# Spatial similarity and features of contents on online social media

--Based on data from Sina（Weibo：a popular social media like Twitter and Facebook in China) microblog in Beijing, and the LDA-VSM model

## Table of Contents

# 1. Introduction

The online social media (OSM) was defined as ''a group of Internet-based applications that are built on the ideological and technological foundations of web 2.0, and that allow the creation and exchange of user-generated content'' (Kaplan & Haenlein, 2010). The widespread of web 2.0 has accelerated and facilitated the penetration of social media including Facebook, Twitter, and Instagram. In addition, the creation of geo-referenced social media information has accelerated lots of studies on the relationship between the geo-information and other attributes of social media. The amount of geo-referenced social media will continue to increase with the rapid prevalence of smartphones with GPS (Zheng, 2014).

The social media is creating massive data every day all over the world. The data with geo-information own three significant attributes – location, time and the subset of the population that the publishers belong. There are three branches of exploring the social media: First, a subset of the population is extracted according to the contents they post on the social media; Second, an identified subset of the population' attributes are explored to profile them. Thirdly,

the events' (or keywords') location, time, and publishers are aggregated to understand the diffusion or current situation of the events. Besides, the trajectories of users have also been explored (Gao et al. 2015). However, after the release of the EU General Data Protection Regulation (GDPR) in 2018, it is impossible to trace any user by social media in EU. Similarly, China and the US also have enacted laws to protect the data of social media users (Liu et al. 2018). Therefore, the research of social media will focus on collective attributes instead of individuals' movement in the future. Although extensive studies about social media have been done in the last years, there are still considerable topics waiting to explore.

Online contents are not entirely representative of the precise social situation and real-time events, but merely provide a public display of personal and collective identity (Kaltenbrunner et al. 2012). A better picture of users' collective behaviors can be profiled by taking into account the collective contents, yielding insights into the features of a subset of the population. Among factors influencing social media contents, geographic distance plays a significant role in users' content, which was proved to be true in London by Ostermann (2015). In a similar vein, Andreas (2012) had analyzed the relationship between online user interactions and geographic proximity and observed that the geographic distance significantly affect social links and the spatial proximity plays a sight influence on users' interaction. While, how the distance affects the similarity of contents on social media and what the spatial characteristic of collective social media contents is are both remains unexplored.

The dataset analyzed in the report contains extensive records from the Weibo (the most popular OSM like Facebook and Twitter in China) including contents, coordinates where the users generate the materials, and other messages. It presents an opportunity to analyze the relationship between OSM contents and geographic location. This paper applies the algorithm proposed by Andrienko et al. (2013) to extract places and offers a model combining LDA (Latent Dirichlet Allocation) and Vector Space Model (VSM) to analyze the similarity of OSM contents between places. Besides, topics that represent the feature of places are to be mined by LDA. The distribution of topics' probability of places is to be analyzed by Moran's I, Geary's C, and DBSCAN.

## 2. Research Objective

The report takes the assumption that there is more similarity in social media's contents between closer places. Therefore, the paper is to explore how the Tobler's first law work in social media contents. The OSM contents in a place extracted by the algorithm might own specific features due to the spatial difference. How to extract places and their characteristics (topics), and how to apply the features reflected by the topic probability vectors to cluster social media users and contents are also to be discussed in the report.

### 2.1 Motivation

To cluster the subset of the population by OSM is universal nowadays, a few applications and systems have made use of user-generated information to offer a better recommendation, to advertise products to customers who might be of interest by algorithms of machine learning. The data of user-generated contents and the news feed technology are the fundamental of some leading internet firms like Google, Facebook, and Toutiao (a data mining-based recommendation engine products in China). These applications that take advantage of users-generated contents are all to extract features of an individual user or a group of users.

It is essential to understand the collective features of the population at different scale places for many researchers and organizations. The conventional methods to explore the collective features are the questionnaire, census, and sampling. The access to extensive OSM data with location and GIS technology provides an opportunity to analyze aggregate measures and features through OSM data. OSM contents are not always representative of the real world, and the data from OSM may be noisy depending on the purpose of collecting data. However, the data also represents a rich and multi-faceted insight into the semantics and views of a space comprising a subset of the population (Ostermann et al. 2015). Therefore, the planners, government and, other firms are to analyze the collective features and transformation of a place through OSM contents. However, the relationship between geographic distance and aggregate functions of OSM are not clear. Also, many studies (G. SUN et al. 2017; Liben-Nowell et al. 2005) collected users' location by geocoding their places from profiles, which is commonly recognized in operating exploration of geography-related OSM information. It is not always

precise and reliable to collect location through users' profiles. Besides, less than 5% of users on OSM provide location posting information. Features extraction is improving approximating users' location by matching users' features with places' characteristics and verifying if the results match the places users provide to make it more precise and reliable. The algorithms trained through OSM information even can be applied to other UGC (User Generated Content) websites (Quora, Zhihu and, Youtube) where users do not upload places. The algorithms and approximated location of users are facilitating the studies about the spatial and temporal characteristics of OSM information and the characteristics of places and sub-population. The work is also motivated by studies provided by Wang et al. (2013), they proposed that the spatial distance influenced the features of OSM users' spatial distribution although the factor functioned less and less and the network on OSM is consistent with physical, social relationship based on spatial neighborhood, social status, and economic development. Therefore, to understand the network and contents on OSM is meaningful to discover the situation of the urban development, population, and crisis.

The original motivation of this report is the assumption that an improved understanding of the spatial relationship of OSM contents' semantics would in turn help explore the collective features of users in various spatial scale. Besides, the results could be applied in exploring other texts or contents' spatial characteristics on the internet even though these texts do not contain location information initially.

## 2.2 Scope of study

The primary dataset contains extensive geo-referenced Tweets (Message sent by Sina microblog like tweets sent by Twitter) from Sina Microblog's API. The dataset is anonymized for privacy. The data posted in Beijing has been extracted from the primary dataset. The spatial scope of this research is in the 16 districts in Beijing. The reason why Beijing is chosen to be the research object is that Beijing owns over 20 million of the population which is the world's third most populous city and most populous capital city. It is the city that has the most significant number of people who use the internet and OSM; most importantly, Beijing has the

most significant number of users of Sina microblog that is the most popular OSM for the public. The spatial scale in this report is the place. How to extract places from cities is to be discussed in the remainder. All Tweets were generated from 1st February 2018 to 1st May 2018 in Beijing because the Sina's API only provides data posted in the last six months

On Sina microblog, the users can upload the latest status, feelings towards affairs and forward others' post. The contents sent by users are categorized into three kinds: broadcast, conversion and, retweet that works like Twitter. The paper focuses on the broadcast that is the universal one and open to the public. A piece of the broadcast is called a tweet in the following.

There are three questions to answer in the remainder of the document. First, the similarity and features of Sina microblog tweets' (text) semantic between space and scale are to be discussed. Second, the method of how to extract the features of a place in the city is also to be explored. Thirdly, how to mine topics ( features ) for places is also to be discussed in the reminder.

## 2.3 Report Outline

The remainder of the document is structured as follows:

·Chapter Three provides an overview of the background behind the motivation for this research and an outline of existing research published in this area

·Chapter Four describes the steps and the methodology taken to prepare the data for analysis and the algorithms chosen to conduct spatial and demographic tests with the data.

·Chapter Five details the analysis of the tweets' similarity between space and scale.

·Chapter Six provides the analysis of places features' distribution.

·Chapter Seven presents the findings from the analysis and puts the results within context.

·Chapter Eight provides concluding remarks and suggested next steps.

# 3. Literature Review

## 3.1 Studies using OSM data in the city

Extensive efforts have been made to explore the relationship between OSM information and locations. As G. SUN et al. (2017) summarized, three factors were modeling the OSM information diffusion crucially: geographic distance (Liben-Nowell et al. 2005), recency effect (Leskovec et al. 2009), and cultural proximity as well as linguistic similarity (Hoftede et al. 2010). Although the cyberspace was recognized as boundless, some social interaction behavior, such as befriending and communication, is considered as geographically determined (G. SUN et al. 2017; Liben-Nowell et al. 2005; Peng et al. 2015). The information and network technology have broken the boundary of geography. However, a few studies on OSM still observed that the distance still plays a significant role in interactions on social media. (Liben-Nowell et al. 2005; Peng et al. 2015). The shorter the distance between two users, the more frequent interactions they have (G. SUN et al. 2017). The recency effect also plays an essential role in information diffusion on OSM platforms as the temporal order is implied all information dissemination process (Liben-Nowell et al. 2005). The recency effect is that users have a cognitive bias that they are prone to recall the latest information. Therefore, the contents on OSM significantly reflect the timeliness of events in the physical world. The timeliness of OSM makes it possible to explore the information diffusion, population's semantic and related topics with state of the art technology on data process, machine learning, and data visualization. The work focuses on the role geographical distance play on OSM information. The temporal effect is not to be taken into account in and the data takes one month as the temporal unit to explore the spatial similarity and characteristics in this paper.

The cultural proximity and linguistic similarity facilitate the diffusion of information on OSM with other factors being equal (G. SUN et al. 2017; McPherson et al. 2001). The cultural proximity and linguistic similarity are both profoundly influenced by distance. Therefore, there is an assumption that the spatial characteristics of OSM contents reflect a place's other features and the subset of the population's features. G. SUN et al. (2017) modeled the OSM information

diffusion by the gravity model. They applied a distance decay function into the formula. Peng et al. (2015) also took the distance into models providing the difference and diffusion of OSM information.

As for the data source, Sina microblog is similar to Facebook and Twitter. Ostermann et al. (2015, P.28) stated tweets are very noisy since the kind of OSM information covers a wide range of topics， from emotional express to public events ones. Hahmann et al. (2014) claimed that geo-referenced Tweets were not always related to the place where users post the tweets and some contents of these tweets could not represent the semantics of the space or places, especially at fine-grained scales (Ostermann et al., 2015). However, the users could choose to post with location information or not, which makes them prefer to post with location information when they want to express some feelings or provide some information related to the places, and they are pleasant to expose the location to the public. The users prefer to post their location information when they think the contents are related to the location to some extent. Therefore, this work handles geo-referenced Sina microblog data to extract single places and explore their similarity and difference. Wang et al. (2013) proposed that the network of Sina microblog users still were influenced significantly by the spatial distance. Although the network information itself is created independent on the distance and physical space, human as a significant participant, live by way of social interaction in the traditional geographical entity space that still plays a role in the network. The network is the extension and representation of geographic space to some extent.

Geo-referenced OSM contents link places with attributes of different nature in such locations, thus discovering the situation of citizens aggregated in the same position who post materials or knowledge revealing the characteristics of places (Ostermann et al. 2015). As for studies about OSM contents in Chinese like Sina microblog, Zhang et al. (2011) try to build a model based on MB-LDA to mine topics from Sina's materials. Nature language process has been applied in a few studies in OSM information. Wang et al. (2013) did research on the network of geography based on Sina users' relationship that took Sina microblog as an example, and they propose that the users of Sina who used the OSM frequently mainly were distributed in some super metropolis like Beijing, Shanghai, and Guangzhou that all own over 10 million people. Overall, few of the studies combine spatial analysis with OSM contents in Chinese. How the

Tobler's first law work in Chinese OSM contents is not evident yet. Besides, the relationship between places and users or materials produced by the need to be explored.

### 3.2 Extract places in the city

How to extract places except by administrative boundary is a problem to cluster OSM users at fine-grained scales. Agnew et al. (1987, 2011) provided three descriptions to define a distinct place: distinct location, local, and the sense of place. The site is separated from another location spatially. The locale is distinguished by the attributes and properties of a site like the function and the infrastructure of the place (Teobaldi and Capineri 2014). The sense of place is defined by the citizens' impression and the semantic expression of the site. According to the definition that a distinct place should own the specific concept of a place that can be distinguished from other places especially the neighborhood (Winter and Freksa 2012; Ostermann et al.,2015). The idea of places can be extracted from UGC contents product in such place since the semantic expression of users on OSM represent the concept of places that the users have in mind (Purves et al 2011).

In the spatial OSM data, the users are usually specified as points in studies. However, places are defined as polygons. Thus, it is necessary to cluster points to fit in the sites. Conventionally, the point pattern analysis has been taken by density-based methods. For instance, the Poisson distribution, Ripley's K and quadrat analysis have been mainly applied to test the spatial randomness of points' distribution. The DBSCAN and k-means have also been applied to cluster users (Andrienko, 2013). The density-based cluster method can construct arbitrary shapes and sizes of clusters. Andrienko (2015) observed that the DBSAN was prone to build huge clusters even covering the whole city when the method was applied to extract distinct places in cities using geo-referenced OSM data. The huge clusters merge many distinct places into a polygon. However, the threshold of DBSCAN cluster is increased to decrease the

number of clusters, which would also miss a lot of places. Zhang (2011) similarly stated that the density-based clustering applied in public OSM data failed to extract distinct places. Andrienko (2015) proposed a spatially bounded point clustering algorithm to extract personal and public places in the city by points' location of users. The algorithm is insensitive to density and can fit the arbitrary shape of places but limit the size under certain scale. Ostermann et al. (2015) applied the algorithm in London exploring the Flickr and Twitter data and proved the goodness of the spatially bounded point clustering algorithm. The paper is also interested in how well the algorithm work in Beijing -- a super metropolis like London. Andrienko et al. (2015) stated that the semantic analysis on places -- points could boost the exploration on the mobility of users in cities and a merging features of a places given by the individuals visiting the place could be applied to explore the flow of population in places without violating users' privacy. Wang et al. (2013) illustrated the places' semantic analysis using Sina microblog text data at the scale of cities and found that the Tobler first law works, which also motivated this article.

### 3.3 Vector space model (VSM) and other conventional methods to deal with OSM text data.

The deep learning algorithm has achieved incredible results in the image and audio processing, but it has not seen such exciting results in the NLP (Nature Language Process) field. Collobert and Weston (2007) proposed that the reason is the language (words, sentences, chapters, etc.) belong to the high-level cognitive abstract entities generated in human cognition, while speech and images belong to the lower-level original input signals, so the latter two are more suitable for deep learning to learn features. To some extent, the technology of processing speech data is related to the NLP since the speech also includes numerous the high-level cognitive abstract entities. Raghavan and Wong (1986) claimed that the conventional deep learning of speech could only be categorized speeches without understanding the high-level cognitive meaning. The introduction of the vectorized representation of natural language has facilitated the development of NLP. One-hot Representation was proposed initially to represent words.

However, the one-hot representation product too high dimension of the vectors, which makes it hard to be applied in machine learning and other tasks handling large text documents (Mnih and Hinton, 2007).

The vector space model has been applied in many information retrieval (IR) projects and researches. Ostermann et al. (2015) explored places' similarity of social media text data by transforming the text into a vector in London. It is not clear how well the method would be applied in Chinese text. Besides, Wang et al. (2013) claimed that Chinese linguistic features ask for more pre-process building vector space model and applying other NLP technology instead of fetching algorithms that work well in processing English text. In common documents, there are large functions words like a the. Automatic document indexing frequently is noised by the kind of function words to interference the analysis. Language dependent method like adding a list of words to remove from documents and nonlinguistic methods applying the statistical probability of words to extract function words both have been verified boost the goodness of the vector space model (Chowdhury 2010). Sebastiani (2010) summarized that three aspects weight the term vector: based on term frequency, collection frequency document, and document length normalization. The term vectors of documents generated by the vector space model can be a reference to information retrieval, comparison, and other statistical analysis. The most popular measure is comparing the cosine coefficient while Jaccard and Dice's coefficients have also been applied in some analysis (Salton 1989; Cogsys 2018). Many studies in the places' semantic analysis based on OSM text data built the term vector directly without eliminating the effect of function words and weighing words, which could affect the results in processing Chinese text on OSM (Huang et al. 2013). Weighing words by the text' length and removing function words (stop words) are both necessary to be taken into account in processing the data. Also, it is not so meaningful to compare the spatial distance or cosine value when massive texts are transformed into vectors, and there are too many dimensions. While VSM works well when the number of dimensions is low, that means that high-dimensional vectors have to be transformed into low-dimensional vectors before comparing and operating the vectors. In this paper, the vector is to be built according to the topics vectors product by LDA.

