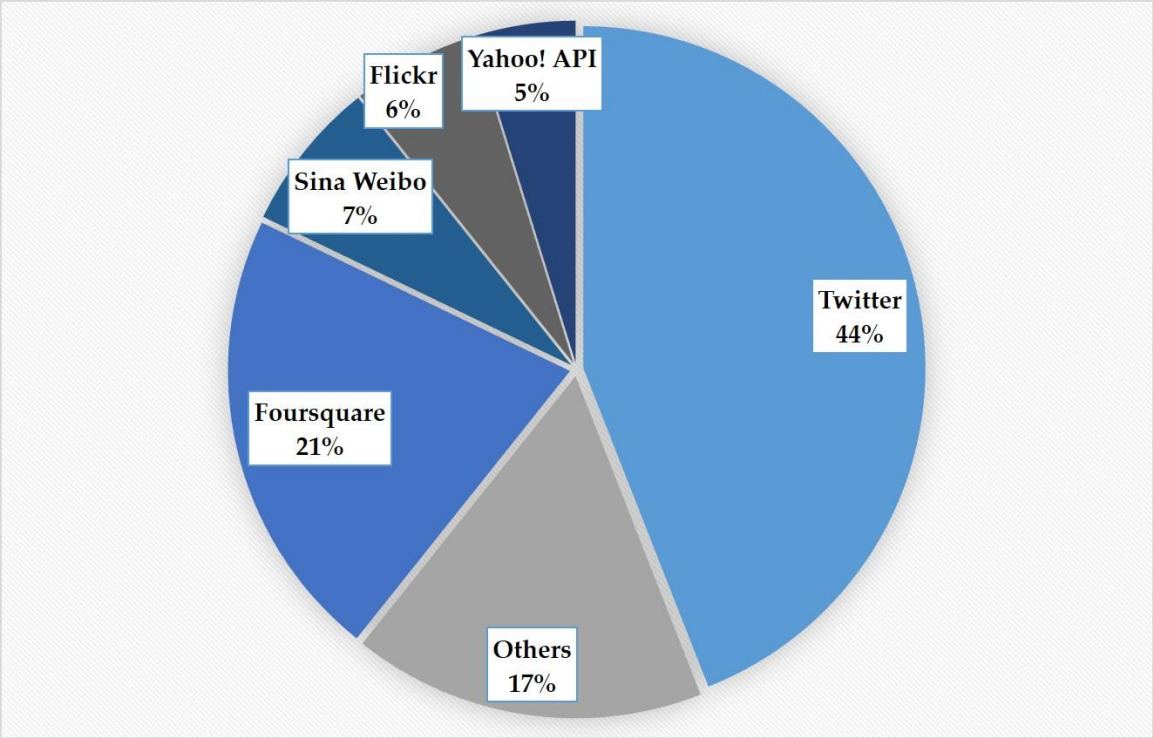
Table 9 The distribution of publication source

Name	N	Type	Name	N	Type
ACM	8	C	Sustainability	1	J
AAAI	4	C	American Behavioral Scientist	1	J
WWW	4	C	SIGCHI	1	C
Computers, Environment and Urban Systems	4	J	GeoJournal	1	J
PloS one	3	J	Annals of the American Association of Geographers	1	J
IEEE	3	C	Arabian Journal of Geosciences	1	J
Cartography and Geographic Information Science	2	J	Proceedings of the first workshop on social media analytics	1	C
Transactions in GIS	2	J	Engineering Applications of Artificial Intelligence	1	J
ISPRS International Journal of Geo-Information	2	J	International Conference on Service Systems and Service Management	1	C
Cartography and Geographic Information Science	1	J	Digital Government: Research and Practice	1	J
Habitat International	1	J	Proceedings of the first international conference on IoT in urban space	1	C
Transportation Research Part C: Emerging Technologies	1	J	Environment and Planning B: Planning and Design	1	J
Social Network Analysis and Mining	1	J	Remote Sensing of Environment	1	J
Neurocomputing	1	J	Applied Geography	1	J
American ethnologist	1	J	People and Nature	1	J
Current Issues in Tourism	1	J	Journal of Spatial Information Science	1	J
Royal Society open science	1	J	Journal of Political Economy	1	J
Landscape and urban planning	1	J	Nature	1	J
Transportation Research Record	1	J	Tourism Management	1	J
Tourism Geographies	1	J	International Journal of Human-Computer Studies	1	J
Environmental Monitoring and Assessment	1	J	Applied Psychology	1	J
Urban Studies	1	J	The Professional Geographer	1	J
Intn'l Forum on Tourism Statistics	1	J	International Journal of Geographical Information Science	1	J
International journal of disaster risk reduction	1	J	International journal of environmental research and public health	1	J
SCOPES	1	C	Journal of Experimental & Theoretical Artificial Intelligence	1	J
arXiv preprint	1	J	Journal of Public Transportation	1	J

Note：J: journal; C: conference. ACM: Association for Computing Machinery; AAAI: Association for the Advancement of Artificial Intelligence; WWW: International World Wide Web Conference; IEEE: Institute of Electrical and Electronics Engineers; SCOPES: International Workshop on Software and Compilers for Embedded Systems; SIGCHI : Special Interest Group on Computer–Human Interaction.

In terms of data sources (**Figure 31**), 45% of the papers utilize Twitter data; and 20% obtained data from Foursquare because both data sources are of a comparatively higher degree of openness. They can provide the retrieving range

from a global level (Fried, Surdeanu, Kobourov, Hingle, & Bell, 2014; Hawelka et al., 2014) to a city level(Q. Huang & Wong, 2016; F. Luo et al., 2016). Sina Weibo is a regional LBSN application which users are mainly located in China, and thus the availability of data is limited. Flickr can provide free geotagged photos for scholars. Some researchers collect webpages (Tsou et al., 2013) or POIs (S. Jiang, Alves, Rodrigues, Ferreira Jr, & Pereira, 2015)through Yahoo! API.



Source: own elaboration

Figure 31 Data sources of reviewed articles

Other data sources include Facebook, Instagram, Yelp, Gowalla[16], Google Places, etc. The full list of data sources is in **Table 10**. Facebook and Instagram are popular LBSN applications, however, the access to data is quiet limited due to the privacy concerns and the commercial policy. Gowalla, Brightkite, Niche and AutoNavi Mpas are also regional applications rather than international software.

[16] https://go.gowalla.com/

Table 10 Data sources of reviewed articles

Data Source	Popular degree	N. of reviewed papers	Data Source	N. of reviewed papers	Popular degree
Twitter	Global	37	Gowalla	2	Europe
Foursquare	Global	17	Brightkite[17]	2	Europe(stopped)
Sina Weibo	China	6	AutoNavi Maps[18]	1	China
Flickr	Global	5	OpenStreetMap	1	Global
Yahoo! API	Global	4	Google Places	1	Global
Facebook	Global	2	Yelp	1	Global
Instagram	Global	2	Niche[19]	1	USA
			Telecom	1	Global

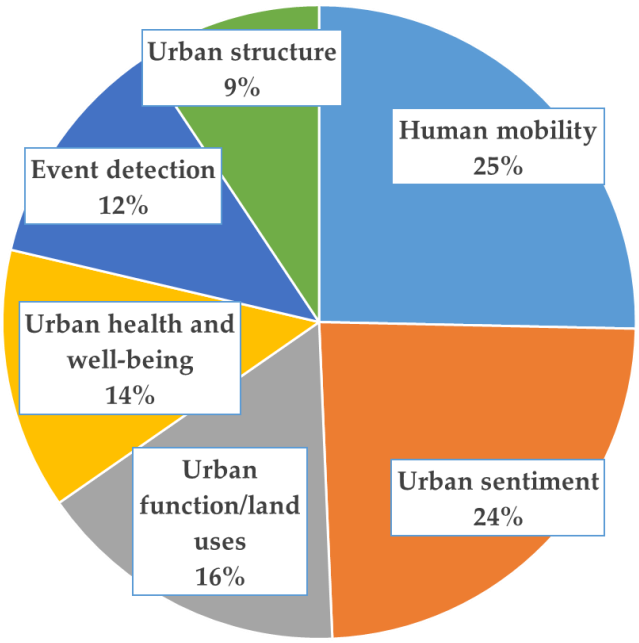
Source: own elaboration

We classify six categories of urban applications based on the previous categorizations (Niu & Silva, 2020; Y. Zheng et al., 2014). These directions cover three main aspects of urban studies, including geospatial, social-economic, and social-semantic layers**.** **Figure 32** described the proportion of reviewed articles according to their research directions.

[17] A place check-in application. Closed in 2010. The dataset that contains all links among users can be found in: http://networkrepository.com/soc-brightkite.php

[18] https://mobile.amap.com/

[19] Niche: Explore Schools, Companies, and Neighborhoods. https://www.niche.com/about/mobile/



Source: own elaboration

Figure 32 Breakdown of reviewed articles in different urban issues

Urban structure is the least reviewed topic because many studies have analyzed the urban structure from perspectives of human mobility (Cranshaw, Schwartz, Hong, & Sadeh, 2012)or urban functional places(S. Jiang, Ferreira Jr, & Gonzalez, 2012). Human mobility is a classic research direction of urban studies meanwhile supports the specific planning decisions, such as transport planning and the detection of locations. Event detection is a relatively new branch of urban studies because LBSN data are characterized by the ability to describe the urban dynamics(Xia, Hu, Zhu, & Naaman, 2015). Before the blooming of social media applications, it was barely to imagine that events could be detected and visualized on a map nearly in real-time. Urban sentiment analysis and urban health actually have been investigated for a long time. LBSN data provides new opportunities to observe these issues by semantic contents or spatiotemporal behaviors of users.

IV.3.Human mobility

Previous studies have proven that LBSN could be considered as a valuable proxy of human mobility at both country and city levels (Abbasi & Alesheikh, 2018; Asgari, Gauthier, & Becker, 2013; Daggitt, Noulas, Shaw, & Mascolo, 2016; Hawelka et al., 2014). At the global scale, Hawelka et al. (2014) studied global patterns of human mobility using geo-tagged Twitter data, which included mobility flows between countries and temporal patterns of human mobility. They found that Twitter flows were strongly correlated with flows of international tourists. At the regional scale, Twitter is a representative proxy for observing inter-regional mobility, especially in regions where have less available other types of data(Blanford, Huang, Savelyev, & MacEachren, 2015). As LBSN data is sensible to temporal variation, X. Hu, Li, and Bao (2017) utilized Weibo data to analyze the mobility flows among provinces of China during the Spring Festival(the most important festival in China). These two researches also indicate that LBSN data could be a powerful source to study domestic migration flows. Moreover, social ties also influence human mobility, which can explain about 10% to 30% of all human movements (Cho et al., 2011). They also found that periodic behaviors (i.e. repeated spatiotemporal movements) accounted for 50% -70% of human mobility.

At the city level, Abbasi and Alesheikh (2018) compared four models of the prediction of human mobility in New York City. They used taxi records as a reference to evaluate the results of LBSN data. They concluded that the predicting performance of LBSN data is better than the population and is more accurate for predicting long trips within the city. Béjar et al. (2016)extracted spatial-temporal characteristics of Twitter and Instagram users in Barcelona and Milan. They found that the main tourist attractions were important connecting nodes of human mobility in both datasets. However, Instagram routes were more related to tourist activities, while Twitter routes tended to be more diverse. F. Luo et al. (2016) studied the effect of demographic backgrounds on human mobility in Chicago using Twitter data. The race/ethnicity had a significant influence on the mobility pattern in Chicago, despite the general mobility pattern followed the distribution of power law.

Furthermore, human mobility can be divided into residents' mobility and tourists' mobility. Ferreira, Silva, and Loureiro (2015) collected Foursquare data in London, New York, Rio de Janeiro, and Tokyo. They classified POIs into nine categories and compared the check-in patterns of tourists and residents in these categories. They found that some categories show a significant temporal difference of check-in pattern between tourists and residents, such as locations classified as Arts and as Transport. The *Livehoods Project* in Pittsburgh utilizes social media data of residents gathered from Foursquare (Cranshaw et al., 2012). They discovered three mobility patterns of citizens and explained the relationship between different functions of places and these mobility patterns. For example, the spilled pattern of mobility usually indicated transit places.

Zhu, Blanke, and Tröster (2014) also utilized Foursquare to extract mobility patterns of different travel purposes ,such as shopping, working, and eating out. The result showed that characteristics of the traveling places and temporal features of trips played vital roles in inferring travel purposes. In other words, citizens' mobility is not random and highly predictable. Moreover, the socioeconomic condition also influences the mobility pattern. The study of Q. Huang and Wong (2016) concluded that poor people traveled the longer distance for working than mid-income people in Washington DC.

Regarding tourist mobility, it seems to have some universal characteristics: the tourist mobility is limited and mainly clustered around tourist attractions (Martí, García-Mayor, & Serrano-Estrada, 2020). Girardin, Calabrese, Dal Fiorre, et al. (2008)identified the differences of spatial activities between tourists and locals in New York via cell phone data and Flickr data. They confirmed that the movement range of visitors is limited, especially for foreign visitors. Vu et al. (2015) combined GPS and Flickr data to cluster tourists' main routes in Hong Kong and showed that tourists tend to travel to and between adjacent areas. Hasnat and Hasan (2018)described tourist movements as clustered around tourist attractions based on Twitter Streaming data in Florida.

Concerning the methodology, it is impossible to list all methods and their details. We only briefly introduce two commonly used approach for analyzing

human mobility according to our pool of reviewed articles. Firstly, although each study has its innovative methods, network analysis is one of the most popular theories to deal with mobility flows (see **Table 11**). It simplifies the complex flows into "edges" and "nodes", and use edges and nodes to represent various flows without the limitation of geographical scales. "Nodes" can be person, specific places, cities, countries, and aggregated places (e.g. restaurants, tourist attractions)." Edges" are the connection between nodes, which have different methods of calculation. The centrality analysis(Blanford et al., 2015) and community detection (Sun, 2016) also belong to the network theory. With the development of various computing software, the network analysis can be conducted straightforwardly, such as NetworkX[20], ArcGIS Network Analyst[21], and QGIS[22]. These software offer mature and manageable packages for spatial network analysis.

Another useful tool is the radius of gyration. Radius of gyration is often exploited for measuring the travel range of a user(Gonzalez et al., 2008; Hawelka et al., 2014), which is originated from the measure of mass distribution of an object. The radius of gyration of a user can be defined as:

$$R_u = \sqrt{\frac{\sum_1^i (P_i - MP_u)^2}{N_u}} \quad (1)$$

Where MP_u can be the geographical mean position or a given center of the user's trajectories. For example, F. Luo et al. (2016) used the activity center of a user to calculate the radius of gyration. P_i is one position of the user and N_u is the number of trajectories. Therefore, the radius of gyration can be understood as the standard deviation of distances between a point and the central point. A lower value of radius of gyration indicates the range of activities is smaller or more locally.

[20] https://networkx.org/

[21] https://www.esri.com/en-us/arcgis/products/arcgis-network-analyst/overview

[22] https://docs.qgis.org/3.4/en/docs/training_manual/vector_analysis/network_analysis.html

Table 11 Summary of representative studies of human mobility

Publication	Journal	Application	Location	LBSN Data			Major methods	Granularity of analysis
				Source	Period	Volume		
Hawelka et al. (2014)	Cartography and Geographic Information Science	Global mobility patterns and characteristics of mobility of different countries	Global	Twitter	2012/01-2012/12	944 M generated by 13 M users	Radius of gyration, network analysis	Country
Blanford et al. (2015)	PLOS One	Regional mobility flows	Kenya, Africa	Twitter	2013/06 – 2014/03	720,194 tweets generated by28,332 users	Radius of gyration, centrality analysis	Region
X. Hu et al. (2017)	IEEE	Inter-province mobility for understanding migration flows.	China	Sina Weibo	2017/02/13 -2017/02/21	30 B weibos generated by 1 B users	Clustering, network analysis	Province
Abbasi and Alesheikh (2018)	Arabian Journal of Geosciences	Testing the prediction of LBSN data in human mobility within the city	New York City, USA	Foursquare , Taxi trips	2012/04 - 2013/09	333,819 check-ins and 800,000 taxi trips	Four models of mobility, Sørensen similarity index	Borough
Zhu et al. (2014)	Proceedings of the first international conference on IoT in urban space	Detecting travel purposes from LBSN data	Puget Sound region of Washington State, USA	Puget Sound Travel Survey 2006, Foursquare POIs	2006, 2014	583 sub-categories of Foursquare POIs	multi-class classification: L1-regularized Linear SVM	Individual
Cho et al. (2011)	Journal of computer-mediated communication	Detecting the characteristics of human mobility in both short and long trips	Europe	Gowalla,Brightkite, mobile	Gowalla:2009/02-2010/10	Check-ins of Gowalla 6.4 M, Brightkite:4.5 M	Network analysis, modelling human mobility based on	Individual

				phone traces	Brightkite: 2008/04-2010/10:		Gaussian distribution	
F. Luo et al. (2016)	Applied Geography	Evaluating the influence of demographic	Chicago, USA	Twitter	2013/01-2013/06	300M tweets generated by 3M users	Radius of gyration, DBSCAN clustering	Individual
Béjar et al. (2016)	Journal of Experimental & Theoretical Artificial Intelligence	Developing general methods of spatiotemporal analysis using LBSN data	Barcelona, Spain; Milan, Italy	Twitter, Instagram	Barcelona: 2013/10 - 2014/09 Milan: 2014/03/20 14/08	Twitter: 3M in Barcelona, 1M in Milan. Instagram:3M in Barcelona, 0.5M in Milan	Clustering analysis, network analysis	Individual
Cranshaw et al. (2012)	International AAAI Conference	Discovering dynamically active clusters of the city	Pittsburgh metropolitan area, USA	Foursquare	2011	42,787 check-ins of 3,840 users at 5,349 venues	Clustering	Neighborhood
Vu et al. (2015)	Tourism Management	Travel patterns of tourists in Hong Kong	Hong Kong, China	Flickr	2011-2013/08	29,443 photos collected from 2100 Hong Kong inbound tourists	Markov Chain	Individual
Hasnat and Hasan (2018)	Transportation Research Part C: Emerging Technologies	Spatial mobility pattern of tourists	Florida, USA	Twitter	2017/03/29-2017/04/24	8,707 Twitter users	Ensemble classifiers, DBSCAN clustering	Individual
Q. Huang and Wong (2016)	International Journal of Geographical Information Science	Different activity patterns of Twitter users and their socioeconomic status	Washington DC, USA	Twitter	2014/01-2014/03; 2015/09 - 2015/11	14066 unique users	Activity pattern analysis, Standard deviational ellipse	Group

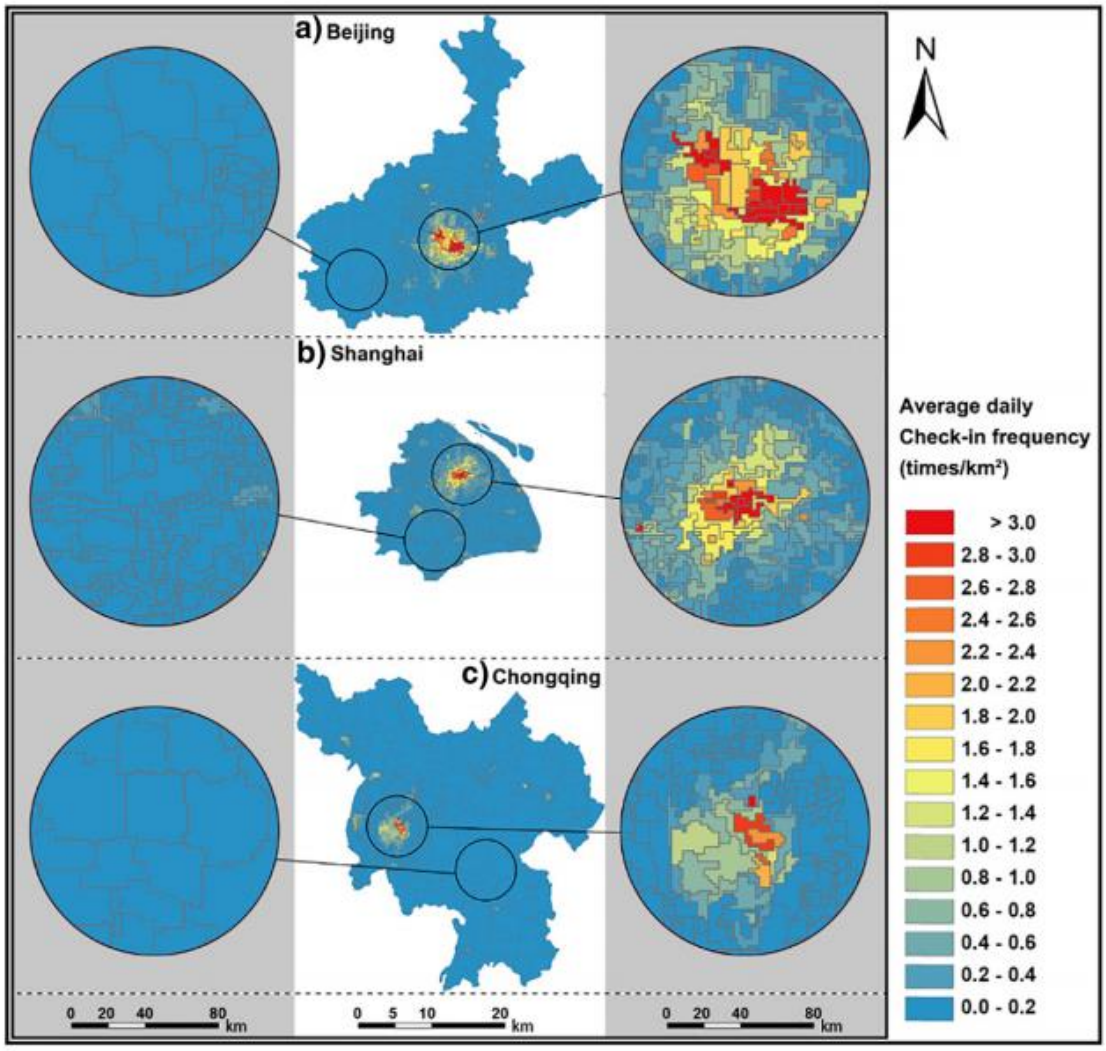
Ferreira et al. (2015)	IEEE	The spatiotemporal behaviors of tourists and residents	London, New York, Rio de Janeiro, Tokyo.	Foursquare	One week of April, 2012	Check-ins: London: 15,671, New York: 86,867, Rio de Janeiro: 27,222 Tokyo 118,788	Network analysis	Individual

Source: own-elaboration. Note: M: million. B: billion

IV.4.Urban structure

Urban structure is a classic studied object of urban issue because it can disclose the growth pattern and urban internal changes using models. Traditionally, it refers to the distribution of the internal social-economic environment, such as the employment sub-centers (Roca Cladera, Marmolejo Duarte, & Moix, 2009) and sub-centers of human activities(Carlos Marmolejo-Duarte & Cerda-Troncoso, 2020). Leveraging LBSN data, the description of urban structure can extend to a dynamic view of human activities (T. Chen, Hui, Wu, Lang, & Li, 2019). The identification of urban structure based on LBSN data is usually related to human mobility and the density of human activities. As many of them are discussed in the previous section, this section only explores studies that focus on the urban structure rather than human mobility(**Table 12**).

On the one hand, as many studies pointed out(Cai, Huang, & Song, 2017; Q. Huang & Wong, 2016; S. Zhang, Liu, Tang, Cheng, & Wang, 2019), the distribution of check-ins is usually concentrated in populated urban areas. On the other hand, as the temporal characteristics of human activities of a place is associated with the function of the place, and thus urban structure can be deduced from the patterns of human activities, such as monocentric city or multi-centric city. For example, Cai et al. (2017) defined the sub-center as a certain place with a high density of human activities (**Figure 33**). They followed the geographically weighted regression to identify sub-centers of three metropolitan cities in China using night light imagery and Sina Weibo data.



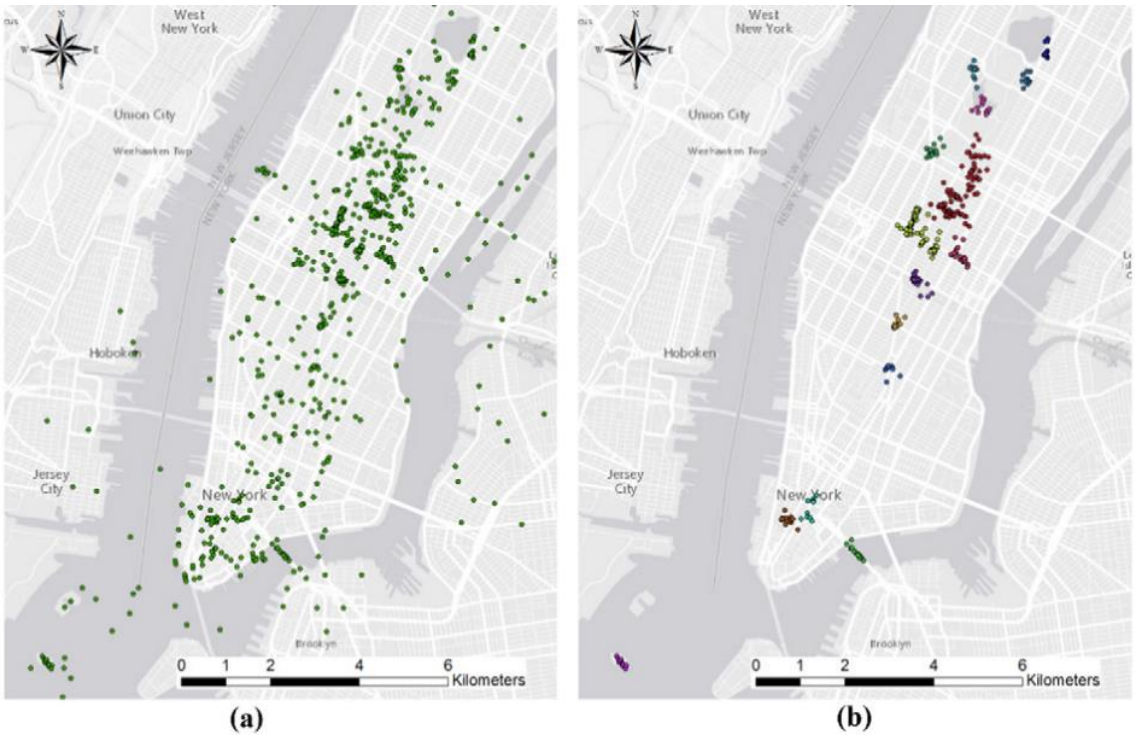
Source: Cai et al. (2017)

Figure 33 Spatial density of Weibo check-ins

T. Chen et al. (2019) utilized Facebook check-ins to identify the urban spatial structure in Hong Kong in terms of people's activities. They confirmed the multi-centric structure of Hong Kong through the temporal patterns of check-ins and the spatial variety of Facebook POIs. The activity patterns of Facebook users tended to be mixed in satellite cities of Hong Kong, which indicated that functions of these cities were various rather than "sleeping" solely. Hollenstein and Purves (2010) used photo tags from Flickr, such as "downtown" and "cbd", to delimit the central areas of six international metropolises. It turned out that these tags were partially able to capture the administrative boundary of the city, though only

between 0.5–2% of tags described city core areas generically. Meanwhile, they found out that the accuracy of LBSN data was enough to describe the dynamics at the scale of neighborhoods. Sun, Fan, Li, and Zipf (2016) precisely captured the city center through clustering Foursquare check-ins in Germany.

Furthermore, Y. Hu, Gao, et al. (2015) added public interest into the investigation of urban spatial structure using geotagged photos from Flickr. They displayed urban areas of interest (areas that can attract people's attention) of six cities from six different countries from 2004 to 2014 using DBSCAN clustering. The urban areas of interest(AOI) are not equal to the urban sub-centers, though there are overlapping in some areas(**Figure 34**). For example, some AOIs were formed by visitors' curiosity. Famous landmarks and commercial places frequently appeared in these areas. In summary, these studies tended to use the spatiotemporal distribution of LBSN data to identify urban structure or city center. They confirmed the availability of LBSN that can delimitate urban centers globally.



Source: Y. Hu, Gao, et al. (2015)

Figure 34 Extracting AOI from Flickr photo in NewYork; (a) locations of Flickr photos; (b) point clusters detected by DBSCAN

Besides, LBSN data also provides a new perspective to observe and define the city. B. Jiang and Miao (2015) utilized the spatiotemporal clusters of human activities to develop a new concept – "Natural Cities". Different from the traditional definition of city, Natural Cities were identified by human settlements or activities that were from "mass geographic information and based on the head/tail division rule ". They presented the changes of natural cities during 30 months in the United States using the data from a LBSN software Brightkite.

For detecting the urban structure, clustering is an important method to figure out hotspots activities: DBSCAN, K-means, K-medoid, Local Moran's I, local Getis-Ord G_i are popular approaches to detect clusters or groups. Local Moran's I (Anselin, 1995)and local Getis-Ord G_i (Getis & Ord, 2010; Ord & Getis, 1995) are classic methods to detect areas with high/low values. Strictly speaking, the function of these two methods is to explore the spatial association rather than spatial clustering.

Regarding to spatial clustering, partitioning and hierarchical algorithms are the two basic types of clustering(Kaufman & Rousseeuw, 1990). The hierarchical algorithm tries to build a hierarchy of clusters that may form from "top-down" or "bottom – top" approaches. However, such a method does not fit the characteristics of LBSN data which are composed by enormous points. Moreover, the performance of hierarchical algorithms is lower than K-mean algorithm(Abbas, 2008).

Partitioning algorithms seek to divide the dataset into clusters/groups without the distinction of hierarchy. They are widely used in researches, especially DBSCAN, K-means, K-medoid, and DBSAN. Both K-means and K-medoid requires a pre-determined number of clusters. However, K-medoid uses actual points as centers while K-means uses calculated average centers. K-means algorithms perform very well with huge datasets (Abbas, 2008). DBSCAN (density –based spatial clustering for applications with noise) is a density-based clustering method(Ester, Kriegel, Sander, & Xu, 1996), which contains two parameters: the search radius and the minimal number of points within a cluster. It can deal with different shapes of clusters without the necessity of pre-defining

the number of divisions. More importantly, it is robust to data noise that often occurs in LBSN datasets(Cai et al., 2017). However, the limitation of DBSCAN is that it cannot identify numeric features or attributes(Sun et al., 2016). It is also not an ideal method to conduct areas with large differences of densities because it is difficult to determine an appropriate search radius(Kriegel, Kröger, Sander, & Zimek, 2011).

Table 12 Summary of analyzed studies of urban structure

Publication	Journal	Summary of application	Location	LBSN Data			Major methods	Granularity of analysis
				Source	Temporal Range	Volume		
T. Chen et al. (2019)	Habitat International	Detecting the urban spatial structure of Hong Kong	Hong Kong, China	Facebook	One week	3.4M-8.4 M Facebook check-ins per day, gathered from over 50,000 unique POIs	Clustering analysis	1 km^2 cells
Y. Hu, Gao, et al. (2015)	Computers, Environment and Urban Systems	Discovering the interesting areas of cities through Geo-tagged photos.	New York City, London, Paris, Shanghai, Mumbai, Dubai	Flickr	2004/06 - 2014/06	7,492,965 photos in total	DBSCAN clustering, Semantic and image clustering	Polygon
Cai et al. (2017)	Remote Sensing of Environment	Detection of sub-centers	Beijing, Shanghai, Chongqing	Sina Weibo	2015/04/25 -2016/05/25	More than 5.6M check-ins	Local Moran's I, geographically weighted regression	Polygon
Sun et al. (2016)	Environment and Planning B: Planning and Design	Identifying city centers using LBSN data	Berlin, Munich, Cologne	Gowalla	2009/02-2010/10	70,000 check-ins	Local Getis-Ord G_i DBSCAN Givan_Newman algorithm	Venue
Hollenstein and Purves (2010)	Journal of Spatial Information Science	Exploring the relationship between Flickr tags and places in city central areas	Zurich, London, Sheffield, Chicago, Seattle, and Sydney	Flickr	2008/05/02-2008.06/27	8M	Comparison analysis, Gaussian kernel density	City

B. Jiang and Miao (2015)	The Professional Geographer	Redefinition of the city	USA	Brightkite	2008/04-2010/10	2.8 M locations	Network analysis	City

Source: own elaboration

IV.5.Urban land uses and functions

The previous section only explores the urban structure from the view of density. This section focuses on introduce studies of land-uses and urban functions using LBSN data. The urban function is another vital perspective to analyze the urban structure, especially for modern cities(S. Jiang et al., 2012; J. Yuan et al., 2012). The distinction of urban functional areas, such as residential areas, commercial centers, leads to the diversification of human activity patterns in different areas. For example, workplaces mainly serve working people rather than students. Their active periods may be different from each other. Therefore, since LBSN data are derived from people's daily life, the function of urban areas could be identified by the patterns of LBSN activities. Utilizing LBSN data to identify land uses or urban usages has been studied by many researchers (Akhmad Nuzir & Julien Dewancker, 2017; Frias-Martinez & Frias-Martinez, 2014; J. Yuan et al., 2012). In concrete, the analysis of urban functions includes two major issues: the spatial distribution of land uses/ urban usages, and the connectivity between land-uses.

Firstly, the temporal pattern of LBSN activities is one of the most prominent indicators of different urban usages. For example, Lei, Zhang, Qi, Su, and Wang (2018) investigated the active degree of four types of land uses during one week through Sina Weibo from Beijing, China. The commercial and public green places showed a peak on Saturday, while the education and residential places tended to be smooth from Monday to Sunday. Frias-Martinez and Frias-Martinez (2014) aggregated the temporal pattern of Twitter users using spectral clustering for detecting land uses in London, New York, and Madrid. Their result indicated that geo-located tweets could be a powerful source for identifying land uses. Under the unsupervised approach, 60%-80% area of commercial -business land uses could be detected. J. Yuan et al. (2012) identified regional functions of Beijing metropolitan area through human mobility(taxi tracks) and data of POIs (points of interests). It took the region as a collection of functions, and a function could be identified by some mobility patterns. The unsupervised result was compared with the official planning. Their method successfully recognized the major commercial/ entertainment areas and residential areas. In addition, different groups of people

could also help to identify the function of places. Kádár (2014a)concluded that the majority of geo-tagged images generated from Flickr tourist users were gathered around tourist attractions or landmarks in Budapest.

Secondly, the aggregation of POIs information is another important method to identify land uses. Noulas, Scellato, Mascolo, and Pontil (2011) classified different urban usages, such as Food, Arts, and Parks, through clustering densities of Foursquare POIs and user's check-ins from New York and London. However, the temporal pattern of users was not mentioned by the article. Besides the general identification of land uses, POIs are also available for detecting the disaggregated land uses (i.e., employment size by category). S. Jiang et al. (2015) combined the official employment data with Yahoo POIs to enhance the estimation of land uses at the city block level. They proposed a method to match the type of POIs with the category of employment and then aggregated types of POIs. These clusters of POIs were compared with the ground- truth data that were from a commercial business provider.

Thirdly, some scholars detect urban functions from the semantic contents of LBSN data. It is also an attempt to find the correlation between the semantic layer and the spatial space. For example, Steiger, Westerholt, Resch, and Zipf (2015) combined Twitter sentiment analysis with official census data to examine the correlation between topics of tweets and their spatiotemporal locations. They concluded that Twitter was a representative proxy for workplace-based investigation because the positive correlation between work-related topic clusters and density of workplace population (the population working in an area during the working days) was statistically confirmed by linear regression. S. Gao, Janowicz, and Couclelis (2017) investigated the relationships between semantic topics and urban functions using Foursquare data from Los Angles. They found out some topics were more related with specific urban functions, such as the "shopping-plaza" topic appeared frequently in places like shopping malls, stores and some shops.

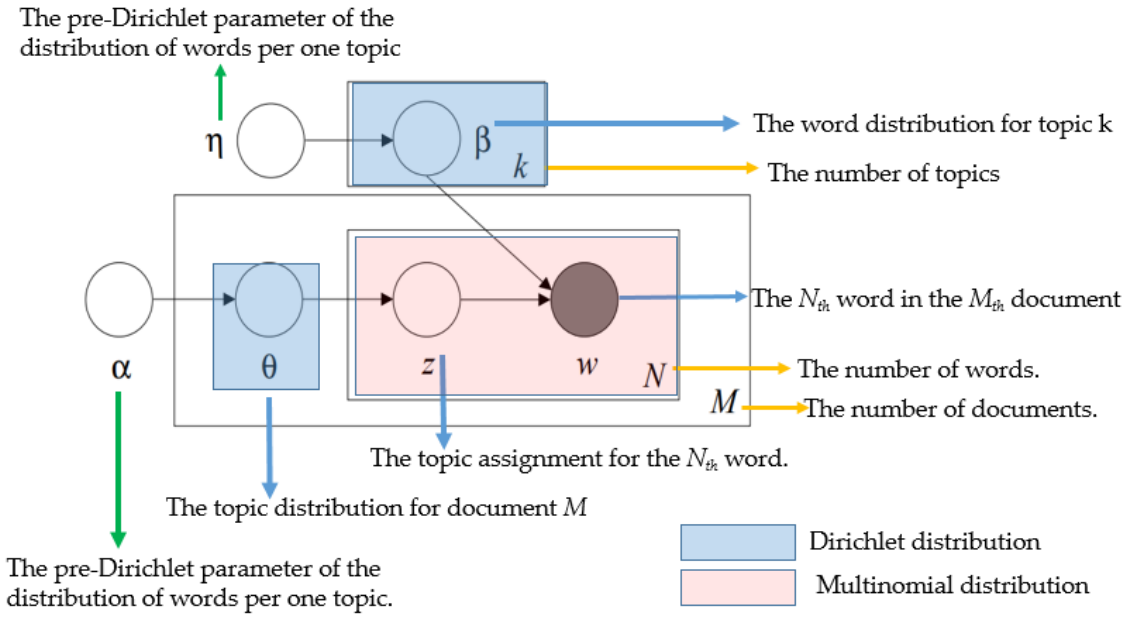
The above studies only show the degree of spatial aggregation of human activities in separate places or the tendency of people's movements. Even though

the quantification of flows between places have been involved in studies of travel patterns and land-use detection, the focuses are still different from functional linkages. Preoţiuc-Pietro and Cohn (2013) discussed the degree of transition between different types of Foursquare POIs using the check-ins of high-frequency users, but did not explain the functional relationships among POIs. Later, Sun (2016)explored the spatial interaction between Foursquare POIs (i.e. the degree that one venue connected to its nearby POIs) using complex networks. Four areas with strong interactions were detected in New York city and their main usages were listed. However, this study still focused on the spatial connections rather than the functional connections.

In summary, LBSN data is a representative source to detect land uses/ urban functions in general. The spatiotemporal behaviors of users can identify different land uses at a region/neighborhood level. Combined with the ground-truth observation, the precision of land use detection has reached a high level, such as the research of (J. Yuan et al., 2012). However, such a completed and high-precision dataset is difficult to access for many scholars. The semantic approach could provide much useful information about urban space. However, two core issues affect the precision of the method. Firstly, LBSN applications only reflect a part of daily life, though we are more and more relying on the Internet than before. Secondly, the spatial space is the place that allows people's activities to happen, rather than the motivation of people's behaviors. Therefore, it is hard to discover the direct connections between the semantic information and the built environment. Besides, the sophisticated algorithm of semantic analysis has to cost many efforts to explain in the article, such as the Latent Dirichlet allocation(LDA) algorithm. The focus of the study could be possibly distorted toward the process of the study rather than the result.

When focusing on the quantitative methods of urban functions/ land uses, clustering analysis is the main approach to aggregate points and polygons with the similar functions (**Table 12**). However, the LDA algorithm is also a pioneer method to discover the functional areas. LDA (Blei, Ng, & Jordan, 2003) is a unsupervised probabilistic model that can discover the latent characteristics of

discrete data such as a limited corpus. It is widely used for topic modelling of semantic analysis(Steiger, Westerholt, et al., 2015).



Source; reproduced from Blei et al. (2003)

Figure 35 Graphical representation of the LDA model.

LDA model consists of two main process **(Figure 35)**:

1) the proportion of topics for a given document (θ)and the word distribution for a topic(β). They are calculated by the Dirichlet distribution that are pre-set by the prior parameter α and η separately.

2) the topic assignment for a word in a document (z)and the probability of the word in the document(w). The mathematic equation is described as:

$$p(\beta_{1:k}, \theta_{1:M}, z_{1:N}, w_{1:M}) = \prod_{i=1}^{k} p(\beta_i) \prod_{i=1}^{M} p(\theta_m) \left(\prod_{n=1}^{N} p(z_{(m,n)}|\theta_m) p(w_{m,n}|\beta_{1:k}, z_{(m,n)}) \right) \quad (2)$$

The result returns two matrices β and θ that describe the assignment and proportion of topics separately.

We take POIs as an example to explain the process of extractions of urban functions using the LDA model. Firstly, an urban function or land use is assumed as a "topic". Correspondingly, the type of POIs is considered as a "word", such as

tourist attraction, restaurant, and workplace. A region or an area that contains POIs is a "document". After running the model, each region/area can obtain a vector that describes the composition of functions. The detail process of execution can be found in S. Gao et al. (2017) . The advantage of the LDA approach is that it can automatically generate "topics" of land uses and imply the latent composition of POIs of the land-use topic. For example, tourist area may include POIs of restaurants, hotels, tourist attraction, and meeting points.

Table 13 Summary of analyzed studies of urban functions/ land uses

Publication	Journal/ conference	Summary of application	Location	LBSN Data			Major methods	Granularity of analysis
				Source	Temporal Range	Volume		
T. Chen et al. (2019)	Habitat International	Detecting the urban spatial structure of Hong Kong	Hong Kong, China	Facebook	One week	3.4M-8.4 M Facebook check-ins per day, gathered from over 50,000 unique POIs	Clustering analysis	1 km2 cells
Steiger, Westerholt , et al. (2015)	Computers, Environment and Urban Systems	Validating the relationship between the semantic space and the physical space	London, UK	Twitter	2013/07/31 – 2014/07/31	476,071 users	Latent Dirichlet Allocation (LDA), clustering, network analysis, spatial autocorrelations	Individual
S. Gao et al. (2017)	Transactions in GIS	Extracting urban functional regions from POIs and human mobility	Los Angles, USA	Foursquare	2016/10	480 different types of POIs	Latent Dirichlet Allocation (LDA),	Region
Sun (2016)	International Journal of Geo-Information	Detecting spatial interactions using residents' movements	New York, USA	Foursquare	2014/03/03 -2014/04/27	148,169 check-ins	Network analysis	Point
Lei et al. (2018)	International Journal of Geo-Information	Identifying the spatial-temporal patterns of Sina Weibo users in Beijing	Beijing, China	Sina Weibo	2014/06	97,000	Local Moran's I	100*100m grids

Kádár (2014a)	Tourism Geographies	Detecting urban usage patterns of tourist	Vienna, Prague, Budapest	Flickr	Photos before 2013	Not mentioned	Correlation analysis	Points
Frias-Martinez and Frias-Martinez (2014)	Engineering Applications of Artificial Intelligence	Identifying urban land uses by Twitter activities	London, New York, and Madrid	Twitter	2010/10/25 - 2010/12/12	Not mentioned	Spectral clustering, Self-Organizing Maps	Polygon
J. Yuan et al. (2012)	18th ACM SIGKDD international conference on Knowledge discovery and data mining	Detecting urban functional areas through the mobility pattern	Beijing, China	POIs data, Taxi tracks (source not mentioned)	POIs: 2010,2011, Taxi tracks: 3 months in 2010 and 2011	12,000 taxicabs	Topic modeling (LDA), Kernel density estimation	Grids
Noulas et al. (2011)	AAAI Conference on Web and Social Media	Classifying urban usages based on Foursquare check-ins and venues	London, New York	Foursquare	2010/05/27 -2010/09/14	12 M check-ins generated by 679,000 users	Clustering	Cells
S. Jiang et al. (2015)	Computers, Environment and Urban Systems	Investigating the spatial interaction using residents' movement	Boston, USA	Yahoo! POIs	Not mentioned	64,133 POIs	Maximum likelihood estimation	Census block, polygon
Preoţiuc-Pietro and Cohn (2013)	ACM conference	Describing the transition probability between different categories of venues	Not mentioned	Foursquare	2011/08/31 - 2011/10/01	959,122 check-ins	Statistical analysis	Points

Source: own elaboration

IV.6. Urban sentiment / public perception analysis

LBSN data naturally has been concerned by researchers in recent years for sentiment analysis. In addition to the geo-spatial information, the contents of LBSN data provide a fast access to understand people's opinions and emotions. It can provide valuable information about the work stress (W. Wang, Hernandez, Newman, He, & Bian, 2016), elections(D. Paul et al., 2017; Yaqub et al., 2020), social movement (Gleason, 2013), even the stock market (Pagolu, Reddy, Panda, & Majhi, 2016), etc. For example, Collins et al. (2013) studied the sentiment variations of passengers of suburban trains near the city of Chicago. They found out that the dissatisfaction to incidents can be measured by social media data

Along with the geographical information, researches also study the relationship between spatial place and mass sentiments. In general, the wealth degree is positively correlated with the sentiment. Mitchell et al. (2013)collected tweets across 50 states of the United States to investigate the degree of happiness among cities and states. The degree of happiness refers to the average happiness score based on the frequency of positive words in tweets. They found that the happiness score is strong associated with increasing household income. Safety and crime issue was statistically correlated with citizen's subjective perception (Y. Hu, Deng, & Zhou, 2019).

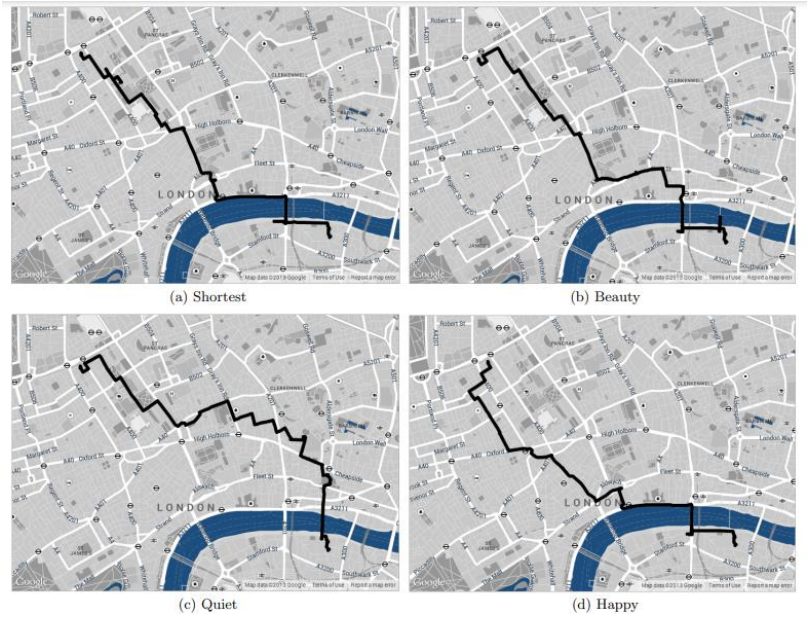
With regard to the sentiment reaction in specific places, Padilla, Kavak, Lynch, Gore, and Diallo (2018) investigated tourists' emotions at tourist destinations via Twitter data in Chicago. Their result showed that seasonal temperature is positively correlated with the positive sentiment in general. Urban parks could also reduce people's negative feelings according to the investigation of Schwartz, Dodds, O'Neil-Dunne, Danforth, and Ricketts (2019) in San Francisco because negation words such as 'no', 'not' decreased in frequency during visits to urban parks.

Gallegos, Lerman, Huang, and Garcia (2016) utilized Foursquare data to study the happier places in Los Angeles. It concluded that the happier places tend to be

observed in census tracts which have more Foursquare check-ins. Moreover, places with more amenities, such as restaurants, gym and beach, tended to be happier than other places. Bertrand, Bialik, Virdee, Gros, and Bar-Yam (2013) generated a sentiment map of New York City using Twitter. They found that public sentiments generally performed higher in public parks and lower at transportation hubs. Cao et al. (2018) studied the relationship between sentiment score and land uses using a linear mixed-effect model. They concluded that sentiment scores were higher in the commercial and public areas in the noon/evening and on weekends. The areas of farmland, transportation and industry tended to show negative sentiment in the midnight and weekdays.

Besides, the public responses to natural disasters can also be detected by LBSN data. For example, Neppalli, Caragea, Squicciarini, Tapia, and Stehle (2017) explored the public sentiment during Hurricane Sandy via Twitter users. The negative tweets clustered around hurricane locations more closely. Moreover, they found that people tended to share more informational contents than personal opinions during the disaster.

LBSN sentiment analysis can also be applied in smart planning. For example, Quercia, Schifanella, and Aiello (2014) proposed more emotional-route options (beautiful, quiet, and happy) for users rather than the shortest routes(**Figure 36**). The evaluation of these routes was based on the user's opinion from the website -- UrbanGems.org. After that, they calculated the happiness score for a location using these evaluations, and thus they could recommend the most beautiful/ happy routes. Interestingly, they found that increasing more walking routes could increase the pleasant feeling of users.



Source: Quercia et al. (2014)

Figure 36 Different paths between Euston Square and Tate Modern

Nik-Bakht and El-Diraby (2016) designed an online social media game to detect people's perspectives on the sustainability of urban infrastructure. It is a kind of half-designed social media study because the specialized game contains intentions of researchers. The mission of game players is to annotate infrastructure-related tweets using a set of indicators of sustainability. The perspective of sustainability is produced spontaneously by players when they play the game. It is a method to greatly improve the effectiveness of the data while the amount of samples is limited.

In summary, Twitter and other LBSN data have provided a more economical way to access public sentiments. However, the majority urban sentiment studies mainly focus on specific places or some groups of people. The relationship between the macro- built environment and mass sentiment still have plenty room for investigation. Moreover, most researches are limited to English texts or single language, lack of comparison of different groups of people who come from different cultural backgrounds.

Regarding the methodology of sentiment analysis, many studies (see **Table 14**) choose professional software to handle it, for example, Sentistrengh (Thelwall,

Buckley, Paltoglou, Cai, & Kappas, 2010)and the IBM Watson Alchemy application program[23]. Few research groups are able to develop their own sentiment classifier without the expert of natural language processing (NLP). The discussion of the sentiment analysis can be found in the part of case study.

[23] https://www.ibm.com/watson/alchemy-api.html

Table 14 Summary of representative studies of urban sentiments

Publication	Journal / conference	Summary of application	Location	Data			Major methods	Granularity of analysis
				Source	Range	Volume		
Quercia et al. (2014)	25th ACM conference on Hypertext and social media	Planning of recommended routes	London,UK	UrbanGems.org; Flickr	2012/09-2012/12	3301 participants; 5M photos (Flickr)	Network analysis	Cell of 200 x 200 meters
Nik-Bakht and El-Diraby (2016)	International Journal of Human-Computer Studies	Exploring the public community's perspective on sustainability of urban infrastructure	North America	Twitter; Sustweetability (a self-designedgame)	Twitter :2012/08 - 2013/06 Game: 2013/06/25-2013/08/10	167 participants; 25853 tweets (782 involved tweets)	Annotation of topics; statistical analysis	Individual
Gallegos et al. (2016)	25th International Conference Companion on World Wide Web	The distribution of public sentiments in the city	Los Angeles, USA	Twitter	2014/07-2014-10	6M	SentiStrength; radius of gyration	Census track
Schwartz et al. (2019)	People and Nature	The effect of greenspace to visitors	San Francisco, USA	Twitter	2016 /05/19-2016/08/02	About 70,000 tweets per day	Hedonometer; Amazon's Mechanical Turk	Park polygon
Collins et al. (2013)	Journal of Public Transportation	Detecting the sentiment variation of passengers of suburban trains	Chicago, USA	Twitter	2011/06/11-2011/06/23	457 relevant tweets	SentiStrength Statistical analysis	Individual
Padilla et al. (2018)	PloS one	The temporal variation of tourist sentiment in tourist attractions	Chicago, USA	Twitter	2014/05/07-2015/05/02	8,034,025 tweets from 225,805 use	SentiStrength	Tourist place
Bertrand et al. (2013)	arXiv preprint arXiv:1308.5010	Measuring the spatiotemporal variation of public sentiments at a fine-grained scale	New York City, USA	Twitter	Two weeks in April 2012	603,954 tweets	Customized corpus-based sentiment classifier	Polygon
Cao et al. (2018)	International journal of environmental research and public health	The relationship between different land uses and public sentiments	Massachusetts, USA	Twitter	2012/11/31 - 2013/06/03	880,937 tweets posted by 26,060 users	IBM Watson Alchemy application program;	Polygon of land uses

							multivariate linear mixed-effects model	
Neppalli et al. (2017)	International journal of disaster risk reduction	Variations of public sentiment during hurricane Sandy	USA	Twitter	2012/10/26-2012/11/12	12.9 M tweets 74,708 geo-tagged tweets	Sentiment classification (Navie Bayes and SVM);standard deviation ellipse; emotional divergence value	Cluster
Y. Hu et al. (2019)	Annals of the American Association of Geographers	Public perceptions towards their living environment	New York City , USA	Niche	On and before 2017/05/02	7673 reviews	LDA, correlation analysis	Neighborhood

Source: own elaboration

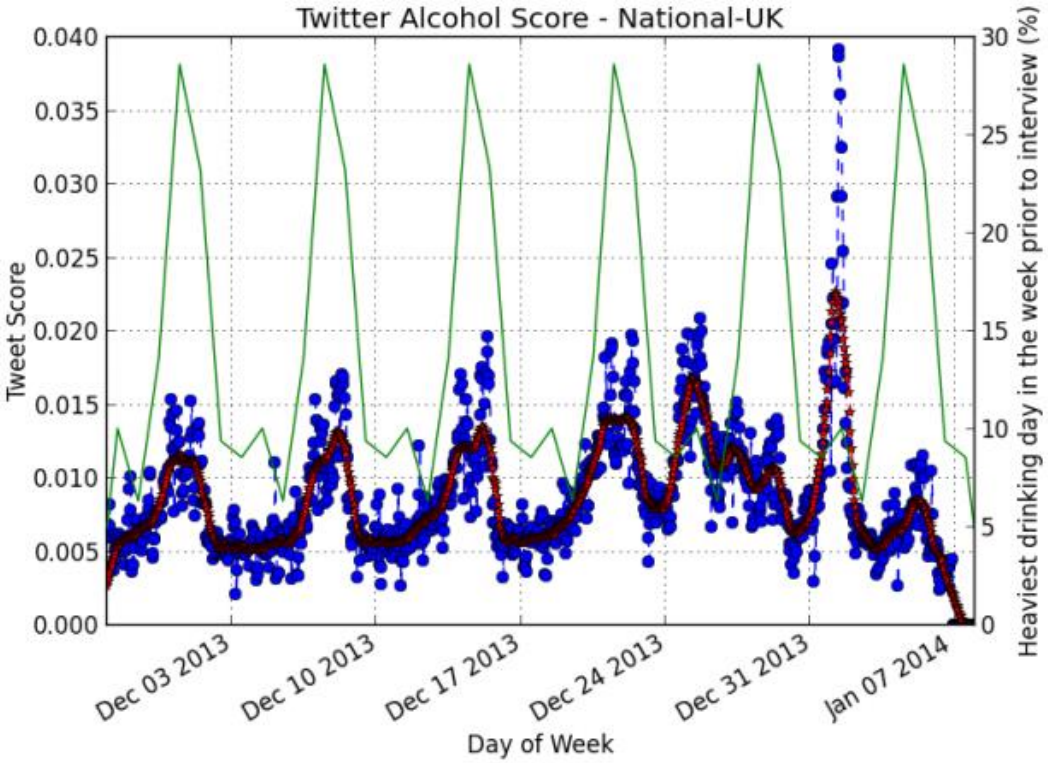
IV.7. Urban health and well-being

Many researches have made efforts to utilize LBSN data to detect health-related topics and social well-being issues, such as disease spread(M. Paul & Dredze, 2011; Samani et al., 2020; Xiao, Jiaqi, & Fuji, 2016),mental problems(De Choudhury, Gamon, Counts, & Horvitz, 2013; W. Wang et al., 2016), obesity(Culotta, 2014; Ghosh & Guha, 2013; Mitchell et al., 2013), and urban deprivation(Venerandi, Quattrone, Capra, Quercia, & Saez-Trumper, 2015). The basic methodology of these studies is to extract specific words that can indicate aliments, such as "I got flu" or "sore throat".

One of the most important applications of LBSN data is the prediction of epidemic spreading. In Influenza/flu analysis, LBSN data, such as Twitter and Weibo, are significantly correlated with the government statistic. For example, the correlation reached 0.78 between influenza-related tweets the official statistics from the U.S. Centers for Disease Control and Prevention (CDC) (Culotta, 2010). The model of M. Paul and Dredze (2011) even gained a 0.958 correlation coefficient with the CDC data. Moreover, they can forecast flu outbreaks about one or two weeks earlier than the government data. For instance, Xiao et al. (2016)detected the flu states and the outbreak time using Weibo data in China. They not only counted the temporal variation of these messages, but they also added sentiment features to the prediction process for improving the precision of flu state detection. Their prediction of peak time was two weeks earlier than the official data.

