International Journal of Geo-Information

MDPI

Article

Analyzing Social-Geographic Human Mobility Patterns Using Large-Scale Social Media Data

Zeinab Ebrahimpour [1,2,*], Wanggen Wan [1,2], José Luis Velázquez García [3], Ofelia Cervantes [4] and Li Hou [5]

1 School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China; wanwg@staff.shu.edu.cn
2 Institute of Smart City, Shanghai University, Shanghai 200444, China
3 Department of Computer Science, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla 72840, Mexico; joseluisvzg@inaoep.mx
4 Department of Computing, Electronics, and Mechatronics, Universidad de las Américas Puebla, Puebla 72810, Mexico; ofelia.cervantes@udlap.mx
5 School of Information Engineering, Huangshan University, Huangshan 245041, China; houli@shu.edu.cn
* Correspondence: z_ebrahimpour@shu.edu.cn; Tel.: +86-151-210-40290

Received: 17 December 2019; Accepted: 20 February 2020; Published: 21 February 2020

Abstract: Social media data analytics is the art of extracting valuable hidden insights from vast amounts of semi-structured and unstructured social media data to enable informed and insightful decision-making. Analysis of social media data has been applied for discovering patterns that may support urban planning decisions in smart cities. In this paper, Weibo social media data are used to analyze social-geographic human mobility in the CBD area of Shanghai to track citizen's behavior. Our main motivation is to test the validity of geo-located Weibo data as a source for discovering human mobility and activity patterns. In addition, our goal is to identify important locations in people's lives with the support of location-based services. The algorithms used are described and the results produced are presented using adequate visualization techniques to illustrate the detected human mobility patterns obtained by the large-scale social media data in order to support smart city planning decisions. The outcome of this research is helpful not only for city planners, but also for business developers who hope to extend their services to citizens.

Keywords: human mobility; location-based social network; geographic mobility patterns

1. Introduction

Modeling human mobility in a city is tightly related to geographical patterns and spatial distributions. Understanding individual movements brings useful insights for a variety of applications, such as urban planning [1], security [2], migration studies [3], disease spread [4], traffic prediction (transportation planning) [5], tourism [6], and recommender systems [7]. Researchers have tried to use surveys [8] from travel or tourist centers in a traditional way to study mobility patterns; however, thanks to new technologies, finding a dataset to analyze people's mobility is not a big concern anymore. Recently, a significant effort has been made with different types of datasets, including phone call records (CDR) [9], WiFi or RFID [10,11], global positioning system (GPS) [12] and location-based social network (LBSN) [13] data, in order to obtain useful information from geographical movements. In this area, researchers have tried to tackle various questions: Does human mobility follow any model or pattern? Is it possible to extract significant patterns to define mobility models? Is it possible to estimate trajectory via home-to-work commutes? Do these trajectory patterns depend on the initial geographical position of individuals?

Traditional survey records are based on questions and answers which require human resources and a lot of time to obtain and to be analyzed statistically. On the other hand, GPS or CDR data have privacy limitations and lack a large amount of data. More recently, the explosive use of social media with geo-locations shared by users launched new research fields using location-based social networks (LBSNs) like Twitter, Facebook, Foursquare, or the Chinese microblog Sina Weibo, revealing unique opportunities to track mobility behaviors. These user-generated data provide extensive knowledge in three dimensions, namely, space, time, and textual content. Then, human mobility research can be accomplished using a different and complementary perspective, depending on the goal of the analysis.

With the rapid improvement in technology and widespread use of smart-phones in contrast to other sources of data, more and more people use applications to share their moments with others. This huge amount of check-in data can be considered as big data. It contains not only geo-tagged data, but also people's social profiles, enabling researchers to conduct a comprehensive study on geographical positions, mobility behavior, social connections or travel demands. Moreover, the check-in data from social networking sites are publicly available in big sizes. For example, the Chinese Sina Weibo microblog allows users to send texts, photos, and videos while sharing their location online. According to the Weibo industry research and development center [14], in the first half of 2018, the number of mobile microblog users in China was about 316 million, showing an increase of 29.23 million compared with the end of 2017. On the other hand, recent findings in [15] have proven that China is one of the most rapidly developing countries, thus city planners have focused more on implementing urban planning designs for spatial urban pattern development. Constructing a bridge between urban infrastructure layout and planning is a momentous objective. Therefore, studying urban spatial schemes has gained huge attention, aiming to design impactful commercial business districts (CBD) in smart and sustainable cities. The European regional cohesion policy determines capacity, quality, viability timeline and history and landscape as four dimensions of the objective of achieving all aspects of urban life through expert planning in smart metropolitan areas. Focusing on metropolitan spatial structures, city planners play an important role in monitoring and managing a better and safer mobility, topography, economic development, consumer preferences, population growth, and sustainable transportation, such that all these factors together aim to improve the quality of inhabitant's lives. Using technologies as an emerging tool to analyze the spatio-temporal aspects of human mobility in newly developed cities such as Shanghai helps the government to better understand citizen's needs and the heartbeat of movement in order to provide efficient traffic flow between movements due to supply and demand. To achieve this aim, it is important to identify the frequently congested regions in CBD areas and the reasons these areas are congested. Understanding citizen's demands helps managers to make better decisions to spread and balance different types of services in all regions in order to maximize movement flow, optimize resources, and reduce congestion. Researchers have offered different descriptions of commercial central districts. McColl et al. [16] described an area with a main centralization of business land use as CBD, with a similar point of view to Drozdz et al. [17], who pointed out that a significant CBD area contains enormous financial activities. Following [18], we agreed that a CBD area contains a dense distribution of commercial resources and has high-density road networks. Every open data source point of interest (POI) that represents an entity in the geographical space is a reflection of the CBD.

Our main motivation here is to test the validity of geo-located Weibo data as a means of elucidating human mobility and activity patterns. In addition, we aim to identify important places in people's lives with the support of location-based services in order to improve understanding of general human movement patterns and support the creation of realistic and practical models of human mobility. We also tried to adopt a geo-tagged Weibo dataset in order to analyze the behavioral characteristics of users from two perspectives, namely, their activity time and check-in point of interest types.

Despite the different values of check-in social media data, there are some issues that affect the quality of research in this field. When using social networks, an Internet connection is required to allow users to share their posts and generate spatio-temporal data, using the geolocation from where

the user posts the message and the moment when they did. However, what if the user does not have an Internet connection at the specific posting time? The post will not have a tagged geo-location and time stamp and will not be useful for data analytics. In addition, a fake check-in happens when a user shares a location which is above the venue threshold of the actual location. Due to various reasons, a user may share a post claiming they are "enjoying food in a restaurant", but in fact, they have made the post staying at home. Thus, fake check-in records should be excluded from the analysis.

In this paper, we aim to analyze human mobility in the city of Shanghai from spatio-temporal big data extracted from the Chinese microblog Sina Weibo. As a major city of China, Shanghai was chosen as the case study of this research applying several data analytics algorithms to validate the use of a one-year period of Weibo data as a means of discovering human mobility patterns. The data available from the Sina Weibo dataset were enriched with a POI dataset to associate a more meaningful interpretation of the geo-location of each Weibo post. Several questions were posed, such as: How does one find out significant locations in people's lives as an important aspect of characterizing human mobility? Is there an impact due to gender differences? How does one discover and describe mobility trajectories?

The remainder of this paper is organized as follows. In Section 2, relevant research work related to social media data analytics is presented. Section 3 gives a brief introduction of our case study and the dataset used in our analysis. In activity-based mobility studies, the generation of origin-destination (OD) matrices for analyzing movement behavior is essential. People's movement is explored from an intra-urban perspective to generate OD travel matrices to identify trips from home to workplaces. In Section 4, we explain the details of the cleaning data in order to identify accurate OD matrices. Section 5 presents the methodology applied and provides a detailed description of the algorithms implemented to detect trip patterns, discover the most popular visited places, explore the temporal dimension, and estimate trajectories through the calculation of origin–destination travel matrices. In particular, different groups of people (male and female) were considered in the analysis of 11 different districts of Shanghai. The results obtained are visualized using adequate techniques to support decision-making.

2. Related Work

Human mobility refers to the movement of an individual or a group of people from their origin geolocation to their destination. Recently, people's mobility has become a hot topic not only for the academic community for fundamental research purposes, but has also as an important research field for policymakers in smart cities in order to provide better services for citizens (e.g., in urban planning, public health, transportation). Human mobility has become an interesting topic for different research fields, ranging from computer science to social science [19] and geographical science [20]. The concept can be applied to areas of study such as urban planning, decision-making, migration, epidemic control, and transportation services. Understanding human movement becomes a crucial research question, and the purpose behind it is to integrate different data sources available for a variety of topics. Different topics can be obtained and analyzed from two perspectives of spatio-temporal analysis.

Liu et al. [20] worked on taxi trajectory data and introduced a number of point pair vectors. They analyzed the population distribution assuming that each trip has two points: a pick-up point which was considered as the origin and a drop-off point considered as destination. They applied the Monte Carlo method to Shanghai data to validate their model. In another study, Liang et al. [5] showed that the average population who visits a zone in a city is a good data source to model human mobility and the resident population. Thus, they calculated the daily population visiting a zone, using a radiation model to predict the flows of crowds. These two research works focused on the spatial distribution of flows, however, to highlight a difference with another research, Wu et al. [21] calculated intra-urban human mobility by combining activity-based and movement-based approaches and focused more on temporal variation. Compared to previous studies, Wu et al. [21] used temporal properties to find out whether people's activities were in a fixed area or not and introduced two different types of activities, namely, locationally mandatory activities (LMA) and locationally stochastic

activities (LSA). Since agent-based modeling (ABM) techniques are good for capturing actions in a time series, Bonabeau et al., [18] introduced different applications using ABM related to dynamic human behavior patterns in the real world. The most important fields for the application of flow studies include traffic, human mobility management, and evacuation. The importance of this field is revealed when the number of crowd disasters (disasters caused by crowds) rises. A suitable system for supporting transport system planning based on people's movement analysis would present the transportation network and simulate an individual's activity, origin, the destination of trips, and timing properly. A further step in this system would be to predict how the changes in people's behavior and activity could affect trips and movement in cities. Another interesting application of human mobility analysis is for city management. Carpooling has been introduced in order to tackle some transportation problems in big cities, with the goal of sharing part of the ride or expenses with other users. Human mobility analysis supports the implementation of mobility management policies to control the number of vehicles on the road. Furthermore, it reduces traffic and CO2 emissions, and, thus, air pollution. Another interesting study was published by Cezar [22], who worked on GPS data to find the routine trips in Pisa [23] and applied the density-based clustering method to social media data to find out the number of users in different events, like cultural and religious conferences or festivals, and detect which ones attract a large number of people.

Another useful data source to analyze people's movement is call detailed record (CDR) data. Cellular network providers collect the call records and text messages of each person to provide better services. Each CDR data contain the duration and location of each call, as well as associated text messages if they are available. These data can be used to obtain the location of the owner of a cell phone and, according to several surveys, CDR data are considered a better data source. For instance, Isaacman et al. [24] used CDR data to confirm individuals' movements that can be detected by analyzing the footprint left by the user in the places they called from. They chose Hartigan's leader algorithm [25] to cluster the cell towers, which were first sorted in a descending order. The reason they chose this algorithm is that it does not require the definition of the number of clusters in the analysis. Since the distance among cell towers in suburban areas is about 1 to 3 miles, while in urban areas it is approximately 200 meters, they found that one mile is an appropriate equalization of the two kinds of area. Moreover, logistic regression was applied to determine the likelihood of the importance of a cluster.

Nowadays, due to the easy use of social media services like Foursquare, Twitter, Facebook, and Sina Weibo, especially on phones, they have become very popular and have attracted the attention of researchers. There are billions of people who use social media to share their knowledge, ideas, daily life routines and photos while sharing their location. These location-based services (LBS) not only attract people's attention towards sharing their experiences but also allow researchers to take advantage of this significant source of data for analyzing people's behavior. Cheng et al. [26] collected check-in records from 220,000 users from different sources of location sharing services (LSS) to analyze aspects of spatio-temporal and textual mobility patterns. They proposed different ideas based on the density of the datasets in New York City, Los Angeles, and Amsterdam. They found out that human mobility is influenced by economic and geographical constraints and sentiment analysis of text messages enriched understanding of their research. Ullah et al. [27] also analyzed spatio-temporal data from LBSNs to show the impact of people in green spaces. Yan et al. [28] investigated three different datasets, including Sina Weibo data, to analyse individual's decision-making regarding the places they tend to go and the influence of the economic aspects of crowds on hot spot destinations. They proved that the gravity model is well-suited to predicting the effects of mobility on destinations. They believe the results are applicable for decision-making policies. In another research, Liu et al. [29] focused on examining the mobility differences between four different communities of Wuhan City, based on Sina Weibo data. They classified communities according to their check-in activities at specific areas and investigated their spatio-temporal behavior in six groups of categories. Wang et al. [30] analyzed the relationship between human trajectory and Weibo access locations, additionally combining mobile

phone records. They first classified the locations visited by users based on phone records and then found out the most frequently visited places by the Weibo distribution data. They concluded that time and speed patterns are beneficial for the classification of trajectories. In a similar work, Chao et al. [31] studied Sina Weibo data to analyse students' activities on campus. Using the Chinese University of Geosciences Wuhan (CUG Wuhan) community, as a case study, they found out the influence of distance on students' spatio-temporal mobility patterns. Differing by the gender of users, they made a comparison of the results. They discovered that the community, in their research, tended to make more Weibo posts related to entertainment activities. In another work, Hasan et al. [32] analyzed a Twitter dataset to categorize the spatio-temporal dimension of mobility patterns. Categorizing people's activity was addressed in this research, illustrating the frequency of visited popular places. Kernel density estimation (KDE) [33] was used to find out the distribution of activities over three major cities in the U.S. by splitting the temporal aspect into 3 hour intervals. They discovered that people choose their destination based on the popularity of visited places, not randomly. Moreover, there is a greater likelihood of making a destination based on the influence of other users that have selected the same place before and shared their experiences through social media. One year later, Hasan et al. [34] worked more on social media check-in data from Twitter to classify individual activity patterns. They applied the latent dirichlet allocation (LDA) [35] algorithm to find the distribution of specific words in tweets, that represent home and working places.

Yilan et al. [36] proposed a three-step methodology. First, in order to find the unique venues, they applied a distance-based clustering algorithm on 'days' of the check-in records instead of the total number of user's check-in. They used a distance-based clustering algorithm to discover the unique venues, then filtered the important clusters by frequently visited places in total days, and finally find out the 'work' and 'home' places, which were labeled based on check-in records and POI data. A gravity model provides an estimate of the volume of flow, for example, goods, services, or people between two or more locations. This could be the movement of people between cities or the volume of trade between countries. Gao et al. [37] proved that check-in data are a suitable and efficient model to predict human mobility using their gravity model, despite some other research works [29] where authors believe the gravity model [30] is not well-fitted to elucidate spatial interactions. In order to solve the challenge of low graph density, they used the particle swarm optimization (PSO) method to obtain the best fit. They revealed the underlying distance of inter-urban travel demand based on social media check-in data and flight passenger data.

