Identifying city center using human travel flows generated from location-based social networking data

Yeran Sun, Hongchao Fan, Ming Li, Alexander Zipf

Abstract

Since cities become more complex and some of large cities are likely to be polycentric, a better understanding of cities requires a clear topology that reveals how city centers are spatially distributed and interacted. The identification of city center that aims to find out accurate location of city center or delineate city center with a precise boundary becomes vital. This work attempts to achieve this by using a new type of movement data generated from location-based social networks, whereby three different methods are deployed for clustering and compared regarding identification of city centers and delineation of their boundaries. Experiments show that city centers with precise boundaries can be identified by using the proposed approach with location-based social network data. In further, it finds out that the three methods for clustering have different advantages and disadvantages during the process of city center identification, and thus seem to be suitable for cities with different urban structures.

Keywords: center identification; human mobility; local Getis-Ord $\boldsymbol{G_i^*}$; DBSCAN; Grivan-Newman

1. Introduction

The structure of cities is one of the most vital issues in urban studies. A better understanding of the structure of a city facilitates urban planning, policymaking, resource allocation and traffic monitoring, among other things. City centers are core of cities and are areas with clustering of socio-economic activities (Anas et al., 1998). Previous studies have revealed that large cities, which can be considered complex systems, tend to be polycentric (e.g., Roth et al., 2011). Since cities become more complex and some of large cities are polycentric, a better understanding of cities requires a clear topology that reveals how city centers or sub-centers are spatially distributed and interacted (Kloosterman and Musterd, 2011). The spatial arrangement of city centers and how individuals interact with these centers is a crucial problem with many applications ranging from urban planning to epidemiology (Roth et al., 2011). The most prominent and visible effects of such spatial organization of economic activity in large and densely populated urban areas are characterized by severe traffic congestion and the strong possibilities of rapidly spreading viruses, biological and social, through the dense underlying networks (see Eubank et al., 2004; Balcan et al., 2009; Wang et al., 2009). Therefore, the identification of city center that aims to find out accurate location of city center or delineate city center with a precise boundary becomes vital. On the other hand, since 'central places' theory is widely used, some models are proposed to simulate some urban phenomena (e.g., population distribution, supply of transport infrastructure, etc.) (see Clark, 1951; Alonso, 1960). These simulation models are based on the distance from city center, enabling a reasonable identification of city center to become vital in many urban applications or urban studies (e.g., population density estimation, site selection of service facilities, traffic monitoring, etc.).

Conventionally, city centers are detected using socio-economic data collected by authorities (e.g., Thurstain-Goodwin and Unwin, 2000). In the recent years, more and more researchers (e.g., Ratti et al., 2006; Roth et al., 2011; Jiang et al., 2012) tried to analyze urban structures by using movement data, because it can reveal the interactions between different urban centers in a polycentric city (Roth et al., 2011). However, the research works using movement data are limited so far to detect city centers for medium-sized and monocentric cities (e.g., Thurstain-Goodwin and Unwin, 2000; Borruso and Porceddu, 2009; Lüscher and Weibel, 2012). Roth et al. (2011) tried to apply subway record data for a city with more than one centers. But their approach is not able to delineate precise boundaries of city centers. Therefore, how to use a variety of mobility data (e.g., subway record, taxi GPS traces, social media geo-data, etc.) to identify city centers with accurate positions or precise boundaries in relatively complex cities becomes a new research direction.

The presented work attempts to fill the abovementioned research gaps, namely, to detect city centers in polycentric cities and determine the boundaries of city centers in this work, whereby location-based social networking (LBSN) data is used, because geo-referenced and time-stamped 'check-in' (sometimes referred to as a type of volunteered geographic information (VGI) can be used to indicate user mobility (e.g., Noulas et al., 2011; Scellato et al., 2011; Cheng et al., 2011; Wei et al., 2012; Bao et al., 2012). First of all, human mobility information and also the travel flow count for a venue are generated using LBSN check-ins. Then travel flow clusters are detected by means of three distinct methods and significant clusters are selected to differentiate potential city centers from non-centers. The approach is deployed to identify city centers for three German cities (Berlin, Munich and Cologne) using the movement data generated by LBSN check-ins. At the same time, this study empirically tests the validity of using LBSN data for city center identification. To better test the validity, three distinct methods are used to identify city centers. Obviously, the validity of using LBSN data for city center identification will be better proved by good performances of distinct types of methods than that by only one type of method. Moreover, we will compare the three distinct methods, to discuss the advantages and disadvantages the three methods have for city center identification and which types of urban structures they are likely to be suitable for.

The remainder of this paper is organized as follows. Section 2 introduces previous studies, while Section 3 introduces the approach used in this study. Section 4 presents how the empirical analysis was carried out for this case study and the relevant results, and lastly, Chapter 5 presents the conclusion and makes recommendations for future work.