Some scholars also found that food-related words or POIs from LBSN data can estimate the obesity rate at a larger scale(Ghosh & Guha, 2013; Mitchell et al., 2013). Culotta (2014) analyzed tweets from the 100 most populous counties in the U.S and found out the distribution of six health issues that were significantly positive correlate with the relevant tweet words. Fried et al. (2014) also demonstrated that the spatial characteristics of food-related posts could predict the overweight rate and the diabetes rate. Moreover, some large-scale habits can also be overserved by LBSN data. For example, H. Ullah et al. (2020) studied the general temporal pattern of people's activities in green parks, Shanghai. Kershaw, Rowe, and Stacey (2014) examined patterns of alcohol consumption in the UK using a six-week Twitter

(**Figure 37**). Similarly, they retrieved the frequency of alcohol terms in tweets and generated a Social Media Alcohol Score (SMAS). The result from Twitter showed a 0.97 accuracy at the regional level against the ground truth from the official statistics.



Note: Green line: Ground Truth data. Blue line: daily SMAI. Red Line: 7 point moving average.

Source: Kershaw et al. (2014)

Figure 37 Daily SMAI for whole of UK over six week

Furthermore, LBSN data can also be beneficial for studying mental health issues. For example, W. Wang et al. (2016) studied the work stress across the United States using a whole year's tweets. Although they only investigated high-frequency words related to work stress, they proved the potentials of LBSN data for urban health research. For example, their results indicated the relationship between types of daily activitie6s and stress. For instance, compared with Friend or Food, Money, Achievement, and Work brought more negative feelings to people.

Regarding urban well-being, it could be divided into urban built environments and social-economic situations. On the one hand, LBSN data can be exploited as fast access to public sentiment about the surroundings(Santos, Silva, Loureiro, & Villas, 2020). As we mentioned in the section of urban sentiment, green space and some leisure places could arouse happier feelings. On the other hand, LBSN data could also participate in the evaluation of the urban environment. Venerandi et al. (2015)proposed a supervised approach to compute urban deprivation using Foursquare and OpenStreetMap. They calculated significant urban features of a neighborhood by densities of different categories of POIs. After that, the comparison between the official Multiple Deprivation Index(MDI) and urban features could find the POIs that were associated with the deprivation. Finally, a Naive Bayes classifier was used to identify the deprived areas automatically.

Shelton, Poorthuis, and Zook (2015) investigated mobility segregation using Twitter data from Louisville. Based on the neighborhood unit, they grouped Twitter users by their traces of activities. They concluded that LBSN data could describe the phenomenon of inequality. D. R. Davis, Dingel, Monras, and Morales (2019) utilized Yelp data to explore the segregation of urban consumption in New York City. They observed the segregation from five aspects: spatial distance, racial and ethnic demographics, income, and crime level. The result demonstrated that social friction has a larger impact on consumption segregation that the spatial friction, after controlling the preference of tastes.

Table 15 Summary of representative studies of urban health and well-being

Publication	Journal / conference	Summary of application	Location	Data			Major methods	Granularity of analysis
				Source	Range	Volume		
W. Wang et al. (2016)	Applied Psychology	Studying the weekly variation of work stress	USA	Twitter	2009/05/25 - 2010/10/18	2,102 M Tweets from 46,908,115 users	LIWC(linguistic inquiry word count)	Day of week
Kershaw et al. (2014)	ACM	Investigating the temporal patterns of alcohol consumption at national and regional level	UK	Twitter	2013/11/27-2014/01/09	31.6M	Correlation analysis	Region , Nation
Venerandi et al. (2015)	ACM	Creating urban feature metrics of neighborhood to discover the areas of urban deprivation	London, Manchester, West Midlands, UK	Foursquare , OpenStreet Map	2009-2014/04/08	32,003,852 Foursquare check-ins, 131,549 nodes of OpenStreetMap	Offering advantage metric; Spearman's rank correlation	Ward polygon
Culotta (2014)	ACM	Investigating the correlation between Twitter activities and health issues	USA	Twitter	2012/12/05-2013/08/31	130M	Correlation analysis	County
Fried et al. (2014)	IEEE	Analyzing the spatiotemporal patterns of language of food via Twitter	USA, global	Twitter	2013/10/02-2014/05/29	3,498,749 tweets with specific food terms of hashtags	LDA	State and city level
Xiao et al. (2016)	Neurocomputing	Predicting the temporal pattern of influenza spreading in China	China	Sina Weibo	2014/01-2015/01	4M flu-related tweets	Support vector machine (SVM) with restricted Boltzmann machine	Province
M. Paul and Dredze (2011)	AAAI	The applications of Twitter data in the study of public health	USA	Twitter	2009/05-2010/10	0.5M tweets	Ailment Topic Aspect Model	State

Culotta (2010)	Proceedings of the first workshop on social media analytics	The relationship of temporal variation between flu-related tweets and CDC data	USA	Twitter	2010/02/12 – 2010/04/24	574,643 tweets	Linear regression	Nation
Shelton et al. (2015)	Landscape and urban planning	The segregation of residents' spatial activities	Louisville, Kentucky, USA	Twitter	2012/06-2014/07	5.7 M tweets	Frequency of spatial activities	Neighborhood
D. R. Davis et al. (2019)	Journal of Political Economy	Investigating the segregation of urban consumption	New York City	Yelp	2005-2011	18,015 reviews	Logit regression	Census tract

Source: own elaboration

IV.8. Event detection

In the current, the large volume of LBSN data makes the identification of events easier than never before. The event detection based on LBSN data can be classified into two general types: one natural disaster, such as earthquakes or hurricanes; another is the social event, such as stock market or public events. Becker, Naaman, and Gravano (2011) proposed a definition of "event " under the background of Twitter data : "a real-world occurrence *e* with (1) an associated time period T_e and (2) a time-ordered stream of Twitter messages M_e, of substantial volume, discussing the occurrence and published during time T_e." They also established four features for detecting real-world events within a certain period at a cluster level: volume of messages, social interaction between users, topical coherence, and Twitter-centric. Their ontological method is well summarized the core concepts of the event detection.

Regarding social movements, Tsou et al. (2013) designed a Spatial Web Automatic Reasoning and Mapping System (SWARMS) to analyze the spatial distribution of web pages and social media. They utilized the 2012 U.S Presidential Election as a case study to reveal the correlation between online activities and real campaign events. It validated that LBSN data could be used to measure social activity quantitatively. Juris (2012) studied Occupy Boston movement 2012 and #Occupy movement on Twitter in terms of a social-political view. The author argued that social media have contributed to aggregate individuals in the physical space. Although it does not provide any quantitative models, the field observation of such social event is vital for understanding the interaction between the cyberspace and the real world. Moreover, rather than a tool of observation, AlSayyad and Guvenc (2015) considered that social media and communication networks played a role in the ripple effect of Arab Spring.

LBSN can also be a useful source for detecting natural disasters. one of the most cited research is from Sakaki et al. (2010). They investigated the diffusion of earthquake-related tweets and developed a probabilistic model to predict the location of earthquakes. Their approach could detect 96% of recorded earthquakes from the Japan Meteorological Agency. However, their study does not discuss the

temporal variation of responses to earthquakes. This question was solved by Crooks, Croitoru, Stefanidis, and Radzikowski (2013). They analyzed the temporal pattern of Twitter's response to the Mineral, VA earthquake in 2011. They concluded that Twitter users could be a "sensor" to report the instant location of the earthquake, though the percentage of irrelevant tweets appeared soon after a few minutes.

Table 16 Summary of analyzed studies of event detection

Publication	Journal / conference	Summary of application	Location	Data			Major methods	Granularity of analysis
				Source	Temporal Range	Volume		
Crooks et al. (2013)	Transactions in GIS	Investigating the effectiveness of Twitter data as a source to detect the location of earthquake.	USA	Twitter	Eight hours in 2011/08/23	21,362 geo-located tweets	Temporal analysis, signal-to-noise ratio	Point
Tsou et al. (2013)	Cartography and Geographic Information Science	Mapping social events in the case of US election	30 US cities, USA	Twitter, Yahoo! API, Bing API	Twitter: 2012/06/25-2012/11/05 ,Webpages: 2011/12/18-2012/11/07	16.7M tweets, 44,200webpages	Temporal analysis, words clustering	State
Becker et al. (2011)	AAAI	Introducing an ontological model for online event detection	New York City, USA	Twitter	2010/02	2.6 M	Incremental clustering; Classification(Naive Bayes, SVM, logistic regression)	City
Juris (2012)	American ethnologist	Explaining the interaction between online social movement and physical assembling of people	Boston USA	Twitter Ground Truth observation	2011	Not mentioned	Qualitative analysis	City
AlSayyad and Guvenc (2015)	Urban Studies	Analyzing the active interaction between media, social movements, and urban space	Middle East	Twitter	2011	Not mentioned	Qualitative analysis	Nation
Sakaki et al. (2010)	IW3C2	Detecting earthquakes by Twitter	Japan	Twitter	2009	Thousands	SVM, Kalman filter, Marcov process, Bayes rules	Nation

Source: own elaboration

IV.9. Summary

The section provides an overview of the applications of LBSN data in urban studies. An increasing number of related publications can be postulated, especially in the field of computing. About 20% of reviewed articles are generated from top computing conferences. Thanks to the growing availability of LBSN data, a new disciplinary field – urban computing is evolving rapidly(Y. Zheng et al., 2014). Therefore, unsurprisingly, more studies focus on the advancement of computing techniques than urban issues.

Meanwhile, our result also shows a variety of journals that utilize LBSN data to investigate different urban questions. They provide hundreds of observations for urban problems, whereas the diversity of each observation could lead to the deconstruction of a universal theory, just as Jessop, Brenner, and Jones (2008) said. However, it probably opens a new paradigm for urban studies and urban management. Firstly, the spatiotemporal reliability of LBSN data is proved to be enough in disaster management and urban health issues, as comparing with the official data. The large volume of spatiotemporal behaviors of LBSN users could reach a fine-enough precision in predicting urban functional regions(J. Yuan et al., 2012) and citizens' mobility(X. Hu et al., 2017). Although the precision of location from LBSN data exists errors(Valls Dalmau, 2019), many studies have confirmed the accuracy at the scale of neighborhood and city level(Abbasi & Alesheikh, 2018; Cho et al., 2011).

Secondly, LBSN data provides a new perspective to analyze the urban structure through nearly real-time human activities. The social-spatial difference among different groups of people can be captured by LBSN data. It promotes the study of urban structure toward a segment view in social and spatiotemporal dimension. Moreover, various LBSN data sources (**Table 12**) indicate the effectiveness of LBSN data in the study of human mobility and urban structure.

Thirdly, leveraging LBSN data, the urban sentiment/ perception analysis can generate a sketchy evaluation of public perception towards events or the urban environment. We could observe some nonrandom relationship between semantic information and places, including the positive correlation between green space and public sentiments, and the relationship between wealth and positive sentiments.

Regarding data sources, Twitter plays a dominant role among reviewed papers, especially in the field related to semantic analysis. Moreover, Twitter data are mainly retrieved from the United States. Several reasons may lead to the result. Firstly, our searching range only focuses on English articles, and therefore many studies using other languages are neglected. Secondly, as previously stated, the accessibility of Twitter data is higher than other data. For example, the database of W. Wang et al. (2016) collected a random sample of 10 percent of the Tweets across the US over one year. Moreover, the richness of data contents and degree of popularity of Twitter is higher than Foursquare. Thirdly, the availability of other data sources, such as the official statistics and cartography, also affects the feasibility of the study because more and more scholars attempt to combine LBSN data with other sources.

Concerning methods, almost all reviewed papers adopt the spatiotemporal analysis according to the aim of the study, the characteristics of the data and the study areas. It is difficult to compare and summarize a set of common approaches for spatiotemporal analysis. However, several methods are popular among all urban issues: network analysis, clustering analysis, and topic modelling. Besides, thanks to the development of computing software, more sophisticated algorithms and data fusion are involved in researches. It enhances the precision of LBSN data meanwhile requires scholar to manipulate more computing skills.

Without doubts, utilizing LBSN data also faces several challenges and potential bias, such as the representativeness of the population and the heterogeneity of methods. These issue will be discussed in the next chapter.

Chapter V.

Limitations and representativeness of LBSN data

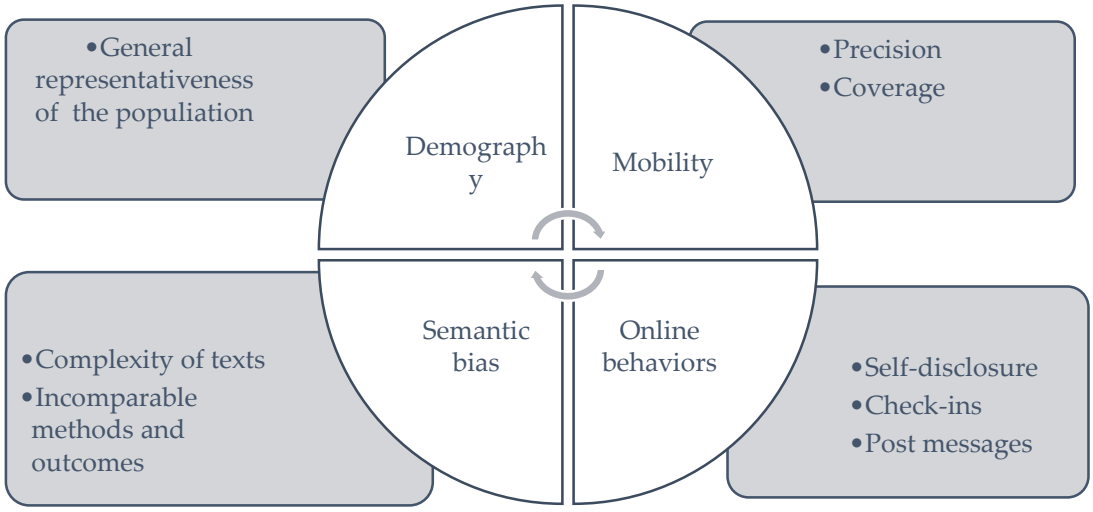
Despite the powerful advantages of LBSN data, it is necessary to mention the limitations of LBSN data that has been under debate for a long time. The representativeness of LBSN data refers to the degree to which the data can represent the underlying facts, such as the numbers of population, the trajectories of human mobility, and the public perception. Mounting academic investigations have suggested the bias and representativeness of LBSN data that could distort the real situation of the physical world (Ruths & Pfeffer, 2014); Tufekci (2014)and affect the result of the investigation(De Choudhury et al., 2010; Kossinets, 2006; Murthy, Gross, & Pensavalle, 2016). The controversy is caused by the peculiarity of LBSN data itself.

Nevertheless, numerous studies have proven that LBSN data could be representative to some degree. Statistically, on the one hand, the "law of large number" (average of the results obtained from samples converges to the average of the population as more samples involved) guarantees a large-enough random sample dataset that can describe characteristics of the population. Therefore, LBSN data could generate satisfying results from a statistical view because these data usually consist of thousands of observations and the monitoring period could last several years.

On the other hand, Kruskal and Mosteller (1980) defined nine meanings of "representative samples" and "representative sampling" that can be summarized as the following: 1) it is a general data without selective forces; 2) the sample describe the population in miniature, which contains the important characteristics of the population; 3) the sample with good coverage that reflects the variations of the population; 4) the sample serves well a particular purpose; 5) the sample is derived from specific sampling methods. In conclusion, representative samples

are not equal to probability samples, and thus it may not require a properly probabilistic sampling. The definition of "representativeness" more depends on specified characteristics of the population and the purpose of the investigation.

Therefore, it is more appropriate to examine the representativeness of LBSN data under specific situations and domains(**Figure 38).** Except for the general representativeness of the data, this section discusses typical bias of LBSN data that are related to urban spatiotemporal analysis: demography, spatial distribution, human mobility, human online behaviors, and semantic analysis. Concerning the complexity of different types of LBSN application, we highlight the three studied applications of the dissertation: Twitter, Foursquare, and Sina Weibo.



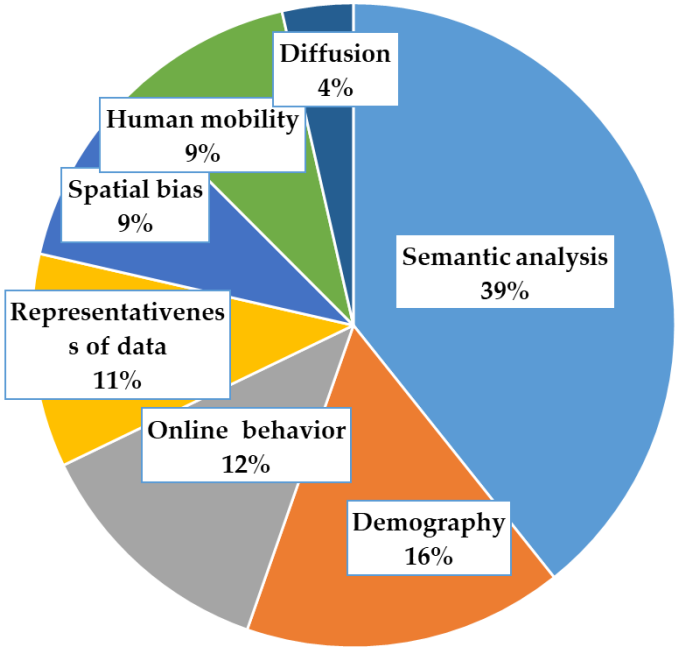
Source: own elaboration

Figure 38 Division of representativeness of LBSN data

V.1. A survey of the literature review

This dissertation reviews 56 papers that are related to the limitation and representativeness of LBSN data. **Figure 39** demonstrates the composition of these articles, of which 39% discusses the semantic analysis. Diffusion is about how

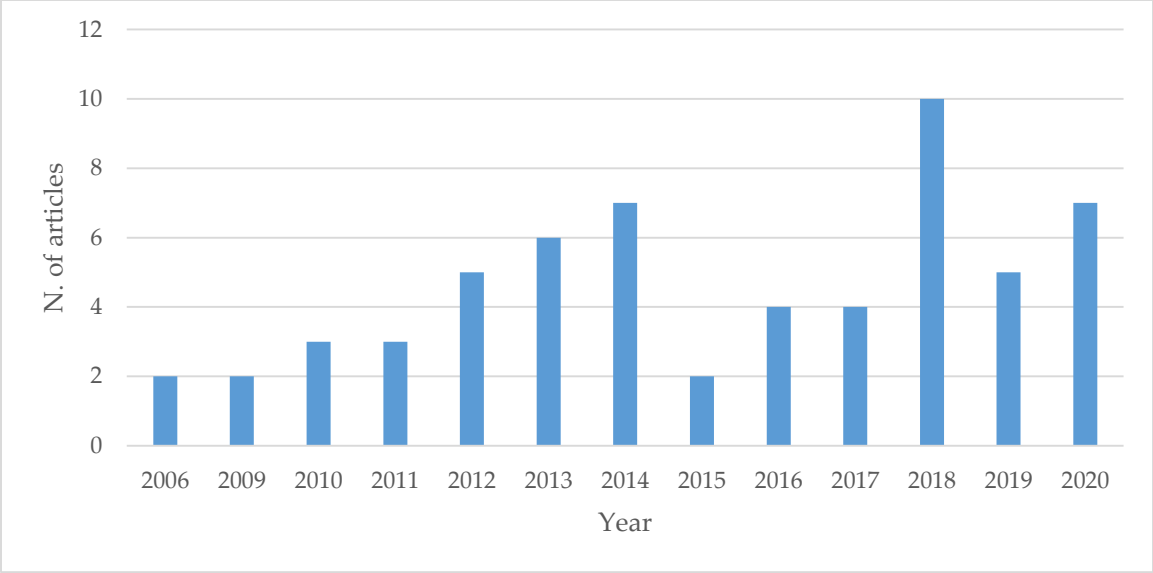
information spread via social media networks, which is not equal to personal online behaviors. Six articles analyze the general representativeness of LBSN data, such as the sampling issue and the inconsistency of data(Martí et al., 2019).



Source: own elaboration

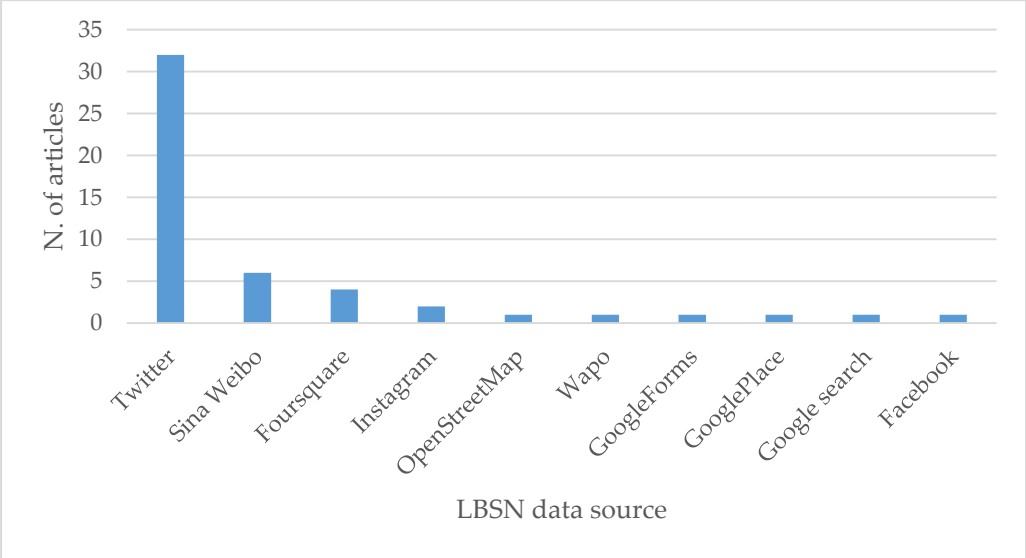
Figure 39 Breakdown of reviewed articles regarding LBSN limitations

In general, the study of LBSN limitations increases with time(**Figure 40**), which indicates that more and more scholars noticed these problems and devoted efforts to solve them. However, the available data are still mainly from Twitter(**Figure 41**) that accounts for more than 50% of all data sources. The following is Sina Weibo and Foursquare.



Source: own elaboration

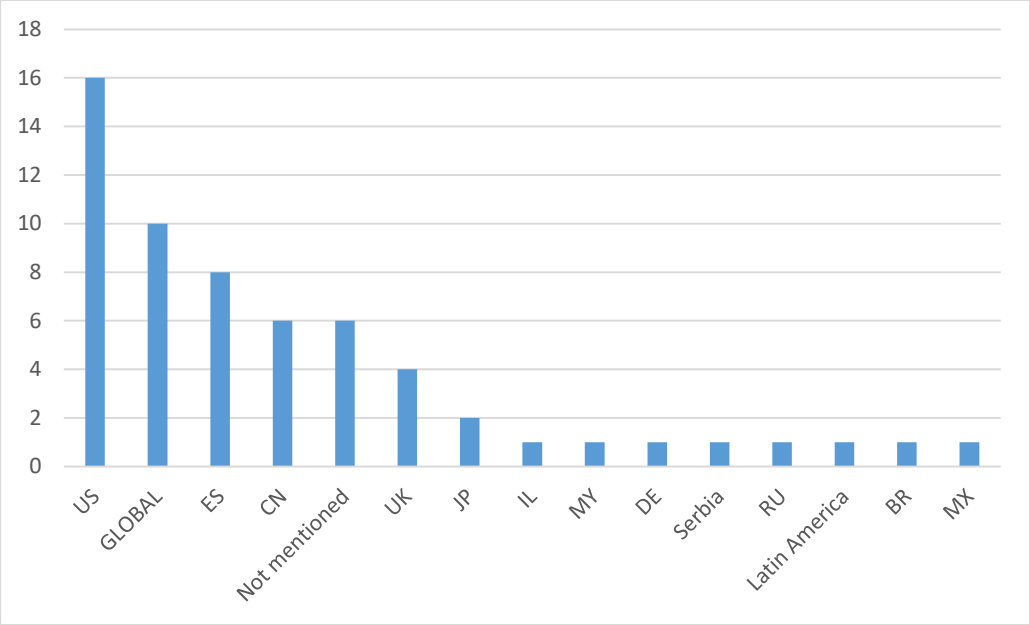
Figure 40 Year-distribution of reviewed papers



Source: own elaboration. Note: Wapo application: http://wapa-app.com/index_en.html

Figure 41 Frequency of LBSN data sources

Figure 42 shows the geo-spatial distribution of these studies. Most datasets are located in the United States. Ten investigations cover the global scale. Six studies related to semantic analysis do not disclose their study locations and range because they mainly focus on linguistic problems.



Note: CN: China. IL: Israel. MY: Malaysia. RU: Russia. Latina America: five countries of Latin America, see Hubert, Estevez, Maguitman, and Janowski (2018).

Figure 42 Distribution of studied countries

For further investigating the correlation between geo-locations and these studies, **Figure 43** displays the distribution of each topic in different countries. Although most semantic analyses are located in the United States, this domain presents the most diverse places of case studies. One possible explanation is that it does need to cooperate with other types of data in some cases. For other issues, such as human mobility or spatial bias, they require a more completed dataset as the reference to compare the difference. Therefore, it is no wonder that the United States is the leading country in LBSN data studies.

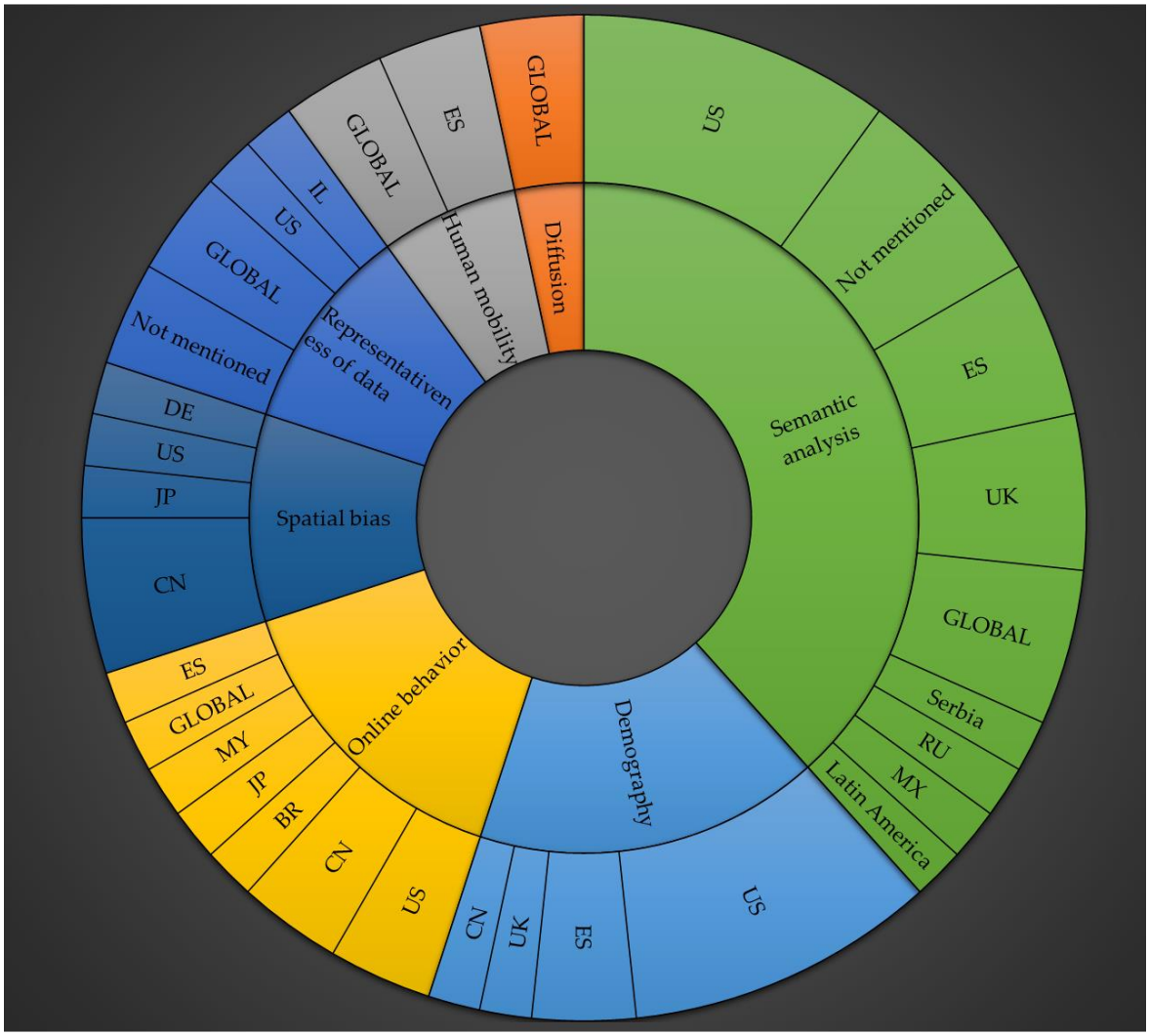
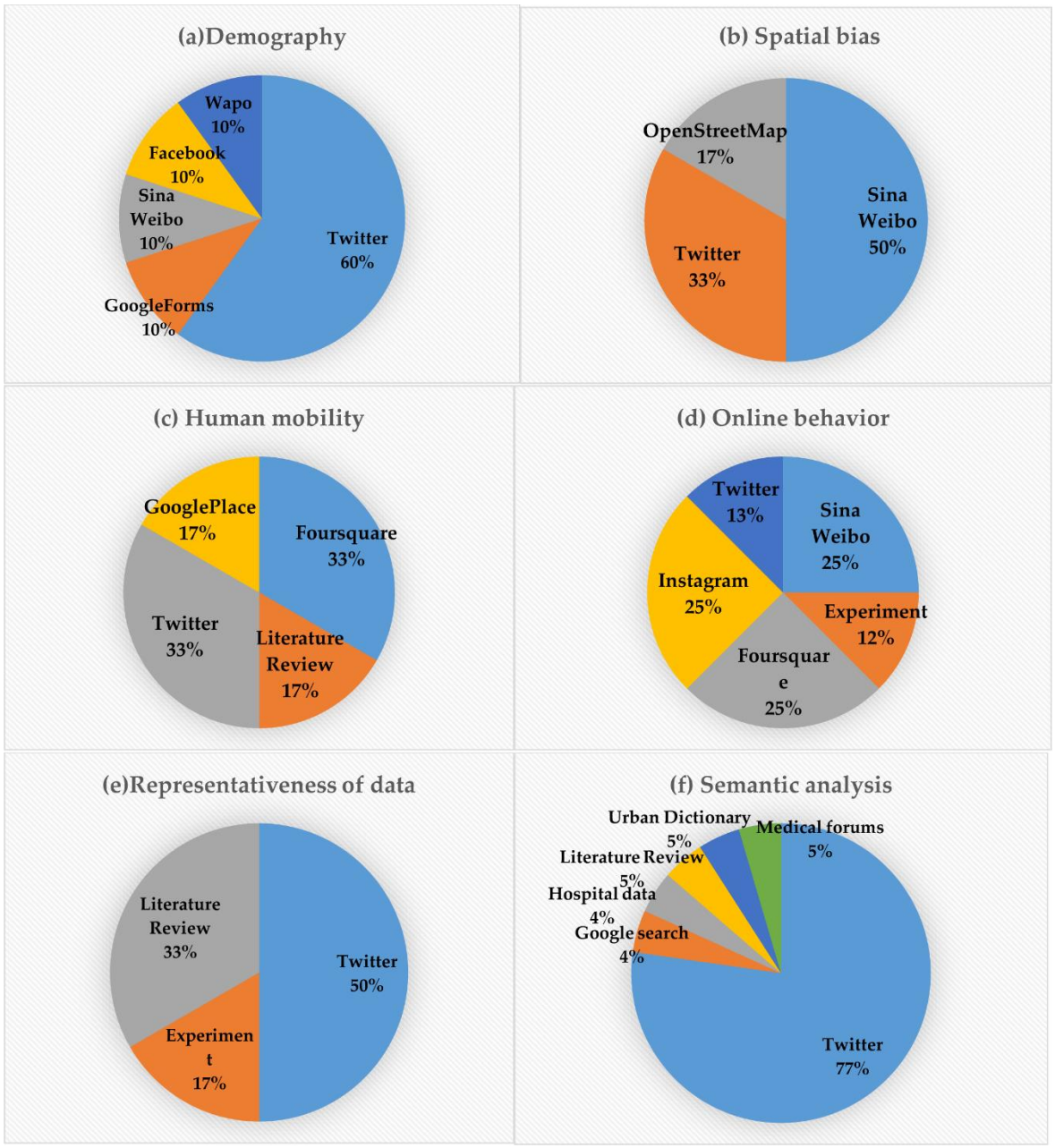


Figure 43 Country-distribution of study cases

Figure 44 lists the distribution of data sources among different sub-topics. Popular LBSN data appears in all domain and large-scale studies. By contrast, some high-specialized datasets are only utilized in semantic analysis(**Figure 44** – (f)). On the one hand, the widely used Twitter datasets indicate that the reliability of Twitter data has been validated to some degree. On the other hand, it means that other types of LBSN data are waiting for further investigations.



Source: own elaboration

Figure 44 Distribution of LBSN data in each domain

V.2. General limitations of LBSN data

2.1. Limitations of the data

Firstly, LBSN applications are constantly changing at a fast pace, in all aspects of properties: the degree of popularity, coverage of data, demographic structure of users, and among others. Unlike the national census or other official survey, LBSN data could reflects the dynamics of users instantly, and thus it is sensitive to the time spam and social changes. For example, the popularity of LBSN applications will impact on the precision of event detection. The rising and declining of numbers of active users affect the representativeness of the data. On the other hand, the characteristics of content have a great impact on the quality of sampling in the study of diffusion(De Choudhury et al., 2010) because it strongly influences whether the user diffuse it or not. The spreading distance of international topics is obvious longer than local news.

Secondly, the large volume of LBSN dataset reduces its representativeness to some degree. Because of the huge volume of data, it is impossible to access the completed dataset for researches. For instance, Twitter has a rate of approximately 6000 tweets per second[24]. It means high economic cost for data collection and storage. Therefore, academic studies often use APIs to access the limited public data[25]. The sampled Twitter API roughly provides 1% of all data to the public. Sina Weibo does not mention how many messages it offers, however, it set the limitation on the number of accessing the API. However, the algorithm of public data is not available for most of LBSN companies. The mechanism of filtering and providing data is left in the dark, though Morstatter et al. (2013) found out that the free API of Twitter almost returned completed geotagged tweets.

The complexity also reflects in the characteristics of the population that LBSN data possibly represent, given the growth of social media applications. For example, different groups of people may express contradict opinions and online

[24] https://www.internetlivestats.com/twitter-statistics/

[25] https://developer.twitter.com/en/docs/twitter-api/tweets/sampled-stream/introduction

behaviors, according to different periods, places, and events. Therefore, the inconsistency of data is another highlighted limitation(Martí et al., 2019). Besides, the complexity of daily language is a great challenge to conduct semantic analysis for investigating the motivations, perceptions, and explanations behind the online behaviors. People increasingly use pictures, short abbreviations, and emoticons (Sampietro, 2016)when they are communicating on social media platforms. For example, 75.5% of millennial respondents would like to use emoji during online communication(Bosch & Revilla, 2020). Although many researches considered the issue in semantic analysis, it still requires more effort to identify and coherent them with the texts appropriately.

2.2. Incomparability of datasets and methods

The heterogeneity is the most prominent characteristic for LBSN datasets. Firstly, proprietary information and algorithm affect the quality of the dataset and the result. Researchers who can get advanced access to the dataset are probably able to obtain a high-quality dataset. It raises sampling problems in the analysis of hashtags of tweets (Morstatter, Pfeffer, & Liu, 2014). For example, Morstatter et al. (2013) compared the ranking of hashtags between two datasets from Twitter: free Twitter API and the Firehose – the commercial API that provides 100% of all public tweets. They found out that the two rankings are not consistent in sorting the top hashtags.

Secondly, we can only retrieve part of the whole data in a specific period and spatial region, and thus the dataset depicted the details of the corresponding place and period. Therefore, each specific study is limited to a unique social and spatial environment. It is strenuous to compare results from different datasets, especially for semantic analysis. Although other features of LBSN data, such as places and users, are also limited by the diversity, the content is more sensitive to the diversity of everyday life. From social events(Saura, Reyes-Menendez, & Palos-Sanchez, 2019) to the weather of the day(Padilla et al., 2018), all factors might influence the semantic performance. For example, Twitter sentiment is strongly related to the socioeconomic bias in the United States(Mitchell et al., 2013). However, it needs more studies from different places to prove if the conclusion is transferrable.

V.3. Demographic representativeness

The first doubt of the representativeness is whether a large LBSN dataset of users is enough to reflect the demographic characteristics of people, such as age, gender, race, income, among others. Obviously, it is impossible to connect every user's account with his/her true identity of the offline world. Therefore, the demographic background of LBSN users is not as accurate as census or specific surveys.

Up to current investigations, the gender balance is quite different among different social media applications due to the function of the application. For example , **Table 17** shows the demographic preferences among different groups of people in US. The first row is the average use rate of social media applications among U.S. adults. The rest rows are the average use percentage regarding different groups. Platforms with more female users are Pinterest, Instagram, Facebook, which tend to be more focused on life styles and daily life. One the other hand, applications contains many discussions of social and political issues attract more male users, such as Reedit and Twitter.

Table 17 Demographic compositions of popular social media platforms of U.S.

Use of different online platforms by demographic groups

% of U.S. adults who say they ever use the following online platforms or messaging apps

	YouTube	Facebook	Instagram	Pinterest	LinkedIn	Snapchat	Twitter	WhatsApp	Reddit
U.S. adults	73%	69%	37%	28%	27%	24%	22%	20%	11%
Men	78	63	31	15	29	24	24	21	15
Women	68	75	43	42	24	24	21	19	8
White	71	70	33	33	28	22	21	13	12
Black	77	70	40	27	24	28	24	24	4
Hispanic	78	69	51	22	16	29	25	42	14
Ages 18-29	91	79	67	34	28	62	38	23	22
18-24	90	76	75	38	17	73	44	20	21
25-29	93	84	57	28	44	47	31	28	23
30-49	87	79	47	35	37	25	26	31	14
50-64	70	68	23	27	24	9	17	16	6
65+	38	46	8	15	11	3	7	3	1
<$30,000	68	69	35	18	10	27	20	19	9
$30,000-$74,999	75	72	39	27	26	26	20	16	10
$75,000+	83	74	42	41	49	22	31	25	15
High school or less	64	61	33	19	9	22	13	18	6
Some college	79	75	37	32	26	29	24	14	14
College+	80	74	43	38	51	20	32	28	15
Urban	77	73	46	30	33	29	26	24	11
Suburban	74	69	35	30	30	20	22	19	13
Rural	64	66	21	26	10	20	13	10	8

Note: Respondents who did not give an answer are not shown. Whites and blacks include only non-Hispanics. Hispanics are of any race.
Source: Survey conducted Jan. 8-Feb. 7, 2019.

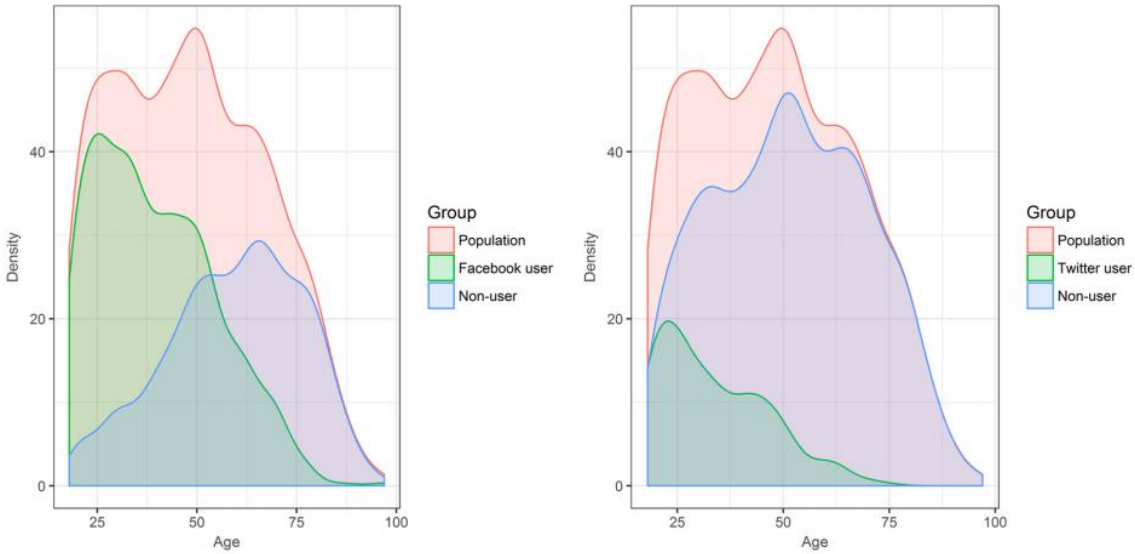
PEW RESEARCH CENTER

Source: PEW RESEARCH CENTER. Retrieved from: https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/ft_19-04-10_socialmedia2019_useofdifferent/ .

The general demographic information is published by these social media companies, which is summarized by the users' registration information. Users of TikTok are the youngest among current popular applications, 41% of them are aged between 16 and 24[26]. Twitter users are older than TikTok, in which 57% of global Twitter users are between 25 and 49 years-old. 75% of users of Sina Weibo

[26]Source: https://www.omnicoreagency.com/tiktok-statistics/#:~:text=41%20percent%20of%20TikTok%20users,26%25%20between%2018%20and%2024.

is between 18 and 30. From these data, we can infer that young people are more represented than elder people (over 50 years-old) on social media data. Therefore, these LBSN data do not have general representativeness of the population. For example, Mellon and Prosser (2017) studied the political attitudes and the demographics of British social media users using 2015 British Election Study (BES)[27]. They concluded that neither Twitter nor Facebook was demographically representative of the population(**Figure 45**). The average age of Facebook and Twitter users were younger that the average age of the population.



Source: Mellon and Prosser (2017)

Figure 45 Age distribution of Twitter and Facebook users from UK

Moreover, LBSN users are not randomly distributed over population. Mislove, Lehmann, Ahn, Onnela, and Rosenquist (2011) also mentioned that Twitter data over-presented the populous countries in the United States because the ratio between the number of Twitter users and the number of population was increasing with the increment of population.

To cure the above problems, it is possible to deduce the demographic information of a dataset through extracting a sampling dataset and identifying sample users' age, gender, and ethnicity by fully monitoring their tweets, photos,

[27] Source: https://www.britishelectionstudy.com/data-object/version-3-0-2015-face-to-face-post-election-survey/

and related social media accounts (e.g. Facebook and Instagram) (Murthy, et al. 2016; Hargittai and Litt ,2011). However, the identification totally depends on the subjective estimation, not to mention its potential privacy and moral concerns.

Another method to correct the issue is to narrow down the study scope to a certain group of people. The identity of these specific groups of people can be deduced by their public information and online behaviors, such as tourists and locals. For example, Da Rugna *et al.* (2012) showed that geotagged photos on Flickr could identify the original country of tourists. They counted the number of countries that each user visited from 2010 to 2011 and calculated users' total length of stay in those countries. The country that a user stayed longest was considered as his/her original country. However, as the paper noticed, the method would fail if users did not make enough check-ins in their home countries. Using data from specific applications is another strategy to distinguish a specific group. For example, some dating applications is good way to study the activities of homosexual people(Castelló, Baeza, & García, 2018) .

V.4. Bias of human online behaviors

Online behaviors refer to people's personal and interpersonal actions on the Internet, such as sharing location, using hashtags, and online communications. Rom and Alfasi (2014) concluded that people are likely to perform the same behaviors in both online and offline social interactions in Israel. Furthermore, Individuals may tend to share their intimate thoughts (Schouten, Valkenburg, & Peter, 2009) online while they are bound to be affected by social desirability(Amichai-Hamburger & Vinitzky, 2010).

First of all, as stated previously, it is related to the demographic background at a macro level. Young female users tend to disclosure their geo-location on LBSN platforms, according to studies in China(Y. Yuan, Wei, & Lu, 2018) and the United States(Haffner, Mathews, Fekete, & Finchum, 2018). The higher ratio of female online-performance is positive correlated with women's social status and income(Y. Yuan et al., 2018).However, it seems that gender do not have an evidential influence on the spatial density of Twitter and Flickr(Goodchild & Li,

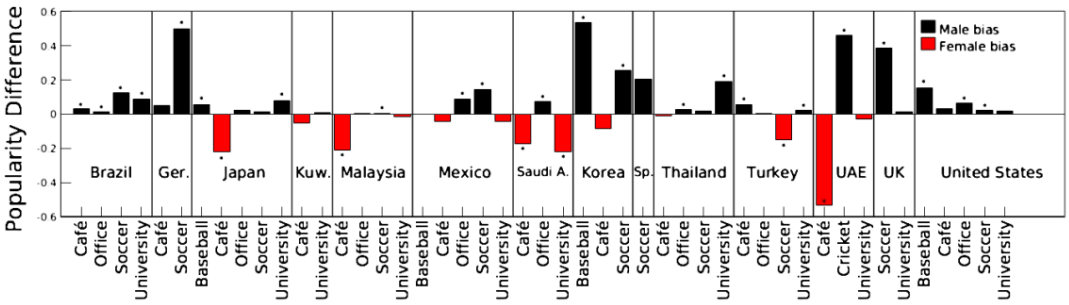
2012). Moreover, gender does not present significant difference among the total population in developed nations and big cities(Mellon & Prosser, 2017; Mueller, Silva, Almeida, & Loureiro, 2017).

Regarding expressing opinions, politically active Twitter users slanted to the males who lived in urban areas with extreme political preferences, according to the study of the 2011 Spanish elections and the 2012 US presidential election(Barberá & Rivero, 2015). However, it is worth noting that there are no significant differences between social media users and non-social media users on political attitudes and behaviors, after controlling for age, gender, and education (Mellon & Prosser, 2017). The gender difference also reflects in the habitual of expression. For example, in Malaysia, male users likely use informative hashtags to comment on the food, such as the style and taste; while female users tend to use adjective words to express their feelings about the food, such as great, nice, and delicious.(Ye, Hashim, Baghirov, & Murphy, 2018)

Secondly, the regional difference also seems to influence online behaviors(Q. Gao, Abel, Houben, & Yu, 2012; Mueller et al., 2017). Q. Gao et al. (2012)tried to use Hofstede's cultural dimensions theory[28] to explain different microblogging behaviors on Weibo and Twitter, such as the frequency of using hashtags and the tendency of disclosing activities. They found that Sina Weibo users were more active on weekends than Twitter users, vice versa. However, it is quite debatable that the culture difference would cause the difference in posting frequency because it is even difficult to define the culture difference. Mueller et al. (2017) showed the gender difference in term of popular places of Foursquare check-ins in fourteen nations (**Figure 46**). At the first glance, it seems that the gender preference of check-in behaviors indeed exists among different countries. However, the same category, for example, cafeteria, appear different gender orientations in Japan and the United States. If we limit the scope of study to a

[28] The theory builds up a framework of cultural values to analyze different cultures, and relate these values to people´s behaviors. It includes consists of six major dimensions: social power distance, individualism/ collectivism, the tolerance for uncertainty, masculinity/femininity, long-term orientation/short-term orientation, and social norms orientation (Indulgence/restraint). See more details: Hofstede (2011)

country, the result might be different. Therefore, it is not a gender difference but a regional difference.



Source: Mueller et al. (2017)

Figure 46 Difference of check-ins between gender

Thirdly, the design of social media platforms also influences users' behaviors. Different applications have different digital behaviors due to the function and characteristics of applications (Silva et al., 2013). For example, Users of Instagram tend to publish more photos than Twitter users because the aim of Instagram is to share pictures.

In a nutshell, daily life is full of details and complex social relations, so does online behavior. LBSN data reminds us that the complexity. Therefore, it is vital to clarify the scale of the study and the research object for controlling the demographic and socioeconomic heterogeneity.

Table 18 Summary of representative studies of demographic bias

Publication	Journal/ conference	Summary of the study	Location	LBSN data
L. Chen et al. (2019)	Personality and Individual Differences	Evaluating the degree of self-disclosure in Sina Weibo	CN	Sina Weibo
Castelló et al. (2018)	WIT Transactions on The Built Environment	Identifying active areas of online dating	ES	Wapo
Barberá and Rivero (2015)	Social Science Computer Review	Understanding the political representativeness of Twitter users in Spain and US election	ES, US	Twitter
Mellon and Prosser (2017)	Research & Politics	The representatives of political attitudes of social media users in UK	UK	Twitter, Facebook
Haffner et al. (2018)	Geographical Review	The demographic difference of LBSN users among college students	US	Survey of LBSN users
Hargittai and Litt (2011)	New media & society	The effect of race in Twitter adoption among a diverse group of young adults	US	Twitter
Malik, Lamba, Nakos, and Pfeffer (2015)	ICWSM	How demographic factors influence the distribution of geotagged users	US	Twitter
Mislove et al. (2011)	ICWSM	Comparing the Twitter population and U.S. population in terms of (geography, gender, and race/ethnicity.	US	Twitter
Murthy et al. (2016)	Journal of Computer-Mediated Communication	Studying urban Twitter demographic profiles in U.S.	US	Twitter

Source: own elaboration

V.5. Representativeness of spatial human mobility

5.1. Precision of LBSN location

The representativeness of human mobility can be evaluated in terms of precision and coverage. LBSN applications determine a message/user's instant positions through Wi-Fi and GPS localization which perform better than traditional questionnaires and cellular networks (**Table 19**). On the other hand, the positioning method also contributes to the heterogeneous distribution of LBSN data because the communication facilities are usually well-equipped in urban areas.

Table 19 Comparative summary of different data collection techniques

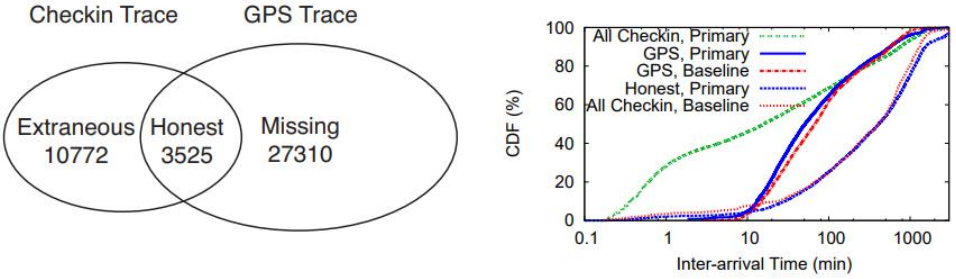
Methods	Advantages	Disadvantages
Survey	• Multi-purposed use	• Expensive to collect observations • Not accurate
Wi-Fi localization	• Accuracy • Energy usage ~ 50% GPS	• Low coverage area • Providing access point is expensive
GPS localization	• Highly precise ~ 5m error - • Can distinguish between transportation modes	• High battery (energy) usage • Expensive • Low quality signals in indoor environment
Cellular network (passive) (Call Data Records)	• Automatically generated	• Sparse in time localization • Needs more filtering • Less accuracy (~ 175 m error)
Cellular network - localization (active)	• More accuracy than passive localization • Less expensive than previous methods	• More expensive than passive form • Arise the issue of large database

Source: Asgari et al. (2013)

In terms of technical level, the positioning precision of LBSN data is better than traditional surveys and cellular networks in cities. However, LBSN data is less accurate when other factors are involved. Firstly, since LBSN data is user-generated without demographic validation, the representativeness is weaker than surveys and records of cell phones which are based on reliable ground-truths. For instance, according to the result of Z. Zhang et al. (2013), 90% of actually visited

GPS records were missing from the check-in records of Foursquare **(Figure 47),** due to users' concern of privacy and their personal preferences of check-ins. However, it is worth noting that they only extracted 244 sample users from worldwide Foursquare users. This small sample might affect the creditability of their result. Meanwhile, 75% of Foursquare check-ins did not match their actual GPS traces. They found out that the rewarding rules of Foursquare was the major reason that caused "falsely" check-ins. Similarly, G. Wang, Schoenebeck, Zheng, and Zhao (2016) also observed the phenomenon that users tended to make more "fake" check-ins on places where they have not visited.

Secondly, the change of privacy policy also affects the accuracy of open data. For example, according to Twitter's official documents, they encourage users to add their location in the form of a general location label, instead of the GPS position[29]. The number of GPS –tagged tweet reduced significantly after the policy change(Tasse et al., 2017).



（1）Matching result of check-ins （2）Cumulative Distribution Function(CDF) of inter-arrive time.
Source: Z. Zhang et al. (2013) Note: inter-arrive time is the time interval between two footprints.

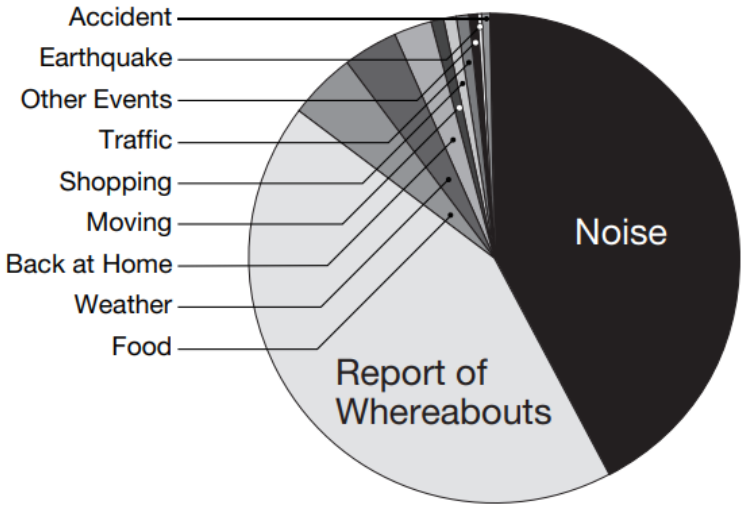
Figure 47 Results of Foursquare check-ins against GPS traces

Therefore, compared with the census or traditional questionnaire, LBSN data is a useful source to unveil the urban dynamics in an aggregated way rather than an individual level (W. Luo et al., 2019; Valls Dalmau, 2019).

Secondly, LBSN data usually has a higher percentage of data noise than the other dataset (G. Wang et al., 2016). The noise of data consist of: 1) messages

[29] https://help.twitter.com/en/using-twitter/tweet-location

generated from spammer and robots (Stringhini, Kruegel, & Vigna, 2010; X. Zheng et al., 2015); 2) non-meaningful contents(Hiruta, Yonezawa, Jurmu, & Tokuda, 2012); 3) non-meaningful traces that heavily affect the process of analysis. For example, Hiruta et al. (2012) investigated a 2,000 sample of geotagged tweets from Japan, of which 42.5% did not contain any topic contents (**Figure 48**), for instance, only wrote: "Hello!" or "Morning".



Source: Hiruta et al. (2012)

Figure 48 Result of survey of geotagged tweets in Japan

Regarding to the mobility pattern, Lenormand et al. (2014) compared mobility patterns of Madrid and Barcelona based on three different sources: Twitter, census, and cellphone data. They concluded that the results obtained with the three data sources comprise comparable despite the representativeness of Twitter is lower than the others. The three data source detected almost the same relationship between population density and mobility pattern. Therefore, they are interchangeable under a certain spatiotemporal scale. However, in order to obtain a reliable result, Twitter needs to cost more time on data processing. In summary, using Twitter to investigating urban dynamics in metropolis has potential to generate reliable result when the dataset is large enough. "Geo-located Twitter data still yields generalizable results when studies are restricted to populated

metropolitan areas with a high percentage of smartphone users" (Plunz, et al. 2019).

5.2. Bias of uneven spatial distribution

The phenomenon that LBSN data skews toward urban areas has been observed by many scholars(Fan et al., 2020; Hecht & Stephens, 2014). Hecht and Stephens (2014) evaluated the potential urban bias of LBSN data using Foursquare, Flickr, and Twitter in the United States. The result concluded that LBSN data were denser in Metropolitan areas than rural areas, especially Foursquare data. One possible reason is that Foursquare is designed for finding places in the urban area. Fan et al. (2020) found out that disaster-related tweets were concentrated in populous areas of the United States, regardless of socioeconomic conditions. Sina Weibo also seemed to present the similar distribution because the majority of studies using Sina Weibo were target on urban areas (W. Y. Wang, 2020).

In fact, the bias sounds reasonable because LBSN data are generated from users. The equal distribution of LBSN data is against the actual density of human population. The uneven distribution between urban and rural areas is of universality all over the world(Jackson et al., 2013; Morstatter et al., 2013). It is no wonder that most researches related to LBSN data are concentrated in urban areas for obtaining reliable results.

The socioeconomic factors also influence the distribution of LBSN data. Malik et al. (2015) concluded that the non-random distribution of geotagged tweets was correlated with income, ethic, distance to coast, and urban/rural area in the United States. Twitter users were more commonly found in wealthier areas.