### 3.4 Method to mine features of places.

To extract the features of places is to extract the topic (characteristics) of the location, which is initiated with extracting topics of texts. The topics extracting has been widely applied in information retrieval, AI and feeds. In the era of information explosion, it is especially important to mine the topic information from the extensive knowledge and analyze the intimate semantic association. Microblog is an unstructured message with some structured social network information. This kind of social network association plays a supporting role in features (topic) mining. For the other side, every tweet contains a fragment of information (usually one or two sentences). The small structure makes it hard to mine the topic and features. However, a place, an individual or a group of people have posted more than one tweets so that the massive amount of texts can, in turn, help mine features of a place or a group of users (Sun et al. 2011).

Mining the topics of a series of texts through the vector space model is to transform unstructured text into point data in the multidimensional space. The distance between points in the multidimensional space is the signal if the documents are similar, which work well in comparing points between neighborhood. However, the output of vector space model does not provide the topic of a place or text (Sun et al. 2011), and the distance in extensive text data is not meaningful, which makes it hard to explore the difference among places. Latent semantic analysis (LSA) was proposed by Deerwester (1990) to mine the topic of text based on singular value decomposition that explores the text's grammatical structure and discovers the potential relevance of documents. The limitation of LSA is that it cannot match a word with various meaning due to the uniqueness of words in multidimensional space (SUN et al. 2011). Besides, the singular value decomposition involved in matrix operations makes the results usually have negative numbers that are hard to be explained.

Also, a few methods of the topic model were proposed. PLSA and LDA are two representative methods applied in mining the topic of texts. Probabilistic latent semantic analysis (PLSA) was introduced by Hofmann (2017) based on LSA. The TF-IDF（term frequency－inverse document frequency）is usually used to represent texts, and the results are high-dimensional data. LSA and PLSA are both involved in dimensionality reduction to decrease the difficulty to

calculate (Sun et al.). LDA (latent Dirichlet allocation) is an extension of PLSA applying the Dirichlet distribution. Blei and Jordan (2003) who proposed LAD stated that PLSA was prone to encounter the issue -- overfitting and PLSA could not allocate the probability of the text out of training data. Therefore, LDA introduces hyper-parameters to build a three layer 'Document - Topic-Words' Bayes, probability model. LDA has been applied in topic mining (Wei and Bruce, 2006), citation analysis (Dietz and Scheffer 2007), and OSM contents analysis (Mei et al. 2008). Especially, Sun et al. (2011) had built an MB-LDA model based on LDA to extract topics of tweets on Sina microblog, and the model worked well. MB-LDA takes at (@) and non-original tweets (forward other users' tweets) into account building a Bayesian network.

The vector space model has been applied in many information retrieval (IR) projects and researches. Ostermann et al. (2015) explored places' similarity of social media text data by transforming the text into a vector in London. It is not clear how well the method would be applied in Chinese text. Besides, Wang et al. (2013) claimed that Chinese linguistic features ask for more pre-process building vector space model and applying other NLP technology instead of fetching algorithms that work well in processing English text. In common documents, there are large functions words like a the. Automatic document indexing frequently is noised by the kind of function words to interference the analysis. Language dependent method like adding a list of words to remove from documents and nonlinguistic methods applying the statistical probability of words to extract function words both have been verified boost the goodness of the vector space model (Chowdhury 2010). Sebastiani (2010) summarized that three aspects weight the term vector: based on term frequency, collection frequency document, and document length normalization. The term vectors of documents generated by the vector space model can be a reference to information retrieval, comparison, and other statistical analysis. The most popular measure is comparing the cosine coefficient while Jaccard and Dice's coefficients have also been applied in some analysis (Salton 1989; Cogsys 2018). Many studies in the places' semantic analysis based on OSM text data built the term vector directly without eliminating the effect of function words and weighing words, which could affect the results in processing Chinese text on OSM (Huang et al. 2013). Weighing words by the text' length and removing function words (stop words) are both necessary to be taken into account in processing the data. Also, it is not so meaningful to compare the spatial distance or cosine value when massive texts are transformed into vectors, and there are too many dimensions. While VSM works well when the number of dimensions is

low, that means that high-dimensional vectors have to be transformed into low-dimensional vectors before comparing and operating the vectors. In this paper, the vector is to be built according to the topics vectors product by LDA.

# 4. Methodology

As outlined in the introduction, the approach follows the five phases which are described more details in this section.

1. Data Source: Collect data and clean data (Including removing stop words and text segmentation).
2. Extract Places: Extract all potential places from collected tweets (points) in Beijing.
3. LDA-VSM model: Mine topics of places by LDA and build a topics' vector of topics for each place.
4. Analyze Similarity: Calculate vectors' cosine similarity between places and analyze similarity between space and scale.
5. Analyze Features: Cluster the points by DBSCAN based on topics simplified vectors, and analyze the distribution of topics as well as the difference of places between neighbor places based on Moran's I, Geary's C.

## 4.1 Data source

Millions of tweets have been collected through Sina's API and crawlers. Each tweet owns attributes as shown in Table4.1.

| Name | Example and details |
|---|---|
| User' ID | The publisher' ID, i.e. 1196166632 |
| Time | Time when tweets are published i.e. 2018-07-19 21:04 |
| Contents | The text contents of tweets (most in Chinese) i.e.'今天天气很好，后海公园很美'（'The weather is nice today, the Houhai Park is |

| | |
|---|---|
| | beautiful ) |
| Coordinate | The coordinate where tweets are published i.e. [39.94006,116.438492]<br>The coordinates are based on WGS84 |
| Place | The name of the place provided by users |
| Up | The number of thumb up |
| Retweet | The number how many times the tweets are forwarded |
| Comments | The number of comments following the tweet. |
| Device | The device that is used to publish the tweet. |

**Table 4.1: The attributes of tweets**

The paper focus on the coordinate and text contents. In addition, the places' name provided by users are stored to help analyze the topics of places in the remainder. Most of the contents are in Chinese that are to be transformed into English in this paper. After collecting data, all tweets that were published in Beijing from 1st February 2018 to 1st May 2018 have been extracted to be the dataset used in this paper. The operations of dataset in the remained is based on MySQL and python.

The raw data source collected from Sina microblog includes 3,525,600 piece of tweets. Usually, only 3% - 5% of information on OSM is attached with location data. In order to get tweets with location information, the code (attached in the Appendix 1) collects data from the places of interest (POI) on the Sina website hence there are 1,527, 808 piece of tweets after removing tweets that were not located in city of Beijing and without location information. The shape file of Beijing is collected from Open Street Map. In some researches, identifying the users who live in the city instead of traveling at the place at first has been accomplished in case the visitors' tweets affect the result. However, this paper focus on the contents posted at places hence it is not necessary to identify local users before analysis.

After building the dataset, text segmentation package -- "JIEBA" is applied to segment the words of text on Python. Stop words (i.e. I, you, as) and function words (i.e. have, want, get) in segmented texts is to be removed based on the stop words lists that were proposed by Wei et al (2006) and Wang et al (2013). In addition, the words without specified meaning used on Sina microblog (SUN et al 2011) has also been removed from the segmented texts before the LDA process. The stop words list that includes stop words and words without specified meaning is provided in the Appendix 2.

## 4.2 Extract places.

### 4.2.1 Algorithms to extract places

Places are extracted based on all tweets coordinates. There are three phases -- a, b and c based on the algorithm proposed by Andrienko (2013). The input of the algorithm is a set of all points (tweets) $T = \{(X, Y)\}$

**Parameters:** Maximal radius of the place$R_{max}$, Minimal radius of places--$R_{min}$, and minimum number of tweets (points) in a place--$N$ . Minimal number of connected points--$C_{min}$, Maximal radius of larger places--$R^{l}_{max}$.

**Phase a:** Extract raw places.

Step 1: Randomly choose one points from $T$ and choose its closest points. A circle whose centre is the average of coordinates $\{(X, Y)\}$ of the two points and radius is $R_{max}$is built.

Step 2: Points in the circle are added into the circle. After one point is added, the center is re-calculated. When there is no points in the circle anymore, the circle is completed and repeat step 1 and 2 until all points are processed.

After phase a, a list $P$ of resulting groups points has been created. In fact, the places could be bigger or smaller than the circle whose radius is $R_{max}$. Therefore phase b and c are designed to extract smaller and bigger places respectively.

**Phase b**: Subdivide places

Go through $P$ . For each group of points $G \in P$, apply the following steps.

Step 1: Set R the radius of the set of points $G$, if R < $R_{min}$, go to next group.

Step 2: Otherwise, grids like the left of Figure 1 divides the circle into 9 parts. If the central part own fewer points than any other parts (the central density is not the highest one), do phase again inside $G$ with R/2 instead of $R_{max}$ as shown in Figure 1. If not, go to the next group.

Step 3: Add new resulting groups from step 2 into $P$ and remove $G$ from $P$.

Step 4: Recalculate the center for each new groups and if the points in neighbor groups are closer to the now group than that to the current group, allocate the points into the new group.

Step 5: Remove the groups from $P$where the number of points is less than $N$ .

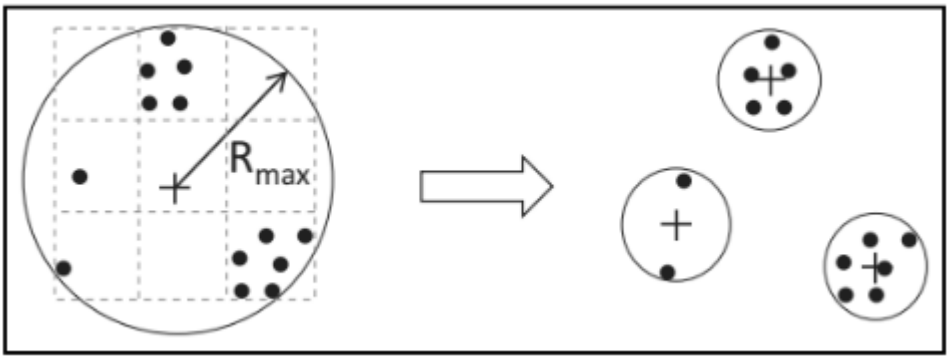



**Figure 4.1: A group of points are subdivided into smaller places.**

**Phase c**: Merging places

Sometimes, the place is larger than the circle whose radius is $R_{max}$. Phase c is to merging neighbor places that belong to a place together. Between two neighbour groups in $P$ , there would be some points belong to both circles whose radius is $R_{max}$. These points are called as connected points. The following steps can identify larger places as shown in Figure 5.

Step 1: Calculate all pairs of neighboring places in $P$ and their connected points. Make a list of all pair of places that own at least $C_{min}$ connected points. Sort the list in the descending order of the number of connected points.

Step 2: Go through the descending list. For each pair:

If the radius of the circle covering the two places exceeds $R^{l}_{max}$, skip the pair.

Else, merge the two place to a new place. Replace the two old places by the new merged place in the sorted list. Go to the next pair.

Step 3: Return the results into $P$

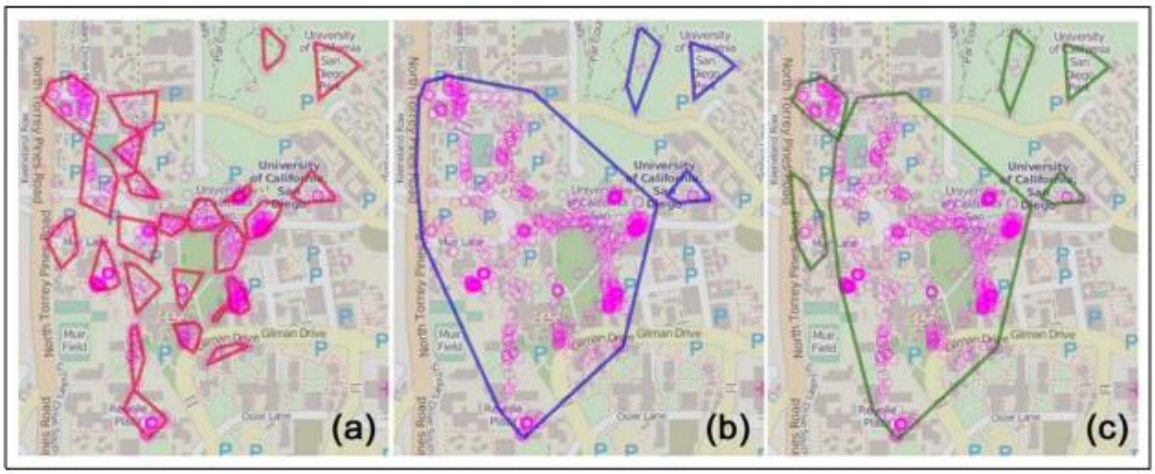



**Figure 4.2: Merging larger places**

The result after the steps -- $P$ is the list of places used in the following calculation and analysis. In addition, the centro-coordinates of all places are also calculated and added into the dataset. Due to that MySQL does not support geo-information, the results of extracting places are stored in MySQL in the format of centro-coordinates of places and each tweet (points) own one more attribute -- places' name or "no" that represents that the point does not belong to any place.

### 4.2.2 Set parameters of the algorithm.

According to the studies where the algorithm has been applied to extract personal and public places based OSM data (Andrienko et al ,2013; Ostermann et al. 2015;Andrienko 2015; Zhang et al. 2011), the parameters involved in the algorithm are set as Table 2.

**More details and methods about testing and adjusting are to be explained.**

| Parameters | Value (**To be tested and adjusted according to the practical**) |
|---|---|
| $R_{max}$ | 300m |
| $R_{min}$ | 100 m |
| $N$ | 270 |
| $C_{min}$ | 70 |
| $R_{max}^{l}$ | 1300 m |

**Table 4.2: Parameters to extract places**

## 4.3 Build topics vector for places by LDA

In this paper, LDA works based on the online LDA model built by the Sklearn on Python. In this paper, the TF-IDF has not been used before extracting topics for each tweet as the LDA is based on the frequency of words in text hence it is not necessary to do LDA with TF-IDF. The parameters and the hype-parameters of LDA used in this paper are shown in Table 3 based on a few studies and testing (Steyvers and Griffiths 2007; Sun et al. 2011; Blei 2003):

| Parameters | Detail and value |
|---|---|
| $\alpha$ | 1 |
| $\alpha_c$ | 1 |
| $\beta$ | 0.01 |
| $\lambda$ | 1 |
| $n$ | 50 |

**Table 4.3: The parameters and input of**

To build the topics vector, the approach follows 4 steps:

Step 1: Apply LDA to calculate topics probability distribution for every tweet. Each tweet has:

$$Pro1 = \{topic\ 1: X_1, topic\ 2: X_2 \ldots\ldots Topic\ n: X_n\}$$

*n :the number of topics as a input parameters.*

Step 2: Sort the $Pro1$ by the order of decreasing probability. Take the first three topics and their topics. i.e. for certain tweet :

$$Pro2 = \{topic\ 9: X_9, topic\ 6: X_6 Topic\ 18: X_{18}\}$$

Step 3: For each places, get the average of all $Pro2$ of tweets belong to the place based on the same topic. Each place get:

$$Pro3 = \{topic\ 1: p_1, topic\ 2: p_2 \ldots\ldots Topic\ n: p_n\}$$

Step 4: Transform the $Pro3$ for each place into the vector. Each place get a vector

$$V_j = [p_1, p_2 \ldots\ldots\ldots p_n]$$

## 4.4 Analyze the similarity

In the following, the cosine value between two places' vector $V_j$ represents the similarity of OSM contents between two places. Therefore, all cosine between all pair of places' vector $V$ has been calculated and stored for the remainder analysis. The reason why the cosine similarity is selected to be the measurement is its feasibility computationally and that it is a clear and well-understood tool to compare term vectors based on text. The index is the cosine value of the angle between two vectors. The value of cosine range from -1 to 1. But the value in vectors that usually represent the number of words (topics in the paper) is always non-negative hence the resulting cosine range from 0 to 1. If the vectors are the same completely, the cosine is 1 and the angel is $O^O$.As for two orthogonal vectors, the cosine is 0. The index functions as an approximation of semantic similarity between places and has been applied successfully in Geographic Information Retrieval (Vockner et al. 2013; Ostermann et al. 2015). After

calculating all pairs of places' cosine, the distance of all pairs of places has also been calculated after building places and their vectors.

After the calculation of all pairs' cosine and distance, the Shapiro-Wilk with the null hypothesis that the sample data comes from a normal distribution is used to test the normality of the distribution of cosine and distance. Furthermore, Kendall's Tau correlation test is to be applied in all pairs of places hence the general relationship between distance and correlation is to be explored. Kendall's tau is a measure of the correspondence between two samples. The tau statistic closer to 1 indicate stronger positive correlation, closer to -1 shows stronger negative correlation.