2. Related works

Since the 1950s, there have been research works to theoretically define or delineate city center using limited data sources (see e.g., Murphy and Vance, 1954; Carol, 1960; Alonso, 1964; Murphy, 1972). Over the last decade, a number of approaches have been made available to quantitatively delineate the city center using a variety of data sources. In Thurstain-Goodwin and Unwin (2000) a city center was delineated by creating an index called ‘index of town centeredness’, which is composed of a series of indicators to represent the typicality of a city center. The ‘kernel density’ estimation method is then used to transform the discrete geo-referenced data created by the relevant indicators into continuous surfaces denoting spatial densities. In terms of the individual density values of the indicators, a continuous surface for the ‘index of town centeredness’ was generated. A similar study was made by Borruso and Porceddu (2009). To present a behavioral science method for determining the referents of vague spatial terms, and particularly vague regions, Montello et al. (2003) asked pedestrians to draw the city center with 100% confidence and 50% confidence respectively. Using remote sensing data, Taubenboeck et al. (2013) presented a conceptual framework to define the CBD using physical and morphological parameters, and tests the approach using 3D city models of three European test sites. A transferable method was developed to detect and delineate CBDs over larger areas from a combination of Cartosat-1 digital surface models and multispectral Landsat ETM+ imagery.

With the development of mobile devices and the popularity of social media (Twitter, Facebook, Flickr and Foursuqare, among others), user generated geo-data (including geo-referenced images and geo-referenced check-ins) shows the potential to help delineate a city center. For instance, Hollenstein and Purves (2010) used geo-referenced images from Flickr to describe a city’s core, using image tags such as ‘downtown’, ‘central’, ‘cbd’, ‘inner city’ and ‘city center’. Similarly, they used the kernel density estimation method to transform these typical geo-referenced images, which were represented as spatial points, into continuous spatial density surfaces. To better quantify the typicality of a city center, Lüscher and Weibel (2012) sought to combine the approaches of Thurstain-Goodwin and Unwin (2000), and Montello et al. (2003). They created an index to represent the typicality of a city center which is composed of frequency-based characteristics, landmark-based features and area-like characteristics. The weight of each characteristic was determined based on a questionnaire. Specifically, participants were asked to classify the facilities (such as restaurants, museums, bars, shops and railway stations) into three types of characteristic, and further assign a weight to each. Empirical studies have demonstrated that their proposed approach performs well when delineating city centers.

In the above-mentioned works cities for test purpose are not typically polycentric cities (e.g., Thurstain-Goodwin and Unwin, 2000; Montello et al., 2003; Lüscher and Weibel, 2012). More recently, movement data has been used to identify city centers or sub-centers in typically polycentric cities. Compared to socio-economic data, movement data can not only identify urban centers but also reveal the interactions between different urban centers in a polycentric city. For instance, Roth et al. (2011) reveal that the majority of subway flows were distributed among a small number of stations, indicating that human activity was concentrated in a small number of centers dispersed across the city, indicating a polycentric structure in London.

3. Detection of city centres using LBSN data

In this section, the proposed approach of detection of city centres using LBNS data is presented. First of all, LBSN data (mainly check-in) is introduced in terms of its general characteristics and representativeness. Check-ins will be generated into movement data firstly. Then clusters will be detected for the identification of city centers. After that, city centers will be identified. At last, the identified city centers will be validated.

3.1 Introduction of check-in data

Chowell et al. (2003) mapped a city into a network composed of nodes, each of which represented a physical location such as a building. In location-based social media such as Foursquare, each venue also represents a physical location (see Figure 1), and in this case a 'venue' can be considered a Point-Of-Interest (POI), one widely used. Common types of venue include restaurants, offices, apartments, hotels, bus stops, shops and gyms.

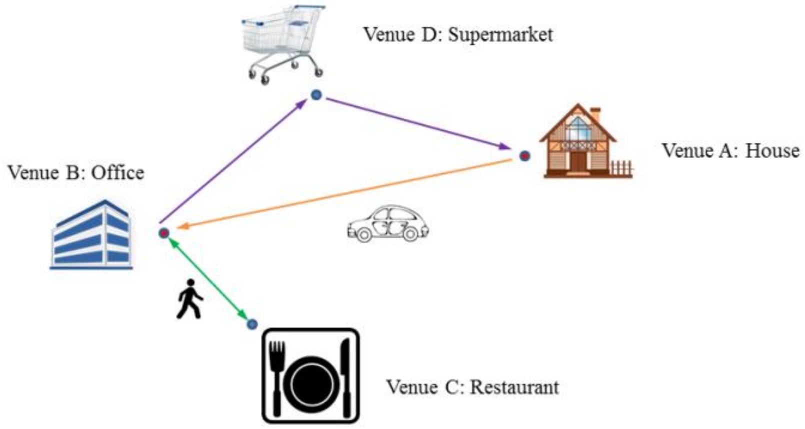


Figure 1: Examples of user mobility

In Figure 1, for instance, a user checks in at venue *A* (a house) and venue *B* (an office) consecutively. We can therefore deduce that there has been a 'movement' from venue *A* to venue *B*, irrespective of the specific route taken by the user between the two venues. Sometimes, a user might move twice between two venues, in opposite directions, such as moving from an office to a restaurant before lunch, and moving from the restaurant to the office after lunch. In studies of human mobility, 'displacement' is widely used to measure the length of a user movement (e.g., González et al., 2008). In this context, 'displacement' is defined as a vector, 1) whose length equals the distance between two consecutive positions (venues), and 2) whose direction is from the initial position (original venue) to a final position (destination venue). In Figure 1, there is a 'displacement' from venue *A* (house) to venue *B* (office). Therefore, pairs of consecutively geo-referenced and time-stamped check-ins can constitute displacements of users. Moreover, a displacement can be considered a travel flow. From the original venue (starting position) the displacement is an outflow, while to the destination venue (final position) the displacement is an inflow. For a venue, the total flow count is equal to the sum of the outflow and inflow counts. In Figure 1, the outflow and inflow counts for venue *A* are both one, thus the total flow count is two.