Recently, many efforts have been made to analyze human mobility from very distinct datasets. A great effort has been made to review recent works in [38]. Different methods and models have been used to understand human movement. The gravity model is one of the most used models for analysis in this regard. Since the gravity model is static, the parameters must be adjusted with real mobility data from additional sources of data. Beiro et al. [39] designed a hybrid model with a classical gravity model by training a real dataset from Flickr to find people's traces in the U.S.

As a summary, different data types and methods used for human mobility analysis are presented in Table 1.

In addition to previous approaches, there are also research works on human mobility for geographical studies and urban planning whose goal is to find people's important places of movement and, specifically, to analyze their mobility between home and work with the purpose of location tracking. This kind of analysis allows the creation of origin–destination (OD) matrices. There are three categories for OD matrices that can be useful in transportation applications. Collecting data for OD household surveys in a traditional way is time-consuming, does not provide continuous data, and is basically for a specific study area. Another method is count-based OD from traffic detectors. Although this method reduces estimation time and expenses, it has a high cost in terms of installing infrastructure on roads, plus it presents coverage issues. With the emergence of new technology, new sources of data have been introduced to collect traffic flow data, such as GPS and Bluetooth data, which both have their own limitations, for example, due to privacy concerns. When collecting user's

trajectory data by GPS, the user's agreement is required, and since users tend to turn off Bluetooth on their devices to save battery, this influences the sample rate. LBSN services overcome all these limitations in the era of new technologies as a dynamic mobility solution. Yang et al. [40] obtained an OD matrix using a novel mixed combining regression and gravity model using Foursquare check-in data. They classified eight categories of venues using an agglomerative hierarchical clustering method. Using Chicago Metropolitan Agency for Planning (CMAP) data helped to compare and evaluate their model estimating OD travel matrix for non-community trips. In another study, Wang et al. [41] worked with taxi GPS data to obtain travel patterns from OD flows. They proposed a chord diagram plot to illustrate the spatio-temporal patterns of residents in seven-day taxi trajectories in Beijing.

Table 1. Comparative summary of different methods in human mobility analysis.

Data Type	Pros	Cons
Traditional survey [8]	Multipurpose usage	A human source is required, time-consuming, not accurate, and static
WiFi [10]	Energy usage is ~50% of GPS	Low coverage area
GPS [12]	More precise (~5 m error) and can distinguish between transportation modes	High energy usage, signal problems, and expensive
CDR [20]	Auto-generated	Lack of a big data size and privacy issues
Social media [33]	Cheap, easy access, big size	Challenges in extracting useful insights

In another work, Kurkcu et al. [42] proposed a density-based clustering algorithm, elucidating the most common human mobility features of home-to-work travel, validating Twitter data as a ubiquitous and suitable data source to elucidate travel demands. Identifying home places, they clustered the most visited locations of a user based on the number of tweets and assumed the most visited place as a home or origin. In addition to time filtering, they applied some keyword filters on tweets to verify the origin of a user's trip. Based on those conditions, they chose the strongest density as a home place for a given user. Their research shows that Twitter data are a potential suitable source of data to analyze activity patterns. On the other hand, Xuan et al. [43] analyzed different group's behavior using the smart card data (SCD) of Shenzhen combined with social media data. Focusing on students and travelers, they revealed useful insights into travel flows in both aspects of the spatial and temporal characteristics. They applied the K-means algorithm to divide travelers into different groups and check their specific temporal activities on different days of the week. The results show a major hub in the metro system based on the number of tap-in and tap-out times in the dataset.

All these works have shown the relevance and potential of using geo-located data from different sources to analyze human mobility. In particular, this confirms the interest in exploiting social media data in this regard. In this paper, a selected number of algorithms inspired by previous works are implemented to discover human patterns, starting from traditional statistical analysis and progressing to dynamic data analysis of social media.

3. Materials

A description of the data used in this study and their main features are presented in this section, explaining their origin and characteristics. In particular, a description of the Sina Weibo dataset is provided, as well the Shanghai POI dataset.

3.1. Case Study

Shanghai is one of four municipalities under the direct administration of the central government of the People's Republic of China. The city is located on the Yangtze River, on the east coast of China, shown in Figure 1. The municipality's area includes the city itself, surrounding suburbs, and an agricultural hinterland. Shanghai is China's most populous city, covering a land area of 6340.5 square

kilometers, with 24.1970 million inhabitants living in 16 districts at the end of 2016, according to the 2017 Shanghai Statistical Yearbook.

Figure 1. Location of Shanghai in China.

In 2017, 16 districts (Baoshan, Changning, Fengxian, Hongkou, Huangpu, Jiading, Jingan, Jinshan, Minhang, Pudong New Area, Putuo, Qingpu, Songjiang, Xuhui, Yangpu, and Chongming) constituted the city of Shanghai (see Figure 2). Chongming is the largest island and has an area of 489 square miles. The Pudong district was originally established in 1958. The district borders, the Huangpu River, separating it from the central business district of Puxi. Pudong is one of the earliest industrial areas. Hongkou District is another important industrial area that lies to the north east of the Suzhou River.

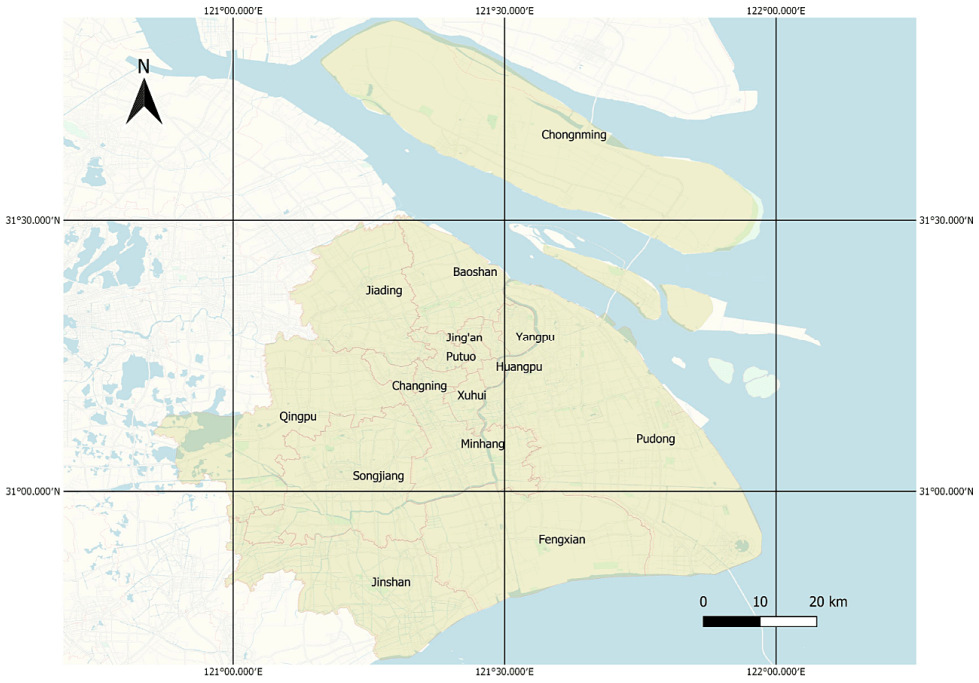


Figure 2. Administrative boundaries of different districts.

3.2. Dataset

Following an approach that explores geo-located social media data, this paper combines different sources of data to analyze human mobility. These sources of data include Sina Weibo data, to detect geo-locations corresponding to home and work locations and to enrich the analysis. POIs are also used to illustrate destinations frequented by people. First, we present the original dataset, then we describe how the data were delimited into selected boundaries and how they were combined with the POI database.

3.2.1. Sina Weibo Data

The initial data were collected from the microblog network Sina Weibo using the Baidu API. Sina Weibo is a micro-blogging social media platform that is considered as the Chinese version of Twitter, as Western social media platforms like Facebook, Twitter, and YouTube are blocked in China. As the second most popular site in China, it provides media and news updates, and users can follow their favorite celebrities. A McKinsey survey reported that 95 percent of Chinese people use Sina Weibo in their everyday lives, compared to 70 percent in South Korea and 67 percent in the United States [44]. (The full report, 'China's social-media boom', can be read on McKinsey's Greater China web-site, mckinseychina.com). In Q1 2018, it was reported that there was a 20.7% increase over the last year in use of Weibo, which had reached 411 million Monthly Active Users (MAUs), opening a potential data source to analyze movement patterns. Data were collected in the period of 2014 and 2015, with a total of 325,713 Weibo posts in the whole dataset, covering a part of China. Considering the importance of Shanghai as a cosmopolitan city and the financial center of China, we focus here on data belonging to this city. The number of Weibo posts belonging to Shanghai city is 248,339. Table 2 shows a short description of our dataset. Each record of the data contains different types of information about users, such as their user ID, gender, location (longitude and latitude), time, message, etc. After storing and filtering data, we used OGIS for visualization. Table 2 shows the initial dataset statistics.

Table 2. Dataset description.

Dataset	Number of Users	Number of Check-Ins
Original Dataset	76,603	325,713
Shanghai Case Study	48,594	248,339

As Table 2 shows, the original dataset comprises 325,713 check-ins, and after data preprocessing it was reduced to 248,339 check-ins. The cleaning and pre-processing stages applied to the source data are described with more detail in Section 4.

3.2.2. Boundaries Data

In order to understand the spatial aspect of check-ins, we used shapefiles representing the administrative area boundaries of Shanghai to determine the district from which the Weibo posts were published. The resources were taken from an open geo-data source called global administrative areas (GADM). GADM is a high-resolution database of administrative areas (boundaries) for all countries in the world, at all levels of sub-division. Administrative areas in this database include countries, provinces, departments, counties, etc. For each area, it provides some attributes, like name, variant names, and "spatial features" about the location of the areas. The GADM data are publicly available to be downloaded by country or the whole world in different formats, such as shapefile, RData, Google Earth kmz format, and ESRI geodatabase. The GADM database allowed us to confirm the origin (district and province) from which the Sina Weibo posts were published. Figure 3 shows the number of published Weibo posts in 16 districts within Shanghai.

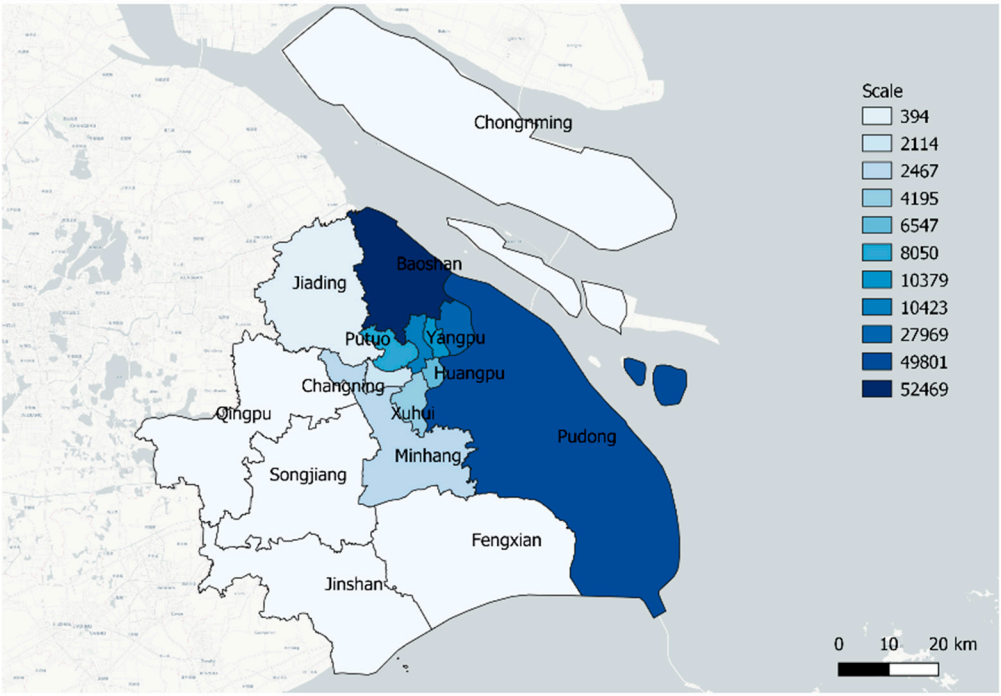


Figure 3. Weibo count by district with administrative boundaries.

3.2.3. Points of Interest

In addition to the Sina Weibo dataset, we used a POI dataset. The definition of a point of interest is a specific physical location that someone may find interesting, such as a restaurant, retail store, hospital, bank, etc. Some researchers have classified travel demand for intra-urban human mobility by POI data [45]. In a similar approach, this work first used the Baidu API to translate Chinese text into English and validate it with a Chinese native speaker. Figure 4 shows an example of the source POI data that were completed with the English translation for the category and name of the place, focusing only on Jiading district.

ID	Longitude	Latitude	Category	Category_En	Name	Name_En	Activity
B2094654DB65A4FD419A	121.310980	31.29164	医院	Hospital	南翔医院	Nanxiang Hospital	Healthcare
B2094757D568AAFB4099	121.315720	31.296480	住宅区	Residential area	南翔镇南华路，民主街解	Nanxiang Town, Minzhu Street	Residential
B2094750D26BA7F9429A	121.317808	31.29337	医院	Hospital	上海南翔医院	Shanghai Nanxiang Hospital	Healthcare
B2094654D46BA5FC419D	121.309330	31.28952	银行	Bank	中国工商银行	Industrial and Commercial Bank of China	Services
B2094757D06AA5FF449C	121.313786	31.282249	火车站	Train station	南翔北站	North Nanxiang Station	Transportation
B209465DB65A0F8469C	121.310300	31.287600	经济型连锁酒店	Economical Chain Hotel	锦江之星	Jinjiang Inn Hotel	Services
B2094654D46BABFC499F	121.304960	31.291130	洗浴推拿场所	Bath and massage center	玉指健康会所	Yuzhi Health Center	Services
B2094757D06FA3FD4799	121.325304	31.296882	收费站	Toll gate	S5 沪嘉高速南翔收费站	S5 Hujia High-Way Nanxiang Toll Gate	Transportation

Figure 4. Source data for points of interest (POIs), including Chinese and English versions of the activities.