The relationship between distance and cosine value (similarity) and the analysis of topics vectors correlations between space and scale are to be illustrated in Chapter 5.

## 4.5 Analyze the distribution of places features (topics).

The analysis of features' spatial autocorrelation is based on Moran's I and Geary's C. And DBSCAN is used to cluster places topics.

Step 1: Features autocorrelation.

After building the vector $V_j$ for each place, there is a vector for the topic $i$:

$$T^i = [P_1^i, P_2^i \ldots . P_m^i] \quad m : \textit{the number of places}$$

For each topic, Moran's I is calculated to check if the topic cluster and Geary's C is calculated to check if there is low value cluster or high value cluster, namely, the spatial autocorrelation.

There is a point need to explain before applying Moran's I and Geary's C. First, Moran's I and Geary's C both are algorithms to measure the spatial autocorrelation usually. The first step is to compute the neighborhood matrix before calculating the weight matrix. Different from handling area data where the polygon has fixed neighbors, the point data's neighborhood matrix can be calculated by a few means i.e. within a distance criterion, k-nearest neighbors, or triangulation. In this paper, the neighborhood matrix of point (places) is to be computed based

on the distance criterion and the k-nearest neighbors (KNN) and the results with different means and parameters are to be discussed. The thresholds are to be given in the remainder.

Step 2: Topics cluster

According to the results of Moran's I and Geary's C, the topics that cluster significantly are to be explored by DSCAN to discover their distribution. How to set parameters in DBSCAN, and the analysis of clustering results are to be discussed in the remainder.

Step 3: Spatial characteristics cluster

For each place, place $j$ has a vector $V_j$

$$V_j = [p_1, p_2 \ldots\ldots p_n] \quad n: the\ number\ of\ topics$$

$V_j$represents the place j's coordinate in 50 - dimensional space. Thus DBSCAN is being used to cluster places, how to set parameters in DBSCAN, and the analysis of clustering results are to be discussed in the remainder.

## 5. Analysis of tweets' similarity

### 5.1 Analysis of places

In terms of the algorithm extracting places from tweets (the point data), the maximal radius of the place$R_{max}$, minimal radius of places--$R_{min}$, and maximal radius of larger places--$R_{max}^{l}$ are fixed comparatively in a specific city as the three parameters represent the scale of places. Even though the collective behaviours and cities' features in Beijing and London are different in many fields, the scale of places are usually similar, and the places' radius mostly arranges from 50m to 200m (Andrienko. 2013; Wang et al. 2013). In fact, the scale of places in metropolises varies slightly. In contrast, the minimum number of tweets (points) in a place--$N$ and the Minimal number of connected points--$C_{min}$ are mainly influenced by the data and the spatial

difference that need to be explored before extracting places. The -$N$ is usually set to be about three times than $C_{min}$ hence $N$ is the only parameter need to be tested applying the algorithms every time. It is almost impossible to get the complete users data from OSM except for the company who runs the platform and has access to all data. Therefore, it is hard to extract all places even regardless of the private places. The $N$ functions as a threshold that limits the least number of tweets in places that own enough contents to be analysed. In addition, the high threshold can select the kind of public places where more people visit and there are more related contents on the internet, which is necessary for easy the calculation that grows exponentially.

The left in Figure5.1 illustrates the histogram of the number of points in each places extracted with the parameters in Table 2, but the $N$ is set 40. There are 31052 places entirely and the number of most places range from 40 to 200. In practice, the number of places 31052 is too significant that would make the calculation slow. In addition, the places own fewer tweets are usually private places, and the shortage of tweets could result in bias. Thus, the $N$ is set 275 and the distribution of places' tweets is provided in the right of Figure1. There are 2850 places totally that are to be used in the remainder. Besides, the distribution of $N$ -275 is evener. Figure 5.2 illustrates the distribution of places and the number of attaching tweets. Then centrality is obvious and the places in the central area own more visitors than places the outside. The reason why the places with low density outside the central area are remained to analyse is that the distance between the central areas is generally short.

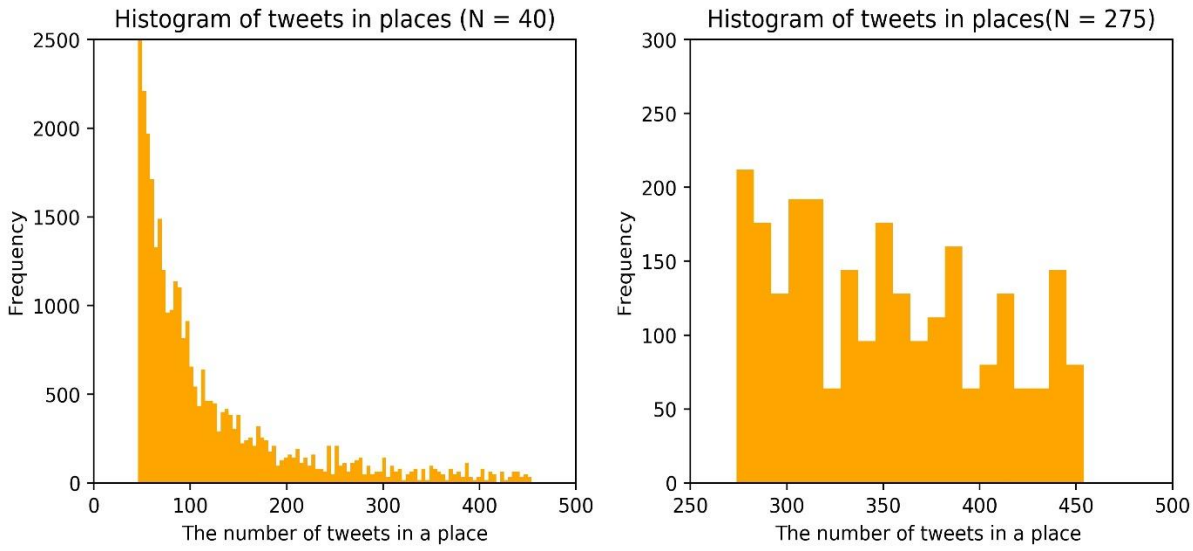



**Figure5.1: The histogram of places tweets**

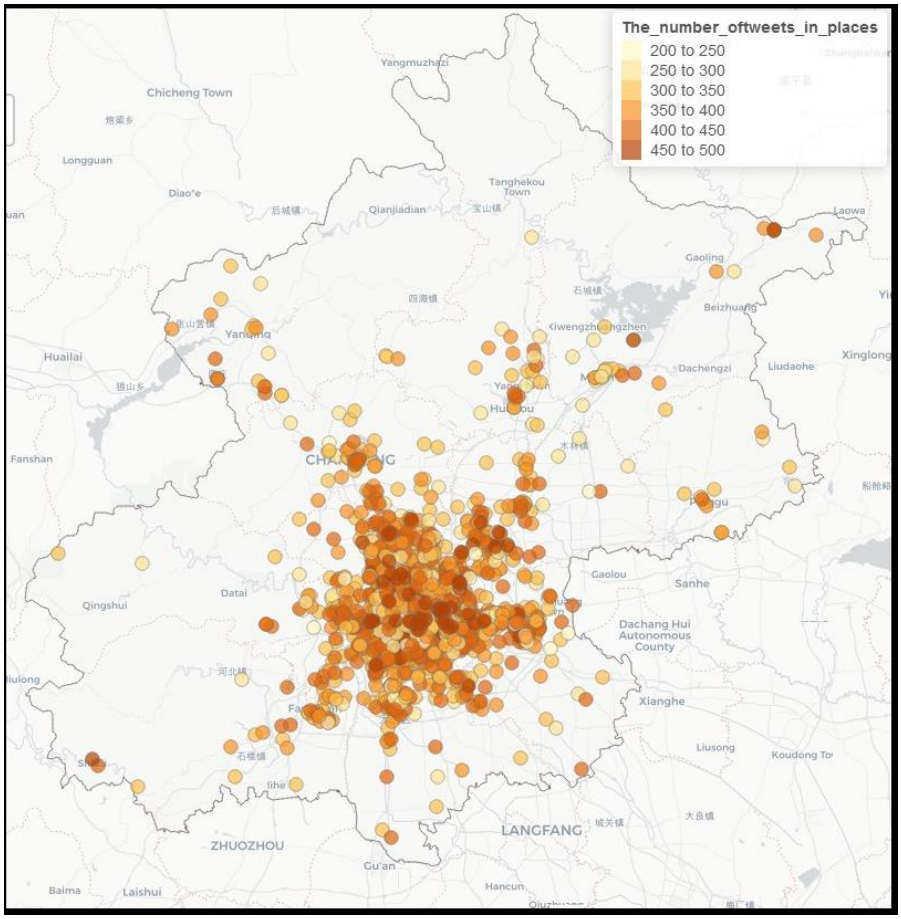
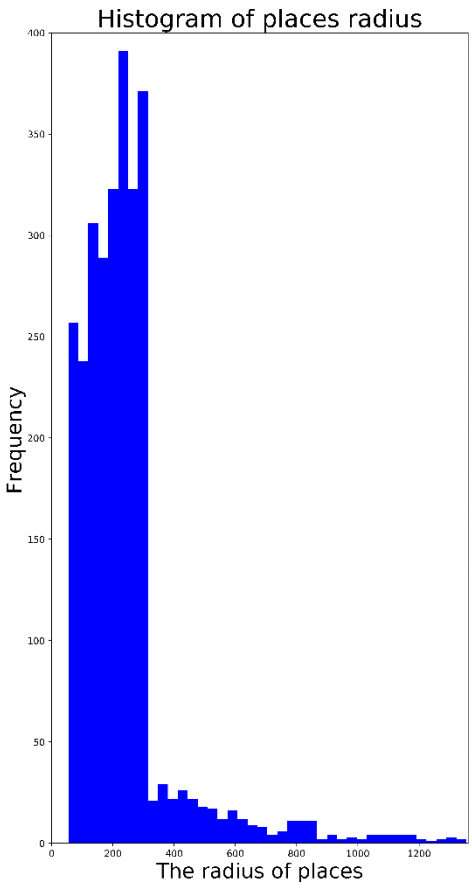



**Figure 5.2: The number of tweets in places**

Figure 5.3 provides the distribution of places' radius, 67.7 % of all places own radius from 50 to 300 m that are ordinary places and occupy the vast majority. 15.6% are small places hold radius below 100, and the least is 50 m. As for more prominent places, huge sites account for 14.2 that is similar with small places while there are few super places own radii more than 700. However, the bar chart shows that places whose radius are bigger than 1000m do exist, which means that the algorithm has extracted places at various scale. It is need to be noticed that places with a small radius less than 50 m have not been found yet, which is to fixed by the data quantity and improving algorithms. The number of places owns radius more than 300m decrease dramatically, it means that the merged sites are not as widespread as usual places. Figure 5.4 shows that there is no apparent bias as for the radius between the rural and central area.



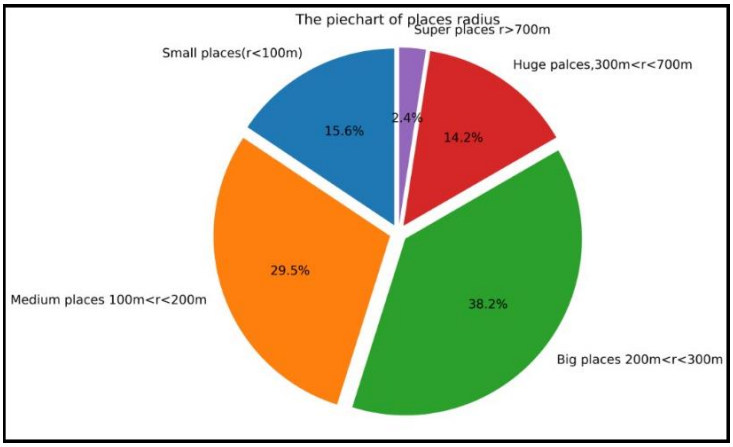




**Figure 5.3: The places' radius**

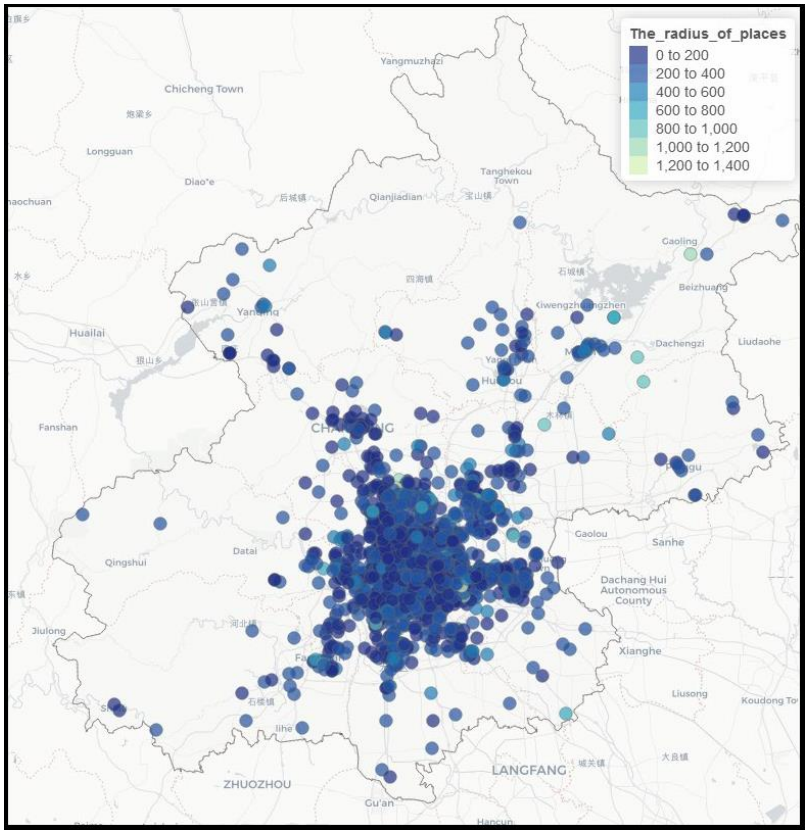



**Figure 5.4: The distribution of radius of places**

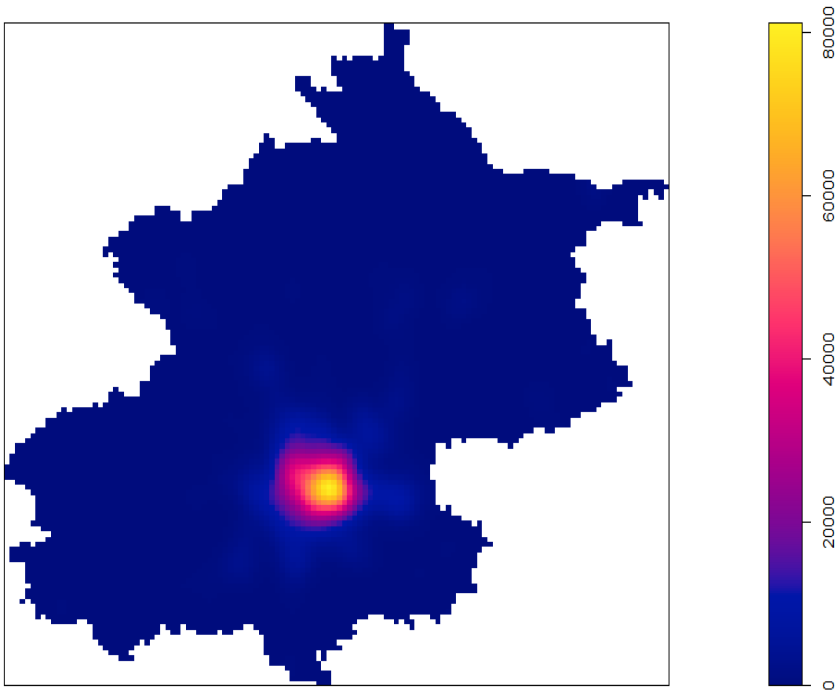



**Figure 5.5: KDE density with (Sigma = 500 m)**

Kernel Density Estimation (KDE) is used to explore the density of tweets in Beijing, and the result is shown in Figure 3 that illustrates the centrality is obvious. Even inside the central area, the density increase dramatically as it is closer to the centre.

