sustainability

MDPI

Article

Mapping Dynamic Urban Land Use Patterns with Crowdsourced Geo-Tagged Social Media (Sina-Weibo) and Commercial Points of Interest Collections in Beijing, China

Yandong Wang [1,*], Teng Wang [1], Ming-Hsiang Tsou [2], Hao Li [1], Wei Jiang [1] and Fengqin Guo [1]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; wangtengvas@whu.edu.cn (T.W.); lihao@whu.edu.cn (H.L.); jiangweigis@whu.edu.cn (W.J.); 2013286190095@whu.edu.cn (F.G.)

[2] Department of Geography, San Diego State University, San Diego, CA 92182, USA; mtsou@mail.sdsu.edu

* Correspondence: ydwang@whu.edu.cn; Tel.: +86-27-6877-8969

Academic Editor: Tan Yigitcanlar
Received: 29 August 2016; Accepted: 15 November 2016; Published: 21 November 2016

Abstract: In fast-growing cities, especially large cities in developing countries, land use types are changing rapidly, and different types of land use are mixed together. It is difficult to assess the land use types in these fast-growing cities in a timely and accurate way. To address this problem, this paper presents a multi-source data mining approach to study dynamic urban land use patterns. Spatiotemporal social media data reveal human activity patterns in different areas, social media text data reflects the topics discussed in different areas, and Points of Interest (POI) reflect the distribution of urban facilities in different regions. Human activity patterns, topics of discussion on social media, and the distribution of urban facilities in different regions were combined and analyzed to infer urban land use patterns. We collected 9.5 million geo-tagged Chinese social media (Sina-Weibo) messages from January 2014 to July 2014 in the urban core areas of Beijing and compared them with 385,792 commercial Points of Interest (POI) from Datatang (a Chinese digital data content provider). To estimate urban land use types and patterns in Beijing, a regular grid of 400 m × 400 m was created to divide the urban core areas into 18,492 cells. By analyzing the temporal frequency trends of social media messages within each cell using K-means clustering algorithm, we identified seven types of land use clusters in Beijing: residential areas, university dormitories, commercial areas, work areas, transportation hubs, and two types of mixed land use areas. Text mining, word clouds, and the distribution analysis of POI were used to verify the estimated land use types successfully. This study can help urban planners create up-to-date land use patterns in an economic way and help us better understand dynamic human activity patterns in a city.

Keywords: social media; Sina-Weibo; urban land use; POI; text analysis

1. Introduction

The increasing popularity of social media services and smartphones has enabled the public to share their daily activities online and to leave their digital footprint in urban areas. Collecting geo-tagged social media messages with GPS coordinates within urban areas could help researchers understand dynamic spatial-oriented human activities and urban spatial patterns. For example, Tsou et al. (2013) [1] demonstrated a research framework for tracking and analyzing spatial content of social media (Twitter) that can facilitate the tracking of social events (2012 U.S. presidential election) from a spatial-temporal perspective. Liu et al. (2014) [2] used location-based social media data to analyze the underlying patterns of trips and spatial interactions in cities, and revisited spatial interaction and distance decay in spatially-embedded networks. Several scholars used social media, as a crowdsourced spatiotemporal data content, to understand emergency events, enhance emergency situation awareness, and improve the efficiency of emergency response [3–6].

Geo-tagged social media messages and GPS tracking data (from mobile phones and vehicles) have been used to understand human movement behavior and spatiotemporal patterns [7–10]. For example, studying social media "check-in" patterns can provide a better explanation of urban dynamics, as well as a deeper understanding of land use pattern changes [11]. Using mobile telephone positioning data, researchers can analyze the diurnal rhythms of city life and its spatiotemporal differences [12]. Mobile telephone-based sensor data can be used for detecting dynamic urban activities in different time in different cities (Harbin, Paris, and Tallinn) [13]. Using GPS trajectory data from taxi drivers, Liu et al. (2010) [14] revealed taxi drivers' spatial selection of routes and their operation behaviors. Deville (2014) [15] introduced a new approach using mobile phone data to estimate dynamic population densities in near real time.

From an urban planning perspective, researchers have begun to focus on urban area characterization and dynamic patterns associated with various human activities and behaviors [16–19]. Liu et al. (2012) [20] used a seven-day taxi trajectory data to study the relationship between the urban land uses and traffic patterns. Their study shows that human mobility data from smartphones can provide a good estimation for urban land use patterns in a timely fashion, which can help urban planners design better routes for mitigating traffic and improving public services. Frias-Martinez et al. (2014) [21] presented another good case study of land use pattern detection by using location-based Twitter data in Manhattan, London, and Madrid, showing that geo-located tweets can constitute a complementary data source for urban planners. However, there may be a certain bias when using single-source data to detect the urban land use types of the city, especially for large cities in developing countries. The urban land use types of these fast-growing cities are changing rapidly, and different types of land use are mixed together. Analysis from multiple angles is necessary to accurately infer urban land use types of these cities.

In this paper, we introduced a comprehensive analysis framework for detecting dynamic urban land use patterns by using multiple crowdsourced information services, including a popular social media service in China (Sina-Weibo) and a commercial-based Points of Interest (POI) collection (Datatang). By using grid-based aggregation methods for analyzing the spatiotemporal patterns of social media messages, our research goal is to discover the unique characteristics of urban land use types, such as residential areas, commercial areas, work areas, and transportation hubs. Spatiotemporal data analysis algorithms and text mining methods were adopted to identify different types of urban land use patterns. We used Sina-Weibo application programming interfaces (APIs) to collect Chinese geo-tagged social media messages in Beijing City from January 2014 to July 2014 and downloaded commercial POI data collection in Beijing from Datatang.com. By applying a clustering algorithm (K-means), different areas were classified based on the variations in social media message frequency (hourly) temporal trend patterns. Then we estimated urban land use patterns using multiple procedures. First, social media message (Sina-Weibo) temporal trend patterns were used to estimate land use types, such as residential, commercial, or business (office) areas. Second, using text mining algorithms, we can validate the estimated land use patterns by comparing the popular keywords between different categories of land use areas. Third, the distribution of different categories of POI within each land use category can be used to verify detail urban activities associated with different land use types.

2. Related Works in Mapping Urban Dynamics and Land Use

2.1. Mapping Urban Dynamics with Social Media and Social Sensor Data

Social media, as crowdsourced data content providers, have been used in business analytics, knowledge discovery, event detection, and dynamic mapping applications [22]. Some researchers adopted a new term, *social sensor data*, to indicate individual-level crowdsourced geospatial data, such as check-ins (Foursquare), social media messages (Twitter and Sina-Weibo), and online location-based service reviews (Yelp), which contain rich information about spatial interactions and place semantics in local communities. These social sensor data can provide an opportunity for us to understand our socioeconomic environments in urban areas [23]. Spatio-temporal analysis of social sensor data can provide additional insights into how collective social activity shape urban systems [24].

Currently, a large number of scholars study urban dynamics by using taxi GPS dataset [25], cell phone data [26] or social media data [27]. Researchers have found that there is a significant association between commuting activity and land use types [28]. Using Twitter messages, Han et al. (2015) [29] proposed a new analytic method to identify spatiotemporal differences in the level of geographical awareness of Twitter users living in each U.S. city. In order to discover the hidden logic of connections between areas of a city, a new kind of pattern called the C-pattern was revealed by analyzing frequently co-occurring changes in population densities [30]. Yuan et al. (2012) [31] adopted a topic-based inference model to for discovering areas of different functions in a city using both GPS trajectory datasets and POI datasets. Using cell-phone traffic data, a technique was developed for real-time monitoring of population density in an urban area, which could improve the efficiency of urban systems management and planning [32].

2.2. Clustering Algorithms for Land Use

Clustering algorithms can be used to aggregate objects of a collection into groups (classes) based on their similarities [33]. In the field of data mining, there are several representative clustering algorithms, such as density-based spatial clustering of applications with noise (DBSCAN), Expectation-Maximinzation (EM), and K-means. DBSCAN is a well-known implementation of the density-based clustering algorithm that can divide an area at a sufficiently high density into clusters [34]. It requires that the number of objects within a certain area is not less than a given threshold value. DBSCAN algorithm can effectively deal with noises and spatial clustering of arbitrary shapes. However, if the density of the clustered objects is uneven, the clustering effect is poor. For high-dimensional objects, the definition of density is difficult to select.

On the other hand, the K-means algorithm is appropriate for clusters of high-dimensional objects. The K-means algorithm [35] is a typical clustering algorithm based on distance of objects, as the similarity evaluation index. The closer the distances between two objects, the more similar the two objects are. The K-means algorithm is a fast clustering algorithm that can handle large amounts of data efficiently. The number of clusters (k) is an input parameter for K-means algorithm and very important for the quality of clustering results. Gap statistic approach [36] can be used to identify the appropriate number of clusters k (number) for targeted objects. In this research, since we needed to identify the clusters of 24-dimensional objects (using hourly-based social media temporal trend in each day, 24 h), we used K-means for the clustering algorithm to identify various types of land use patterns.

2.3. Text Mining in Social Media

The advance of computational technology over the past decade has enabled the dramatically progress of text mining methods and tools. Text mining is a computational process to understand the content and meaning of text corpora with rich semantics. Text categorization is an important part of the text mining and a process for determining the category of text according to its content. Several representative approaches can be used for text categorization, such as Latent Dirichlet Allocation (LDA), Support Vector Machine (SVM), and Deep Learning.

LDA [37] is a document theme generation model, which is an unsupervised machine learning technique to identify the underlying themes in a large-scale document collection. These themes (topics) can be used to classify each document item into different categories.

SVM [38] is a supervised learning model, which can be used for text classification. For SVM, the low dimensional space of points (keywords) is mapped into a highly dimensional space, so that points (keywords) become linearly separable. The linear division principle can be used to judge classification boundaries for each document (corpus).