Then, knowing the POI activity, we classified the POI data into eight different categories including: 'dining' (restaurants, cafes, food courts, and tea houses), 'education center' (schools, universities, kindergartens, libraries, institutes), 'entertainment' (KTVs, parks, cinemas, museums, temples, concert halls, and art galleries), 'healthcare' (hospitals, clinics, sport centers), 'residential' (buildings and independent home residential areas), 'services' (beauty salons, hotels, post offices, stores, and shopping centers), 'transportation' (bus stations, train stations, airports, and metro stations), and 'work places' (companies, offices, industrial areas, and banks). Table 3 shows statistical results providing insight into the different categories.

Table 3. Sina Weibo venue statistics by category.

Category	Number of Check-Ins	Percentage	Number of Venues	Percentage	Avg. Check-Ins
Dining	3049	5.85%	2637	7.49%	1.15
Education center	3614	6.93%	2259	6.42%	1.59
Entertainment	15,730	30.17%	12382	35.17%	1.27
Healthcare	2832	5.43%	1901	5.4%	1.48
Residential	1324	2.54%	567	1.61%	2.33
Services	5081	9.74%	2651	7.53%	1.91
Transportation	12,050	23.11%	9744	27.68%	1.23
Work places	8461	16.23%	3065	8.7%	2.76

The statistical results in Table 3 indicate that the 'entertainment' category attracted the greatest number of users (check-ins), followed by the 'transportation' category. The highest average of check-ins belongs to the 'work places' category. Figure 5 shows the distribution of POIs organized by category, considering their geolocation in the different districts of Shanghai. Most POIs belong to the 'services' category, followed by 'education centers' and 'residential areas'. The smallest amount of POIs corresponds to the 'transportation' and 'healthcare' categories. Different colors illustrate the concentration of each category of POI in each district of the city of Shanghai.

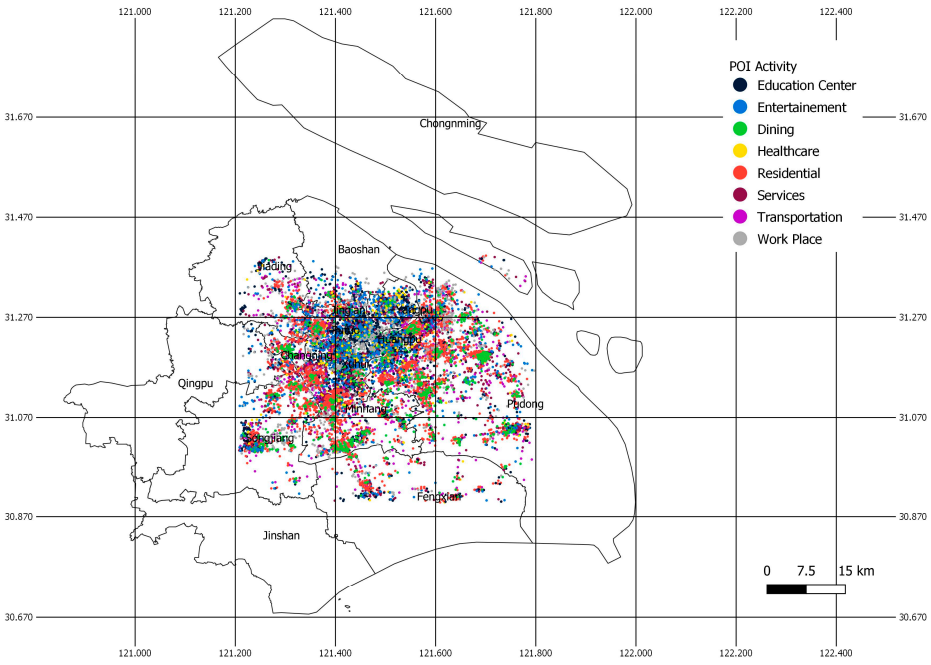


Figure 5. Distribution of POI categories.

4. Data Preparation/Pre-processing

In order to get a reliable and consistent dataset before data can be further explored using data analysis techniques, the raw data were converted into an appropriate format to ensure that they belonged to the space and period of interest in order to prevent the inclusion of biased results. A set of scripts was developed in the Python programming language, ready to be applied to the dataset. All steps and transformations applied are described in the following sub-sections.

4.1. Data Cleaning

In this part of the process, incorrect data were detected, filtered and corrected in order to prevent false conclusions. The pre-processing consists of filtering out irrelevant columns, removing duplicate rows, resolving inconsistencies, and performing data harmonization, allowing better data interpretation in the analysis process. After deleting nine columns of data that did not contain related content, we were left with 48,594 unique users and a dataset size of 248,339 Weibo posts.

4.1.1. Filtering Check-in Records

Fake check-ins exist, despite some mechanisms to prevent them in social media services. They may happen when a user chooses a geotag for a post when, in fact, it does not exist or they are not there. The place or location from where tweets or messages are sent is assumed to be 'home' here. In our analysis, we assumed that a place or location from which the user posted a Weibo, or the location from where most messages are sent during non-working times, is the 'Home'. In order to detect origins and destinations and eliminate fake check-ins, we filtered the data. Since the data have two aspects of space and time, a major filter was applied to each aspect:

- Temporal filtering:

For the time feature, we assumed those users who published only within a period of 10 days were possibly (potential) tourists. Moreover, we extracted the relationship of different check-in timestamps to calculate check-in duration, considering the start and end time. We also created a timestamp for each published Weibo, adding the day of the week to better analyze the temporal feature;

- Removing spatial outliers from the data:

The outliers here are defined as extreme values not relevant to the current study area. Considering the location, we chose Weibo posts that had been published within the area of our case study, taking into account the territorial boundaries, as shown in Figure 6. On the other hand, to find users who usually send messages from a similar location, we filtered the users with an average distance of less than 50 m by calculating the distances of the dispersion deviation of each user. Only about 100 Weibo posts were excluded, and some of them came from the shore of Shanghai.

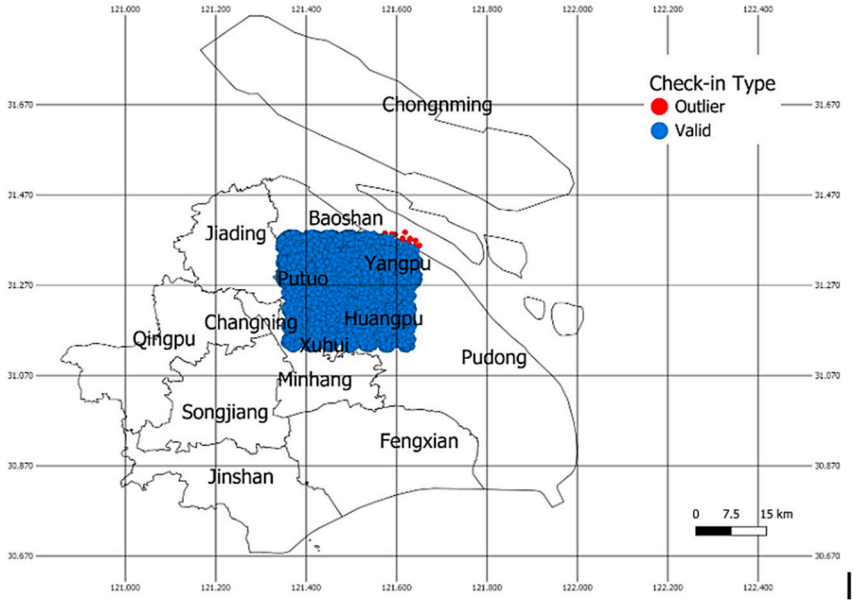


Figure 6. Outliers outside the focus of the study (Shanghai districts) were not considered.

4.1.2. Additional Filtering

The purpose of adding more filters to our analysis was to increase the quality of data and to work with more relevant data. Additional filters were applied in order to produce precise results from this analysis:

- Different accounts identified by the same user ID were eliminated. We removed any user ID with more than 100 Weibo posts per day at the same location, considering it to be a bot account;
- In order to find users who have very little activity among the users in the case study, we eliminated those who have published only five Weibo posts, identifying users with low activity;
- Users with 4 days or less of consecutive check-ins were also erased.

According to the above criteria, the final check-in records totaled 233,467 Weibo results. We used these data in the next step to extract meaningful insights into our research.

4.2. Data Transformation

In order to take advantage of the collected data, new fields were generated using the POI and boundary datasets, giving, as a result, a whole enriched dataset. The fields that were added were:

- Day of week: determined by mapping between timestamp and Chinese calendar;
- District name: determined from the GPS location column and boundaries dataset;
- District of publication: determined from the GPS location column and boundaries dataset;
- Activity category: determined using the nearest POI.

5. Methodology and Results

The research follows the lifecycle of an average data analysis/mining project. It starts by formulating questions, then moves on, collecting the data, preprocessing the data, exploring the data and communicating the data findings. Figure 7 shows the steps of the methodology applied.

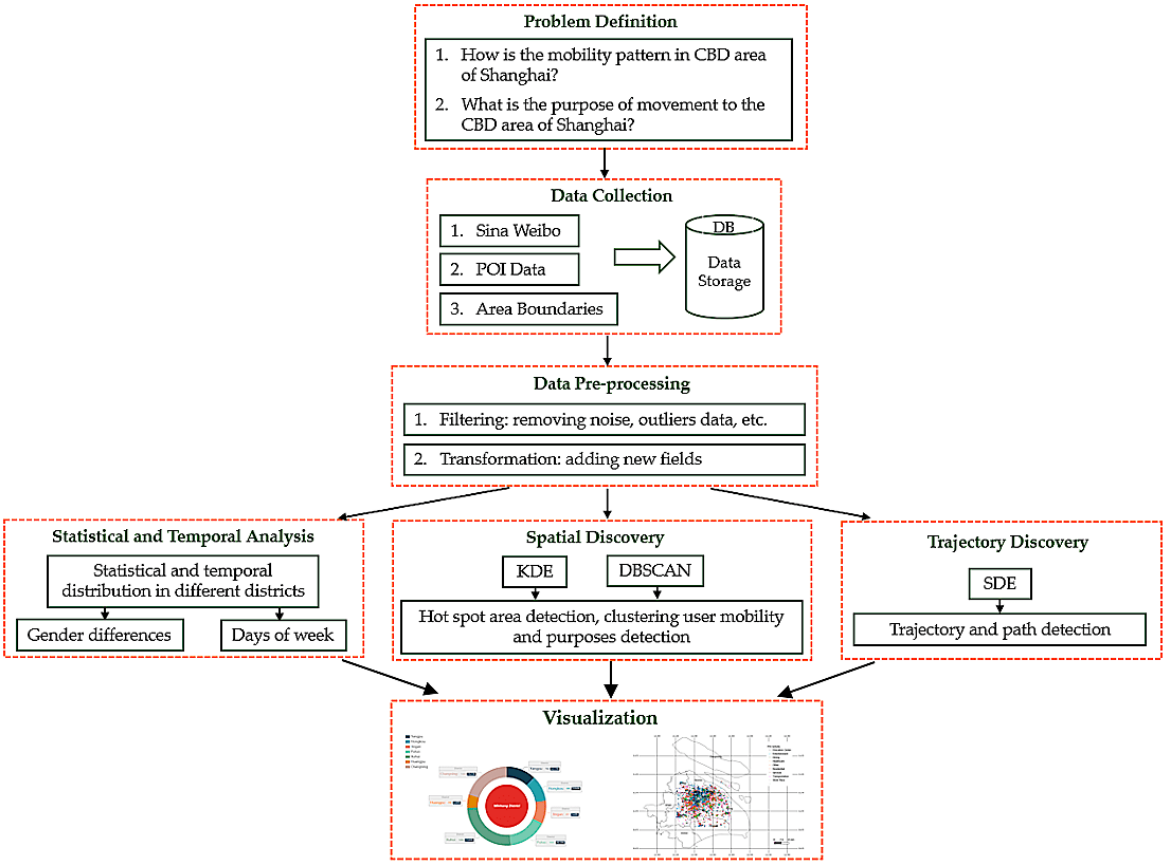


Figure 7. Methodology.

5.1. Exploratory Analysis and Visualization

In our experiments, we divided the analysis into different sections. In this section, we investigate and compare the statistical analysis regarding the Sine Weibo data from between 2014 and 2015 from 11 districts of Shanghai, including the CBD area. After cleaning the data, we obtained 233,467 check-in, geo-tagged results and analyzed these from different perspectives, as explained below:

- Gender perspective

Regarding the analysis of different communities, we analyzed the data based on gender. The results in Figure 8 show female and male activities in two periods of time from 2014 and 2015. We found that, in both years, females were more active than males. Analyzing the total check-in numbers differentiated by gender, we found that Sina Weibo, as a social media platform, is more popular among females, with 22% more female users than males;

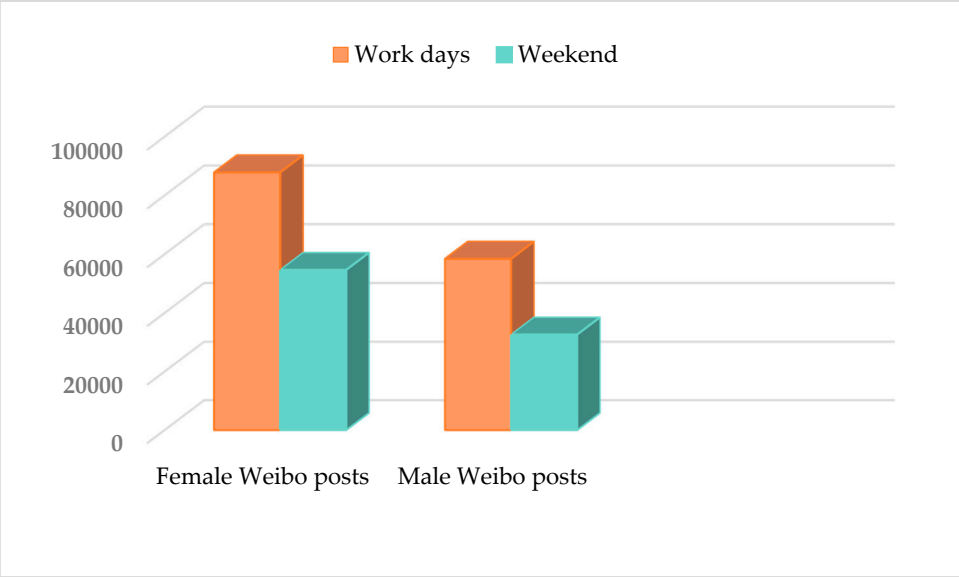


Figure 8. Gendered Sina Weibo usage frequency from the period 2014–2015.