## 5.2 Analysis of semantical similarity

After calculating all pairs of points' distance and cosine, the distance and cosine between every point and its nearest point have also been extracted, and they are asymmetrical. Thus, the relationship between the distance and cosine is to be checked. As the assumption is that the Tobler's first law of geography works in the relationship, the stronger negative correlation between the distance and cosine exists, there is more similarity between OSM contents in places semantically. Figure 5.6 illustrates the relationship.

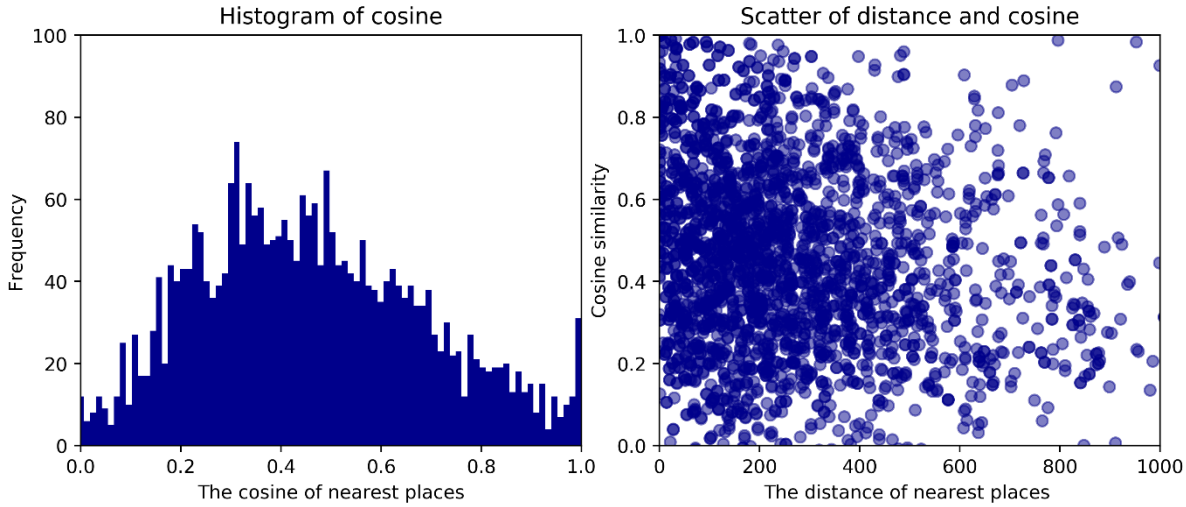



**Figure 5.6: The relationship between nearest places' distance and cosine similarity**

According to Figure 5.6, the place pairs owning high similarity from 0.6 to 1 are most distributed within 500 m. However, many place pairs hold low similarity below 0.5, which indicates there is a distinct diversity of neighbor places' contents in Beijing and the variety is also to be explored in Chapter 6.

Thus, before further analysis about spatial similarity, the correlation of the distance and cosine similarity needs to be checked. The Shapiro-Wilk test checks the normality, and the results are shown in Table 5.1. It is evident that they are not normally distributed.

| Shapiro-Wilk test | Distance | Cosine Similarity |
|---|---|---|
| W | 0.330437 | 0.926325 |

| | | |
|---|---|---|
| P-value | 0.0000 | 5.1057e-35 |

**Table 5.1: Test the normality by Shapiro-Wilk**

The non-parametric Kendalls's Tau correlation tests are used to check the correlation and the results are shown in Table 5.2. The negative Tau value indicates the correlation is in accordance with Tobler's first law of geography as closer places own higher cosine similarity than distant places.

| Correlation | Kendalls's Tau | P-value | z-value |
|---|---|---|---|
| Results | -0.0326 | 0.00732 | -5.37614 |

**Table 5.2: Test the correlation for nearest pairs' distance and cosine**

Figure 5.7 illustrates the distribution of cosine with the nearest place. It is evident that places which are more similar to their nearest neighbor places cluster in the central area. However, the distribution is not even, and the sites in the west of Beijing are more similar to the nearest places than that in the east.

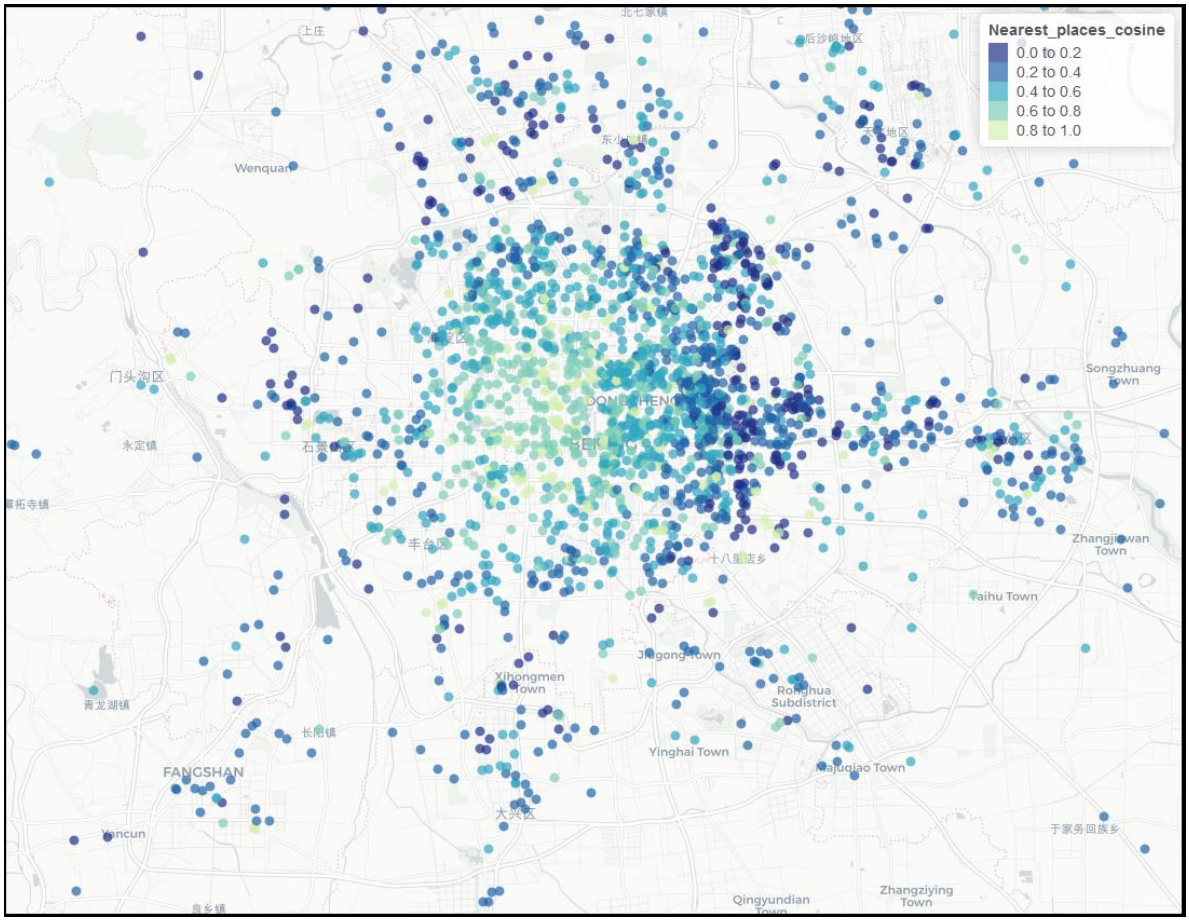


**Figure 5.7: The nearest places' cosine**

After the computation of Kendall's Tau tests that check the correlation of nearest pairs' distance and cosine, for each place, the correlation between distance and cosine with all other sites is calculated, and the result is shown in Figure 5.8. The results are distributed normally (Shapiro-Wilk test: W=0.99518, p-value < 0.0000). The normality indicates the spatial similarity of OSM contents is universal in Beijing. The distribution of Tau for each place with all other locations is shown in Figure 5.9. The points distributed mainly by three clusters and it shows the relationship between distance and cosine similarity varies spatially. The groups in Figure 5.9 are consistent with the skewness in Figure 5.7. To some extent, the skewness and clusters show that the places in the west have stronger semantically similarity as for OSM contents generally.

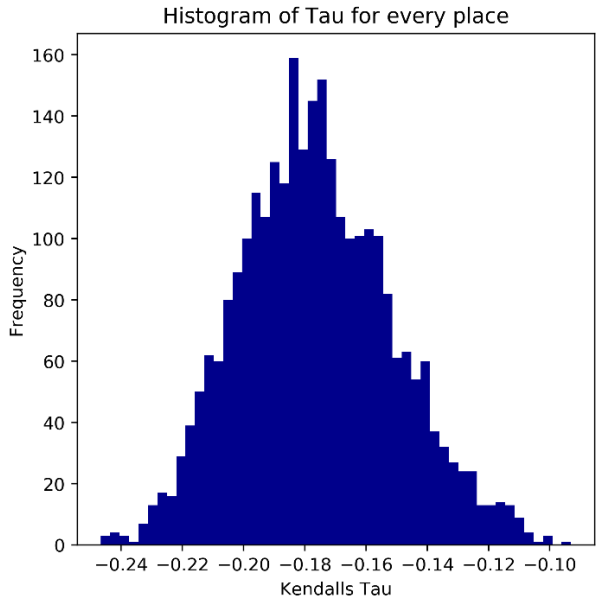



**Figure 5.8: The results of the Kendall's Tau of every places**

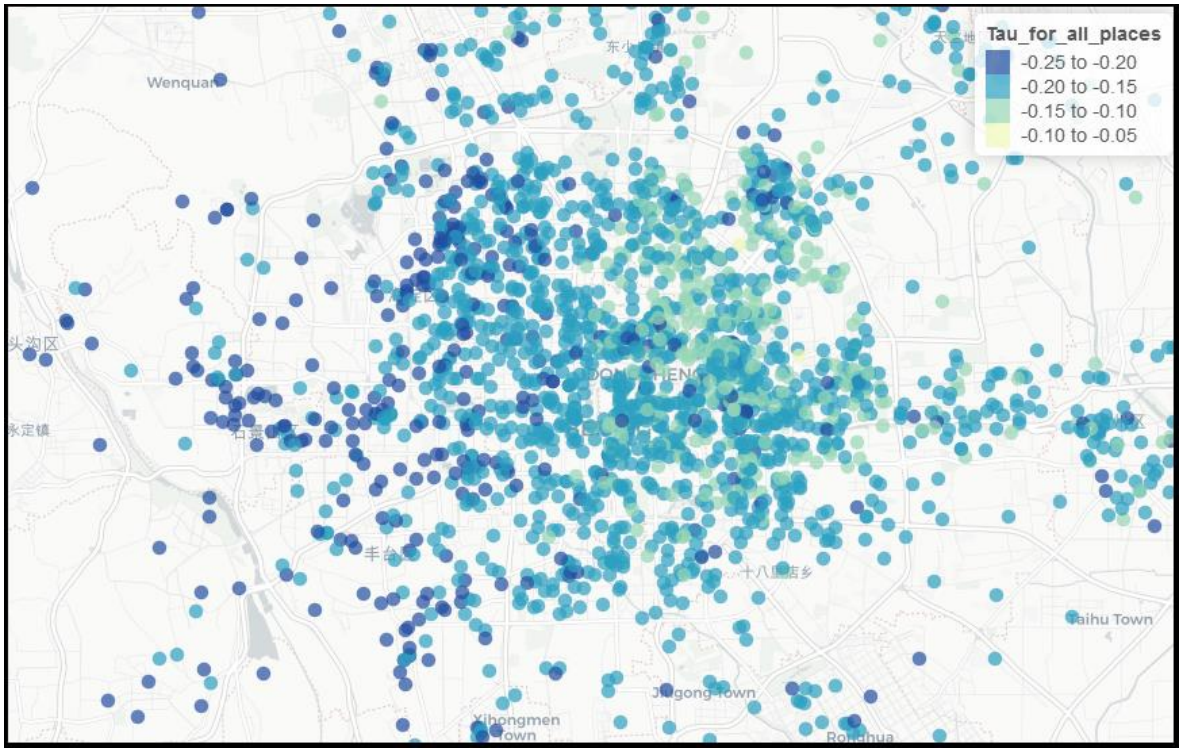



**Figure 9: Kendall's Tau for all places**

Figure 5.10,5.11,5.12 provide the distribution of average cosine within circles whose radius are 50m, 500m, and 1000m. The reason why some places disappear in these figures is that there is no neighborhood within the specified circle for these places. In the central area of Beijing – Xicheng District, the average cosine similarity is higher than that outside. The gradual centrality in Figure 5.11 and 12 prove that the places' OSM contents are more similar to near places than distant places. However, it also shows that the correlation in a range of scale as the values in Figure 5.11 and 5.12 are similar. Besides, the correlation is influenced by the density of places strongly, hence the completeness is a significant limitation of exploring the contents of OSM and its spatial characteristics. Besides, the central cluster of average cosine similarity and the negative correlation between distance and cosine similarity both make it possible to extract places OSM contents' characteristics, identify locales that relate certain topic, i.e., activities, education. Even more, identifying the location based on the users-generated contents is feasible to be accomplished. According to the results in this paper, it is necessary to take account into the spatial features exploring the OSM contents as the correlation between the distance and semantically similarity vary even inside a city. The contents on the outside of the central area where the density of places is low are distinctive comparatively. However, in the central area where the density of places is high, and the distance between places is short, the high average of cosine similarity might be due to the algorithm extracting places divides a place into several small places in the central and merges several places outside. The spatial features that make the Tobler's first law of geography does not fit the spatial similarity might be triggered by the urban plan, natural environment, and cultural-social factors. The features of the places would be discussed in the next chapter.