Table 20 Summary of representative spatial bias studies

Publication	Journal/ conference	Summary of the study	Location	LBSN data
Lenormand et al. (2014)	PloS one	Cross-checking different sources of mobility: Twitter, census, and cell phone	ES	Twitter
Martí et al. (2019)	Computers, Environment and Urban Systems	Proposing a descriptive framework for the study of urban phenomena through LBSN data	ES	Twitter, Foursquare, GooglePlace
Asgari et al. (2013)	arXiv preprint	Comparing different data types used in tools and applications of human mobility	GLOBAL	Literature Review
Z. Zhang et al. (2013)	Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks	Studying the validity of Foursquare data against GPS traces in terms of human mobility	GLOBAL	Foursquare
Jendryke, Balz, and Liao (2017)	Transactions in GIS	Presenting a framework of LBSN data collection and visualization leverage Sina Weibo data	CN	Sina Weibo
W. Y. Wang (2020)	Handbook of the Changing World Language Map	Identifying the spatial distribution of Cantonese Weibo in Guangzhou	CN	Sina Weibo
Jackson et al. (2013)	ISPRS International Journal of Geo-Information	Assessing completeness and spatial error of features in volunteered geographic information	DE	OpenStreetMap
Hecht and Stephens (2014)	AAAI	Detecting urban Biases in Volunteered Geographic Information	US	Foursquare, Flickr, Twitter
Hiruta et al. (2012)	ACM	Detecting and classifying place-triggered geotagged tweets	JP	Twitter

Source: own elaboration

V.6. Representativeness of semantic analysis

6.1. The uncertainty of evaluation of sentiments and opinions from texts

First of all, in essential, understanding the sentiments and intentions of other people is a subjective judgment. We are unable to precisely certify the true meaning and emotions behind the linguistic expressions. As Ferdinand de Saussure claimed, language is a social fact which is constructed and constrained by the social norms and social environment(De Saussure, 2011). On the one hand, it is hard to confirm what we speak out is precisely what we are feeling because our expressions are also restricted and trained by society. Each individual has a unique way to express emotions and opinions. On the other hand, our personal experiences and social life also affect the prediction of other people's expressions. Therefore, the degree of agreement of sentiment classification between human evaluations is just about 80%.*(Ogneva, 2010)*

Regarding the semantic analysis from LBSN data, it is also built upon human evaluations because the emotion of words and expressions are primarily estimated by linguistic experts. The algorithm of sentiment analysis only maps these estimations to a larger dataset. The computer cannot predict any of the backgrounds of texts, for example, the location and motivation of the text, unless they are "told" by human supervisors. It "understands" the texts through probability. Therefore, the ambiguity that exists in the manually semantic analysis also happens in the automatic semantic analysis. Moreover, The mechanism of self-disclosure on social media platforms has not been fully understood(L. Chen et al., 2019). As online information is highly anonymous, it is difficult to deduce the creditability of these messages.

Thirdly, many active and influential accounts do not belong to non-individual accounts, such as organizations, companies, and news agencies. These accounts should be excluded from the dataset to avoid the distortion of the result if we

attempt to investigate individuals' perceptions. So far, this process heavily relies on human inspection. These problems altogether become a very challenging task in the area of sentiment analysis.

6.2. Problems of LBSN expressions

One of the biggest obstacles to conducting semantic analysis is data preparation. Unlike the edited texts (e.g. books and articles), LBSN texts usually contain many misspelling, grammatical errors, emoticons, colloquial expressions, and slang. Hence, these flaws increase the complexity and noise of the text and influence the precision of expression.

Table 21 lists typical problems in semantic analysis, which require more researches in the future. Emoticons can convey emotions effectively. However, only some simple emoticons can be recognized by analyzing algorithms(Narayanaperumal, 2020; Ullah, Marium, Begum, & Dipa, 2020), such as "☺", ":-(" , and ":-D", because they are characters or have internal codes that computer can read (Bosch & Revilla, 2020). Moreover, as increasingly using emoticons and pictures instead of texts to express emotions in online communications, the semantic analysis that solely relies on textual contents is not sufficient. To date, most related analyses focus on textual data, whereas multimedia (e.g. images, emoticons, videos) has barely explored.

The problem of negation usually appears in non-English texts. For example, Spanish has many nuanced expressions of negation. The use of negation does not mean negative sentiment in many cases. The usage of negation also can express uncertainty or surprised tones. Lima, Perez, Cuadros, and Rigau (2020) introduced a corpus of health field that collects patterns of negation in Spanish corpus texts, including sarcasm, jokes, and other creative users of words that frequently occurred in Twitter texts. In summary, the effect of negation on sentiment is hard to be estimated mathematically. Therefore, most researches can only obtain confidential results under some topics (Joshi, Bhattacharyya, & Carman, 2017) and relatively small scales of dataset(Maynard & Greenwood, 2014).

The rest expressions, repetitions, abbreviations, terms, and slangs, are more investigated and comparatively easier to be solved by algorithms because their

sentiment orientations are clearer. Many algorithms have included special rules or corpus to deal with the(Park et al., 2018; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011; L. Wu, Morstatter, & Liu, 2018). For example, L. Wu et al. (2018) built a corpus of slang that had labeled the sentiment orientation, and cooperated with Sentistrength(Thelwall et al., 2010) to improve the performance of sentiment classification.

Table 21 Typical intricate elements of semantic analysis

Element of expressions	Example	Related studies
Emoticons		Bosch and Revilla (2020); Narayanaperumal (2020); M. A. Ullah et al. (2020)
Negation	I do not say that I do not like it.	Ljajić and Marovac (2019);Lima et al. (2020)
Creative uses of languages	sarcasm, paradox, simile, jokes, irony	Maynard and Greenwood (2014);Joshi et al. (2017); B. Li, Kuang, Zhang, Chen, and Tang (2012)
Repetition	"Soooo", "toooo "	Taboada et al. (2011);Park et al. (2018)
Abbreviation	AFK, AKA	Yadav, Ekbal, Saha, and Bhattacharyya (2018)
Different contexts	Special terms and slangs	Manuel, Indukuri, and Krishna (2010);L. Wu et al. (2018)

Source: own elaboration

Secondly, the majority of studies focus on English, which causes a kind of language bias. The modern metropolis usually consists of people from various regions and countries. A single language may not enough to represent the whole population. Furthermore, the classification of languages is also an intricate problem, when texts are mixed languages, especially when they have the same proportion of different languages. For example, in our case study from Barcelona, many tweets that contain few words or hashtags, are identified as Catalan tweets simply because the geotags are local names of Catalonia.

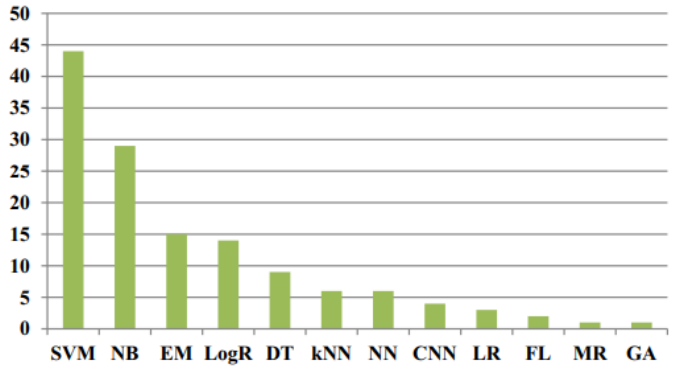
6.3. Incomparability of data and methods

Among methods of semantic analysis, topic mining and hashtag analysis are less controversial than sentiment analysis. They are based on individual words and not

required to understanding the meaning behind the words. According to the study of Morstatter et al. (2013), topic analysis has the least difference between Twitter free API and the Firehose API. In other words, a Twitter sample dataset is representative in conducting the topic analysis. Urban researches that utilize word-based semantic analysis, can reach a high precision, for example, prediction of flus using Twitter(Culotta, 2014; M. Paul & Dredze, 2011; Sinnenberg et al., 2017) and disaster management(Horita et al., 2013; Neppalli et al., 2017).

Conversely, the sentiment analysis involves many features that affect the identification of the sentiment, such as the uppercase, punctuation, proprietary words, conjugation, and among others. It is difficult to evaluate which method is better as semantic data are complex and heterogeneous, though all approaches claim they are better than the others.

Regarding the computing models, the majority of work of the sentiment analysis used models based on the support vector machine (SVM) and Naive Bayes (NB)classifier (**Figure 49).** Other advanced techniques with a higher capacity of manipulating multi-variables, such as Deep learning, fuzzy logic model, and Evolutionary computing, are few applied in sentiment analysis. As an emerging discipline, sentiment analysis still has a large-enough to improve. Besides, with the increasing of commercial potentials in sentiment analysis, little software with excellent performance is free to the public. As Kumar and Jaiswal (2020) pointed out, "the tools and software are useable and affordable only by organizations (both private and government, but currently unavailable to generic users for assisting intelligent and personalized data analysis. "



Note: SVM: support vector machine. NB: Naive Bayes. EM: Ensemble Methods. LogR: Logistic Regression; DT: Decision Tree. kNN: k Nearest Neighbor. NN: Neural Networks . CNN: convolutional neural network. LR: Linear Regression. FL: Fuzzy logic.MR: Multiple Regression GA: Genetic Algorithm.
Source: Kumar and Jaiswal (2020)

Figure 49 Quantitative methods of sentiment analysis from 2010 to 2020

As many problems mentioned above, the representativeness of sentiment analysis should be based on a concrete domain. Small and high-specialized datasets usually could obtain a reliable result because the uncertainty and noise of data was largely removed. For example, Hubert et al. (2018) explored the interaction between the government and citizens in five Latin American countries using Twitter data. They collected tweets from the official Twitter accounts of ministries or departments, and thus the noise of data greatly reduced. Neppalli et al. (2017) studied the Twitter sentiments toward Hurricane Sandy within a two-week period. They retrieved data by searching related keywords, which was helpful to enhance the correlativity of the data. Moreover, domains that have a direct correlation with emotions are easier to cooperate with the sentiment analysis, such as social events(Hürlimann et al., 2016; Kovacs-Györi et al., 2018) and topics(Meier, Mutz, Glathe, Jetzke, & Hölzen, 2019).

In summary, semantic analysis is a challenging topic for researchers who are not linguistic experts. In order to obtain a stable result, the research design should be very clear and well-defined to reduce the uncertainty factors. Meanwhile, for enhancing the accuracy of sentiment classification, it has to cost tremendous efforts to clean the dataset and introduce hybrid algorithms of sentiment classification.

Table 22 Summary of representative studies of semantic analysis issues

Publication	Journal/ conference	Summary of the study	Location	LBSN data
Lima et al. (2020)	arXiv preprint	Studying the problem of negation in Spanish clinical Texts	ES	Hospital data
Sampietro (2016)	Doctoral thesis	Studying usages and diffusion patterns of emoticons y emojis	ES	Twitter
Oriol (2020)	Quality and Quantity	A survey of using emojis in mobile web surveys for millennials	ES,MX	Twitter
Joshi et al. (2017)	ACM Computing Surveys	A survey of automatic sarcasm detection	Not mentioned	Twitter
Manuel et al. (2010)	2010 second Vaagdevi international conference on information Technology for Real World Problems	Analyzing internet slang for sentiment mining	Not mentioned	Twitter
Kovacs-Györi et al. (2018)	ISPRS International Journal of Geo-Information	Classifying park visitors in London based on spatiotemporal and sentiment analysis of Twitter data	UK	Twitter
Maynard and Greenwood (2014)	LREC 2014 Proceedings	Investigating the impact of sarcasm on sentiment analysis	UK	Twitter
Narayanaperumal (2020)	Doctoral thesis	Deep Neural Networks for Sentiment Analysis in Tweets with Emoticons	US	Twitter
Neppalli et al. (2017)	International journal of disaster risk reduction	Sentiment analysis during Hurricane Sandy in emergency response	US	Twitter
Taboada et al. (2011)	Computational linguistics	Lexicon-based methods for sentiment analysis	US	Twitter
Kumar and Jaiswal (2020)	Concurrency and Computation: Practice and Experience	Systematic literature review of sentiment analysis on Twitter using soft computing techniques	Not metioned	Literature Review

Source: own elaboration

V.7. Summary

This chapter analyses the limitations of LBSN data in terms of five domains, which reviews the major problems and bias of LBSN data in the urban spatiotemporal analysis. The restraint of LBSN data stems from two aspects. Firstly, not everyone uses LBSN applications and post everything online, and thus LBSN data only represents part of society. Secondly, technically, it is possible to connect each user's real identity with his /her online activities. However, privacy and legal issue control the boundary of data that we can access. Hence, it is difficult to improve the confidence level of LBSN data, just like census or traditional survey. To specific domains, the representative issues of LBNS data can be summarized as the following:

- The demographic bias and uneven distribution of LBSN data can be verified in different regions. It implies that these phenomena do not appear by chance. LBSN data is more likely linked with people who are younger and have better education and income.

- The online behaviors are affected by socioeconomic conditions. LBSN data performs higher representativeness in developed urban areas.

- LBSN data can yield a reliable result of describing human mobility at an aggregated level in urban areas.

- The semantic analysis leveraging LBSN data could reach a higher accuracy in urban health domain and disaster management.

In summary, the "chaotic" LBSN data may exactly reflect the real world that we just have not observed before. Therefore, as Morstatter et al. (2013) suggested, wisely choose the data and study scope is essential for the representativeness of

LBSN data. Meanwhile, it is a trending that an investigation utilizes multiple datasets to get a reliable result. LBSN data becomes one of the regular data sources in urban studies.

Chapter VI.

Case study

VI.1.Case study I

Analysis of the Spatial Structure of Beijing from the point view of Weibo Data

1.1. Introduction

The urban structure has long been one of the essential topics in urban geography and planning (Carlos Marmolejo-Duarte, Echavarría Ochoa, & Biere Arenas, 2016). Traditional methods mostly utilized official demographic census(Carlos Marmolejo-Duarte, Núñez, & Roca Cladera, 2013)or land-use data to analyse the urban structure. However, these indicators are insufficient to capture the dynamics and complexity of urban systems. They only considered employment and residential activities(De Ureña, Pillet, & Marmolejo-Duarte, 2013). However, all the other activities that people carry out outside the workplace or home are frequently ignored (Carlos Marmolejo-Duarte & Cerda-Troncoso, 2012).

Nowadays, thanks to the development of geo-technologies, a novel channel that shows the changes and operational mechanisms of urban society is opened. The data generated by the Internet or GPS-tracking can capture the detailed individual activities within cities. Since 2009, the explosion of social media around the world, such as Twitter and Facebook, prompted academic studies based on social media. Despite the ability of social media could break through the limitation of physical distance to some degree, it retains a strong relationship with physical, cultural, and linguistic boundaries(Stephens & Poorthuis, 2015). Therefore, social

media data is a new data source to study the urban structure. For example, Green (2007) discussed the urban structure in terms of the functional perspective, which is based on social network analysis.

Therefore, this research tries to detect the temporal-spatial structure of the Beijing metropolitan area via Weibo data. Weibo is one of the most popular social media platforms in China where western social media such as Twitter is not used due to censorship. Weibos' function is similar to Twitter which allows users to exchange news or information. According to its earnings report in 2015, the daily active users of Weibo reached 106 million persons in China mainland. In addition, because of the issue of transparency, it is not easy to get the related data to follow the traditional methods for identifying the sub-centers of Beijing, such as data of localized employment and detailed land-uses. Therefore, using the data of daily activities is a way to overcome such data limitations.

Beijing is a very typical megacity that faces many tough challenges and the reform of urban structure. Beijing was one of the top 10 biggest world cities in 2015[30]. The urban scale has been expanded 12 times in 55 years. The total area of modern Beijing has reached to 16,410.54 km^2, which consists of 16 administrative districts. Such rapid urban expansion brought heavy pollution and terrible traffic congestion to the city. Although the Beijing government stressed the reconstruction of urban structure through building sub-centers, the model of urban expansion still like ripples which diffused from the downtown. Whether such planed sub-centers effectively affect the urban structure of Bringing deserves to be explored from a broader perspective as it is analyzed in this paper. Moreover, the urbanization of Beijing has aroused attention from researchers and officials since many years ago. However, their investigation mainly focused on the urban area within the six districts in the central region, rather than the whole Beijing area. In addition, a few of them introduced social media data to investigate the urban structure of the city.

Therefore, this study prepares to combine urban contexture with Weibo data to detect the sub-centers of Beijing. The studied area extends to "peripheral

[30] Source: City metric, http://www.citymetric.com/skylines/where-are-largest-cities-world-1051

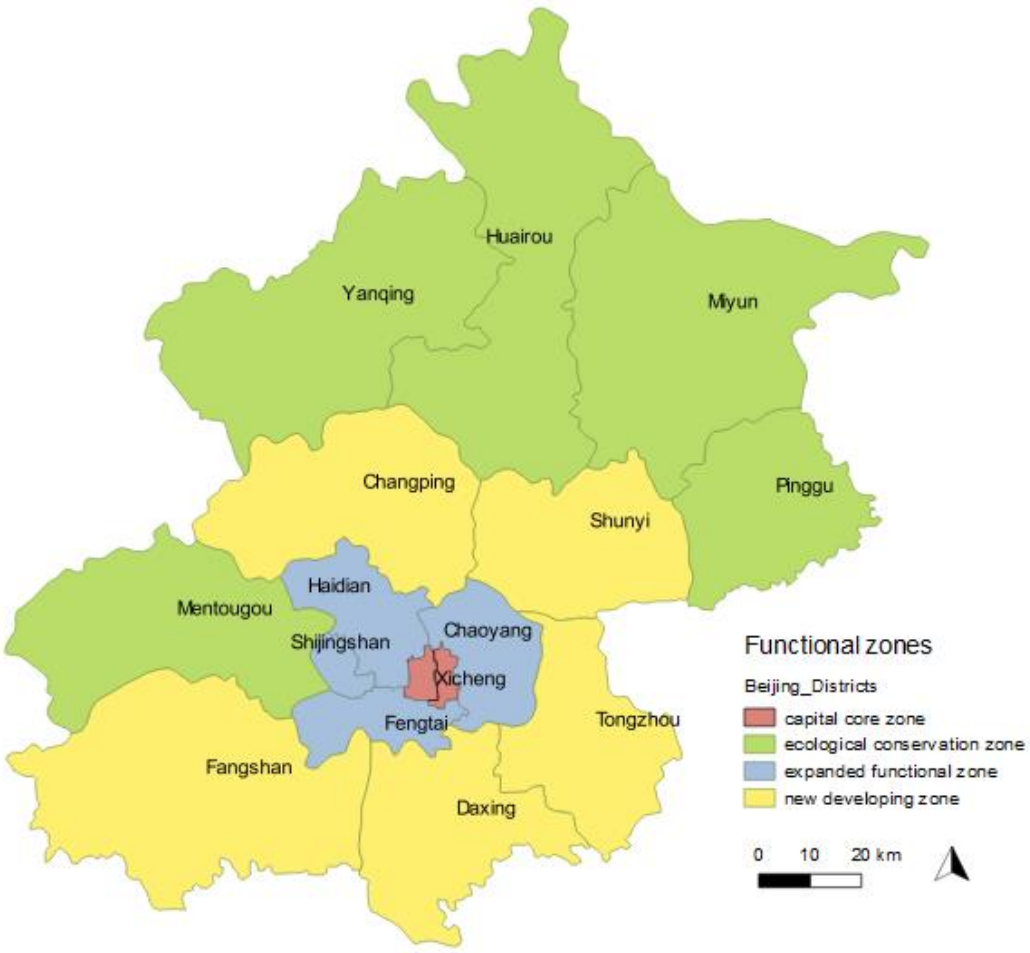
districts" of Beijing, thus we could evaluate the developmental situation of the whole Beijing area. The part of the literature review summarizes pioneer urban researches that cooperate with the time-geography framework and Weibo data. A brief retrospection on the urban structure of Beijing is offered in the third part. The fourth section introduces the exponential model to identify Weibo sub-centers in terms of different periods of a week. The only clear spatial structure of Weibo distribution is confirmed by the model is the period of the weekend. However, the urban structure of Beijing indeed shows a polycentric structure according to the distribution of potential Weibo sub-centers, though the high-density clusters of Weibo activities mainly gathered in the northern area of Beijing. Despite its inherent limitations, Weibo data successfully identify sub-centers that are not considered by the traditional methods of identification of urban structure, such as university and recreational places.

1.2. Literature review

1.2.1 Brief evolution of the urban structure of Beijing

Beijing is the second-largest city in China, also the political, cultural and technological innovation center of China. The current Beijing consists of 16 districts and more than 20 million permanent residents. Meanwhile, according to the Beijing City Master Plan 2004, the functional structure of Beijing city consists of the capital core zone, the expanded functional zone, the new developing zone and the ecological conservation zone (**Figure 50**).

The capital core zone includes Dongcheng and Xicheng district, which is the national political center. The expanded functional zone undertakes the largest proportion of employment and urban services. The new developing zone is the region for future urban development, as well as the agricultural basement. The ecological conservation zone accounts for 53.3% of the total area of Beijing, which aims to preserve the water resources and the ecological environment.



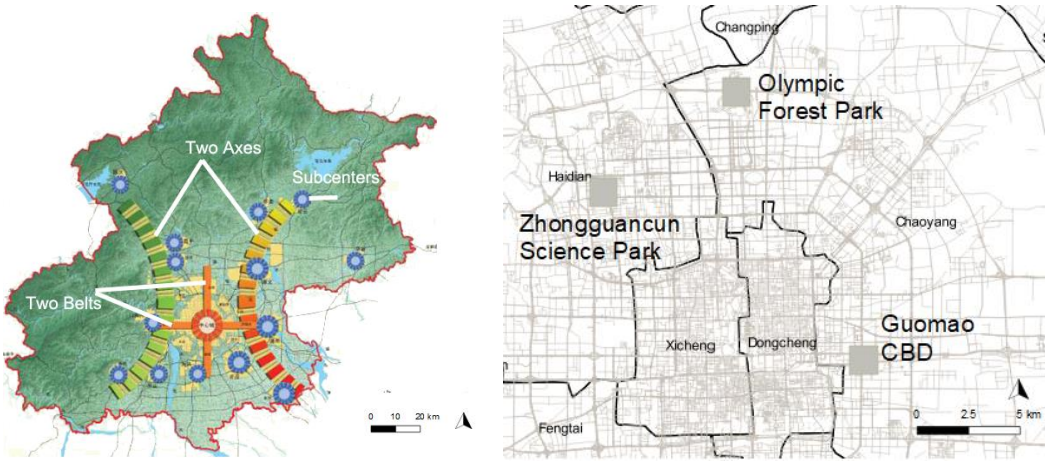
Source: self-elaboration

Figure 50 The functional zones of Beijing according to the 2004 City Master Plan

The modern urban structure of Beijing basically inherits the structure of the Qing Dynasty – centered on the Forbidden City, organized on the north-south axis. In 1949, at the beginning of the People's Republic of China, Beijing was just a city where had 1.65 million people. In 1953, the Beijing government brought up *Key Points of Draft Plan for Reconstructing and Expanding Beijing*. Based on the plan, urban expansion followed a monocentric model based on four concentric ring roads and several radial ones. However, the heavy industrial pollution in the late 1970s urged the government to change the urban plan. In 1983, the Beijing government published the *Master Plan of Beijing Urban Construction*. Since then, the southern part of Beijing has gradually declined, due to the removal of heavy industrial factories. On the contrary, the northern part of the city, especially

Zhongguancun, has been developed rapidly, because several famous universities were located in this area, such as Tsinghua University and Peking University.

Ten years later, the *Beijing Master Plan (1991—2010)* emphasized the importance of controlling the scale of the central city. It planned to reconstruct the urban structure from monocentric mode to "dispersed clusters". Beijing government prepared to build 10 satellite cities to disperse the population. After the new millennium, the model of dispersed clusters indeed helped to disperse the population in the downtown to some degree, whereas the population was soaring in the surrounding areas. The Beijing government decided to reform the urban structure in 2001. The monocentric model was replaced by a multi-centric development scheme. The layout of *Beijing City Master Plan 2004* was "Two Axes – Two Belts – Multi-centers" **(Figure 51**). "Two axes" refers to the traditional north-south axis and the horizontal line along with Chang'an Street, which aims to protect the traditional spatial structure and its cultural values. "Two-belts" divides Beijing city into an eastern development belt and western environmental belt. "Multi-centers" plans to build several functional centers that could adapt to the requirements of globalization, for example, Zhongguancun Science Park, GuoMao CBD, Olympic center, etc. It planned to build 11 new cities to disperse the population.



(a) (b)

Source: (a)Beijing Master Plan, 2004; (b) own elaboration

Figure 51 The layout of Beijing Master Plan 2004

In 2016, according to the *Beijing Municipal Commission of Urban Planning*, the urban structure of Beijing will be transformed into "one central city, one deputy city, two-axes and multi-sub-centers". The deputy city undertook part of the administrative function of Beijing.

In summary, as the speed of urban expansion of Beijing is so fast the controversy of the urban structure of Beijing remains, this paper using Weibo data to detect the sub-centers can provide a new sight to analyze the urban structure of the city. Moreover, because the data coverage covers New developing zones and ecological zones, it can evaluate the developmental situation of the whole Beijing area.

1.2.2 Previous studies of the Beijing urban structure

Since the end of the 1990s, the urbanization issues of Beijing city have drawn much attention from researchers. Many studies investigated the Beijing metropolitan area from various perspectives and methods. Besides the traditionally documental study(Ai, Zhuang, & Liu, 2008), remote sensing and GIS have become the prevalent techniques to analyse the urban structure and land use of Beijing(Kuang, 2012; Q. Wu et al., 2006; Xie, Fang, Lin, Gong, & Qiao, 2007). The conclusion of these coincides in suggesting that urban sprawl mainly eroded areas of cropland around Beijing. Another important approach is to utilize census data to investigate the social-spatial structure of the city. Feng and Zhou (2003) utilized the second and the fifth Chinese national census in 1982 and in 2000 to compare the changes of the social-spatial structure of Beijing. D. Huang, Liu, and Zhao (2015)used companies registered data to investigate the potential employment sub-centers in Beijing. Interestingly, the former stated that the social-spatial structure of Beijing already shows some characteristics of the polycentric model, but Huang's study showed that the city was still monocentric.

In recent years, "Big Data" has also involved in studies of Beijing city. Beijing City Lab launched a project—SinoGrid to collect various aspects of city life based on 1 km^2 fishnet. Long and Liu (2013) combined remoting images, POIs data, and Weibo check-in data to explore the degree of mixed land-use in Beijing. They concluded that land-use mixing is higher in the central areas of the city than the peripheral ones. Microsoft Research utilized taxi data combined with POIs to discover the functional regions in Beijing (J. Yuan et al., 2012). After comparing with the actual urban contexture, they concluded that the POIs and mobility patterns could be a powerful tool to identify the land uses of a city. Y. Wang et al. (2016) identified seven types of land use clusters in Beijing using Weibo and POIs data. They divided the Beijing area into cells with 0.4 km^2 and cluster these cells using the hourly frequency of Weibo on each cell. Based on the temporal active trend of the cluster, they estimated the type of land use and verified it by mining the contents of these Weibos. In summary, the urban structure of Beijing has been explored by various perspectives and technics. However, these studied primarily focus on the land-uses or urban functions, rather than the urban structure. Therefore, our research aims to investigate the urban structure of Beijing based on Weibo activities for filling the gap.

1.3. Methodology

1.3.1 Research design

We calculated the cumulative density of Weibo-messages in Beijing to identify Weibo sub-centers. Meanwhile, a comparative study will help us to identify the function of the sub-center. Because it is hard to obtain the latest official document of cadastral information and land use, we made a direct comparison with Google map and Baidu map which is the most popular electronic map in mainland China. The current version map can help us to observe the latest status of land-uses. The third step was to analyze the frequency of Weibo messages in terms of different time periods. Based on the results, this paper built a regression model to distinguish the Weibo sub-centers during the week, workdays and weekends, then make an urban contexture analysis of Beijing city. It tried to discuss in which

degree Weibo data could reflect the urban tissue, and the type of land uses of these sub-centers.

1.3.2 Data collection

Weibo is one of the most popular social media in China, whose functions are similar to Twitter. A user can post words, pictures, mention or talk to other people via this platform. Every message sent by a user is called a "Weibo". Users can create a "nickname" for their Weibo accounts, but all users should register with true identities.

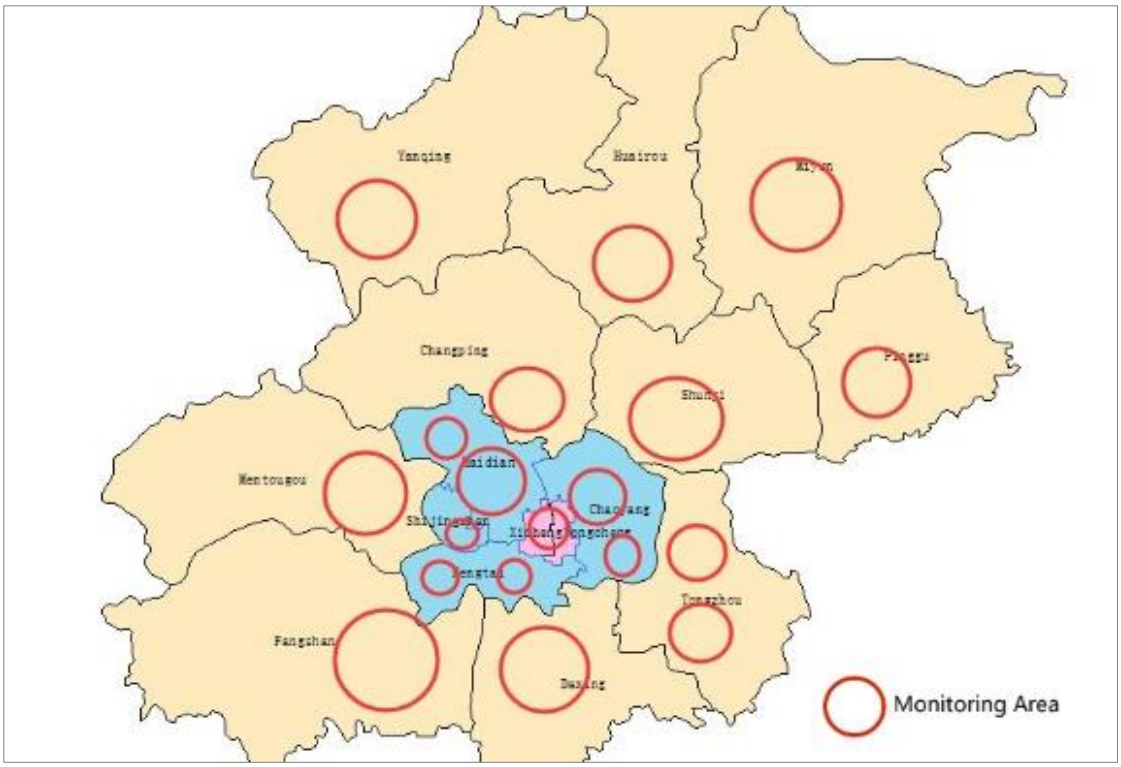
The total users of Weibo are over 500 million in China. According to *Weibo Development Report 2014*, the main users were among 19-35 years old, accounted for 72% for total users. The monthly active users have reached 167 million, the whole Beijing area took up 8.7% of them. Considering the total population was 21.52 million in 2014, it suggests that half of the people in Beijing use Weibo time to time. Therefore, it is reasonable to utilize Weibo data to explore the urban space of the Beijing area.

Weibo API also provides free resources for researchers, though there are many restrictions. For instance, the maximum number of inquiring data is 1,000 times per day, and one inquiry only can get 2,000 messages. Despite the restrictions, Weibo API offers various open ports of data to researchers: Weibo of users, user's basic information, network relationship of accounts, geographic information, etc. Therefore, we could get a lot of useful information about citizens through Weibo API, especially the port of geo-information allows researchers to explore the dynamic activities in a city.

Table 23 Parameters of monitoring circles

District	Latitude	Longitude	Monitor-Radius(km)	Monitoring Areas(km2)
Districts of Dongcheng and Xicheng	39.915974	116.385956	4.31	58.359
Chaoyang1	39.910231	116.526678	7.63	182.894
Chaoyang2	40.024454	116.473704	7.55	179.079
Haidian1	39.996334	116.27558	7.56	179.553
Haidian2	40.089426	116.151917	3.6	40.715
Shijingshan	39.919546	116.205842	4	50.265
Tongzhou1	39.914332	116.704191	6	113.097
tongzhou2	39.76988	116.73304	10	314.159
Changping	40.15804	116.355207	10.5	346.361
Shunyi (airport included)	40.114097	116.662694	11	380.133
Daxing	39.696295	116.402051	11	380.133
Fengtai1	39.835938	116.373859	3	28.274
Fengtai2	39.852911	116.282106	4	50.265
Miyun	40.403709	116.866321	10	314.159
Fangshan	39.676711	116.090903	11	380.133
Mentougou	39.964773	115.977905	11	380.133
Pinggu	40.175009	117.121087	11	380.133
Huairou	40.332106	116.633393	7.5	176.715
Yanqing	40.459453	115.97512	11	380.133
Total monitoring areas				4314.692

Source: own elaboration



Source: own elaboration

Figure 52 Schematic diagram of the monitoring range

In this study, we utilized the access port of "nearby Weibo" to collect data, which supported to record data around a "monitoring point" (the maximum radius is 11 km). The port can track those Weibo messages that users agreed to share their immediate geo-location. The message flow included the original posts and forwarded messages in the monitoring area by chronological order. Because of the limitation of accessing the port per day, it is impossible to cover all areas. We set 19 central points to collect the data (see **Table 23** and **Figure 52**), considering the technical limitations of the application and the current situation of build-up areas. After the pre-test of data collection in March and April of 2016, we decided to deliver more tracking frequency and scale to those areas where have had a higher volume of data. Each monitoring circle was restricted in one district and was avoid overlap. We tried to cover the central areas as much as possible. For the outer-districts, such as Huairou, Pinggu, Miyun and the others, the monitoring center-point was set in their central town. The measure of monitor-

radius is accomplished by Google Maps' tool which can measure the distance between two points on the map.

The temporal range of data is from 00:00 on the 11th of April to 24:00 on the 17th of April. This time-period avoided vacations and festivals of China, thus it could represent the normal situation of Beijing city. After eliminating null data, the total Weibos of the study were 53,967. Three districts appeared data loss in some periods. The amount of Weibo in Haidian district has fallen suddenly on Wednesday due to the data lost from 17:07 of Wednesday to 3:00 on Thursday. Shunyi lost data from 17:52 on Saturday to 16:52 on Sunday. Fengtai lost 16 hours' data from Wednesday to Thursday. However, this loss of data would not affect the analysis greatly, because Shunyi and Fenngtai district generated fewer Weibos in the pre-test, which were below than 15 Weibos per hour. Although Haidian district had a higher frequency, it only lost 10 hours' data. In order to compensate for the loss, the hourly frequency of Weibo of each of the three districts was filled in the lost hours separately.

1.3.3 Quadrat analysis

In order to identify the Weibo sub-centers, it is necessary to divide the city into equal zones to calculate the density of Weibos. Considering that the urban shape of Beijing city is squared and composed of orthogonal roads (**Figure 53**), it is reasonable to adopt quadrats to calculate the Weibo density.

(a) the urban shape of Beijing within the 5th belt (b) A piece of the urban tissue of Beijing

Source: own elaboration

Figure 53 Detail of Beijing urban tissue

The more points gathered in a quadrat; the higher the density of the quadrat. Therefore, the size of the grid plays a pivotal role in the analysis. McDonald and Prather (1994)used 1.29 km^2 of a grid as their research unit in Chicago. Giuliano and Small (1991)used 'transportation analysis zones' which is defined by the California Association of Government (SCAG). The average area of analysis zone is about 8 km^2. McMillen and Lester (2003)used a quarter square mile (about 0.65 km^2) as the basic unit of observation in Chicago, because the small unit can find out those low-density regions in the midst of high-density areas. Our unit of the grid is determined by a statistical method. According to the study of Griffith, Amrhein, and Desloges (1991), the formula of optimized quadrat area is:

$$\text{Quadrat area} = \frac{2\text{A}}{\text{q}} \tag{1}$$

A represents the surface of the studied area, q is the number of points. The studied area is calculated by the actual monitoring area, not the whole urban area. Therefore, the area of studied range is calculated by the formula of the circle:

$$\text{Quadrat area} = \frac{\sum_{i=1}^{19} 2\pi r_i^2}{q} \tag{2}$$

where *r* is the radius of the monitoring radius, *i* is the serial number of monitoring circle.

The total area of monitoring circles equals 4,314.70 km^2, and the number of points (*i.e.* Weibo messages) is 53,967, thus the statistical optimum quadrat area is 0.16 km^2 (i.e. the length of the side is 0.4km). Based on this size, we also expand the length of side triple and sextuple respectively for comparative reasons, because the size of the units in previous researches is larger than the one we calculated (**Table 24**). The medium one is similar to previous studies. The largest one significantly reduces the percentage of minimum value, however, it cannot match the boundary of the monitoring circle very well. It enlarges the total monitoring area. Therefore, the largest size is excluded.

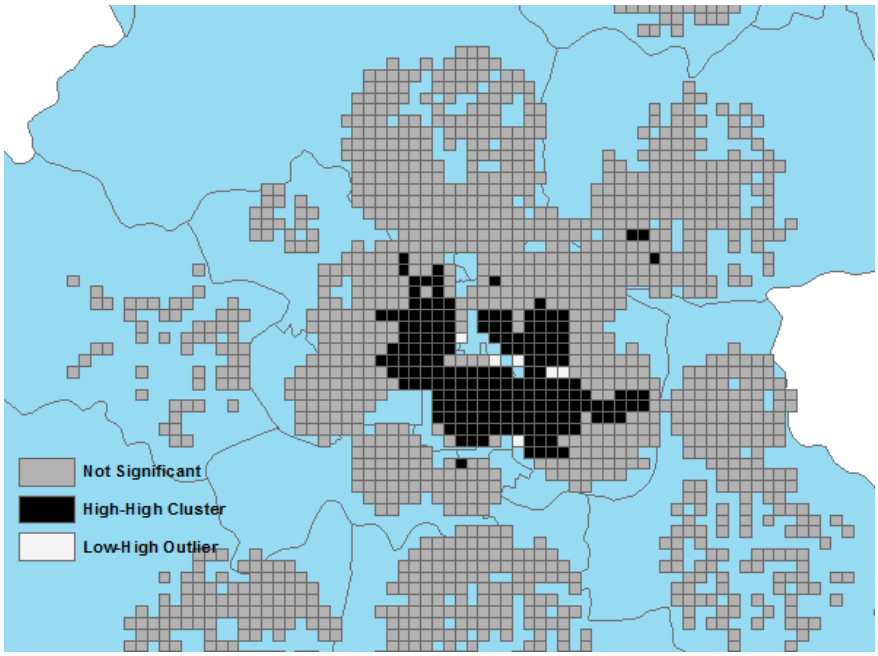
Table 24 Three different sizes of quadrat

Quadrat side (km)	Number of Quadrats	Minimum number of Weibo in one quadrat Percentage		Maximum number	Mean	Standard Deviation
0.4	6952	1	33.5%	442	7.76	16.75
1.2	1814	1	23%	887	29.53	69.22
2.4	705	1	16%	1831	75.98	196

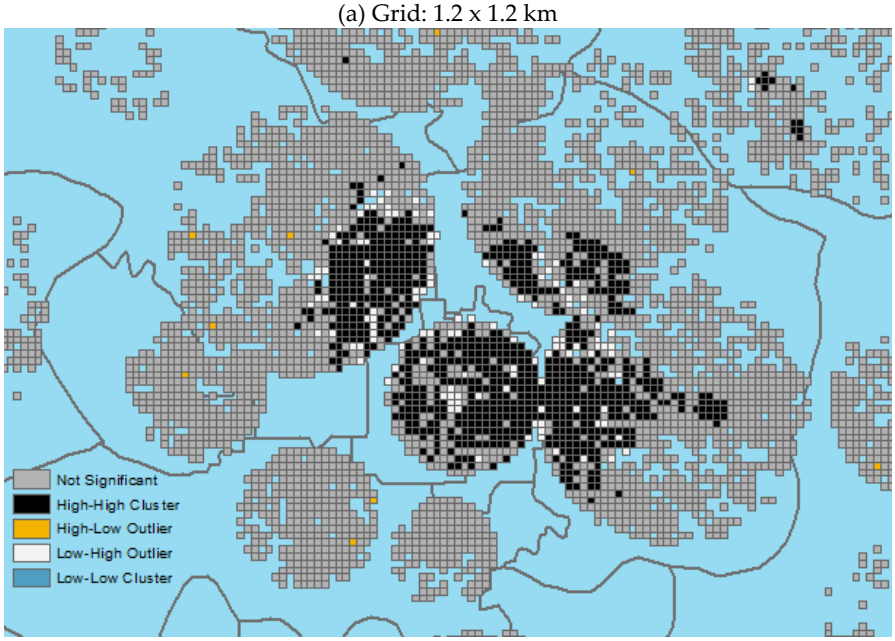
Note: the quadrats that do not contain Weibo messages were excluded from the sum of quadrats.

Source: own elaboration

Next, Anselin Local Moran's I (Arcgis 10.2) is introduced to test the clustering degree of points based on medium and small grids respectively. If there is no significant clustering area, or the distribution is dispersed, it indicates that Weibo data is probably useless in detecting the urban structure. Both grids (**Figure 54**) show strong trends of concentration, however, the medium grid (1.2 x 1.2 km) (**Figure 54**-a) fails to distinguish those areas of lower density in the central city. Therefore, we use the small size of the grid as the unit to calculate the density.



(a) Grid: 1.2 x 1.2 km



(b) Grid: 0.4 x 0.4 km

Source: own elaboration

Figure 54 Anselin Local Moran's I Result

1.4. Identification of Weibo sub-centers

This paper tries to identify Weibo sub-centers in four periods – workdays (Monday to Friday, from 8:00 to 20:00), nights (Monday to Friday, from 23:00 to 6:00), the weekend and the whole week. A potential sub-centers should satisfy two criteria: 1) it should be located in the High-High cluster which is based on the Anselin Local Moran's I Result; 2) the density of the potential area should be above the mean Weibo density of the HH cluster plus one Standard Deviation. The first condition selects out those high active grids. The second condition excludes those grids of lower density in the HH clusters. After the selecting out those cells that satisfy the criteria, we aggregate these cells into potential sub-centers through nearest-neighbor principle. The standard of aggregation is that the distance between two qualified grids is equal or less than 0.4 km.

As we detect sub-centers using the density approach, there are three major methods(Roca Cladera et al., 2009): reference thresholds, parametric methods, non-parametric methods. The reference threshold is to set a statistical or numerical threshold to select out sub-centers, such as a threshold of the density of employment or density of workplaces(Garcia-López & Muñiz, 2010; Giuliano & Small, 1991). This method can be used to compare the results of a single city in different periods. However, when the stricter statistically rigorous is demanded, the parametric or non-parametric modelling has more advantages. Since the monitoring range of our dataset does not cover the whole of Beijing, it is hard to estimate a continuous surface of density to introduce the non-parametric model. Therefore, we adopted the classic negative exponential model (McDonald & Prather, 1994)to identify Weibo sub-centers. Firstly, we assume that the distribution of density of Weibo also follows the similar pattern of the employment distribution - it decreases as the distance to the city center or CBD increases. Secondly, our model only restricts to the potential Weibo sub-centers:

$$D(x_i) = D_0 e^{-\alpha x_i + b} \tag{3}$$

$D(x_i)$ corresponds to the Weibo density of a given i potential sub-center, and x_i is defined as the Euclidian distance from the mean center of the i potential sub-center to the sub-center which has the highest Weibo density (which in turn is assumed to be the Beijing' CBD). α is the slope of the density that is reduced departing from the Weibo CBD. Since the negative exponential model is nonlinear, we transform it into the linear form for calculation purposes:

$$\ln D(x_i) = -\alpha Xi + C \quad (4)$$

where $C = \ln D_0 + b$.

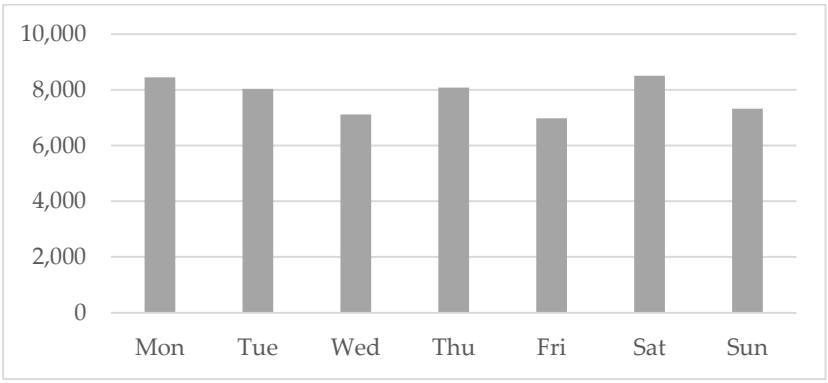
Finally, based on unstandardized residual values from the model, two thresholds are introduced to the test of identification of the sub-centers: 1) the residual value of a confirmed sub-center should be positive and is equal or above the mean value of all residual values plus one standard deviation; 2) the residual value of a confirmed sub-center is positive and above the mean value of all residual values. The final threshold is decided by the analysis of the social-economic profile and the actual urban land use.

1.5. Results for the identification of Weibo sub-centers

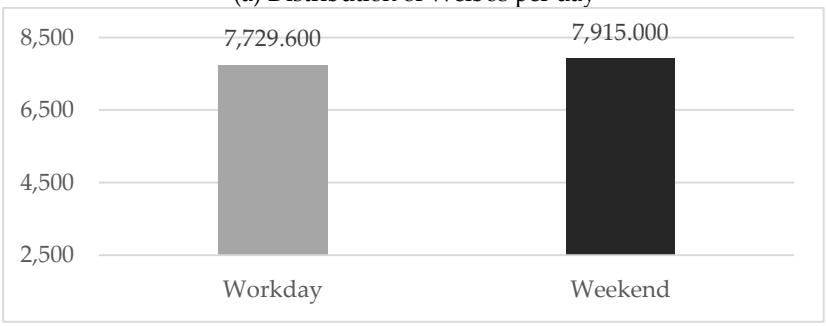
1.5.1 Temporal variation of Weibo messages

Figure 55-(a) shows the variation of Weibo activities along one week. The Saturday and Monday has the highest number of Weibos. The least active day is the Friday. The Weibo activity is slightly higher at the weekend than during workdays (**Figure 55**-(b)). According to *the 2015 Weibo Search Engine White-Book*, the first and second rank of searching frequency are news and celebrities. It indicates that people mainly use Weibo as a source of information and as an entertainment application. In terms of the hourly Weibo activities (**Figure 55** -(c)), the most active period is from 21:00 to 0:00 on workdays and the weekend. Therefore, it is reasonable to deduce that the majority of users enjoy their leisure time after 21:00 pm. The

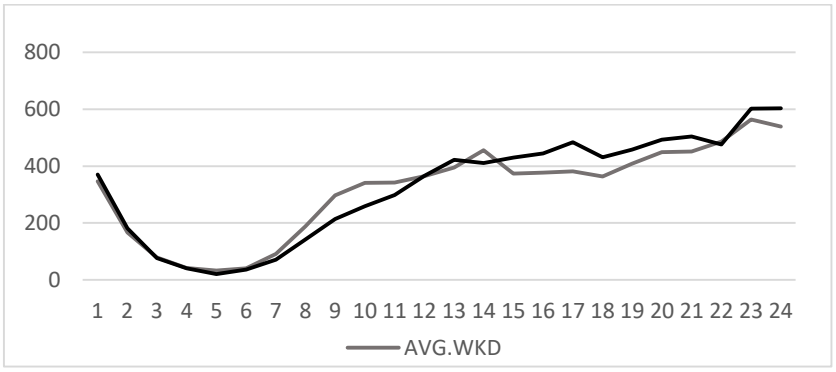
average Weibo activities on workdays are higher than the weekend before 12:00 am. After 14:00, the average frequency of weekend is higher than the workdays. It reflects the habitual differences between workdays and weekend.



(a) Distribution of Weibos per day



(b) Daily average frequency of Weibos during workdays and the weekend



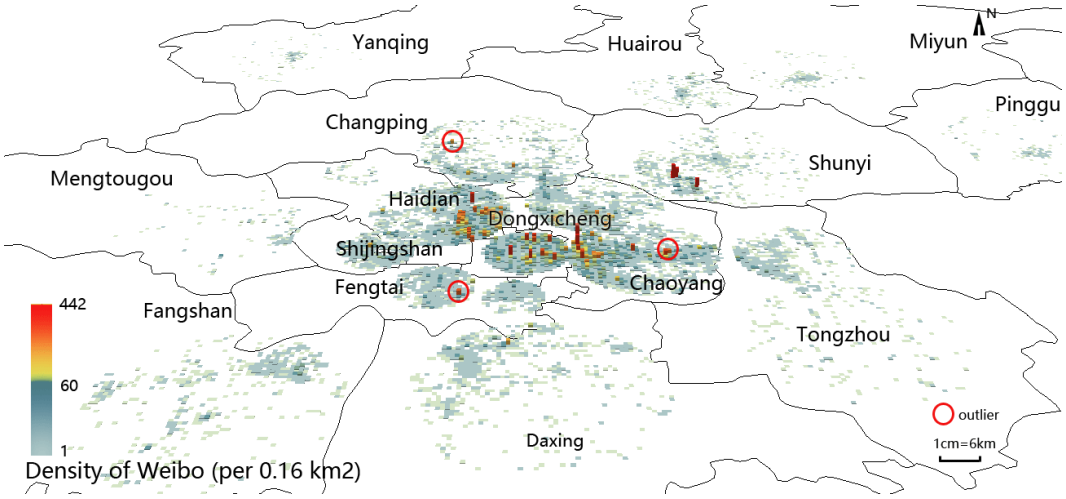
(c) Average frequency of Weibo messages per hour

Source: own elaboration

Figure 55 Temporal distribution of Weibo activities

1.5.2 Spatial distribution of Weibo density

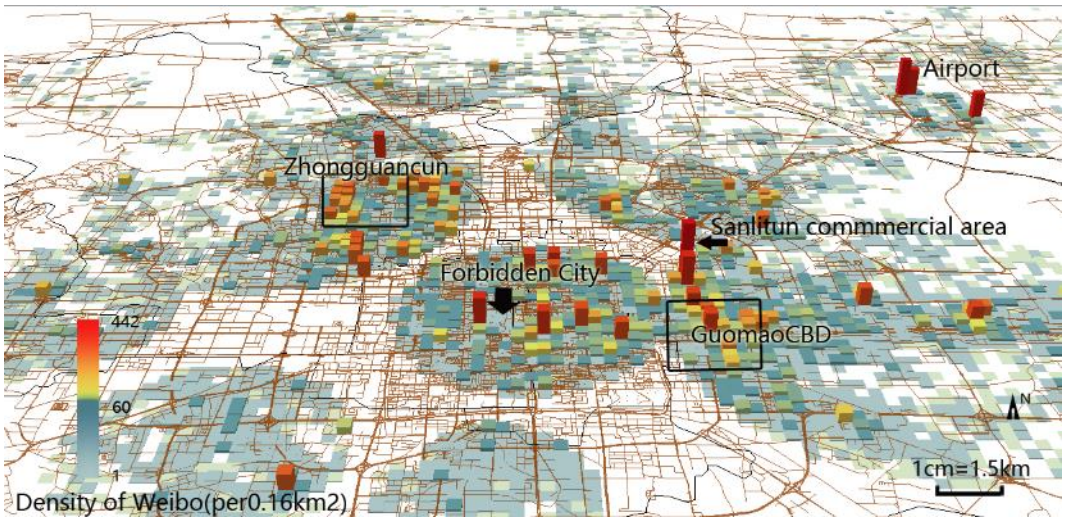
Figure 56 shows that grids with high-density mainly concentrated in the central zone and in the expanded functional zone. There is a decreasing tendency of Weibo density from the central area to the periphery. The density of the developing zone and ecological conservation zone is lower than the central city in general. It is coincident with the lower density of population and urban constructions in these areas.



Source: own elaboration

Figure 56 3D Weibo density panorama

Besides the airport in the Shunyi district, there are three places of higher density which surrounded by low-density cells (**Figure 57**) Compared to the actual urban contexture, they are located in universities, such as Capital University of Economics and Business in Fengtai district, the Communication University of China and Beijing International Studies University in Chaoyang district, and the zone of a university in Changping district.



Source: own elaboration

Figure 57 3D view of Weibo density in the central area of Beijing

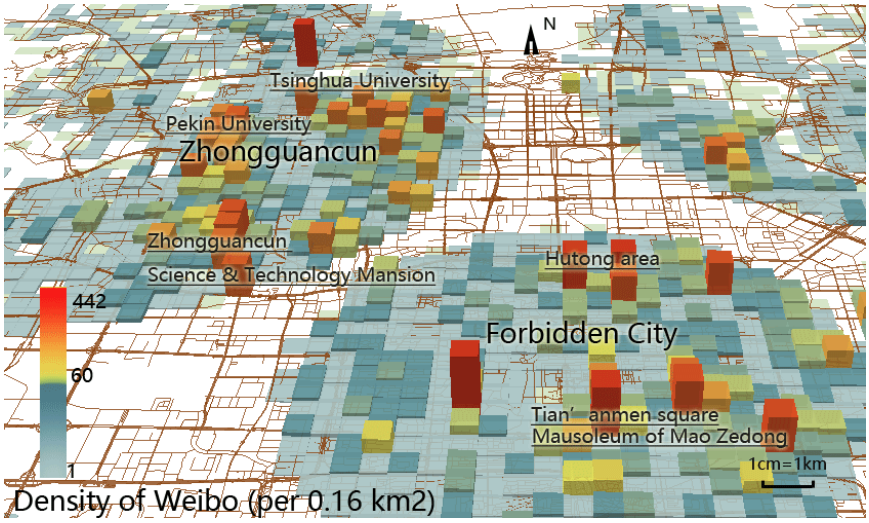
If we check the density map closely (**Figure 57**), the distribution of density in the central area is actually variable from place to place, rather than the same intensity. From the Guomao CBD to Sanlitun, there is a high-density area along with the third ring road, because many office towers are constructed in the area. **Figure 58** contains a partial view of Guomao CBD along with the third ring road. The CCTV headquarter, China Merchants Tower and many other famous towers all gather in these streets.

Source: Google earth

Figure 58 3D view of urban contexture in the Guomao CBD area

Within the second ring road, the density around Forbidden City actually was not very high, because the central government and Beijing government are located in the area (**Figure 59**). Tian'anmen square and Mausoleum of Mao Zedong are in front of the Forbidden City. Thousands of tourists visit these two places every day, thus the density rises[31]. Similarly, the higher density in the northern part of the traditional center was formed by tourism, because many Hutong (a kind of typical residential construction and culture of Beijing) are located here.

[31] As the roaming service of mobile is very cheap even free (depends on the type of contracts) in China, the willingness that tourist use Weibo is not affected by mobile roaming.



Source: own elaboration

Figure 59 3D view of Weibo density in traditional city center and Haidian

In the Haidian district, universities and office areas are the main types of land use. In the past two decades, Zhongguancun has gathered nearly 20,000 high-technology enterprises, including Lenovo and Baidu. The higher density grids distribute along with Zhongguancun Streets and Chengfu Road in general. It matches the distribution of technological companies in this area. According to the introduction on Zhongguancun Science Park's official website, the distribution of these companies approximately starts from Zhongguancun Street, then goes up to Chengfu Road, forms an "F"-shape area.