Deep Learning refers to an approach that builds a learning neural network to simulate the human brain. Word2vec, provided by Google, is a Deep learning tool example [39]. Using the word2vec (word to vector) web tool, words can be converted into a vector according to the relationship between words and context; the similarity of vector space calculated can indicate the similarity of text semantics. The word2vec tool can help find similar words in a group of documents that together constitute a topic. In our research, word2vec is adopted to reveal the topic hidden in the social media texts.

3. Data Collection and Land Use Type Analysis

3.1. Data Collection

Sina-Weibo, a Twitter-like microblogging system, is the most popular microblogging service in China with 176 million active users monthly in 2013. The amount of daily active Weibo users can reach 4.6 million with about 100 million messages posted every day (Sina-Weibo, 2013). Similar to Twitter's messages, which are called "tweets", Sina-Weibo users can only post their messages with a 140-Chinese-character limit. Each posted message in Sina-Weibo is called "weibo" (microblogs). Sina-Weibo has very similar functions of Twitter, such as retweet (RT), mentioned (@), and hashtags (#). However, the major differences between Sina-Weibo and Twitter are the available content in user profiles. Sina-Weibo collects more optional personal information, such as gender, user locations, birthday, and blood type, in user profiles. Sina-Weibo also provides powerful search engine application programming interfaces (APIs) for collecting and analyzing their microblog messages from third parties. The Sina-Weibo data used in this study was collected during the period from January 2014 to July 2014 within the city of Beijing. Totally, 9.5 million Sina-Weibo messages containing geo-tagged information are collected using Sina-Weibo APIs.

Figure 1 illustrates the average daily (24 h) temporal trend of Sina-Weibo message frequency in Beijing aggregated at one-hour time intervals. There are clear daily fluctuations in the frequency of Sina-Weibo messages. Very few Sina-Weibo messages were generated in the early morning (between 2:00 a.m. and 6:00 a.m.). The lowest frequency time is between 3:00 a.m. and 4:00 a.m. The highest frequency time is at 10:00 p.m. There are relatively high volumes of Sina-Weibo messages were sent between 8:00 a.m. to 6:00 p.m. Lastly, there is a dramatically decreased frequency between 9:00 p.m. to 2:00 a.m. when people are going to sleep at night.

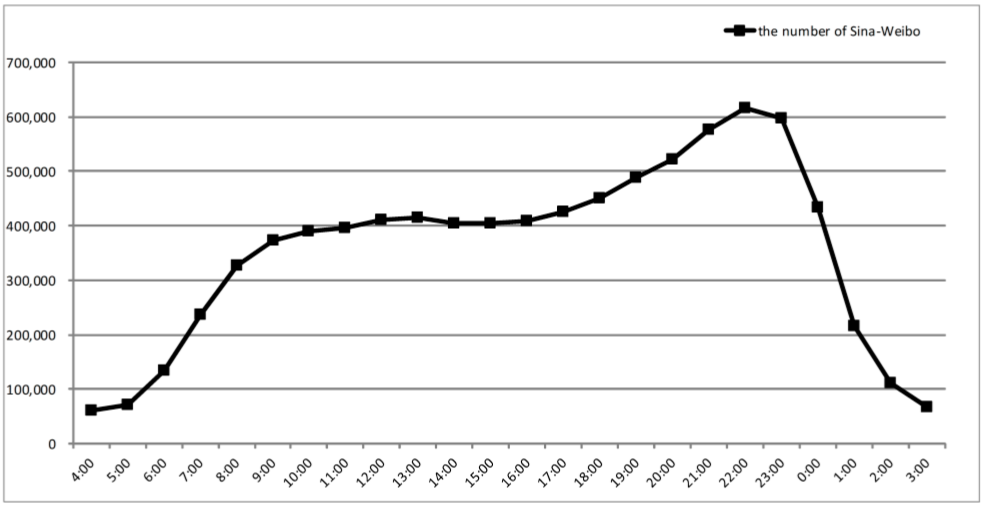


Figure 1. The average daily temporal trend (24 h) of Sina-Weibo message frequency in Beijing (aggregated by one hour time intervals).

3.2. Grid-Based Land Use Segmentation and Aggregated Temporal Trends

In order to estimate the spatial variation of land use patterns, we divided our study area (the central part of Beijing within the red rectangle in Figure 2) into regular grids of 400 m × 400 m. There are two reasons to choose this rectangle as our study area. First, the majority of collected geo-tagged Sina-Weibo messages in Beijing were located within the Sixth Ring Road (the ring road nearby the rectangle box in Figure 2). Second, the area within the Sixth Ring Road is the densely populated urban core of Beijing.

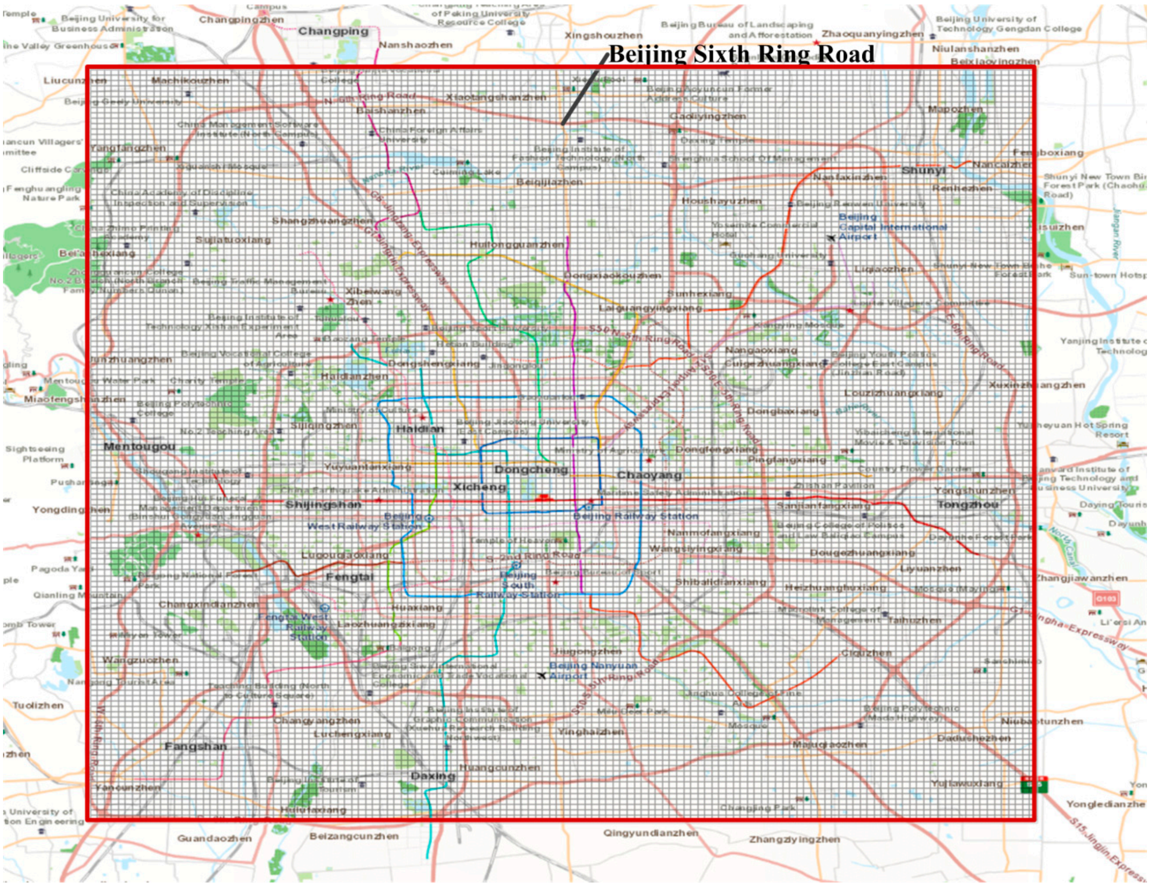


Figure 2. Building a grid-based land use segmentation with 400 m × 400 m grids for the urban core of Beijing.

As shown in Figure 2, our study area consists of 18,492 (134 × 138) cells within the rectangular box. These cells are sequentially numbered from 0 to 18,491 from left to right and from bottom to top. After the grid-based land use segmentation, we calculated the daily temporal trends of the Sina-Weibo messages within each cell. Based on the characteristics of daily temporal trends (24 h time periods as 24 eigenvalues) in each cell, we used the K-means clustering algorithm to classify 18,493 cells into seven different categories (clusters). Cells with similar temporal trend characteristics were classified in the same cluster (land use category).

Within each cell, we counted the total number of Sina-Weibo messages generated within each time period (one hour). Then, we calculated the proportion of the number of Sina-Weibo messages generated within the cell during each time period divided by the total number of messages generated within the cell in all time periods. Therefore, for each cell, we can build a unique daily temporal trend of social media messages for each cell. However, some cells have very few Sina-Weibo messages within each time period and it will be difficult to calculate their daily temporal trends. We decided to remove the cell which containing less than 100 Sina-Weibo messages in total. There are 12,277 cells (out of 18,492 cells) removed from our study and only 6215 cells with high numbers of social media messages are used for our urban land use mapping task. Figure 3 illustrate the distribution of 6215 high frequency cells (color pixels) in the urban core areas. The cells with high frequency of social media messages may indicate the high population density areas in Beijing.