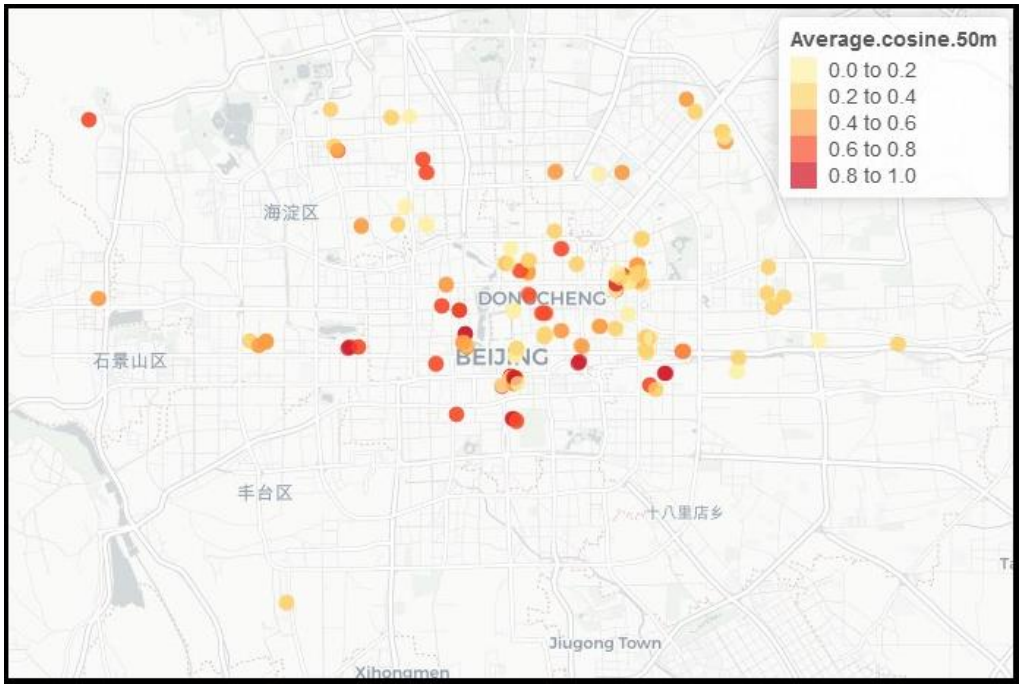



**Figure 5.10: Average cosine within 50m**

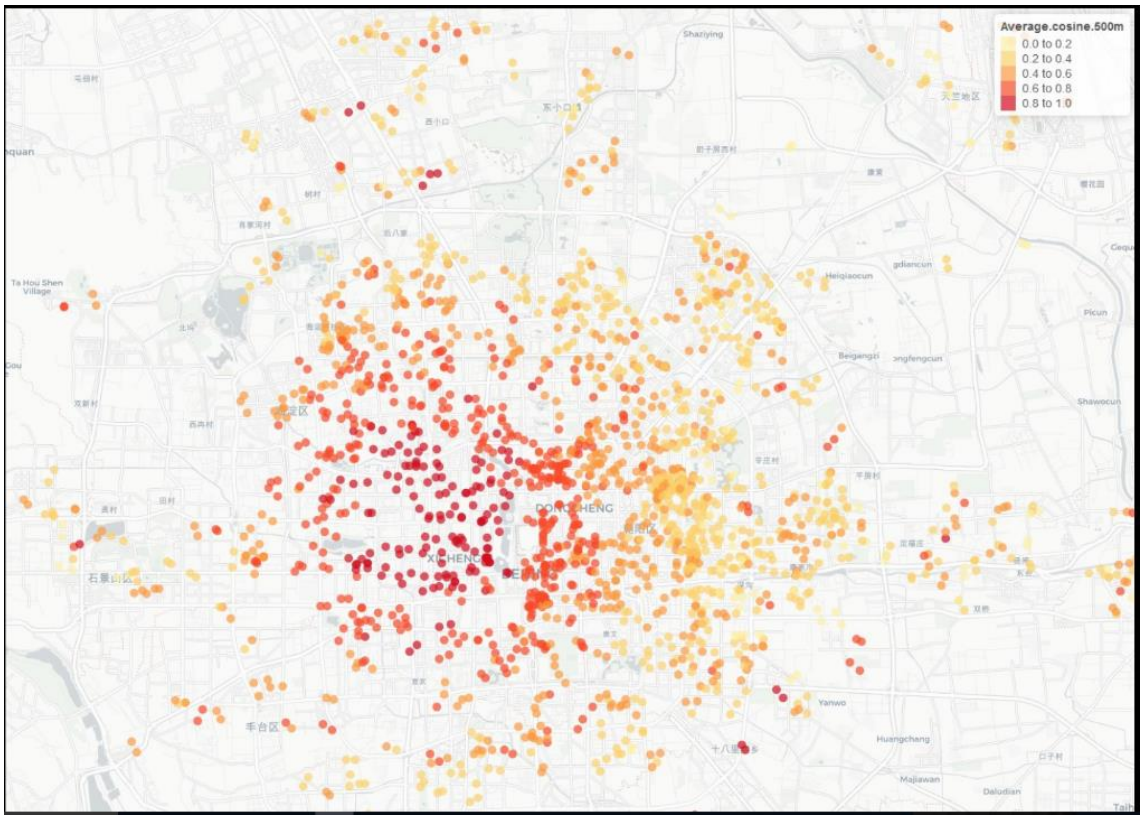



**Figure 5.11: Average cosine within 500m**

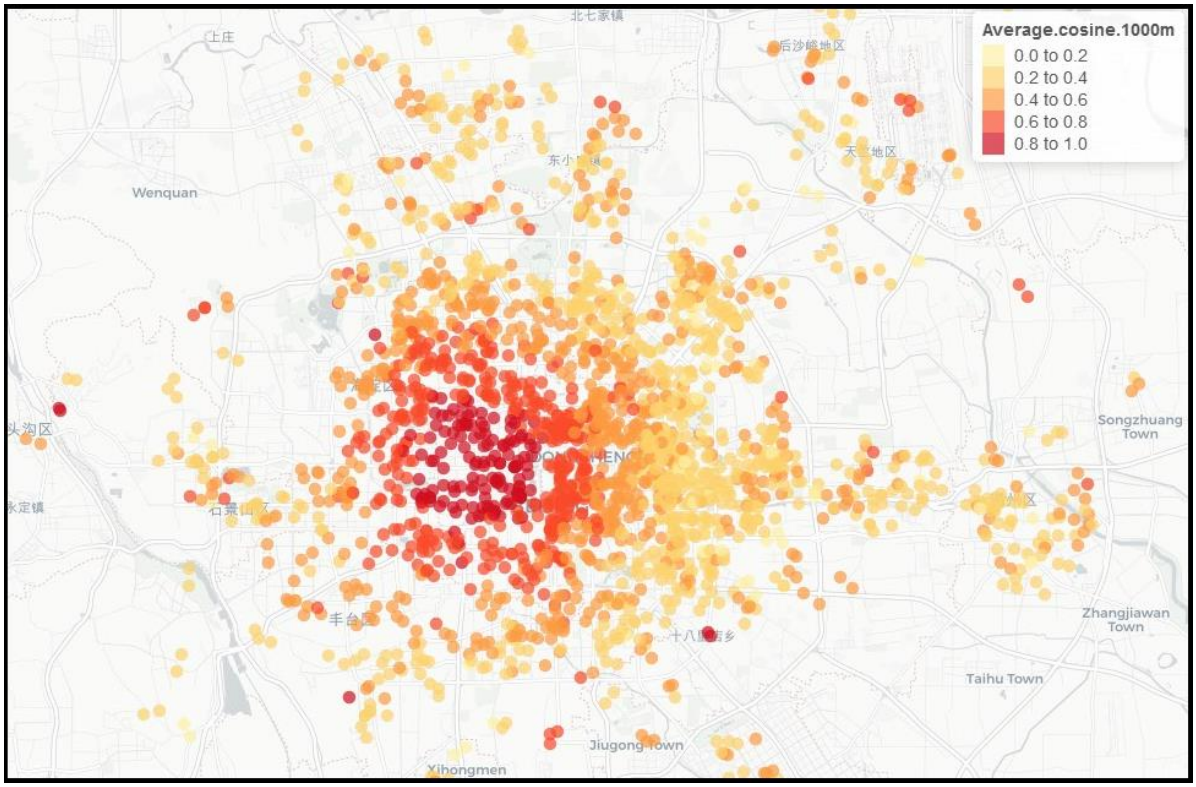



**Figure 5.12: Average cosine within 1000m**

## 6. Analysis of places' spatial features.

To explore the spatial distribution of places' features (topics), first, Moran's I, and Geary's C are used to check the autocorrelation for each topic. The neighborhood is based the on distance criterion and KNN. The distance criterion represents the radius of a circle where all points inside are considered as neighbor points. It also functions as the scale of clusters expected to identify. Thus, the distance criterion is set 50m, 500m, and 1000m as the three distance criterion divide the points in Peking differently according to Figure 6.10, 6.11 and 6.12. KNN weights make sure that all locations have the same number of neighbors more than 1. The k-nearest neighbor criterion can define the neighbors for given points. In this paper, the k is set as

1, 3 and 5 respectively and the difference among the three criterions can be analyzed to discover the spatial features of places.

The results of Moran's I and Geary's C are shown in Figure 6.1 and 2. According to the Figure 6.1, when the distance criterion of neighbors is set as 100m, the distribution of places' topics show high autocorrelation, which means that places prone to own similar topics with other places within 100m far from them. It is obvious that the autocorrelation with distance criterion 100m is far stronger than that with 500m and 100m. To some extent, it proves that the places are more similar with near places than distant places and the diversity of topics does not work at the scale of 100m. It is worth being noticed that half places are considered as isolated with distance criterion 100m, which makes the results might be influenced by the density of places and the limitation of the data source. But the similar trends between different criterions prove that the autocorrelation and cluster do exist. It also can prove that the Tobler's first law function in the similarity of places' topics. A topic can be considered as a parameter or part or an index of a place. Therefore, the places' features are influenced by their neighborhood than distant places. In terms the Geary's C results, Figure 6.2 illustrates similar trends as Figure 6.1, those topics with high autocorrelation most own low C value that represents positive correlation. Besides, the trends in the three criterions are also similar in Figure 6.2. More than half of topics represent positive correlation, which provides there is a strong similarity of places' OSM contents in the scale of places.

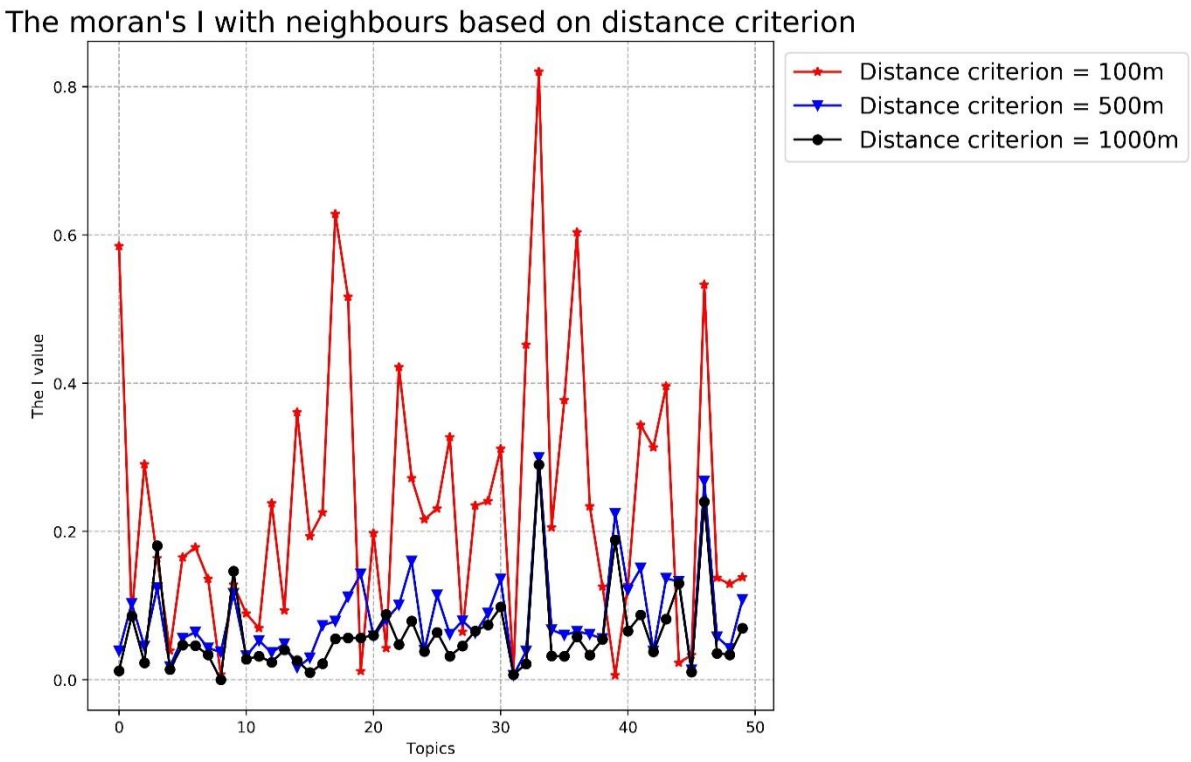



**Figure 6.1: The Moran's I results with distance criterion**

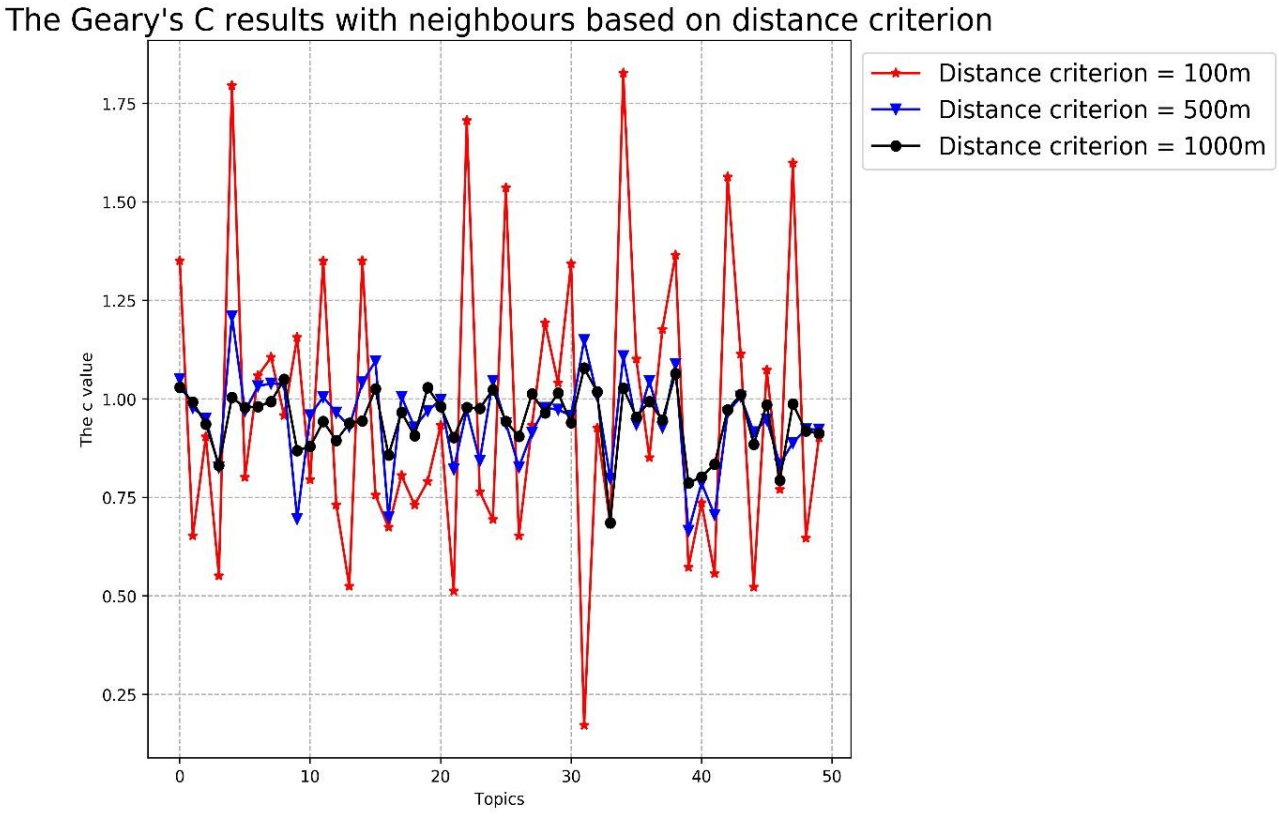


**Figure 6.2: The Geary's C results with neighbours based on distance criterion**

The results of Moran's I and Geary's C with KNN are shown in Figure 6.3 and 4. The trends are both similar with that in Figure 6.1 and 6.2 for those topics that own positive correlation. As the K-n increase, the autocorrelation reflected by the I and C value decrease slightly that is different with Figure 6.1 and 6.2. The KNN makes every point have at least one neighborhood. Most of the places are distributed in the central area of Beijing, and the density of places in the central region is high hence the places' 5 or 3 – nearest places are usually less than 100m away from it even closer. The OSM contents 5 nearest places reflect strong similarity. However, the similarity and autocorrelation are not as significant as that within 100m. The difference means that, inside a circle within 100m, the diversity of places features reflected by topics probability is obvious among the nearest neighbors. The autocorrelation reflected in the Moran's I and Geary's based on KNN are far lower than the value based on the distance criterions.