1.5.3 Identification of Weibo sub-centers in Beijing

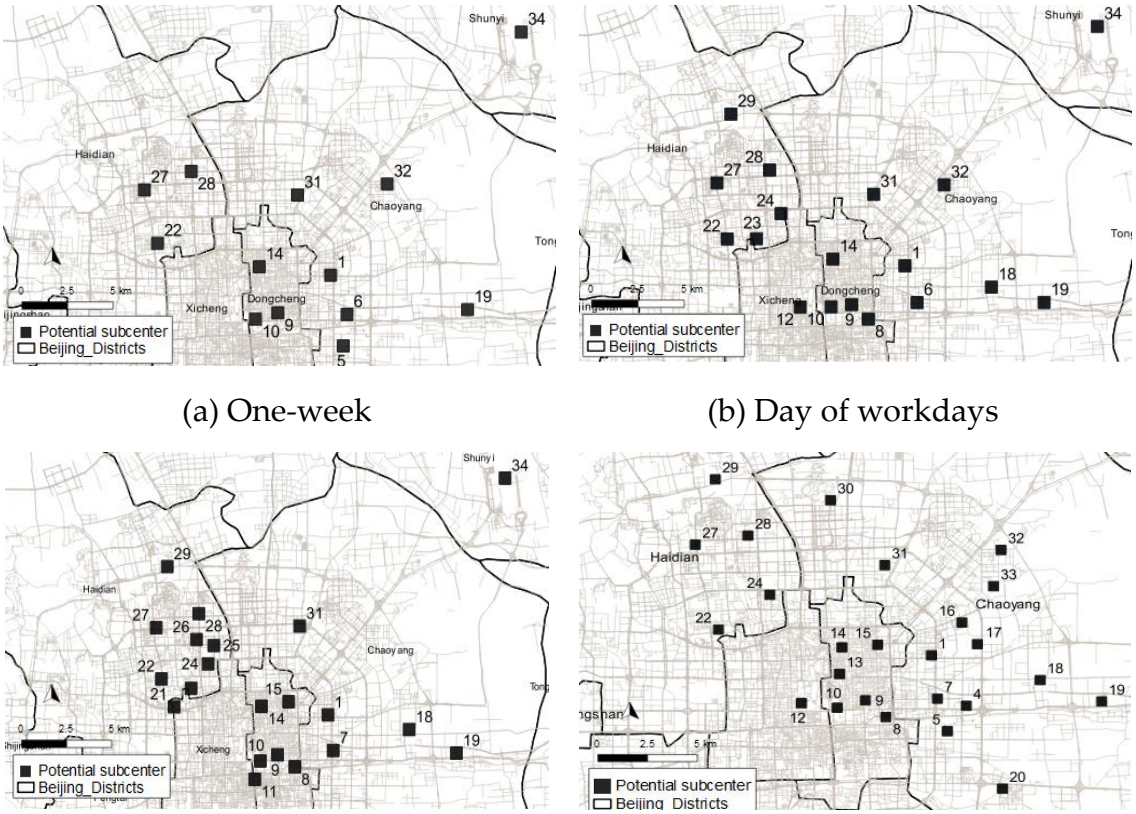
Figure 60 shows that all potential Weibo sub-centers are located in Chaoyang, Haidian, Dongcheng, Xicheng and Shunyi Districts. The potential Weibo sub-centers of one week are less than other periods because the average density is the highest (**Table 25**) so that only 72 grids satisfy with the criteria abovementioned, which only account for 2% of total grids. The potential sub-centers of days (**Figure 60**-(b)) and nights (**Figure 60**-(c)) of workdays are quite close to each other. However, the density of Weibo activities is weaker in the night.

Table 25 Statistical description of potential Weibo sub-centers and distance

	Ln_Density					Distance to Weibo-CBD(km)				
	Mean	Std.	Range	Min	Max	Mean	Std.	Range	Min	Max
One week	6.569	0.264	0.921	6.271	7.193	6.626	3.952	13.321	0.000	13.322
Weekend	5.682	0.376	1.647	5.267	6.913	6.394	3.774	14.882	0.000	14.882
Day_workday	5.887	0.318	1.014	5.546	6.560	7.245	3.964	14.882	0.000	14.882
Night_workday	5.002	0.212	0.722	4.666	5.388	7.602	3.934	14.834	0.000	14.834

Source: own elaboration

In terms of land uses (**Table 26**), it changes with different periods. Mixed area, university, transportation hub, and tourist attraction show more active in workdays, while sub-centers of university decrease in the weekend. None of the potential sub-centers belong to the residential area in the time span of one week (**Figure 60**- (a)) and day-period of workdays (**Figure 60**- (b)). Commercial and office areas (1, 5, 9, 22, 27) are the major types of land use for Weibo sub-centers considering the complete week period. In the day–period of workdays, two important railway stations—Beijing Northern Station (23) and Beijing Eastern Station (8) appear as potential Weibo sub-centers. It is possibly caused by their function as transportation hubs that connect with Beijing suburban trains. Especially, many Interprovincial trains pass through Beijing Eastern Station, thus it is also an active area during the night. Compared with the day-period, besides the two universities in the eastern part (18,19), there are more potential Weibo sub-centers in the night of workday concentrated in the northwest part where many universities are located. It implies that the main active users during the night-period are students of universities.



(a) One-week (b) Day of workdays

(c) Night of workdays (d) The weekend

Source: own elaboration

Figure 60 Distribution of potential Weibo sub-centers in Beijing

During the weekend (**Figure 60**-(d)), tourist attractions (10,13,14) and recreational areas (16,17,20,30,32) account for 40% of total potential sub-centers; for example: 10 is Tian'anmen Square, 13 and 14 are Hutong areas, and 16 is the 798 Art Zone. Three residential places (5, 16, 31) are shown as potential sub-centers. Some recreational places only appear active at the weekend. For example, 20 is an amusement park, and 30 is the Olympic Forest Park. Both of them are far away from the city center, thus such Weibo activity is probably generated from the users who live in Beijing and visit such venues.

Moreover, it is worth noticing that the highest density area locates at the Beijing International airport (**Figure 60**-(a)(b)(c), 34) during the one week-period and workdays-period. It may be caused by its enormous passenger flows.

According to ACI´s data[32], Beijing airport got the No.2 rank of total passenger traffic among 1,144 airports worldwide in 2015. The enplaning and deplaning passenger was nearly 90 million in 2015. Thus, Beijing airport is a functional hub on a national scale without any doubt.

However, the airport is excluded from our regression model aimed at confirming Weibo sub-centers, because it is unreasonable to set the airport as the central point of the model and its importance corresponds to the national interests, not just the intra-urban one. Furthermore, in a preliminary model test, the inclusion of the airport reduces the coefficient of determination (R square) significantly, due to its highest density and further distance to the central city (about 25.5 km).

Table 26 The land use type of overlapping potential sub-centers

Code of Sub-center	Period				N. of overlapping sub-center	Land use
1	Day	Night	One week	Weekend	4	Commercial area
4				Weekend	1	Commercial area, subway station
5				Weekend		Residential area
6	Day				1	Guomao CBD
7				Weekend		Guomao CBD
8		Night			1	Beijing East railway station
9	Day	Day				Commercial area
10	Day	Night	One week		3	Tourist Attraction
11		Night				Residential and tourist area
12	Day			Weekend	2	Tourist Attraction
14		Night	One week		2	Tourist Attraction
15		Night			1	Residential and tourist area
16				Weekend	1	Residental area

[32] ACI : Airports Council International. Data source link: http://www.aci.aero/News/Releases/Most-Recent/2016/04/04/ACI-releases-preliminary-world-airport-traffic-rankings-

17				Weekend	1	Chaoyang park
18	Day				1	Commercial area
19	Day	Night	One week		3	University
20				Weekend		Amusement park
21		Night			1	Mixed area(office, commercial and university)
22	Day		One week	Weekend	3	Mixed area(office, commercial and university)
23	Day				1	Commercial area, Beijing Northern Railway Station
27	Day	Night	One week	Weekend	4	Mixed area(office, commercial and university)
28	Day		One week		2	University
29	Day	Night		Weekend	3	University
30				Weekend	1	Olympic Forest Park
31				Weekend	1	Residential area
32				Weekend	1	798 Art zone
33				Weekend	1	Mixed area(office, commercial and residential)
34	Day	Night	One week		3	Beijing International Airport

Source: own elaboration

Now, we adopt the logarithmic linear regression model to test these potential Weibo sub-centers. Rather than the traditional central point, the potential Weibo sub-center which has the highest average density of Weibo (Weibos per km^2) is chosen as the center to calculate the distance (*i.e.* Weibo-CBD). Therefore, the most relevant potential Weibo sub-center - the Sanlitun commercial area is set as the central point to calculate the distance. It has the highest density of all periods after the airport, which is a famous landmark in Beijing which consists of popular culture hubs, commercial centers, and more than 200 bars.

Table 27 is the summary of models for each different period. At the 90% confidence level, the only model that resulted in an acceptable statistical significance is that of the weekend Weibo density. The distance could explain 14.3% of the variation of Weibo density during the weekend. In terms of one-week period, although the model distance from the point of comparative highest Weibo density

could explain 21% of the variation of Weibo density, the statistical significance rejects the model result. However, the model is completely ineffective for the two periods of workdays, which adjusted R^2 is negative. It is caused by the R^2 is almost zero which means that the distance cannot explain the variance of Weibo activities during the workday. One possible explanation is that those potential sub-centers in Haidian districts impact the model because they were almost absent in the weekend period (**Figure 60**(d)).

Table 27 Model Summary of the four periods

(a) Model summary

Model	R^2	Adjusted R^2	Std. Error of the Estimate	F	Sig.
One week	0.210	0.132	0.246	2.671	.133
Day_Workday	0.000	-0.660	0.328	.004	.953
Night_workday	0.030	-0.022	0.221	1.616	.230
Weekend	0.143	0.106	0.356	3.853	.062

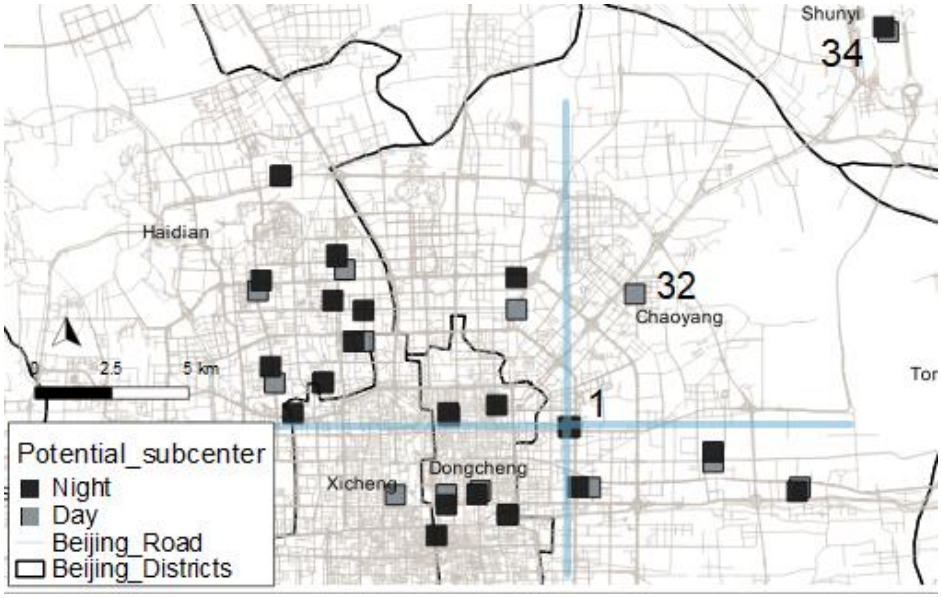
(b)Coefficients

Model		Unstandardized Coefficients		Standardiz-ed Coefficient-s	t	Sig.
		B	Std. Error	Beta		
One week	(Constant)	6.773	.143		47.264	.000
	Distance(km)	-.031	.019	-.459	-1.634	.133
Day_workday	(Constant)	5.896	.170		34.721	.000
	Distance(km)	-.001	.021	-.016	-.060	.953
Night_workday	(Constant)	3.250	.106		30.601	.000
	Distance(km)	-.011	.012	-.195	-.844	.410
Weekend	(Constant)	5.924	.142		41.630	.000
	Distance (km)	-.038	.019	-.379	-1.963	.062

Source: own elaboration

Such a lower explanatory percentage possibly indicates that the urban structure of Beijing belongs to a polycentric model from the perspective of Weibo activities. However, another possibility is that the directions of the distribution of these

potential points may influence the result. In other words, the development of subcenters is not exactly symmetrical. Thus, we set point 1 as the center to divide the map into four quadrants (**Figure 61**). This division separates the Guomao CBD, Zhongguancun area, traditional center and airport into different quadrants, thus it can alleviate the effect of opposed directional development trends. Because the 34th - airport is not involved in the model test, the 32nd is added into the second quadrant. The confirmed sub-center is the sum of each quadrant's result.



Source: own elaboration

Figure 61 Division of workday's potential sub-centers

However, none of the models is shown statistical significance (**Table 28**). The model is still completely ineffective during the day-period. The adjusted R^2 of the night period is improved after the classification because it eliminates the effect of different directional development trends. The potential candidates of the fourth quadrant have a lower density of Weibo than the candidates of other quadrants which are located at a similar distance. The first and second quadrant suffer the least influence from the distance. Because their average density is very high, more or less equals to the Guomao CBD area, but they are far away from the central point. Compared with the fourth quadrant, the third quadrant also presents that the distance has weak influence. One possible reason is that the third quadrant

belongs to the traditional center of the city, and the fourth quadrant is near to the suburban area.

Table 28 Model Summary of Workdays and nights of workdays

Model		1st&2nd quadrant	3rd quadrant	4th quadrant
Day workday	R^2	0.008	0.051	0.203
	Adjusted R^2	-0.116	-0.265	-0.196
	Std. Error of the Estimate	0.378	0.435	0.389
	F	.063	.163	.509
	Sig.	.809	.714	.550
Night workday	R^2	0.128	0.472	0.356
	Adjusted R^2	0.049	0.296	0.034
	Std. Error of the Estimate	0.218	0.227	0.262
	F	1.616	2.679	1.107
	Sig.	.230	.200	.403

(b)Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
Day_workday 1st&2nd quadrant	(Constant)	5.900	.263		22.451	.000
	Distance(km)	-.007	.028	-.088	-.250	.809
Day_workday 3rd quadrant	(Constant)	6.251	.395		15.829	.001
	Distance(km)	-.033	.081	-.227	-.403	.714
Day_workday 4th quadrant	(Constant)	6.232	.312		19.973	.002
	Distance(km)	-.040	.056	-.450	-.714	.550
Night_workday 1st&2nd quadrant	(Constant)	3.362	.144		23.385	.000
	Distance(km)	-.019	.015	-.358	-1.271	.230
Night_workday 3rd quadrant	(Constant)	3.558	.213		16.720	.000
	Distance(km)	-.074	.045	-.687	-1.637	.200
Night_workday 4th quadrant	(Constant)	3.344	.209		16.020	.004
	Distance	-.039	.037	-.597	-1.052	.403

Source: own elaboration

In summary, the only clear spatial trend of Weibo density is the weekend in terms of the model results. It is probably caused by the spatial diversity of Weibo density in these periods. Except for the weekend, the variation of Weibo density is too small to present the evident correlation with the distance. The Weibo density has the largest variation in the weekend (**Table 27**), which standard deviation is nearly double that of other periods. On the contrary, the mean distance of all potential sub-centers to the Weibo center is the lowest at the weekend.

Table 29 The tested sub-centers of weekend based on two standards

	Standard	**Sub-center**
1	Positive Residual ≥ Mean of all residuals +Std. Deviation	1,12,31
2	Positive Residual ≥ Mean of all residuals	1,4,12,14,18,22,27,29,31

Source: own elaboration

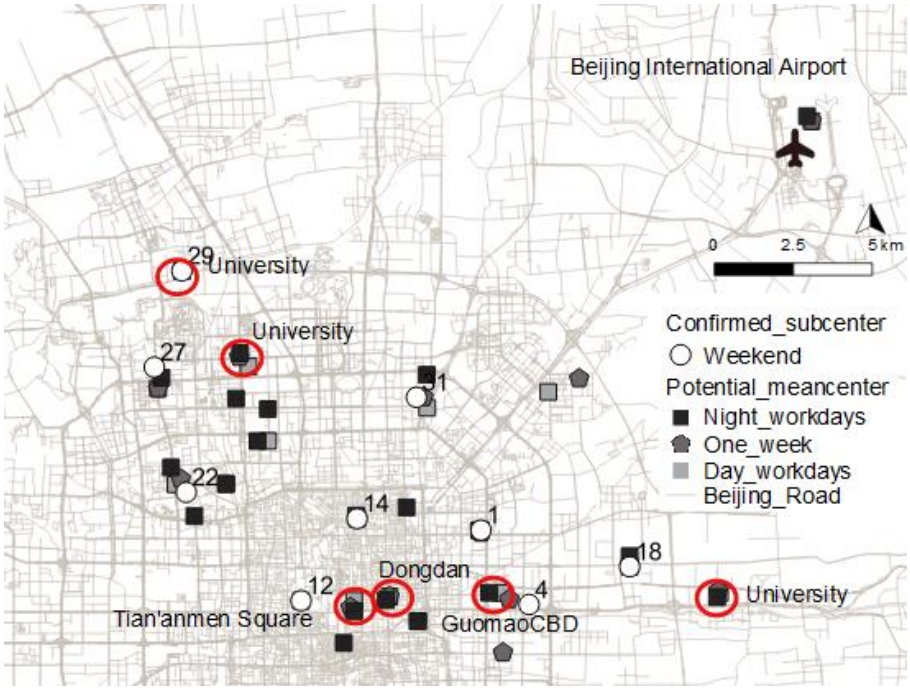
Therefore, we only select the final Weibo sub-centers at the weekend (**Table 29**). The confirmed Weibo sub-centers at the weekend are the numbers of 1, 12, and 31 following the stricter standard, while there are eight sub-centers if we use the mean-standard. The stricter one appears one commercial areas (1), one tourist attraction (12) and a residential area (31). Combined with the actual social-economic profiles, we tend to adopt the mean value as the standard of selecting sub-center. The number 18 (**Table 30**) is a shopping mall. The number 12 and 14 are in the traditional urban center. The number 22 and 27 are areas of mixed land uses and located in Zhongguancun, Haidian district. The Haidan district accounts for 17.9% of total permanent residents of Beijing in 2014. Moreover, Zhongguancun is a huge employment center as well as an important commercial area of Beijing. According to *Zhongguancun Report 2014*, 1,899 million people worked in Zhongguancun area. Considering many companies still operate at the weekend and, it is reasonable to include this area as a sub-center at the weekend.

Table 30 The land use type of confirmed sub-centers at the weekend

Code of Sub-center	Land use
1	Commercial area
4	Commercial area, subway station
12	Tourist Attraction
14	Tourist Attraction
18	Commercial area
22	Mixed area(office, commercial and university)
27	Mixed area(office, commercial and university)
29	University
31	Residential area
34	Beijing International Airport

Source: own elaboration

The distribution of Weibo sub-centers of the weekend are not located at the very center of Beijing city (**Figure 62**), many of them are out of the second ring road and concentrated in the northern part of Beijing. The figure also overlaps the potential sub-centers of the other three periods for comparison. The overlap between weekend sub-centers and other potential areas are only two of them: the number 1 is located in Chaoyang district and the number 27 are in the Haidian district. However, most of the potential sub-centers in different periods are quite close to each other. There are seven places overlap in three different periods: the airport, universities, Dongdan commercial area, and Tian'anmen Square and Guomao CBD. All types of these land use could maintain a higher intensity of human activities.



Source: own elaboration

Figure 62 Confirmed sub-centers at the weekend and overlapping potential areas

1.6. Discussion and conclusion

In summary, the urban structure of Beijing is indeed polycentric, though the high-density clusters mainly gathered within the fifth ring road of Beijing. Statistically, according to the result of the model, the period of weekend is the only one that presents the polycentric structure. However, the poor performance of the negative exponential model also indicates that the Weibo density only has a weak correlation with the distance to the center.

Secondly, as the Local Moran I's results show, the Weibo activities actually varies from place to place within the central area. Those Weibo sub-centers can present the major urban active areas, such as Zhongguancun, the traditional city center and the airport, etc.

It is necessary to stress that Weibo sub-centers are not employment centers. The employment center only considers the working activities rather than other human activities, such as recreational activities or tourist activities. The failure of the model may indicate that the empirical analysis of urban spatial structure requires more than pure employment analysis.

Thirdly, all Weibo sub-centers of the weekend and those potential areas of other periods are located in the northern part of Beijing. It is probably caused by the development of the educational zone since the 1950s and the falling of the first industries in the southern area due to the change in planning restrictions produced by the air pollution evident since the 1970s. Accordingly, the lack of Weibo sub-center in the southern part of Beijing is associated with the economic structural reform, because many heavy industry factories were located in that area.

Secondly, Weibo data is proved to be a useful data source to study the urban structure and urban functional areas. Those confirmed Weibo sub-centers actually belong to regions where are vivid zones, such as commercial areas and places of interest. Compared with the employment data, Weibo data can reflect the various types of land use more directly. Moreover, the employment data exclude the possibility that zones of universities could be potential sub-centers. Because most of the undergraduates and graduates live in the dormitories of universities in China, it forms another type of center of human activities.

Thirdly, the Weibo potential sub-centers imply the difference of human activities in different periods. It can be observed the difference in active Weibo areas between workdays and the weekend. For example, Zhongguancun as a major employment center of Beijing has more active grids during workdays than at the weekend. Meanwhile, it is reasonable to infer that working people post fewer Weibos in working places during workdays, and thus tourist attractions display a higher degree of activities in workdays by contrast. Conversely, people tend to go to recreational places or stay at home at the weekend, hence, commercial and residential areas are more active in the weekend.

Nevertheless, Weibo data exhibit limitations when it is applied to study the urban structure. Firstly, it cannot identify land uses without using exogenous data. The distribution of Weibo density is not enough to describe the whole situation of the city. It relies on other data that can provide the details of the land-use and social-economic information of a specific area. Secondly, the bias that is caused by types of users and users' preferences could affect the representativeness of Weibo data. For example, the major users of Weibo are aged from 18 to 35 years-old who tend to stay in other places longer than in their homes, thus the distribution of Weibo could not reflect residential areas very precisely.

VI.2.Case study II

Identifying functional relations of urban places through Foursquare from Barcelona

2.1. Introduction

With the increasing mobility among cities, visitors are also becoming an important part of the city; because cities provide some permanent services for them, such as hotels and tourist information centers. From the perspective of human activities, visitors usually occupy certain areas of a city. As C. Hall and Page (2003, p. 49) noted that "… tourism is subsumed and integrated into the postmodern city …it is one aspect of the form of the city." Therefore, the investigation of the difference between visitors and residents in urban usages is beneficial for the arrangement of urban facilities. Besides, it is crucial to understand the co-living situation of tourists and inhabitants in a postmodern city because the development of tourism may contribute a dominant factor of urban development as well as a source of conflict.

Moreover, previous studies of activities between visitors and locals mainly focus on the spatiotemporal comparison, such as the spatial scale of activities(Kádár, 2014b; Vu et al., 2015) or patterns of movement (García-Palomares, Gutiérrez, & Mínguez, 2015; Hasnat & Hasan, 2018). However, the spatial relation is not sufficient to explain the functional relationship of places. The active degree of human activities in some specific areas, such as employment centers or commercial areas, only indicates what kind of land uses tends to be used more often, rather than connections between different urban usages.

What's more, the description of urban usages the city is usually based on practical experience. For example, it is well known that some places, such as hotels

and tourist attractions mainly belong to facilities of tourism. Meanwhile, locals often use workplaces, gyms, etc. However, can these urban usages be quantified?

Therefore, this research chooses Barcelona as the case study, to investigate different behaviors and activities between visitors and local people using Foursquare data. The case study proposed in this paper investigates functional relations between tourist places in Barcelona using 18 months of Foursquare (an LBSN) check-in activity.

Firstly, according to users' behaviors, we separate Foursquare users into two groups: tourists and residents. Secondly, it investigates the two groups' movements and activity space analytically and graphically, aiming to identify their patterns of behaviors in Barcelona.

Thirdly, this study aims to quantify the important functional relations between urban places, which are usually not noticeable on spatial proximity relations. Here, the functional relation between places is defined as the flows of tourist/local users between Foursquare POIs. Further, we use the functional interaction value to measure the functional proximity of different groups of POIs which are classified by different categories of usages.

The main contribution of the study lies in its practical method of quantifying the functional relations of places, rather than trying to depict the dynamics of activities in a specific city. The study process mainly comprises four steps: identification of tourist users, classification of POIs based on usage, calculation of flows among different categories of POIs, and analysis of functional relations. A network graph is used to present the significant relations among different usages. Finally, we also compare the spatial proximity of POIs with their functional proximity for finding out their differences.

The remainder of this paper is structured as follows. The literature review summarizes the main methods to identify tourist users and related studies of tourist and local activities using LBSN data. The part of methods introduces the methodology and implements the identification of tourists within the entire

sample. The result analyzes the spatial characteristics of movement of both groups and the functional relationships among places in Barcelona. Finally, the conclusion discusses the limitations and future work.

2.2. State of the art

2.2.1 The characteristics of tourist and local activities

In general, the scale of tourist activities is more spatially concentrated than locals' movements. Urban centers, airports, and tourist attractions, such as famous churches or museums, are typical clustering places for tourists. For example, Girardin, Calabrese, Dal Fiore, Ratti, and Blat (2008) identified the differences of spatial activities between tourists and locals in New York via cell phone data and Flickr data. They confirmed that the movement range of visitors is limited, especially for foreign visitors. Béjar Alonso (2014) extracted the spatial-temporal characteristics of Twitter and Instagram users in Barcelona; they found that the main tourist attractions were important connecting nodes in both datasets, though their study did not distinguish tourist users from locals. Kádár (2014b) concluded that the majority of geo-tagged images generated from Flickr tourist users were gathered around tourist attractions or landmarks in Budapest.

Secondly, the frequent visiting places of tourists also has some universal patterns in cities. For example, Y. Li, Steiner, Wang, Zhang, and Bao (2013) identified the integration degree between tourists and locals in ten US cities. Although the level of integration was different from city to city, tourists showed the similar preference on visiting places, such as landmarks, historical building, etc.

Regarding local mobility, many studies have disclosed that local people and tourists share urban spaces, though tourist movements as clustered around tourist attractions. Schwitzguébel and Bartomeus (2018)compared the spatial patterns of business usages between tourists and residents in 11 metropolitan areas through Yelp data. They concluded that both of them consumed the similar urban spaces

and venues with leisure functions. Béjar et al. (2016) extracted the spatial-temporal characteristics of Twitter and Instagram users in Barcelona; they found that the main tourist attractions were important connective nodes in both datasets, though their study did not distinguish tourist users from locals.

On the other hand, on a macro level, some similar spatial characteristics of residents' activities have shown in different cities. First of all, people's movement seems to be geographic proximity. the "locality of social media behaviors" has been mentioned in several researches(Hasan & Ukkusuri, 2015; Sun, 2016). Hasan, S., & Ukkusuri, S. V. (2015) studied the patterns of daily mobility in New York city and found that Foursquare users tended to choose nearby places for satisfying different demands. Based on three different LBSN datasets, Cho et al. (2011)found that periodic behaviors (moving between home and workplaces) account for 50%-70% of all human movements.

However, such spatial-temporal characteristic fails to uncover how places interact with each other functionally, since it only shows the degree of spatial aggregation of human activities in separate places or the tendency of people's movements. Although the quantification of flows between places have been involved in studies of travel patterns and land-use detection, the focuses are still different from functional linkages. Travel patterns focus on the temporal patterns of human mobility(F. Luo et al., 2016; Sagl, Resch, Hawelka, & Beinat, 2012; Thuillier, Moalic, Lamrous, & Caminada, 2017). Land-use detection aims to reveal the relation between spatial distribution of POIs and urban land-uses(S. Gao et al., 2017), rather than the connections between different urban usages.

To date, few studies have investigated the functional linkages among POIs; so, it still offers room for exploration. The most closely related study to the present work is from Ferreira et al. (2015). They collected Foursquare data in London, New York, Rio de Janeiro, and Tokyo. They classified POIs into nine categories and compared the check-in patterns of tourists and residents in these categories. They found that some categories show a significant temporal difference of check-in pattern between tourists and residents, such as locations classified as Arts and as Transport. However, they focused on the mobility patterns between specific places,

rather than the functional closeness between different categories. Preoţiuc-Pietro and Cohn (2013)discussed the probability of transition of users between different types of Foursquare POIs, but did not explain the functional relationships among POIs. Therefore, our study actually proposes a new perspective to analyze the functional relations between places.

2.2.2 Identifying tourists from locals

Field surveys are a traditional approach to identify tourists. For example, Mckercher and Lau (2008)identified tourists by conducting questionnaires in hotel lobbies. Obviously, it is difficult to implement such method when large number of tourist samples are needed. LBSN data offers the potential to identify tourists from locals at low cost.

According to previous studies, the identification mainly relies on the geo-location(Da Rugna, Chareyron, & Branchet, 2012; F. Luo et al., 2016; Vu et al., 2015); or the time threshold(García-Palomares et al., 2015; Girardin, Calabrese, Dal Fiorre, et al., 2008; Kádár, 2014b). For example, F. Luo et al. (2016) residential Twitter users from visitors in Chicago by determining their locations during the night. Users are identified as locals if most of their check-ins during nights are in residential areas. However, this method is effective only if hospitality services are segregated from residential areas. It is also difficult to apply the approach in a compact city which has large amounts of mixed-use land, such as Barcelona.

Manca, Boratto, Roman, i Gallissà, and Kaltenbrunner (2017) combined the time threshold and the geo-location of Twitter users to distinguish tourist users from locals. Users who posted tweets less than 20 days in Barcelona as considered as tourists. In general, a longer time threshold could be more reliable; however, it requires a dataset with more long-term data. Moreover, the determination of the time threshold is usually derived from empirical experiences or the advice of tourism experts, and so lacks objectivity.

Above all, this study sought to classify users as tourists or local residents by examining users' behaviors on Foursquare and applying a threshold based on

statistical analysis of the dataset. A semi-supervised model, described in the next section, was adopted to identify tourist users. This method allowed further exploitation of the dataset, because both active and inactive users are included.

2.3. Methods

2.3.1 Study scope

Barcelona is an ideal city for an analysis of tourism; it has been known as a tourist destination since the early 1900s. The government established a Commission for the Attraction of Foreigners and Tourists in 1906 and aimed to build up the city as a tourist destination known as the "Pearl of Mediterranean". Its importance as a tourist destination only increased after the Summer Olympic Games of 1992. According to the Annual Tourism Sector of Barcelona Report 2014, the total number of overnight tourists who stayed in hotel accommodation reached more than 7.5 million, ranking as the 20th most-visited city in the world. In 2017, this number rose to nearly 9 million. Currently, according to the local government of Barcelona, tourism accounts for about 12% of the city's GDP and generates approximately 9% of employment in Barcelona.

This paper extracted data from a global Foursquare check-in dataset (Yang, Zhang & Qu, 2016). The studied area includes the inner area of the Barcelona Metropolitan Region (RMB), due to the monitoring range of Foursquare data(**Figure 63**), which is slightly larger than the first zone of the RMB. The first zone is also called the Metropolitan Area of Barcelona (AMB), which comprises 36 municipalities. According to the official statistics, the population was 3,239,337 in 2014, about half of whom lived in the city of Barcelona. The land uses of Barcelona city is highly mixed. Most check-ins, 57,764 items, occurred in Barcelona city (**Figure 63**). Other than Barcelona, only four cities in the region have more than 1,000 check-ins: L'Hospitalet de Llobregat, El Prat de Llobregat, Badalona, Sant Cugat del Vallés and Conellà de Llobregat.

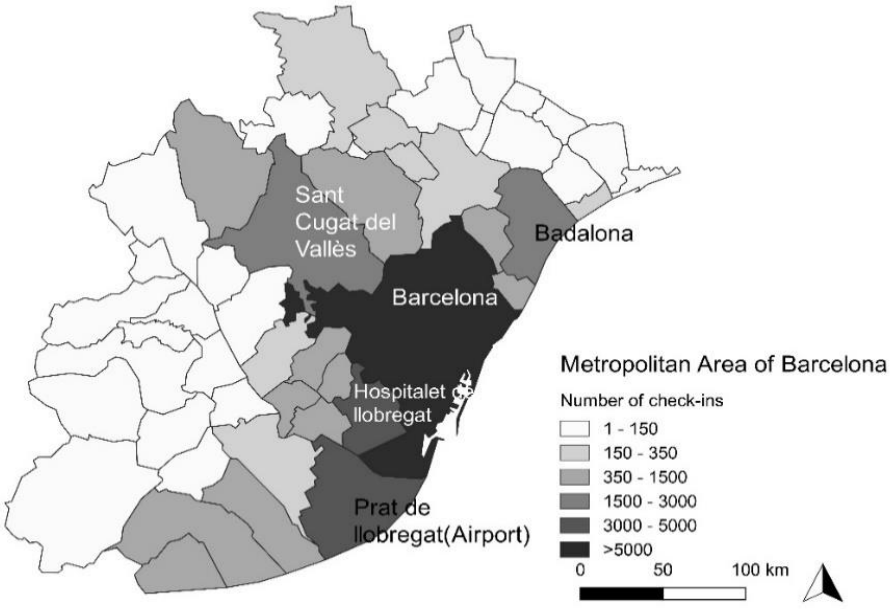


Figure 63 Distribution of check-ins in Barcelona Metropolitan Region

2.3.2 Desciption of Foursquare dataset

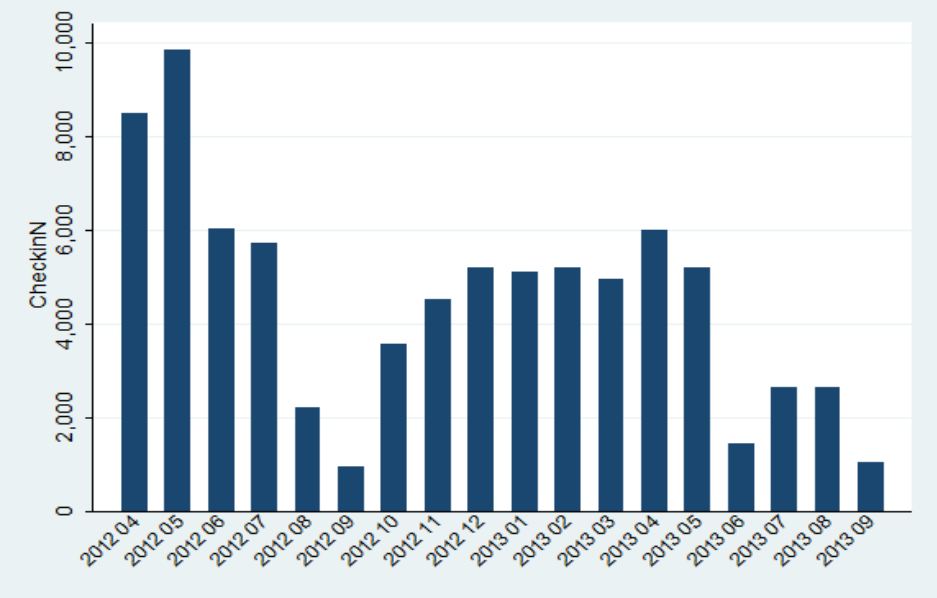
Created in 2009, Foursquare is a local search-and-discovery service application, which could provide practical information of living for users. On Foursquare, users make "check-ins" at venues which are predefined or created by themselves. Because Foursquare data provides the attributes of POIs and their geographic location, it is a popular source of information for mining online-user behaviors(Preoţiuc-Pietro & Cohn, 2013), urban mobility(Hasnat & Hasan, 2018), improvement of categorization systems(Y. Hu, McKenzie, Janowicz, & Gao, 2015), among other applications. The main components of Foursquare data include venue (i.e. a place), Foursquare users, and their activities on the platform(**Table 31**) .

Table 31 Components of Foursquare data

Attribute	Venue	Users ID	Check-in	Timestamp
Description	Points of Interest (POI) in a certain area. It also includes some details of the places, such as tips, likes.	The total number of users in a venue	The number of check-in in a certain place of interest	The created time of the check-in

Source: own elaboration

This paper extracts data from a global Foursquare check-ins dataset that collected data between 2012-04-03 and 2013-09-16(Yang, Zhang, & Qu, 2016). A total of 80,936 check-ins were made by 4,527 users in the area of Barcelona. The number of check-ins declines significantly after June of 2013, this period only has 6931 check-ins(**Figure 64**). It may be caused by the declining of active users of Foursquare, or some adjustments of privacy policy, or some unknown technic problems. There were also missing data from 25 of August to 03 of September, and 25 of September to 16 of October in 2012.



Source: own elaboration

Figure 64 Monthly Check-ins of Foursquare in Barcelona

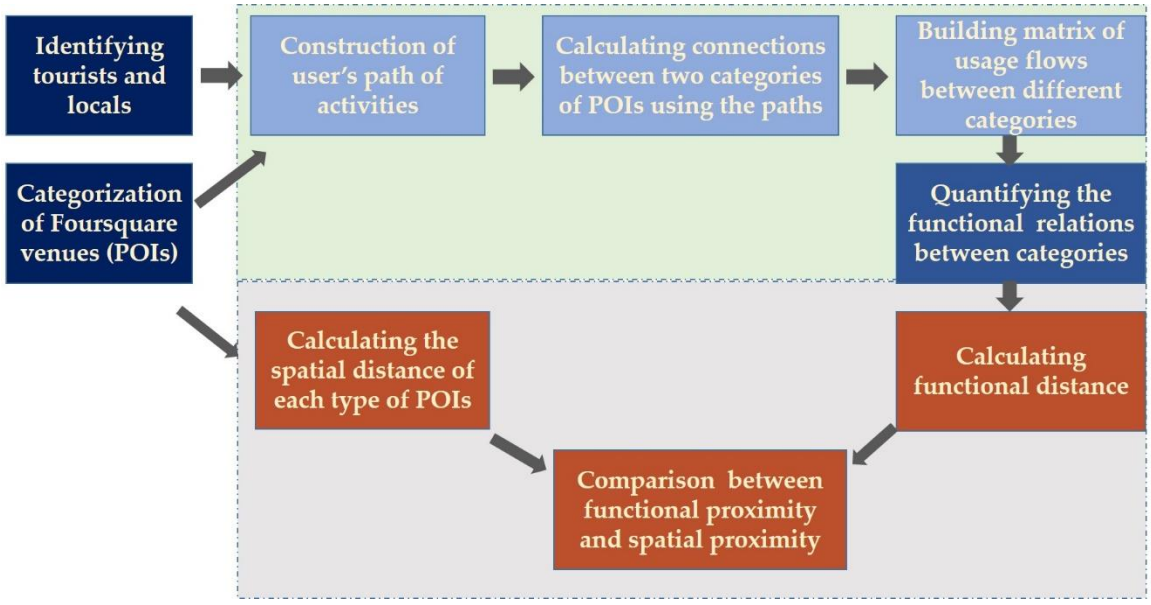
To eliminate noise, users who only checked-in one time in the area were excluded from the dataset. Therefore, 3,350 "valid" users with a combined total of 79,798 valid check-ins were included (**Table 32**). The duration means that the period that the user stayed in Barcelona, which is difference between the last timestamp and the first timestamp of the user.

Table 32 Summary of valid users

Total Users		Mean	Std. Dev.	Min	Max
3,350	Check-ins	23.8203	54.14576	2	1,182
3,350	Duration (days)	112.5728	179.0711	0	531

Source: own elaboration

2.3.3 Process of analysis



Source: own elaboration

Figure 65 Outline of analzying process

The method for measuring the functional relations of POIs included eight steps: 1) identifying tourist users as a group distinct from local users and classifying all Foursquare POIs into 22 different categories for purpose of analysis; 2) constructing the chronological path of activity of each tourist user based on their check-ins; 3) calculating the number of connections between two categories of POIs using the paths; 4) building a heat map of the matrix of flows; 5) introducing a model of interaction values to investigate the functional relation of these categories; 6) representing the matrix of interaction values in two dimensions through PROXCAL multidimensional scaling (MDS); 7) visualizing the prominent functional relationships via a network graph;8) comparing the functional proximity with the spatial proximity based on different categories of POIs.

2.3.4 Identification of tourists and locals

This study uses a semi-supervised method to distinguish tourist users from locals using Foursquare data. This method consists of K-means clustering with manual improvement. Departing from previous research, it is based on the assumptions that the number of users' check-ins, total travel distance, and duration of stay. The method of calculation of duration is the following:

$$\text{Total duration of stay} = T_{last} - T_{first} \tag{1}$$

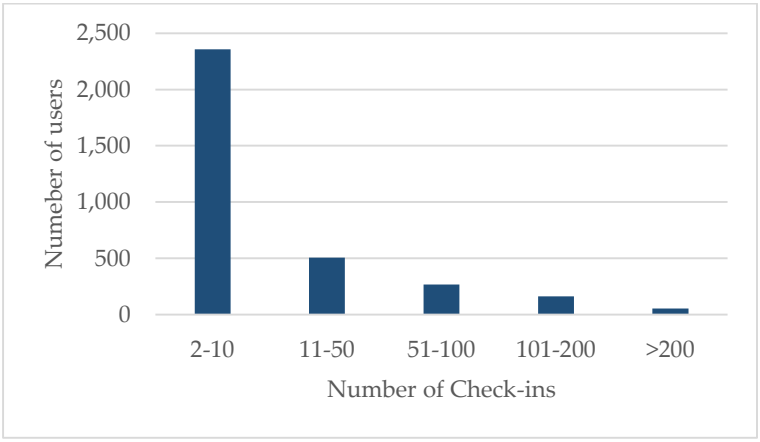
where $T_{last}$is the latest timestamp of a check-in of a user and T_{first} is the earliest timestamp.

The travel distance is calculated via ArcGIS software:

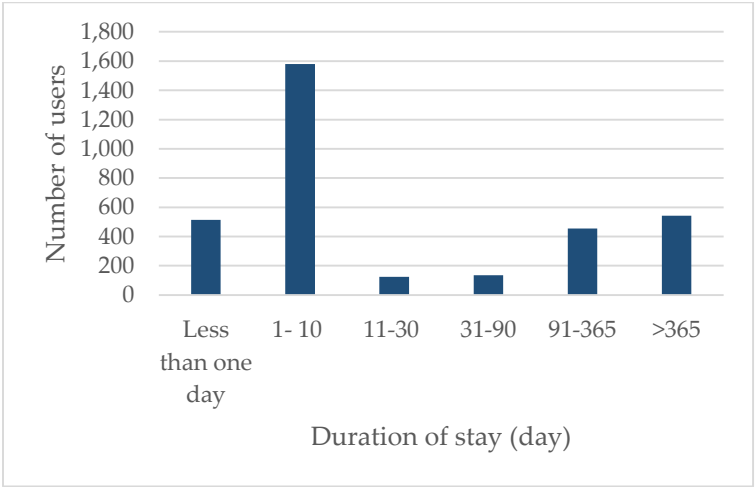
$$\text{Total Travel Distance} = \sum_{i}^{n} \sqrt{\left(X_{P_{t+1}} - X_{P_t}\right) + \left(Y_{P_{t+1}} - Y_{P_i}\right)} \tag{2}$$

where $P_t$is the position of a given check-in at timestamp i, with coordinates (X_{P_t}, Y_{P_t}) of the Universal Transverse Mercator (UTM) system. The distance between P_{t+1} and P_t is calculated by the straight-line distance between two points.

Before clustering, several users were chosen to form a sample to test the validity of the K-means clustering by checking whether they are classified correctly. One user was initially randomly selected from each group of different frequency of check-ins (**Figure 66(a)**), combined with the different duration of stay (**Figure 66(b)**). In other words, each selected user was from different groups of check-ins and duration.



(a) User's check-ins



(b) Duration of stay

Figure 66 Foursquare users' check-ins and stay duration

Therefore, except null values of inquiries, 23 users were selected out and their identities were manually verified. According to the numbers of check-ins, the

places where they checked-in and their duration of stay, we identified whether the selected user is a tourist or a local. For example, if a user checked-in 10 times and the duration of stay was two days, and all places where the user checked-in are places of amusement or tourist attractions; we classified this user as a tourist.

Next, we utilized K-means clustering to divide all users into the two groups. For discrete data, algorithms of grouping data are classification and clustering. Classification requires a training dataset which contains samples whose category is known. As the characteristics of tourist behavior were unknown in our case, clustering was the better approach to divide users. K-means clustering is widely applied due to its simplicity. Moreover, K-means algorithms perform very well with huge datasets (Abbas, 2008). The Z-score was used to standardize the three indicators for clustering:

$$z = \frac{(x - \mu)}{\sigma} \quad (3)$$

where z is the standardized score of each of the indicators, x is the value of indicator, μ is the mean of x, and σ is the standard deviation.

Manual examination showed that two local users of the sample group were included in the tourist group because they had comparatively lower duration of stay. On the other hand, four tourist users of the sample group were included in the local group because their duration of stay were too long. Such difference is partly caused by our method of calculation of duration. It is possible that the duration of stay was incorrect for some visitors returning in the second year. Therefore, it was necessary to use a threshold of check-ins and duration to improve the classification. Four different combinations of threshold were tested (**Table 33**). Those users whose indicators were above the threshold were categorized as locals.

Table 33 Four thresholds of improvement

Threshold	Total duration (days)	Total check-ins (times)	Description

1	84	18	Mean values of duration and check-ins based on all 4,527 users
2	90	24	Empirical threshold
3	98	21	Mean values of duration and check-ins based on threshold 1 and 4
4	113	24	Mean values of duration and check-ins based on 3,350 valid users

Source: own elaboration

The results of classification were more stable when time span was more than 84 days. Therefore, the strictest threshold, 4, was adopted as the final standard for improvement. Users who stayed more than 113 days and made check-ins over 24 times are classified as locals. In total, 2,770 tourist users were identified. These users generated 19,180 check-ins during the monitoring period (**Table 34**). 580 residents created 76% of check-ins.

Table 34 The summary of locals and tourists

	Number	**Check-ins**	**Average number of check-ins**
All Users	3,350	79,798	23.82
Residents	580	60,618	104.51
Tourists	2,770	19,180	6.92

Source: own elaboration

2.3.5 Classification of Foursquare POIs

With regard to the features of Foursquare POIs, it seems that the distribution of check-ins is mainly clustered into a few categories: Travel & Transport, Food, and Shopping. Abbasi and Alesheikh (2018)mentioned that shopping and eating places contributed 59% of check-ins in Manhattan. Y. Li et al. (2013) investigated the popularity of Foursquare POIs in 14 urban regions all over the world. Their results show that the Travel &Transport category occupies the highest frequency of check-ins. Preoţiuc-Pietro and Cohn (2013)collected Foursquare data of frequent

users worldwide for one month. They used the basic category from Foursquare website and learned that Shopping & Services, Food, and Travel &Transport accounted for 53.8% all check-ins.

There are 13,887 unique Foursquare POIs in Barcelona, which are labeled by 385 sub-categories. Restaurants form a large portion of all types of POIs. According to the official website, the venue categories of Foursquare classifies them by nine major categories: Art and entertainment, Faculty and University, Event, Food, Night spots, Outdoor Recreation, Professional and others, Store and services, Travel and transportation. However, these categories need to be improved for conducting the tourist-functional analysis. For example, Hotel is under the category of Travel and transportation, but Hotel actually belongs to the category of accommodation. Moreover, some names of the category are too vague. For example, "Event" actually contains temporary and permanent markets. It is clearer to use "market" rather than "event".

Therefore, this paper assembles these categories into 22 main types considering their usages, for example, all kinds of restaurants are grouped as "Restaurant". **Table 35** lists the new classification with brief descriptions. The Transport, Restaurant, Hotel and Outdoor Resorts make up a combined 47.5% of check-ins. It is worth noting that we extracted Plaza as a separate category, because it is a compound urban place which mixes multiple functions, such as leisure, transport hub, food, shopping services, etc. The volume of check-ins in the Plaza category also indicates that it is an important functional hinge for tourists in Barcelona.

Table 35 New category of Foursquare POIs

Types of POIs	N. of POIs	N. of tourist check-ins	Avg. tourist check-in per POI	Description
Restaurant	3,318	2,726	0.82	Mediterranean Restaurant, Japanese Restaurant, Food, Diner, etc.
Public services	1,321	7,46	0.56	Medical, Finance, Post Office, Bakery, Salon, Barbershop, Spa, Tattoo, etc.

Workplace	1,174	585	0.50	Building, Campaign Office, Co-working Space, Design Studio, etc.
Bar	1,138	1,031	0.91	Bar, Beer Garden, Cocktail Bar, Jazz Club, Nightclub, etc.
Outdoor resorts	1,061	2,017	1.90	Tourist attractions, Rest Area, Park, Scenic Lookout, etc.
Shop	940	789	0.84	Bike Shop, Dessert Shop, Frozen Yogurt, Gift Shop, etc.
Store	839	1,398	1.67	Kids Store, Pet Store, Paper / Office Supplies Store, Video Store, etc.
Transport	668	2,792	4.18	Train Station, Subway, Airport, Boat, Airport Terminal, Light Rail, etc.
Education places	634	428	0.68	University, College, Elementary School, Student Center, etc.
Café	540	330	0.61	Café, Tea Room, Cafeteria, etc.
Residential place	506	563	1.11	Neighborhood, Residential Building (Apartment / Condo), etc.
Hotel	478	1,576	3.30	Motel, Hotel, etc.
Sports center	308	581	1.89	Athletic & Sport, Baseball Field, Basketball Court, Football Stadium, Golf Course, etc.
Museum, Art, Historical place (MAH)	206	1,156	5.61	Public Art, Performing Arts Venue, Museum, Historic Site, Castle, etc.
Opera, Concert, Cinema (OCC)	192	186	0.97	Indie Movie Theater, Concert Hall, Movie Theater, Opera House, etc.
Gym	148	154	1.04	Gym Pool, Gym, Yoga Studio, etc.
Infrastructure	131	443	3.38	Bridge, Harbor / Marina, River, etc.
Conference center	88	412	4.68	Conference Room, Meeting Room, Convention Center, etc.
Market	74	177	2.39	Fair，Farmers Market，Flea Market，Fish Market, etc.
Plaza	57	1,062	18.63	Stables, Track, Planetarium, etc.
Others	51	24	0.47	Plaza
Tourist info center	5	4	0.80	Tourist Information Center

Source: own elaboration

2.3.6 Construction of paths of Foursquare users

In this study, a path is defined as the time-sequential check-ins of a user; each user has a unique path. Based on the study of Scholtes (2017), the construction of the paths is built on two assumptions:

(1) Each user's path has a chronological order, hence, it is directed.

(2) Paths are not transitive. Only direct connections count. For example, assuming there is a path: a →b →c, [a,b] and [b,c] are valid connections and [a,c] is not counted.

This method avoids the duplicate calculation of connections among nodes. We constructed the functional paths of a user in terms of the category of usages.

2.3.7 The matrix of usage-flows

Based on the paths, we calculated the number of direct connections between each pair of POIs, including the connections within the same category. The "inflow" from the category i to j is the total number of connections from i to j. The reverse is the "outflow". The "flow" of paired usage is the sum of the "inflow" and the corresponding "outflow":

$$f_{ij} = C_{ij} + C_{ji} \tag{4}$$

where f_{ij} is the flow between i and j, and $C_{ij}$is the number of connections from i to j.

2.3.8 Visualization of spatial paths

To visualize the spatial distribution of tourist paths, all paths taken by tourist between two POIs were counted. There are several methods to aggregate paths, such as edge bundling(Graser, Schmidt, Roth, & Brändle, 2019) or aggregating paths through characteristic points(Adrienko & Adrienko, 2010). However, the former method requires very high-capacity computation, because it needs to compare the similarity of each path and then implement aggregation. The latter method extracts the characteristic points of each path first, and then calculates the centroids of these points in term of the distance between points to reduce the cost of computation. The aggregated flows are then generated from these centroids. The shortfall of this method is that the centroids are not the actual places, and thus it is hard to reflect the precise spatial relationships between places on small scales. Therefore, this study used the original traces to display the spatial relationships

between POIs and tourist movement. To delimitate the major flows of users' activities, only traces that repeated more than once are visualized.

2.3.9 The interaction values analysis

This study introduces the improved model of interaction value from Roca Cladera and Moix Bergadà (2005)to depict the functional interactions among different usages. The first model of interaction value was created by Smart (1974), and was developed from the gravity model. The advantage of that model is that it explores the functional relation between two areas or objects without the inference of physical distance. It also eliminates distortion caused by differences in the "masses" of objects (*i.e.* the number of tourists visiting each of the POIs in our case), because the function divides the product of total flows of "sender" and "receiver" POI. Thus, this model can uncover the interaction relation between two objects effectively. It has been invoked in different studies of interactions, such as commuting flows between two areas(Roca Cladera et al., 2009), immigration flows(Dou, Arellano Ramos, & Roca Cladera, 2018) and air passenger flows (Burns, Cladera, & Bergada, 2008). It takes the form:

$$IV_{ij} = \frac{f_{ij}^2}{O_i \cdot I_j} + \frac{f_{ji}^2}{I_i \cdot O_j} \quad (5)$$

where $IV_{ij}$is the interaction value between the category i and j ; f_{ij} is the existing flow from category i to j ; $O_i$is the sum of outflows of category i, I_i is the sum of inflows of category i. Moreover, a statistic threshold was set to delimit the prominent relations of interaction:

$$\text{Prominent interaction value} = \text{Mean value} + 1 \text{ standard derivation of IVs} \quad (6)$$

where mean value and standard derivation are the values of the whole matrix of interaction value.

To visualize the functional proximity between different usages, we used a PROXCAL multidimensional scaling (MDS) method to reduce the original matrix of interaction values to only two dimensions. MDS is a fast way to visualize the

level of similarity of objects. In our case, the similarity is understood as the strength of interaction values between usages. The closer is one usage to other, the stronger their functional relationship. The closer is one usage to the center of the graph, the stronger its relationship with all other usages.

2.3.10 Comparison between spatial proximity and the functional proximity

To further analyze the difference between functional proximity and the spatial proximity of POIs, we also plotted the average geo-distance and the "functional distance" of each types of POIs. The functional distance is represented as the distance to the zero point of the PROXCAL plot:

$$Functional_{dist\,i} = \sqrt{PX_i^2 + PY_i^2} \tag{7}$$

where (PX_i , PY_i) are the coordinates of i usage on the PROXICAL two-dimension plot. The spatial distance is calculated using the Euclidean distance from the weighted median center of all POIs containing tourists' check-ins. Considering Foursquare check-ins are highly concentrated in Barcelona city (**Figure 63**), it is reasonable to use the weighted median center as the central point to measure the spatial distribution of different types of POIs approximately. The advantage of median center is that it could indicate the spatially central tendency meanwhile is robust to outliers. Based on the algorithm of ArcGIS, the weighted median center(Burt, Barber, & Rigby, 2009; Kulin & Kuenne, 1962) is given as:

$$D_p^t = \sqrt{(X_p - X^t)^2 + (Y_p - Y^t)^2 + (Z_p - Z^t)^2} \tag{8}$$

where (X_p, Y_p) is the geo-coordinates of a POI, Z_p is the weight of a POI which is the number of tourist check-ins in our case, (X_t, Y_t) is a candidate median center at the t step of the iterative process, D_p^t is the distance between the candidate center and other POIs at step t. The final median center minimizes the Euclidean distance to all other points in the dataset. Therefore, the average geographic distance of each type of POI is the corresponding number of POIs divide by the total distance of the corresponding POIs:

$$Spatial_{dist_i} = \frac{\sum_i^n \sqrt{\left(X_{P_i} - X_M\right) + \left(Y_{P_i} - Y_M\right)}}{N_i} \quad (9)$$

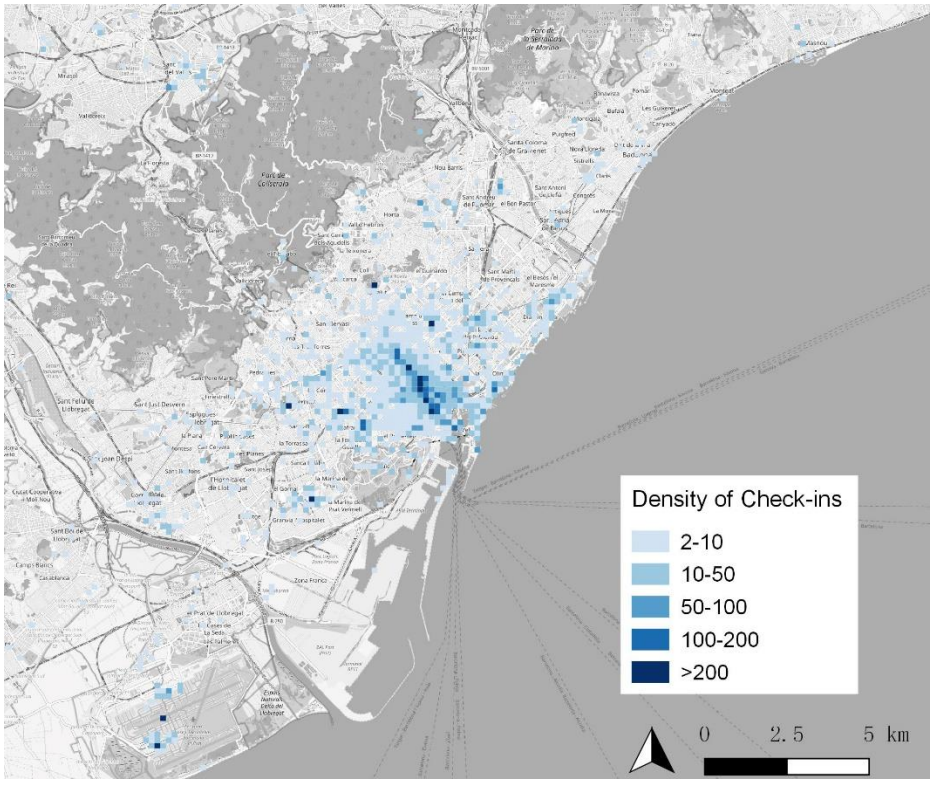
where (X_{P_i}, Y_{P_i}) is the geo-coordinates of a POI of the i type of usage, (X_M, Y_M) is the median center, N_i is the total number of POIs of the i type.