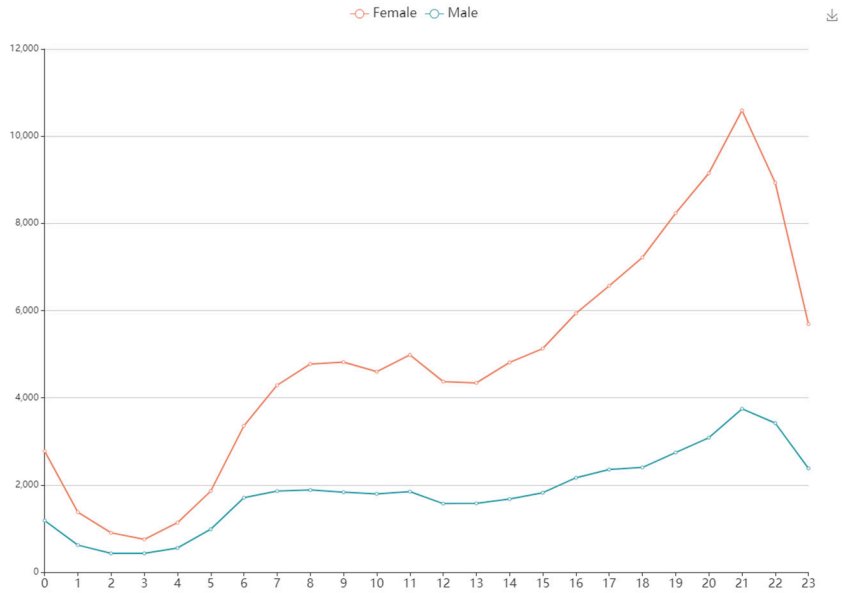
- Temporal perspective

In order to find the peak hour of activities and the patterns of human mobility, we analyzed the data based on time. We analyzed the data of two different periods of the years separately and found that the activity trend was almost the same; thus, we merged the data of the two year period and illustrated the results in Figure 9. Despite other studies that have analyzed mobility based on periods of time (every six or four hours), we chose hourly movement to figure out more details because of the rapid movement of people, and because we believe that people do not stay in one place for several hours and check-in several times. The horizontal axis represents the time of day, starting from midnight (0) and moving to 11 PM, and the vertical axis shows the number of check-ins at that specific time or hour. The distribution of check-ins for females (red color) and males (blue color) is denoted by two different colors. First, we sorted the time series in descending order by each hour timestamp, then checked the user check-ins in those series by gender

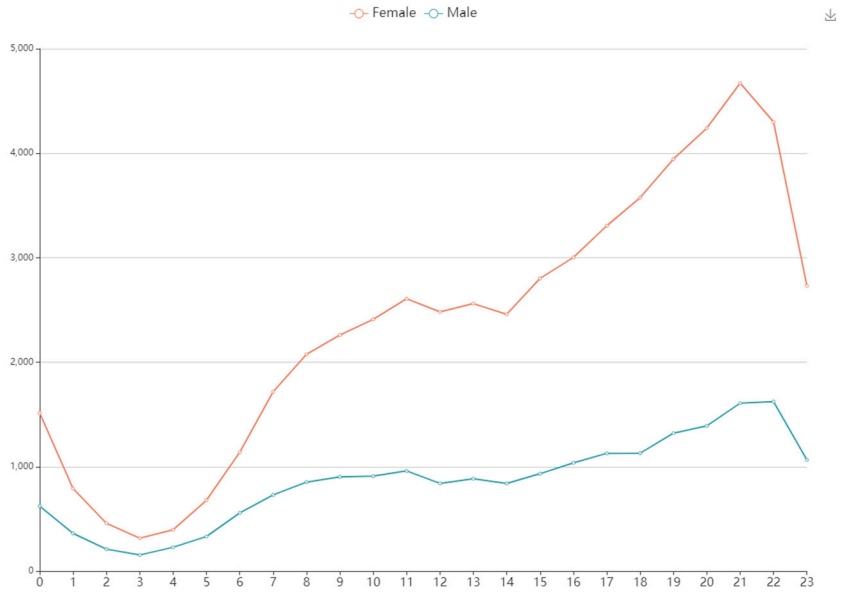
$$P_u^{d,t} = \{(x,y)|(x,y) \in (female, Male)\} \quad (1)$$

where $P_u^{d,t}$ denotes the set of time t per day d for the group of users u by gender.

The results show that both females and males are more active at the end of working days in the period of 18:00 to 21:00, while fewer activities happened between 11:00 to 14:00, which are considered as lunch and resting times during working days. However, it was also observed that the check-in frequency of females was almost consistent, with a slight increase during the day, while it was the opposite for males in the same period of time;



(**a**)

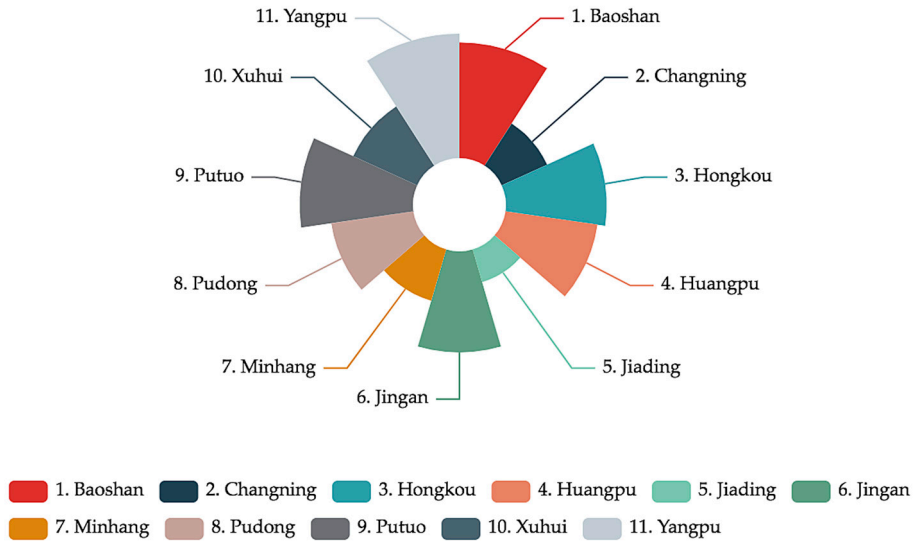


(**b**)

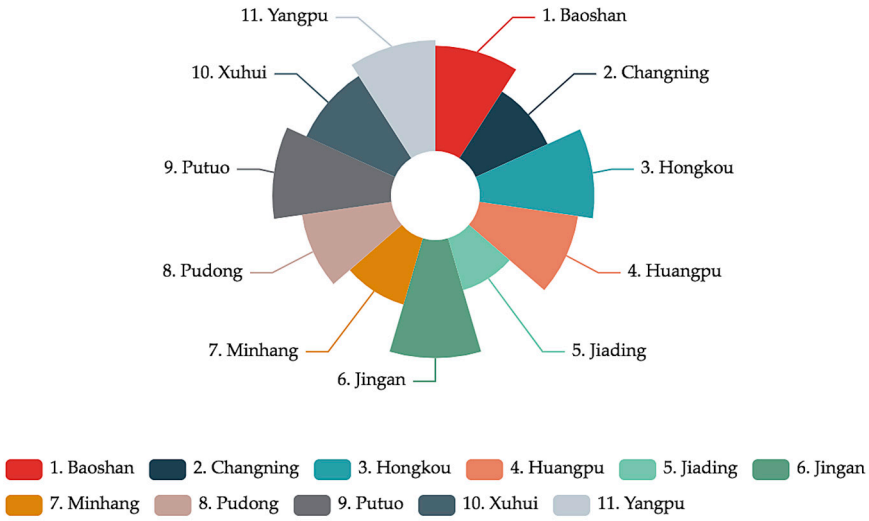
Figure 9. (**a**) Hourly check-in trends of workdays. (**b**) Hourly check-in trends of weekend days.

- Spatial perspective:

Based on the analysis, it is clear that human activities vary over time, but it is of great significance to analyze human mobility in different regions. For this reason, we analyzed user activities in different districts to illustrate spatial dependencies. The goal here is to help urban authorities, who hope to understand human mobility patterns in different regions within the city. Figure 10 shows human activities based on 11 districts of Shanghai from 2014 and 2015. From Figure 10, it is obvious that Yangpu is the most famous district among users because it is now the location of the city hall, The Bund, and the most famous shopping malls and tourist sites.



(**a**)



(**b**)

Figure 10. (**a**) Check-in records in different districts in 2014. (**b**) Check-in records in different districts in 2015.

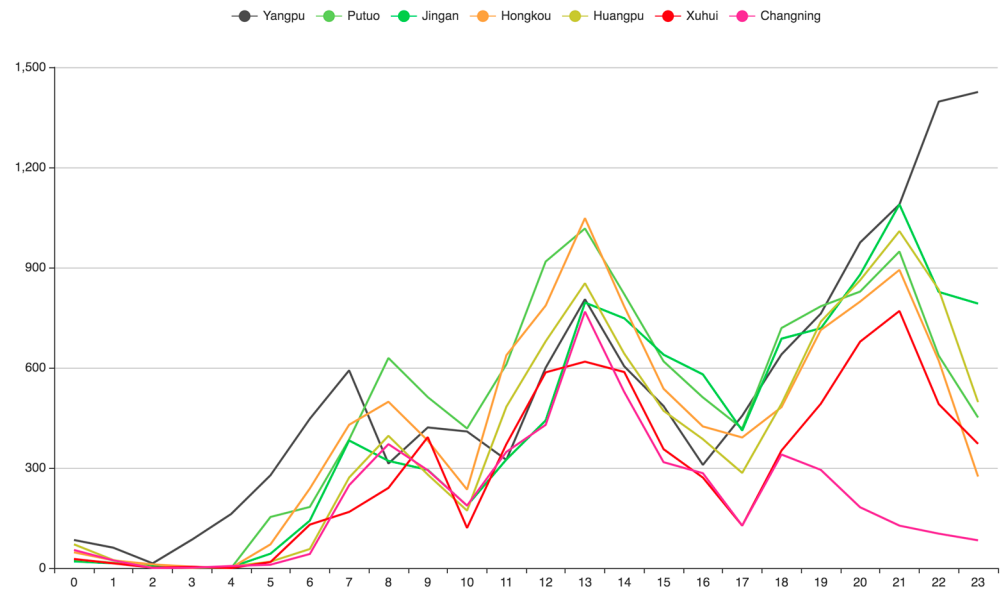
Table 4 shows the most visited districts in Shanghai. Among these top six districts, five are located in the CBD area of Shanghai;

Table 4. The six highest density districts from the two years.

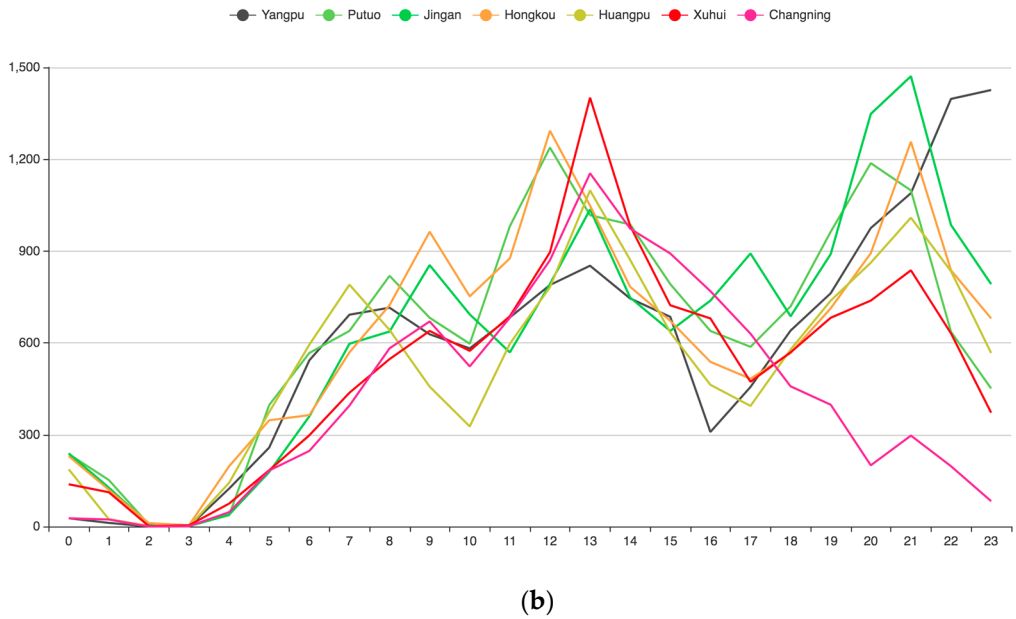
2014	2015
Putuo—11.23%	Yangpu—13.25%
Jingan—11.15%	Baoshan—12.32%
Hongkou—10.86%	Putuo—12.08%
Yangpu—10.48%	Jingan—10.765%
Baoshan—9.95%	Hongkou—10.74%
Huangpu—9.44%	Huangpu—9.91%

- Spatio-temporal perspective

Moreover, for further and in-depth analysis we investigated big geo-data and applied mining techniques to discover knowledge concerning spatio-temporal relations, as shown in Figure 11. We analyzed the check-in activities at different hours of the day and compared the two years of data for seven districts of Shanghai, considered as a CBD. The horizontal axis represents the time of day, starting from midnight (0) and moving to 11 PM, and the vertical axis shows the number of check-ins at that specific time or hour. The distribution of check-ins is denoted in eleven districts of Shanghai, including the CBD area and its four neighboring districts, which are shown in a different color.



(**a**)



(**b**)

Figure 11. (**a**) Hourly check-in activities in different districts in 2014. (**b**) Hourly check-in activities in different districts in 2015.

The analysis shows that people tend to publish more Weibo posts within the period of 1100 to 1400 h, which is considered as a lunch and resting time. Each color shows a different district and the size of the graphic reflects the number of Weibo posts posted in that district. We can see that the trends are almost the same over the two years. There are clear increases in the number of Weibo posts posted during the lunchtime break period and when people are normally back home at night.

5.2. Kernel Density Estimation (KDE)

As discussed above, the temporal information of movement in geographical space is important to detect the spatio-temporal trends of underlying human mobility. However, with the increasing number of aggregated human/vehicle trajectories in urban space, the space–time path representation model will be hard to interpret because of the overlapping and cluttering issues. To solve this problem, we applied the kernel density estimation (KDE) [46], which has been widely used in spatial analysis to characterize a smooth density surface that shows the geographic clustering of point or line features in 2D space. KDE is a non-parametric algorithm used to calculate the density of features in a given dataset. Applying KDE helps us to create a smooth curve in a given dataset and find the strongest density, which represents the most important places in people's lives. KDE is formulated as follows

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad (2)$$

where h is bandwidth and n is the number of points. The bandwidth affects the smoothness of the resulting distribution. With the kernel function k, this algorithm weights the points in each location to calculate the distance. The bandwidth has a direct influence on the shape of the curve, meaning that the bigger the value of the bandwidth parameter, the smoother the curve, which contains more points (a large bandwidth leads to a very smooth density distribution and a small bandwidth leads to an unsmooth distribution). Since we applied KDE, we can show a curve line showing the underlying distribution, which is expressed as

$$\mathrm{pk(y)} = ((\mathrm{y} - \mathrm{x_i})/\mathrm{h}) \quad (3)$$

As shown in Equation (3), k is the kernel function estimate of the density at a point y within a group of points, x_i, and $i = 1 \ldots N$ and h is the bandwidth. We used KDE to find the nearest neighbor here.