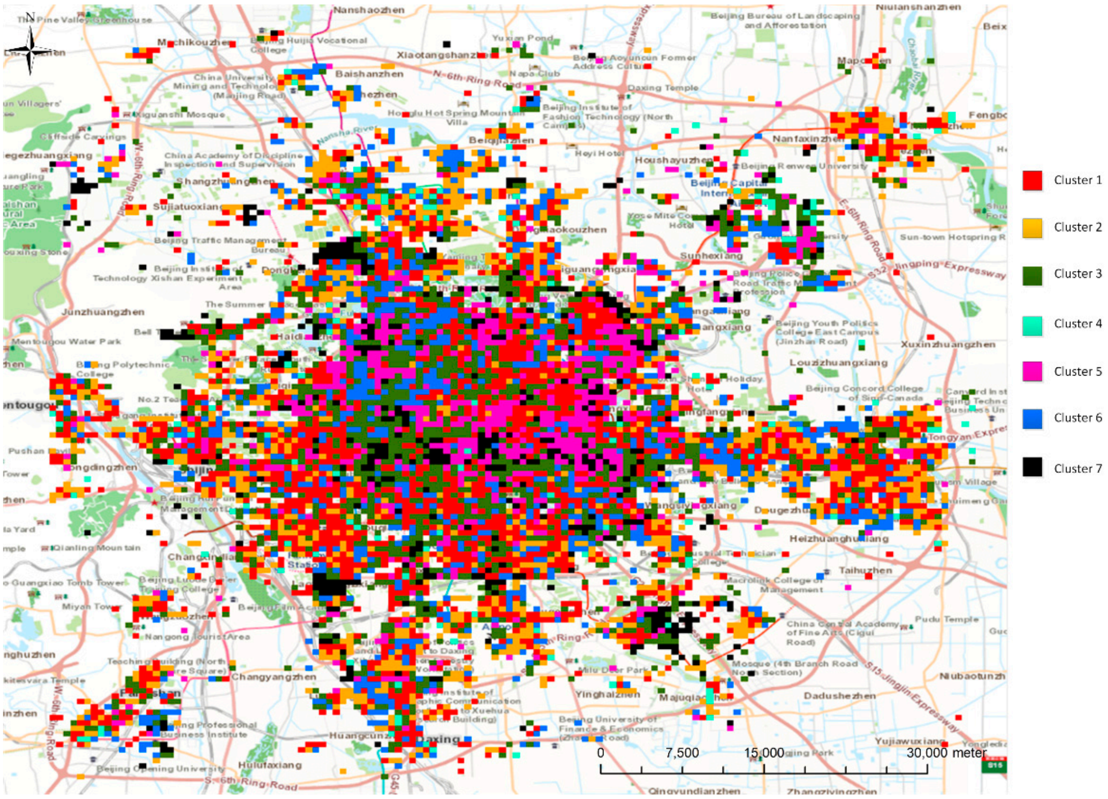


Figure 3. Geographical distributions of the 400 m × 400 m land use cells with high frequency social media messages (containing more than 100 messages in each cell). The color of each cell indicates the clustering results (from Cluster 1 to 7) with in the urban core of Beijing.

To analyze the temporal change of social media messages (posting times) in each cell zone, we divided one day into 24 segments (one hour as one segment). In each cell, we calculated the proportion of posting numbers in each time period to the total number of posting in one day (24 h). Therefore, we can observe the temporal patterns of posting activities in each grid during 24 h. In order to group the cells with similar daily temporal trends for Sina-Weibo message frequency, K-means clustering algorithm was adopted to classify the grids based on the 24 h time periods (one hour per unit), considered as 24 eigenvalues. In order to identify the appropriate number of clusters k, we followed the gap statistic approach. As a result, when the number of clusters was seven, the gap between different clusters is more recognizable, and the gap between the cells in the same cluster is relatively small. Therefore, k = 7 was selected as the input parameter for the K-means algorithm in this study.

These activity temporal patterns, or the Sina-Weibo temporal patterns, for different clusters are shown in Figure 4.

Figure 3 shows the geographical distributions of different clustering results. The distribution of the cells corresponding to each cluster is relatively decentralized and mixed. To further explore the characteristics of each cluster, we used the Fifth Ring Road as a line to divide the urban core areas and the surrounding areas of Beijing. We calculated the percentages of cluster cells in each category and compare the inner and outer urban areas. The results are shown in Table 1. We found that the areas corresponding to Cluster 3 and Cluster 5 are mainly concentrated inside the urban core areas. The areas corresponding to Cluster 2 and Cluster 4 are mostly located in the outer urban area.

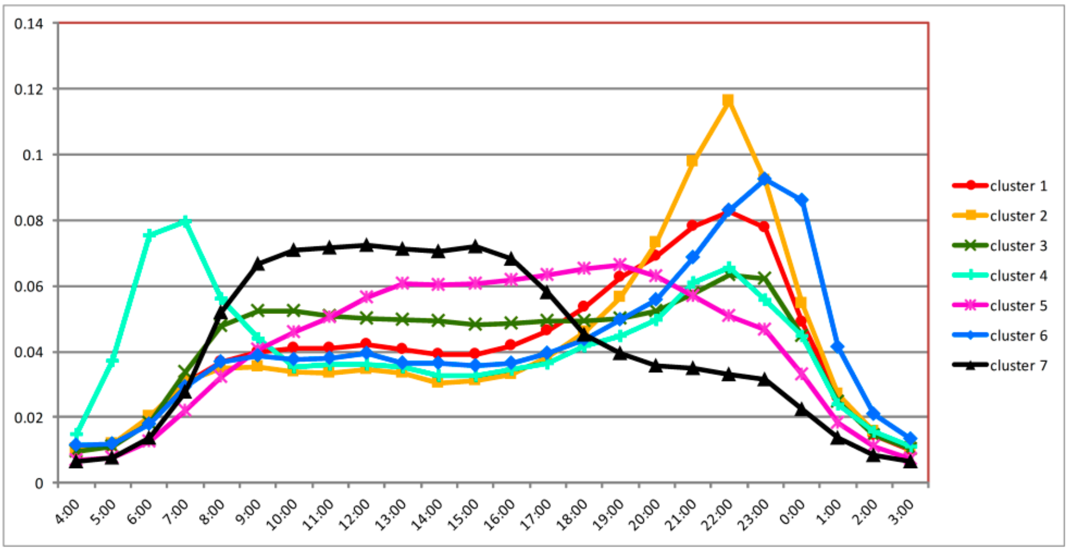


Figure 4. Sina-Weibo daily (24 h) temporal patterns of different clusters using K-means cluster algorithm (k = 7).

Table 1. Comparison of inner and outer urban core areas (divided by the Fifth Ring Road) with different land use clusters in Beijing.

	Total Number of Cells	Total Number of Sina-Weibo	Inner Fifth Ring Road (Urban Core)	Outer Fifth Ring Road
Cluster 1	1738	2,561,789	0.552	0.448
Cluster 2	1034	783,409	0.313	0.687
Cluster 3	1319	2,048,465	0.634	0.366
Cluster 4	136	58,112	0.257	0.743
Cluster 5	634	1,198,237	0.751	0.249
Cluster 6	908	1,071,741	0.504	0.496
Cluster 7	446	577,702	0.596	0.404

3.3. Analysis of Different Clusters with Associated Land Use Types

By analyzing social media message (Sina-Weibo) temporal trend patterns in different areas, we can estimate different types of urban land use in the corresponding areas. In Figure 4, the social media messages from Cluster 6 are mostly generated from 7:00 p.m. to 12:00 a.m., which indicated that most social activities in these regions are relatively happen in the evening. Therefore, we estimated Cluster 6 areas as residential areas. Cluster 7 has a completely different temporal pattern compared to Cluster 6. Social media messages in Cluster 7 are mainly generated from 8:00 a.m. to 5:00 p.m. Therefore, we estimated Cluster 7 as work areas. Another example is Cluster 4, where many social media messages are generated in the morning time (6:00 a.m. to 8:00 a.m.) and evening time (two peaks). We estimated Cluster 4 as transportation hub related areas, such as airports and train stations.

To verify our estimations of each cluster, we used text mining methods to analyze the key messages among each cluster. Ideally, people in work areas are more likely to discuss their work topics or business items in their social media messages. People in the commercial areas are more likely to discuss topics about shopping stores or restaurants. In order to find the key topics within hundreds of social media messages in each cluster, we applied the word2vec for text mining, provided by Google, a learning tool based on Deep Learning. When a word was selected as core vocabulary, some related keywords and correlation coefficients can be obtained by applying word2vec to all Sina-Weibo texts. Collections of these words and core vocabularies represent a topic. The Sina-Weibo texts were originally analyzed in Chinese. In order to make it easier to read for non-Chinese readers, we translated the text mining results from Chinese into English in this paper.

For example, we chose the keyword, "Airport Terminal (note that all keywords have been translated from the original Chinese)", as a core vocabulary. Then we obtained the related keywords and the correlation coefficient from word2vec, as shown in Table 2. The correlation coefficient represents the degree of correlation between the keyword and the core vocabulary: the higher the value of the correlation coefficient, the closer the relationship between the keyword and the core vocabulary.

Table 2. The related keywords associated with the core vocabulary, "Airport Terminal", and their coefficient (keywords were translated from Chinese to English).

Keyword	Coefficient	Keyword	Coefficient	Keyword	Coefficient
Parking apron	0.807	Air China	0.638	Beijing Airport	0.610
Gate	0.803	Lounge	0.637	Alternate	0.608
Border	0.752	Air France	0.636	Waiting	0.607
Executive Lounge	0.744	Airport	0.628	Night flight	0.607
Capital Airport	0.735	Navigation	0.627	Large aircraft	0.604
International Flights	0.680	Aircraft fault	0.627	Waiting room	0.602
Direct flight	0.664	Capital International Airport	0.626	Inbound	0.602
Beijing Capital Airport	0.652	Terminal	0.624	Flight	0.600
High	0.647	Elevator	0.619	Terminal building	0.600
Boarding	0.643	Shandong Airlines	0.617	South Station	0.600
Flights	0.638	Hainan Airlines	0.614	Hainan Airways	0.599
Flight number	0.638	Business Class	0.611	...	...

With the keywords related to the core vocabulary, a Sina-Weibo message can be evaluated for relevance to the core vocabulary (the core topic). If one of these keywords was found in a Sina-Weibo text, the Sina-Weibo message will be identified as relevant to this core topic. The proportions of messages for each topic in different clusters can be calculated. As shown in Figure 5, people in the Cluster 4 areas are more likely to mention keywords related to airport terminals. This result matches our previous estimation that Cluster 4 areas are transportation related areas, such as airports and train stations.