2.4. Results

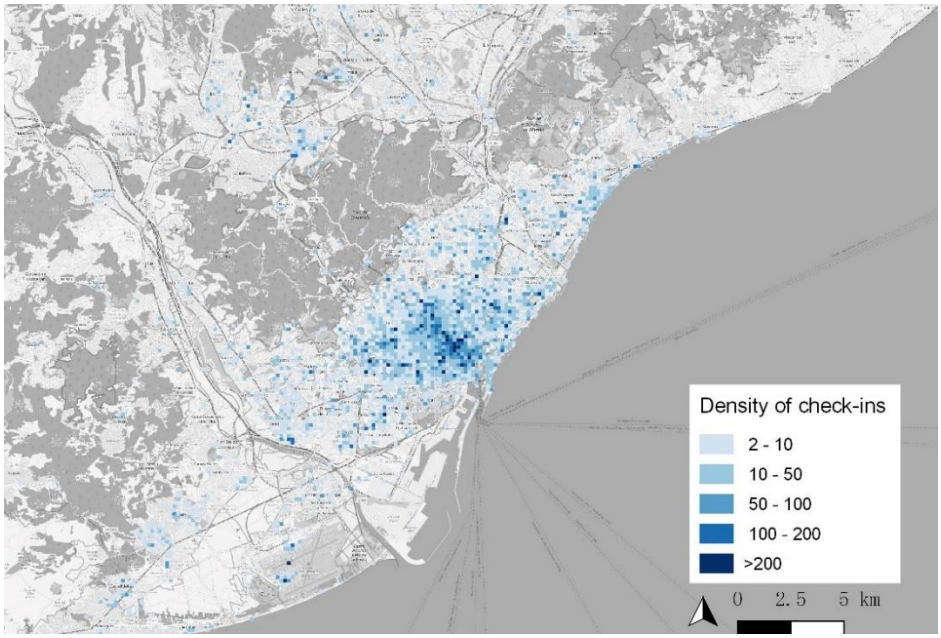
2.4.1 Spatial distribution of tourists and locals

Regard to the spatial distribution of check-ins, **Figure 67** aggregated check-ins by 0.003 km^2 squared cells, which shows that tourists' activities mainly concentrated in Barcelona city and the airport area. The place with the densest population of tourists was the airport(**Figure 67** (a)), which garnered about 8% of total check-ins. Residents' activities(**Figure 67**(b)) spread of several municipalities of Barcelona Metropolitan area, though the densest area is also located in the central part.

77% of check-ins were concentrated in the municipality of Barcelona. Nearby municipalities contributed the left part of the check-ins(**Table 36**). Especially, the densest area basically follows two famous avenues -- Passeig de Gràcia and La Rambla. It indicates that the city center of Barcelona is the most active urban space for both locals and tourists.



(a) Tourists' activities



(b) Residents' activities

Source: own elaboration

Figure 67 Spatial activities of tourists and residents

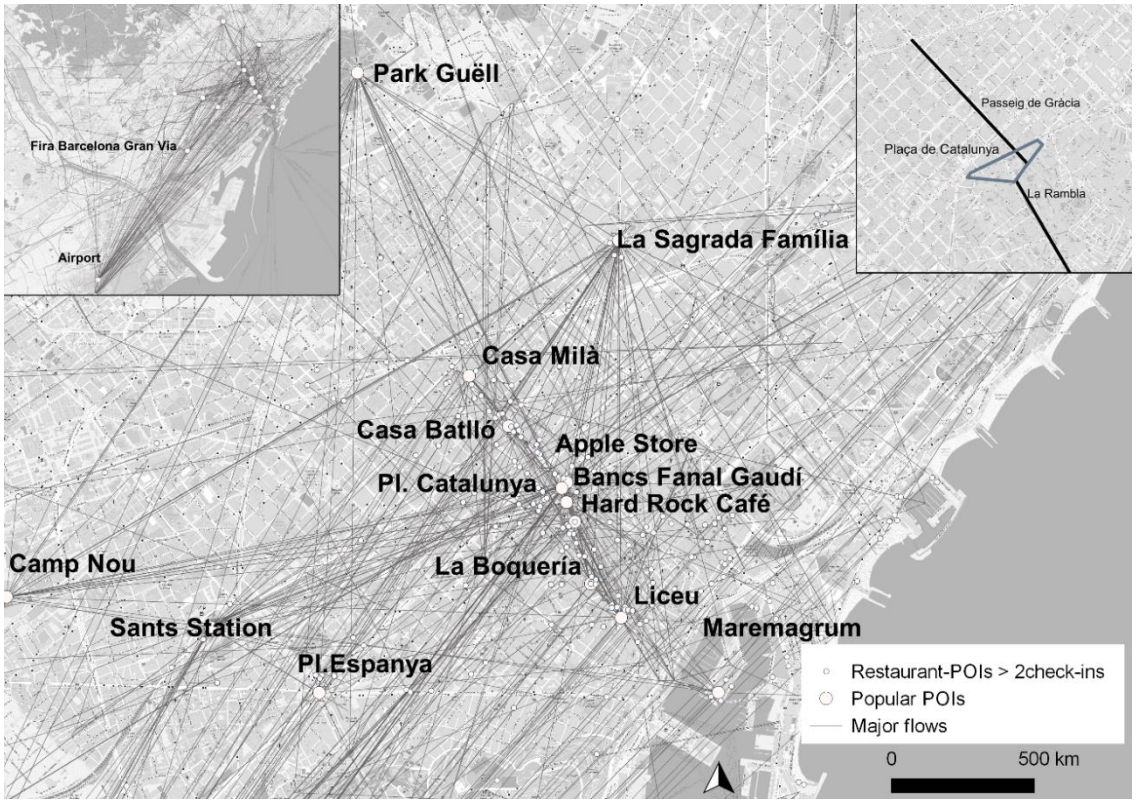
Table 36 Spatial distribution of check-ins among municipality

Municipality	Locals	Tourists	N. of check-ins	Percentage of check-ins
Barcelona	42940	14826	57766	72.39%
Hospitalet de Llobregat	2975	781	3756	4.71%
Prat de Llobregat (El)	2640	2071	4711	5.90%
Badalona	2528	119	2647	3.32%
Sant Cugat del Vallès	2134	343	2477	3.10%
Cornellà de Llobregat	673	148	821	1.03%
Cerdanyola del Vallès	651	62	713	0.89%
Gavà	617	37	654	0.82%
Castelldefels	495	109	604	0.76%
Viladecans	492	24	516	0.65%
Sant Just Desvern	423	8	431	0.54%
Esplugues de Llobregat	421	85	506	0.63%
Santa Coloma de Gramenet	400	15	415	0.52%
Sant Joan Despí	363	43	406	0.51%
Sant Adrià de Besòs	352	62	414	0.52%
Montgat	266	5	271	0.34%
Rubí	264	55	319	0.40%
Sant Feliu de Llobregat	233	9	242	0.30%
Sant Boi de Llobregat	181	37	218	0.27%
Ripollet	171	58	229	0.29%
Barberà del Vallès	156	10	166	0.21%
Sabadell	146	25	171	0.21%
Masnou (El)	116	37	153	0.19%
Sant Vicenç dels Horts	106	4	110	0.14%
Montcada i Reixac	96	27	123	0.15%
Santa Perpètua de Mogoda	88	3	91	0.11%
Tiana	86	3	89	0.11%
Sant Andreu de la Barca	74	6	80	0.10%
Sant Fost de Campsentelles	74	0	74	0.09%
Molins de Rei	72	21	93	0.12%
Torrelles de Llobregat	46	2	48	0.06%
Alella	45	32	77	0.10%
Sant Quirze del Vallès	39	55	94	0.12%
Cervelló	38	1	39	0.05%

Santa Coloma de Cervelló	37	3	40	0.05%
Vallirana	33	37	70	0.09%
Palma de Cervelló (La)	31	0	31	0.04%
Castellbisbal	25	8	33	0.04%
Llagosta (La)	22	3	25	0.03%
Begues	18	3	21	0.03%
Papiol (El)	14	0	14	0.02%
Martorelles	13	0	13	0.02%
Badia del Vallès	9	1	10	0.01%
Pallejà	6	2	8	0.01%
Sant Climent de Llobregat	4	0	4	0.01%
Mollet del Vallès	3	0	3	0.00%
Corbera de Llobregat	1	0	1	0.00%
Santa Maria de Martorelles	1	0	1	0.00%

Source: own elaboration

As to the specific locations of active Foursquare users' activities, 64% of POIs do not have records from tourists, suggesting that the range of tourists' activity is limited. Only 18 POIs had more than 100 check-ins (**Figure 68**). Except for the airport, the left of them are located in the central area of Barcelona. Plaça de Catalunya takes the second rank of check-ins after La Sagrada Família. It is one of most important transport hubs and public spaces in Barcelona city, and it connects to the historical center of Barcelona. The Apple store, the main building of El Corte Inglés (the main department store in Spain), and the Hard Rock Café are located around Plaça de Catalunya. It is probable that each of these locations are meeting points for tourists since they have an outstanding position and/or buildings. Hence, the Plaça de Catalunya can be considered another activity center for tourists. Two other commercial places received high numbers of check-ins. Maremagrum is a shopping mall near the Aquarium of Barcelona that provides multiple services for tourists. Fira Gran Via Barcelona is an international conference center which hosts many important conferences every year, such as the World Mobile Conference.



Source: own elaboration

Figure 68 Major tourist spatial flows and POIs' check-ins

Moreover, tourist attractions are clearly important nodes driving the spatial flows of visitors. La Sagrada Família, Casa Milà, Casa Batlló and Park Güell belong to famous historical heritage of Barcelona, while Camp Nou is a world-famous soccer center. These POIs were also on the official list of the top 15 most-visited places in Barcelona during 2012-2015 (**Table 37**). Such consistency means that tourist attractions indeed organize the spatial movement of tourists.

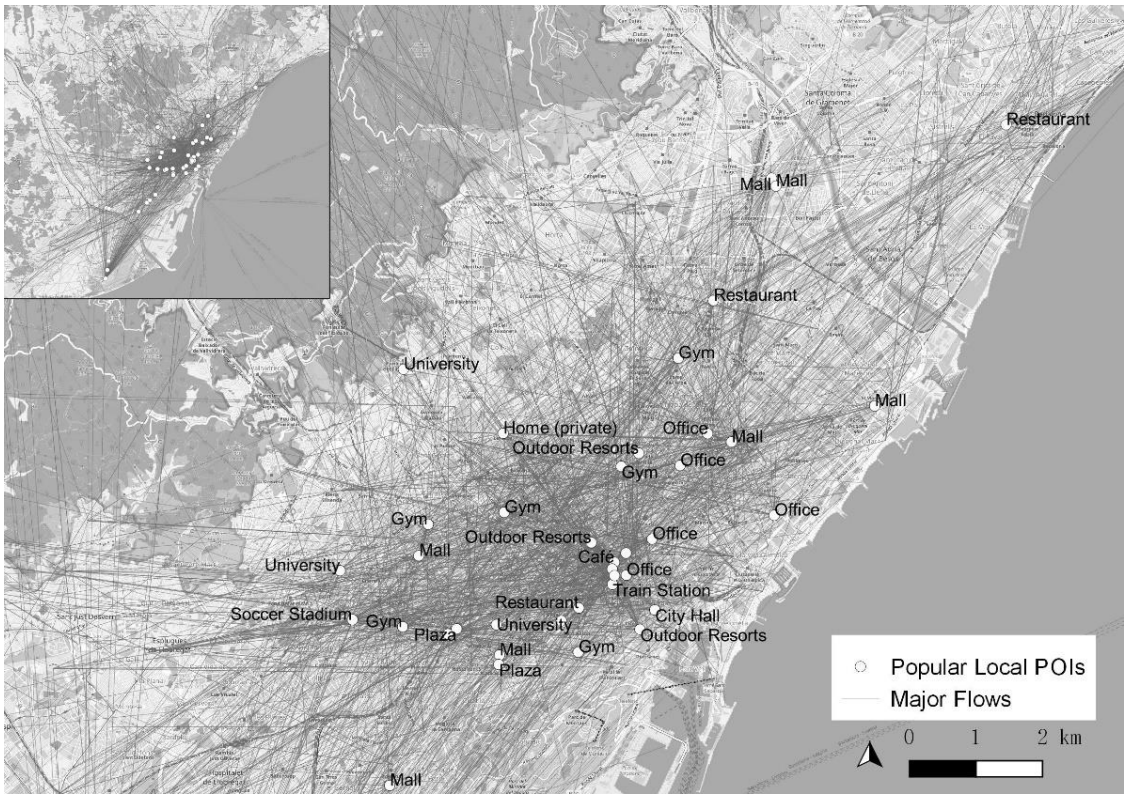
Table 37 Number of visitors to major attractions in Barcelona

Place	2012	2013	2014	2015
La Sagrada Família	3,233,526	3,176,970	3,260,880	3,722,540
Park Güell	-	-	2,598,732	2,761,436
Museum FC Barcelona (Camp Nou)	1,540,648	1,506,022	1,530,484	1,785,903
Aquarium of Barcelona	1,647,163	1,718,380	1,590,420	1,549,480
El Born Centre cultural	-	675,726	1,894,400	1,486,228
Poble Espanyol de Montjuïc	1,223,875	1,258,645	1,236,664	1,221,647
Museum of Picasso	948,869	915,226	919,814	1,008,125

Zoo Park of Barcelona	1,080,187	1,070,104	1,057,188	1,004,069
Casa Batlló	-	796,301	930,000	992,126
La Pedrera (Casa Milà)	861,583	944,509	932,356	990,112
History Museum of Barcelona	-	556,730	973,034	916,517
CaixaFòrum Museum of Barcelona	971,101	686,151	775,068	775,020
CosmoCaixa Barcelona	788,176	716,877	739,649	733,778
Castell de Montjuïc	1,159,042	1,072,000	577,639	670,526

Source: http://www.bcn.cat/estadistica/castella/dades/anuaris/anuari14/cap13/C1304010.htm

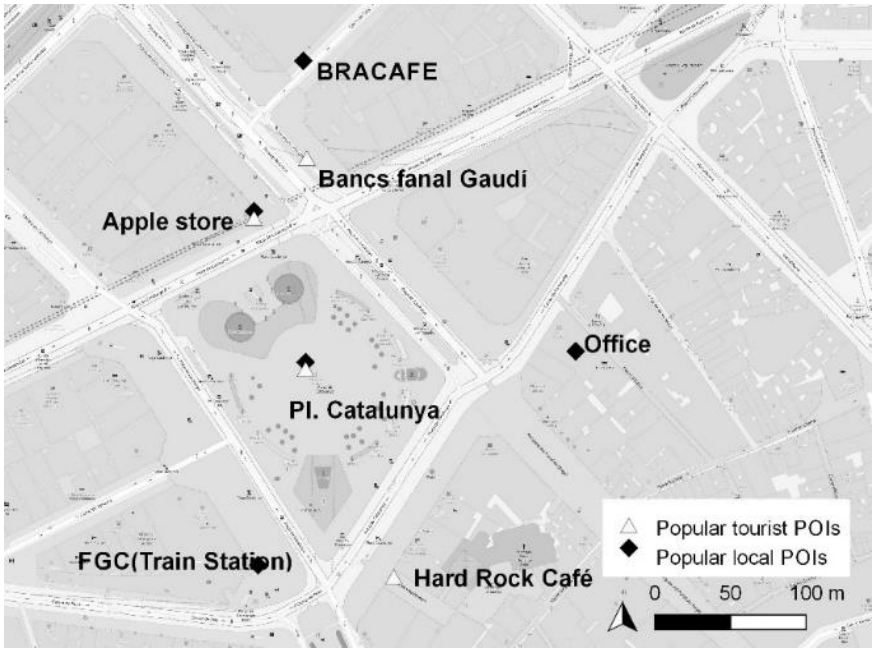
By contrast, the popular POIs of locals are 49 points which belong to diverse categories of usages(**Figure 69**), such as workplace, gym, education place, train station, etc. The hot spots of locals do not show a clear pivotal role in local-flow trajectories, except airport. It is caused by the vast and complex movements of residents. Motivations of local mobility are also more complicated than tourists'.



Source: own elaboration

Figure 69 Major local spatial flows and POIs' check-ins

Although tourist and local activities intersected in the central area, the function of popular POIs still reveals the difference that the two group used the city. For example, Plaça de Catalunya(**Figure 70**)is one of most important transport hub and public space for Barcelona city, and it connects to the historical center of Barcelona. Pl.de Catalunya takes the second rank of tourist check-ins and fifth of local check-ins. The coincident points are the Apple store and Pl. Catalunya. The Apple store is just located around Plaça de Catalunya. It is probably that it belongs to meeting points for tourists and locals. Bancs fanal Gaudí is a tourist attraction. There is a tourist bus stop just in front of the Hard Rock Café, thus it is also a meeting point for tourist. The rest popular POIs of locals are FCG train station, one office building and Coffee Bar, which belongs to typical places of local living.



Source: own elaboration

Figure 70 The distribution of popular POIs around Pl. Catalunya

2.4.2 The frequency of utilization of different usages between locals and tourists

The differences of POI-usages illustrate the variance of utilization of POIs between tourists and locals. The value of difference is equal to the percentage of check-ins of each category in tourist group subtracts the corresponding percentage of local group (**Table 38**).

Hotel and Workplaces have the largest differences between locals and tourists. Transport and Outdoor resorts are positive for tourists, because the intensity that tourists checked-in these places is higher than locals. It is worth to noticing that the conference center is also positive for tourists because Barcelona holds many international conferences every year. According to the Barcelona Convention Bureau, 433 congresses were held in Barcelona in 2017. On the other hand, Education place, Gym and Public services usually belong to typical places of residents, thus the differences are positive for locals. It matches with the results from Carlos Marmolejo-Duarte and Cerda-Troncoso (2012). They pointed out that centers of leisure and education also take a large portion of people's timeline. In general, such differences are matched with their actual living patterns in a city.

Sports center, Store, Shop and Bar account for almost equal percentage in both of groups, because shopping and looking for pleasures are common motivations for locals and tourists. However, the local hot spots of Store and Sports center are distinct from those of tourists. According to the results of spatial distribution above, the shopping places of tourists are close to the tourist attractions, while locals usually go to malls where are far away from the city center. Related to sports centers, except Camp Nou, locals also go to RCDE Stadium for supporting the RCD Espanyol football club.

Table 38 Difference of urban usages between tourists and locals

Usages	Local check-ins	%	Tourist check-ins	%	Difference
Hotel	698	1.15%	1576	8.22%	7.07%
Transport	4724	7.79%	2792	14.56%	6.76%
MAH	907	1.50%	1156	6.03%	4.53%

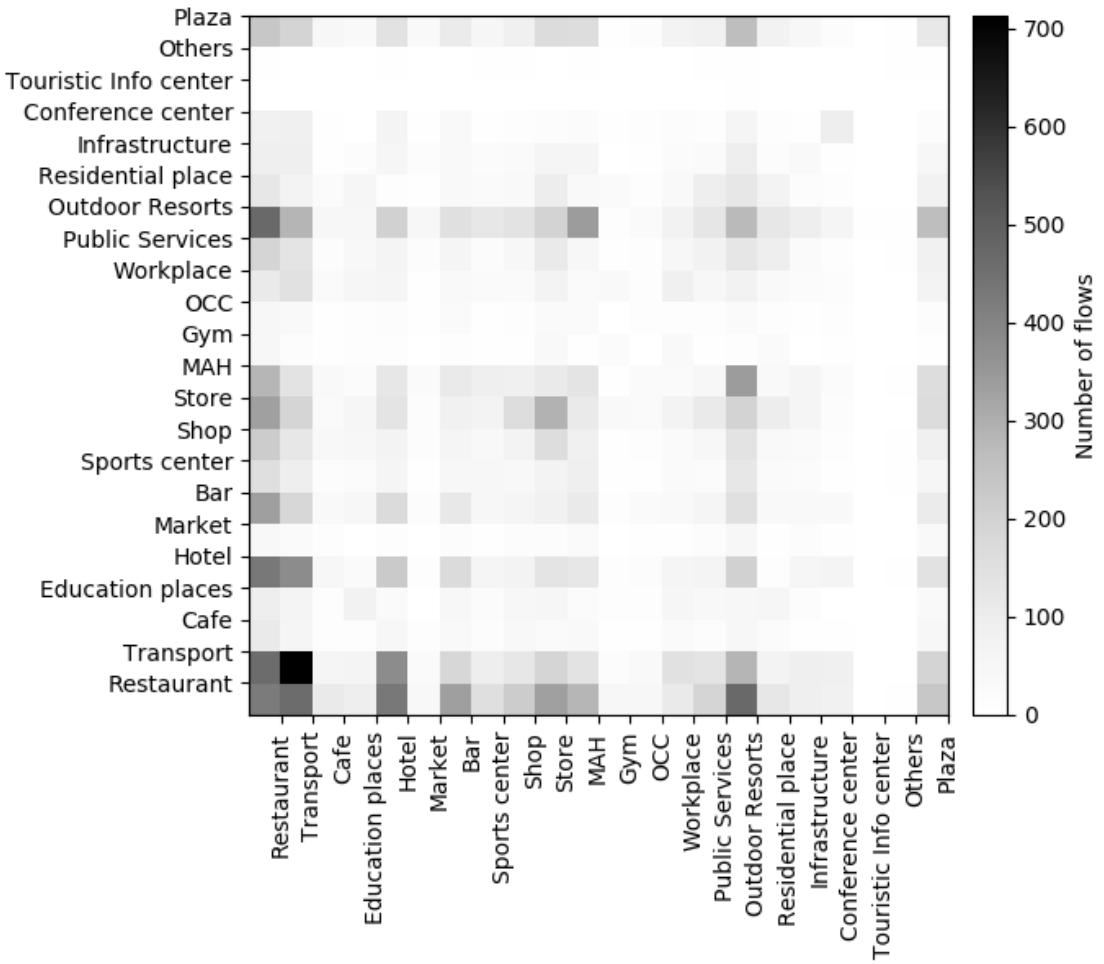
Outdoor resorts	3772	6.22%	1917	9.99%	3.77%
Plaza	1996	3.29%	1162	6.06%	2.77%
Conference center	333	0.55%	412	2.15%	1.60%
Infrastructure	691	1.14%	443	2.31%	1.17%
Sports center	1358	2.24%	581	3.03%	0.79%
Market	238	0.39%	177	0.92%	0.53%
Touristic Info center	3	0.00%	4	0.02%	0.02%
Store	4428	7.30%	1398	7.29%	-0.02%
Others	157	0.26%	24	0.13%	-0.13%
Bar	3398	5.61%	1031	5.38%	-0.23%
Shop	2804	4.63%	789	4.11%	-0.51%
Cáfe	1763	2.91%	330	1.72%	-1.19%
OCC	1367	2.26%	186	0.97%	-1.29%
Restaurant	9533	15.73%	2726	14.21%	-1.51%
Residential place	3237	5.34%	563	2.94%	-2.40%
Public services	4234	6.98%	746	3.89%	-3.10%
Gym	3161	5.21%	154	0.80%	-4.41%
Education places	4982	8.22%	428	2.23%	-5.99%
Workplace	6834	11.27%	585	3.05%	-8.22%

Source: own elaboration

2.4.3 The matrix of usage flows

The heat map displays the matrix of bilateral usage-flows, which presents the intensity of connections between different categories of POIs. Each cell represents the number of direct connections between two categories or within the same category.

Among tourist flows (**Figure 71**), Restaurant-Transport, Restaurant-Hotel, Restaurant-Outdoor resorts have the highest flows between each other, which are over 400. Transport has highest flows with itself. This result fits with the spatial distribution of popular tourist POIs. Those higher flows of usages, such as Transport, Tourist attractions, also present on the popular venues. The flow of Tourist Info Center is sparse, because people usually would check-in at a special or interesting place, rather than an information office.

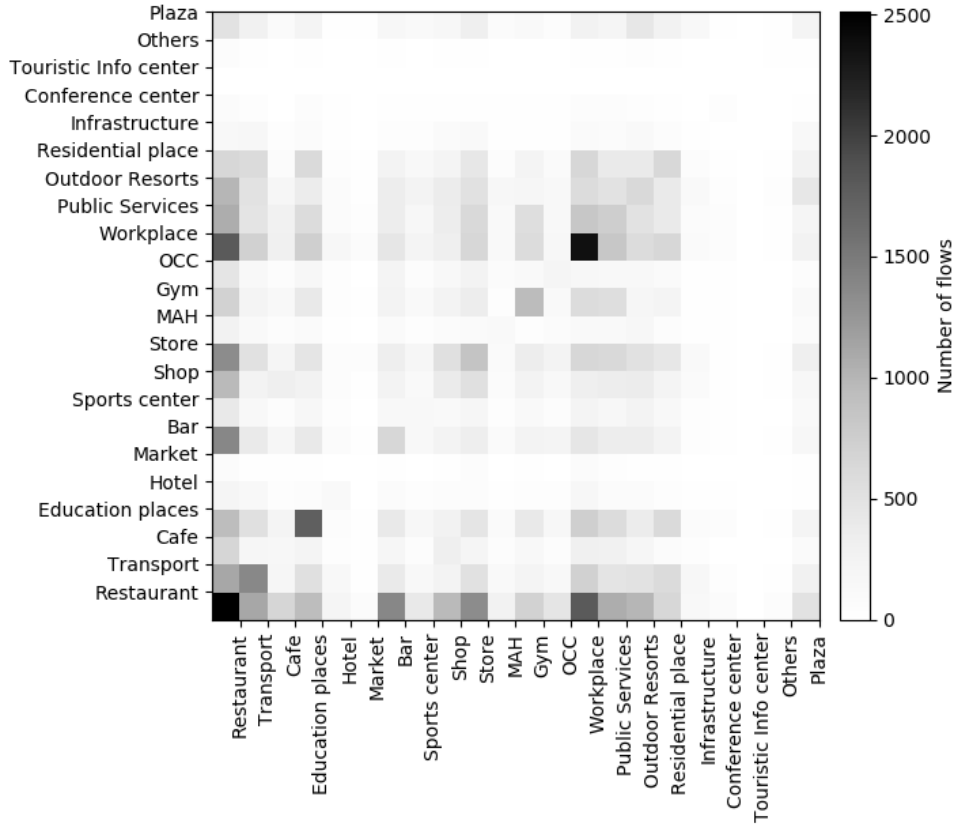


Source: own elaboration

Figure 71 Heat map of tourist usage-flows matrix

This result fits with the spatial distribution of check-ins, as well as it uncovers some relationships among different categories. First, those higher flows of usages, such as Transport, Tourist attractions and Plaza, were also present in popular venues. Meanwhile, as **Figure 68** shows, those POIs of restaurants mainly concentrate along with Passeig de Gràcia street and La Rambra street which are pedestrian streets and gather many tourist attractions. Secondly, it shows the typical usage of tourist activities. Restaurant, Transport, Hotel and Outdoor Resorts have more intense connections with each other than with non-tourist locations such as Gyms, Opera, Concert, Cinema (OCC) and Workplace.

The heat map of local flows (**Figure 72**) shows that the densest paired flows are Workplace-Workplace and Restaurant-Restaurant. Restaurant has stronger flows with other categories, such as Education places, Workplace, Bar, Store, etc. The distribution of flows partly clarifies local users' movement between POIs which seems chaos on the spatial map of flows(**Figure 69**). Since restaurants widely spread in the city and are used frequently, it is understandable that the trajectories of residents seem to be mess.



Source: own elaboration

Figure 72 Heat map of tourist usage-flows matrix

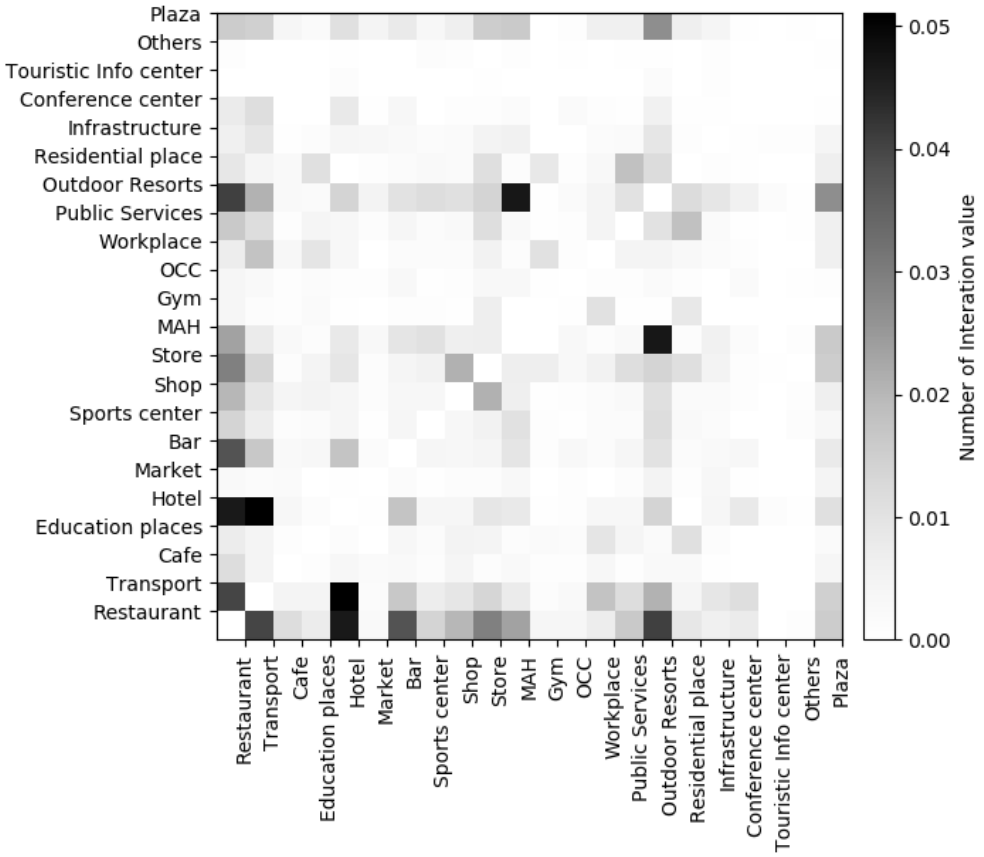
In general, the flow matrix presents the typical usages of both groups. Restaurant, Transport, Hotel and Outdoor Resorts have more intense connections with each other in tourist flows. On contrary, typical local usages, such as Workplace, Gym and Education places, present evident local flows passed through these places.

2.4.4 Interaction Values

Similarly, the bilateral interaction values are visualized via a heat map. The highest interaction value of tourist usages(

Source: own elaboration

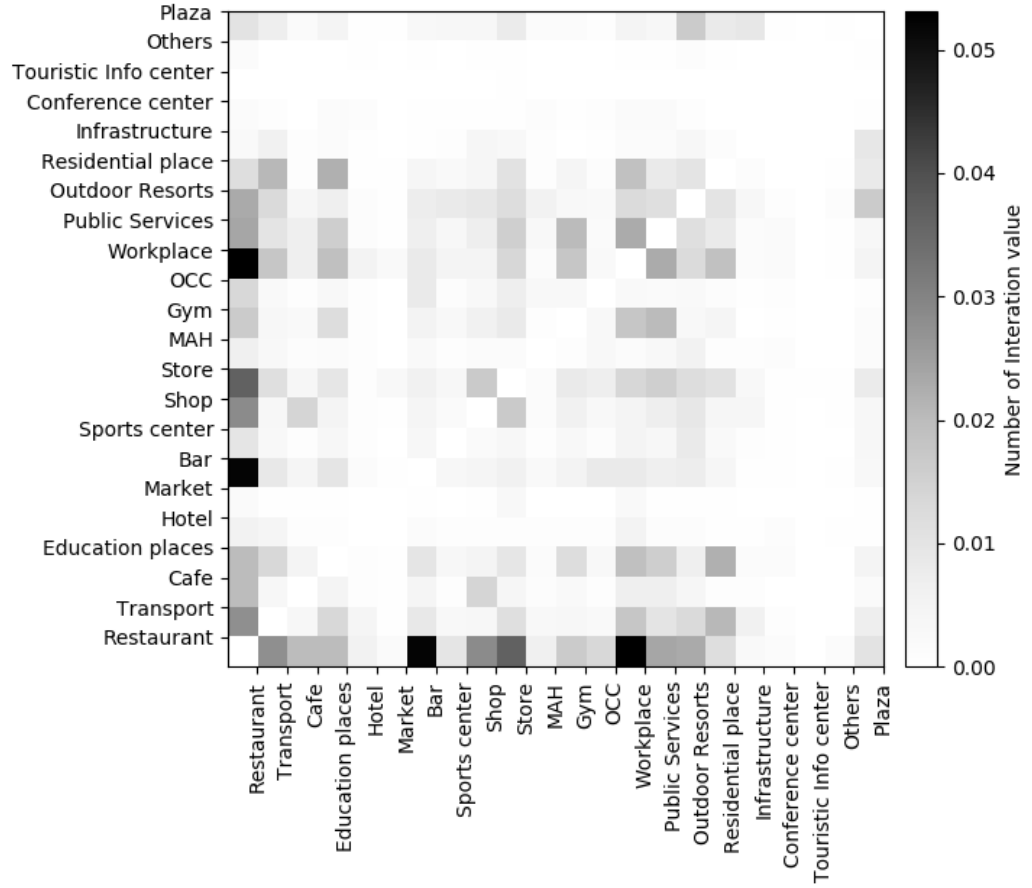
Figure 73)is between Hotel and Transport, followed by Outdoor Resorts-MAH, and Restaurant-Hotel. This result indicates that tourists tend to move directly between Transport-Hotel more than other categories. The number of movements between Museum, Arts and Historical place (MAH) and Outdoor resorts is also higher than between other categories.



Source: own elaboration

Figure 73 Matrix of interaction values among tourist usages

As to local usages(**Figure 74**), Restaurant – Workplace and Restaurant – Bar shows intense interactions. It seems understandable because the relationship indicates the lifestyle of local people.



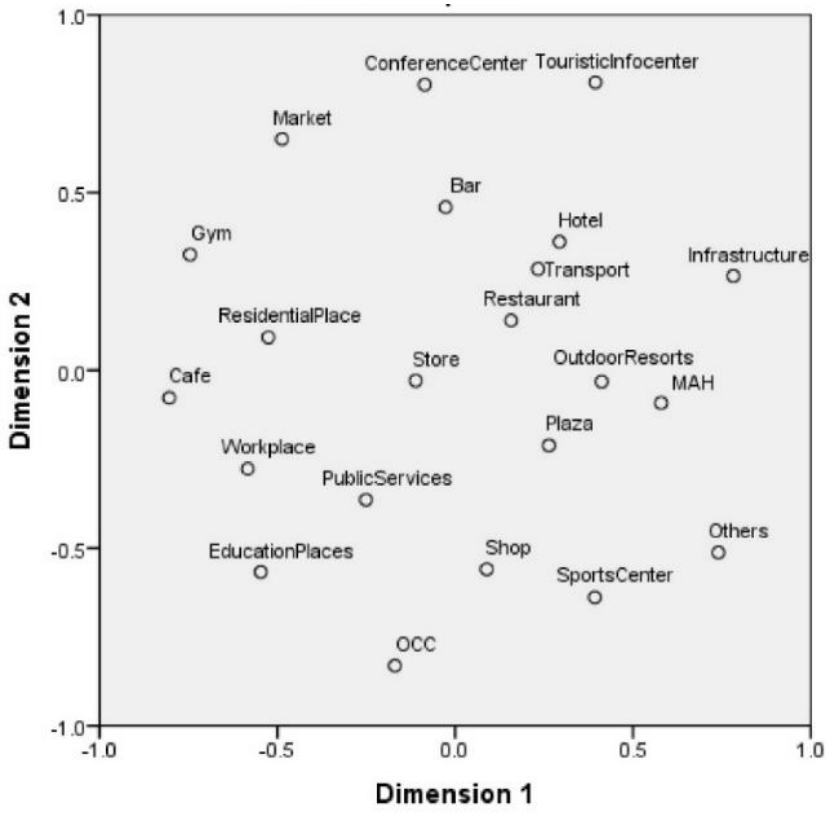
Source: own elaboration

Figure 74 Matrix of interaction values among local usages

Next, we use the PROXCAL-MDS to reduce the original number of dimensions (i.e. 22^{22}) to only two dimensions (Regarding tourist usages(**Figure 75**), it is clear that Hotel, Transport and Restaurant have the closest functional interaction with each other, while Outdoor Resorts, MAH and Plaza are closer to each other. Both of these groups are located at the central part of the graph, indicating that they dominate the tourist flows. The usages located at the peripheral positions, such as Educational places, Gym, and Workplace, have minimal relations with other usages. This may imply that inside our tourist sample

there are different kind of temporal visitors to the city: the first clearly attracted by heritage, cultural and leisure venues; the second, more linked to places intensively used by local population.

,). MDS is a fast way to visualize the level of similarity of objects. In our case, the similarity is understood as the strength of interaction values between usages. The closer is one usage to other, the stronger their relationship.



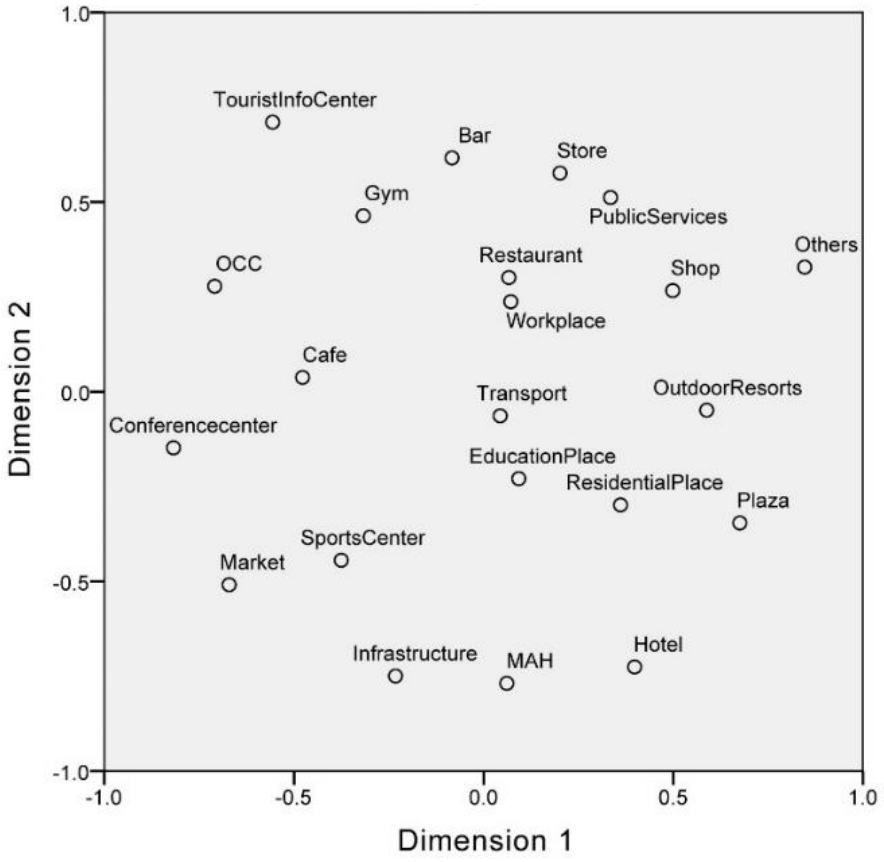
Source: own elaboration

Figure 75 Proxcal plot of interaction value matrix (tourist group)

Regarding tourist usages(**Figure 75**), it is clear that Hotel, Transport and Restaurant have the closest functional interaction with each other, while Outdoor Resorts, MAH and Plaza are closer to each other. Both of these groups are located

at the central part of the graph, indicating that they dominate the tourist flows. The usages located at the peripheral positions, such as Educational places, Gym, and Workplace, have minimal relations with other usages. This may imply that inside our tourist sample there are different kind of temporal visitors to the city: the first clearly attracted by heritage, cultural and leisure venues; the second, more linked to places intensively used by local population.

By contrast, the plot of local group reveals that Restaurant- Workplace has the closest bilateral interaction (**Figure 76**). Transport -Educational place and Educational place –Residential place are closer to each other.



Source: own elaboration

Figure 76 Proxcal plot of interaction value matrix(local group)

For better understanding the interaction, **Table 39** lists the prominent interaction values (**Formula 6**) of each group. There are 25 pairs of tourist usages and 28 pairs of local usages are qualified respectively. In general, Restaurant as

one category of usages, plays an important position in the functional relations of both groups, especially in local usages; though few restaurants are on the list of popular POIs.

The highest interaction value of tourist group is between Hotel and Transport, followed by Outdoor resorts-MAH, and Restaurant-Hotel. The number of movements between Museum, Arts and Historical place (MAH) and Outdoor resorts is also higher than between other categories.

In the local group, the Restaurant- Workplace pair takes the first rank of interaction values, followed by six pairs of usages which are all related with Restaurant. Education place, Transport and Gym are also shown on the list of prominent interaction values.

Table 39 Prominent interaction values

Tourists			**Locals**		
Type	Type	Interaction Value	Type	Type	Interaction Value
Transport	Hotel	0.051061	Restaurant	Workplace	0.053080
MAH	Outdoor Resorts	0.046973	Restaurant	Bar	0.052096
Restaurant	Hotel	0.046562	Restaurant	Store	0.036538
Restaurant	Outdoor Resorts	0.040870	Restaurant	Shop	0.028454
Restaurant	Transport	0.039965	Restaurant	Transport	0.027591
Restaurant	Bar	0.038003	Restaurant	Public services	0.023833
Restaurant	Store	0.029328	Restaurant	Outdoor resort	0.023204
Outdoor resorts	Plaza	0.026766	Workplace	Public services	0.022989
Restaurant	MAH	0.023513	Education place	Residential place	0.022143
Shop	Store	0.021034	Transport	Residential place	0.020567
Transport	Outdoor Resorts	0.020784	Gym	Public services	0.020234

Restaurant	Shop	0.019955	Restaurant	Education places	0.020052
Public services	Residential place	0.017956	Restaurant	Cafe	0.019939
Transport	Workplace	0.017750	Education place	Workplace	0.019181
Hotel	Bar	0.017413	Workplace	Residential place	0.018763
Transport	Bar	0.016533	Transport	Workplace	0.017773
Restaurant	Public services	0.016266	Gym	Workplace	0.017726
MAH	Plaza	0.016008	Shop	Store	0.016589
Restaurant	Plaza	0.015661	Outdoor resort	Plaza	0.016579
Store	Plaza	0.015453	Restaurant	Gym	0.016414
Transport	Plaza	0.014737	Education place	Public services	0.015775
Restaurant	Sports center	0.013810	Store	Public services	0.015717
Hotel	Outdoor Resorts	0.013807	Cafe	Shop	0.014227
Store	Outdoor Resorts	0.013800	Store	Workplace	0.013711
Transport	Store	0.013465	Restaurant	OCC	0.013461
			Transport	Education places	0.013396
			Transport	Outdoor resort	0.012792
			Workplace	Outdoor resort	0.012553

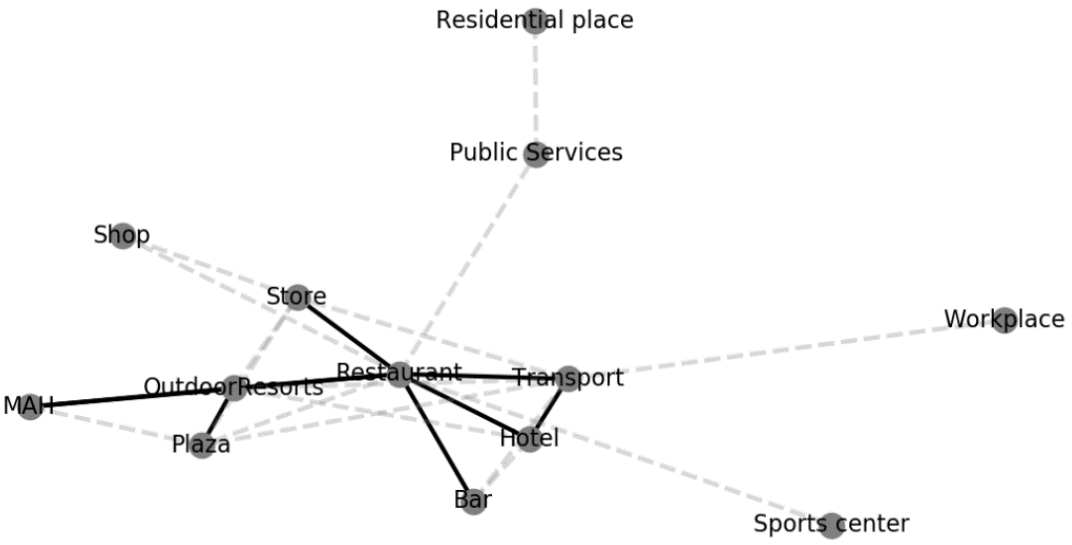
Source: own elaboration

Compared with the matrix of flows, the matrix of interaction values clearly shows the functional proximity of connections between two categories of usages, without the distortion of the scale of flows. For example, as to tourist group, Outdoor Resorts-MAH does not have a high volume of flows(**Figure 72**), but their interaction value is at the second-highest rank of interaction values (**Table 39**). Residential place and Workplace are also on the list; however, they are only connecting with Services and Transport separately. It is possible that Residential appears on the list due to trips to visit friends and relatives, or due to an error in the classification of tourists. According to the official survey, the motivation of travel for business accounted for 40% of all visitors in 2014. Thus, it is reasonable that Transport-Workplace is shown on the list.

To depict the main relations in a simple way, **Figure 77** and **Figure 78** exhibit the prominent interaction values. The nodes are the category of usages and edges are the interaction values. The graph is visualized by Networkx program using Fruchterman-Reingold force-directed algorithm(Fruchterman & Reingold, 1991). Because the algorithm of visualization is aimed to reduce the crossing edges as few as possible, the position of nodes and the distance between them do not have specific meanings. However, the nodes with less edges tend to be placed at the periphery.

The more edges one node has, the higher functional centrality it has. The value is represented by different colors. The black lines represent values equal or above the mean value of all prominent interaction values, and the grey dashed lines represent values below the mean value.

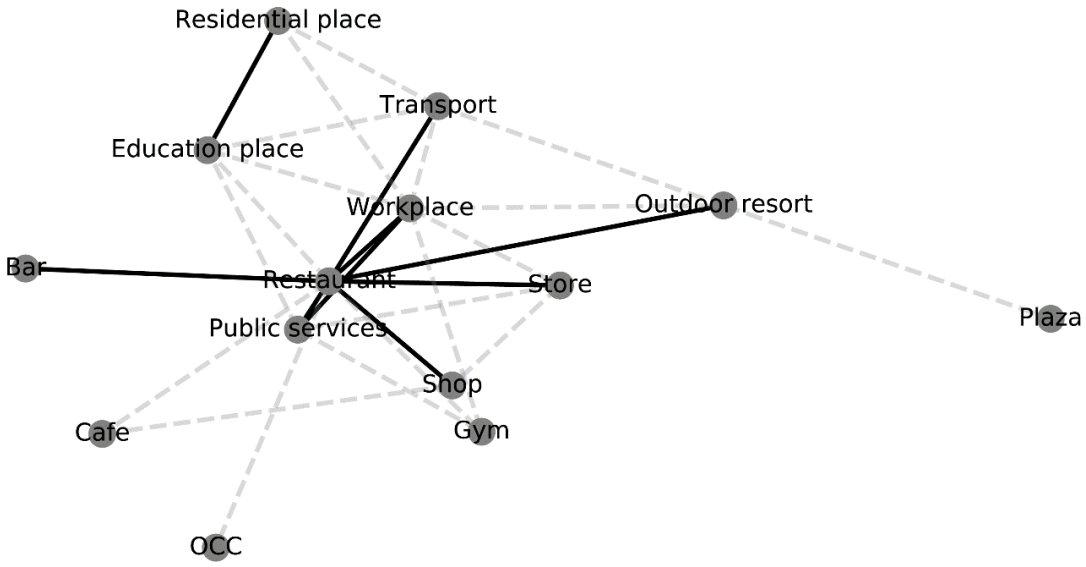
It is clearly show that "Restaurant" has the largest number of connections with other categories in both groups, i.e. it is the central vertex in both networks. **Figure 77** embodies the basic tourist activities: eating, travelling, visiting, shopping and getting accommodation. Those categories that do not correspond to typical tourist places are at periphery. For tourists, Transport, Hotel and Outdoor resorts also have many edges with other usages. They also belong to the core tourist usages according to the practical experience.



Source: own elaboration

Figure 77 Graph of paired usages with prominent interaction values (tourist group)

On the graph of local usages(**Figure 78**), Workplace has the second largest number of edges, followed by Transport, Education place and Public services. These places are typical usages of locals. According to the study of Daily mobility survey of Catalonia (Carlos Marmolejo-Duarte & Cerda-Troncoso, 2012), "home-work-home", "home-shopping-home" and "home-study-home" accounted for 34.5%, 11.6% and 9.6% of all the chains of local activities during weekdays, respectively. All these activities are also presented on the graph. However, residential place does not take the central place in the network. One possible explanation is that the issue of self-representation because people may feel boring to check-in at home every day (Lindqvist et al., 2011) or they want to check-in at places where they want to share with other people(Cramer, Rost, & Holmquist, 2011).



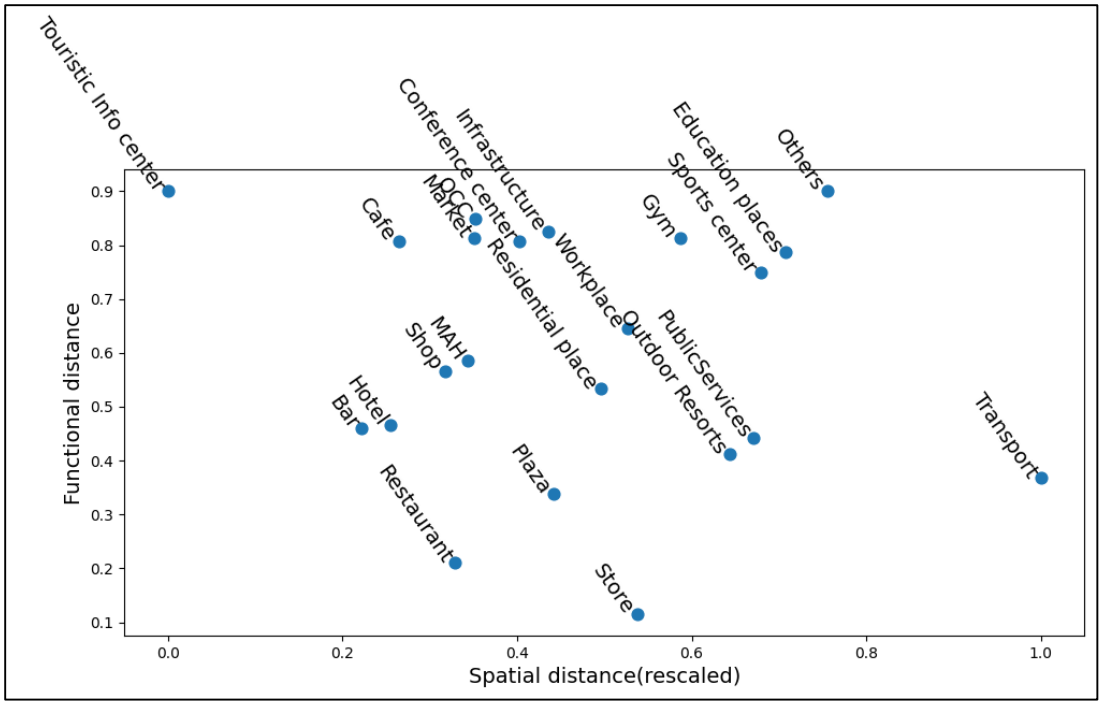
Source: own elaboration

Figure 78 Graph of paired usages with prominent interaction values (local group)

2.4.5 Comparison between spatial proximity and functional proximity

For outlining the difference between spatial relationship and functional relationship, we plot the functional distance(Y-axis) based on PROXCAL graph and the spatial distance of each type of POIs to the median center(**Figure 79**,**Figure 80**). For easier reading, the average spatial distance is rescaled by min-max normalization.

As to the tourist group, **Figure 79** clearly shows that the Transport-POIs has the largest average spatial distance to the median center, but it has highly importance in tourist-functional relations. Hotel - Transport has the highest interaction value of all paired categories. Tourist information center is nearest to the median center, but its functional distance is distant due to few check-ins was generated in here. One possible explanation is that tourists would not like to check-in at places which only provide practical information.



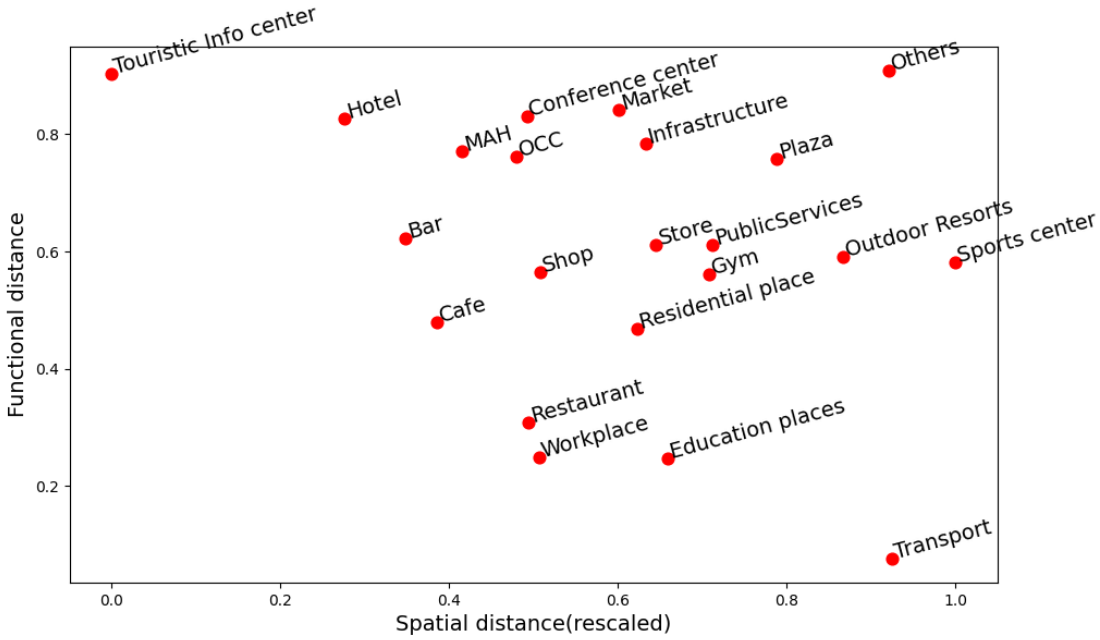
Source: own elaboration

Figure 79 Comparison between functional proximity and spatial proximity(tourist group)

Secondly, some important tourist attractions, such as La Sagrada Família, Park Güell and Camp Nou, are at "peripheral" area of tourist activities. However, as **Figure 68** shows, these tourist attractions play a pivot role among spatial flows of tourists, thus, their functional importance are highlighted. Although it seems that Public services is at the similar position with Outdoor Resorts, the functional distance of the two categories is at the opposite direction (**Figure 75**). It indicates that the spatial proximity does not always take a dominant role. Similarly, Residential places is closer to the PROXCAL center, but is far away from POIs of Outdoor Resorts, Transport etc.

Regarding the comparison of the local group, the spatial distance is farther away than the tourist group in general. Clearly, touristic info center and other lesuerment services are farther than quotidian POIs in terms of functional distance. Restaurant, Workplace, Educationa places and residential places are comprartively close to each other, in both mesures of functional and spatial distance. Naturally, Transport POIs as hubs of city mobility, it does have the

strongest relations with the rest of POIs. Meanwhile, the fartheest distance of Tranport POIs means that these POIs are speading to a larger area of the city.



Source: own elaboration

Figure 80 Comparison between functional proximity and spatial proximity(local group)

2.5. Discussion and conclusions

In summary, this study analyzes the difference of urban usages between tourists and locals from the functional relations of places, meanwhile presents the spatial flows of the two group. Compared with using diversity degree to measure the different urban usages of tourists and locals, the interaction value provides a more comprehensive way to observe how different group of people use the city. It unfolds the functional relations between places and the degree of closeness

between different usages. Therefore, the core usages of locals and tourists can be delimitated and those non-core usages are peripheralized.

The main contribution of the research is that it explores the functional relations of different usages of Foursquare POIs using a quantitative method. It also provides a functional view to analyze how locals and tourists use the city. The differences between locals and tourists not only embody the spatial movement, but also the functional relations of urban places.

Secondly, this study confirms that the functional centroids of activities differ from the centroids of the spatial distribution. The spatial distribution of tourist activities is concentrated in the airport and the central area of Barcelona city where most of the landmarks are located. The spatial relations of tourist activities, in essence, are decided by the locations of tourist attractions. Tourists by their nature tend to make check-ins around tourist attractions. However, their core functional usages are led by daily habits. Eating out is one of the most important activities, and thus POIs in the Restaurant category become the hinge of all functions.