The process consists of finding the POI nearest neighbor to each user status using an efficient spatial search algorithm called KD-Tree, which creates a binary search tree (index) to match against the lookups, which is carried out quickly, reducing the search time to *O (d log n)* where d is the number of dimensions (in this case, two latitude and longitude). If a POI is within a predefined umbral, then we associate the activity category to the user's status. The KD-Tree algorithm consists of partitioning the space along one dimension at a time (for example latitude), finding the median of the data that conform with the selected dimension, splitting the data based on the median, and changing the axis in a cyclic fashion at each partitioning step until, in each partition, we have the highest M number of points in each partition (leaf size 10).

As shown in Figure 12, most user activities are located in the center of Shanghai, or, more specifically, in seven districts, including Jingan, Hongkou, Huangpu, Xuhui, and Putuo, which are considered as central business districts (CBD). Focusing on these specific areas, we analyzed the mobility of people in different aspects. First, in order to remove noise and achieve smoother data, we applied the density-based spatial clustering of applications with noise (DBSCAN) algorithm.

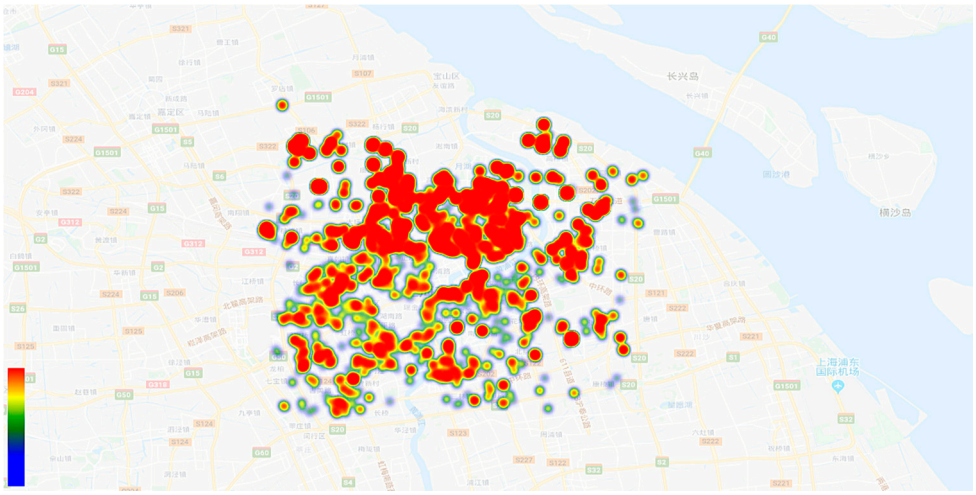


Figure 12. Heat map of Sina Weibo data in Shanghai.

5.3. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Spatial clustering analysis is a well-known data mining technique. It groups objects into clusters according to their similarities in both the location and attribute aspects. DBSCAN [47] was used in this study. We used the DBSCAN algorithm to find clusters of important locations in the user's mobility, which are densely connected in the region given in the dataset. As a clustering method, DBSCAN can remove noise on a set of points and verify whether, based on the Euclidean distance, a group of points are close to each other. Two main parameters are required by DBSCAN:

- Epsilon (*eps*): In order to find close neighborhood points, *eps* is established as the maximum distance needed between points to define a cluster;
- Min Points (*MinPts*): This parameter denotes the minimum number of points to shape a cluster.

For example, if *eps* is equal to 0.5 meters and *MinPts* is equal to 7, DBSCAN will start with a random point and discover the seven points around its region of up to 0.5 meters, forming the first cluster, and if it could not find the minimum number of points to define the dense region, it would consider this random point as noise. We chose 5% of user's published Weibo posts in the dataset as *MinPts* with a 200-meter *esp* value. Choosing a suitable value for *eps* is crucial, because if the value is too high, the majority of the points will be merged in a same cluster, while, on the other hand, if a small value will be selected for *eps*, there will not be enough points for the cluster or it will not see the minimum number of points, and thus it will be considered as noise. A k-distance graph can be used to find a suitable *eps* value. In the implementation of the DBSCAN algorithm, the data were filtered and users that had less than or equal to 20 Weibo posts were chosen. We set the parameters of DBSCAN to find five samples within 500 meters to make a cluster. Figure 13 shows an example of the clustering with DBSCAN.

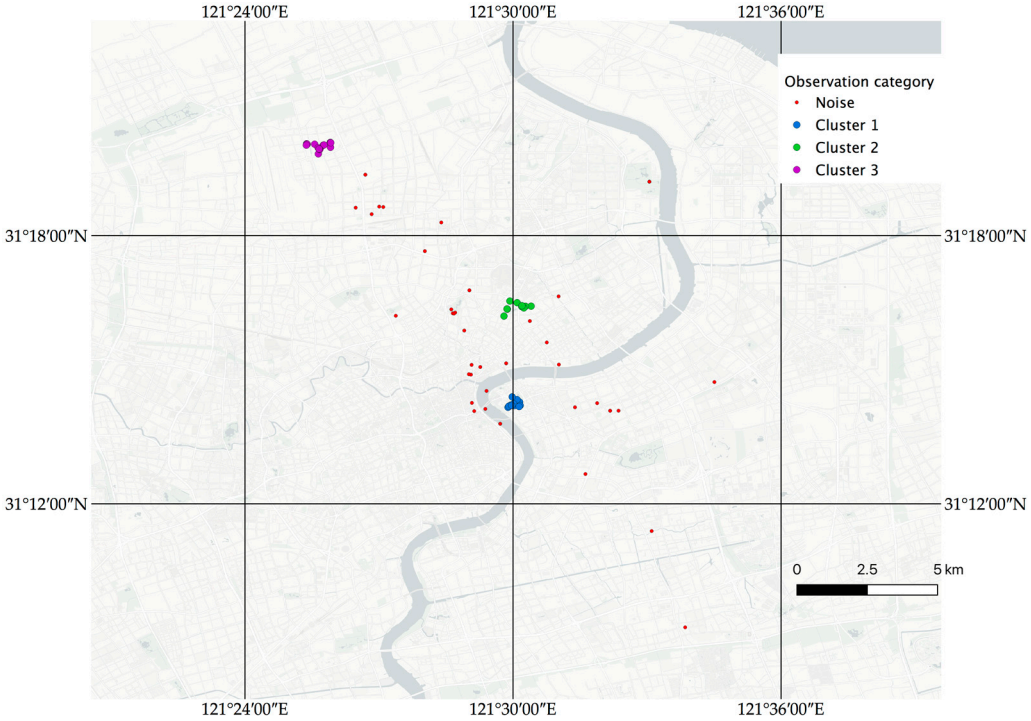


Figure 13. Density-based spatial clustering applications with noise (DBSCAN) clustering implemented to a random user check-ins.

One of the major challenges in analyzing LBSN data is the bias in human mobility, because users can update their location at any time and place at home or work. By applying DBSCAN, we found the most important places in the user's mobility and consider the strongest location of a user, that is, the densest area with the most points, as 'home', since people have more free time at home and we believe that there is a relationship between free time and posting on Weibo. The results were verified with the spatio-temporal aspects of published Weibo posts. For the spatial aspect, we explored the longitude and latitude of each district, in addition to the boundary data. For the temporal aspect in each district, we assumed that the time period between 6 to 11 PM, during Monday to Saturday representative of 'home' activities. However, we eliminated values larger than 10 miles and lower than 0.1 miles due to the limitation of boundaries in the study area, as shown in Figure 13. The results were verified randomly by checking the user's profile, where they chose their residential district while creating their account.

Figure 14 shows a general view of total movement between districts. In general view, In-degree mobility was calculated by the percentage of sum of all Weibo posts from non-CBD districts divided to the total number of Weibo posts, and Out-degree mobility is the percentage of number of Weibo posts from all CBD districts divided by total number of Weibo posts and, in order to give better detail of the idea of mobility, we visualized the statistical results as well as the geospatial map separately for each district for the users whose 'home' was in the same district and had check-ins during the day time from 8:00 a.m. to 6:00 p.m. in the CBD area. To do so, we should calculate the displacement metrics. We investigated the displacement between users' iterative check-ins, representing the mobility distance between the user's likely home locations.

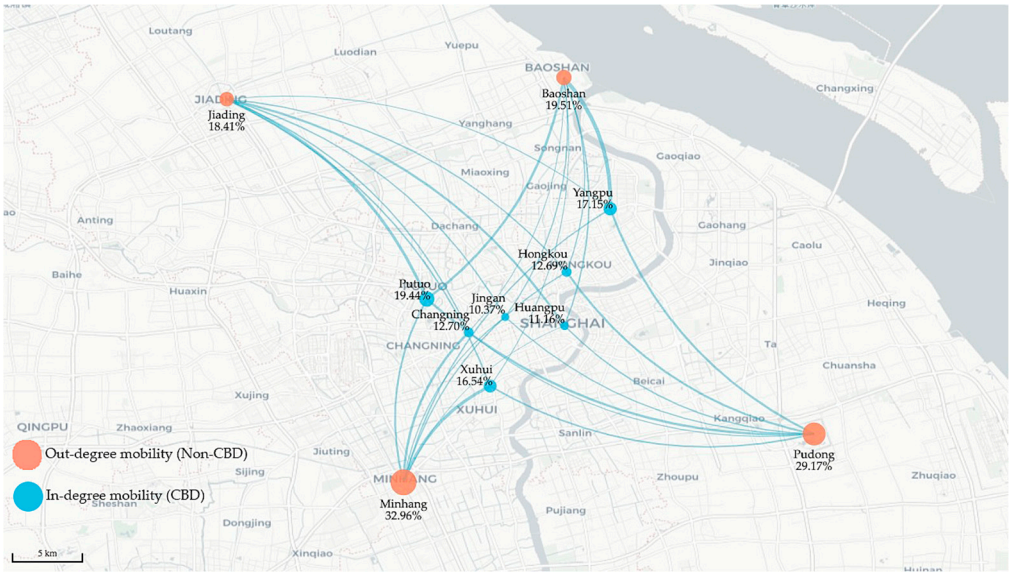


Figure 14. Overall human mobility in the Shanghai commercial business districts (CBD).

Although the gravity model is the most commonly used method to estimate the distance of consecutive check-ins, there are some limitations that need to be balanced, as mentioned in [30]. Therefore, following the previous studies [48], we proposed the Haversine formula to compute the great circle distance between a pair of points by feeding the longitude and latitude of two published Weibo posts within a three-hour interval into the equation below

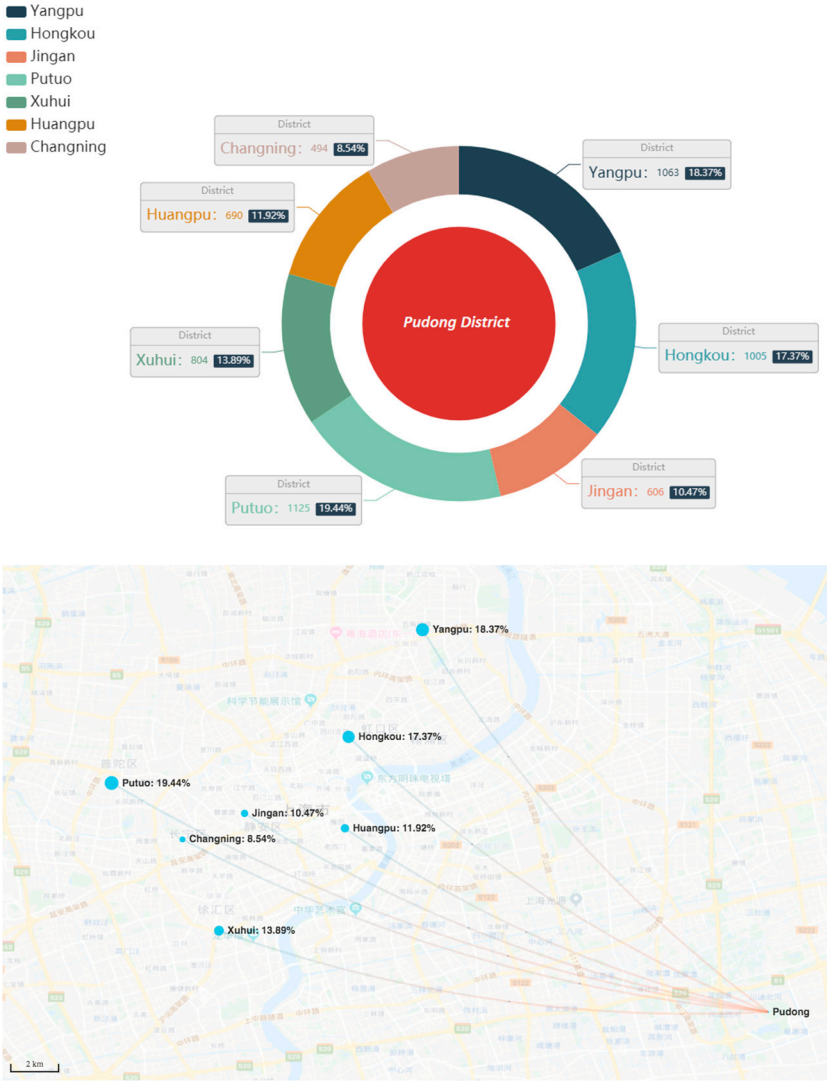
$$C = 2r\,arcsin\left(\sqrt{sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right)+\cos(\varphi_1)\cos(\varphi_2)sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right) \tag{4}$$

where r is the radius of earth, φ is longitude and λ is latitude. At each three-hour interval, the displacement is calculated by dividing the sum of the displacements by the total number of unique users.

Then,

$$D^{d_i,\,t_p} = \left\{\frac{\sum_{d_i,t_p}^{d_i,t_{p+\Delta t}} S^{d,t}}{\sum_{d_i,t_p}^{d_i,t_{p+\Delta t}} U^{d,t}}\right\} \tag{5}$$

where $D^{d_i,\,t_p}$ indicates the average displacements from two period of time, t_p , and $t_{p+\Delta t}$ for each day d_i. The term $S^{d,t}$ represents the displacements for $U^{d,t}$ users contributing in each day and at specific time intervals. In this study, a Δt value = of 3 hours was used to calculate the mobility between districts. As shown in the equation, the average displacement is calculated by dividing the summation of all displacements by the total number of users, and the results of this are shown in Figure 15. The 'geopy' Python package was used to run the code and calculate the distance between Weibo posts.