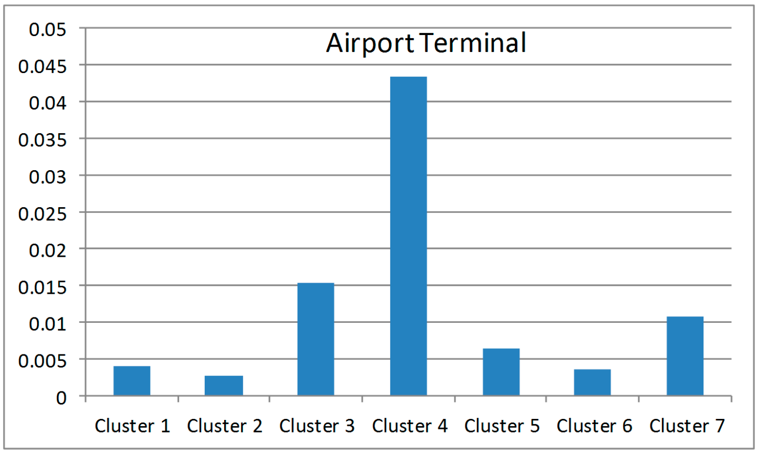


Figure 5. The probability distribution of the airport-terminal-related keywords in different clusters.

Using the same method, we examined several other topics. As shown in Table 3, the keywords in the first column of the table are our core vocabularies which are selected from the high frequency vocabulary of social media messages in each cluster, and those words in the second column of the table are the keywords related to the corresponding core vocabulary generated by word2vec. We discarded the words whose correlation coefficient was less than 0.6. The results are shown in the next section. Due to space limitations, we only show part of the related keywords associated with a core vocabulary.

Table 3. Core vocabulary and related words for different topics.

Core Vocabulary	Related Keywords
Property	Tenants, Property fee, Owner, Sharing, Resident, Construction team, Property Company, Power Supply Bureau, Landlord, Homeowners, Directly Rent, Gas, Rental, Load-bearing walls, Water and electricity, Developers, Illegal buildings, Arbitrary charges, Rental housing, etc.
Dormitory	Study room, Lights Out, Dormitory building, Roommate, Laboratory, House, Aisle, Self-study, Power outage, Classroom, Corridor, Waterhouse, Power failure, Bed, Bedclothes, Back to sleep, Heater, Office, Washbasin, etc.
Library	Study room, Classroom, Reading room, Small classroom, Teaching Building II, Laboratory, Three school buildings, Library, Dormitory, Laboratory building, Light readings, Teaching Building, Peking University Library, School, etc.
Campus	University Campus, School gate, Beijing University, Alma mater, Beijing University of Posts and Telecommunications, Beijing Institute of Technology, Tsinghua University, Beijing University of Science and Technology, Tsinghua Park, etc.
Eating	Dinner, Lunch, Tired of eating, Too hungry, Vegetable dish, Half full, Change to eat, Too full, Quite full, Satiate, Each meal, Supper, Noodles, Eat less, Bowl, Bun, Rice, etc.
Bar	Bar Street, Singing, Cafe, Houhai (place name), Nightlife, Street, Stopover, Drink, Cafe, Pub, Bistro, Stroll, Never sleeps, Good place, Beer, Drum, Belfry, Nightclub, Ambience, Food Street, Quadrangle, Disco, Play, Barbecue, Sachs, etc.
Manager	Executives, Administrative Assistant, Headhunter, Commissioner, Employ, Customer manager, Clerk, Reception, Office, Customer, Project Manager, Foreman, Lobby, Staff, Recruitment, Business Manager, Market, Deputy Chief, etc.
Boss	Proprietress, Recruiting, Colleague, Helper, Furlough, Foreman, Money, The competent, Work number, Leadership, Store manager, Waiter, Staff, Clerk, Service, Cash register, CEO, Plus wages, Company, etc.
Airport Terminal	Parking apron, Gate, Border, Executive Lounge, Capital Airport, International Flights, Direct flight, Beijing Capital Airport, High, Boarding, Flights, Flight number, Air China, Lounge, Air France, Airport, Navigation, Aircraft fault, etc.

3.4. Commercial POI Analysis for the Verification of Land Use Types

Points of Interest (POI) are specific point locations which can be used for location-based services (LBS), such as commercial shops, post offices, and restaurants. Different land use areas may contain different types of POI. For example, commercial areas will have more restaurant and shopping stores POI comparing to the residential areas. We use the distribution of different types of POI among the seven clusters of land use to verify the estimated land use types.

We collected 17 types of POIs relevant to land use types in Beijing from the Datatang, a Chinese online platform and service provider for Big Data sharing and trading. The collected dataset includes 95,588 POIs. Each record contains seven attribute values; CITYCODE, NAME, ADDRESS, TEL, TYPE, X-coordinates and Y-coordinates. There are 182 types of POI found in the attribute TYPE. Table 4 illustrates the list of 17 types of POIs and their associated total numbers for each land use cluster. We also calculated the proportion of each type of POI in each cluster as shown in Equation (1), where N_i represents the number of POI of category i.

$$P_i = \frac{N_i}{\sum_{j=1}^{17} N_j} \tag{1}$$

Table 4. The proportion distribution of 17 types of POI in different land use type clusters.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**
Comprehensive Market	912 (0.0333)	337 (0.0411)	489 (0.0223)	56 (0.0573)	183 (0.0101)	421 (0.0373)	80 (0.0104)
Supermarket	934 (0.0341)	526 (0.0641)	490 (0.0224)	49 (0.0502)	312 (0.0172)	422 (0.0374)	95 (0.0124)
School	1673 (0.0611)	606 (0.0739)	1359 (0.062)	57 (0.0583)	768 (0.0423)	946 (0.0838)	270 (0.0352)
Life service establishments	2168 (0.0792)	701 (0.0855)	1405 (0.0641)	110 (0.1126)	854 (0.047)	796 (0.0705)	274 (0.0358)
Convenience store	2429 (0.0887)	857 (0.1045)	1531 (0.0698)	99 (0.1013)	958 (0.0527)	907 (0.0803)	285 (0.0372)
Beauty salon	2924 (0.1068)	973 (0.1187)	1570 (0.0716)	101 (0.1034)	1437 (0.0791)	1014 (0.0898)	314 (0.041)
Residential region	6491 (0.2371)	2409 (0.2938)	3706 (0.1691)	220 (0.2252)	2047 (0.1127)	2569 (0.2275)	545 (0.0711)
Casual Dining	46 (0.0017)	9 (0.0011)	27 (0.0012)	1 (0.001)	85 (0.0047)	17 (0.0015)	18 (0.0023)
Theater	109 (0.004)	15 (0.0018)	52 (0.0024)	1 (0.001)	139 (0.0077)	23 (0.002)	25 (0.0033)
Cold drink store	99 (0.0036)	20 (0.0024)	73 (0.0033)	5 (0.0051)	208 (0.0115)	44 (0.0039)	41 (0.0054)
Cosmetics shop	154 (0.0056)	42 (0.0051)	83 (0.0038)	3 (0.0031)	189 (0.0104)	45 (0.004)	39 (0.0051)
Mall	289 (0.0106)	66 (0.008)	191 (0.0087)	9 (0.0092)	322 (0.0177)	64 (0.0057)	75 (0.0098)
Sporting goods store	365 (0.0133)	87 (0.0106)	255 (0.0116)	17 (0.0174)	487 (0.0268)	116 (0.0103)	91 (0.0119)
Cafe	279 (0.0102)	36 (0.0044)	348 (0.0159)	7 (0.0072)	714 (0.0393)	138 (0.0122)	246 (0.0321)
Foreign Restaurants	461 (0.0168)	45 (0.0055)	423 (0.0193)	11 (0.0113)	1046 (0.0576)	162 (0.0143)	231 (0.0302)
Company	8003 (0.2924)	1450 (0.1768)	9872 (0.4503)	230 (0.2354)	8396 (0.4622)	3574 (0.3165)	4976 (0.6496)
Industrial Park	37 (0.0014)	21 (0.0026)	48 (0.0022)	1 (0.001)	20 (0.0011)	33 (0.0029)	55 (0.0072)

In Table 4, the proportion of some types of POI is large in each cluster, such as "Company" and "Residential Region". In order to eliminate this discrepancy, for each type of POI, we calculated the proportion of the POI belonging to each cluster for all cases of this type of POI. As shown in Equation (2), where P_i represents the probability of POI belonging to the cluster i for one type of POI.

$$Q_i = \frac{P_i}{\sum_{j=1}^{7} P_j} \tag{2}$$

Thus, we can compare the relative probability between different types of POI. The results will show in the discussion section.

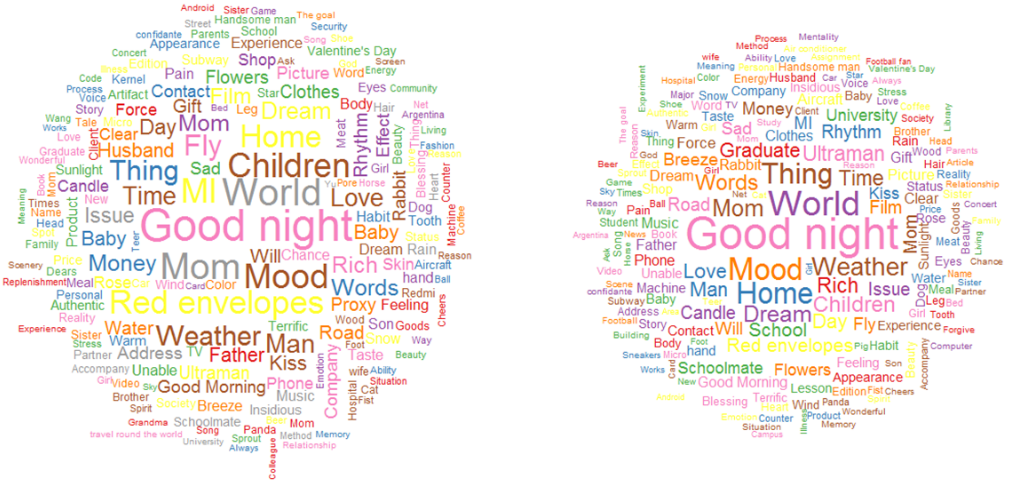
4. Results and Discussion

4.1. Residential Areas (Cluster 2 and Cluster 6)

Among the seven land use clusters, as shown in Figure 4, we found that Cluster 2 and Cluster 6 has similar temporal trends, the social media messages have higher activity frequency from 7:00 p.m. to 12:00 a.m., with a very large activity peak at night between 10:00 p.m. (Cluster 2) and 11:00 p.m.