Note that despite some limitations on representing human mobility, e.g., the bias of age group and bias of place category, check-in data has the ability to represent human mobility. According to the statistics of Foursquare (www.factbrowser.com/tags/foursquare/), a large proportion of its registered users are young and the users are likely to check in at particular places such as airports (Liu et al., 2014). Users of LBSNs check in at commercial locations much more often than residential locations. At the same time, some existing literature show that the commercial locations (e.g., restaurant, retail, clothing shop, pub, etc.) are more important to represent city center than residential locations (e.g., apartment, private home, etc.) (Borruso and Porceddu, 2009; Lüscher and Weibel, 2012). In this case, popular locations in LBSNs are also likely to be important locations used to represent city center. For instance, obviously retails and clothing shops contribute more to the identification of city center than apartments and residential houses. And retails and clothing shops also have more check-ins made by LBSN users than apartments and residential houses. This implies that the heterogeneity in the popularities of LBSN venue categories is somewhat consistent with the heterogeneity in the contributions of venue categories to the identification of city center. Thus, the important venues for center identification are likely to be included by LBSN data, reducing the influence of LBSN data bias on the city center identification.

3.2 *Cluster detection*

In fact, flow count in space is strongly impacted by environmental characteristics e.g., population density, land use, etc. (Liang et al., 2013). If there is a spatially local cluster composed of venues 1) which are located closely to each other and 2) which have higher number of flows, this cluster might indicate the existence of a city center. The following three methods representing three types of clustering methods are more widely used in related works to detect clusters.

1) Local Getis-Ord G_i^* (LGOG)

The local Getis-Ord G_i^* statistical method (Getis and Ord, 1992; Ord and Getis, 1995) is a clustering method for either high values or low values. It is used here to identify significant clusters of high values, as constituted by venues with a large number of flows in a city. The local Getis-Ord G_i^* statistical method is widely used to identify clusters of high values (“hot spots”) or low values (“cold spots”) (O'Sullivan and Unwin, 2010). Indeed, there are two steps needed to finally identify cluster.

a) Detecting hotspots of travel flow

As a statistically significant hotspot, a venue must have a high value of flows and be surrounded by other venues with high values also. The value of the G_i^* statistic is used to indicate if a venue is a hot spot or cold spot. Apart from the flow count of the venue, the flow counts of its neighbors are taken account of by the calculation of the G_i^* statistic for the venue. Thus, neighbor distance *threshold* is used to identify the neighbors of a given venue, so that if the distance between a venue *i* and *j* was less than the distance *threshold*, *j* was a neighbor of *i*; otherwise, it was not. With a specific value of neighbor distance *threshold*, the value of the G_i^* statistic for each venue will be calculated.

The value of the G_i^* statistic is essentially a Z-score, hence no further calculations are required. Z-score = $(x - \mu) / \sigma$, where x is the observation, μ is the mean of population and σ is standard deviation of population. A Z-score is used to indicate statistical significance. For statistically significant and positive Z-scores, the larger the Z-score is, the more intense the clustering of high values. As Z-score is originally used in the characterization of normal distribution, a Z-score greater than 1.96 indicates the occurrence of an abnormal event at a significance level of 0.05. Thus, in this study, a statistically significant hotspot is a venue of a high value (a large number of flows) with a corresponding Z-score greater than 1.96. In other words, the hotspots are the venues with Z-scores greater than 1.96.

b) Detecting clusters

Compared to the other two methods, the originally LGOG method offers hotspots instead of clusters that are actually hotspot sets or hotspot groups. Thus, how to group spatially close hotspots into a hotspot set (i.e., cluster) necessarily needs to be determined. Here, a cluster is a hotspot set composed of hotspots 1) are ‘spatially close’ to each other; and 2) are not ‘spatially close’ to hotpot of other cluster. If two hotspots are neighbors of each other, i.e., the distance between them is less than the pre-defined neighbor distance *threshold*, they are considered to be ‘spatially close’ to each other. Using these two rules, distinct clusters will be finally distinguished.

2) Density-based spatial clustering of applications with noise (DBSCAN)

DBSCAN algorithm (Ester, et al., 1996) is a density-based clustering algorithm. In addition to DBSCAN, OPTICS, DENCLUE and their modified versions (Ester, et al., 1996; Ankerst et al., 1999; Hinneburg and Keim, 1998) are typical density-based clustering algorithms. DBSCAN algorithm can detect arbitrary-shaped cluster.

It starts with an arbitrary starting point that has not been visited. This point's ε-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized ε-environment of a different point and hence be made part of a cluster.

If a point is found to be a dense part of a cluster, its ε-neighborhood is also part of that cluster. Hence, all points that are found within the ε-neighborhood are added, as is their own ε-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

To run DBSCAN algorithm, there are two parameters required: *eps* (the maximum radius of a neighborhood) and *minPts* (the minimum number of points required to form a cluster). These two parameters can restrain the spatial range and size (point count) of cluster respectively.