(**a**)

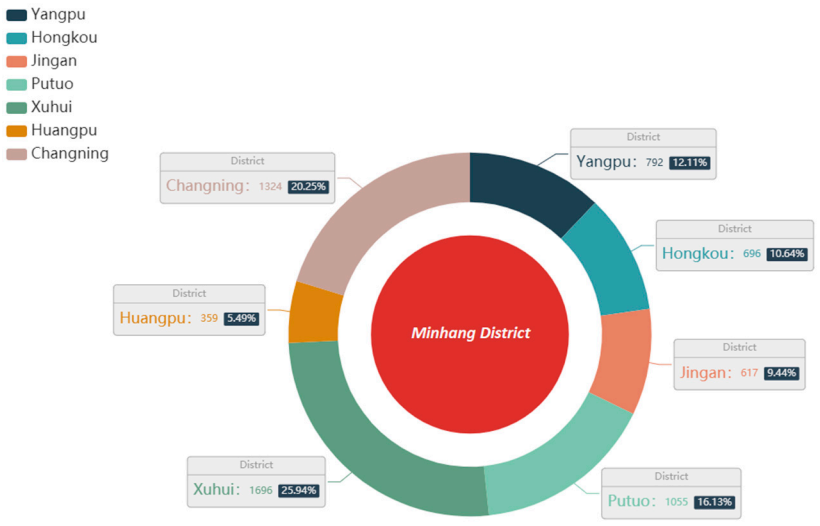
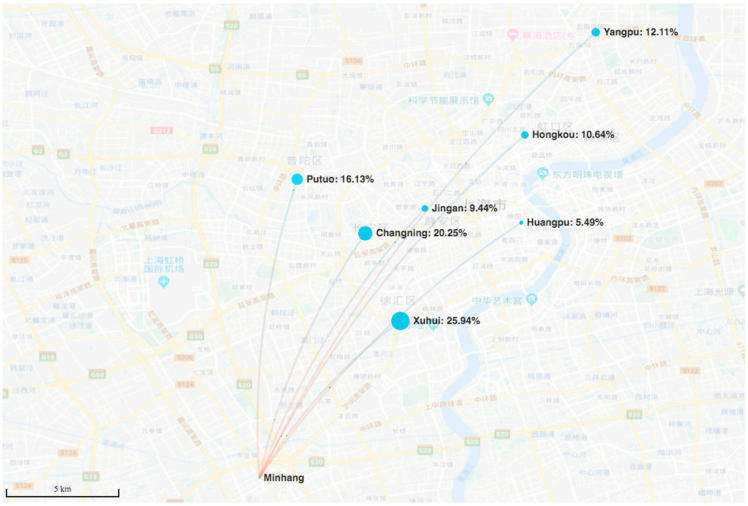
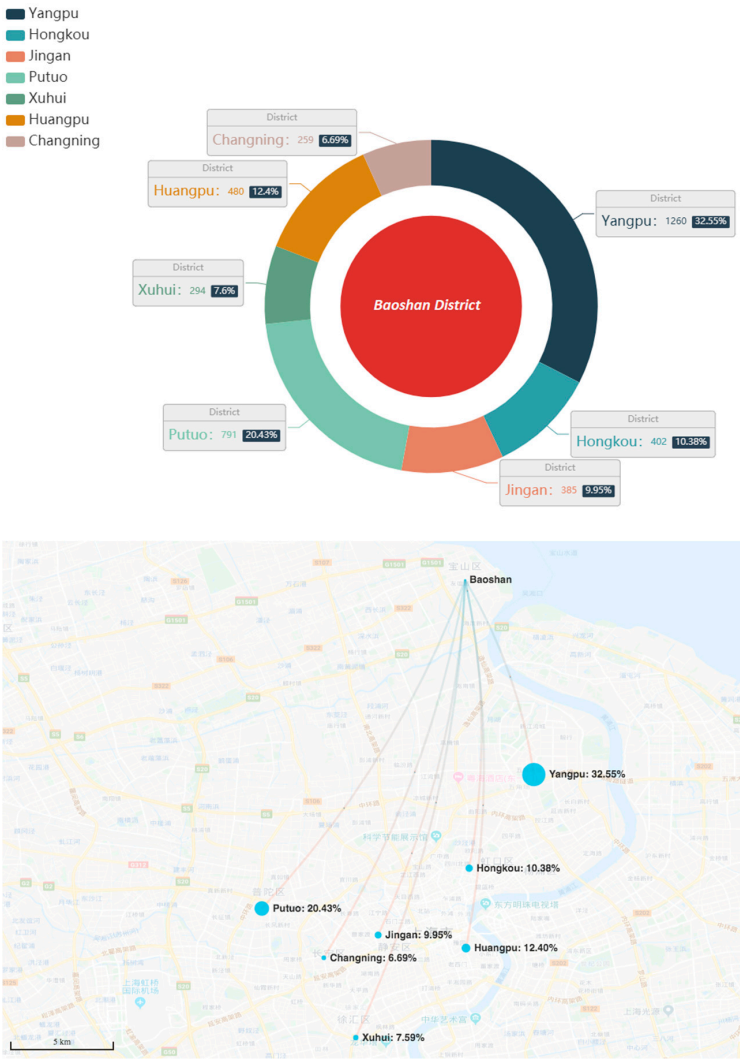


Figure 15. *Cont.*

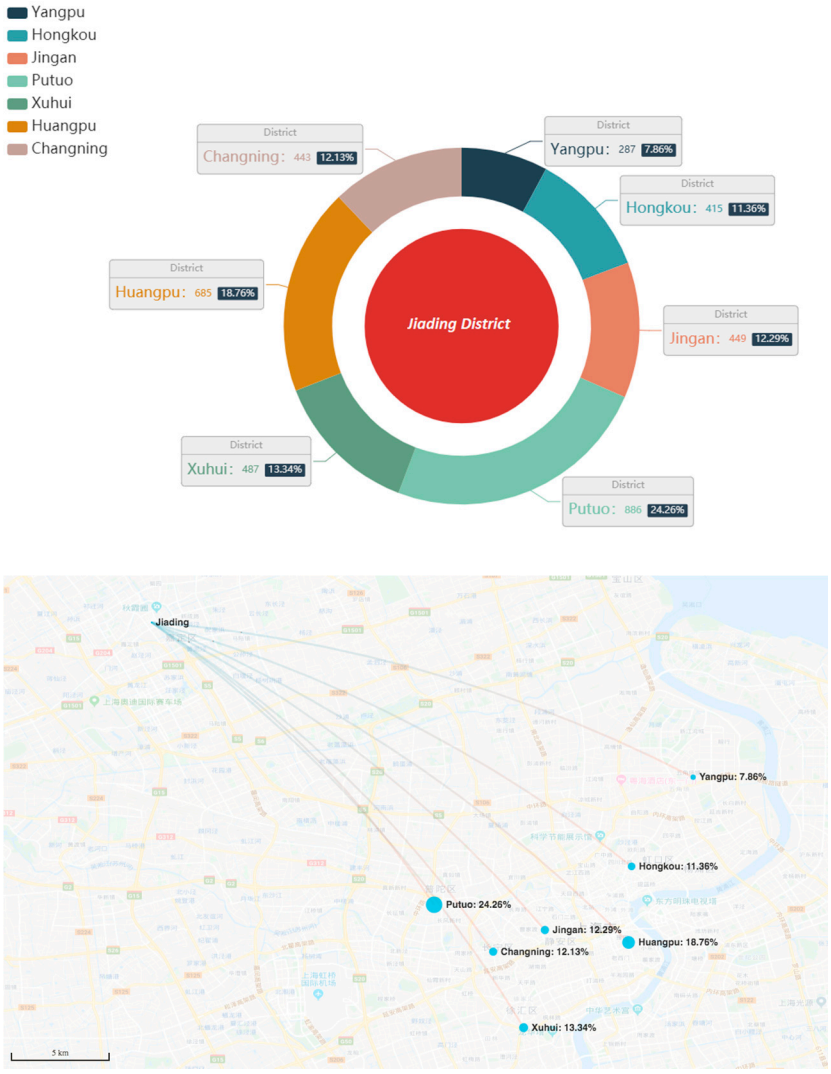


(**b**)



(**c**)

Figure 15. *Cont.*



(**d**)

Figure 15. Human mobility from neighboring districts to the Shanghai CBD: (**a**) Pudong District; (**b**) Minhang District; (**c**) Baoshan District; (**d**) Jiading District.

We made four groups (clusters) for the four neighboring districts of Baoshan, Pudong, Minhang, and Jiading that include users whose 'home' belongs within the longitude and latitude of each district. We investigated the mobility of users in days of week to the central business district, as shown in Figure 15 (here, check the supplementary files to see the dynamic mobility of people).

5.4. Standard Deviational Ellipse (SDE)

The standard deviational ellipse (SDE) was proposed by Baojun to analyze the distribution characteristics of discrete point data, considering a rotated ellipse with a long axis that denotes the main orientation distribution [49]. There are some SDE functions available in software, for example, ArcGIS, used to analyze spatio-temporal data, which helps to visualize the orientation distribution ellipse of discrete points of data. As an effective tool, it visualizes a realistic model for human movement with less sensitivity to outliers. It had been used for research fields, such as the geo analysis field [50]. At a detailed level, we also used this helpful method to investigate human mobility tagged with travel demand. As shown in Figure 16, the major movement of people is along the CBD area in the west-east direction. The eccentricity of the ellipse is larger in the period of 4 to 8 p.m. and decreases gradually

by night from the period between 10 p.m. and midnight. It is tightly connected to the geographic distribution of the city's infrastructure, because the CBD area, which is also called Puxi, is bisected by the Huangpu River on the west part, connecting with the east part of river by the Pudong New Area, where one of the major tourist attractions is located.

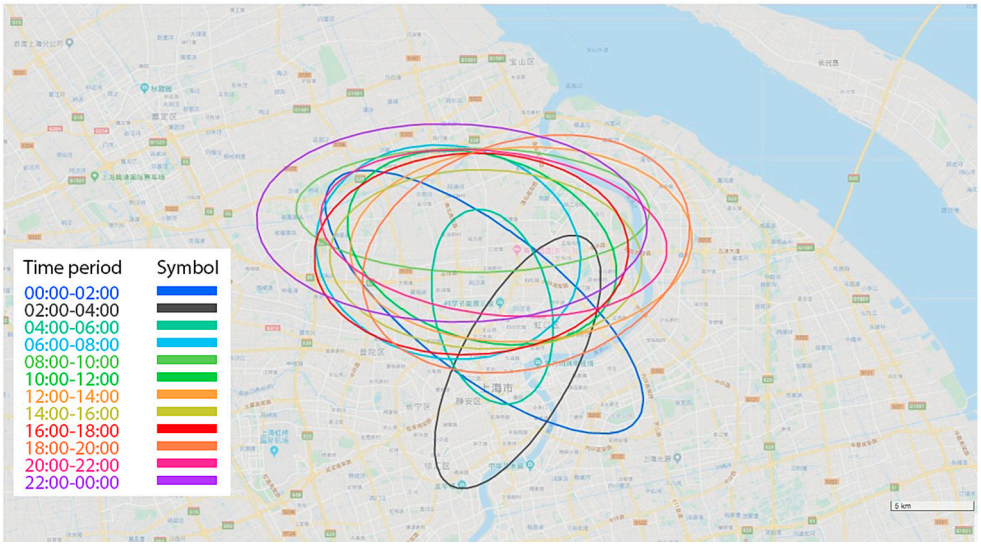
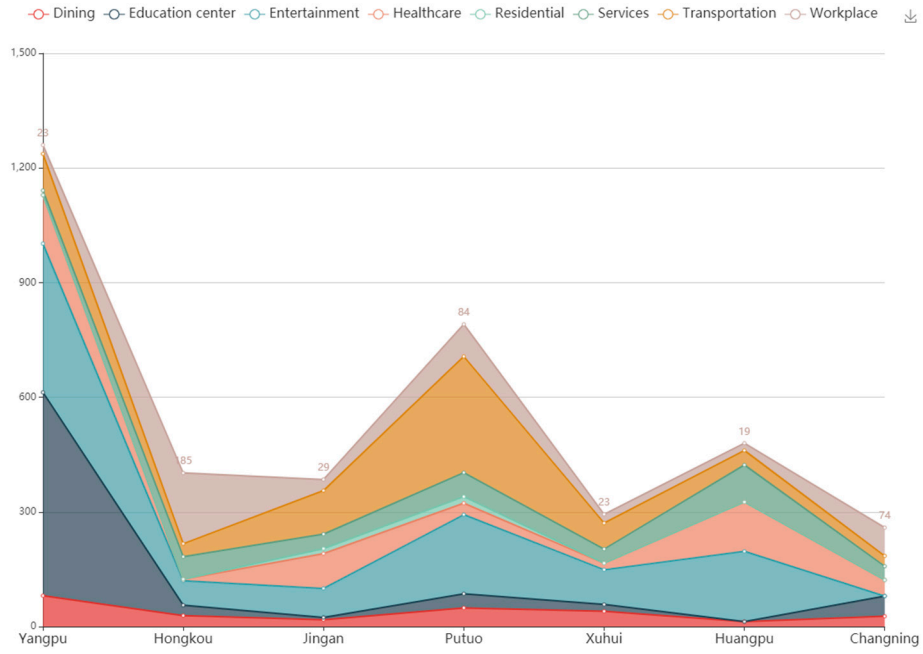


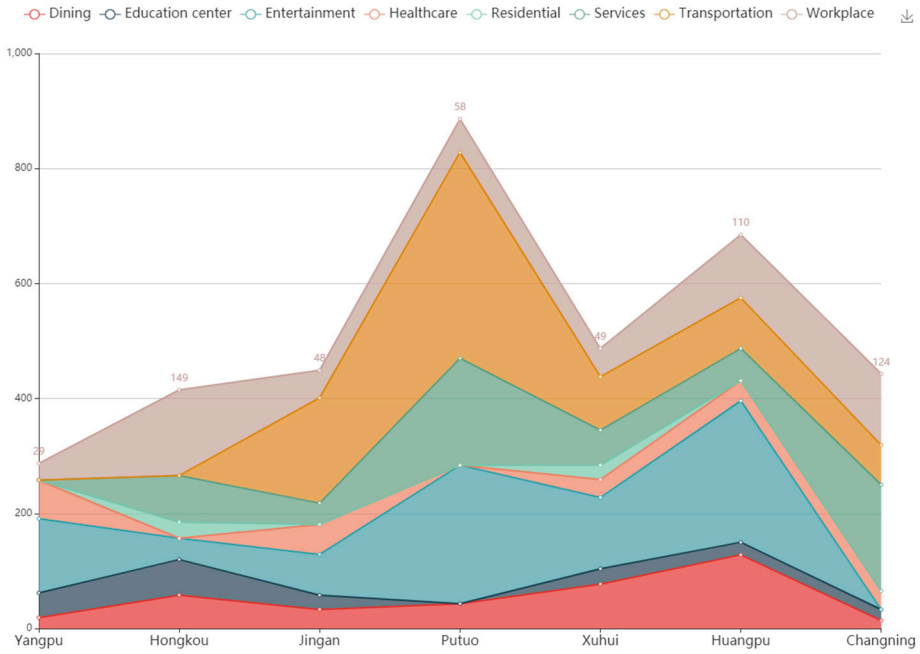
Figure 16. The standard deviational ellipse (SDE) result for travel mobility in the city and moving trajectories.