(Cluster 6). The activity peak in Cluster 6 is an hour later than the activity peak of Cluster 2. We estimated that the two cluster areas 2 and 6 are more likely to be associated with residential areas.

Figure 6 shows the word cloud from all the Sina-Weibo text messages generated within the areas corresponding to Cluster 2 and Cluster 6. The word cloud indicated many residence-related vocabularies, such as “Good night”, “Mood”, “Mom”, and “Home”.

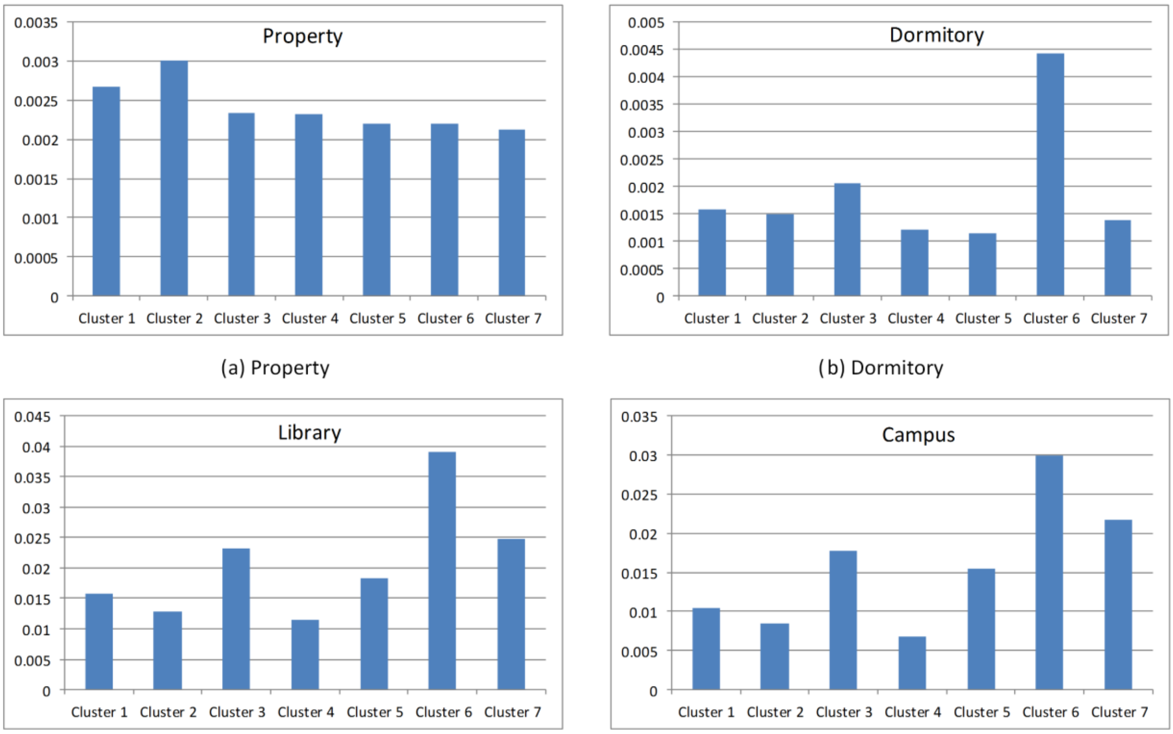


(a) Cluster 2

(b) Cluster 6

Figure 6. The word cloud analysis of: Cluster 2 (**a**); and Cluster 6 (**b**).

In addition, we used text mining methods (word2vec) compare the probability distribution for several core vocabularies (Property, Dormitory, Library, and Campus) in each land use cluster, as shown in Figure 7.



(a) Property

(b) Dormitory

(c) Library

(d) Campus

Figure 7. The probability distribution of different keyword topics in different clusters. (**a**) Property; (**b**) Dormitory; (**c**) Library; and (**d**) Campus.

We found out that Cluster 2 has higher probability of topic “Property” related keywords (see Table 3 for more related keywords). Cluster 6 has higher probability of topics related to “Dormitory” and “Campus”. Therefore, we estimated that Cluster 6 might be university dormitory areas or campus-related residential regions.

We also analyzed the probability of different types of POI within the areas corresponding to the Cluster 2 and Cluster 6. Figure 8 illustrated the POI distribution probability results. The two clusters (2 and 6) have significantly more POIs associated with residential areas, convenience stores, and supermarkets. This result validated our previous estimation that the cluster 2 and 6 are corresponding to residential areas.

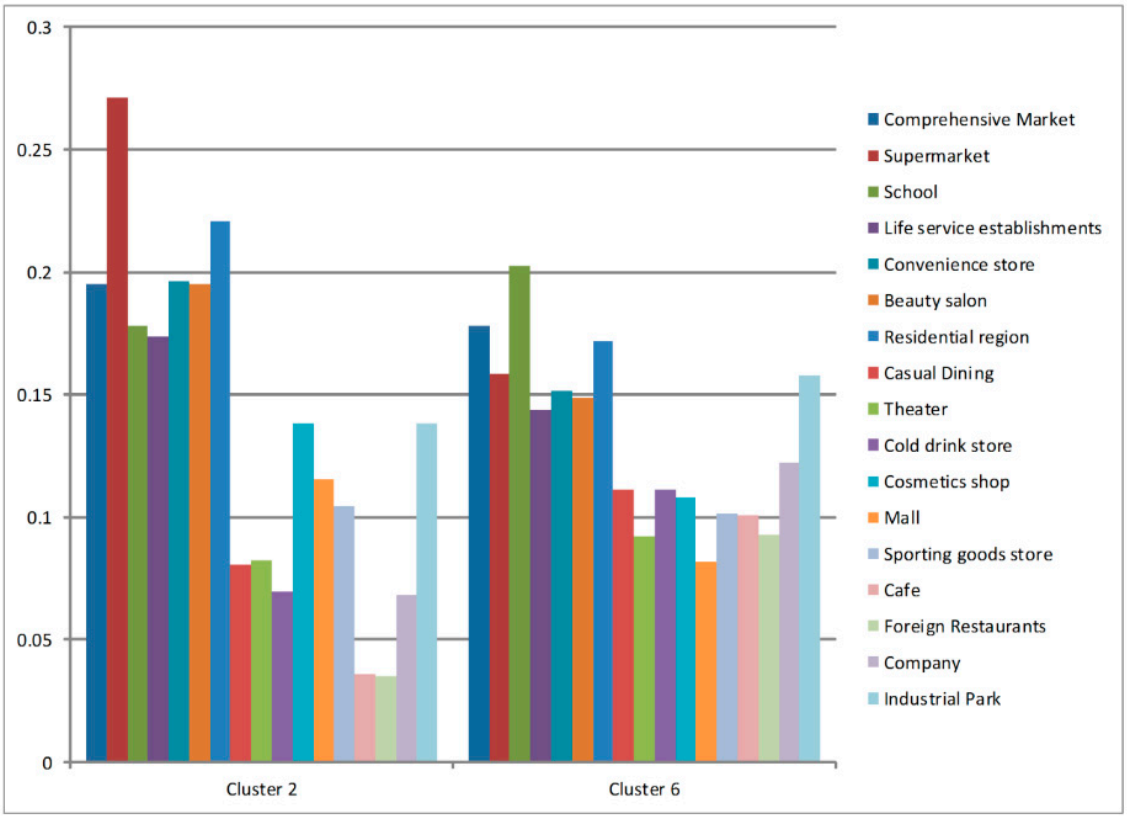
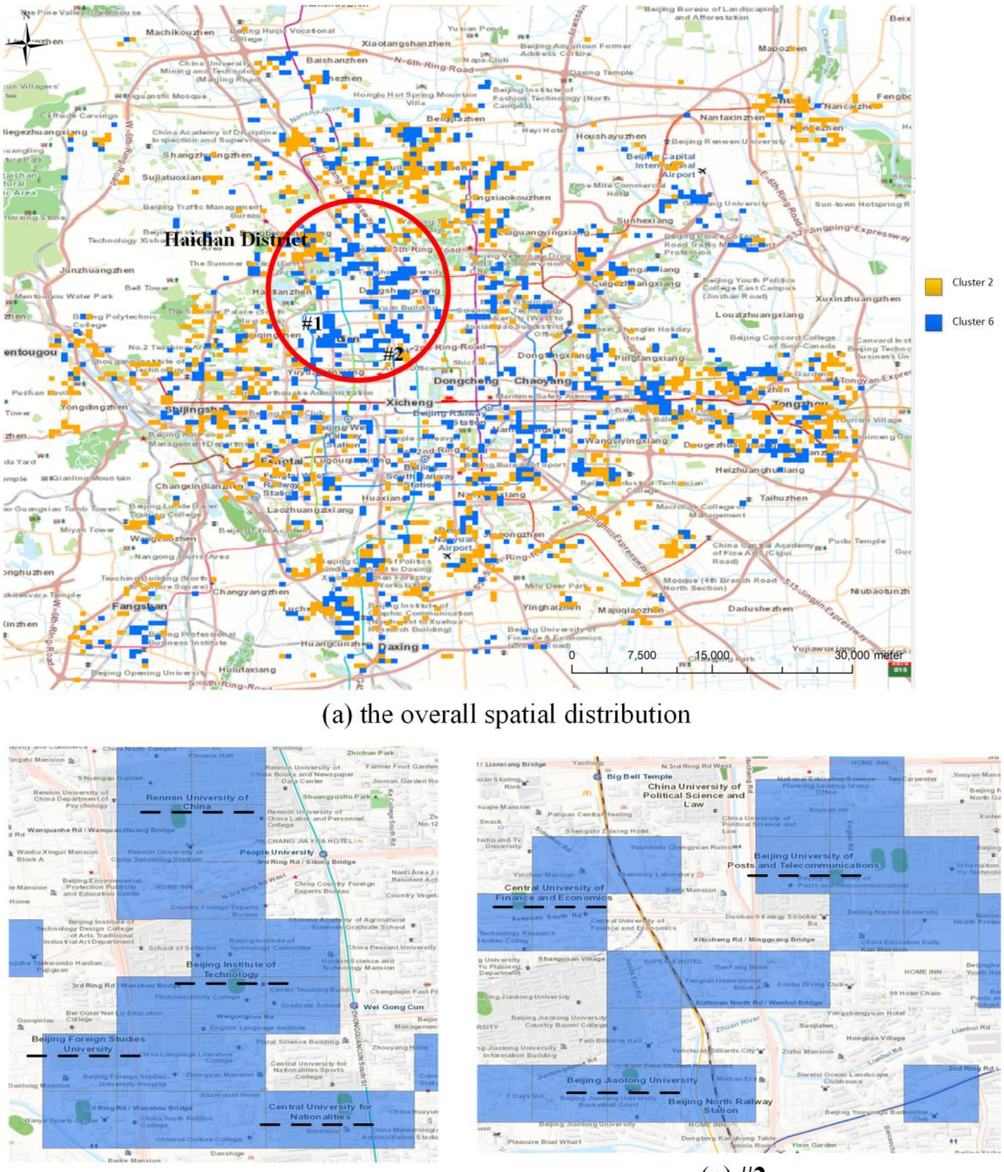