3) Grivan-Newman (GN)

Grivan-Newman (GN) algorithm (Girvan and Newma, 2002) is originally proposed as a community detection algorithm. Some studies use community detection algorithms to investigate spatial relationship, particularly to detect spatially interacted places (e.g., Thiemann et al., 2010; Liu, et al., 2014). Like other community detection algorithms, GN algorithm is based on network or graph theory (Girvan and Newma, 2002). Imagine a venue can be considered as a ‘node’. If there are trips between two nodes (venues), these two nodes are connected. A connection between two nodes is an ‘edge’. In a city, the venues (nodes) and edges (connections) can constitute a network or graph. GN algorithm aims to identify a partition P of nodes into k modules (sub networks or sub graphs) so that the intra-connectivity of the modules in the partition is high and inter-connectivity is low. A fast algorithm of Newman (2001) is used to find out the optimum partition with the highest ‘modularity’. The ‘modularity’ is used to measure the partition P. A larger ‘modularity’ indicates a better partition with higher intra-connectivity and lower inter-connectivity. Therefore, GN is technically a connectivity-based method.

Specifically, this study assumes that venues are more likely to be connected with venues within an identical city center area than venues outside. A community composed of venues with a high intra-connectivity and a limitedly spatial range is likely to be a potential city center. Therefore, GN algorithm is used to identify city center as well. Since city center should have a limited spatial range, this study took only short trips into account. Although some studies detect successfully continuous and compact region with a high intra-connectivity without removing long trips (e.g., Thiemann et al., 2010; Liu, et al., 2014), they make use of inter-urban mobility data while this study uses intra-urban mobility data that is relatively sparse and heterogeneous. Therefore, long trips should be removed to generate a good result, i.e., venues that constitute an identical community should be spatially close to each other. There is an issue, i.e., how to determine the length of the ‘long’ trips here. This study, thus, adopts the GN algorithm by adding a parameter, i.e., *maxLen*, which means the maximum trip length. With a specific value of *maxLen*, the trips with a length larger than this value will be removed and the corresponding connections (edges) will be removed from the network or graph as well.

3.3 Identification of city centers

Based on the results of clustering, city centers are detected in three steps as follow:

1) Matching typical cluster to candidate landmark

Typical clusters need to be matched with the closest candidate landmarks. Here candidate landmarks are landmarks (e.g., central plaza, railway station, etc.) that are considered as central locations of potential city centers according to local knowledge. Such landmark is mapped as a point. The centroids of typical clusters will be matched with landmarks by using the shortest distance rule. Here the centroid of a cluster is the mass center of the points in the cluster. Thus, among several candidate landmarks, the one, who is the closest to the centroid of a typical cluster, will be matched with the cluster. If a landmark becomes the closest landmark of more than typical cluster, among these clusters the one with the shortest distance to this landmark will be chosen as the unique one matched with this landmark.

2) Determining city center

Among the typical clusters that are matched with candidate landmarks, the largest one i.e., the one with the largest flow count, will be initially considered as a city center since 1) this cluster could be matched with a candidate landmark; 2) and it is the most prominent cluster. After that, if another typical cluster could be considered as another city centers will be determined in terms of the *Rate(C)* defined as

$$Rate(C_j) = \frac{flow\_count(C_j)}{flow\_count(C_1)}, j = 2, \dots, k \qquad (1)$$

Where C_j (j=2,…,k) is the jth most typical cluster. *Rate(C)* is used to show the extent of the difference in flow counts between distinct clusters. A higher *Rate(C)* indicates a smaller difference of flow count between the cluster and the first largest cluster, meaning this cluster could be considered as another city center with a higher probability.

3) Delineating city center

After a cluster is considered as a city center, this city center with a precise boundary will be delineated. In this research, Voronoi diagram was used to divide a city into polygons. Each venue corresponds to a polygon. Figure 2 shows the Voronoi polygons generated from venues in a city. In Figure 2, a group of points (red points) constitute a cluster, and thus a group of polygons (red polygons) constitute a 2-dimensional object. Therefore, each cluster of venues (points) corresponds to a 2-dimensional object composed of polygons. The boundary of a cluster could be represented by the boundary of a 2-dimensional object (the blue outline). If a venue cluster represents a city center, the boundary of the cluster will be used to represent a precise boundary of the corresponding city center. Thus, the final result of identification is an area with a precise boundary (i.e., a Voronoi polygon group).