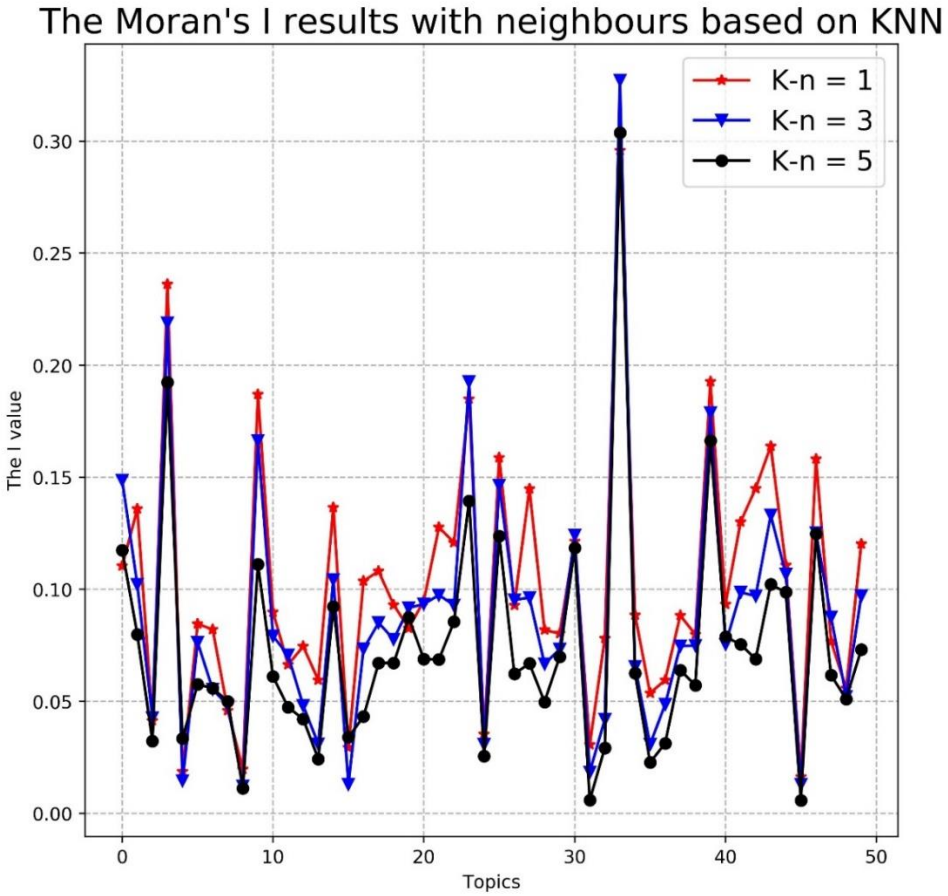



**Figure 6.3: The Moran's I results with neighbours based on KNN**

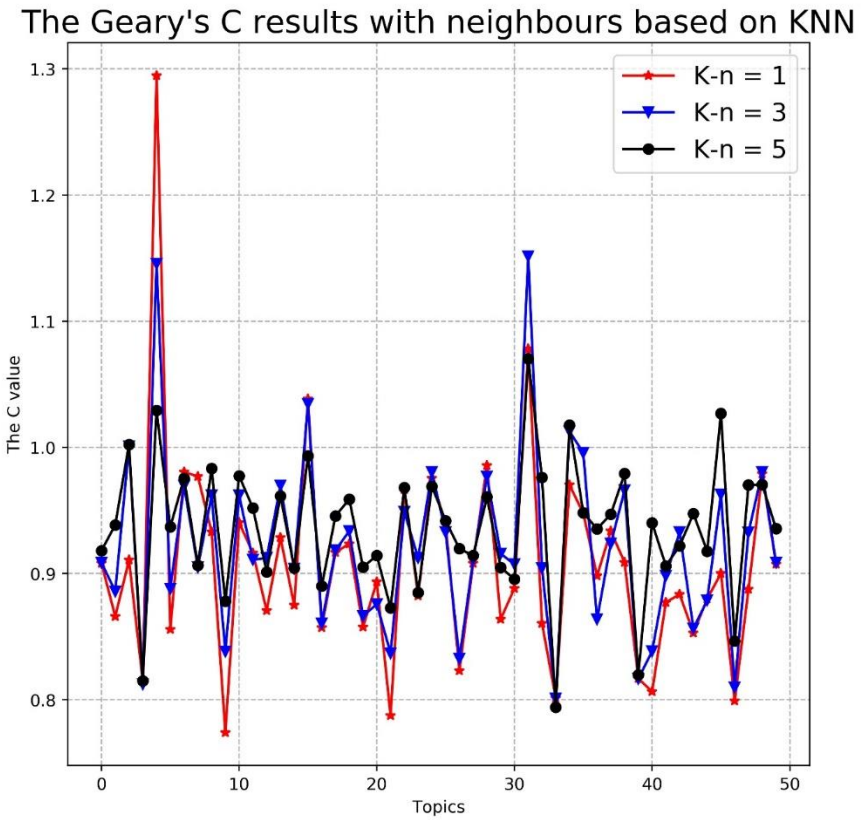



**Figure 6.4: The Geary's C results with neighbours based on KNN**

Taking account in all the results above, the four topics' probability that shows the most significant spatial autocorrelation are extracted to explore the distribution of places topics. The key words are given in Figure 6.5, and these topics' local Moran's I is given in Figure 6.6, 7, 8 and 9. According to these graphs, topics #17 is related to travel, and Figure 6.6 shows that the topic #17 in the central area shows strong spatial autocorrelation. Similarly, the topic #33 is related to the graduation of students, and the topic gives strong spatial autocorrelation in the central area of Beijing as there are more schools in the central area than outside. In contrast, the topic #21 that is related to the superstars and entertainment do not show strong spatial autocorrelation and the centrality as shown in topic #17 and #33. Besides, the topic #31 related to food and restaurants shows strong spatial autocorrelation not only in the central area but also in the corner of the urban area. As discussed above, topics can be considered as places' features or property. This distribution and spatial autocorrelation can reveal some elements of the city. For instance, as for the topic #17 related to the travel and the topic #33 graduation, places highly associated with these topics are usually distributed in the central area as the essential attractions and schools are distributed in the central region of Beijing. Besides, the topic #21 about superstars and entertainment usually does not show significant centrality, and namely, the people who concern the superstars and entertainment are distributed evenly. According to

the spatial autocorrelation of the topic #31 related to the food and restaurants, it can be concluded that the restaurants and place where we can have food are also distributed evenly generally in Beijing, however, in each finer granularity like a small district, the distribution shows strong spatial autocorrelation, namely restaurants are scattered in particular part of the small region.