In order to answer one of the purposes of this study, that is, to elucidate people's interests or reasons to move to other districts, we focused our analysis on the total movement of each of four neighbor districts, namely, Baoshan, Jiading, Minhang and Pudong, to the CBD area and toward the points of interest in each district. The results are shown in Figure 17.

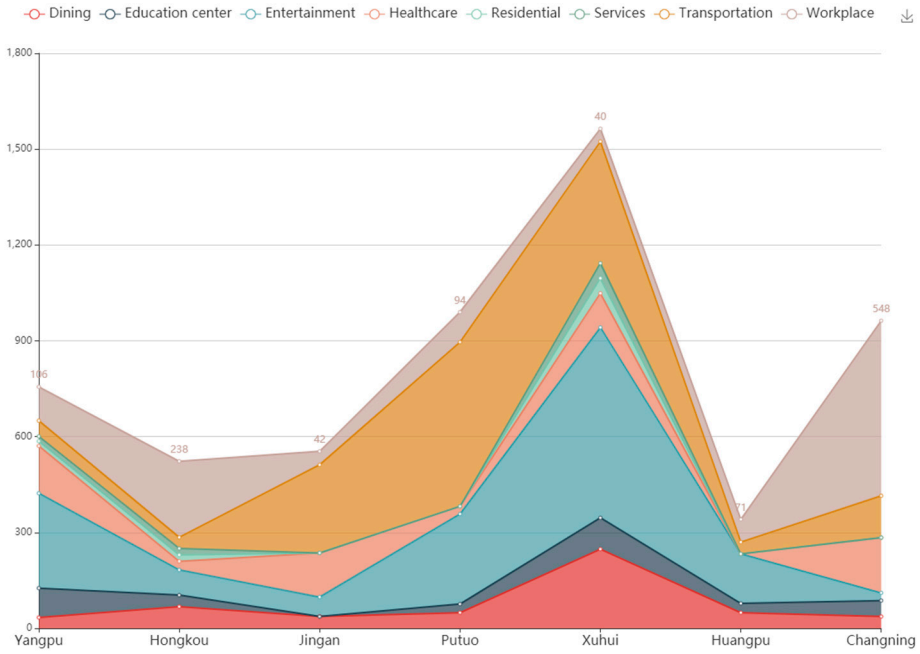


(a)

Figure 17. *Cont.*

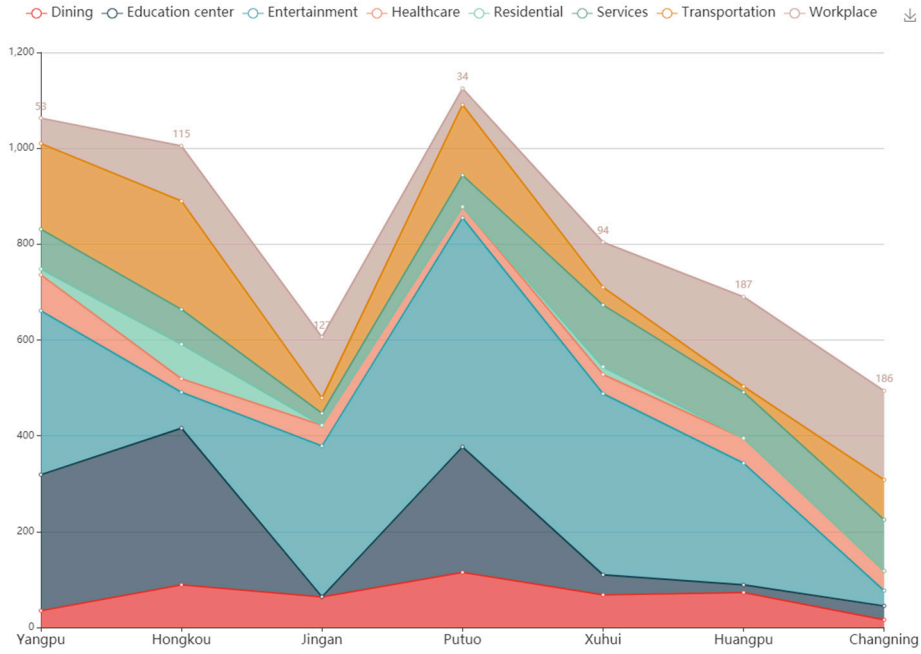


(**b**)



(**c**)

Figure 17. *Cont.*



(**d**)

Figure 17. The POI distribution in the CBD area of Shanghai: (**a**) Baoshan mobility; (**b**) Jiading mobility; (**c**) Minhang mobility (**d**); Pudong mobility.

From the figures, we can understand the following facts that prove the validity of geo-tagged social media data:

1. Transportation-related activities are bold in Putuo (because the western railway station is in this district) from the neighbor district and are less than in Pudong, and this is reasonable because there is a major airport located in Pudong;
2. It can be observed that fewer people move from Baoshan to the CBD, in spite of Pudong. The reason for this is that the population and land area of Baoshan are smaller, and more people move from Pudong.
3. Education activities mostly happen in Yangpu (there are some famous universities there);
4. Another interesting result is that from each neighbor district, the mobility is increased in the border district compared to districts further away. For example, less mobility is seen from Baoshan to Xuhui or Changning.

Moreover, another interesting result achieved by this analysis is the confirmation of different spatial patterns of people moving towards points of interest in the CBD area of Shanghai. Our data demonstrated the tropism of travel in eight categories of the seven districts. As shown in Figure 18, there is a matrix for each trajectory, denoted as $T_{x,y}$, where $x \in \{1,2,3, \ldots, 7\}$ categories of points of interests and $y \in \{1,2,3, \ldots, 8\}$ districts. $T_{x,y}$ in each district represents the travel distance in that specific boundary of the region within a time interval that we chose, namely, two hours, which we considered to be representative for an activity in a day and in order to avoid data redundancy.

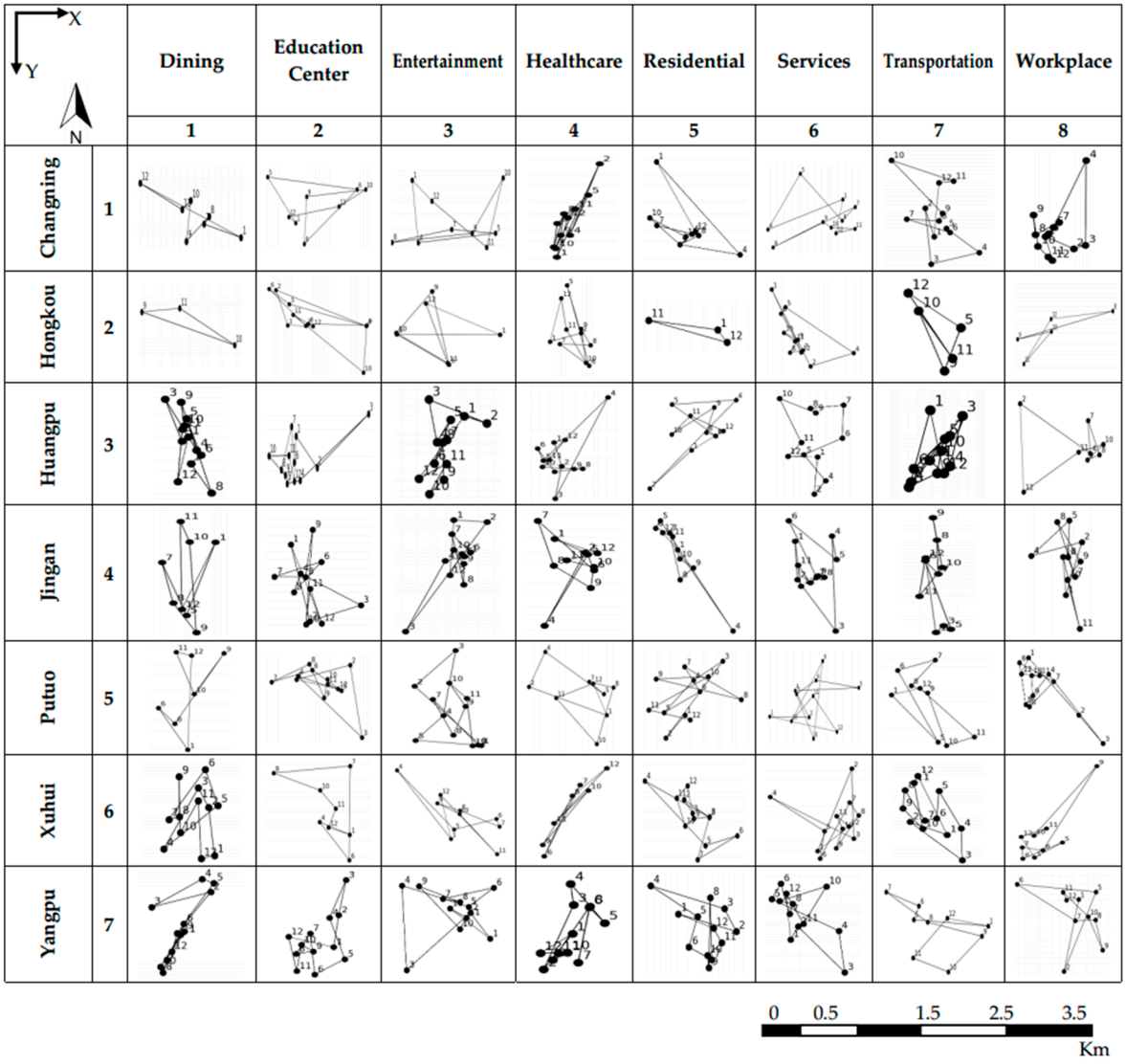


Figure 18. People's movement trajectories of the centroid of the SDE.

The test was successful, as it was able to identify the following findings:

1. Among all travel trajectories, the matrix points out that a larger movement happened in Yangpu ($T_{7,y}$) and Putuo ($T_{5,y}$), but not in Hongkou ($T_{2,y}$), which indicates smaller movement. The reason for this may be that the land area in Yangpu and Putuo is larger than the rest of the districts. ($T_{1,7}$) is a good illustration for the use of transportation in Changning district, as evidenced in the train stations or international airport in this district, followed, respectively, by Putuo ($T_{5,7}$) and Jingan ($T_{4,7}$);
2. Another interesting finding is the travel demand for entertainment activities which indicates that Huangpu ($T_{3,3}$) is a preferred choice compared to Changning and Hongkou, which have a smaller scope. $T_{2,8}$ illustrates this point clearly, where workplace-related check-ins in the Hongkou District require less travel effort. However, it is the opposite case for the Xuhui and Yangpu districts, with a larger $T_{x,y}$ regarding the 'dining' category;
3. The effectiveness of the SDE technique is exemplified in Figure 18, representing the evident direction of movement. For example, the direction of people's movement in Xuhui is almost north-west and movement in Huangpu is nearly toward the south-east. In spite of the fact that LBSN data are not precise enough to estimate the origin–destination of citizens, it is guaranteed

from the results provided by this analysis that they are extremely helpful for urban planning and, additionally, the design of city infrastructure.

6. Discussion

Statistical analysis offers a robust set of techniques and tools for understanding data independently of the topic and domain. In the case of mobility patterns associated with users in social media, this kind of tool provides insights to understand the distribution of users (gender analysis), how and when they interact with each other (temporal analysis), and where they move in a specific region (spatio-temporal analysis). Although these techniques help to understand the way people move and interact, the use of semi-supervised learning algorithms (like clustering) provides a deeper understanding of mobility patterns and how users distribute themselves among different points of interest, ultimately leading to the estimation of trending places and trajectory paths.

The combination of basic statistical tools and grouping algorithms allows researchers and planners to make assumptions about how users interact, which, after using adequate visualization tools, can help to boost decision-making quality in smart cities. Additionally, the adequate duration of location-based social media data provides real-world information that supports decision-making in a fast-growing environment. Most of the applications today analyze the textual content of information for understanding what people think and how they deliver this information to others, but the analysis of spatio-temporal data is gaining momentum, taking into account the benefits associated with the extraction of mobility patterns and how they can be used to increase user experience in a specific region. This presents the following questions: Who is more active? Where do people tend to go? How do they group among each other? These are questions that can be answered with the exploratory experiments presented in this paper, which highlights its relevance. Studying Shanghai, as it is one of the most developed cities, reveals crucial facts that help not only city planners to make better decisions, but also business developers. Revealing the correlation of spatio-temporal distribution helps managers to provide better services in different regions in order to reduce the congestion in a specific area and improve the quality of life. For example, when results reveal a specific percentage of people are moving from district A to district B for healthcare or medical services, that shows people tend to get better services in district B or they are facing a lack of specific services at the origin. Analyzing eight different activities in seven major districts of Shanghai in this research is helpful for urban planners to make precise decisions in designing a better structure for a city. On the other hand, identifying hot spots in the city opens a new aspect for business opportunities. Imagine a company wants to invest its funds to open a new branch in the city; this research helps them to reduce their risk of loss. Compared to recent work by Rizwan et el. [51], in this research work we analyzed human mobility behavior. Different methods and algorithms had been applied on POI data and Sina Weibo Social media data to find the CBD area of Shanghai. This analysis proved to be a useful source of data and validated geo-tagged location-based social media data as suitable research tool for researchers. By clustering similar user's mobility behavior, we found out a variety of citizens' behavior based on their gender and distance from CBD area. Furthermore, we investigated the reasons people tend to go to the CBD area of Shanghai. By creating eight groups of activities, such as dining, entertainment and work-places, we revealed the hidden purposes of mobility. These new insights are useful not only for business developers but also for the government to improve the quality of life in developed cities. Analyzing Weibo data in both aspects of the spatio-temporal human mobility to the CBD area of Shanghai had not been focused on in the previous work. As well as this, users' iterative check-ins from 'home' to CBD area were calculated to measure displacement matrices by Haversine formula. In contrast to [51], which focused only on analyzing the frequency of using LBSN based on gender differences in ten districts of Shanghai, we made four groups of neighboring districts and analyzed user's mobility behavior and their purposes to the CBD area, plus the moving trajectories in Shanghai.