Figure 8. The probability distribution of 17 types of POI in Cluster 2 and Cluster 6.

Figure 9 displays the spatial distribution of Cluster 2 and Cluster 6 cells in Beijing. As shown in Figure 9a, the cells corresponding to the Cluster 2 (yellow color) are distributed in the outer of Fifth Ring Road, while the cells corresponding to the Cluster 6 (blue color) are evenly distributed throughout the inner and outer fifth ring. Moreover, the cells corresponding to the Cluster 6 are concentrated in the Haidian District (the red circle area), which is a well-known region for major universities and science research institutes in Beijing. Figure 9b,c shows the large-scale spatial distribution of areas 1 and 2 in Figure 9a. As can be seen from the figure, the cells corresponding to the Cluster 6 are consistent with the geographical scope of several universities (the black dotted line in the figure, including Renmin University of China, Beijing Institute of Technology, Beijing University of Posts and Telecommunications and so on), which shows that our inference is reliable.


Haidian District
#1
#2
Dongcheng
Chaoyang
Xicheng
Fengtai
Shijingshan
Daxing
0 7,500 15,000 30,000 meter
Cluster 2
Cluster 6
Renmin University of China
Beijing Institute of Technology
Beijing Foreign Studies University
Central University for Nationalities
China University of Political Science and Law
Beijing University of Posts and Telecommunications
Central University of Finance and Economics
Beijing Jiaotong University
Beijing North Railway Station

(a) the overall spatial distribution

(b) #1

(c) #2

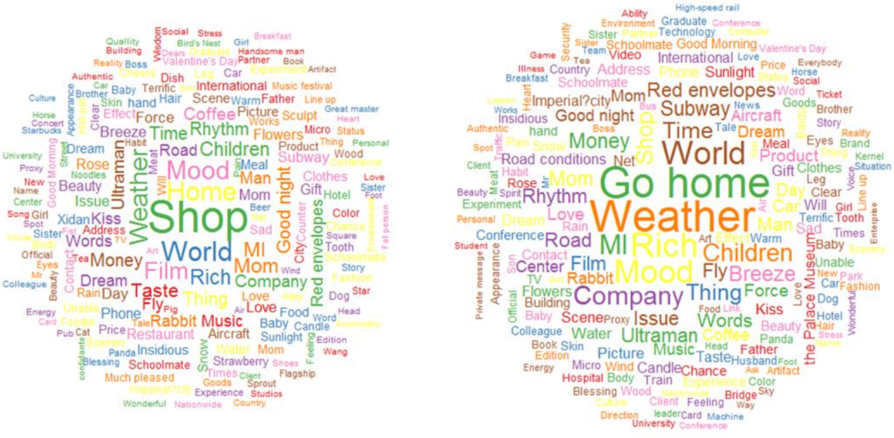
Figure 9. The spatial distribution of Cluster 2 and Cluster 6 areas in Beijing. (**a**) The overall spatial distribution; (**b**) areas 1; (**c**) areas 2.

4.2. Commercial Areas and Work Areas (Cluster 5 and Cluster 7)

In Figure 4, we can find that Cluster 5 started the higher message activities from 8:00 a.m. to 12:00 p.m. and Cluster 7 has a sharper increase of activities between 6:00 a.m. to 9:00 a.m. Then, the number starts to decline at 4:00 p.m. Considering the temporal patterns of social media messages in Cluster 5 and Cluster 7, we estimated that Cluster 5 is more likely to be associated with the commercial areas for dining, shopping and entertainment. Cluster 7 is more related to work areas or business office areas.

Figure 10 illustrates the word cloud from all the Sina-Weibo text generated within the areas corresponding to Cluster 5 and Cluster 7. This figure shows that commerce-related vocabularies, such as “Shop”, “Restaurants”, and “Taste”, were often mentioned in areas corresponding to Cluster 5 (Figure 10a). The work-related vocabularies, such as “go home”, “company”, and “rich”, were frequently discussed in Cluster 7 areas.

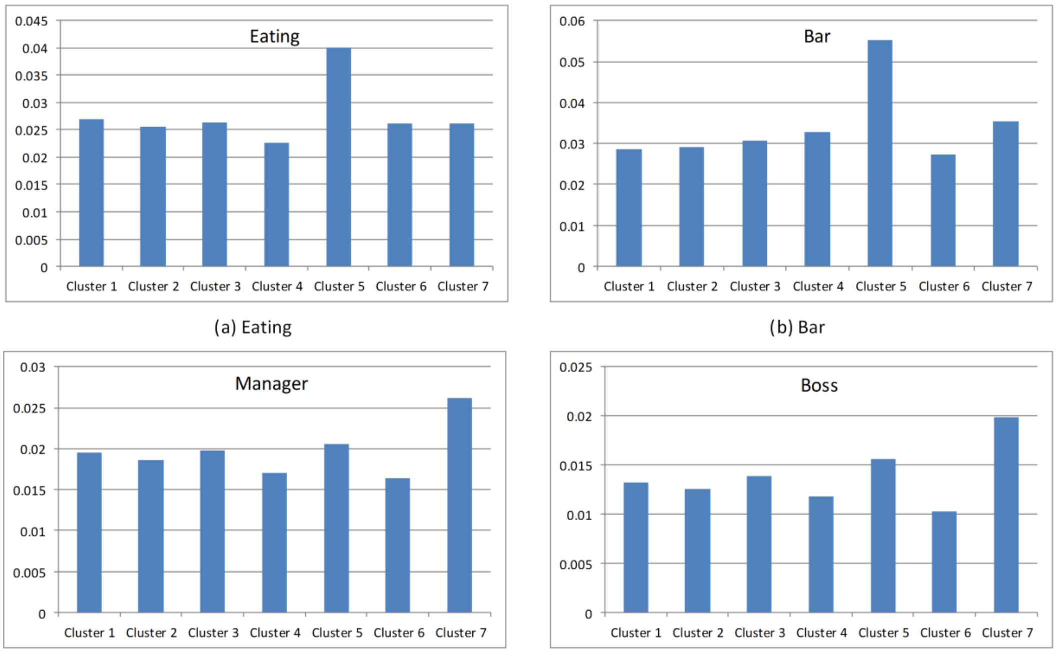
We also used text mining methods (word2vec) to compare the probability distribution for several core vocabularies (Eating, Bar, Manager, and Boss) in each land use cluster, as shown in Figure 11. We found that Cluster 5 messages have more association with the two vocabulary topics, "Eating" and "Bar".



(a) Cluster 5

(b) Cluster 7

Figure 10. The word cloud analysis of: Cluster 5 (**a**); and Cluster 7 (**b**).



(a) Eating

(b) Bar

(c) Manager

(d) Boss

Figure 11. The probability distribution of different keyword topics in different clusters. (**a**) Eating; (**b**) Bar; (**c**) Manager; (**d**) Boss.

As shown in Figure 12, Cluster 5 has higher proportions of POI numbers for "Casual dining", "Malls", "Sporting goods stores", "Cinema", and "foreign restaurants". On the other hand, Cluster 7 has higher proportions of POI related to work areas, such as "Company", "Industrial Park", in the areas corresponding to Cluster 7. This analysis result is consistent with our previous estimation of land use types for Clusters 5 and 7.