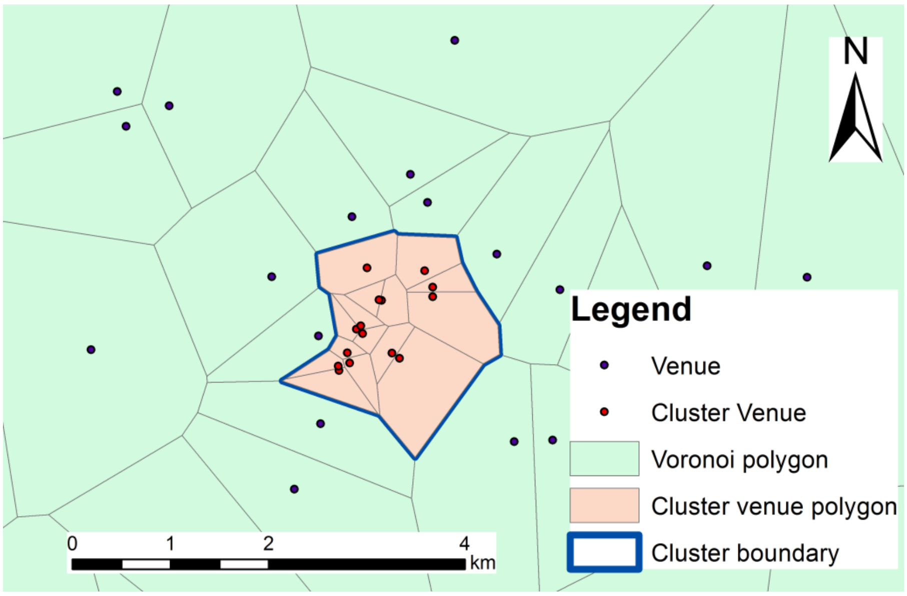


Figure 2: Voronoi polygons of venues and venue cluster

3.4 Validation

In this study, the final result of identification is an area (Voronoi polygon group) with a precise boundary. However, ground truth data is limited. Although some existing literature might report the approximate location of a city center, normally they do not offer a precise boundary of the city center. Conventionally, a landmark (e.g., central square, train station, etc.) is used to represent an approximate location of a city center. In this case, the identified city center can be characterized by an area with precise boundary while the actual one can only be characterized by a point (landmark). Since the precise boundaries of actual city centers are unavailable, we are not able to accurately estimate the accuracy of the boundary delineation.

Due to the limitation of ground truth, we will simply validate the identification. Specifically, the landmark, which is considered to be an approximate central location of a city center by the existing literature, is used to represent the central location of the ‘actual’ city center; while the landmark, which is matched with the cluster representing a city center, is used to represent the central location of the identified city center. If the landmark representing the central location of the actual city center is the same with the one representing the central location of the identified city center, the validation result will be true.

4. Experimental results and discussions

The proposed approach is applied to detect city centers using LBSN check-in data in three German cities such as Berlin, Munich and Cologne. This section demonstrates the experimental results and gives discussions about the detection of city centers.

4.1 Study case

4.1.1 Check-in data set

In this paper, the check-in dataset was collected from an LBSN called Gowalla, which is similar to Foursquare, by Cho et al. (2011). The positional accuracy of the data is from 10 to 15m. In this work, we chose three most important cultural and economic German cities (Berlin, Munich and Cologne) as our test beds. Within the administrative boundaries of these cities, there were approximately 31,000, 20,000 and 19,000 check-ins respectively, and Table 1 shows the user count and venue count in the three cities. The population (2012) and areas of the three cities are also listed in Table 1. The total check-in count for each city is basically proportional to the total population of the city.

Table 1: Statistical description of the sub-dataset for the three study cities

City	Population (millions)	Area (km^2)	Check-in count	Venue count	User count	Displacement count
Berlin	3.375	891.85	31185	2770	1611	4696
Munich	1.388	310.43	19636	1068	1037	2810
Cologne	1.024	405.15	18658	1626	1154	3368

4.1.2 Data pre-processing

The user displacements used in our study were daily displacements, meaning each displacement was composed of two consecutive check-ins made by the same user on the same day. Therefore, we need to filter noise in two situations. Situation 1: when a user generated more than one check-in at the same position, the earliest check-in was kept and the others were discarded. Situation 2: displacements with abnormal speed (for instance 250km/h within a city) will be discarded. Finally, we obtained 4696, 2810 and 3368 displacements (travel flows) for Berlin, Munich and Cologne respectively.

4.2 Distribution of travel flows

Figure 3 shows that the distributions of the travel flow counts among venues in the three cities follow a power law, this being: $P(x) \sim x^{-\alpha}$. This has been verified by a Kolmogorov-Smirnov (KS) test (for more detail see Clauset et al., 2009). In the log-log plot, the slope of the scatter equals the exponent coefficient of the cumulative distribution function (CDF), i.e., $-\alpha$. All slopes are negative in the plot, demonstrating the heterogeneity of the flow count among venues, namely, few of the venues have a relatively large number of flows while the vast majority of the venues have a small number of flows.

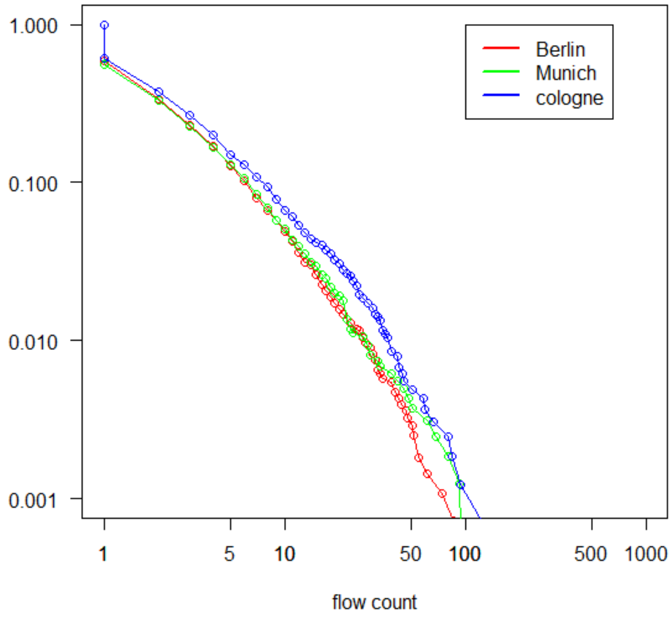


Figure 3: Distribution of travel flow counts among venues shown in a log-log plot. The y-axis represents the cumulative distribution function (CDF), i.e., $F(x)=P(x>=X)$.The CDF of the travel counts for venues in the three cities all follow a power law: $P(x) \sim x^{-\alpha}$. The slope of the scatter (i.e., the exponent coefficient $-\alpha$) is -1.41, -1.34 and -1.24 for Berlin, Munich and Cologne respectively.