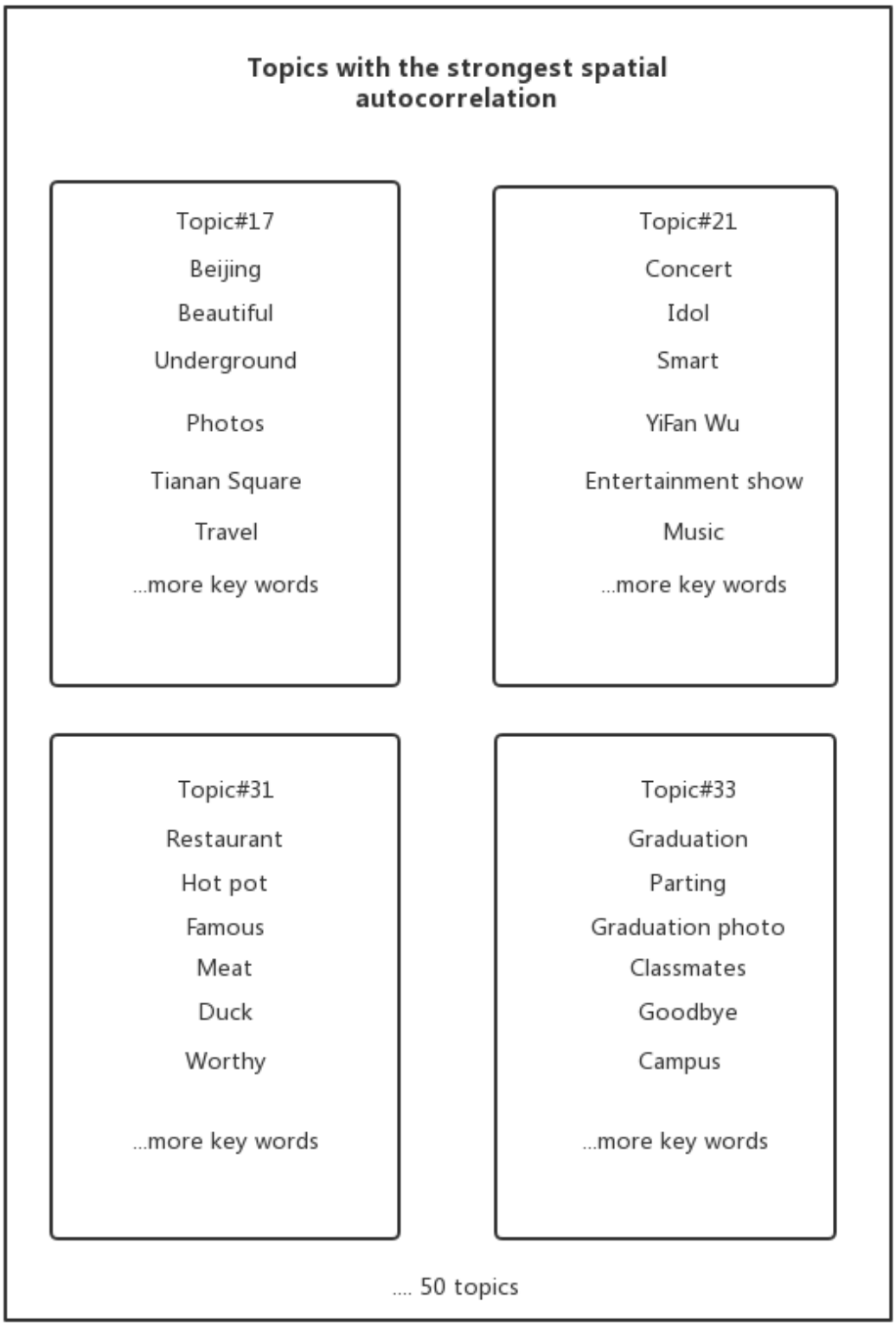



**Figure 6.5: Topics' key words**

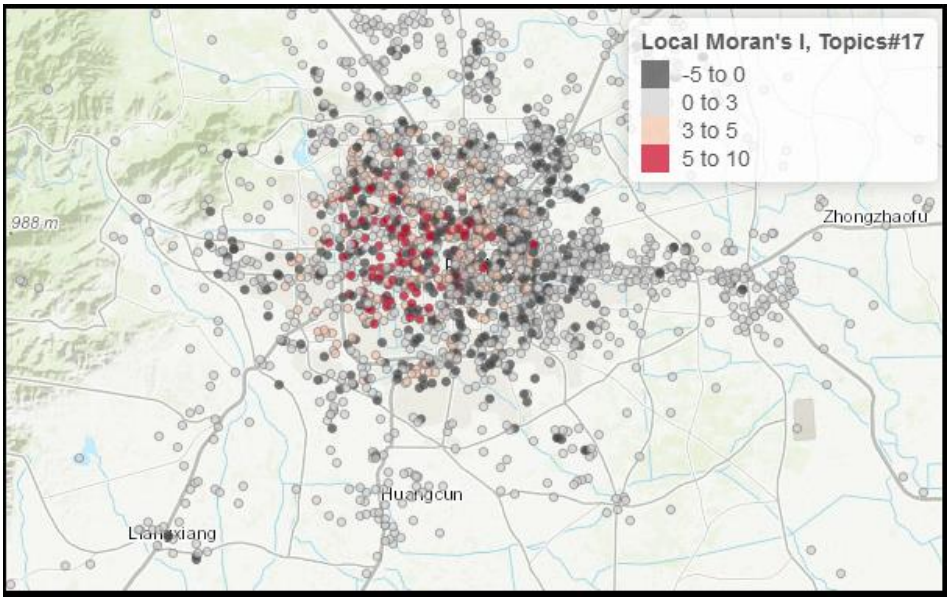



**Figure 6.6: Local Moran's I of topic #17 probability**

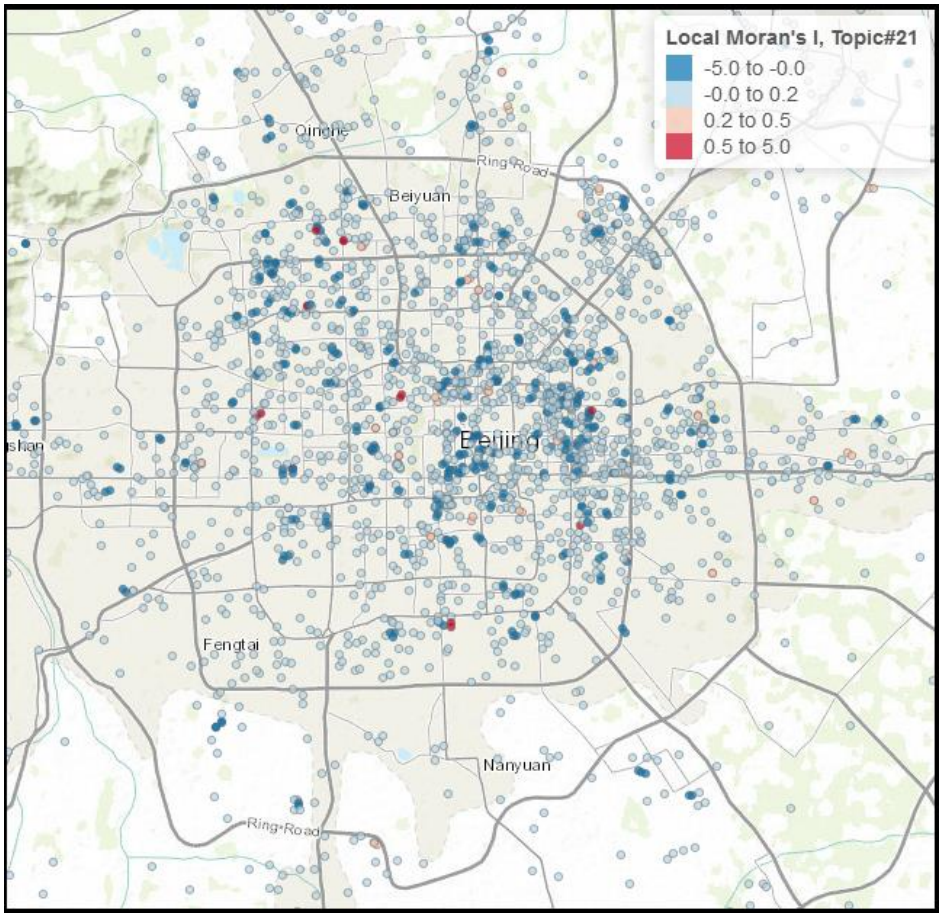



**Figure 6.7: Local Moran's I of topic #21 probability**

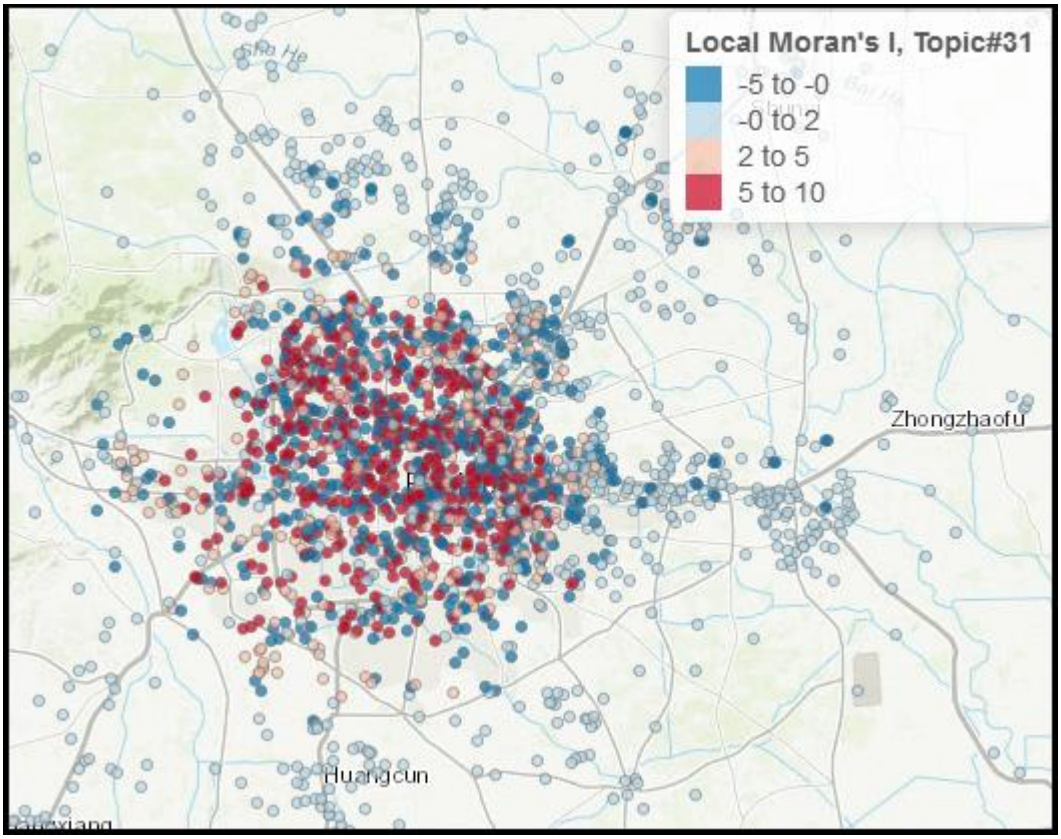



**Figure 6.8: Local Moran's I of topic #31 probability**

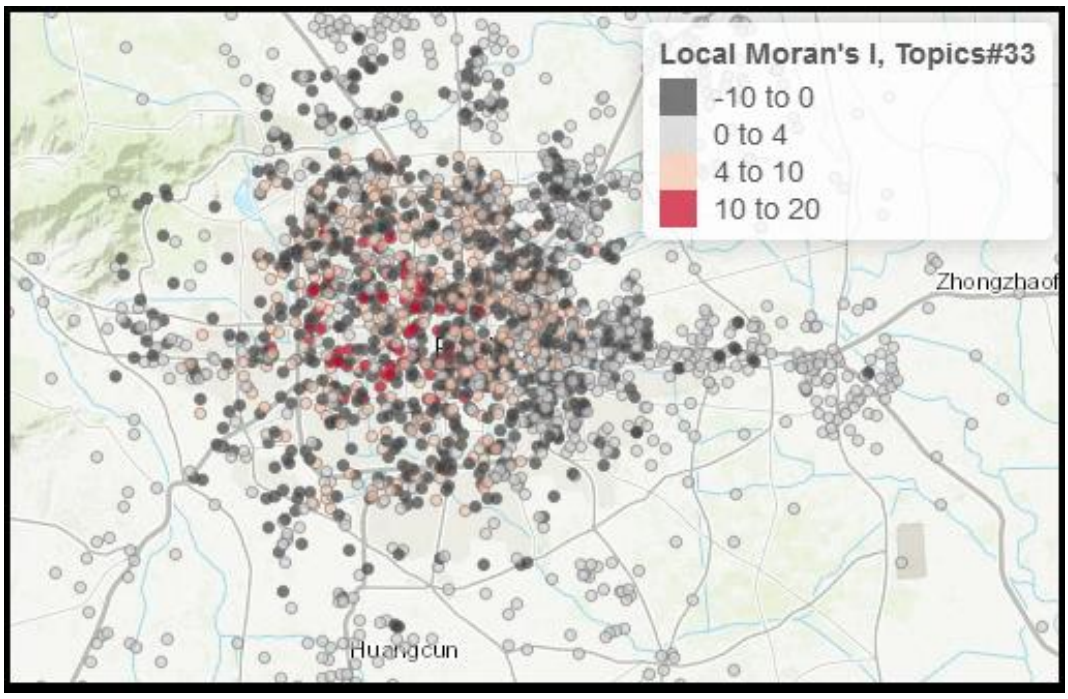



**Figure 6.9: Local Moran's I of topic #33 probability**

DBSCAN is being used to cluster all places, and the 50 topics probability (LDA) for each location represent the coordinates of places in multi-dimensional space. Before applying the

clustering algorithm, two parameters are critical and to be set. First, the epsilon that represents the maximum distance between two samples for them to be considered as in the same neighborhood. Secondly, the min-points, the number of points (or total weight) in a circle within the maximum distance for a position to be considered as a core point. As for the epsilon, all pairs of places' distance in multi-dimensional has been calculated, and the result is shown in Figure 6.10. Most pairs of places distance are distributed from 0.25 to 0.45. The distance to the nearest neighbor has been estimated, and the result is shown in Figure 6.11. The Epsilon is set as 0.2.

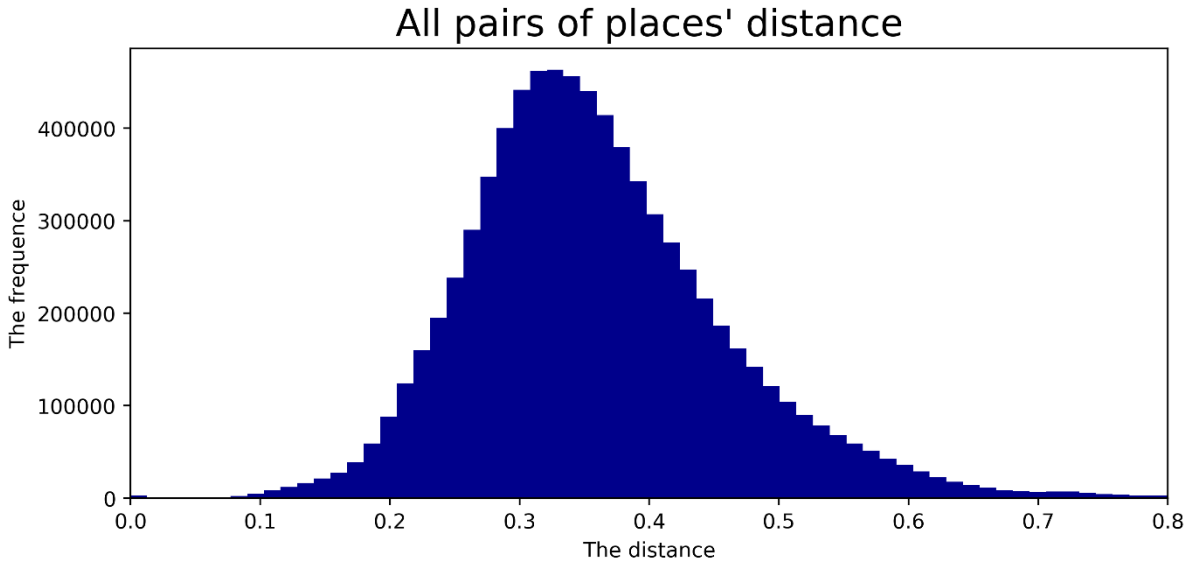



**Figure 6.10: All pairs of places' distance**

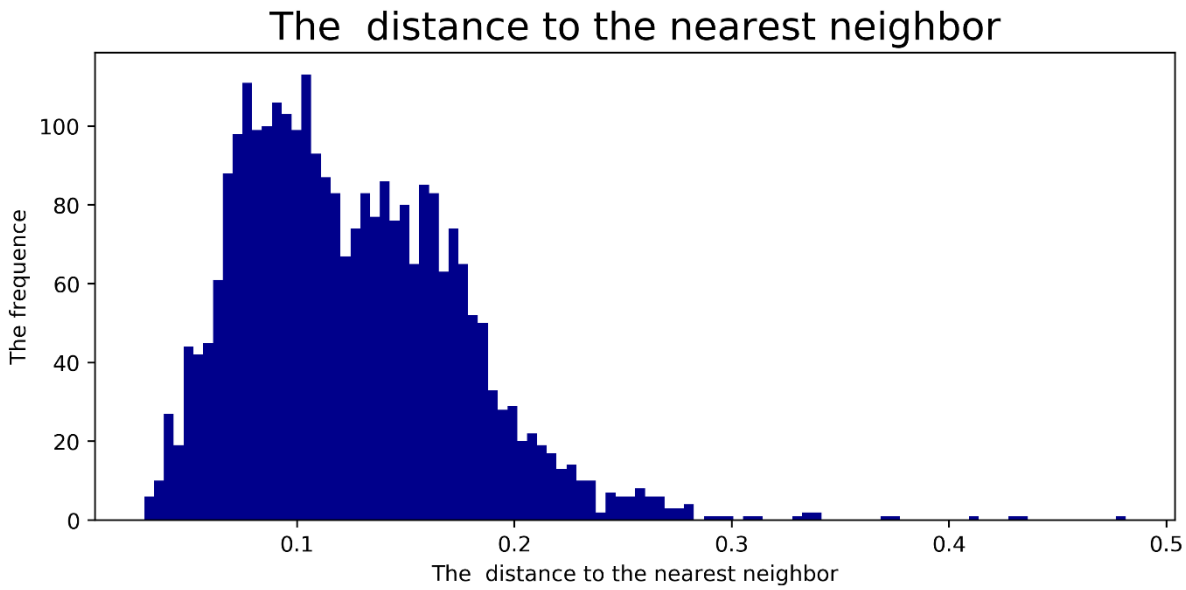



**Figure 6.11: The distance to the nearest neighbour.**

After the estimation of epsilon, how many points lie within each point's epsilon-neighborhood has been computed and the result is given in Figure 6.12. It means that some positions (about 230, which is 7.93% of all points) have too few neighbors (less than 30), probably are noise points. A more significant fraction (about 330, which is 11.4% of all positions) own neighbors from 30 to 100 and starting at 80, the number of neighbors begins to be stable. Based on the histograms above, the epsilon and min-points are set as 0.25 and 100 respectively at first, and a few tests around the value are also to be accomplished to explore the cluster of places.

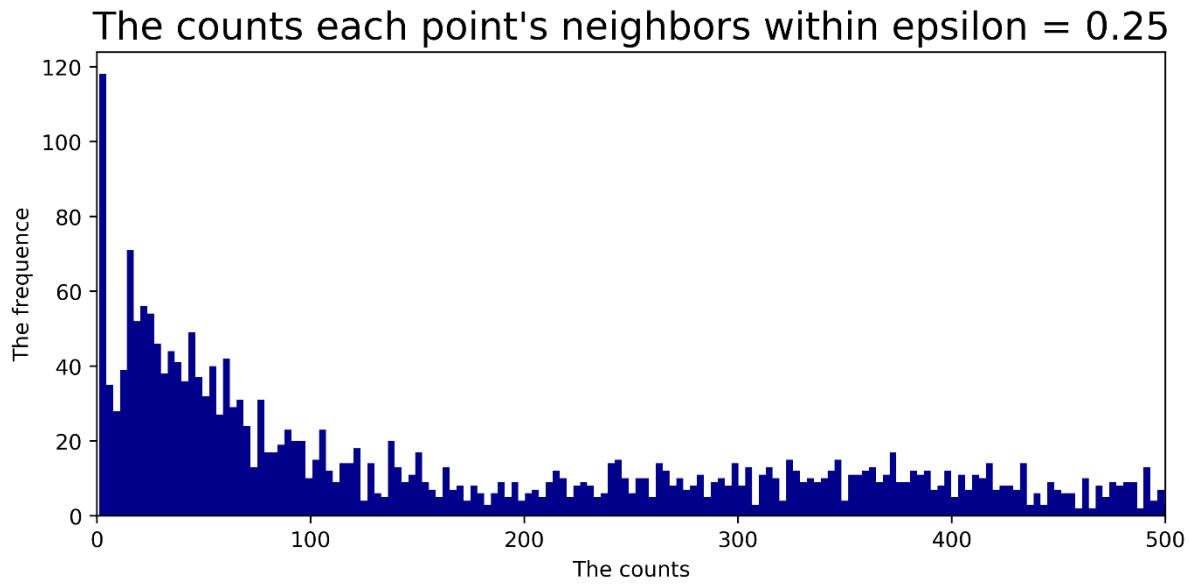



**Figure 6.12: The counts each point's neighbours (epsilon = 0.25)**

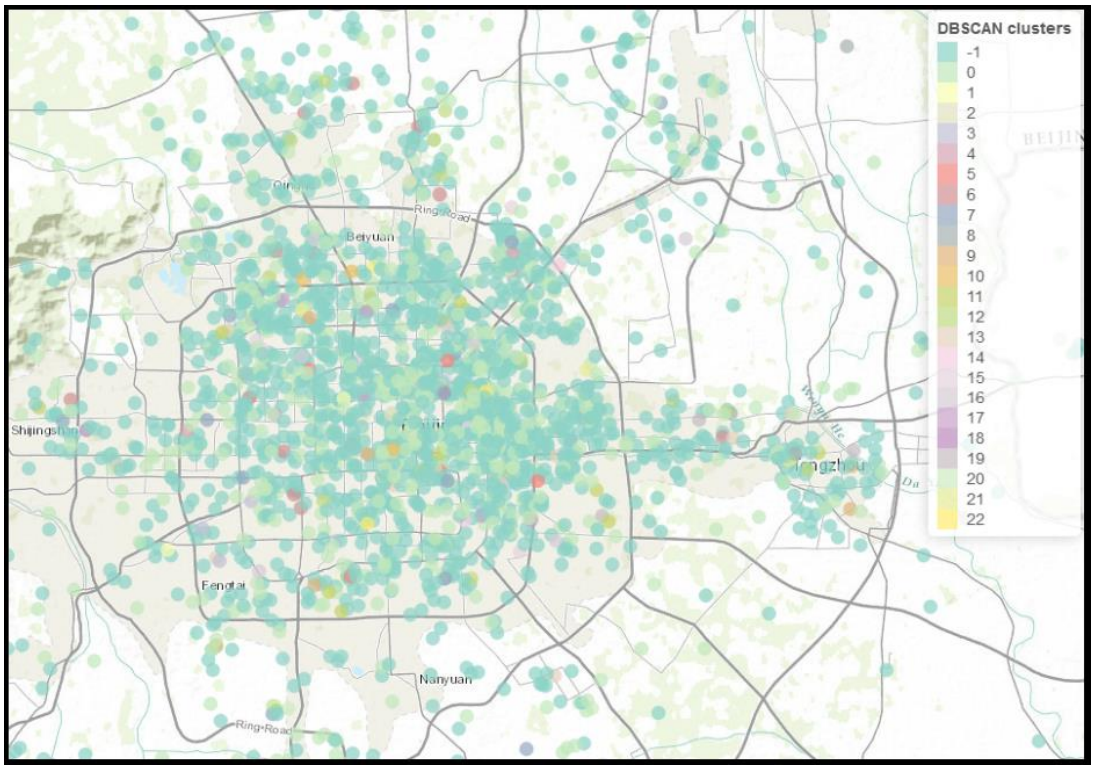


**Figure 6.13: The DBSCAN result (epsilon = 0.25, min-points = 100)**

The result of DBSCAN based on the 50 topics probability is shown in Figure 6.13, about 1000 places are being classified as noise (-1), and 400 sites being classified as 0. Other sites are classified as 26 clusters. Generally, the distribution in Figure 6.13 does not show a distinct group and centrality. To check the DBSCAN, the average cosine for all pairs of places in the specific cluster is computed, and the result is shown in Figure 6.14. By comparison, based on the 4-dimensional space (Topic #17, 21, 33 and 33 probability), DBSCAN is applied, and the average cosine for all pairs of places in each cluster is shown in Figure 6.15. It was evident that the average cosine for places in all groups is higher than that in the group -1 (noise places) and 0. And the average cosine similarity is all about 0.9, which means that the clusters produced by DBSCAN based on the 50 topics probability relatively successfully include places that are more similar than places in other groups, and far more similar than places in the cluster -1 and 0. As for the cluster based on the four topics that show strong spatial autocorrelation, the average cosine similarity for places in the specific group is lower than that in clusters of 50-dimensional space. Also, the average cosine for some clusters is lower than that for the bunch -1(noise) and 0, it means that the groups do not include these similar places and the four topics do not cluster places appropriately.