However, local activities are far more complex due to the variety of motivations. The range of tourist activities are limited within the central area of Barcelona where most of the landmarks are located. The range of local activities extended to the whole Barcelona. Moreover, the spatial distribution of our research also coincides with the pattern of distribution in some other cities, as described in the literature review. Although tourists and locals share the space of city center, the different popular POIs indicate that they have difference on utilization of the city.

In terms of functional relations, the differences of POI usages reflect their typical urban usages. Hotel and Workplace present the largest difference on the percentage of check-ins. Secondly, the hinge of functional relation differs from the spatial centroids. As we analyzed above, tourist attractions construct centroids of spatial tourist flows; meanwhile, the spatial centroids of local activities are not easy to identified, due to the massive local trajectories. By contrast, the core functional usages by nature are led by daily habits of each group. For instance,

eating out is one of the most important activities, and thus POIs in the Restaurant category become the hinge of all functions.

Indeed, the function of Foursquare application causes a high proportion check-in at restaurants, which is also appeared in other studies(Abbasi & Alesheikh, 2018; Preoţiuc-Pietro & Cohn, 2013). However, the average number of check-ins at a restaurant is only 3.69, which is the antepenultimate position on the ordination of the average check-ins of all categories. Hence, such bias does not affect the general results in this study.

Thirdly, the tourist-functional proximity of POIs highlights the major nodes of places with tourism functions. As **Figure 77** shows, the chain of functional places appeared when we raised the threshold for interaction values. The categories of Restaurant, Transport, Outdoor Resort, Hotel, etc. are the places of basic functions for tourists. Moreover, such functional proximities are distinct from the spatial proximities of POIs. The closer spatial proximity between two categories of POIs does not lead to closer functional proximity with each other.

Moreover, the interaction values of these categories highlight the major nodes of places with tourist or local functions. The categories of Restaurant, Transport, Outdoor Resort, Hotel, etc. are the places with basic tourism functions. Restaurant, Workplace, Transport, Education place, Gym, etc. undertake the basic functions of daily life of residents. It proves that the functional closeness of urban places can be quantified by interaction values.

Furthermore, the interaction values depict the main patterns of users' movement. For example, Outdoor Resorts has higher interaction value with Museums, Arts and Historical places. It indicates that tourists tend to travel from one tourist attraction to another directly. Transport-Hotel has the highest interaction value because people usually need to drop their luggage at the hotel when they arrive and carry their luggage from the hotel to transport hubs as they leave. On the local graph, Education place has four prominent interactions with other usages, however, the one with Residential place are the highest among them.

It accords with the situation that most of students live at home and take home-made lunch to school or university.

Indeed, it is undeniable that Foursquare data has a bias. Because the function of Foursquare is to provide users with practical information about places, a high proportion check-ins are in the category of restaurants. However, the average number of check-ins at a restaurant is only 0.82, while the mean number of check-ins at Outdoor resorts and MAHs is much higher. Hence, such bias does not affect the general results in this study. Moreover, the result reflects the typical land uses of tourists in a city, such as hotel, transport and tourist attractions; these categories have a higher intensity of flows. Essentially, Foursquare POIs are able to reflect tourist activities to some degree.

In summary, this paper reveals how POIs functionally interact with each other. The method of evaluation of tourist -functional relations is possible to spread to other groups of people, e.g. a comparison between locals and tourists based on the same functional analysis. Moreover, the spatial distribution of our research also coincides with the pattern of tourist distribution in some other cities, as described in the literature review. This result is similar to the conclusion of Béjar et al. (2016) which utilizes Instagram data, a finding that indicates different datasets may be comparable, as long as the volume of the dataset is substantial.

It should be noted that the defect of data itself limits the scope of the present study. The meta-database that we received eliminated the personal profiles due to the potential privacy issues. Therefore, it lacks any analysis of the background of tourists due the potential privacy issues. For example, we cannot utilize the user's profile to identify their home country or to discuss whether the Foursquare behavior of foreign visitors differs from that of domestic visitors. Secondly, with the decline of popularity of Foursquare, the availability of its data has shrunk in Barcelona. The representative nature of more recent Foursquare data is questionable. The use of new data sources from currently popular LBSNs, such as Instagram and Twitter, will be necessary to examine and understand the latest dynamics of tourists. Last but not the least, it is worth to noting that the availability of Wi-Fi would impact the density of check-ins. Despite of most of the public

spaces do have free Internet Access that provided by the City Council and Private Firms, there is a small risk of overrepresentation in best-serviced premises such as the Airport.

VI.3.Case study III

Quantifying the relationship between public sentiment and urban environment in Barcelona

3.1. Introduction

As Lewis Mumford pointed out, "the city… is the point of maximum concentration for the power and culture of a community"(Mumford, 1970). In this sense, the city is the production of the intricate relationship between the environment and human activities. Such relationships could be presented as routes of people's mobility, urban functional areas, and employment centers. It also could be measured by the public sentiment and perception toward the city. The public sentiment refers to a congregation of people's feelings and perceptions, which is "an attitude which is based on their thoughts and feelings", according to the Collins English dictionary. For example, urbanisation has coincided with a rise of mental disorders due to the polluted environment, crime violence and decreased social support (Gruebner et al., 2017; Srivastava, 2009). However, the exposure to urban parks and green areas is beneficial to reduce people's negative feeling (Schwartz et al., 2019). Therefore, the importance of understanding the public sentiment lies in that it can reveal the quality of an urban environment, which is crucial for informing public policies and creating a better city and quality of life.

In current, Location-based social network (LBSN) data, such as Twitter or Facebook, provides a tangible vision to present those "invisible" public sentiments. The common feeling is possible to be observed and aggregated through social media data. Sentiment analysis via Location-based social network (LBSN) data has been a popular topic in urban studies, such as work stress(W. Wang et al., 2016), the sentiment of railway passengers (Collins et al., 2013), and mapping sentiment (M. Li, Ch'ng, Chong, & See, 2016).

However, previous studies related with the urban environment and public sentiment mainly focus on a single topic of land-use, such as tourist attractions(Barbagallo, Bruni, Francalanci, & Giacomazzi, 2012; Park et al., 2018) or green spaces (Chapman et al., 2018; Schwartz et al., 2019). Another perspective is to detect the relationship between sociodemographic characteristics (e.g. income, gender, and unemployment) and the public sentiment (Ballas, 2013; Blanchflower & Oswald, 2008; Helliwell, 2003). Fewer research investigates the relationship between the whole urban environment and public sentiments. Moreover, most research has been limited to English texts or a single language due to the studied area or the technical problems of analysing different languages. In fact, immigrants and visitors usually hold an important portion in international metropolises. The analysis based on single language is not sufficient to reveal perceptions about the same city from people who use other languages.

Therefore, in order to address these gaps, the intention of this research is to explore the relationship between the urban environment and public sentiment extracted from Twitter. The urban environment is conceived of a polymer that contains four layers: sociodemographic, built-environment, and human mobility and socioeconomic activities. It utilises thirty months of Twitter data to analyse the public sentiments in Barcelona. Specifically, English and Spanish language are involved in the sentiment analysis. Computational sentiment analysis is employed to quantify and classify the sentiment orientation of tweets, *i.e.*: positive, negative, and neutral. After exploring whether it was feasible to model at a single tweet level, we explored the relationship between urban environment indicators and the variation of public sentiments based on the division of the basic statistical area (AEB) of Barcelona. The public sentiment is measured by two aspects: the spatial density of sentimental tweets (*i.e.* tweets with positive and negative sentiment) and variation of sentiment scores. Further, a multivariate regression model is used

to reveal the variation of Twitter sentiment across different AEBs with different urban environment characteristics.

The results demonstrate that the variation of public sentiments could be partially understood by some urban indicators in Barcelona; though the fluctuation of Twitter sentiment score is not strongly correlated with the indicators of the urban environment. Firstly, the distribution of sentimental tweets tends to concentrate on places that usually maintain a high intensity of human activities. Secondly, the wealthier areas show a more positive correlation with a higher public sentiment. The result of Pearson's correlation shows that the public sentiment score is positively correlated with places with a higher socioeconomic level. The model statistically confirms that the density of storefronts and the percentage of commercial shops has a positive impact on the sentiment score. Conversely, the density of the population, the lower wages, and the area of railway system influence the public sentiment negatively. Thirdly, the disruptive events are directly associated with the public negative sentiments, which can be observed by the impact of Barcelona terrorist attack in 2017. In summary, this research has unequivocal implications in analysing the interaction between public sentiment and the urban environment and enriching the governance of a Smart City.

The remainder of the research is organised as follows: Section 2 reviews the issue of representativeness of Twitter and previous research on detection of urban sentiment; Section 3 introduces the methodology and presents the result of Twitter sentiment classification; Section 4 analyses the interrelationship between urban environment and public sentiment via correlation analysis and regression models; Section 5 summarises the research and discusses the potential directions for future work.

3.2. Literature review

3.2.1 Geographical studies of urban sentiments

Mapping urban emotion could date back to the 1960s. Lynch (1960)brought up the concept of "mental map" to represent people's perception of their build

environment. Based on the framework, (Kuipers, 1978) elaborated a theoretical model to state a person's cognitive map (i.e. how people store their spatial surroundings in their mind). However, the cognitive centered conception was challenged by an interactive theory in the 1990s. Space is a material that the body engages and works with(Lupton, 1998), rather than an objective existence solely. Therefore, "emotions can be conceptualized as the felt and sensed reaction that arise in the midst of the (inter) corporal exchange between self and world" (Davidson, Smith, & Bondi, 2012). In other words, the spatial environment influences people's emotion to some degree. Although it is hard to measure the individual's emotional reaction to a spatial place, the common feeling is possible to be observed and aggregated. For example, Molz (2005)analyzed travelers' emotional responses when they were eating at McDonald's. It extracted contents from forty websites when these travelers were traveling around the world. The result showed that McDonald's evoked travelers' emotion of familiarity that mixed contentment and contempt.

Personal questionnaire is the traditional way to understand the relationship between spatial place and people's feeling. Matei Matei, Ball-Rokeach, and Qiu (2001) visualized a first digital map of emotion in Los Angeles, USA. They aggregated 215 participants' perception about the residential communities into a map of fear feeling. Obviously, such a method has great limit on the number of samples and places. Therefore, the LBSN data, such as Twitter or Facebook, naturally has been concerned by researchers in recent years. In addition to the geo-spatial information, the contents of LBSN data provide a fast access to understand people's opinion and emotion. It can provide valuable information about the work stress(W. Wang et al., 2016), elections(H. Wang, Can, Kazemzadeh, Bar, & Narayanan, 2012), social movement (LeFebvre & Armstrong, 2018), even the stock market Pagolu (Pagolu et al., 2016), etc.

Along with the geographical information, researches also study the relationship between spatial place and mass sentiments. Gallegos et al. (2016) utilized Foursquare data to study the happier places in Los Angeles. It concluded that the happier places tend to be observed in census tracts which have more Foursquare check-ins. Collins et al. (2013) studied the emotion of passengers of

suburban trains near the city of Chicago. They found out that the dissatisfaction to incidents can be measured by social media data. With regard to the emotional reaction in specific places, Padilla et al. (2018) investigated tourists' emotions at tourist destinations via Twitter data in Chicago. Their result showed that seasonal temperature is positively correlated with the positive sentiment in general. Urban parks could also reduce people's negative feelings according to the investigation of Schwartz et al. (2019) in San Francisco because negation words such as 'no', 'not' decreased in frequency during visits to urban parks.

3.2.2 Twitter sentiment analysis

The sentiment classification is derived from psychological theories (Ekman & Cordaro, 2011; Vytal & Hamann, 2010)which study the basic emotions like happiness, fear, sadness, etc. The sentiment classification of "positive - neutral-negative" is derived from basic emotions. Further, language as the most direct expression of emotions, people express their own feelings and evaluate other people's emotions through words. One of the most famous related studies is conducted by Shaver, Schwartz, Kirson, and O'connor (1987)which classified emotional words into six emotion categories: love, anger, joy, sadness, surprise, and fear.

Therefore, it is reasonable to observe the mass sentiments using Twitter texts. The simplest way to measure emotion automatically is to calculate the word-frequency (López-Ornelas & Zaragoza, 2015; Quercia, Ellis, Capra, & Crowcroft, 2012). However, single words usually cannot represent the completed emotional trend of a phase or paragraph. Therefore, various algorithms are developed to solve the problem automatically, such as Naïve Bayes classifier (A. Pak & Paroubek, 2010), Graph based Semi-Supervised Learning (Chapman et al., 2018) , Latent Dirichlet Allocation(LDA) (Kovacs-Györi et al., 2018), etc. It has become an emerging academic field of natural language processing which is aim to let the computer "understand" human language.

As a widely used micro-blog platform, Twitter has become an object and data source of recent academic research. Especially, Twitter sentiment analysis, which has become a significant branch of natural language processing, that aims to understand natural languages through a computer. The analysis of Twitter sentiments can provide valuable information on stock market movements (Pagolu et al., 2016), political election (D. Paul et al., 2017; Yaqub et al., 2020), disaster management (Neppalli et al., 2017), air quality (W. Jiang, Wang, Tsou, & Fu, 2015), and disease surveillance(Sinnenberg et al., 2017).

The essence of sentiment analysis is to investigate the rationale of expression of sentiments in texts and estimate the sentiment orientation (*i.e.* positive, neutral or negative) toward the subject of the text(Nasukawa & Yi, 2003). Therefore, the framework of sentiment analysis includes two major parts: the extraction of emotional expressions and the algorithm that quantifies the sentiment score of texts based on certain rules.

With regard to analysing sentiment from Twitter (unlike formal edited texts) Twitter texts is unstructured, informal and usually include emoticons, hashtags, web links and other symbols. The simplest algorithm is based on word-matching. For example, Larsen et al. (2015) analysed the patterns of emotional expression all over the world through searching and extracting emotional words from their Twitter dataset, which is based on two sentiment lexicons: Linguistic Inquiry and Word Count (LIWC) and Affective Norms for English Words (ANEW). Obviously, the method ignores the complex expression of texts and the interrelationship between words. To compensate for this flaw, corpus-based and lexicon-based approaches are two major methods utilised to conduct sentiment analysis.

Corpus-based sentiment analysis uses specific corpora as the training data to extract the feature of sentiments and classifies the polarity of an input text (M. Li et al., 2016). It measures the similarity of the whole sentence rather than single words. For example, Candelieri and Archetti (2015) utilised Support Vector Machine to classify the tweets using a dataset which was classified by human supervisors. The algorithm identified the polarity of a tweet by comparing the similarity of the features between the tweet and the human-labelled tweet. The feature of a tweet was defined by the frequency of positive and negative terms.

The corpus-based approach can perform well for specialised domains, such as medical articles or political speech, because the source of the corpus points to that domain. However, it lacks sufficient evaluation of the quality of the algorithm and requires model training for each domain.

Lexicon-based method(Gilbert & Hutto, 2014; Mitchell et al., 2013) contains a list of words that are pre-defined in terms of sentiment polarity and sentiment strength. It requires high quality of lexical resources for good performance. Based on the dictionaries, the machine sentiment classifier searches the matched emotion expression in texts and assign sentiment scores to them. For instance, Gilbert and Hutto (2014) established a lexicon using three sentiment lexicons --LIWC, ANEW, and G*eneral Inquirer* (GI), and supplemented the lexicon with commonly used expression in social media data, such as acronyms, slang, and emoticons. They applied heavy human inspection in the production of the lexicon in order to improve the quality. Compared with the corpus-based method, the algorithm of lexicon-based method is easier to understand and adopt.

3.2.3 Sentiment analysis in urban contexts

Using personal questionnaires is the traditional way to understand the relationship between spatial place and people's feelings. Matei et al. (2001) visualised the first digital map of fear emotion in Los Angeles, USA through aggregating 215 participants' perceptions about residential communities. Brereton, Clinch, and Ferreira (2008) explored the relationship between individual well-being and urban environment conditions through collating participants' opinion of life-satisfaction. The result confirmed that specific locations are shown to have a direct impact on life satisfaction. For example, the distance to landfill has a negative correlation with subjective well-being.

Along with geospatial information, the contents of the location-based social network (LBSN) data provide fast access to understand public sentiment. Mitchell et al. (2013) collected tweets across fifty states of the United States to investigate

the degree of happiness among cities and states. The degree of happiness refers to the average happiness score based on the frequency of positive words in tweets. They found that the happiness score is strongly associated with increasing household income. Moreover, Gallegos et al. (2016) concluded that places with more amenities, such as restaurants, gyms and beaches, tended to be happier than other places in Los Angeles. Bertrand et al. (2013) generated a sentiment map of New York City using Twitter. They found that public sentiments generally performed higher in public parks and lower at transportation hubs. Cao et al. (2018) studied the relationship between sentiment score and land uses using a linear mixed-effect model. They concluded that sentiment scores were higher in the commercial and public areas during noon/evening and on weekends. The areas of farmland, transportation and manufactory tended to show negative sentiment at midnight and weekdays.

With regard to specific places, Collins et al. (2013) studied the sentiment of passengers of suburban trains near the city of Chicago using Twitter data. They found out that the dissatisfaction to incidents can be detected by the variation of Twitter sentiment during the 24-hour timeline. Padilla et al. (2018) investigated tourists' sentiments at tourist destinations via Twitter data in Chicago. Their result showed that seasonal temperature is positively correlated with the positive sentiment in general. The winter sentiment was the lowest compared with other three seasons. Urban parks could also reduce people's negative feelings according to the investigation in San Francisco(Schwartz et al., 2019) because negation words such as 'no', 'not' decreased in frequency during visits to urban parks. Larger and greener parks delivered more happiness to people than smaller ones.

In summary, the environment indeed influences urban sentiment/perceptions to some degree, no matter if in terms of an individual level or a level of larger geographical scale. Hitherto, most research only explored either the relationship between Twitter sentiment and urban environment or the correlation between demographic characteristics and the sentiment. Few of them analyse the relationship in terms of a completed perspective which includes factors of socioeconomic, built-environment, and human mobility and socioeconomic activities. Therefore, our research is trying to put all these indicators into one

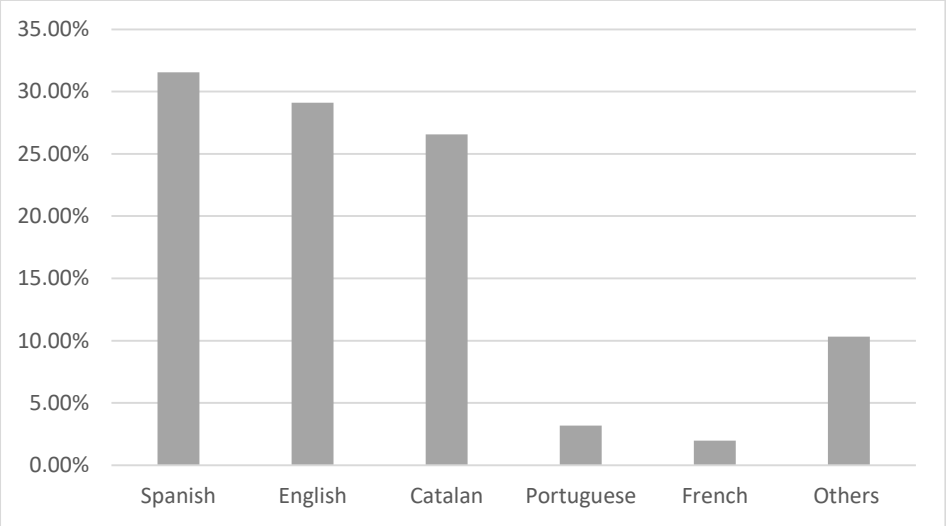
model to discuss the relationship between urban environment and sentiments embedded in tweets.

3.3. Methods and materials

3.3.1 Description of datasets

The study scope is restricted to Barcelona city, Spain. As a world-famous tourism metropolis, according to the *Annual Tourism Sector of Barcelona Report 2017,* the total number of overnight tourists who stayed in hotel accommodation rose to nearly nine million. Meanwhile, Barcelona city is a typical compact city whose area is 101.9 km² and supports 1.6 million residents. The Twitter dataset that we used is over one million tweets, acquired from September 2016 to April 2019. In order to reduce non-active users and improve the quality of data, retweets, and users who only appeared once in the studied area are excluded from the dataset, which account for 44.82% of all 123,437 Twitter users during the period. Meanwhile, non-individual accounts and their tweets, such as weather information, companies, websites, etc. were also removed. The cleaned Twitter dataset contains 707,549 tweets which were generated by 63,178 users.

With regard to languages of cleaned dataset, 47 languages were detected by Twitter automatically. Spanish, English and Catalan account for nearly 80% of all tweets (**Figure 81**). The high percentage of English tweets are caused by the huge volume of visitors. Catalan and Spanish are the official languages of Catalonia. According to the investigation from the Institute of Statistics of Catalonia in 2018, 97.5% of people in Catalonia understood and used Spanish, and 76.1% of them wrote mobile messages in Spanish or combination of Spanish and other languages. Only 11.7% of people used Catalan to write mobile messages solely.



Source: own elaboration

Figure 81 Distribution of tweets' language

The initial proposal was to analyse Twitter sentiments of the three languages. However, after the text cleaning, a human sampling inspection (under 95% confidence level and 4.6 confidence interval) showed that over 50% of 453 samples of Catalan tweets did not contain any valid text except the information of geo-location. It is largely caused by the geolocation that refers to the Catalan local name of places. Firstly, half of these tweets were generated from automatic geolocation of posted pictures or users 'check-ins'. Secondly, the text cleaning removed all hashtags, links, unrecognized characters (emoticons, symbols such as ). Therefore, tweets that only contained these contents would be cleaned except for the geo-information. Since Twitter's algorithm of language detection is not publicly available, we only could estimate that the machine detected them as Catalan tweets because the algorithm considered that hashtags and links could be used for all language users. Therefore, considering the poor quality of Catalan tweets, the sentiment analysis only takes English and Spanish tweets into consideration.

3.3.2 Sentiment analysis

3.2.1 Sentiment classification

The process of sentiment analysis includes two steps. Firstly, as few free open programs support sentiment analysis of multi-languages, Spanish tweets were translated into English via Google Translate. Secondly, the sentiment analysis is conducted by two widely used lexicon-based programs: Vader, that specifically focus on the social media texts (Gilbert & Hutto, 2014); and Sentistrength, that focus on the estimation of sentiments in short informal texts(Lenormand et al., 2014; Thelwall et al., 2010). The intersection of the results of both algorithms is the final classification, *i.e.:* a tweet is confirmed as positive only if it is in the positive category of Vader and SentiStrengh. The classification of neutral and negative tweets follows the same rule. In addition, we also extracted a sample dataset and manually inspected the effect of our method and whether the translation greatly affected the sentiment detection.

Vader rates the sentiment of words based on sentiment orientation and the intensity of the emotion, using a range from -4 (maximum negative) to 4 (maximum positive). For example, the word "great" has a higher positive valence than "good", due to its intensity. The normalised compound score of a sentence, on a scale from -1 to 1, is the sum of scores of each word. Similarly, SentiStrength uses an integral range from- 5 to 5 to indicate the sentiment orientation except for 0. The 1 and -1 represent “no positive emotion” and “no negative emotion” separately. The binary sentence score is given by the maximum value of positive score and negative score. For example, the text “I love you but your room is horrible and nasty.” would be classified as follows, “I love (3) you but your room is horrible (-4) and nasty (-3). <sentence score: 3, -4>.” The sentiment orientation of a sentence is the sum of the two scores. **Table 40** lists thresholds of sentiment classification of both algorithms for a given sentence:

Table 40 Thresholds of sentiment classification

Sentiment	Vader	SentiStrength
Positive	Compound score >= 0.05	Senti_sum > 0
Neutral	-0.05< Compound score < 0.05	Senti_sum = 0
Negative	Compound score <= -0.05	Senti_sum < 0

Source: own elaboration

3.2.2 Human inspection of Spanish-English translation and sentiment classification

It is necessary to emphasise the aims of this inspection: 1) measuring the impact of the translation to the sentiment classification, rather than the quality of translation; 2) verifying the improvement of the mixed use of the two programs. Therefore, we measure the impact of translation by comparing the results of sentiment classifications between the human and the machine. Considering the expense of human evaluation, only tweets that are allocated in different categories of sentiments would be further investigated in their texts. A native Spanish and Catalan speaker with advanced English level was invited to classify the original Spanish tweets into positive, negative, and neutral.

Table 41 Result of Spanish sampling sentiment evaluation

Method of Classification	Total number of agreement tweets	% agreement	Positive tweets	%	Negative tweets	%	Neutral tweets	%
Human (original Spanish tweets)	453	-	204	45.03%	17	3.75%	230	50.77%
Vader	453	-	187	41.28%	38	8.39%	228	50.33%
SentiStrength	453	-	192	42.38%	28	6.18%	233	51.43%
Human + Vader	301	66.45%	127	42.19%	8	2.66%	166	55.15%
Human + SentiStrength	298	65.78%	127	42.62%	8	2.68%	163	54.70%
Vader + SentiStrength	343	75.71%	145	42.27%	17	4.96%	181	52.77%
Human + Vader + SentiStrength	253	73.76%	110	43.48%	6	2.37%	137	54.15%

Source: own elaboration

Table 41 lists all possible combinations of sentiment classifications. The total agreement is defined as the number of tweets that are classified into the same category of sentiment using different methods. The agreement level between the

human and a single machine classification (Vader or SentiStrength) is about 66%. The total agreement of the two software reaches 75%. It is worthwhile to mention that the degree of agreement between human evaluations (made by different persons) is just about 80%(Ogneva, 2010). The total agreement among human, Vader, and Sentistrengh is 55.84%. However, based on the result of the intersection of Vader and Sentistrengh, the agreement increases to 73.76%. It proves that the quality of sentiment classification is improved after intersecting the results.

The difference of each sentiment category between software and human classification is listed in **Table 42**. The biggest difference is between positive and neutral tweets. After investigating the concrete texts of these tweets, we concluded three reasons that led to the unmatched classification. In the group of Neutral (H)-Positive (VS), 75% of these tweets belong to commercial advertisements which only can be detected by manual examination. The advertising tweets are considering as a neutral sentiment. Although the aforementioned cleaning process has already removed many commercial accounts, however, it was hard to identify them when the name of account does not have any characteristics, such as ".com", "studio", and "shop". The problem of understanding the texts, which includes idioms, lack of emotional words, metaphor, satire, and jokes, is the biggest problem of the sentiment classification. In fact, the understanding of texts is the core issue of sentiment analysis, regardless of using human or software methods. However, such an issue is beyond the discussion of our research. The error of translation is specifically defined as the wrong translation of word-to-word or untranslated words. The error of word-translation is the major cause of the different classification between positive and negative. However, the total number of this category is just 1% of all samples. In brief, word-translation does not greatly affect the sentiment analysis.

Table 42 Number and Reasons of different sentiment classification

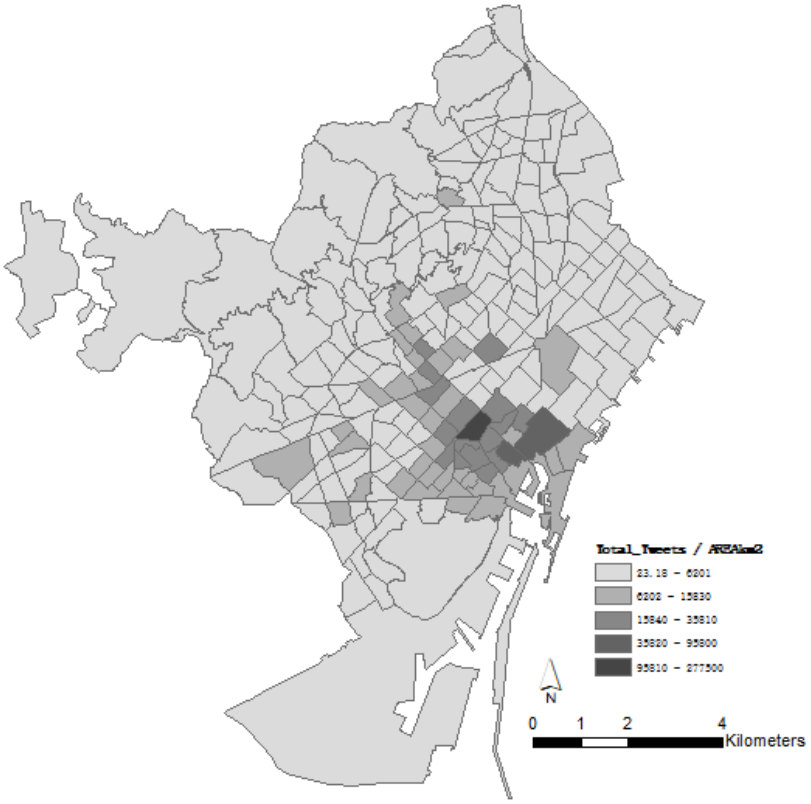
Difference	Number of tweets	Reason of differences
Neutral (VS)-Positive (H)	43	Problem of understanding the meaning of text: 58.14% Commercial advertisement: 30.32% Word-translation error: 6.98%
Neutral (H) – Positive(VS)	32	Commercial advertisement: 75.00% Problem of understanding the meaning of text: 18.75% Word-translation error: 6.25%
Positive (H) – Negative (VS)	5	Problem of understanding the meaning of text: 60% Word-translation error: 40%
Negative (VS) – Neutral (H)	6	Problem of understanding the meaning of text: 100%
Positive (VS) – Negative(H)	3	Problem of understanding the meaning of text: 34% Word-translation error: 66%
Negative (H) – Neutral (VS)	1	Problem of understanding the meaning of text: 100%

Note: H: human, VS: Vader + SentiStength.

Source: own elaboration

3.3.3 Evaluation of the relationship between urban indicators and Twitter sentiment

Initially, the ideal model was to evaluate the influence of urban indicators on Twitter sentiments directly. However, we found that a logit model based on a single tweet level failed to generate reliable outcomes. The neuron network is another way to measure the correlations between two variables by comparing the performance of classification. However, it cannot quantify which indicator is more relevant to the dependent variable and its process of calculation is unexplainable(Abiodun et al., 2018). Therefore, we aggregated Twitter data using the unit of Basic Statistical Area (AEB) and measured the relationship by the most commonly used linear regression. The AEB is a territorial unit for pure official statistical purpose in Barcelona city (**Figure 82**) that has homogeneous socioeconomic population within it, which divides Barcelona into 233 sectors whose average area is 0.44 km^2



Source: own elaboration

Figure 82 Total density of tweets of Barcelona based on AEBs

For better capturing the variation of the public sentiment, the studied tweets were only restricted to positive tweets and negative tweets. The public sentiment is evaluated by two aspects: the sentiment density and the net sentiment score. The former measures the spatial distribution of the sum of positive and negative tweets (*i.e.* sentimental tweets). The latter describes the net variation of Twitter sentiment among AEBs. The sentiment density is given by the average density of all positive and negative tweets on each AEB *i.e :*

$$Sentiment\ Density\ = \frac{Sum\ of\ postive\ and\ negative\ tweets_i}{Area}_i \qquad (1)$$

The sentiment score of a tweet is solely assigned by the Vader sentiment score because the score of SentiStrength, as previously discussed, does not produce a continuous scale. The net sentiment score of each AEB is calculated by the average score of positive tweets and negative tweets:

$$Net\ Sentiment_i = \frac{\sum Vader\ score_{posneg}}{\sum Tweets_{posneg}} \quad (2)$$

The urban environment indicators consist of four layers: sociodemographic, built-environment, and human mobility and socioeconomic activities (see

Table 43). Sociodemographic variables include the demographic density, education attainment, and the composition of professional positions, which could reflect the wealth of the AEB indirectly. The urban built environment contains specific land uses and residents' perceptions of the surrounding environmental quality (e.g. noise, contamination). The most visited places are also introduced into the model as categorical variables.

Human mobility and socioeconomic activities are represented by two data sources. The first one is the citizens' activities and socioeconomic indicators that were built by (Carlos Marmolejo-Duarte & Cerda-Troncoso, 2020; Marmolejo & Cerda Troncoso, 2017) using the origin-destination mobility data from the Metropolitan Transport Authority survey.

The time density is the number of hours that citizens expended in a given transport zone. The diversity is computed considering the activities that citizens perform out of their home (e.g. working, shopping, visiting friends), and the socioeconomic diversity of the people performing the aforementioned activities. Additionally, such authors built a composite centrality index using the DP2 methodology. The larger centrality index indicates the larger time density and diversities. Secondly, Foursquare POIs are exploited to indicate more detailed spatial information and places that people tended to use, which categories of classification are based on the previous case study.

All indicators are involved in the OLS regression model to predict the relationship between the urban environment and public sentiment. The first one introduces the logarithmic density because the ordinary least square approach requires the distribution of the dependent variable approaches to the normal distribution:

$$LnSentimentDensity_i = \beta_0 + \gamma_i * S_i + \delta_i * B_i + \mu_i * H_i + \varepsilon_i \qquad (3)$$

Where i is an AEB. β_0 is the constant of the regression; S_i, B_i, H_i represent the variables of socioeconomic, built environment and human activities respectively; γ ,δ ,μ are gradients associated with these three group of variables separately; ε_i is the error term with usual properties. Correspondingly, we use the same explanatory variables to estimate the variation of the net average sentiment:

$$Net\ Sentiment\ _i = \beta_0 + \gamma_i * S_i + \delta_i * B_i + \mu_i * H_i + \varepsilon_i \qquad (4)$$

Table 43 Urban environment indicators

1. Sociodemographic

Original statistical scale : census tract

Density of population: habitants/km²

1) total population
2) total Spanish population

Education attainment: (University population/ Population above 16 years-old)

3) people with university degree(%)

Percentage of job positions:

(population of specific occupations/ the total population of occupations)

4) managers in companies and public administrations
5) scientific and intellectual technicians and professionals
6) technicians and support professionals
7) administrative employees
8) workers of catering services, personal security, and salesman
9) skilled workers in agriculture and fishing
10) craftsman and skilled workers of the manufacturing
11) facility and machinery operators and assemblers
12) unskilled workers

Socio-professional index: (principal components extracted from the percentage professional positions)

13) High
14) Medium
15) Low

Data source: Padron INE. 2017; Census INE. 2001

2. Built-environment

Original statistical scale : 1)-5) ,11)-19): AEB; 6)-10): census tract

Specific land use :

1) historical area (%)
2) urban parks and garden (%)
3) Area of railway system (%)
4) area of water front and beach (%)
5) urban dense area (%)
6) density of storefront (storefronts/km²)
7) commercial equipment accounts for the total storefront (%)

Percentage of household consider there is a pollution issue: (number of households / all census households)

8) noise
9) dirty
10) contamination

Whether the AEB contains top visited places (buffered 0.4 km,polygon):

11) Casa Milà and Casa Batlló
12) La Rambla Street
13) Passeig de Gràcia avenue
14) Camp Nou stadium
15) La Sagrada Família temple
16) Sants- Montjuïc hill
17) Cosmo La Caixa Museum
18) El born historical zone
19) Park Güell

Data source: Statistical Department of Barcelona ; Census, INE. 2001 ; Own –elaboration

3. Human mobility and socioeconomic activities

Original statistical scale : 1)-6)Transport Zones; 7)-15) AEB

1) Centrality index of activities
2) Diversity (total/working days/ non-working days)
3) diversity of activities (total /working days/ non-working days)
4) socioeconomic diversity (working days/ non-working days)
5) time density (total/working days/ non-working days)
6) density of people performing activities (total /working days/ non-working)

Density of Foursquare POI (points/km²):

7) Bar
8) Restaurant
9) Outdoor resorts
10) Workplace
11) Residential place
12) Hotel
13) Transportation

Data source: Enquesta de Mobilitat Quotidiana (EMQ 2006) origin–destination household surveys, Autoritat del Transport Metropolità's (ATM) ; Yang & Marmolejo-Duarte(2019)

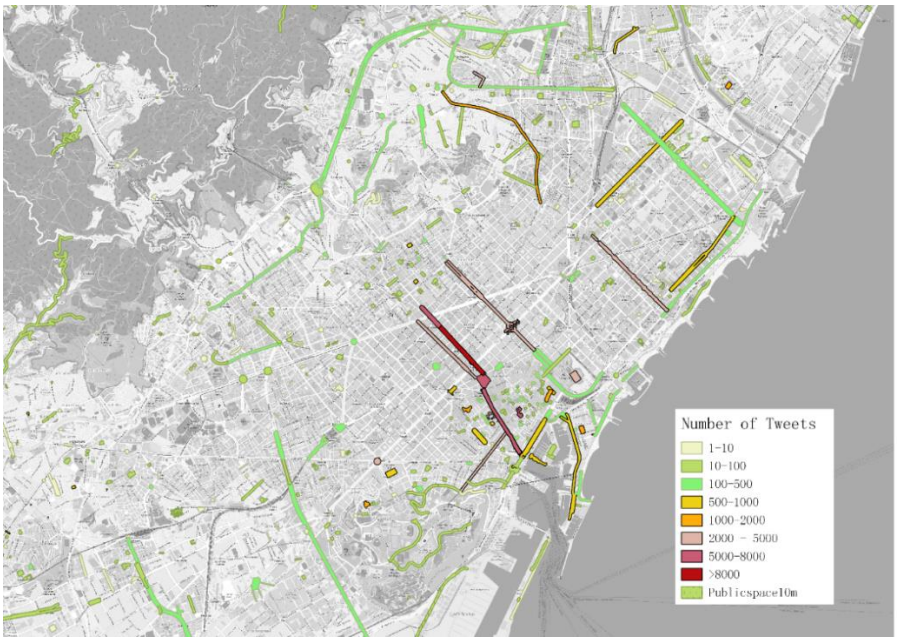
Note: 1. The 2011 census is not statistical significant at census tract level. 2. Barcelona is divided into 1491 census tracts and 63 transport zones. For each AEB, values of indicators that are originated from the two

statistical levels are recalculated by the area -weighted average number. The weight is the percentage of areas of each census tracts within the AEB.

3.4. Results

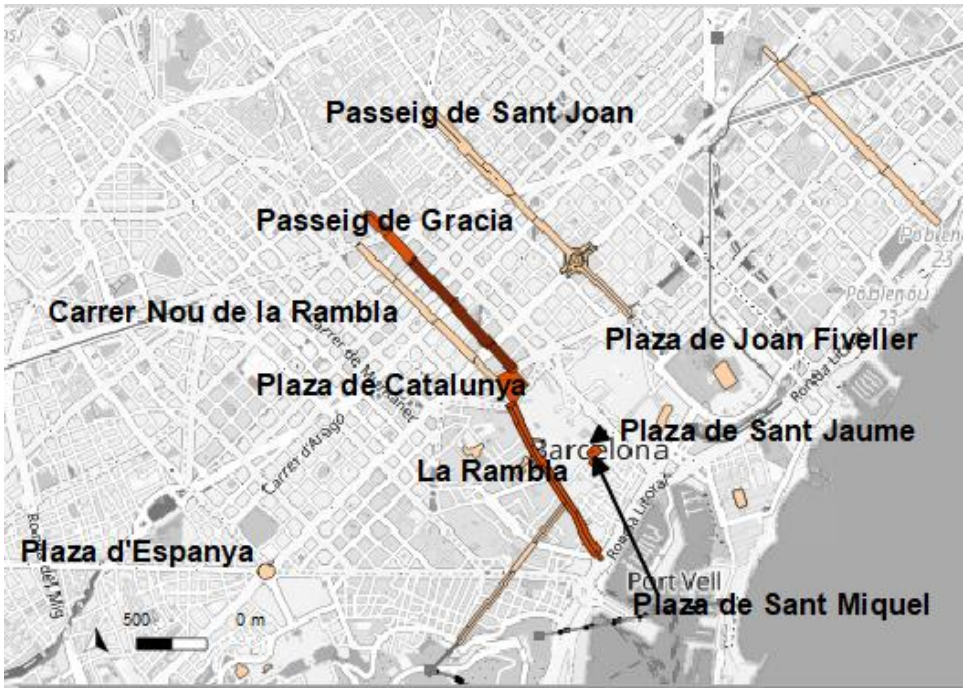
3.4.1 Spatial variation of Twitter activities

According *to London Assembly's 2011*'s report, public space refers to "all spaces including streets, squares and parks that everyone can use and access in principle." Therefore, based on OpenStreetMap, we choose plazas and major pedestrian avenues in Barcelona as the representatives of public space (**Figure 83**), because plazas and pedestrian avenues belong to the most common public space in European cities. What's more, all public areas were expanded by a 10-meter buffer for collecting tweets, as the boundary of these areas are not very precise.



Source: own elaboration

Figure 83 Spatial distribution of tweets in public spaces of Barcelona city



Source: own elaboration

Figure 84 Spatial distribution of tweets in central area of Barcelona

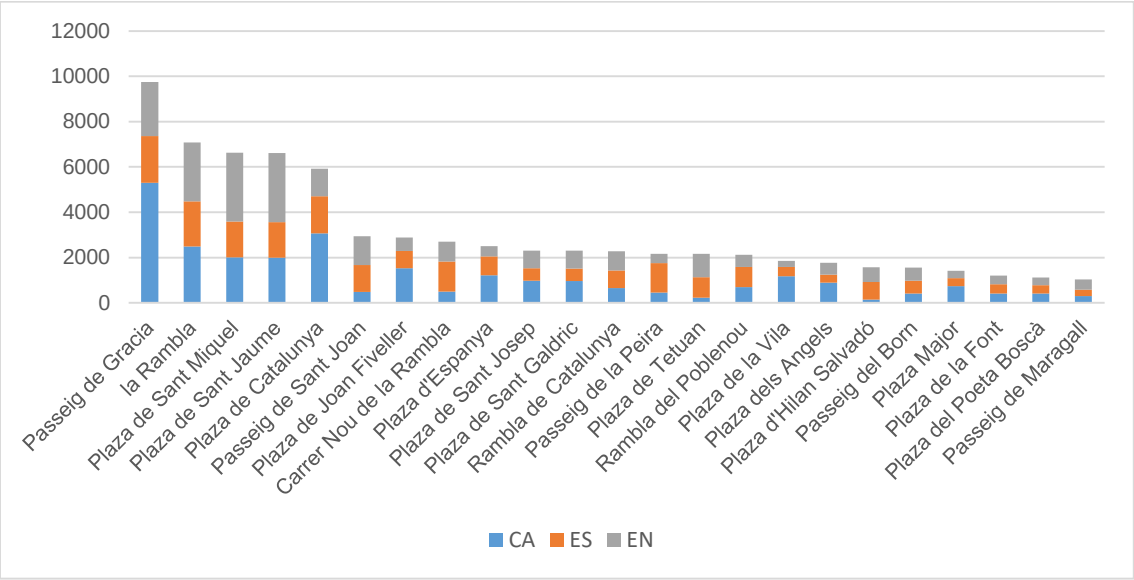
After aggregation, the total polygons of the public space are 1478 in Barcelona. There are 581 polygons contain Twitter messages. However, 75% of polygons contains less than 100 tweets. The majority of tweets were gathered in polygons of the central area of Barcelona (**Figure 84**). The densest public space is along with Passeig de Gracia- la Rambla.

Table 44 Spatial distribution of Tweets in public spaces

N. of tweets	N. of polygon	%
10<	266	45.78%
10-100	182	31.33%
100-500	98	16.87%
500-1000	11	1.89%
1000-2000	8	1.38%
2000-5000	11	1.89%
5000-8000	6	1.03%
>8000	1	0.17%

Source: own elaboration

Figure 85 lists the distribution of languages (English, Spanish, and Catalan)in 23 public spaces which contain more than 1000 tweets. In terms of language distributions, several areas show evident different components of languages. For example, Passeig de Gracia aggregated a lot of Catalan tweets, however, the proportion of Catalan reduced gradually when the space moves down to la Rambla and two plazas of the historical center (Plaza de Sant Miquel, Plaza de Sant Jaume). Spanish tweets take the largest proportion in Passeig de la Peira and Plaza de Tetuan.



Source: own elaboration

Figure 85 Distribution of the three languages in public spaces of Barcelona

3.4.2 Word-Frequency analysis

In terms of high-frequency words(**Table 45**), English and Spanish appear more emotional words than Catalan tweets, such as good, love, beautiful, etc. Catalan high-frequency words contain more names of places of Barcelona, such as 'sant','cugat', 'plaça'(square in English), which are accord with the finding of

previous sample test. English tweets contained more words which are related to tourism, such as photo, drinking, hotel, etc. Spanish tweets seem to be in between.

Table 45 The top forty high-frequency words

English		Spanish(translated)	
Word	Frequency	Word	Frequency
barcelona	7599	barcelona	6066
posted	1611	spain	1521
photo	1606	day	1426
gothic	1494	today	1277
quarter	1117	good	1240
spain	1032	new	1035
day	738	plaza	918
one	711	catalunya	875
love	661	beach	837
new	633	barceloneta	810
city	589	photo	783
ES	553	love	780
time	529	one	780
night	529	happy	695
good	526	time	674
like	519	see	672
last	509	like	619
city	475	best	618
drinking	470	great	583
beautiful	441	rambla	556
happy	437	city	549
best	430	want	548
barceloneta	417	life	533
beach	415	boqueria	531
back	412	night	514
mercat	400	always	490
boqueria	397	thanks	482
playa	393	go	474
hotel	391	ramblas	471
barcelona.	370	apolo	461
great	368	bcn	458
amazing	365	sala	456
rambla	364	know	441
catalunya	357	us	441

see	352	casa	439
running	349	year	437
finished	333	morning	435
get	332	club	426

Source: self-elaboration

3.4.3 Descriptive analysis of sentiment tweets

Table 46 shows the result of the automatic sentiment classification of the whole English and Spanish dataset. The rate of agreement is over 70% in both languages. The percentage of neutral tweets accounts for the largest proportion in both Spanish and English Tweets; negative tweets are less than 5%. The similar proportion of negative tweets could also be observed in the sentiment classification of tourist destinations in Chicago (Padilla *et al.*, 2018). Considering the pivot role of tourism in Barcelona, it could imply that the percentage of negative tweets is lower in tourist places and cities.

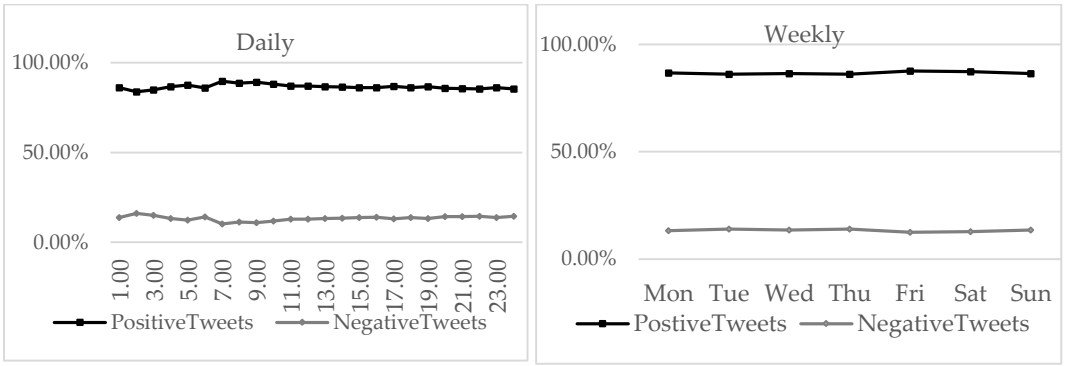
Table 46 Results of Sentiment classification

Category of sentiment	English Tweets	Percentage	Spanish Tweets	Percentage
VS_pos	47,307	22.96%	68,356	30.62%
VS_neg	6,756	3.28%	11,039	4.94%
VS_neu	108,160	52.51%	83,923	37.59%
Total agreement tweets	162,223	78.75%	163,318	73.15%
Total Tweets	205,997	100.00%	223,274	100.00%

Source: own elaboration. Note: VS: Vader + SentiStength. pos, neg, neu means positive, negative, and neutral respectively.

Figure 86 depicts the temporal variation of sentimental tweets in the form of proportion. The most volatile period is at midnight (01:00 am to 6:00 am) of one day. The rest of the periods are comparatively stable. In terms of the weekly variation, the positive tweets slightly rise on Fridays and the percentage of negative tweets is higher on Tuesdays and Thursdays. **Figure 87** shows the variation of the net sentiment score. The average negative score was multiplied -1 for the sake of variation clearly. Similarly, the midnight appears with extreme variation from the positive to the negative. Such fluctuation is understandable

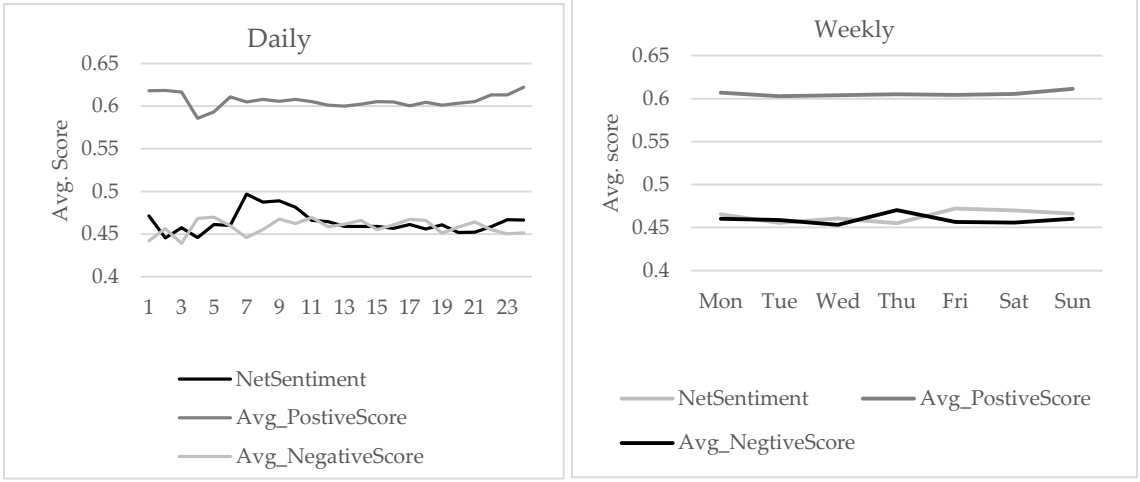
because users who still were awake and sent tweets at midnight probably encounter some disrupting events or issues. The most positive period is in the early morning because the negative sentiment drops steeply. The weekly negative sentiment score appears a spike on Thursday, though the average positive score maintains stable.



(a) (b)

Source: own elaboration

Figure 86 Variation of percentage of sentimental tweets



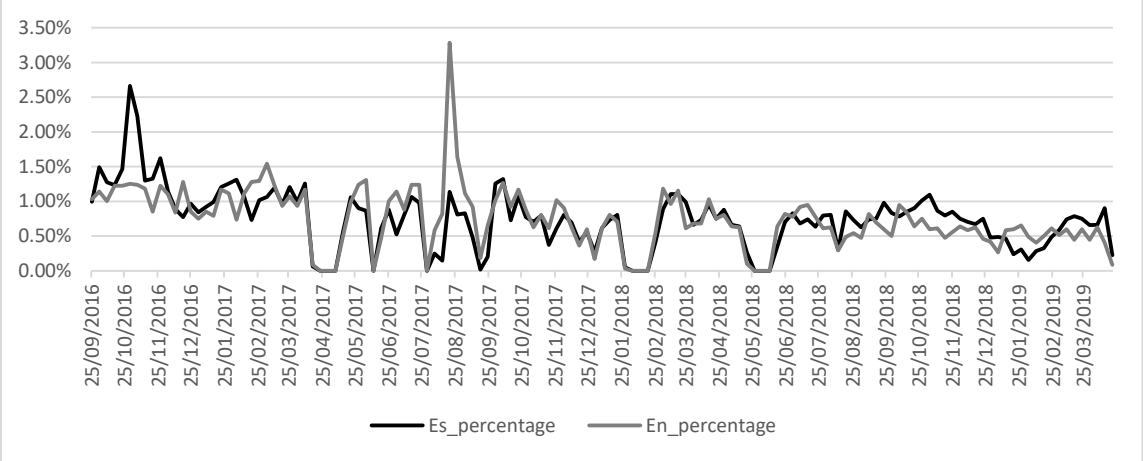
(a) (b)

Source: own elaboration

Figure 87 Variation of the net sentiment score

For exploring the reason of variation of negative sentiment, **Figure 88** depicts the weekly tendency of all Spanish and English negative tweets. Several weeks did

not have negative tweets, which were probably caused by the data loss of the original dataset or the process of data cleaning. If we removed those null values, the variation of negative tweets did not exceed 0.5% in general. There is one evident peak in Spanish and English Tweets separately. The two peaks are correlated with two important events. The negative peak of Spanish tweets happened on the first of November of 2016 (Tuesday), when an important football game of UEFA Champions League was held-- Manchester City won against FC Barcelona with 3-1. The importance of football in Spain could be reflected by the top five Twitter accounts with the largest number of followers – all of them either are football clubs or football players. The negative peak of English tweets appeared on the day of the Barcelona terrorist attack, the 17th of August 2017 (Thursday). It explains the reason that the weekly variation of sentiment scores was lower on Tuesday and Thursday. In summary, the public sentiment indeed can be disclosed from Twitter data.



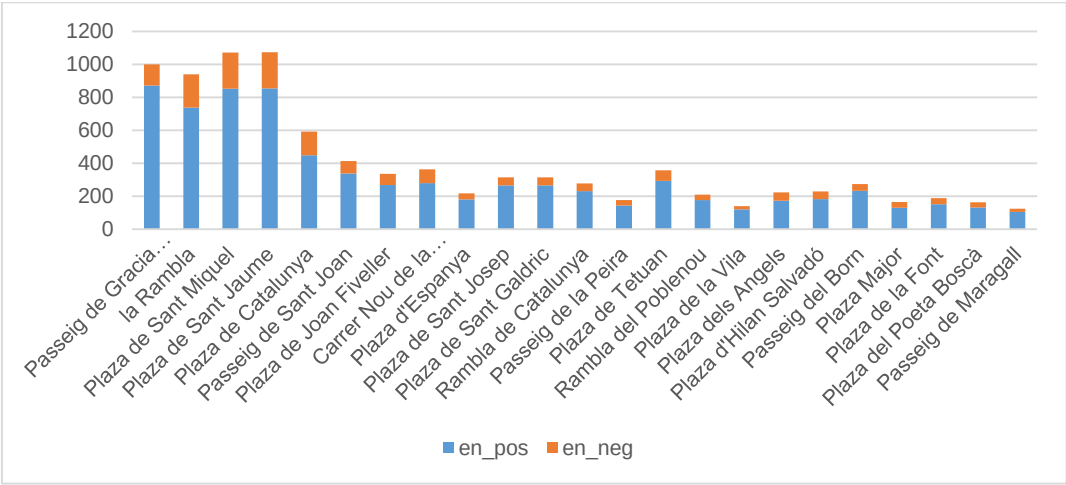
Source: own elaboration

Figure 88 Weekly variation of negative tweets

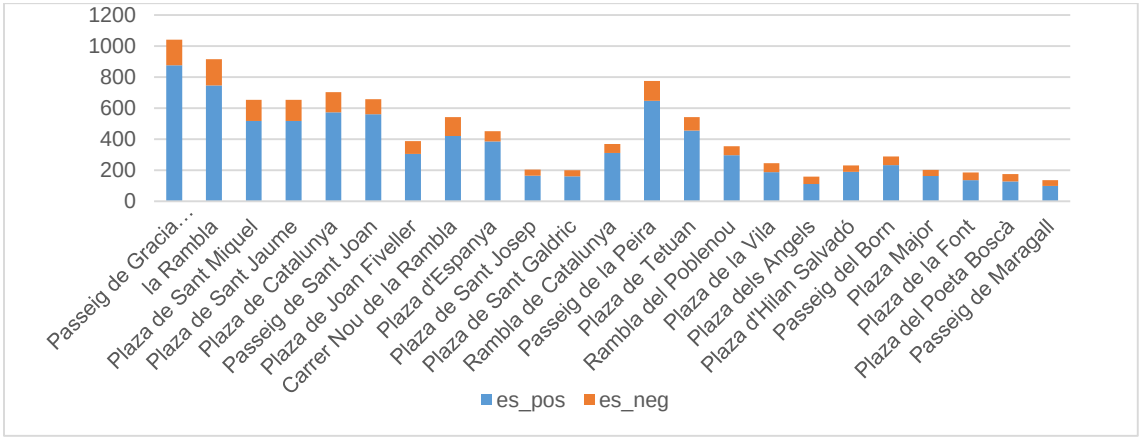
Source: own elaboration

Figure 89 displays the distribution of positive and negative sentiments of Spanish and English tweets in the public spaces separately. The negative tweets only account for a small portion of all these places. Positive emotion takes the dominant role. In general, negative tweets decrease as the total number of tweets decrease. However, the proportion of English negative tweets is higher in Plaza

de Sant Miquel and Plaza de Sant Jaume. One possible explanation is that the government of Catalonia and Barcelona are located in these two places, thus there are more political events that happened and could evoke negative emotion in these two places.