4.3 City center identification

4.3.1 Cluster detection

1) Cluster detection using LGOG

a) Detecting hotspots of travel flow

The calculations of local Getis-Ord G_i^* in this paper were conducted using ESRI ArcMap 10.1. We chose the distance-based method to identify the neighbors of a given feature, so that if the distance between a feature (venue) *i* and *j* was less than the distance *threshold*, *j* was a neighbor of *i*; otherwise, it was not.

An appropriate value for neighbor distance *threshold* is vital. Here we chose 1km as the neighbor distance *threshold* on the experiences of city center size from some related studies (e.g., Borruso and Porceddu, 2009; Lüscher and Weibel, 2012). Thus, 1km was used as the neighbor distance *threshold* of the hotspot detection in the three cities.

b) Detecting clusters

We selected the statistically significant hotspots with Z-scores greater than 1.96. With a distance *threshold* of 1km, a few clusters were distinguished using the method based on 'spatial closeness' (see sub section 3.2). Because they are located in sparsely populated areas, only the clusters with relatively large total numbers of flows and densely populated positions are considered as candidates for city centers.

2) Cluster detection using DBSCAN

The calculations of DBSCAN in this paper were conducted using *R* software (http://www.r-project.org/). Since DBSCAN is a purely clustering algorithm rather than an abnormality detection method or a burst detection method, DBSCAN is not able to identify numeric features or attributes. Instead, DBSCAN only counts points. This means that a venue with a certain flow count should be transformed into a set of separate points with an identical position before being input into DBSCAN. For instance, a venue with a flow count of ten should be transformed into ten distinct points with the identical position (coordinates) with the venue. Each point corresponds to a destination or origin of a certain trip.

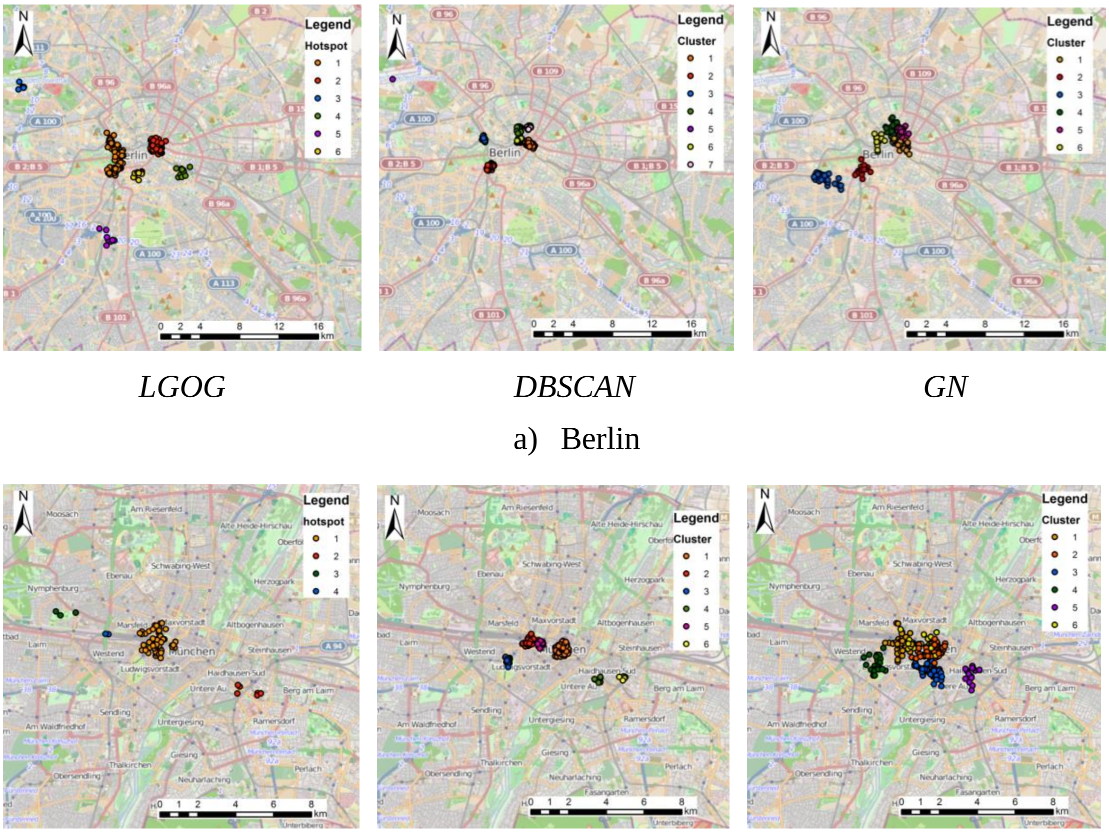
DBSCAN was run a couple of times with distinct pairs of parameters (*eps* and *minPts*). The result of DBSCAN corresponding to a pair of parameters, i.e., *eps* = 150m and *minPts* = 100, were chosen in this study since the centroids of the two largest clusters, i.e., the two clusters with the largest number of points (the largest flow counts), are relatively close to those identified by LGOG in Berlin.