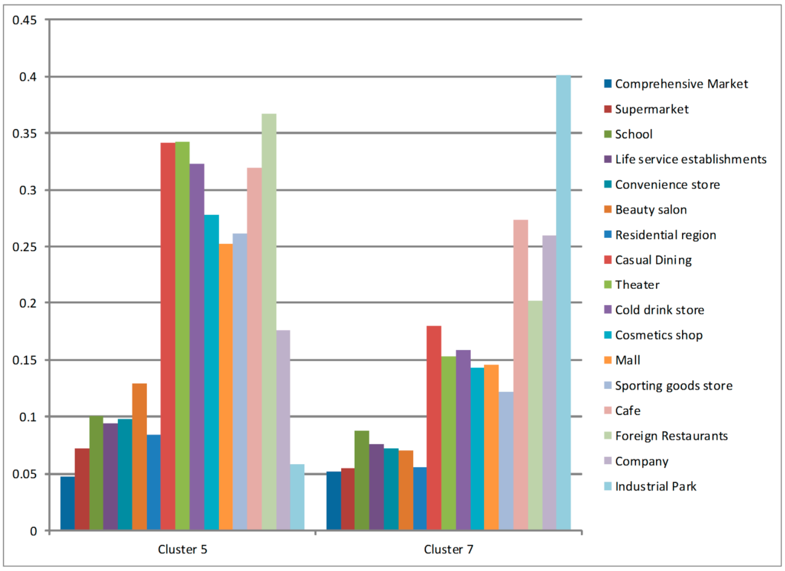


Figure 12. The probability distribution of 17 types of POI in: Cluster 5; and Cluster 7.

Figure 13 displays the spatial distribution of Cluster 5 cells in Beijing. As shown in Figure 13, we can find that the cells corresponding to the Cluster 5 contain the core business district of Beijing, such as Xidan business district, Wangfujing business district and Chaowai business district. Xidan business district and Wangfujing business district are well-known traditional business district with a long history in Beijing, while Chaowai business district is the representative of Beijing's emerging business district. This further confirms our inferences that Cluster 5 cells are more likely to be associated with the commercial areas for dining, shopping and entertainment.

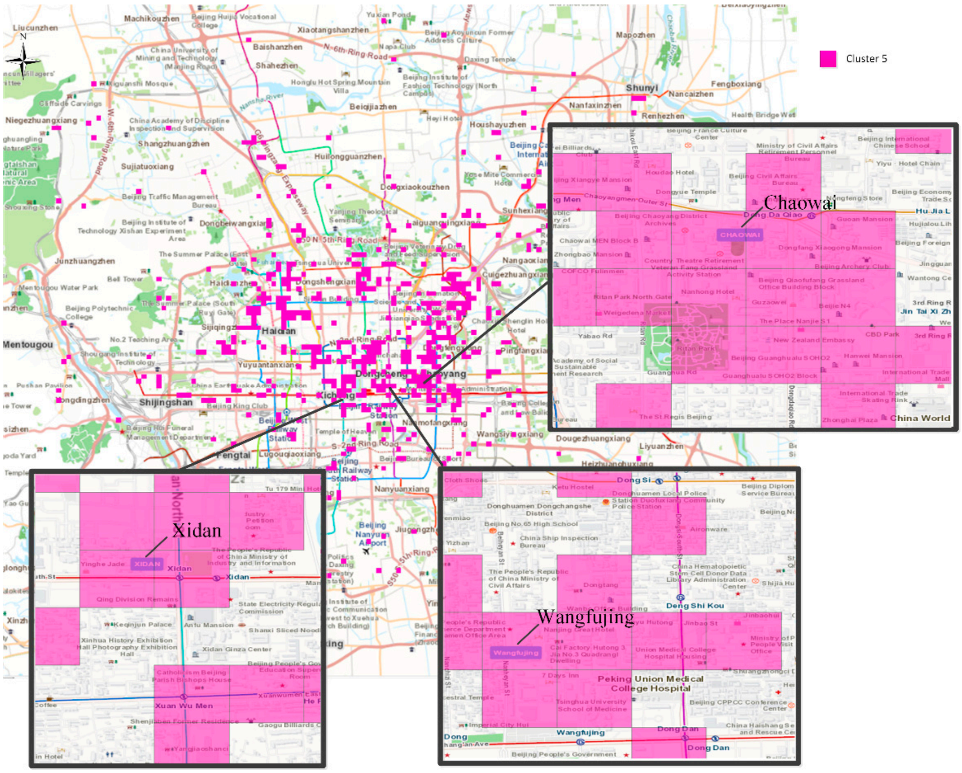


Figure 13. The spatial distribution of Cluster 5 areas in Beijing.

4.3. Transportation Hub Areas (Cluster 4)

In our previous discussion (Section 3.3), we already estimated that Cluster 4 is related to transportation hubs, such as airports and train stations. Figure 4 shows that the activity peaks of Cluster 4 are between 7:00 a.m. to 8:00 a.m. and between 8:00 p.m. to 11:00 p.m. The spatial analysis of the Cluster 4 locations revealed that they are around the airport, railway station or subway stations.

4.4. Mixed Land Use Areas (Cluster 1 and Cluster 3)

Many cities in China have mixed land use types in urban regions. For example, some shopping malls can be built in residential areas and some residents can live in the industrial areas or commercial areas. This type of mixed land use can be found in Cluster 1 and Cluster 3. We estimated that Cluster 1 and Cluster 3 are mixed land use areas. The characteristics of the temporal patterns for Cluster 1 and Cluster 3 are shown in Figure 4. The probability distribution of 17 types of POI (Figure 14) also shows that the various types of POI are relatively mixed and equally distributed in both Cluster 1 and Cluster 3 areas.

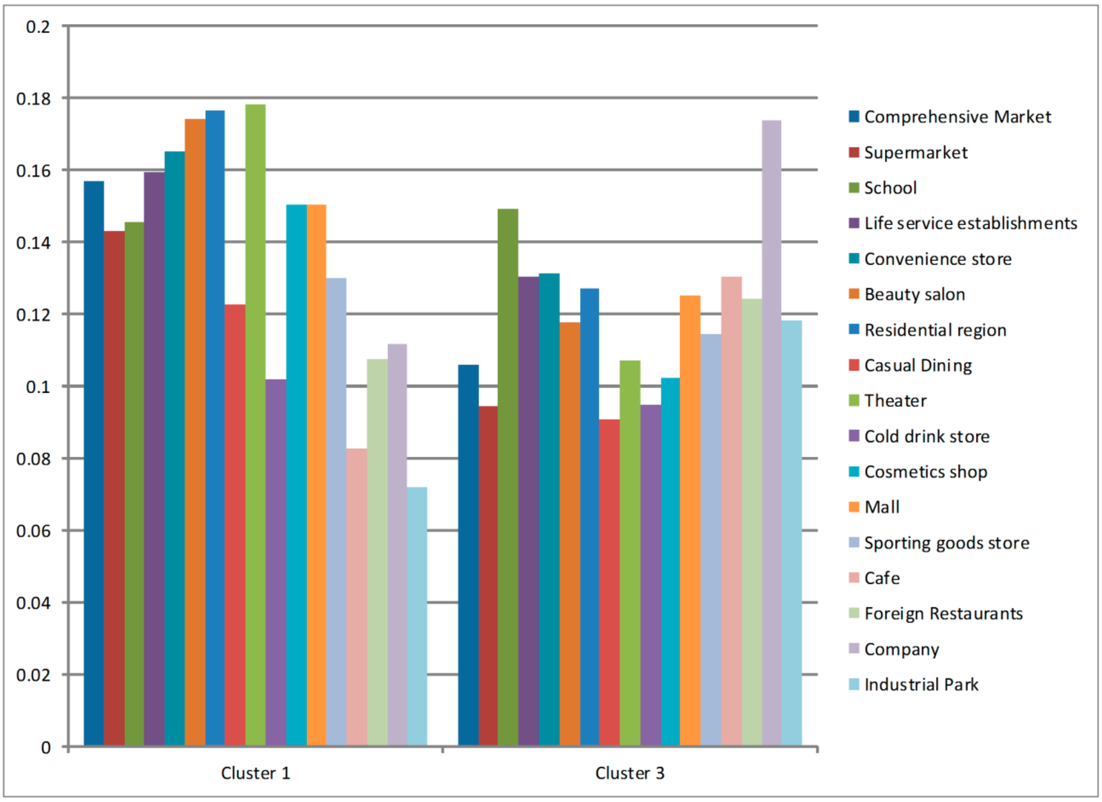


Figure 14. The probability distribution of 17 types of POI in Cluster 1 (mixed land use) and Cluster 3 (mixed land use).

5. Conclusions and Future Works

Classifying urban land use areas is an important task for urban planning and urban design fields. Along with the rapid development of urban regions and the quick increase of population in cities, it is very difficult and expensive to acquire up-to-date land use maps for urban planners or city design specialists, especially in many developing counties, such as China. The proposed crowdsourced mapping frameworks in this paper present a new perspective to explore urban land use patterns and to display up-to-date urban activity spatial patterns.

Social media data is time-sensitive, which can reflect the status in near real-time. More importantly, it is popular participation, cheap and easy to obtain. The spatial-temporal characteristics of social media data can be used to explore up-to-date urban activity spatial patterns. Furthermore, hiding topic in the social media text data and a commercial collection of POI can be combined to reveal the urban land use patterns.