(a) English



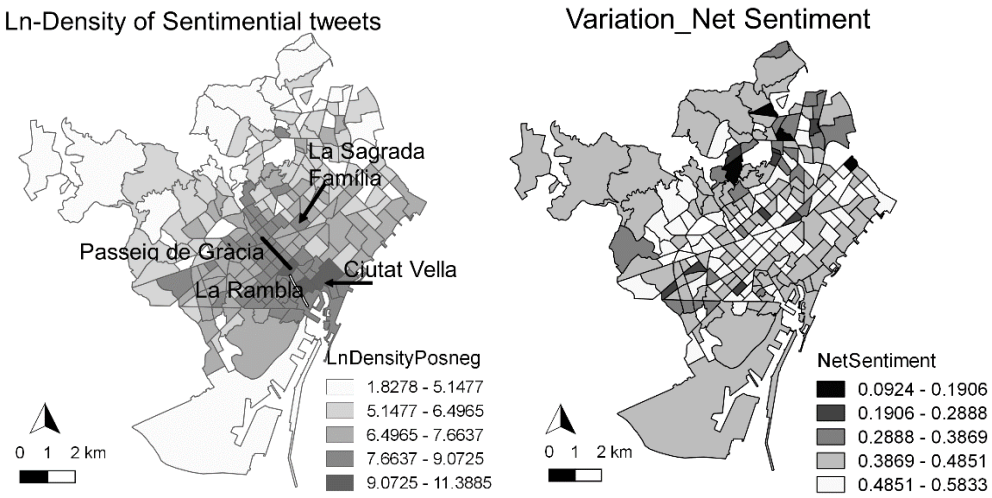
(b) Spanish

Source: own elaboration

Figure 89 Distribution of positive and negative sentiments

3.4.4 Correlation between urban environment and Twitter sentiment

For guaranteeing the representativeness, AEBs that were introduced into the analysis of relationship should have both positive tweets and negative tweets of both languages. Following the criterion, 200 AEBs were entered in the analysis. 33 AEBs that were excluded are mainly located in the peripheral areas of Barcelona. **Figure 90**(a) displays that sentimental tweets are mainly concentrated in the city center which contains most of the famous tourist attractions, such as La Sagrada Família, Ciutat Vella (the medieval neighbourhood), Passeig de Gràcia Avenue and La Rambla Street. Both of streets are the major commercial and tourist avenues in Barcelona and contain some most celebrated architectural works, such as Casa Milà, Casa Batlló and La Boqueria Market. Compared with the sentiment density, the spatial variation of sentiment score (**Figure 90**(b)) is dispersed, though the central area performs higher positive scores in general.



Source: own elaboration

Figure 90 Spatial distribution of sentimental tweets and net sentiment

To fully explore the relationship between urban environment indicators and public sentiment, **Table 47** visualises all paired Pearson's correlation values of indicators which do not filter out statistically insignificant variables (p-value <0.05). The density of sentimental tweets and the sentiment score do not have a strong

interrelationship with each other. This finding broadly suggests that dense-tweet AEBs exhibit a positive score. In general, the sentiment density is highly correlated with human and socioeconomic activities and less associated with sociodemographic indicators. The total centrality index (**Figure 91**(a)) has a higher positive correlation with the density, which indicates these tweets are concentrated in the central area of Barcelona. Places that people tend to stay longer (**Figure 91**(c)(d)(e)), such as restaurants, bars, outdoor resorts, workplaces, appear more sentimental tweets than transportation and residential places. The relationship between net sentiment and the intensity of human and socioeconomic activities is not very intense, though they are positively correlated. The variation of sentiment score has the weakest correlation with urban environment indicators.

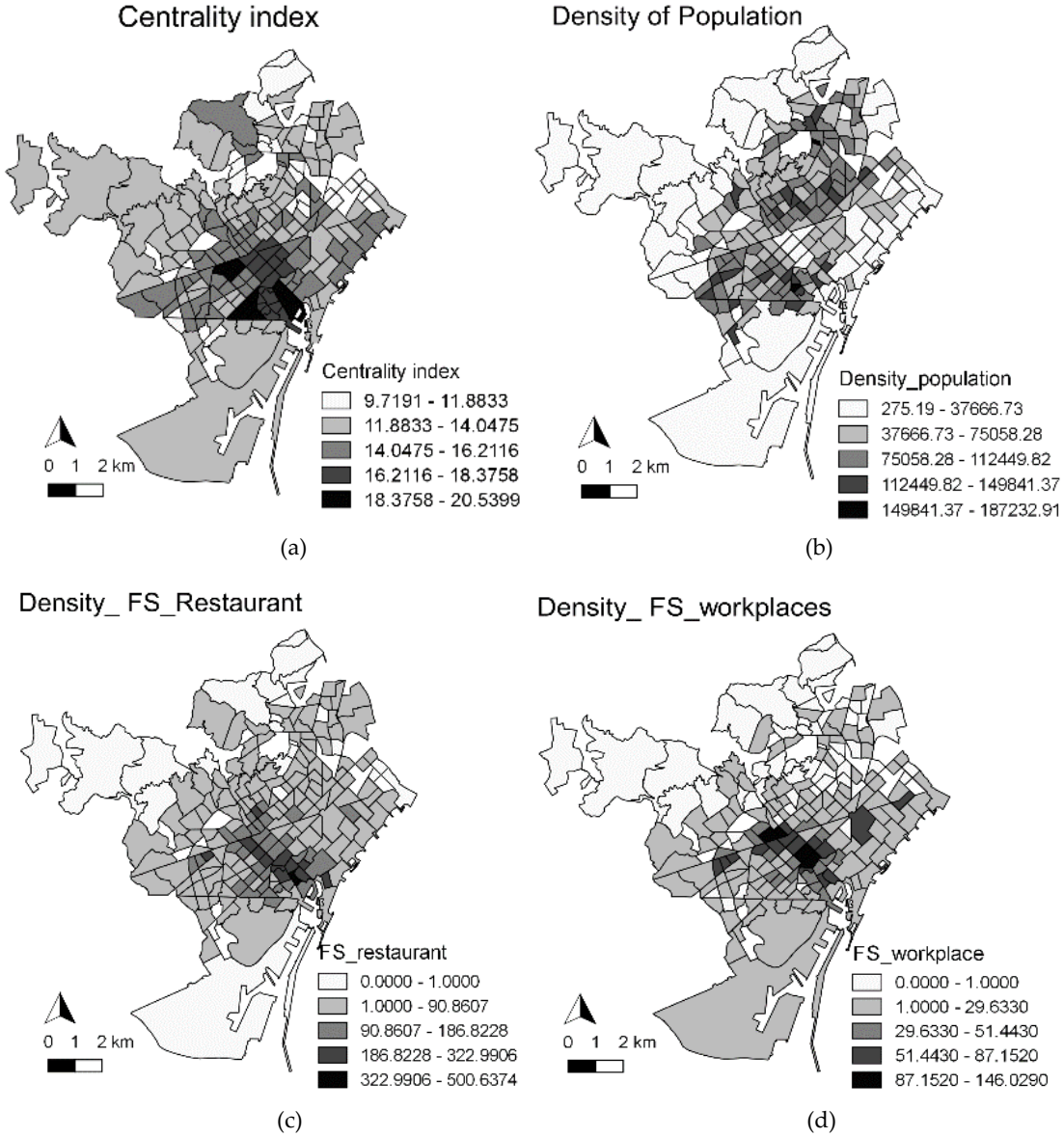
Table 47 Pearson correlation between urban indicators and sentiment indicators

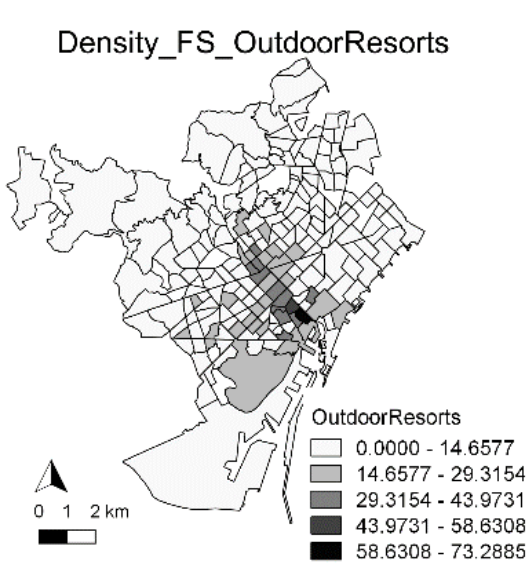
Sociodemographic	NetSentiment	Lndensity posneg	Built enviroment	NetSentiment	Lndensity posneg	Human mobility and socioecomoic	NetSentiment	Lndensity posneg
D_population	-.1486*	.1994**	%_Historical area	.0305	.4020**	Total Centrality	.2229**	.6165**
D_Spanish people	-.1502*	.1852**	%_Park_garden	-.0934	-.0590	Total Diversity	.1876**	.4947**
% _high_education	.3267**	.2234**	%_Railway	-.0894	-.0214	Total_Diversity_wk	.1876**	.4947**
Pr_manager	.2747**	.1166	%_watefront_beac	.0565	.0141	Total_Diversity_nw	.2137**	.4784**
Pr_sc_technican	.3173**	.2502**	%_Urbandense	.0756	.1299	Diversity_activities_w	.1371	.4753**
Pr_sp_technican	.0413	.1287	D_storefront	.2323**	.5432**	Diversity_activities_n	.1830**	.4549**
Pr_adiminstrative	-.2626**	-.2203**	%_commercial_sto	-.1329	.0999	Diversity_se_wk	.1363	.2081**
Pr_services	-.2025**	.0804	O_dirty	-.0431	.0903	Diverstiy_se_nw	.2177**	.1966**
Pr_agriculture	-.0011	-.0060	O_noise	.2772**	.5361**	Time density_total	.1315	.4475**
Pr_artisan	-.3387**	-.3662**	O_contamination	.2300**	.4218**	Time density_wk	.1577*	.4635**
Pr_operator	-.2849**	-.4216**	Casa_Mila_Batillo	.1519*	.3292**	Time density_nw	.0753	.3970**
Pr_noskill	-.1323	-.0298	LaRambla	.0492	.3339**	D_activepeople_total	.1213	.4784**
PC_profession_high	.3396**	.4559**	PgGracia	.0679	.3102**	D_activepeople_wk	.1446*	.4861**
PC_profession_mid	-.2633**	-.1701*	CampNou	-.0442	.0354	D_activepeople_nw	.0802	.4480**
PC_profession_low	-.2119**	.0142	SagradaFamilia	-.0006	.0542	D_FS_bar	.1521*	.6266**
NetSentiment	1	.2424**	Sants_Monjuic	.0338	-.0341	D_FS_restaurant	.2375**	.6851**
Lndensity_posneg	.2424**	1	El born	.0652	.3308**	D_FS_transport	.1167	.3335**
D:Density	O: Opinion		Cosmolacaxia	.0726	-.1147	D_FS_outdoor	.1280	.6234**
Pr:Profession	FS:Foursquare		Park Güell	-.2081**	-.0916	D_FS_hotel	.1297	.5410**
wd: non-workingdays ; wk:workding days;						D_FS_residential	.0625	.2434**
**. Correlation is significant at the 0.01 level (2-tailed).						D_FS_workplace	.2089**	.5419**
*. Correlation is significant at the 0.05 level (2-tailed).								

Source: own elaboration

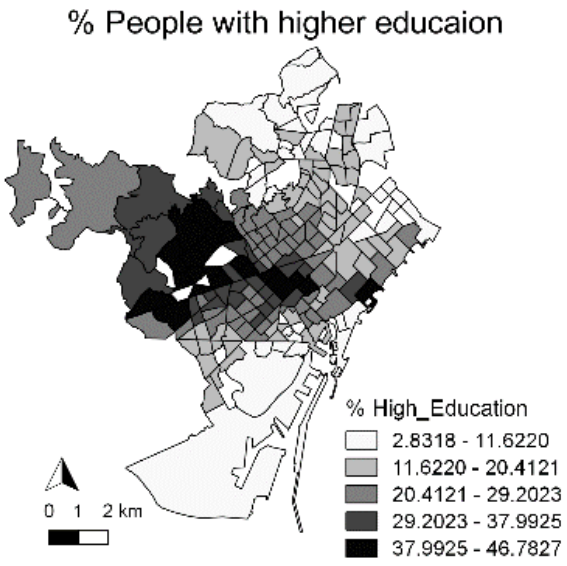
Regarding to sociodemographic indicators, wealthier ABEs exhibit a higher density of sentimental tweets and positive sentiments. Areas with a higher percentage of high-income positions and high-educational people (**Figure 91**(f)) show a higher density of tweets and optimism. Conversely, AEBs with a higher percentage of lower-income positions (**Figure 91**(i)), such as artisans, operators,

and clerks, perform a negative correlation with the density and the sentiment score. The positions related to catering services (**Figure 91**(h)), personal security, and salesman presents a positive association with the sentiment density but a negative relationship with positive scores, remarkably these people do not live in wealthy areas. The density of the population also (**Figure 91**(b)) appears with a similar relationship.

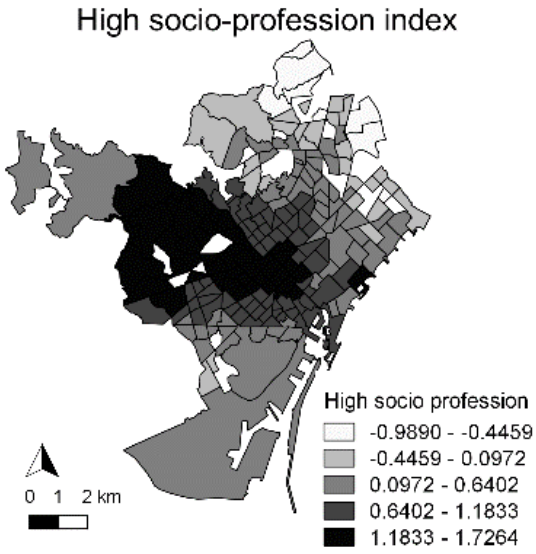



Density_FS_OutdoorResorts
OutdoorResorts
0.0000 - 14.6577
14.6577 - 29.3154
29.3154 - 43.9731
43.9731 - 58.6308
58.6308 - 73.2885
0 1 2 km

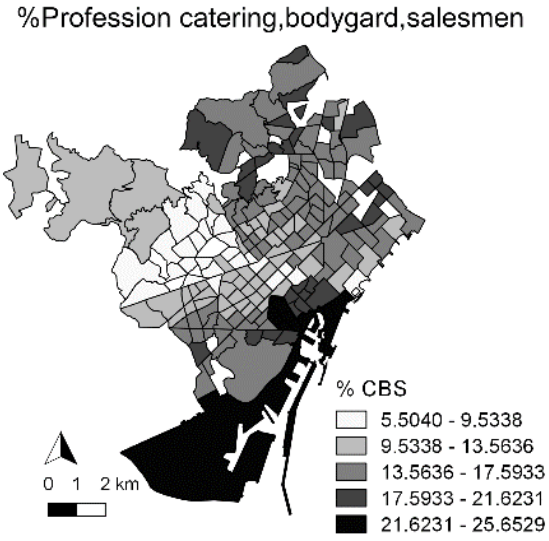
(e)


% People with higher educaion
% High_Education
2.8318 - 11.6220
11.6220 - 20.4121
20.4121 - 29.2023
29.2023 - 37.9925
37.9925 - 46.7827
0 1 2 km

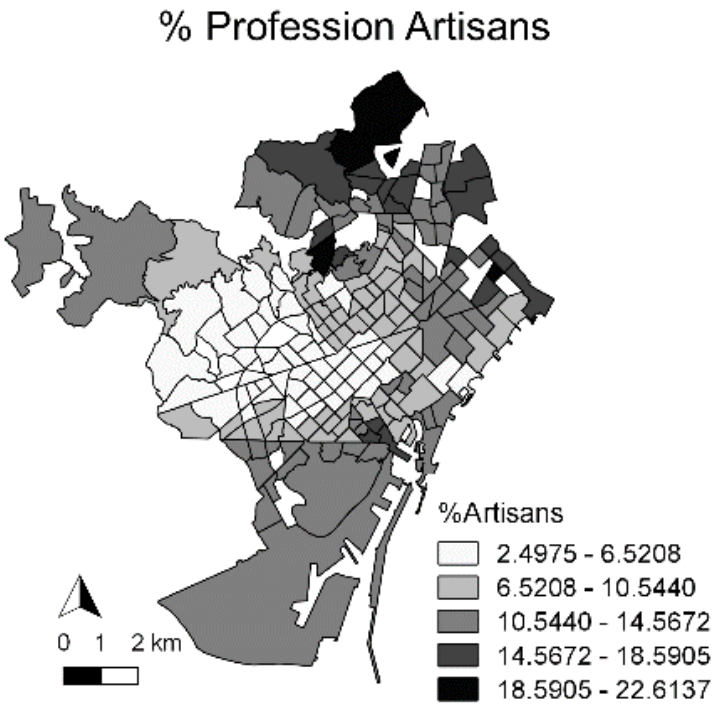
(f)


High socio-profession index
High socio profession
-0.9890 - -0.4459
-0.4459 - 0.0972
0.0972 - 0.6402
0.6402 - 1.1833
1.1833 - 1.7264
0 1 2 km

(g)


%Profession catering,bodygard,salesmen
% CBS
5.5040 - 9.5338
9.5338 - 13.5636
13.5636 - 17.5933
17.5933 - 21.6231
21.6231 - 25.6529
0 1 2 km

(h)



(i)

Source: own elaboration

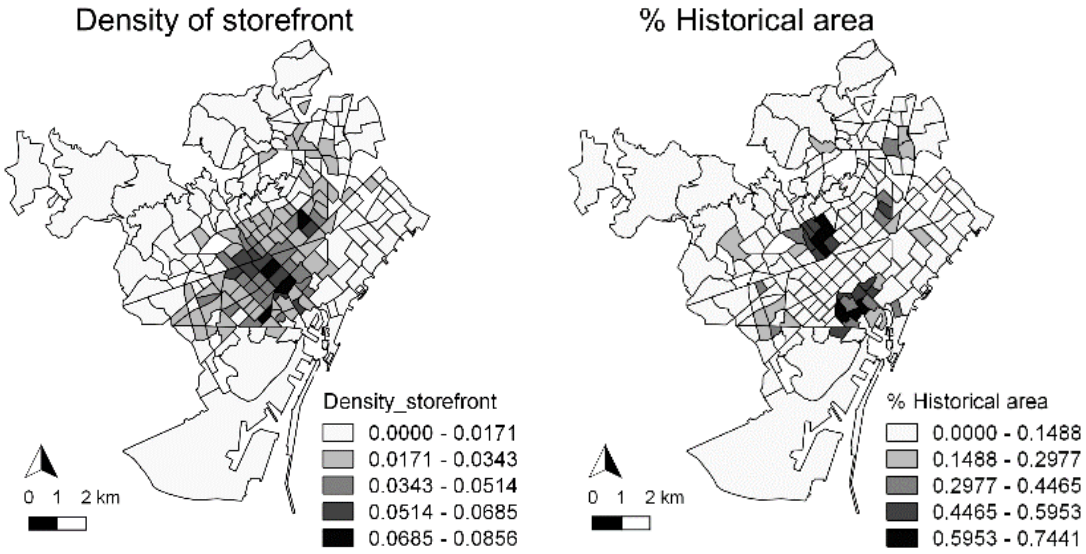
Figure 91 Spatial distribution of sociodemographic, human mobility and socioeconomic activities

The built environment indicators perform a complex and subtle relationship with public sentiment. Historical areas, the density of storefronts, and beach areas are positively correlated with the sentiment density and scores. The percentage of railway areas is negatively correlated with the two sentiment variables because large empty railway areas only appear in suburban districts (**Figure 92**(g)). The percentage of urban parks and gardens appears negatively correlated due to the fact that some central-located AEBs lack such greenery (**Figure 92**(h)). Similarly, the percentage of commercial storefront actually is lower at the urban center because the other activities and personal oriented services, such as health care and culture centers, dominate the central landscape.

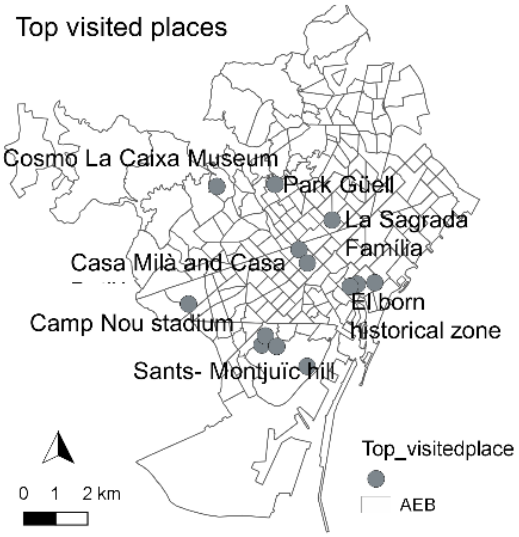
Both noise and contamination are positively correlated with public sentiment because these two indicators are mainly gathered in the city center where various services, intense (despite noisy) street activities and amenities are located, except

for the industrial zone in the southeast of Barcelona (**Figure 92** (d),(e)). Thus such indicators are actually proxying for such active environments. The perception of dirtiness is negatively correlated with positive sentiment because the opinion of dirty areas is higher in the industrial zone of Barcelona as well as in certain tourist populated areas (**Figure 92**(f)).

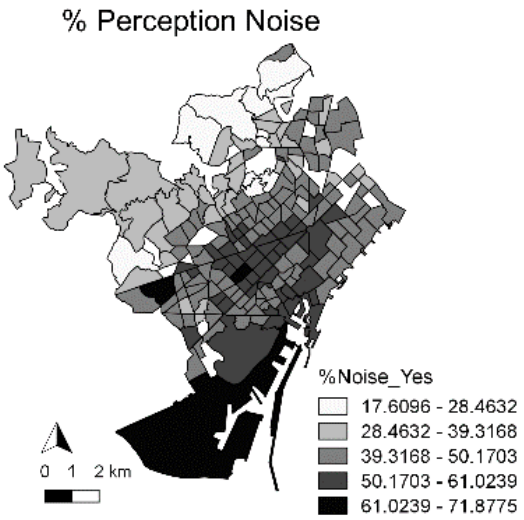
The location and characteristics of tourist attractions (**Figure 92** (b)) affect their correlation with the two sentiment indicators. La Sagrada Família and Camp Nou are positively associated with the density but negatively correlated with the sentiment score. La Sagrada Família is surrounded by a higher percentage of residents who might feel bothered by tourists. The public sentiment in Camp Nou was probably influenced by the results of soccer games. Cosmo Caixa Museum -- a local-famous science museum and Park Güell, are far away from the city center, hence it shows a negative relationship with the sentiment density comparatively and the capacity limited.



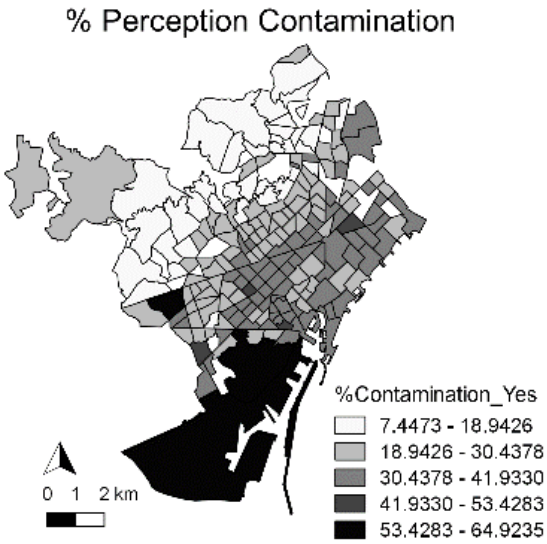
(a) (b)


Top visited places
Cosmo La Caixa Museum
Park Güell
La Sagrada Família
Casa Milà and Casa
El born historical zone
Camp Nou stadium
Sants- Montjuïc hill
Top_visitedplace
AEB
0 1 2 km

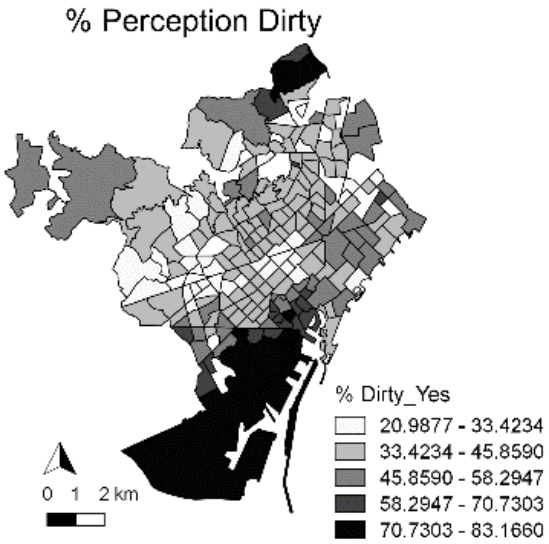
(c)


% Perception Noise
%Noise_Yes
17.6096 - 28.4632
28.4632 - 39.3168
39.3168 - 50.1703
50.1703 - 61.0239
61.0239 - 71.8775
0 1 2 km

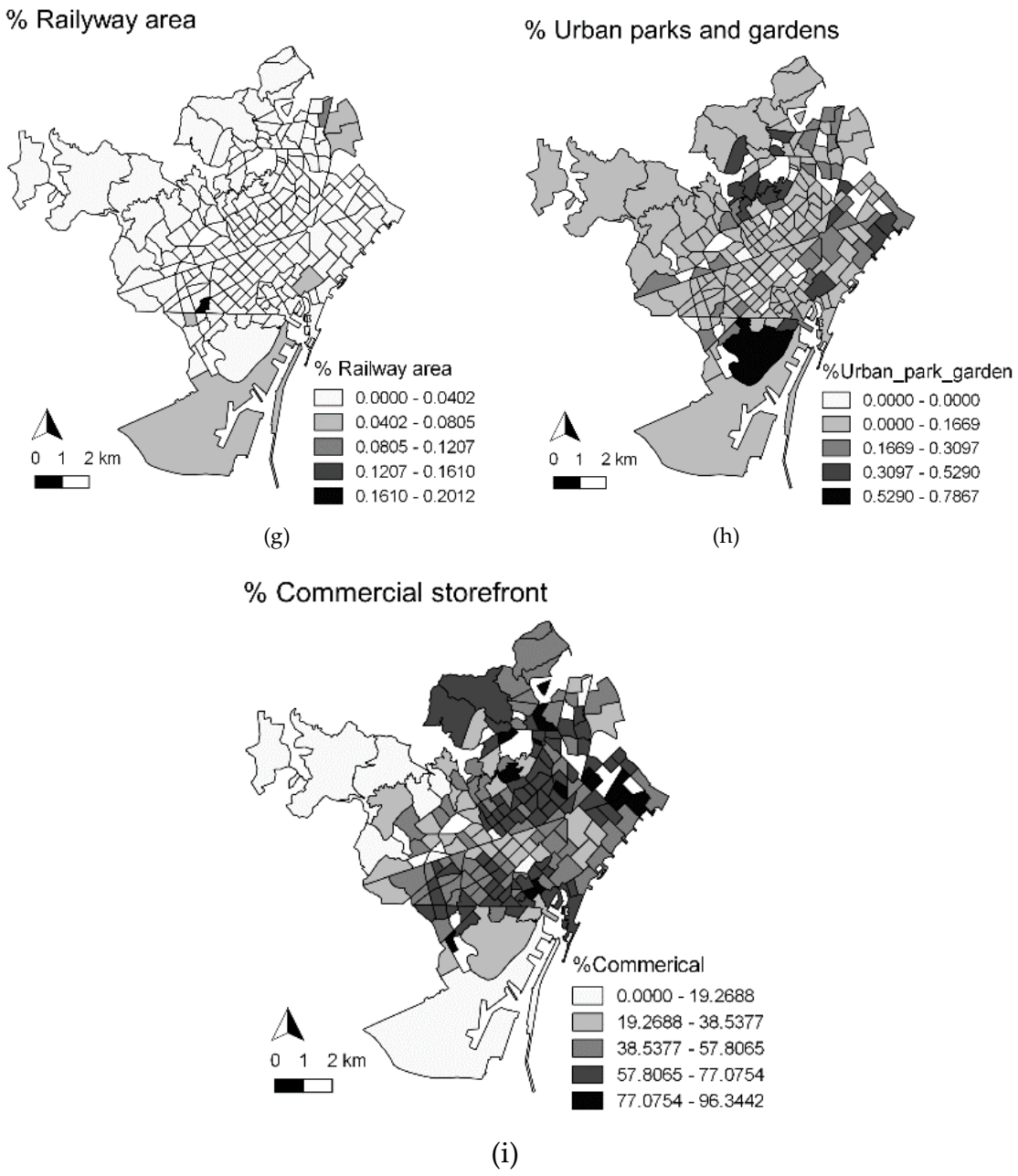
(d)


% Perception Contamination
%Contamination_Yes
7.4473 - 18.9426
18.9426 - 30.4378
30.4378 - 41.9330
41.9330 - 53.4283
53.4283 - 64.9235
0 1 2 km

(e)


% Perception Dirty
% Dirty_Yes
20.9877 - 33.4234
33.4234 - 45.8590
45.8590 - 58.2947
58.2947 - 70.7303
70.7303 - 83.1660
0 1 2 km

(f)



Source: own elaboration

Figure 92 Spatial distribution of built environment

3.4.5 Regression analysis

The model of sentiment density essentially confirms the interrelationship discussed previously, which R^2 reaches to 0.689 (错误!未找到引用源。). The

density of the population and the density of Foursquare POIs are introduced in their logarithmic form for fitting reasons. Since urban settings vary in different AEBs, the robust model is adopted to solve the problem of heteroscedasticity. For example, the historical area is a positive indicator in the model, however, several AEBs with a higher percentage of historical areas (**Figure 92**(b)) contain less sentimental tweets because they are at the peripheral districts of the city. This fact is the result of the urban aggregation of Barcelona which absorbed formerly independent towns nearby. The spatial autocorrelation has a negligible effect on the model, according to the result of the Moran I test of spatial correlation in Geoda software (See **Appendix A**). The assumption of spatial correlation was denied, which indicates that the spatial autocorrelation has a negligible effect on the model.

Table 48 Model summary of sentiment density

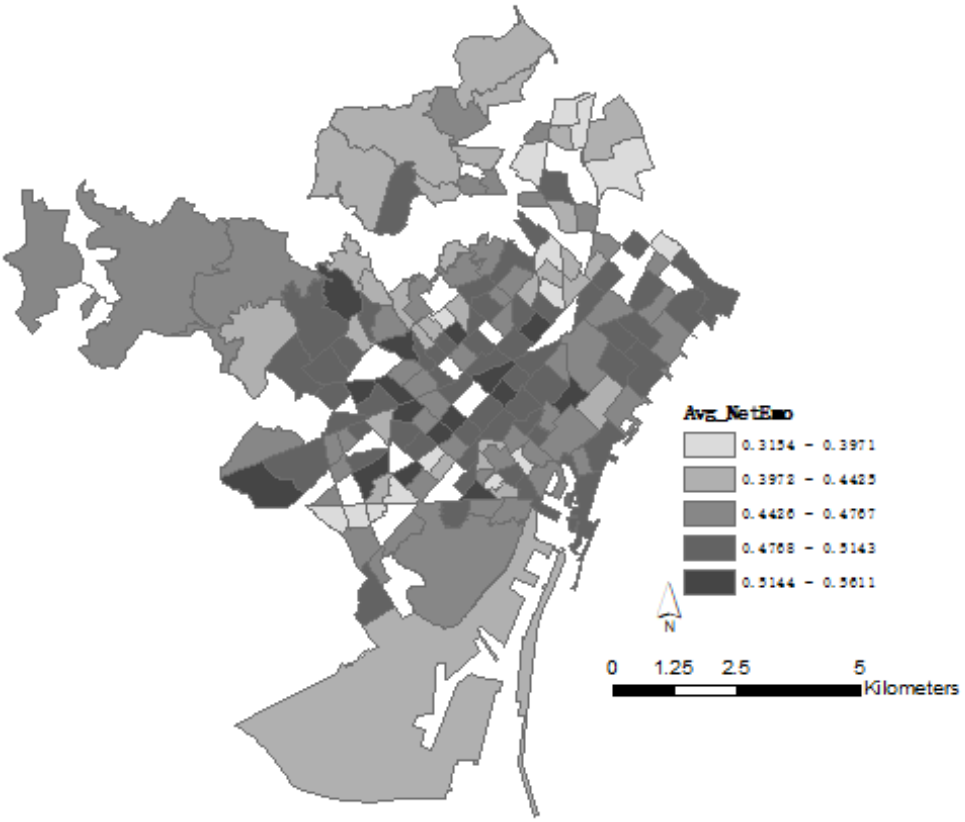
Linear regression	Robust		Durbin-Watson		1.792		
Number of observation	200		R-squared		0.689		
F(10, 189)	35.07		Root MSE		0.760		
Prob > F	0						
	Coef.	Std. Err.	T	Sig.	[95% Conf. Interval]		VIF
	B				lower	upper	
Centrality index	0.112	0.031	3.650	0.000	0.051	0.173	1.950
Historical area(%)	1.110	0.256	4.350	0.000	0.607	1.615	1.240
LnD FS outdoor resort	0.029	0.009	3.470	0.001	0.012	0.045	1.130
LnD total population	0.371	0.089	4.170	0.000	0.195	0.546	1.440
PgGracia Avenue	1.136	0.399	2.840	0.005	0.348	1.923	1.360
El born historical area	0.975	0.253	3.850	0.000	0.475	1.475	1.190
Urban park garden	2.402	0.675	3.560	0.000	1.070	3.733	1.430
PC profession high	0.492	0.127	3.870	0.000	0.241	0.742	1.550
Contamination opinion	0.036	0.008	4.320	0.000	0.020	0.053	1.240
Density storefront	12.748	5.019	2.540	0.012	2.847	22.649	2.690
_Constant	-0.806	1.057	-0.760	0.447	-2.892	1.279	

Source: own elaboration

The most influential variable is the density of the population. Functionally, indicators are related to leisure activities and are coherently correlated with a higher sentiment density, such as storefronts, urban parks, historical areas, and

commercial equipment. The density of outdoor resorts of Foursquare POIs is the second dominant indicator, which includes famous tourist attractions, resting areas, scenic lookouts, amongst others. The urban parks and gardens become a positive indicator in the model because other variables compensate for the deficiency of parks and gardens in the central area of Barcelona (**Figure 92** (h)). The influence of urban parks and gardens is shown. In addition, the high socio-professional index as a positive indicator entered the model, which suggests that high-income AEBs are associated with a higher frequency of utilisation of social networking software.

The model of net sentiment only contains 168 cases due to the OLS requirement of normal distribution. **Figure 93** displays the 168 AEBs that enter in the regression model. It can be observed that some extreme positive and negative deviances were removed. After the test of heteroscedasticity and spatial correlation, the confirmed model (错误!未找到引用源。) shows that the urban environment has a lower impact on the variation of sentiment scores, being the adjusted R^2 equivalent to 0.272.



Source: own elaboration

Figure 93 Distribution of net sentiment at AEBs level used in the model

The most positive impact factor is the density of storefronts. The storefront consists of places of sanitary services, education, social welfare, business, office, culture among others. It could be considered as a proxy of service provision, jobs and bustling zones of the city. The commercial storefront became positive in the model because several AEBs at the north-western areas have been eliminated out of the model. The removed AEBs are of a higher percentage of commercial storefront at the peripheral areas (**Figure 92** (i)). The rest of AEBs with a high percentage of commercial store fronts are mainly close to the city center. Therefore, the higher social-economic is associated with higher net sentiment score. The density of population, as well as low income related positions, and the percentage of railway area are negatively correlated with the net sentiment score. It can be observed that AEBs with larger railway areas also have higher percentages of artisans who belong to lower income class. Such a result is coincident with the

research of Quercia et al. (2012), which revealed that the community's socio-economic well-being and the sentiment score are highly correlated.

Table 49 Model summary of the net sentiments

Number of obs.	168				Durbin-Watson		2.008
F(6, 161)	11.391				R-squared		0.298
Prob > F	0				Adjusted R Square		0.272
					Root MSE		0.038
	Coef.	Std. Err.	T	Sig.	[95% Conf. Interval]		VIF
	B				lower	upper	
Profession artisan	-0.004	0.0010	-4.690	0	-0.006	-0.002	1.617
%Railway area	-0.466	0.1550	-3.003	0.003	-0.773	-0.160	1.056
Density population	-0.448	0.1070	-4.206	0	-0.659	-0.238	1.726
Density storefront	0.729	0.230	3.170	0.002	0.275	1.183	1.849
% Commercial_st	0.001	0	2.603	0.010	0	0.001	1.763
_Constant	0.496	0.011	45.460	0	0.474	0.517	

Source: own elaboration

3.5. Conclusion

This research provides a panoramic perspective to inspect the relationship between public sentiment and the urban environment using Twitter data in terms of the spatial density of tweets and the variation of sentiment scores. Moreover, the sentiment analysis of Spanish and English amplifies the study groups of people. In addition, the translated texts and the mixture of different algorithms of sentiment classification could be an economic alternative plan for researchers who are not linguistic experts.

Firstly, it reveals the sentiment score and the spatial sentiment density is not strongly correlated. The sentiment density has a closer relationship with human mobility and socioeconomic activities. The density of population and the density of people's daily activities are dominant factors. In Barcelona, sentimental tweets tend to assemble in places where there are tourist attractions or leisure places. Such

a correlation indicates that the density of Twitter activities can be a proxy variable to observe physical human activities.

Secondly, the variation of sentiment score is mainly influenced by disruptive events and sociodemographic indicators, as strongly suggested by the analysis of negative sentiment. The socioeconomic indicators could partially explain the variation of Twitter sentiment in a macro-view. The result of Pearson's correlation shows that the wealthier AEBs tend to present higher sentiment scores. The model statistically confirms that the density of storefronts and the percentage of commercial shops has positive impact on the sentiment score. Both indicators could be read as the active degree of economy and lively streets thus the higher economic activities there are, the higher positive Twitter sentiment presents. The positive connection between urban sentiment and wealth have also been observed in the United States (Mitchell et al., 2013).

Moreover, although the regression model only confirmed a few statistical interrelationships between the net sentiment and the built environment, the Pearson's correlation still reveals some interesting relationships. For example, the coastal area, urban parks, bars and restaurants, tourist attractions are positive correlated with the sentiment score. The similar relationships can also be found in previous research(Gallegos et al., 2016; Schwartz et al., 2019). In addition, related to the temporal characteristics of net sentiment, midnight is the most fluctuate period of sentiments, though the amount of negative tweets is few compared with the whole dataset. The nadir of sentiment at midnight was also observed in research of Bertrand et al. (2013) and Gruebner et al. (2017) .Therefore, combing results of previous research, it could conclude that Twitter data actually provides some reliable results that could depict the sophistication of public sentiment.

It is undeniable that some flaws exist in the research. The lack of Twitter demography, such as gender, age, identity (tourist or residents), leads to the result having only scratched the surface of the connection between public sentiment and urban environment. However, like the literature review mentioned, the privacy boundary is a sensitive issue in the research. Secondly, the dataset of Catalan tweets is unavailable to sentiment analysis due to the poor quality of the data. It

does affect the representativeness of our results. In future research, the separation of tourists from locals will improve the accuracy of the analysis of public sentiments. Moreover, the social and spatial inequalities have presented from our results to some degree. Such issues should be investigated further combining with official statistical information.

Chapter VII. Discussions and Conclusions

VII.1. Summary of findings

This dissertation comprehensively investigates the applications and limitations of LBSN data in terms of its development, ontology, usages, and empirical studies. As each chapter has already clarified its conclusion and contribution, this chapter only makes a brief for integrating them.

Chapter II introduces the definition and characteristics of the LBSN data, which nature is massive user-generated information. LBSN data is originated from the daily activities of human beings. Although the motives of using LBSN applications are varying, entertainment, self-presentation, and sociality are principal motives of using these applications. It indicates that LBSN data has limited general representativeness.

Except for high spatial precision and low cost of data collection, the distinctive characteristic of LBSN data is the enrichment of social information. It makes uncountable ordinary human activities countable and refreshes the recognition of urban space and human mobility. It provides a novel outlook for inquiring about human activities behaviors and understanding cities in a dynamic view. The movement of human traces, including travel purposes, can be calculated and represented on maps.

The arising of Big data, including LBSN data, also challenges the epistemology (Ekbia et al., 2015)and the research paradigm(Hey et al., 2009). The digitalization of human movements and thoughts embodies complex social relationships and spatial displacement. For example, the conspicuous observation of human movements(Wilson, 2012) ended the metaphysical exploration of the urban space and spatial relationship (see Chapter II). The particular spatial position and timestamp become accessible, and thus we can detect the spatiotemporal behaviors and measure various spatial relationships in geo-social space.

The dominant role of theoretical exploration in the research paradigm has given way to data-driven investigation. The literature review summarizes urban studies leveraging LBSN data in the past ten years, which shows an extensive range of urban issues that could cooperate with LBSN data, such as urban health, urban functions, and public perceptions. The cooperation of models and data empowers urban analysis to produce a plethora of academic researches. The geo-space has been successfully correlated with social space. However, these researches are mainly concentrated in the United States, which contributed to about half of the literature pool.

On the other hand, challenges and limitations of LBSN data have also been noticed by many scholars. The significance of delimitating bias and representativeness lies in clarifying the boundary of availability and effectiveness of LBNS data. The dissertation explores these problems in Chapter V, which provides a comprehensive comparison among related researches regarding four major domains of urban studies. LBSN dataset usually does not have general representativeness except some powerful mega datasets.

The demographic bias and uneven spatial distribution of LBSN data can be observed in different regions. For example, the degree of inequality of wealth tends to link to the density of LBSN data. It also implies that social inequality also shapes the generation of these data (Shelton et al., 2015), no matter in which scale we collect them. The spatial precision of LBSN data was alleviated due to the protection of privacy. User's preferences also impact the place where they choose to make a check-in. In general, LBSN data can yield a reliable result of human movements at an aggregated level in urban areas. In the sphere of semantic analysis, the limitation mainly comes from the analytical techniques and establishing proper relations between the results of contents and the corresponding contexts. In summary, these limitation reminds us that the proper research design is crucial for utilizing LBSN data.

The chapter of case studies investigates three different urban issues using three types of LBSN datasets, which explores innovative applications of LBSN data in urban studies. In the Weibo data case, the spatiotemporal variation of Weibo activities reflected how people occupied the urban space dynamically. The result indicated that the polycentric structure of the Beijing metropolitan area might be concealed by the high density of human activities.

The Foursquare case study calculated and confirmed the functional relationship between places in Barcelona. It is beneficial for understanding the connections among urban facilities regarding different groups of people. Such functional relations, essentially, is human mobility, and thus it offers a social explanation for places. Furthermore, this perspective also discloses that human activity plays the dominant role in the relationship between the built environment

and human activity. Places and connections between them can be constructed, managed, and recreated by human activities.

The Twitter data project demonstrates that the urban environment has little impact on public sentiment. Socio-economic profiles have a larger influence on Twitter sentiments. However, some specific environmental conditions, such as green space and commercial areas, are positively associated with Twitter sentiments. These findings are useful for informing the construction of smart cities because "smart cities are collections of numerous sentient and connected built environments, which possess components that learn from patterns of daily activity and adapt automatically to changes in such behaviours"(Kandt & Batty, 2021).

VII.2. Discussions and implications

2.1. The indulgence of data: whether LBSN data was over-used?

The widely use of LBSN data also raises the criticism of the indulgence of data (Wyly, 2014). Urban researchers can become very productive leveraging the abundant LBSN data while the causality behind the data is ignored. Wyly argued the phenomenon as "the speedy pseudopositivism of tweet-space analysis" that only can provide a shallow view of the real world. Without a doubt, LBSN data and other big data are not perfect and contain bias. It is important to recognize these limitations and the complex interaction between LBSN and society. It is also important to aware that the data do not explain anything unless people endow it.

However, firstly, there is no flawless and completeness dataset. Even the official data is just an authorized narrative. The official geographical data are not always more accurate than local and indigenous knowledge. For example, regarding cultural values and social communities, the "in-discriminated" official data are less accurate (Williams & Dunn, 2003)because indigenous knowledge cannot be summarized and standardized.

Secondly, there is a traditional prejudge that researchers tend to give a higher evaluation to thoughts than phenomena, though thoughts are subjective and full

of biases. In fact, the theoretical approach also cannot seize the so-called "nature" of the complex society because our thinking ability is not infinite. Social theories are just partially correct and limited in concrete circumstances.

Indeed, data itself is not of meaning, which is just an approach to cognize things. However, the description based on LBSN data is neither shallow nor deductive. The value of LBSN data lies in the user-generated scope that allows researchers to observe tangled human activities. Even those spammer messages that are usually considered as data noise, they play a vital role in monitoring criminal activities(Chakraborty, Pal, Pramanik, & Chowdary, 2016). The accumulation of such studies will mobilize to some stable conclusions and universal phenomena. Without doubts, these phenomena do not definitely construct some causal relations, as David Hume suggested hundreds of years ago. However, these phenomena from LBSN data could offer observable dynamic evidence for social hypothesises.

The real problem of LBSN data lies in the materialization of people, social relationship, and socio-spatial relationship, which is also the reason that we consider these data are valuable. As Shelton et al. (2015) summarized, "LBSN data is not only interesting for its ability to shed light on relatively mundane geographic processes; it is also being used to directly shape the way we live in cities today." These virtual information becomes a means of production into the economy. For example, the emerging of location-based services served for locating commercial goods at first. Nowadays, paid commenters have become a profession for generating fake data with particular contents. Data itself becomes a creator of capital, which could be separated from the physical world and daily life. Such a process will lead to the further aggregation of wealth and homogenization of urban daily life.

Therefore, for researcher and urban policymakers, the awareness of the nature of LBSN data is vital for avoid of the pseudopositivism of data analysis. Since it is impossible to go back to the pre-digitalized world, we have to develop new strategies to utilize and manage LBSN data for creating a liveable environment.

2.2. Discussion: toward citizen science and scientific urban planning

The initial proposal of scientific urban planning can be found in the description from Ford (1913): "Except on the aesthetic side, city planning is rapidly becoming as definite a science as pure engineering." In his opinion, the best group of urban planners are composed of" one an engineer, one an architect, and one, perhaps, a social expert." Such cooperation can provide a comprehensive view to enact a successful long-term city plan. He was half right. The correct prediction is the quantification of urban planning and multi-disciplinary cohesion. However, he might not anticipate that urban planning would become a dynamic "unfinished craft" (Ratti & Claudel, 2016)rather than a static framework.

Future urban planning is an ongoing process that can respond to the rapidly changing city. On the one hand, in the current, the speed of technological developments and mobility is faster than the change of socio-economic structures. Globalization is not only a political ideology but also a necessity that is driven by the productive force of technologies. As the regional surplus is increasing, domestic and international exchanges are inevitable. The degree of globalization is farther deeper than we assumed. The globally COVID-19 epidemic has proven it. On the other hand, the digitalization tends to cultivate a "belief in abstract quantity in every department of life"(Mumford, 1970). Everything seems to be evaluated by a series of numbers: work, amusement, physical exercises, art, and treasure. The power of the machine and commercial expansion seems to be put in a supreme position.

For responding to these changes and making a balance between data and real life, one approach is to back to daily life and particular communities. It is vital to realize that society is more like a mosaic continuum rather than a uniform system. The ultimate goal of a society does not pursue a faster systematic machine, instead, a model of comprehensive development that focuses on human beings and social diversity will be sustainable and long-lasting(Mumford, 1970).

Therefore, the appearance of **citizen science** will be an important perspective to understand and design future cities. According to *Green paper on Citizen Science*, "citizen science refers to the general public engagement in scientific research activities when citizens actively contribute to science either with their intellectual

effort or surrounding knowledge or with their tools and resources". The core concept is to consider all groups' interests and create a livable and sustainable environment. As to urban planning, policymakers and urban planners can utilize these citizen-contributed data to customize the planning according to the particular situation meanwhile keep the related construction standards of planning.

Regarding the role of LBSN data in future urban planning, it could be an effective tool for monitoring the urban dynamics and participating in urban planning. Researchers and urban planners can use LBSN data as a kind of social reference data to investigate various relationships in the urban space. Moreover, the indigenous knowledge provided by LBSN data is beneficial for disclosing the local daily life and the embedded social and cultural contexts. Moreover, it also can be an active instrument for urban management and development. The government can utilize social media platforms to communicate with citizens and manage public space intentionally. The following are some future urban applications in which LBSN data can involve.

Re-understanding urban structure

In the current, the detection of spatiotemporal urban structure has become a cliché to some degree. Many analyzing tools could easily visualize the dynamic urban structure. However, compared with the traditional urban structure based on employment data, the urban structure based on LBSN data shifts the focus from an economic perspective to a daily life perspective. It provides a new perspective to define the city, such as the "natural" city of B. Jiang and Miao (2015) that used the range of LBSN activities to define a city. Some researchers also use LBSN data to delimitate the city center(Sun et al., 2016). These findings could help us to understand the spatial displacement of people and their spatial interests. Just as our Beijing case study shows, the active areas of the Beijing metropolis are not only the employment center. Some active centers are non-employment centers, such as the university zone and amusement parks. Therefore, the traditional definition

based on employment data should be renewed by related outcomes using LBSN data.

Moreover, the semantic information can greatly enrich the information of human activities in urban space. For instance, Steiger, Westerholt, et al. (2015) found that tweets could be a reliable source to reflect the working-related activities in London. As remote work is increasing after the COVID-19 pandemic, LBSN data could be a more accurate source to describe the location of workplaces.

Urban functions and rearrangement of public facilities

Using POIs and users' check-ins to detect urban functions is one of the main applications of LBSN data. It could help policymakers to adjust and rearrangement public facilities and services, such as transport stations, culture centers, health facilities, etc. Meanwhile, comments on these facilities can be a good reference to improve these places.

Managing public space and creating sustainable city

By hotspots of LBSN check-ins, we can check the implementation effect of a plan and manage the public space. For example, Martí, García-Mayor, Nolasco-Cirugeda, and Serrano-Estrada (2020) compared green infrastructure elements between the current urban plan and Foursquare venues in Valencia, Spain. The result from Foursquare data could be a useful reference to check how people use and evaluate these green facilities. Moreover, LBSN data can also be an indicator of urban deprivation(Venerandi et al., 2015).

Culture and custom study

Culture and customs identify a city from another, which are the reflection of different geographical conditions and social backgrounds. However, they are usually hidden behind daily life and not easy to be quantified. Leveraging LBSN data, such as Instagram(Phan & Gatica-Perez, 2017) and Twitter(Fried et al., 2014), it can be easy to map the food habits and trending all over the world.

Tourism development

The digital records of tourists are distinct from locals, and thus tourist spatiotemporal analysis using LBSN data is able to display the moving patterns, consumption preferences, and the urban space that tourists are frequently occupied in a city. Moreover, recommendations of tourist attractions and related services have a great influence on tourist choices. Therefore, LBSN data can be incredibly for helping the government enact plans of tourism development.

Social segregation and inequality

Although part of the bias of LBSN data comes from the socio-economic and demographic conditions, it happens to provide a quantitative evidence of the social segregation and social inequality(D. R. Davis et al., 2019; Longley, Adnan, & Lansley, 2015; Shelton et al., 2015). In the *Cities of Tomorrow*, the ultimate goal of urban studies is to eliminate the economic inequality(P. Hall, 2014). In this sense, LBSN data can be dynamic evidence of these problems that prompt citizens and policymakers to make changes.

VII.3. Conclusion

According to United Nations Population Fund, the global urban population is increasing at the speed of 1.3 million per week till 2030[33]. Such a fast increment rate, combined with the higher speed of human mobility among cities and the development of various technologies, indicates a high expectation for urban planning and management. Despite the complex impetus of urbanization, essentially, people move into cities for a better living. Urban designers have to consider the interests of different groups of people, such as locals, visitors, and immigrants. Meanwhile, it is also necessary to consider the environmental sustainability of the city.

[33] https://www.unfpa.org/icpd/urbanization

For satisfying various needs, designers should not only focus on urban land management and control, but also understand the urban dynamics and the mechanism of urban growth that could be disclosed by particular events, mobility, communities, organizations, people.

This dissertation tries to summarize how LBSN data can be applied in studies of various urban issues. It reveals the potentials and limitations of LBSN data at the level of non-government research. LBSN data as a data bridge, connect social activities with geo-space. In the future, cooperating with a keen understanding of society and other datasets, LBSN data can create more possibilities for urban daily life and urban studies.

Appendix A. Completed table of regression model

The Moran I test shows that there is no evident spatial dependence in both model (**Table A 1**), which is based on weighting matrices of the queen-continuity. The spatial lag model still cannot improve the issue of heteroscedasticity(**Table A 3**). Therefore, we decide to keep results from the robust OLS model (**Table A 2**).

Table A 1 Moran I test for spatial dependence

Ln sentiment tweets Model	MI/DF	VALUE	PROB
Moran's I (error)	-0.0011	0.6209	0.53467
Lagrange Multiplier(lag)	1	4.6768	0.03057
Robust LM (lag)	1	11.2023	0.00082

Lagrange Multiplier(error)	1	0.0006	0.98091
Robust LM (error)	1	6.526	0.01063
Lagrange Multiplier(SARMA)	2	11.2029	0.00369
Net sentiment model			
Moran's I (error)	-0.0159	0.0714	0.94308
Lagrange Multiplier(lag)	1	0.8259	0.36345
Robust LM (lag)	1	9.8088	0.00174
Lagrange Multiplier(error)	1	0.091	0.76287
Robust LM (error)	1	9.0739	0.00259
Lagrange Multiplier(SARMA)	2	9.8998	0.00708

Source: own-elaboration

Table A 2 OLS model summary of sentiment tweets

Linear regression	Robust		Durbin-Watson	1.792			
Number of observation	200		R-squared	0.689			
F(10, 189)	35.07		Root MSE	0.760			
Prob > F	0						
	Coef.	Std. Err.	T	Sig.	[95% Conf. Interval]		VIF
	B				lower	upper	
Centrality index	0.112	0.031	3.650	0.000	0.051	0.173	1.950
Historical area(%)	1.110	0.256	4.350	0.000	0.607	1.615	1.240
LnD FS outdoor resort	0.029	0.009	3.470	0.001	0.012	0.045	1.130
LnD total population	0.371	0.089	4.170	0.000	0.195	0.546	1.440
PgGracia Avenue	1.136	0.399	2.840	0.005	0.348	1.923	1.360

El born historical area	0.975	0.253	3.850	0.000	0.475	1.475	1.190
Urban park garden	2.402	0.675	3.560	0.000	1.070	3.733	1.430
PC profession high	0.492	0.127	3.870	0.000	0.241	0.742	1.550
Contamination opinion	0.036	0.008	4.320	0.000	0.020	0.053	1.240
Density storefront	12.748	5.019	2.540	0.012	2.847	22.649	2.690
_Constant	-0.806	1.057	-0.760	0.447	-2.892	1.279	

Source: own-elaboration

Table A 3 Spatial-lag model summary of sentiment tweets

R-squared	0.699	Log likelihood	-220.531
Sq. Correlation	-	Akaike info criterion	465.062
Sigma-square	0.527	Schwarz criterion	504.642
S.E of regression	0.726		

Variable	Coefficient	Std.Error	z-value	Probability
W_Lnd_posneg	0.194	0.081	2.394	0.017
_Constant	-1.238	0.852	-1.453	0.146
Centrality index	0.094	0.034	2.775	0.006
Historical area(%)	0.917	0.303	3.029	0.002
LnD FS outdoor resort	0.026	0.008	3.235	0.001
LnD total population	0.345	0.070	4.902	0.000
PgGracia Avenue	1.126	0.352	3.199	0.001
El born historical area	0.759	0.244	3.110	0.002
Urban park garden	2.248	0.490	4.585	0.000
PC profession high	0.382	0.117	3.262	0.001

Contamination opinion	0.029	0.007	4.185	0.000
Density storefront	10.249	4.985	2.056	0.040
Diagnostics for Heteroscedasticity				
Random Coefficients				
Test	Df.	Value	Prob.	
Breusch-Pagan test	10.000	31.615	0.000	

Source: own-elaboration

Table A 4 Model summary of the net sentiments

Number of obs.	168			Durbin-Watson		2.008
F(6, 161)	11.391			R-squared		0.298
Prob > F	0			Adjusted R Square		0.272
				Root MSE		0.038
	Coef.	Std. Err.	T	Sig.	[95% Conf. Interval]	VIF
	B				lower upper	
Profession artisan	-0.004	0.0010	-4.690	0	-0.006 -0.002	1.617
%Railway area	-0.466	0.1550	-3.003	0.003	-0.773 -0.160	1.056
Density population	-0.448	0.1070	-4.206	0	-0.659 -0.238	1.726
Density storefront	0.729	0.230	3.170	0.002	0.275 1.183	1.849
% Commercial_st	0.001	0	2.603	0.010	0 0.001	1.763
_Constant	0.496	0.011	45.460	0	0.474 0.517	

Source: own-elaboration

References

Uncategorized References

Abbas, O. A. (2008). Comparisons Between Data Clustering Algorithms. *International Arab Journal of Information Technology (IAJIT), 5*(3).