3) Cluster detection using GN

The calculations of GN in this paper were conducted using *R* software as well. As noted in the sub section 3.2, long trips should be removed to generate a good result, i.e., venues that constitute an identical community should be spatially close to each other.

Similarly, GN was run a couple of times with distinct parameters (i.e., *maxLen*). The result of GN corresponding to a specific parameter, i.e., *maxLen* = 500m was chosen in this study since the centroids of the two largest clusters are relatively close to those identified by LGOG in Berlin.

Figure 4 maps the typical clusters (i.e., clusters with relatively large flow counts) detected by three methods in the three study cities.

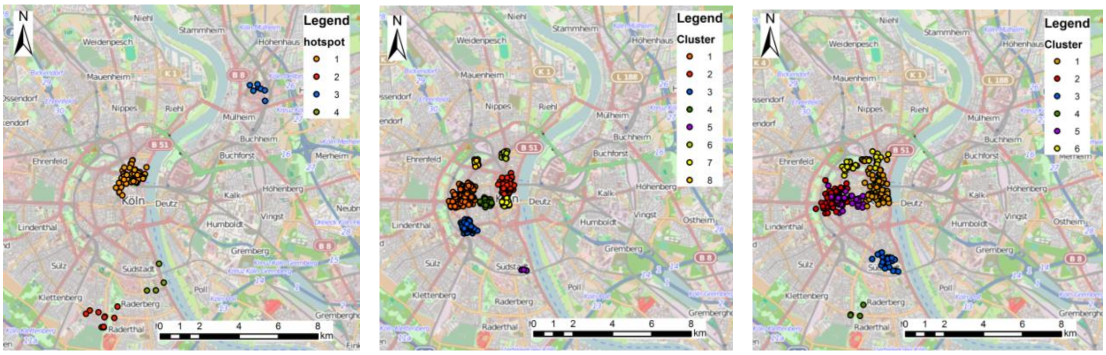


LGOG *DBSCAN* *GN*

a) Berlin

LGOG *DBSCAN* *GN*

b) Munich



LGOG *DBSCAN* *GN*

c) Cologne

Figure 4: Typical clusters detected by three different methods (Basemap: OpenStreetMap, contributors: CC-BY-SA)

4.3.2 City center identification

To be simple, it is assumed that there are no more than two city centers in all the three cities since 1) there are two city centers in Berlin as some literature report and 2) the other two cities that have smaller population and areas than Berlin are likely to have no more city centers than Berlin. Therefore, for the three methods, the two most typical clusters identified are considered as candidates for city centers with an assumption that there are no more than two city centers in each study city.

As introduced in sub section 3.3, three steps were taken to identify city centers.

1) Matching typical cluster to candidate landmark

The two most typical clusters identified in three cities were matched with candidate landmarks representing the central locations of potential city centers based on some socio-economic knowledge of study cities (e.g., economic structure, industrial distribution, history and culture, etc.). Table 2 list the candidate landmarks matched with the two most typical clusters detected by three different methods.

Table 2: Candidate landmarks matched with the two most typical clusters detected by three different methods

Method	Cluster	Matched candidate landmark		
		Berlin	Munich	Cologne
LGOG	1	Potsdamer Platz	Central railway station	Central railway station
	2	Alexanderplatz	Business park	Zollstock
DBSCAN	1	Alexanderplatz	Central railway station	Convention Bureau
	2	Potsdamer Platz	Marienplatz	Hahnen Gate
GN	1	Alexanderplatz	Central railway station	Convention Bureau
	2	Potsdamer Platz	Marienplatz	Hahnen Gate

2) Determining city center

First, the first most typical cluster, i.e., the one with the largest flow count, is considered as a city center. Second, to further determine if the second most typical cluster be a city center, we set the threshold for *Rate* (C_2) as 0.5. If *Rate* $(C_2) > 0.5$, the second most typical cluster will be considered to be another city center.

As a consequence, all three methods identified two city centers in Berlin. However, DBSCAN and GN both identified two city centers in Munich and Cologne; while LGOG identified only one. Table 3 lists the city centers detected by three methods and the flow counts of the clusters which are considered to be city centers. Note that, although the two city centers identified by three methods in Berlin are same, the ranking of city centers in the identification results using LGOG method is different from that using the other two methods. Potsdamer Platz is matched with the *city center 1* (the 1st largest city center) identified using LGOG method; whereas it is matched with the *city center 2* (the 2nd largest city center) identified using the other two methods.