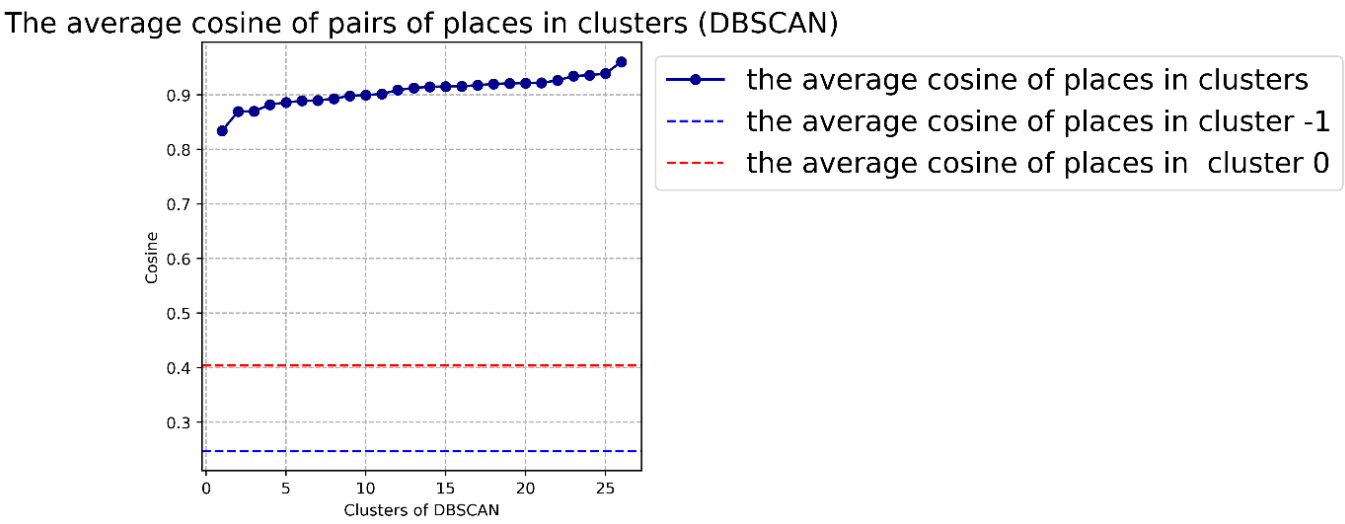



**Figure 6.14: The average cosine for all pairs of place in certain cluster.**

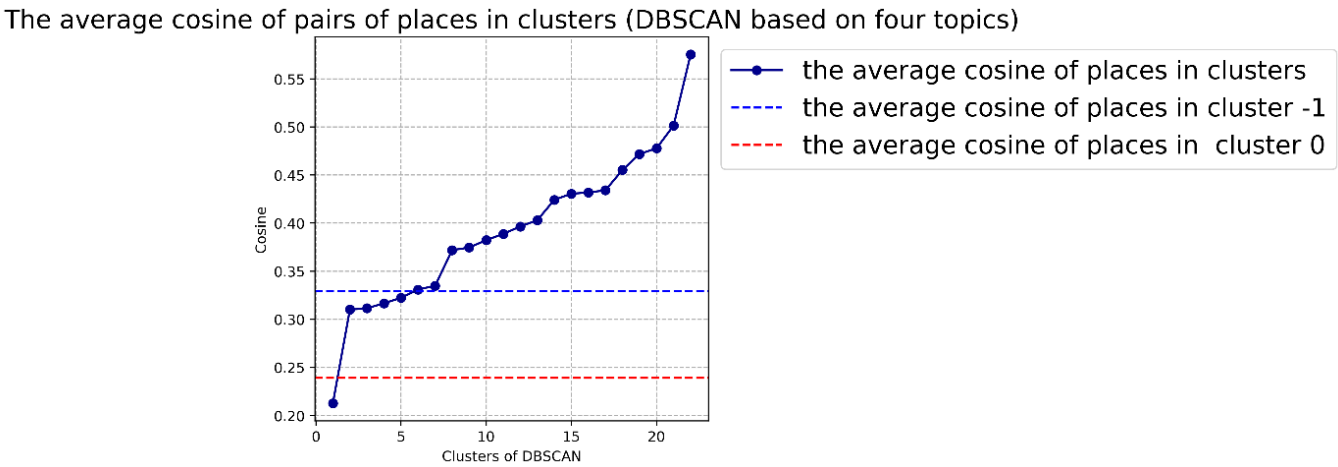



**Figure 6.15: The average cosine for all pairs of place in certain cluster (4-dimensional space)**

# 7. Research findings

## 7.1 Extract places and mine topics for places

As for the algorithm extracting places, the original purpose that Andrienko (2013) proposed the algorithm was to recognize the human behaviors by GPS data. It can be applied in identifying private and public places as if there is abundant data. As discussed above, the radius threshold of sites is fixed comparatively for a specific city. Thus, all places would be extracted once the number of points data exceed a limit. In this paper, if the N is set as 40, there would be more than 30,000 places based on 1.5 million point data. 30, 000 is too big for research as it needs too much computation power. Also, many places of the 30, 000 places cover too few points hence the information is insufficient for extracting and analyzing the topics of places. The data limitation might filter some crucial places as there are too few points collected in the place. Some organizations like Google map and Baidu map provide the map of Place of Interest (POI), it is feasible to combine the places extracted by OSM data with POI to build a data source of places in the urban area to analyze the real-time semantic features. The 2850 places extracted in this paper own at least 270 tweets and places in the central area cover more tweets and a smaller radius than that in outside. These merge places with larger radius are distributed

in the outside of the central region, which reveals that, in Beijing, there is a more broader scale of places in suburbs.

LDA is an unsurprised learning method. Thus, its goodness is significantly influenced by the parameters. The perplexity is usually used to measure the goodness of LDA, PLSA and other algorithms extracting text topics. The training results' perplexity of LDA in this paper is about 10, 000 and the parameters with the lowest perplexity has been adopted. In piratical, the training results of LDA and PLSA are applied to cluster users or texts as one topic's probability in the topic vector can be considered as property or feature of the tweet. Hence the average of topic vectors can be used to represent the places' property and features. The cosine similarity and DBSCAN based on the topic vectors for places appeared to reveal some characteristics of the city and places in this paper, therefore, the LDA – topics extracting algorithm perform well. Similarly, LDA has been used to cluster users of OSM to optimize news feed or advertisement and goods recommendation systems. However, the tweets on OSM like Twitter and Sina microblog are too short to extract topics precisely sometimes, which is a dilemma for clustering or analyzing contents on OSM. Sun et al. (2011) proposed that the relationship at @ could be used to improve the goodness of LDA when clustering contents on OSM as the relationship @ shows strong relativity between two tweets or users. Besides, the tweets pushed by a user also can be used to help extract the topic of contents as there is also strong relativity between materials produced by one user. The two points can be applied to improve the analysis of contents on OSM in the future.

## 7.2 Spatial similarity of places

It is evident that the places are more similar with near places than distant places. The negative correlation between distance and cosine similarity does not show significant centrality but the difference between in the east and west of Beijing. Kendall's Tau tests have proved that Tobler first law of geography work in the OSM contents inside Beijing. Besides, the places' cosine similarity with nearest neighbors or places within 50m, 500m, and 1000m all show strong centrality. It means that in the central area of Beijing, places are more similar to their neighbors than the outside area. The density of places and the number of tweets in a place in the central

area are both higher. It means that the difference between places in the central area of the city is less significant than that in the suburbs, namely the function, collectives or property of places in the central area are more similar.

The places' features reflected by tweets are representative to some extent. There are only a few places own neighbors with 50m, and many places own neighbors with 500m, which is consistent with our threshold of extracting places. In the central area, many places' cosine similarity with places with 500m and 1000m is higher than 0.8. The significant similarity shows the users and contents in the central are more even than other areas. The diversity reflected by the places' cosine similarity is more significant in the suburbs. According to the latest data, more than half of netizens in Beijing, Shanghai and other big cities in China use Sina microblog. What can be concluded from the analysis of cosine similarity is that in the central area of Beijing, the neighbor places' topic vectors are significantly similar and in the suburbs, the diversity at the same scale is obvious. However, the Tobler fist law of geography work overall even though the rules how to extract places and analyze the places' similarity need to adjust in the suburbs.

## 7.3 Places' features

According to Moran's I and Geary's C of topics, some topics (features)'s distribution shows clusters and some topics are distributed dispersedly. With distance criterion of neighborhood 100m, many topics show stronger spatial autocorrelation than that with distance criterion 500m and 1000m. It means that places with 100ms own some shared features reflected by the topic probability and the shared features can be considered as the characteristics of the area. For instance, the topic #33 related to the graduation and #17 related the travel show strong spatial autocorrelation, high probability (places with public property) are clustered in the central area or particular part, which reveal the real these places and topics' spatial-temporal features. It can be an index of the real-time public semantic analysis. It is also worth noticing that the spatial autocorrelation with neighborhood based on the KNN is not as significant as that based on distance criterion 100m, which can be explained as the neighbor places sometimes own different features that may be complementary.

The result of DBSCAN based on the topic probability vector for each places shows that clusters include similar places, namely the group does work well in extracting similar places and mining the features of places because the average cosine similarity of places in specific cluster is far higher than that of noise cluster and the average cosine similarity of all places. Besides, the four topics with strong spatial autocorrelation are insufficient to cluster places by DBSCAN. Therefore, the significant goodness of DBSCAN based on the topic probability vectors can be the fundament of applying LDA-SVM model and DBSCAN to analyze the contents on OSM and explore the features of cities, urban districts and places in the future.

# 8. Conclusion

## 8.1 Conclusion

In this paper, the topics of tweets have been extracted by LDA, and the resulted topic probability in the topic vectors is considered as a property of tweets. Therefore, the average of topic vectors in places that have been extracted by the algorithm based on the density and fixed radius range is considered as the features and property of places. The topic vectors of places work as the term vector in SVM. According to the cosine similarity between places and their neighbors or other places, it can be concluded that the Tobler's first law does work in the likeness of places' topic vectors, which means that the places' contents on OSM are more similar to near places than distant places. However, the negative correlation between distance and cosine varies depending on the area, in the central area, neighbors' contents on OSM are more similar than that in the suburbs. Also, the diversity of places' contents is more significant in the central area than that in the suburbs. As for the shortage of places, some phenomenon found in the central area has not be found in the suburbs. Besides, in the central area, the nearest one or two places' features or functions may be different or complementary even though the average of the cosine of places within 1000m is high.

As for the features reflected by the topic probability in topic vectors, some topics show strong spatial autocorrelation. The distribution of high likelihood can help analyze the distribution of the feature or topic inside cities and at the scale of places. The DBSACN clusters based on the vectors of places work well in extracting similar sites. The significant goodness proves that the

LDA and DBSAN are feasible to be adopted in exploring the contents on OSM in an urban area, which is a trending field as the UGC increase dramatically on the internet nowadays.

## 8.2 Limitation and future work

The apparent deficiency in this research is the data limitation as it is impossible for researchers to collect all data even just in a city. It would be more difficult to take account for the temporal factors according to the data protection and privacy policies. However, the deficiency is also the motivation that encourages this paper. The shortage of tweets in some area make some essential places ignored. Thus, the analysis of correlation and spatial autocorrelation are both influenced to some extent. Besides, lacking sufficient data make it difficult to take account for some temporal factors. Some methods of UGC-LI (user-generated content driven location inference method) have also been proposed. Hence, the results of this paper can be used to improve UGC-LI methods and make it easy to collect places, area or cities' features through contents on OSM for other researchers.

As for the places, the algorithm still might ignore some smaller or bigger places even though the algorithm already take account for those places by some adjustment because the size of sites sometimes vary dramatically and the GPS projection usually is not precise especially when there is no WIFI network. In the future, the data source for POI provided by some organizations can be optimized by the algorithm. Thus, it is not necessary for researchers to extracting places whenever they want to analyze sites inside an urban area.

The relationship at @ and contents posted of one user can both be considered as assistance to improve the topics extracting by LDA in the future. The topics probability is an index of places' property. Therefore, the topic probability produced by LDA can be combined with other attributes or data of places to discover the characteristics, trends, and situation of the urban area.

# Reference:

Kaplan, A.M. and Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of Social Media. Business horizons, 53(1), pp.59-68.

Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: Concepts, methodologies, and applications. ACM Trans. Intell. Syst. Technol. (TIST) 5, 3 (2014), 38.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z. and Soltani, K., 2015. A scalable framework for spatiotemporal analysis of location-based social media data. Computers, Environment and Urban Systems, 51, pp.70-82.

Liu, B., Li, Y., Xue, B., Li, Q., Zou, P.X. and Li, L., 2018. Why do individuals engage in collective actions against major construction projects? — An empirical analysis based on Chinesedata. International Journal of Project Management, 36(4), pp.612-626.

Kaltenbrunner, A., Scellato, S., Volkovich, Y., Laniado, D., Currie, D., Jutemar, E.J. and Mascolo, C., 2012, August. Far from the eyes, close on the web: impact of geographic distance on online social interactions. In Proceedings of the 2012 ACM workshop on Workshop on online social networks(pp. 19-24). ACM.

Ostermann, F.O., Huang, H., Andrienko, G., Andrienko, N., Capineri, C., Farkas, K. and Purves, R.S., 2015. Extracting and comparing places using geo-social media. ISPRS GEOSPATIAL WEEK 2015, 2(W5).

Sun, G., Tang, T., Peng, T.Q., Liang, R. and Wu, Y., 2017. Socialwave: visual analysis of spatio-temporal diffusion of information on social media. ACM Transactions on Intelligent Systems and Technology (TIST), 9(2), p.15

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P. and Tomkins, A., 2005. Geographic routing in social networks. Proceedings of the National Academy of Sciences, 102(33), pp.11623-11628.

Leskovec, J., Backstrom, L. and Kleinberg, J., 2009, June. Meme-tracking and the dynamics of the news cycle. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 497-506). ACM.

Hofstede, G., Hofstede, G.J. and Minkov, M., 2010. Cultures and Organizations: Software of the Mind, and McGraw-Hill USA.

David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. 2005. Geographic routing in social networks. Proc. Natl. Acad. Sci. U.S.A. 102, 33 (2005), 11623–11628.

Tai-Quan Peng, Mengchen Liu, Yingcai Wu, and Shixia Liu. 2015. Follower-followee network, communication networks, and vote agreement of the US members of congress. Commun. Res. (2015), 0093650214559601.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. Annu. Rev. Sociol. (2001), 415–444.

Hahmann, S., Purves, R. and Burghardt, D., 2014. Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. Journal of Spatial Information Science, 2014(9), pp.1-36.

Zhang Chenyi, Sun Jianwei, and Ding Weiqun. "Microblogging Theme Mining Based on MB-LDA Model." Computer Research and Development 48.10(2011): 1795-1802.

Wang Bo, Qi Feng, Xi Guangliang, et al. Research on Network Information Geography Based on Weibo User Relationship——Taking Sina Weibo as an Example[J]. GEOGRAPHICAL RESEARCH, 2013, 32(2): 380-391.

Andrienko, N., Andrienko, G., Fuchs, G. and Jankowski, P., 2016. Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. Information Visualization, 15(2), pp.117-153

Agnew, J., 1987. Place and Politics: The Geographical Mediation of State and Society. Boston and London: Allen and Unwin.

Agnew, J., 2011. Space and place. In: Agnew, J., Livingstone D. (eds.). The SAGE handbook of geographical knowledge, London, SAGE Publications Ltd., pp. 316-330.

Teobaldi, M., Capineri, C., 2014. Experiential tourism and city attractivness in Tuscany. Rivista Geografica Italiana,121, pp.259-274

Huang, H., Gartner, G. and Turdean, T., 2013. Social media data as a source for studying people's perception and knowledge of environments.

Winter, S., Freksa, C., 2012. Approaching the Notion of Place by Contrast. Journal of Spatial Information Science, 5, pp. 31– 50.

Purves, R.S., Edwardes, A., Wood, J., 2011. Describing Place through User Generated Content. First Monday, Volume 16, Number 9 - 5 September 2011.

Andrienko, G., Andrienko, N., Hurter, C., Rinzivillo, S. and Wrobel, S., 2013. Scalable analysis of movement data for extracting and exploring significant places. IEEE transactions on visualization and computer graphics, 19(7), pp.pp-1078.

Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (pp. 160-167). ACM.

Raghavan, V. V., & Wong, S. M. (1986). A critical analysis of vector space model for information retrieval. Journal of the American Society for information Science, 37(5), 279-287.

Mnih, A., & Hinton, G. (2007, June). Three new graphical models for statistical language modelling. In Proceedings of the 24th international conference on Machine learning (pp. 641-648). ACM

Chowdhury, G. G. (2010). Introduction to modern information retrieval. Facet publishing.

Sebastiani, Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR) 34.1 (2002): 1-47.

Cogsys.imm.dtu.dk. (2018). Introduction: Vector Space Model. [online] Available at: http://cogsys.imm.dtu.dk/thor/projects/multimedia/textmining/node5.html [Accessed 13 Jul. 2018].

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of. Reading: Addison-Wesley.

Wang Bo, Yan Feng, Xi Guangliang, Qian Qian, Wu Chengyue, & Zhang Hao. (2013). Research on Network Information Geography Based on Weibo User Relationship——Taking Sina Weibo as an Example. Geography Research, 32(2), 380 -391.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391-407.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. Handbook of latent semantic analysis, 427(7), 424-440.

Mei, Q., Cai, D., Zhang, D., & Zhai, C. (2008, April). Topic modeling with network regularization. In Proceedings of the 17th international conference on World Wide Web (pp. 101-110). ACM

Dietz, L., Bickel, S., & Scheffer, T. (2007, June). Unsupervised prediction of citation influences. In Proceedings of the 24th international conference on Machine learning (pp. 233-240). ACM.

Wei, Xing, and W. Bruce Croft. "LDA-based document models for ad-hoc retrieval." Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research3.Jan (2003): 993-1022

Hofmann, T. (2017, August). Probabilistic latent semantic indexing. In ACM SIGIR Forum (Vol. 51, No. 2, pp. 211-218). ACM

Vockner, B., Richter, A., Mittlböck, M.. 2013. From Geoportals to Geographic Knowledge Portals. ISPRS International Journal of Geo-Information 2 (2): 256–75.