Our approach is to use geo-tagged social media messages (Sina-Weibo) as a crowdsourced map data provider. We utilized the public APIs to collect six months of geo-tagged social media messages in Beijing (total 9.5 million messages collected). There are five key steps in our crowdsourced analysis and mapping framework. First, a regular grid of 400 m × 400 m was created to divide the urban core of Beijing into 18,492 cells. Second, we calculated the numbers of social media messages within each cell and their temporal frequency trends. Third, we only kept the cells with high frequency of social media activities and used the K-means to categorize these cells into seven types of land use clusters. Fourth, we applied text mining approach and word clouds to verify our estimation of land use type for each cluster. Fifth, we used a commercial collection of POI to exanimate the relevance of associate POI types in each land use cluster. We also found some research challenges in this study. First, our methods only focus on 2D dimensions of urban land use rather than 3D dimensions. In a big city, like Beijing, multiple urban land use types (such as residential and commercial) are mixed within high-rise buildings and concentrated housing areas. Urban land use types are more dynamic and mixed in many Chinese cities. By analyzing the fluctuations and message content in social media over time and space, we may be able to monitor the dynamic and mixed features of urban land use patterns in big Chinese cities. Second, we only use a pre-defined grid system (400 m × 400 m) for our urban land use spatial analysis unit. The size was adopted by following previous research works. However, we may need to examine the sensitivity of our methods by using different size of grids, such as 800 m × 800 m or 200 m × 200 m in the future. Third, we only considered the temporal trends of social media messages by combining all messages within a cell. We may need to apply some linguistic methods to classify different types of social media messages first and to remove some errors and noises before creating the temporal trend graph. Finally, we only collected the social media data from January 2014 to July 2014 (six months). The temporal trend patterns might be changed in different seasons or months. If we can collect the whole-year datasets, we can compare the dynamic changes of land use patterns between Summer season and Winter season in Beijing and these dynamic changes might provide more useful information for urban planning.

This research provides a new way to study urban land use patterns in a city from a multidisciplinary perspective. We combine multiple research methods, including GIS, text mining (Deep Learning), K-means, word clouds, and other visualization tools, to explore the dynamic relationships between the temporal trend patterns of social media messages and the land use patterns in Beijing. Social media data, as a major crowdsourced data source, can be linked and aggregated into multiple map layers and GIS datasets for multiple purposes [40]. The fundamental concept of "map overlay" is applied here to combine, integrate, and cross-reference multiple data sources together (including geo-tagged social media, texts, and commercial POI data) and explore their dynamic spatiotemporal patterns in maps. We hope that our new method can be used for other types of applications for urban development in the future, such as hourly population density estimations for disaster responses, site selection for commercial companies, business potential analytics, etc.

Acknowledgments: This work has been supported by the China Special Fund for Surveying, Mapping and Geoinformation Research in the Public Interest (Grant No. 201512015), the National Key Research and Development Program of China (No. 2016YFB0501403) and the National Natural Science Foundation of China (Grant No. 41271399).

Author Contributions: In this paper, Yandong Wang designed research and wrote the paper, Teng Wang performed research, analyzed the data and wrote the paper, Ming-Hsiang Tsou co-designed research and extensively updated the paper, Hao Li, Wei Jiang and Fengqin Guo developed some earlier prototypes. All authors read and approved the final manuscript.

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

References

1. Tsou, M.-H.; Leitner, M. Visualization of social media: Seeing a mirage or a message? *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 55–60. [CrossRef]
2. 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]
3. Yates, D.; Paquette, S. Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. *Int. J. Inf. Manag.* **2011**, *31*, 6–13. [CrossRef]
4. Yin, J.; Lampert, A.; Cameron, M.; Robinson, B.; Power, R. Using social media to enhance emergency situation awareness. *IEEE Intell. Syst.* **2012**, *27*, 52–59. [CrossRef]
5. Sakaki, T.; Okazaki, M.; Matsuo, Y. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 919–931. [CrossRef]
6. Bakillah, M.; Li, R.-Y.; Liang, S.H.L. Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: The case study of typhoon Haiyan. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 258–279. [CrossRef]
7. Song, C.; Koren, T.; Wang, P.; Barabási, A.-L. Modelling the scaling properties of human mobility. *Nat. Phys.* **2010**, *6*, 818–823. [CrossRef]
8. Cho, E.; Myers, S.A.; Leskovec, J. Friendship and mobility: User movement in location-based social networks. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 1082–1090.
9. Phithakkitnukoon, S.; Smoreda, Z.; Olivier, P. Socio-geography of human mobility: A study using longitudinal mobile phone data. *PLoS ONE* **2012**, *7*, e39253. [CrossRef] [PubMed]
10. Xiao, G.; Juan, Z.; Zhang, C. Travel mode detection based on GPS track data and Bayesian networks. *Comput. Environ. Urban Syst.* **2015**, *54*, 14–22. [CrossRef]
11. Preoţiuc-Pietro, D.; Cohn, T. Mining user behaviours: A study of check-in patterns in location based social networks. In Proceedings of the 5th Annual ACM Web Science Conference, Paris, France, 2–4 May 2013; pp. 306–331.
12. Ahas, R.; Aasa, A.; Silm, S.; Tiru, M. Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. *Transp. Res. C Emerg. Technol.* **2010**, *18*, 45–54. [CrossRef]
13. Ahas, R.; Aasa, A.; Yuan, Y.; Raubal, M.; Smoreda, Z.; Liu, Y.; Ziemlicki, C.; Tiru, M.; Zook, M. Everyday space-time geographies: Using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 2017–2039. [CrossRef]
14. Liu, L.; Andris, C.; Ratti, C. Uncovering cabdrivers' behavior patterns from their digital traces. *Comput. Environ. Urban Syst.* **2010**, *34*, 541–548. [CrossRef]
15. Deville, P.; Linard, C.; Martin, S.; Gilbert, M.; Stevens, F.R.; Gaughan, A.E.; Blondel, V.D.; Tatem, A.J. Dynamic population mapping using mobile phone data. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 15888–15893. [CrossRef] [PubMed]
16. Hasan, S.; Ukkusuri, S.V. Urban activity pattern classification using topic models from online geo-location data. *Transp. Res. C Emerg. Technol.* **2014**, *44*, 363–381. [CrossRef]
17. Lee, R.; Wakamiya, S.; Sumiya, K. Urban area characterization based on crowd behavioral lifelogs over Twitter. *Pers. Ubiquitous Comput.* **2012**, *17*, 605–620. [CrossRef]
18. Ferrari, L.; Rosi, A.; Mamei, M.; Zambonelli, F. Extracting urban patterns from location-based social networks. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Chicago, IL, USA, 1 November 2011; pp. 9–16.
19. Adnan, M.; Leak, A.; Longley, P. A geocomputational analysis of twitter activity around different world cities. *Geo-Spat. Inf. Sci.* **2014**, *17*, 145–152. [CrossRef]
20. Liu, Y.; Wang, F.; Xiao, Y.; Gao, S. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landsc. Urban Plan.* **2012**, *106*, 73–87. [CrossRef]
21. Frias-Martinez, V.; Frias-Martinez, E. Spectral clustering for sensing urban land use using Twitter activity. *Eng. Appl. Artif. Intell.* **2014**, *35*, 237–245. [CrossRef]
22. Tsou, M.-H.; Leitner, M. Visualization of social media: Seeing a mirage or a message? *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 55–60. [CrossRef]

23. Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; Shi, L. Social sensing: A new approach to understanding our socioeconomic environments. *Ann. Assoc. Am. Geogr.* **2015**, *105*, 512–530. [CrossRef]
24. Sagl, G.; Resch, B.; Hawelka, B.; Beinat, E. From social sensor data to collective human behaviour patterns: Analysing and visualising spatio-temporal dynamics in urban environments. In GI-Forum 2012: Geovisualization, Society and Learning, Salzburg, Austria, 3–6 July 2012.
25. Qi, G.; Li, X.; Li, S.; Pan, G.; Wang, Z.; Zhang, D. Measuring social functions of city areas from large-scale taxi behaviors. In Proceedings of the Pervasive Computing and Communications Workshops (PERCOM Workshops), Seattle, WA, USA, 21–25 March 2011; pp. 21–25.
26. Soto, V.; Frias-Martinez, E. Robust land use characterization of urban landscapes using cell phone data. In Proceedings of the 1st Workshop on Pervasive Urban Applications, San Francisco, CA, USA, 12–15 June 2011.
27. Fujisaka, T.; Lee, R.; Sumiya, K. Exploring urban characteristics using movement history of mass mobile microbloggers. In Proceedings of the 11th Workshop on Mobile Computing Systems & Applications, Annapolis, MD, USA, 22–23 February 2010; pp. 13–18.
28. Antipova, A.; Wang, F.; Wilmot, C. Urban land uses, socio-demographic attributes and commuting: A multilevel modeling approach. *Appl. Geogr.* **2011**, *31*, 1010–1018. [CrossRef]
29. Han, S.Y.; Tsou, M.H.; Clarke, K.C. Do global cities enable global views? Using twitter to quantify the level of geographical awareness of U.S. Cities. *PLoS ONE* **2015**, *10*, e0132464. [CrossRef] [PubMed]
30. Trasarti, R.; Olteanu-Raimond, A.-M.; Nanni, M.; Couronné, T.; Furletti, B.; Giannotti, F.; Smoreda, Z.; Ziemlicki, C. Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. *Telecommun. Policy* **2015**, *39*, 347–362. [CrossRef]
31. Yuan, J.; Zheng, Y.; Xie, X. Discovering areas of different functions in a city using human mobility and POI. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 186–194.
32. Pulselli, R.M.; Romano, P.; Ratti, C.; Tiezzi, E. Computing urban mobile landscapes through monitoring population density based on cell-phone chatting. *Int. J. Des. Nat. Ecodyn.* **2008**, *3*, 121–134. [PubMed]
33. Mak, K.F.; McGill, K.L.; Park, J.; McEuen, P.L. Valleytronics. The valley Hall effect in MoS(2) transistors. *Science* **2014**, *344*, 1489–1492. [CrossRef] [PubMed]
34. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; AAAI Press: Palo Alto, CA, USA, 1996; pp. 226–231.
35. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A k-means clustering algorithm. *Appl. Stat.* **1979**, *28*, 100–108. [CrossRef]
36. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2001**, *63*, 411–423. [CrossRef]
37. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
38. Suykens, J.A.K.; Vandewalle, J. Least squares support vector machine classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [CrossRef]
39. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the Workshop at International Conference on Learning Representations, Scottsdale, AZ, USA, 2–4 May 2013.
40. Tsou, M.-H. Research challenges and opportunities in mapping social media and Big Data. *Cartogr. Geogr. Inf. Sci.* **2015**, *42*, 70–74. [CrossRef]