There are many benefits associated with this study (like the ones presented in previous chapters), but there are also some natural limitations, like the availability of temporal data, which, in many social

media channels, is restricted (Twitter, for instance). The results are also highly dependent on the user's interactivity being associated with different points of interest, where a low interaction means a poor understanding of mobility patterns, while a higher interaction means a better perception, which can support smart city planning decisions.

Additional limitations in this study include the user's preferences, relating to how often they update lifetime events (sporadic activities lead to biased conclusions). Users also tend to lie regarding their position and preferences, as social media is not a private hub, thus information could be used for malicious purposes.

To overcome previous limitations, we explore different spatio-temporal algorithms and different preprocessing techniques for the data to refine the conclusions, bearing in mind that not only the frequency of actions, but also the grouping factor of data regarding the mobility of users, is important. The idea is to reach more accurate predictions based on automatic pattern processing, rather than making inferences about variable relationships among data (i.e., classical statistical analysis).

7. Conclusions

Different statistical techniques were applied in order to extract valuable mobility patterns from the Internet. The Obtained results highlight the viability of the proposed methodology as a way to extract spatio-temporal knowledge from users in a trending social media site. Considering the theoretical implications of the presented approach, the following conclusions have been reached.

As a social media channel, Sina Weibo provides a large amount of information that has been proven to be valuable in terms of mobility pattern discovery. Specifically, accurate geo-located interactions increase the quality of the information discovered. Gender analysis over a large amount of social media interactions could be used as a starting point for uncovering fine-grained trends that otherwise may be lost due to the homogeneity of data. This could elucidate answers to questions such as: Who works more hours during the week? Who is more active on social media in a specific place? Those are questions that can be answered with this kind of analysis. Temporal analysis produces different statistics associated with a user's activities over time, helping to uncover the peak trendy hours in a region. The obtained patterns also contribute to a better understanding of user's activities over time and how this affects their interactions on a social media channel. Points of interest provide a relevant source of temporal information for finding patterns like interactions among different places and for keeping track of relevant spatio-temporal user interactions and locations [50].

Experimental results with DBSCAN here provide accurate clusters that expose important locations in user's mobility, filtering noisy points that apparently are important, but not well connected based on distance measures. This ultimately leads to patterns that can be used to discover day to day mobility in a specific region. Here, the KDE algorithm produces insightful analysis of DBSCAN clusters that contribute to the detection of fine-grained groups, based on the density of points. The clusters obtained after applying both algorithms enrich the discovery of mobility patterns by detecting hot spots visited by users. Research on the use of statistics for understanding human mobility patterns continues to be in favor of improving obtained results, keeping in mind the complexity/size of social media data. Ongoing and future work includes the following actions: applying the proposed methodology over different social media channels, considering the opportunities and constraints associated, as well as the availability of open data; using other well-known statistical techniques, adapting a semantical point of view for uncovering mobility patterns associated with spatio-temporal aspects of users on social media; and creating a framework of tools/techniques that can be applied generically to different social media channels to produce valuable information independently of the topic and domain of the data, as well as the complexity of the information.

Supplementary Materials: The following are available online at http://www.mdpi.com/2220-9964/9/2/125/s1.

Author Contributions: Conceiving the idea, Conducting the experiments, Writing the original draft and editing, Zeinab Ebrahimpour; Application of statistical and other formal techniques and Visualization José Luis Velázquez García; Supervision and Provision of study materials, Wanggen Wan; Supervision, Project administration and Proofreading, Ofelia Cervantes and Funding Acquisition, Li Hou. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [Science and Technology Commission of Shanghai Municipality] grant number [18510760300] and [Anhui Natural Science Foundation] grant number [1908085MF178] and [Anhui Excellent Young Talents Support Program Project] grant number [gxyqZD2019069].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Noulas, A.; Scellato, S.; Lambiotte, R.; Pontil, M.; Mascolo, C. A tale of many cities: Universal patterns in human urban mobility. *PLoS ONE* **2012**, *7*, e37027. [CrossRef]
2. Kang, C.; Ma, X.; Tong, D.; Liu, Y. Intra-urban human mobility patterns: An urban morphology perspective. *Phys. A Stat. Mech. Appl.* **2012**, *391*, 1702–1717. [CrossRef]
3. Bernard, A.; Bell, M.; Charles-Edwards, E. Life-course transitions and the age profile of internal migration. *Popul. Dev. Rev.* **2014**, *40*, 213–239. [CrossRef]
4. Treiber, M.; Kesting, A. *Traffic Flow Dynamics: Data, Models Simulation*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 3, p. 54.
5. Liang, X.; Zhao, J.; Dong, L.; Xu, K. Unraveling the origin of exponential law in intra-urban human mobility. *Sci. Rep.* **2013**, *3*, 2983. [CrossRef] [PubMed]
6. Domínguez-Mujica, J.; González-Pérez, J.; Parreño-Castellano, J. Tourism and human mobility in Spanish Archipelagos. *Ann. Tour. Res.* **2011**, *38*, 586–606. [CrossRef]
7. Bao, J.; Zheng, Y.; Mokbel, M.F. Location-Based and Preference-Aware Recommendation Using Sparse Geo-Social Networking Data. In Proceedings of the 20th International Conference on Advances in Geographic Information Systems (ACM), Redondo Beach, CA, USA, 6–9 November 2012; pp. 199–208.
8. Maat, K.; Van Wee, B.; Stead, D. Land use and travel behaviour: Expected effects from the perspective of utility theory and activity-based theories. *Environ. Plan. B Plan. Des.* **2005**, *32*, 33–46. [CrossRef]
9. Chen, C.; Ma, J.; Susilo, Y.; Liu, Y.; Wang, M. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transp. Res. Part C Emerg. Technol.* **2016**, *68*, 285–299. [CrossRef]
10. Zhang, Y.; Wang, L.; Zhang, Y.Q.; Li, X. Towards a temporal network analysis of interactive WiFi users. *Europhys. Lett.* **2012**, *98*, 68002. [CrossRef]
11. Cattuto, C.; Van den Broeck, W.; Barrat, A.; Colizza, V.; Pinton, J.F.; Vespignani, A. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE* **2010**, *5*, e11596. [CrossRef]
12. Siła-Nowicka, K.; Vandrol, J.; Oshan, T.; Long, J.A.; Demšar, U.; Fotheringham, A.S. Analysis of human mobility patterns from GPS trajectories and contextual information. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 881–906. [CrossRef]
13. Zhu, W.Y.; Wang, Y.W.; Chen, C.J.; Peng, W.C.; Lei, P.R. A bayesian-based approach for activity and mobility inference in location-based social networks. In Proceedings of the 17th IEEE International Conference on Mobile Data Management (MDM), Porto, Portugal, 13–16 June 2016; Volume 1, pp. 152–157.
14. Weibo Industry Research and Development Center. Available online: https://m.chyxx.com/view/688522.html (accessed on 2 November 2018).
15. McCowage, M. Trends in China's Capital Account. *RBA Bull.* **2019**, *31*, 5–11.
16. McColl, R.W. *Encyclopedia of World Geography*; Infobase: New York, NY, USA, 2014; Volume 1, p. 160.
17. Drozdz, M.; Appert, M. Re-Understanding CBD: A Landscape Perspective. Available online: https://halshs.archives-ouvertes.fr/halshs-00710644/ (accessed on 2 May 2011).
18. Yang, J.; Zhu, J.; Sun, Y.; Zhao, J. Delimitating urban commercial central districts by combining kernel density estimation and road intersections: A case study in Nanjing city, China. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 93. [CrossRef]
19. Rhee, I.; Shin, M.; Hong, S.; Lee, K.; Kim, S.J.; Chong, S. On the levy-walk nature of human mobility. *IEEE/ACM Trans. Netw.* **2011**, *19*, 630–643. [CrossRef]
20. Liu, Y.; Kang, C.; Gao, S.; Xiao, Y.; Tian, Y. Understanding intra-urban trip patterns from taxi trajectory data. *J. Geogr. Syst.* **2012**, *14*, 463–483. [CrossRef]

21. Wu, L.; Zhi, Y.; Sui, Z.; Liu, Y. Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PLoS ONE* **2014**, *9*, e97010. [CrossRef]
22. Times, V.C.; Venturini, R. Mining Human Mobility Data and Social Media for Smart Services. Ph.D. Thesis, Universidade Federal de Pernambuco, Recife, Brazil, 2019; p. 65.
23. Monteiro de Lira, V.; Renso, C.; Perego, R.; Rinzivillo, S.; Cesario Times, V. The ComeWithMe system for searching and ranking activity-based carpooling rides. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016; pp. 1145–1148.
24. Isaacman, S.; Becker, R.; Cáceres, R.; Kobourov, S.; Martonosi, M.; Rowland, J.; Varshavsky, A. Identifying Important Places in People's Lives from Cellular Network Data. In Proceedings of the International Conference on Pervasive Computing, Seattle, WA, USA, 21–25 March 2011; pp. 133–151.
25. Hartigan, J.A. *Clustering Algorithms*; John Willey Sons: New York, NY, USA, 1975; pp. 113–129.
26. Cheng, Z.; Caverlee, J.; Lee, K.; Sui, D.Z. Exploring Millions of Footprints in Location Sharing Services. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011; p. 82.
27. Ullah, H.; Wan, W.; Haidery, S.A.; Khan, N.U.; Ebrahimpour, Z.; Luo, T. Analyzing the Spatiotemporal Patterns in Green Spaces for Urban Studies Using Location-Based Social Media Data. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 506. [CrossRef]
28. Yan, X.Y.; Zhou, T. Destination choice game: A spatial interaction theory on human mobility. *Sci. Rep.* **2019**, *9*. [CrossRef]
29. Liu, Z.; Yang, C. Exploring Group-Level Human Mobility from Location-Based Social Media Check-in Data. In Proceedings of the 2018 Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS), Wuhan, China, 22–23 March 2018; pp. 1–5.
30. Wang, X.; Ding, J.; Uhlig, S.; Li, Y.; Jin, D. Deviations of Check-ins and Human Mobility Trajectory. In Proceedings of the 2019 5th International Conference on Big Data Computing and Communications (BIGCOM), Qingdao, China, 9–11 August 2019; pp. 115–123.
31. Yang, C.; Xiao, M.; Ding, X.; Tian, W.; Zhai, Y.; Chen, J.; Ye, X. Exploring human mobility patterns using geo-tagged social media data at the group level. *J. Spat. Sci.* **2019**, *64*, 221–223. [CrossRef]
32. Hasan, S.; Zhan, X.; Ukkusuri, S.V. Understanding Urban Human Activity and Mobility Patterns Using Large-Scale Location-Based Data from Online Social Media. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, Chicago, IL, USA, 8 August 2013; p. 6.
33. Ihler, A. Kernel Density Estimation Toolbox for MATLAB. Available online: http://ssg.mit.edu/~ihler/code/ (accessed on 2 November 2005).
34. Hasan, S.; Ukkusuri, S.V. Urban activity pattern classification using topic models from online geo-location data. *Transp. Res. Part C Emerg. Technol.* **2014**, *44*, 363–381. [CrossRef]
35. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
36. Cui, Y.; Xie, X.; Liu, Y. Social media and mobility landscape: Uncovering spatial patterns of urban human mobility with multi source data. *Front. Environ. Sci. Eng.* **2018**, *12*, 7. [CrossRef]
37. Liu, Y.; Sui, Z.; Kang, C.; Gao, Y. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PLoS ONE* **2014**, *9*, e86026. [CrossRef] [PubMed]
38. Ebrahimpour, Z.; Wan, W.; Cervantes, O.; Luo, T.; Ullah, H. Comparison of Main Approaches for Extracting Behavior Features from Crowd Flow Analysis. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 440. [CrossRef]
39. Beiró, M.G.; Panisson, A.; Tizzoni, M.; Cattuto, C. Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Sci.* **2016**, *5*, 30. [CrossRef]
40. Yang, F.; Jin, P.J.; Cheng, Y.; Zhang, J.; Ran, B. Origin-destination estimation for non-commuting trips using location-based social networking data. *Int. J. Sustain. Transp.* **2015**, *9*, 551–564. [CrossRef]
41. Wang, H.; Huang, H.; Ni, X.; Zeng, W. Revealing Spatial-Temporal Characteristics and Patterns of Urban Travel: A Large-Scale Analysis and Visualization Study with Taxi Gps Data. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 257. [CrossRef]
42. Kurkcu, A.; Ozbay, K.; Morgul, E.F. Evaluating the Usability of Geo-Located Twitter as a Tool for Human Activity and Mobility Patterns: A Case Study for NYC. In Proceedings of the Transportation Research Board's 95th Annual Meeting, Washington, DC, USA, 10–14 January 2016; pp. 1–20.
43. Yang, Y.; Heppenstall, A.; Turner, A.; Comber, A. Who, Where, Why and When? Using Smart Card and Social Media Data to Understand Urban Mobility. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 271. [CrossRef]

44. Chiu, C.; Ip, C.; Silverman, A. Understanding social media in China. *McKinsey Q.* **2012**, *2*, 78–81.
45. Bagrow, J.P.; Lin, Y.R. Mesoscopic structure and social aspects of human mobility. *PLoS ONE* **2012**, *7*, e37676. [CrossRef]
46. Wilson, C.M.; Gerard, P. Kernel density estimation for hierarchical data. *J. Commun. Stat. Theory Methods* **2019**. [CrossRef]
47. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd* **1996**, *96*, 226–231.
48. Laylavi, F.; Rajabifard, A.; Kalantari, M. A multi-element approach to location inference of Twitter: A case for emergency response. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 56. [CrossRef]
49. Baojun, W.; Bin, S.; Inyang, H.I. GIS-based quantitative analysis of orientation anisotropy of contaminant barrier particles using standard deviational ellipse. *Soil Sediment Contam.* **2008**, *17*, 437–447. [CrossRef]
50. Loo, B.P.; Lam, W.W.Y. A multilevel investigation of differential individual mobility of working couples with children: A case study of Hong Kong. *Transp. A Transp. Sci.* **2013**, *9*, 629–652. [CrossRef]
51. Rizwan, M.; Wan, W.; Cervantes, O.; Gwiazdzinski, L. Using Location-Based Social Media Data to Observe Check-In Behavior and Gender Difference: Bringing Weibo Data into Play. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 196. [CrossRef]