Abbasi, O. R., & Alesheikh, A. A. (2018). Exploring the potential of location-based social networks data as proxy variables in collective human mobility prediction models. *Arabian Journal of Geosciences, 11*(8), 173.

Abdi, H. (2003). Partial least square regression (PLS regression). *Encyclopedia for research methods for the social sciences, 6*(4), 792-795.

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon, 4*(11), e00938.

Adrienko, N., & Adrienko, G. (2010). Spatial generalization and aggregation of massive movement data. *IEEE Transactions on visualization and computer graphics*, *17*(2), 205-219.

Ai, W., Zhuang, D., & Liu, Y. (2008). The variation of urban land use in Beijing in the last one hundred years. *Geo-information Science, 10*(4), 489-493.
Akhmad Nuzir, F., & Julien Dewancker, B. (2017). Dynamic land-use map based on twitter data. *Sustainability, 9*(12), 2158.
Albarran, A. B. (2013). *The social media industries*: Routledge.
Alhabash, S., & Ma, M. (2017). A tale of four platforms: Motivations and uses of Facebook, Twitter, Instagram, and Snapchat among college students? *Social media+ society, 3*(1), 2056305117691544.
Alqurashi, A. F., Kumar, L., & Sinha, P. (2016). Urban land cover change modelling using time-series satellite images: A case study of urban growth in five cities of Saudi Arabia. *Remote Sensing, 8*(10), 838.
AlSayyad, N., & Guvenc, M. (2015). Virtual uprisings: On the interaction of new social media, traditional media coverage and urban space during the 'Arab Spring'. *Urban Studies, 52*(11), 2018-2034.
Ames, M., & Naaman, M. (2007). *Why we tag: motivations for annotation in mobile and online media.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.
Amichai-Hamburger, Y., & Vinitzky, G. (2010). Social network use and personality. *Computers in human behavior, 26*(6), 1289-1295.
Andrade, R., Alves, A., & Bento, C. (2020). POI Mining for Land Use Classification: A Case Study. *ISPRS International Journal of Geo-Information, 9*(9), 493.
Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical analysis, 27*(2), 93-115.
Asgari, F., Gauthier, V., & Becker, M. (2013). A survey on human mobility and its applications. *arXiv preprint arXiv:1307.0814*.
Bagci, H., & Karagoz, P. (2016). *Context-aware friend recommendation for location based social networks using random walk.* Paper presented at the Proceedings of the 25th international conference companion on world wide web.
Ballas, D. (2013). What makes a 'happy city'? *Cities, 32*, S39-S50.
Barbagallo, D., Bruni, L., Francalanci, C., & Giacomazzi, P. (2012). An empirical study on the relationship between twitter sentiment and influence in the tourism domain. In *Information and Communication Technologies in Tourism 2012* (pp. 506-516): Springer.
Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review, 33*(6), 712-729.
Barnard, H. (2008). Maps and mapmaking in Ancient Egypt. *Encyclopaedia of the History of Science, Technology, and Medicine in Non-Western Cultures*, 1273-1276.
Becker, H., Naaman, M., & Gravano, L. (2011). *Beyond trending topics: Real-world event identification on twitter.* Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media.
Béjar Alonso, J. (2014). Mining frequent spatio-temporal patterns from location based social networks. Retrieved from https://upcommons.upc.edu/bitstream/handle/2117/24313/SpatioTemporal.pdf
Béjar, J., Álvarez, S., García, D., Gómez, I., Oliva, L., Tejeda, A., & Vázquez-Salceda, J. (2016). Discovery of spatio-temporal patterns from location-based social networks. *Journal of Experimental & Theoretical Artificial Intelligence, 28*(1-2), 313-329.

Berelson, B. (1949). What'missing the newspaper'means. *Communications Research 1948-1949*, 111-129.

Bertrand, K. Z., Bialik, M., Virdee, K., Gros, A., & Bar-Yam, Y. (2013). Sentiment in new york city: A high resolution spatial and temporal view. *arXiv preprint arXiv:1308.5010*.

Blanchflower, D. G., & Oswald, A. J. (2008). Is well-being U-shaped over the life cycle? *Social science & medicine, 66*(8), 1733-1749.

Blanford, J. I., Huang, Z., Savelyev, A., & MacEachren, A. M. (2015). Geo-located tweets. Enhancing mobility maps and capturing cross-border movement. *PloS one, 10*(6), e0129202.

Blanzieri, E., & Bryl, A. (2008). A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review, 29*(1), 63-92.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Blight, M. G., Ruppel, E. K., & Schoenbauer, K. V. (2017). Sense of community on Twitter and Instagram: Exploring the roles of motives and parasocial relationships. *Cyberpsychology, Behavior, and Social Networking, 20*(5), 314-319.

Bosch, O. J., & Revilla, M. (2020). Using emojis in mobile web surveys for Millennials? A study in Spain and Mexico. *Quality & Quantity*, 1-23.

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210-230.

Brereton, F., Clinch, J. P., & Ferreira, S. (2008). Happiness, geography and the environment. *Ecological economics, 65*(2), 386-396.

Burns, M. C., Cladera, J. R., & Bergada, M. M. (2008). The spatial implications of the functional proximity deriving from air passenger flows between European metropolitan urban regions. *GeoJournal, 71*(1), 37-52.

Burt, J. E., Barber, G. M., & Rigby, D. L. (2009). *Elementary statistics for geographers*: Guilford Press.

Cai, J., Huang, B., & Song, Y. (2017). Using multi-source geospatial big data to identify the structure of polycentric cities. *Remote Sensing of Environment, 202*, 210-221.

Cambria, E., Wang, H., & White, B. (2014). Guest editorial: Big social data analysis. *Knowledge-Based Systems, 69*(1), 1-2.

Candelieri, A., & Archetti, F. (2015). *Detecting Events and Sentiment on Twitter for Improving Urban Mobility*. Paper presented at the ESSEM@ AAMAS.

Cantril, H., & Allport, G. W. (1935). The psychology of radio.

Cao, X., MacNaughton, P., Deng, Z., Yin, J., Zhang, X., & Allen, J. G. (2018). Using twitter to better understand the spatiotemporal patterns of public sentiment: A case study in Massachusetts, USA. *International journal of environmental research and public health, 15*(2), 250.

Carr, C. T., & Hayes, R. A. (2015). Social media: Defining, developing, and divining. *Atlantic journal of communication, 23*(1), 46-65.

Castelló, P. T., Baeza, J. L., & García, C. P. (2018). USE OF APPLICATIONS WITH GEOREFERENCED CONTACTS 'DATING APPS'TO IDENTIFY CREATIVE AREAS. *WIT Transactions on The Built Environment, 179*, 197-207.

Chakraborty, M., Pal, S., Pramanik, R., & Chowdary, C. R. (2016). Recent developments in social spam detection and combating techniques: A survey. *Information Processing & Management, 52*(6), 1053-1073.

Chapman, L., Resch, B., Sadler, J., Zimmer, S., Roberts, H., & Petutschnig, A. (2018). Investigating the emotional responses of individuals to urban green space using twitter data: A critical comparison of three different methods of sentiment analysis. *Urban Planning, 3*(1), 21-33.

Chen, G. M. (2011). Tweet this: A uses and gratifications perspective on how active Twitter use gratifies a need to connect with others. *Computers in human behavior, 27*(2), 755-762.

Chen, L., Hu, N., Shu, C., & Chen, X. (2019). Adult attachment and self-disclosure on social networking site: A content analysis of Sina Weibo. *Personality and Individual Differences, 138*, 96-105.

Chen, T., Hui, E. C., Wu, J., Lang, W., & Li, X. (2019). Identifying urban spatial structure and urban vibrancy in highly dense cities using georeferenced social media data. *Habitat International, 89*, 102005.

Cho, E., Myers, S. A., & Leskovec, J. (2011). *Friendship and mobility: user movement in location-based social networks.* Paper presented at the Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.

Collins, C., Hasan, S., & Ukkusuri, S. V. (2013). A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. *Journal of Public Transportation, 16*(2), 2.

Couclelis, H. (1999). Space, time, geography. In *Geographical information systems* (Vol. 1, pp. 29-38): John Wiley & Sons.

Cox, K. R. (2013). *Making human geography*: Guilford Publications.

Cramer, H., Rost, M., & Holmquist, L. E. (2011). *Performing a check-in: emerging practices, norms and'conflicts' in location-sharing using foursquare.* Paper presented at the Proceedings of the 13th international conference on human computer interaction with mobile devices and services.

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). *The livehoods project: Utilizing social media to understand the dynamics of a city.* Paper presented at the International AAAI Conference on Weblogs and Social Media.

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). # Earthquake: Twitter as a distributed sensor system. *Transactions in GIS, 17*(1), 124-147.

Culotta, A. (2010). *Towards detecting influenza epidemics by analyzing Twitter messages.* Paper presented at the Proceedings of the first workshop on social media analytics.

Culotta, A. (2014). *Estimating county health statistics with twitter.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Da Rugna, J., Chareyron, G., & Branchet, B. (2012). *Tourist behavior analysis through geotagged photographies: a method to identify the country of origin.* Paper presented at the 2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI).

Daggitt, M. L., Noulas, A., Shaw, B., & Mascolo, C. (2016). Tracking urban activity growth globally with big location data. *Royal Society open science, 3*(4), 150688.

Davidson, J., Smith, M. M., & Bondi, L. (2012). *Emotional geographies*: Ashgate Publishing, Ltd.

Davis, D. R., Dingel, J. I., Monras, J., & Morales, E. (2019). How segregated is urban consumption? *Journal of Political Economy, 127*(4), 1684-1738.

Davis, F. D. (1993). User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. *International journal of man-machine studies, 38*(3), 475-487.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). *Predicting depression via social media.* Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media.

De Choudhury, M., Lin, Y.-R., Sundaram, H., Candan, K. S., Xie, L., & Kelliher, A. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media? *Icwsm, 10*, 34-41.

De Saussure, F. (2011). *Course in general linguistics*: Columbia University Press.

De Ureña, J. M., Pillet, F., & Marmolejo-Duarte, C. (2013). Aglomeraciones/regiones urbanas basadas en varios centros: el policentrismo. *Ciudad y Territorio Estudios Territoriales (CyTET), 45*(176), 249-266.

Diddi, A., & LaRose, R. (2006). Getting hooked on news: Uses and gratifications and the formation of news habits among college students in an Internet environment. *Journal of broadcasting & electronic media, 50*(2), 193-210.

Dou, X., Arellano Ramos, B., & Roca Cladera, J. (2018). China's inter-provincial population flow based on the interaction value analysis. *Geographical Research= Dili yanjiu, 37*(9), 1848-1861.

Einstein, A. (1920). *The Theory of Relativity*. London: Methuen and Co.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., . . . Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology, 66*(8), 1523-1545.

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion review, 3*(4), 364-370.

Epstein, J. M. (2008). Why model? *Journal of artificial societies and social simulation, 11*(4), 12.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise.* Paper presented at the Kdd.

Fan, C., Esparza, M., Dargin, J., Wu, F., Oztekin, B., & Mostafavi, A. (2020). Spatial biases in crowdsourced data: Social media content attention concentrates on populous areas in disasters. *Computers, Environment and Urban Systems, 83*, 101514.

Feng, J., & Zhou, Y.-x. (2003). The social spatial structure of Beijing Metropolitan Area and its evolution: 1982–2000. *Geographical research, 4*, 008.

Ferreira, A. P. G., Silva, T. H., & Loureiro, A. A. F. (2015). *Beyond sights: Large scale study of tourists' behavior using foursquare data.* Paper presented at the 2015 IEEE International Conference on Data Mining Workshop (ICDMW).

Ford, G. B. (1913). The city scientific. *Engineering Record, 67*(May), 551-552.

Frias-Martinez, V., & Frias-Martinez, E. (2014). Spectral clustering for sensing urban land use using Twitter activity. *Engineering Applications of Artificial Intelligence, 35*, 237-245.

Fried, D., Surdeanu, M., Kobourov, S., Hingle, M., & Bell, D. (2014). *Analyzing the language of food on social media.* Paper presented at the 2014 IEEE International Conference on Big Data (Big Data).

Frigui, S., Rouibah, K., & Marzocchi, G. L. (2013). CROSS-CULTURAL COMPARISON BETWEEN ARABIC AND WESTERN COUNTRIES IN LOCATION-BASED SOCIAL NETWORKING USAGE ON MOBILE PHONES: THE CASE OF FACEBOOK. *Issues in Information Systems, 14*(2).

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience, 21*(11), 1129-1164.

Fuchs, C. (2017). *Social media: A critical introduction*: Sage.

Gallegos, L., Lerman, K., Huang, A., & Garcia, D. (2016). *Geography of Emotion: Where in a City are People Happier?* Paper presented at the Proceedings of the 25th International Conference Companion on World Wide Web.

Gan, C. (2018). Gratifications for using social media: A comparative analysis of Sina Weibo and WeChat in China. *Information development, 34*(2), 139-147.

Gao, Q., Abel, F., Houben, G.-J., & Yu, Y. (2012). *A comparative study of users' microblogging behavior on Sina Weibo and Twitter*. Paper presented at the International Conference on User Modeling, Adaptation, and Personalization.

Gao, S., Janowicz, K., & Couclelis, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS, 21*(3), 446-467.

Garcia-López, M.-À., & Muñiz, I. (2010). Employment decentralisation: Polycentricity or scatteration? The case of Barcelona. *Urban Studies, 47*(14), 3035-3056.

García-Palomares, J. C., Gutiérrez, J., & Mínguez, C. (2015). Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography, 63*, 408-417.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta, 185*, 1-17.

Getis, A., & Ord, J. K. (2010). The analysis of spatial association by use of distance statistics. In *Perspectives on spatial data analysis* (pp. 127-145): Springer.

Ghosh, D., & Guha, R. (2013). What are we 'tweeting'about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science, 40*(2), 90-102.

Gilbert, C., & Hutto, E. (2014). *Vader: A parsimonious rule-based model for sentiment analysis of social media text.* Paper presented at the Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf.

Girardin, F., Calabrese, F., Dal Fiore, F., Ratti, C., & Blat, J. (2008). Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive computing, 7*(4), 36-43.

Girardin, F., Calabrese, F., Dal Fiorre, F., Biderman, A., Ratti, C., & Blat, J. (2008). *Uncovering the presence and movements of tourists from user-generated content.* Paper presented at the Intn'l Forum on Tourism Statistics.

Giuliano, G., & Small, K. A. (1991). Subcenters in the Los Angeles region. *Regional science and urban economics, 21*(2), 163-182.

Gleason, B. (2013). # Occupy Wall Street: Exploring informal learning about a social movement on Twitter. *American Behavioral Scientist, 57*(7), 966-982.
Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *nature, 453*(7196), 779-782.
Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal, 69*(4), 211-221.
Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial statistics, 1*, 110-120.
Graham, M., Zook, M., & Boulton, A. (2013). Augmented reality in urban places: contested content and the duplicity of code. *Transactions of the Institute of British Geographers, 38*(3), 464-479.
Graser, A., Schmidt, J., Roth, F., & Brändle, N. (2019). Untangling origin-destination flows in geographic information systems. *Information Visualization, 18*(1), 153-172.
Green, N. (2007). Functional polycentricity: A formal definition in terms of social network analysis. *Urban Studies, 44*(11), 2077-2103.
Griffith, D. B., Amrhein, C. G., & Desloges, J. R. (1991). *Statistical Analysis for Geographers* (J. R. Desloges Ed.): Prentice Hall.
Gruebner, O., Rapp, M. A., Adli, M., Kluge, U., Galea, S., & Heinz, A. (2017). Cities and mental health. *Deutsches Ärzteblatt International, 114*(8), 121.
Gwena, C., Chinyamurindi, W. T., & Marange, C. (2018). Motives influencing Facebook usage as a social networking site: An empirical study using international students. *Acta Commercii, 18*(1), 1-11.
Haffner, M., Mathews, A. J., Fekete, E., & Finchum, G. A. (2018). Location-based social media behavior and perception: Views of university students. *Geographical Review, 108*(2), 203-224.
Hägerstrand, T. (1970). What about people in regional science? *Papers in regional science, 24*(1), 7-24.
Hägerstrand, T. (1989). *Reflections on "what about people in regional science?".* Paper presented at the Papers of the Regional Science Association.
Hall, C., & Page, S. (2003). *Managing Urban Tourism*.
Hall, P. (2014). *Cities of tomorrow: an intellectual history of urban planning and design since 1880*: John Wiley & Sons.
Hargittai, E., & Litt, E. (2011). The tweet smell of celebrity success: Explaining variation in Twitter adoption among a diverse group of young adults. *New media & society, 13*(5), 824-842.
Haridakis, P., & Hanson, G. (2009). Social interaction and co-viewing with YouTube: Blending mass communication reception and social connection. *Journal of broadcasting & electronic media, 53*(2), 317-335.
Harris, C. D., & Ullman, E. L. (1945). The nature of cities. *The Annals of the American Academy of Political and Social Science, 242*(1), 7-17.
Harvey, D. (1973). *Social Justice and the City* (REV - Revised ed.): University of Georgia Press.
Hasan, S., & Ukkusuri, S. V. (2015). Location contexts of user check-ins to model urban geo life-style patterns. *PloS one, 10*(5), e0124819.

Hasnat, M. M., & Hasan, S. (2018). Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. *Transportation Research Part C: Emerging Technologies, 96*, 38-54.
Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science, 41*(3), 260-271.
Haworth, B., & Bruce, E. (2015). A review of volunteered geographic information for disaster management. *Geography Compass, 9*(5), 237-250.
Hecht, B. J., & Stephens, M. (2014). A Tale of Cities: Urban Biases in Volunteered Geographic Information. *Icwsm, 14*(14), 197-205.
Heipke, C. (2010). Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing, 65*(6), 550-557.
Helliwell, J. F. (2003). How's life? Combining individual and national variables to explain subjective well-being. *Economic modelling, 20*(2), 331-360.
Hestenes, D. (1997). *Modeling methodology for physics teachers.* Paper presented at the AIP conference proceedings.
Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1): Microsoft research Redmond, WA.
Hiruta, S., Yonezawa, T., Jurmu, M., & Tokuda, H. (2012). *Detection, classification and visualization of place-triggered geotagged tweets.* Paper presented at the Proceedings of the 2012 ACM conference on ubiquitous computing.
Hochman, N., & Schwartz, R. (2012). *Visualizing instagram: Tracing cultural visual rhythms.* Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media.
Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. *Online readings in psychology and culture, 2*(1), 2307-0919.1014.
Hollenstein, L., & Purves, R. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science, 2010*(1), 21-48.
Horita, F. E. A., Degrossi, L. C., de Assis, L. F. G., Zipf, A., & de Albuquerque, J. P. (2013, August 15-17). *The use of volunteered geographic information (VGI) and crowdsourcing in disaster management: a systematic literature review.* Paper presented at the Proceedings of Nineteenth Americas Conference on Information Systems, Chicago, Illinois.
Horton, F. E., & Reynolds, D. R. (1970). Action space formation: a behavioral approach to predicting urban travel behavior. *Highway Research Record, 322*, 136-148.
Hossain, M., Kim, M., & Jahan, N. (2019). Can “liking” behavior lead to usage intention on Facebook? Uses and gratification theory perspective. *Sustainability, 11*(4), 1166.
Hoyt, H. (1939). *The structure and growth of residential neighborhoods in American cities*: US Government Printing Office.
Hu, X., Li, H., & Bao, X. (2017). *Urban population mobility patterns in Spring Festival Transportation: Insights from Weibo data.* Paper presented at the 2017 International Conference on Service Systems and Service Management.
Hu, Y., Deng, C., & Zhou, Z. (2019). A semantic and sentiment analysis on online neighborhood reviews for understanding the perceptions of people toward their

living environments. *Annals of the American Association of Geographers, 109*(4), 1052-1073.

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems, 54*, 240-254.

Hu, Y., McKenzie, G., Janowicz, K., & Gao, S. (2015). *Mining human-place interaction patterns from location-based social networks to enrich place categorization systems.* Paper presented at the Proceedings of the workshop on cognitive engineering for spatial information processes at COSIT.

Huang, D., Liu, Z., & Zhao, X. (2015). Monocentric or polycentric? The urban spatial structure of employment in Beijing. *Sustainability, 7*(9), 11632-11656.

Huang, Q., & Wong, D. W. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science, 30*(9), 1873-1898.

Hubert, R. B., Estevez, E., Maguitman, A., & Janowski, T. (2018). *Examining government-citizen interactions on Twitter using visual and sentiment analysis.* Paper presented at the Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age.

Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in human behavior, 28*(2), 561-569.

Hürlimann, M., Davis, B., Cortis, K., Freitas, A., Handschuh, S., & Fernández, S. (2016). *A twitter sentiment gold standard for the brexit referendum.* Paper presented at the Proceedings of the 12th international conference on semantic systems.

Jackson, S. P., Mullen, W., Agouris, P., Crooks, A., Croitoru, A., & Stefanidis, A. (2013). Assessing completeness and spatial error of features in volunteered geographic information. *ISPRS International Journal of Geo-Information, 2*(2), 507-530.

James, M. L., Wotring, C. E., & Forrest, E. J. (1995). An exploratory study of the perceived benefits of electronic bulletin board use and their impact on other communication activities. *Journal of broadcasting & electronic media, 39*(1), 30-50.

Jendryke, M., Balz, T., & Liao, M. (2017). Big location-based social media messages from China's Sina Weibo network: Collection, storage, visualization, and potential ways of analysis. *Transactions in GIS, 21*(4), 825-834.

Jessop, B., Brenner, N., & Jones, M. (2008). Theorizing sociospatial relations. *Environment and planning D: society and space, 26*(3), 389-401.

Jiang, B., & Miao, Y. (2015). The evolution of natural cities from the perspective of location-based social media. *The Professional Geographer, 67*(2), 295-306.

Jiang, S., Alves, A., Rodrigues, F., Ferreira Jr, J., & Pereira, F. C. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems, 53*, 36-46.

Jiang, S., Ferreira Jr, J., & Gonzalez, M. C. (2012). *Discovering urban spatial-temporal structure from human activity patterns.* Paper presented at the Proceedings of the ACM SIGKDD international workshop on urban computing.

Jiang, W., Wang, Y., Tsou, M.-H., & Fu, X. (2015). Using social media to detect outdoor air pollution and monitor air quality index (AQI): a geo-targeted spatiotemporal

analysis framework with Sina Weibo (Chinese Twitter). *PloS one, 10*(10), e0141185.
Jin, P. J., Cebelak, M., Yang, F., Zhang, J., Walton, C. M., & Ran, B. (2014). Location-based social networking data: Exploration into use of doubly constrained gravity model for origin–destination estimation. *Transportation Research Record, 2430*(1), 72-82.
Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR), 50*(5), 1-22.
Juris, J. S. (2012). Reflections on# Occupy Everywhere: Social media, public space, and emerging logics of aggregation. *American ethnologist, 39*(2), 259-279.
Kádár, B. (2014a). Measuring tourist activities in cities using geotagged photography. *Tourism Geographies, 16*(1), 88-104.
Kádár, B. (2014b). Pedestrian space usage of tourist-historic cities: comparing the tourist space systems of Vienna and Prague to Budapest.
Kandt, J., & Batty, M. (2021). Smart cities, big data and urban policy: Towards urban analytics for the long run. *Cities, 109*, 102992.
Kannangara, S., Xie, H., Tanin, E., Harwood, A., & Karunasekera, S. (2020). *Tracking Group Movement in Location Based Social Networks.* Paper presented at the Proceedings of the 28th International Conference on Advances in Geographic Information Systems.
Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons, 53*(1), 59-68.
Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2018). Advances in social media research: Past, present and future. *Information Systems Frontiers, 20*(3), 531-558.
Katz, E., Blumler, J. G., & Gurevitch, M. (1973). Uses and gratifications research. *The public opinion quarterly, 37*(4), 509-523.
Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis* (Vol. 344): John Wiley & Sons.
Kershaw, D., Rowe, M., & Stacey, P. (2014). *Towards tracking and analysing regional alcohol consumption patterns in the UK through the use of social media.* Paper presented at the Proceedings of the 2014 ACM conference on Web science.
Khan, M. L. (2017). Social media engagement: What motivates user participation and consumption on YouTube? *Computers in human behavior, 66*, 236-247.
Kircaburun, K., Alhabash, S., Tosuntaş, Ş. B., & Griffiths, M. D. (2020). Uses and gratifications of problematic social media use among university students: A simultaneous examination of the Big Five of personality traits, social media platforms, and social media use motives. *International Journal of Mental Health and Addiction, 18*(3), 525-547.
Kitchin, R., & Dodge, M. (2011). *Code/space: Software and everyday life*: Mit Press.
Ko, H., Cho, C.-H., & Roberts, M. S. (2005). Internet uses and gratifications: A structural equation model of interactive advertising. *Journal of advertising, 34*(2), 57-70.
Kolodziej, K. W., & Hjelm, J. (2017). *Local positioning systems: LBS applications and services*: CRC press.

Koncz, N. A., & Adams, T. M. (2002). A data model for multi-dimensional transportation applications. *International Journal of Geographical Information Science, 16*(6), 551-569.

Korgaonkar, P. K., & Wolin, L. D. (1999). A multivariate analysis of web usage. *Journal of advertising research, 39*(2), 53-53.

Kossinets, G. (2006). Effects of missing data in social networks. *Social networks, 28*(3), 247-268.

Kossinets, G., & Watts, D. J. (2009). Origins of homophily in an evolving social network. *American journal of sociology, 115*(2), 405-450.

Kovacs-Györi, A., Ristea, A., Kolcsar, R., Resch, B., Crivellari, A., & Blaschke, T. (2018). Beyond spatial proximity—classifying parks and their visitors in London based on spatiotemporal and sentiment analysis of Twitter data. *ISPRS International Journal of Geo-Information, 7*(9), 378.

Kraak, M.-J. (2003). *The space-time cube revisited from a geovisualization perspective.* Paper presented at the Proc. 21st International Cartographic Conference.

Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(3), 231-240.

Kruskal, W., & Mosteller, F. (1980). Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review/Revue Internationale de Statistique*, 169-195.

Kuang, W. (2012). Spatio-temporal patterns of intra-urban land use change in Beijing, China between 1984 and 2008. *Chinese Geographical Science, 22*(2), 210-220.

Kuipers, B. (1978). Modeling spatial knowledge. *Cognitive science, 2*(2), 129-153.

Kulin, H. W., & Kuenne, R. E. (1962). An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics. *Journal of Regional Science, 4*(2), 21-33.

Kumar, A., & Jaiswal, A. (2020). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience, 32*(1), e5107. doi:10.1002/cpe.5107

Kwan, M.-P. (1999). Gender, the home-work link, and space-time patterns of nonemployment activities. *Economic geography, 75*(4), 370-394.

Kwan, M. P. (2004). GIS methods in time-geographic research: Geocomputation and geovisualization of human activity patterns. *Geografiska Annaler: Series B, Human Geography, 86*(4), 267-280.

LaRose, R., Mastro, D., & Eastin, M. S. (2001). Understanding Internet usage: A social-cognitive approach to uses and gratifications. *Social Science Computer Review, 19*(4), 395-413.

Larsen, M. E., Boonstra, T. W., Batterham, P. J., O'Dea, B., Paris, C., & Christensen, H. (2015). We feel: mapping emotion on Twitter. *IEEE journal of biomedical and health informatics, 19*(4), 1246-1252.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., . . . Gutmann, M. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY), 323*(5915), 721.

LeFebvre, R. K., & Armstrong, C. (2018). Grievance-based social movement mobilization in the# Ferguson Twitter storm. *New media & society, 20*(1), 8-28.

Lei, C., Zhang, A., Qi, Q., Su, H., & Wang, J. (2018). Spatial-temporal analysis of human dynamics on urban land use patterns using social media data by gender. *ISPRS International Journal of Geo-Information, 7*(9), 358.

Leighton, G. R., Hugo, P. S., Roulin, A., & Amar, A. (2016). Just Google it: assessing the use of Google Images to describe geographical variation in visible traits of organisms. *Methods in Ecology and Evolution, 7*(9), 1060-1070.

Lenormand, M., Picornell, M., Cantú-Ros, O. G., Tugores, A., Louail, T., Herranz, R., . . . Ramasco, J. J. (2014). Cross-checking different sources of mobility information. *PloS one, 9*(8), e105184.

Leszczynski, A. (2015). Spatial media/tion. *Progress in Human Geography, 39*(6), 729-751.

Li, B., Kuang, H., Zhang, Y., Chen, J., & Tang, X. (2012). *Using similes to extract basic sentiments across languages.* Paper presented at the International Conference on Web Information Systems and Mining.

Li, J., Zeng, F., Xiao, Z., Jiang, H., Zheng, Z., Liu, W., & Ren, J. (2020). Drive2friends: Inferring Social Relationships from Individual Vehicle Mobility Data. *IEEE Internet of Things Journal*.

Li, M., Ch'ng, E., Chong, A., & See, S. (2016). *The new eye of smart city: novel citizen sentiment analysis in twitter.* Paper presented at the 2016 International Conference on Audio, Language and Image Processing (ICALIP).

Li, X., Zhang, C., & Li, W. (2017). Building block level urban land-use information retrieval based on Google Street View images. *GIScience & Remote Sensing, 54*(6), 819-835.

Li, Y., Steiner, M., Wang, L., Zhang, Z.-L., & Bao, J. (2013). *Exploring venue popularity in foursquare.* Paper presented at the 2013 Proceedings IEEE INFOCOM.

Lien, C. H., & Cao, Y. (2014). Examining WeChat users' motivations, trust, attitudes, and positive word-of-mouth: Evidence from China. *Computers in human behavior, 41*, 104-111.

Lima, S., Perez, N., Cuadros, M., & Rigau, G. (2020). NUBES: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. *arXiv preprint arXiv:2004.01092*.

Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., & Zimmerman, J. (2011). *I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application.* Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Liu, I. L., Cheung, C. M., & Lee, M. K. (2010). Understanding Twitter Usage: What Drive People Continue to Tweet. *Pacis, 92*, 928-939.

Ljajić, A., & Marovac, U. (2019). Improving sentiment analysis for twitter data by handling negation rules in the Serbian language. *Computer Science and Information Systems, 16*(1), 289-311.

Long, Y., Han, H., Tu, Y., & Shu, X. (2015). Evaluating the effectiveness of urban growth boundaries using human mobility and activity records. *Cities, 46*, 76-84.

Long, Y., & Liu, X. (2013). Featured graphic. How mixed is Beijing, China? A visual exploration of mixed land use. *Environment and Planning A, 45*(12), 2797-2798.

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A, 47*(2), 465-484.

López-Ornelas, E., & Zaragoza, N. M. (2015). *Social media participation: A narrative way to help urban planners.* Paper presented at the International Conference on Social Computing and Social Media.

Lucas, K., & Sherry, J. L. (2004). Sex differences in video game play: A communication-based explanation. *Communication Research, 31*(5), 499-523.

Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Applied Geography, 70*, 11-25.

Luo, W., Wang, Y., Liu, X., & Gao, S. (2019). Cities as spatial and social networks: towards a spatio-socio-semantic analysis framework. In *Cities as Spatial and Social Networks* (pp. 21-37): Springer.

Lupton, D. (1998). *The emotional self: A sociocultural exploration*: Sage.

Lynch, K. (1960). *The image of the city* (Vol. 11): MIT press.

Malik, M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). *Population bias in geotagged tweets.* Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media.

Manasijević, D., Živković, D., Arsić, S., & Milošević, I. (2016). Exploring students' purposes of usage and educational usage of Facebook. *Computers in human behavior, 60*, 441-450.

Manca, M., Boratto, L., Roman, V. M., i Gallissà, O. M., & Kaltenbrunner, A. (2017). Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study. *Online Social Networks and Media, 1*, 56-69.

Manovich, L. (2016). Notes on Instagrammism and mechanisms of contemporary cultural identity (and also photography, design, Kinfolk, k-pop, hashtags, mise-en-scène, and состояние). *Sl: Disponible en: http://manovich. net/index. php/projects/instagram-and-contemporary-image*.

Manuel, K., Indukuri, K. V., & Krishna, P. R. (2010). *Analyzing internet slang for sentiment mining.* Paper presented at the 2010 second Vaagdevi international conference on information Technology for Real World Problems.

Marmolejo-Duarte, C., & Cerda-Troncoso, J. (2012). La densidad-tiempo: otra perspectiva de análisis de la estructura metropolitana. *Scripta Nova, 16*.

Marmolejo-Duarte, C., & Cerda-Troncoso, J. (2020). Metropolitan Barcelona 2001–06, or how people's spatial–temporal behaviour shapes urban structures. *Regional Studies, 54*(4), 563-575.

Marmolejo-Duarte, C., Echavarría Ochoa, J. C., & Biere Arenas, R. (2016). El valor de la centralidad: un análisis para la Barcelona Metropolitana. *ACE: Architecture, city and Environment, 11*(32), 95-112.

Marmolejo-Duarte, C., Núñez, C. A., & Roca Cladera, J. (2013). Revisión de la densidad de empleo como medio para detectar sub-centros metropolitanos: un análisis para Barcelona y Madrid. *ACE: Architecture, city and Environment, 8*(23), 33-64.

Marmolejo, C., & Cerda Troncoso, J. (2017). Spatiotemporal behavior of the population as an approach to analyze urban structure: the case of Metropolitan Barcelona. *Cuadernos Geográficos, 56*(2), 111-133.

Martí, P., García-Mayor, C., Nolasco-Cirugeda, A., & Serrano-Estrada, L. (2020). Green infrastructure planning: Unveiling meaningful spaces through Foursquare users' preferences. *Land Use Policy, 97*, 104641.

Martí, P., García-Mayor, C., & Serrano-Estrada, L. (2020). Taking the urban tourist activity pulse through digital footprints. *Current Issues in Tourism*, 1-20.

Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2019). Social media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems, 74*, 161-174.

Matei, S., Ball-Rokeach, S. J., & Qiu, J. L. (2001). Fear and misperception of Los Angeles urban space: A spatial-statistical study of communication-shaped mental maps. *Communication Research, 28*(4), 429-463.

Maynard, D. G., & Greenwood, M. A. (2014). *Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis.* Paper presented at the LREC 2014 Proceedings.

McDonald, J. F., & Prather, P. J. (1994). Suburban employment centres: The case of Chicago. *Urban Studies, 31*(2), 201-218.

Mckercher, B., & Lau, G. (2008). Movement patterns of tourists within a destination. *Tourism Geographies, 10*(3), 355-374.

McMillen, D. P., & Lester, T. W. (2003). Evolving subcenters: employment and population densities in Chicago, 1970–2020. *Journal of Housing Economics, 12*(1), 60-81.

Meier, H. E., Mutz, M., Glathe, J., Jetzke, M., & Hölzen, M. (2019). Politicization of a Contested Mega Event: The 2018 FIFA World Cup on Twitter. *Communication & Sport*, 2167479519892579.

Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics, 4*(3), 2053168017720008.

Miller, H. J. (1991). Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information System, 5*(3), 287-301.

Miller, H. J. (2005). A measurement theory for time geography. *Geographical analysis, 37*(1), 17-45.

Miller, H. J., & Bridwell, S. A. (2009). A field-based theory for time geography. *Annals of the Association of American Geographers, 99*(1), 49-75.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the demographics of twitter users. *Icwsm, 11*(5th), 25.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one, 8*(5), e64417.

Molz, J. G. (2005). Guilty pleasures of the Golden Arches: mapping McDonald's in narratives of round-the-world travel. *Emotional geographies*, 63-76.

Morstatter, F., Pfeffer, J., & Liu, H. (2014). *When is it biased? Assessing the representativeness of twitter's streaming API.* Paper presented at the Proceedings of the 23rd international conference on world wide web.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204*.

Mueller, W., Silva, T. H., Almeida, J. M., & Loureiro, A. A. (2017). Gender matters! Analyzing global cultural gender preferences for venues using social sensing. *EPJ Data Science, 6*(1), 5.

Mukhina, K., Visheratin, A., & Nasonov, D. (2020). *Spatiotemporal Filtering Pipeline for Efficient Social Networks Data Processing Algorithms.* Paper presented at the International Conference on Computational Science.

Mumford, L. (1970). *The Culture of Cities* (Third ed.): Harcourt Brace Jovanovich.

Murthy, D., Gross, A., & Pensavalle, A. (2016). Urban social media demographics: An exploration of Twitter use in major American cities. *Journal of Computer-Mediated Communication, 21*(1), 33-49.

Nagel, A. C., Tsou, M.-H., Spitzberg, B. H., An, L., Gawron, J. M., Gupta, D. K., . . . Lindsay, S. (2013). The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. *Journal of medical Internet research, 15*(10), e237.

Narayanan, M., & Cherukuri, A. K. (2016). A study and analysis of recommendation systems for location-based social network (LBSN) with big data. *IIMB Management Review, 28*(1), 25-30.

Narayanaperumal, M. (2020). *Deep Neural Networks for Sentiment Analysis in Tweets with Emoticons.* (Doctoral dissertation). Nova Southeastern University, Retrieved from https://nsuworks.nova.edu/gscis_etd/1117

Nasukawa, T., & Yi, J. (2003). *Sentiment analysis: Capturing favorability using natural language processing.* Paper presented at the Proceedings of the 2nd international conference on Knowledge capture.

Neppalli, V. K., Caragea, C., Squicciarini, A., Tapia, A., & Stehle, S. (2017). Sentiment analysis during Hurricane Sandy in emergency response. *International journal of disaster risk reduction, 21*, 213-222.

Neutens, T., Schwanen, T., & Witlox, F. (2011). The prism of everyday life: Towards a new research agenda for time geography. *Transport reviews, 31*(1), 25-47.

Nik-Bakht, M., & El-Diraby, T. E. (2016). Sus-tweet-ability: Exposing public community′ s perspective on sustainability of urban infrastructure through online social media. *International Journal of Human-Computer Studies, 89*, 54-72.

Niu, H., & Silva, E. A. (2020). Crowdsourced Data Mining for Urban Activity: Review of Data Sources, Applications, and Methods. *Journal of Urban Planning and Development, 146*(2), 04020007.

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). *Exploiting semantic annotations for clustering geographic areas and users in location-based social networks.* Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media.

Nov, O., Naaman, M., & Ye, C. (2010). Analysis of participation in an online photo-sharing community: A multidimensional perspective. *Journal of the American society for information science and technology, 61*(3), 555-566.

O'cass, A., & Fenech, T. (2003). Web retailing adoption: exploring the nature of internet users Web retailing behaviour. *Journal of Retailing and Consumer services, 10*(2), 81-94.

Ogneva, M. (2010, 2010/04/09). How companies can use sentiment analysis to improve their business. . Retrieved from http://mashable.com/2010/04/19/sentiment-analysis/

Omar, B., & Dequan, W. (2020). Watch, share or create: The influence of personality traits and user motivation on TikTok mobile video usage.

Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis, 27*(4), 286-306.

Oriol, B.-J. (2020). Using emojis in mobile web surveys for millennials? A study in Spain and Mexico. *Quality and Quantity*.

Padilla, J. J., Kavak, H., Lynch, C. J., Gore, R. J., & Diallo, S. Y. (2018). Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter. *PloS one, 13*(6), e0198857.

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). *Sentiment analysis of Twitter data for predicting stock market movements.* Paper presented at the 2016 international conference on signal processing, communication, power and embedded system (SCOPES).

Pak, A., & Paroubek, P. (2010). *Twitter as a corpus for sentiment analysis and opinion mining.* Paper presented at the LREc.

Pak, I., Teh, P. L., & Cheah, Y. N. (2018). Hidden sentiment behind letter repetition in online reviews. *Journal of Telecommunication, Electronic and Computer Engineering, 10*(3-2), 115-120.

Palmgreen, P., & Rayburn, J. D. (1979). Uses and gratifications and exposure to public television: A discrepancy approach. *Communication Research, 6*(2), 155-179.

Papacharissi, Z., & Rubin, A. M. (2000). Predictors of Internet use. *Journal of broadcasting & electronic media, 44*(2), 175-196.

Park, S. B., Kim, H. J., & Ok, C. M. (2018). Linking emotion and place on Twitter at Disneyland. *Journal of Travel & Tourism Marketing, 35*(5), 664-677.

Paul, D., Li, F., Teja, M. K., Yu, X., & Frost, R. (2017). *Compass: Spatio temporal sentiment analysis of US election what twitter says!* Paper presented at the Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining.

Paul, M., & Dredze, M. (2011). *You are what you tweet: Analyzing twitter for public health.* Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media.

Perse, E. M., & Dunn, D. G. (1998). The utility of home computers and media use: Implications of multimedia and connectivity. *Journal of broadcasting & electronic media, 42*(4), 435-456.

Peters, C., Amato, C. H., & Hollenbeck, C. R. (2007). An exploratory investigation of consumers' perceptions of wireless advertising. *Journal of advertising, 36*(4), 129-145.

Phan, T.-T., & Gatica-Perez, D. (2017). *Healthy# fondue# dinner: analysis and inference of food and drink consumption patterns on instagram.* Paper presented at the Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia.

Plunz, R. A., Zhou, Y., Vintimilla, M. I. C., Mckeown, K., Yu, T., Uguccioni, L., & Sutto, M. P. (2019). Twitter sentiment in New York City parks as measure of well-being. *Landscape and urban planning, 189*, 235-246.

Pred, A. (1981). Social reproduction and the time-geography of everyday life. *Geografiska Annaler. Series B, Human Geography, 63*(1), 5-22.

Preoţiuc-Pietro, D., & Cohn, T. (2013). *Mining user behaviours: a study of check-in patterns in location based social networks.* Paper presented at the Proceedings of the 5th annual ACM web science conference.

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2012). *Tracking" gross community happiness" from tweets.* Paper presented at the Proceedings of the ACM 2012 conference on computer supported cooperative work.

Quercia, D., Schifanella, R., & Aiello, L. M. (2014). *The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city.* Paper presented at the Proceedings of the 25th ACM conference on Hypertext and social media.

Raper, J. (2005). Spatio-temporal ontology for digital geographies. *Re-presenting GIS*, 199.

Ratti, C., & Claudel, M. (2016). *The city of tomorrow: Sensors, networks, hackers, and the future of urban life*: Yale University Press.

Rauschnabel, P. A., Sheldon, P., & Herzfeldt, E. (2019). What motivates users to hashtag on social media? *Psychology & Marketing, 36*(5), 473-488.

Rickards, L., Gleeson, B., Boyle, M., & O'Callaghan, C. (2016). Urban studies after the age of the city. *Urban Studies, 53*(8), 1523-1541.

Rizwan, M., Wan, W., & Gwiazdzinski, L. (2020). Visualization, Spatiotemporal Patterns, and Directional Analysis of Urban Activities Using Geolocation Data Extracted from LBSN. *ISPRS International Journal of Geo-Information, 9*(2), 137.

Roca Cladera, J., Marmolejo Duarte, C. R., & Moix, M. (2009). Urban structure and polycentrism: Towards a redefinition of the sub-centre concept. *Urban Studies, 46*(13), 2841-2868.

Roca Cladera, J., & Moix Bergadà, M. (2005). The interaction value: its scope and limits as an instrument for delimiting urban systems. *Regional Studies, 39*(3), 357-373.

Roick, O., & Heuser, S. (2013). L ocation B ased S ocial N etworks–Definition, Current State of the Art and Research Agenda. *Transactions in GIS, 17*(5), 763-784.

Rom, E., & Alfasi, Y. (2014). The role of adult attachment style in online social network affect, cognition, and behavior. *Journal of Psychology, 1*(1), 24-34.

Rondan-Cataluña, F. J., Arenas-Gaitán, J., & Ramírez-Correa, P. E. (2015). A comparison of the different versions of popular technology acceptance models. *Kybernetes*.

Rubin, R. B., Perse, E. M., & Barbato, C. A. (1988). Conceptualization and measurement of interpersonal communication motives. *Human Communication Research, 14*(4), 602-628.

Ruggiero, T. E. (2000). Uses and gratifications theory in the 21st century. *Mass communication & society, 3*(1), 3-37.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346*(6213), 1063-1064.

Sagl, G., Resch, B., Hawelka, B., & Beinat, E. (2012). *From social sensor data to collective human behaviour patterns: Analysing and visualising spatio-temporal dynamics in urban environments.* Paper presented at the Proceedings of the GI-Forum.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: real-time event detection by social sensors.* Paper presented at the Proceedings of the 19th international conference on World wide web.

Samani, Z. N., Karimi, M., & Alesheikh, A. (2020). Environmental and infrastructural effects on respiratory disease exacerbation: a LBSN and ANN-based spatio-temporal modelling. *Environmental Monitoring and Assessment, 192*(2), 1-17.

Sampietro, A. (2016). *Emoticonos y emojis. Análisis de su historia, difusión y uso en la comunicación digital actual.* (Doctor). Universitat de València,

Santos, F. A., Silva, T. H., Loureiro, A. A., & Villas, L. A. (2020). Automatic extraction of urban outdoor perception from geolocated free texts. *Social Network Analysis and Mining, 10*(1), 1-23.

Saura, J. R., Reyes-Menendez, A., & Palos-Sanchez, P. (2019). Are black Friday deals worth it? Mining Twitter users' sentiment and behavior response. *Journal of Open Innovation: Technology, Market, and Complexity, 5*(3), 58.

Scholtes, I. (2017). *When is a network a network? Multi-order graphical model selection in pathways and temporal networks.* Paper presented at the Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining.

Schouten, A. P., Valkenburg, P. M., & Peter, J. (2009). An experimental test of processes underlying self-disclosure in computer-mediated communication. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 3*(2).

Schwartz, A. J., Dodds, P. S., O'Neil-Dunne, J. P., Danforth, C. M., & Ricketts, T. H. (2019). Visitors to urban greenspace have higher sentiment and lower negativity on Twitter. *People and Nature, 1*(4), 476-485.

Schwitzguébel, A. C., & Bartomeus, O. R. (2018). Location-based social network data for exploring spatial and functional urban tourists and residents consumption patterns. *ARA: Journal of Tourism Research/Revista de Investigación Turística, 8*(2), 32-52.

Seidman, G. (2013). Self-presentation and belonging on Facebook: How personality influences social media use and motivations. *Personality and Individual Differences, 54*(3), 402-407.

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology, 52*(6), 1061.

Sheldon, P., & Bryant, K. (2016). Instagram: Motives for its use and relationship to narcissism and contextual age. *Computers in human behavior, 58*, 89-97.

Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and urban planning, 142*, 198-211.

Shih, H.-P. (2004). Extended technology acceptance model of Internet utilization behavior. *Information & management, 41*(6), 719-729.

Silva, T. H., Vaz de Melo, P. O., Almeida, J. M., Salles, J., & Loureiro, A. A. (2013). *A comparison of Foursquare and Instagram to the study of city dynamics and urban social behavior.* Paper presented at the Proceedings of the 2nd ACM SIGKDD international workshop on urban computing.

Silva, T. H., Viana, A. C., Benevenuto, F., Villas, L., Salles, J., Loureiro, A., & Quercia, D. (2019). Urban computing leveraging location-based social network data: a survey. *ACM Computing Surveys (CSUR), 52*(1), 1-39.

Sinn, D., & Syn, S. Y. (2014). Personal documentation on a social network site: Facebook, a collection of moments from your life? *Archival Science, 14*(2), 95-124.

Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., & Merchant, R. M. (2017). Twitter as a tool for health research: a systematic review. *American journal of public health, 107*(1), e1-e8.

Smart, M. W. (1974). Labour market areas: uses and definition. *Progress in planning, 2*, 239-353.

Smock, A. D., Ellison, N. B., Lampe, C., & Wohn, D. Y. (2011). Facebook as a toolkit: A uses and gratification approach to unbundling feature use. *Computers in human behavior, 27*(6), 2322-2329.

Snijders, C., Matzat, U., & Reips, U.-D. (2012). " Big Data": big gaps of knowledge in the field of internet science. *International journal of internet science, 7*(1), 1-5.

Srivastava, K. (2009). Urbanization and mental health. *Industrial psychiatry journal, 18*(2), 75.

Standage, T. (2013). *Writing on the wall: Social media-The first 2,000 years*: Bloomsbury Publishing USA.

Steiger, E., De Albuquerque, J. P., & Zipf, A. (2015). An advanced systematic literature review on spatiotemporal analyses of t witter data. *Transactions in GIS, 19*(6), 809-834.

Steiger, E., Westerholt, R., Resch, B., & Zipf, A. (2015). Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems, 54*, 255-265.

Stephens, M., & Poorthuis, A. (2015). Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems, 53*, 87-95.

Stringhini, G., Kruegel, C., & Vigna, G. (2010). *Detecting spammers on social networks.* Paper presented at the Proceedings of the 26th annual computer security applications conference.

Sun, Y. (2016). Investigating “locality” of intra-urban spatial interactions in New York city using foursquare data. *ISPRS International Journal of Geo-Information, 5*(4), 43.

Sun, Y., Fan, H., Li, M., & Zipf, A. (2016). Identifying the city center using human travel flows generated from location-based social networking data. *Environment and Planning B: Planning and Design, 43*(3), 480-498.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics, 37*(2), 267-307.

Tasse, D., Liu, Z., Sciuto, A., & Hong, J. (2017). *State of the geotags: Motivations and recent changes.* Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American society for information science and technology, 61*(12), 2544-2558.

Thomas, K., Grier, C., Song, D., & Paxson, V. (2011). *Suspended accounts in retrospect: an analysis of twitter spam.* Paper presented at the Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference.

Thompson, N., Wang, X., & Daya, P. (2019). Determinants of news sharing behavior on social media. *Journal of Computer Information Systems*.

Thrift, N., & Pred, A. (1981). Time-geography: a new beginning. *Progress in Human Geography, 5*(2), 277-286.

Thuillier, E., Moalic, L., Lamrous, S., & Caminada, A. (2017). Clustering weekly patterns of human mobility through mobile phone data. *IEEE Transactions on Mobile Computing, 17*(4), 817-830.

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods*, *21*(3), 525-547.

Tsou, M.-H., Yang, J.-A., Lusher, D., Han, S., Spitzberg, B., Gawron, J. M., . . . An, L. (2013). Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election. *Cartography and Geographic Information Science*, *40*(4), 337-348.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400*.

Ullah, H., Wan, W., Haidery, S. A., Khan, N. U., Ebrahimpour, Z., & Muzahid, A. (2020). Spatiotemporal Patterns of Visitors in Urban Green Parks by Mining Social Media Big Data Based Upon WHO Reports. *IEEE Access*, *8*, 39197-39211.

Ullah, M. A., Marium, S. M., Begum, S. A., & Dipa, N. S. (2020). An algorithm and method for sentiment analysis using the text and emoticon. *ICT Express*, *6*(4), 357-360.

Utz, S., Tanis, M., & Vermeulen, I. (2012). It is all about being popular: The effects of need for popularity on social network site use. *Cyberpsychology, Behavior, and Social Networking*, *15*(1), 37-42.

Valls Dalmau, F. (2019). *Digital traces and urban research: Barcelona through social media data.* Universitat Politècnica de Catalunya,

Venerandi, A., Quattrone, G., Capra, L., Quercia, D., & Saez-Trumper, D. (2015). *Measuring urban deprivation from user generated content.* Paper presented at the Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478.

Vu, H. Q., Li, G., Law, R., & Ye, B. H. (2015). Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. *Tourism Management*, *46*, 222-232.

Vytal, K., & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *Journal of cognitive neuroscience*, *22*(12), 2864-2885.

Wang, B., Zubiaga, A., Liakata, M., & Procter, R. (2015). Making the most of tweet-inherent features for social spam detection on Twitter. *arXiv preprint arXiv:1503.07405*.

Wang, F., Wang, G., & Philip, S. Y. (2014). *Why checkins: Exploring user motivation on location based social networks.* Paper presented at the 2014 IEEE International Conference on Data Mining Workshop.

Wang, G., Schoenebeck, S. Y., Zheng, H., & Zhao, B. Y. (2016). *" Will Check-in for Badges": Understanding Bias and Misbehavior on Location-Based Social Networks.* Paper presented at the Icwsm.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). *A system for real-time twitter sentiment analysis of 2012 us presidential election cycle.* Paper presented at the Proceedings of the ACL 2012 system demonstrations.

Wang, W., Hernandez, I., Newman, D. A., He, J., & Bian, J. (2016). Twitter analysis: Studying US weekly trends in work stress and emotion. *Applied Psychology*, *65*(2), 355-378.

Wang, W. Y. (2020). Mapping Cantonese: The Pro-Cantonese Protest and Sina Weibo in Guangzhou. *Handbook of the Changing World Language Map*, 201-213.

Wang, Y., Wang, T., Tsou, M.-H., Li, H., Jiang, W., & Guo, F. (2016). Mapping dynamic urban land use patterns with crowdsourced geo-tagged social media (Sina-Weibo) and commercial points of interest collections in Beijing, China. *Sustainability, 8*(11), 1202.

Whiting, A., & Williams, D. (2013). Why people use social media: a uses and gratifications approach. *Qualitative Market Research: An International Journal*.

Williams, C., & Dunn, C. E. (2003). GIS in Participatory Research: Assessing the Impact of Landmines on Communities in North-west Cambodia. *Transactions in GIS, 7*(3), 393-410.

Wilson, M. W. (2012). Location-based services, conspicuous mobility, and the location-aware future. *Geoforum, 43*(6), 1266-1275.

Wohlin, C. (2014). *Guidelines for snowballing in systematic literature studies and a replication in software engineering.* Paper presented at the Proceedings of the 18th international conference on evaluation and assessment in software engineering.

Wu, L., Morstatter, F., & Liu, H. (2018). SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation, 52*(3), 839-852.

Wu, Q., Li, H.-q., Wang, R.-s., Paulussen, J., He, Y., Wang, M., . . . Wang, Z. (2006). Monitoring and predicting land use change in Beijing using remote sensing and GIS. *Landscape and urban planning, 78*(4), 322-333.

Wyly, E. (2014). The new quantitative revolution. *Dialogues in Human Geography, 4*(1), 26-38.

Xia, C., Hu, J., Zhu, Y., & Naaman, M. (2015). *What is new in our city? a framework for event extraction using social media posts.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining.

Xiao, S., Jiaqi, Y., & Fuji, R. (2016). Detecting influenza states based on hybrid model with personal emotional factors from social networks. *Neurocomputing, 210*, 257-268.

Xie, Y., Fang, C., Lin, G., Gong, H., & Qiao, B. (2007). Tempo-spatial patterns of land use changes and urban development in globalizing China: a study of Beijing. *Sensors, 7*(11), 2881-2906.

Yadav, S., Ekbal, A., Saha, S., & Bhattacharyya, P. (2018). *Medical sentiment analysis using social media: towards building a patient assisted system.* Paper presented at the Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Yang, D., Zhang, D., & Qu, B. (2016). Participatory cultural mapping based on collective behavior data in location-based social networks. *ACM Transactions on Intelligent Systems and Technology (TIST), 7*(3), 1-23.

Yaqub, U., Sharma, N., Pabreja, R., Chun, S. A., Atluri, V., & Vaidya, J. (2020). Location-based Sentiment Analyses and Visualization of Twitter Election Data. *Digital Government: Research and Practice, 1*(2), 1-19.

Ye, Z., Hashim, N. H., Baghirov, F., & Murphy, J. (2018). Gender differences in Instagram hashtag use. *Journal of Hospitality Marketing & Management, 27*(4), 386-404.

Yoo, J., Choi, S., Choi, M., & Rho, J. (2014). Why people use Twitter: social conformity and social value perspectives. *Online Information Review*.

Yuan, J., Zheng, Y., & Xie, X. (2012). *Discovering regions of different functions in a city using human mobility and POIs.* Paper presented at the Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.

Yuan, Y., Wei, G., & Lu, Y. (2018). Evaluating gender representativeness of location-based social media: A case study of Weibo. *Annals of GIS, 24*(3), 163-176.

Zhang, S., Liu, X., Tang, J., Cheng, S., & Wang, Y. (2019). Urban spatial structure and travel patterns: Analysis of workday and holiday travel using inhomogeneous Poisson point process models. *Computers, Environment and Urban Systems, 73*, 68-84.

Zhang, Z., Zhou, L., Zhao, X., Wang, G., Su, Y., Metzger, M., . . . Zhao, B. Y. (2013). *On the validity of geosocial mobility traces.* Paper presented at the Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks.

Zheng, X., Zeng, Z., Chen, Z., Yu, Y., & Rong, C. (2015). Detecting spammers on social networks. *Neurocomputing, 159*, 27-34.

Zheng, Y. (2011). Location-based social networks: Users. In *Computing with spatial trajectories* (pp. 243-276): Springer.

Zheng, Y., Capra, L., Wolfson, O., & Yang, H. (2014). Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST), 5*(3), 1-55.

Zhu, Z., Blanke, U., & Tröster, G. (2014). *Inferring travel purpose from crowd-augmented human mobility data.* Paper presented at the Proceedings of the first international conference on IoT in urban space.