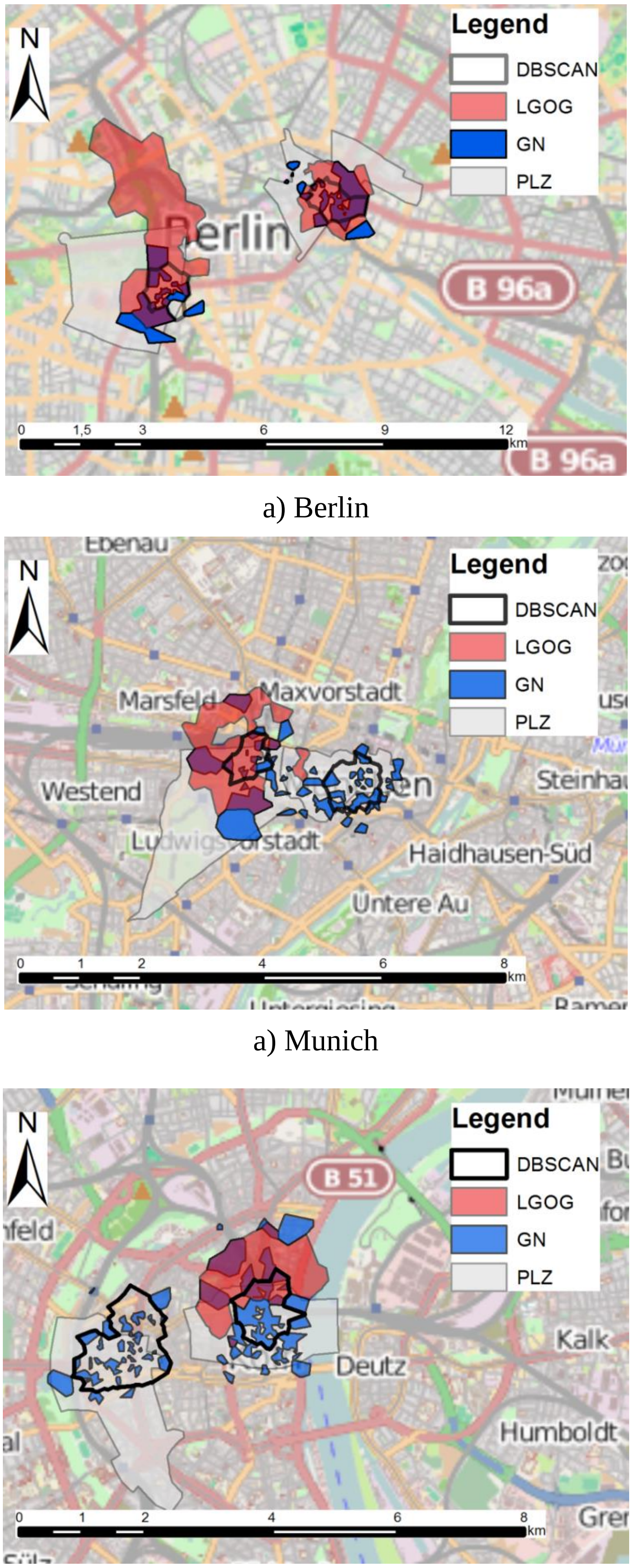
Table 3: The city centers detected by three methods

City	City center	Central location of city center (matched landmark)	Flow count of cluster		
			LGOG	DBSCAN	GN
Berlin	1	Alexanderplatz	425	352	349
	2	Potsdamer Platz	681	260	271
Munich	1	Central railway station	474	395	443
	2	Marienplatz*	-	302	346
Cologne	1	Convention Bureau	444	864	762
	2	Hahnen Gate*	-	554	481

Note: * means that the city center is not identified by LGOG method.

3) Delineating city center

In this part, city centers with precise boundaries will be delineated. Figure 5 maps the identified city centers with precise boundaries using different methods. In cities of Germany, compared to the district or borough data, postal code region (PLZ) data has a larger spatial scale and thus was used as the data source of comparative city center boundary in this study (http://arnulf.us/PLZ). Note that the precise boundaries of actual city centers in the three study cities are unavailable, the comparative boundaries of city centers used here can’t really represent the precise boundaries of actual city centers. A PLZ is a sub region with a unique postal code. Specifically, the PLZ where a certain candidate landmark is located was chosen to represent the administrative boundary of the corresponding city center. For instance, a specific PLZ where the landmark Alexanderplatz is located was used as the administratively comparative boundary of the *city center 1* in Berlin since Alexanderplatz corresponds to the *city center 1* in Berlin (see Table 3). The boundaries of city centers delineated don’t well overlap the administratively comparative boundaries, implying that human activities seems to be not large influenced by the administrative divisions of a city.



a) Berlin

a) Munich

c) Cologne

Figure 5: The identified city centers using three methods (Basemap: OpenStreetMap, contributors: CC-BY-SA).

4.3.3 Validation of the results

Berlin has two central business districts (CBDs) that are located in Potsdamer Platz and Alexanderplatz according to some literature (e.g., Gornig and Häussermann, 2002; Van Criekingen, et al. 2007). Actually, city center has a more broad definition than CBD that is center of economy. Normally, city center could be an economical center, political center or cultural center of a city; while CBD is only an economical center.

Due to the limitation of ground truth, the two CBDs in Berlin are used to represent two actual city centers in this study. The approximate central locations of city centers in Munich and Cologne are not clearly reported in the existing English publications. Therefore, in this study there is no ground truth data for city centers in Munich and Cologne. In this study, the landmark, which is used as an approximate central location of a CBD, is used to represent the central location of an actual city center. While the candidate landmark, which is matched with an identified city center, is used to represent the central location of the identified city center. In Berlin, the two landmarks representing the central locations of two actual city centers are the same with the two landmarks representing central locations of two identified city centers. This means that the two city centers in Berlin are successfully identified. In empirical study, all the three methods successfully identified the two city centers in Berlin. This well proves the validity of using LBSN data for city center identification.

4.4 Discussion

In this part, advantages and limitations of the approach in this study are firstly discussed. On the one hand, compared to the majority of existing approaches, this approach 1) is able to identify city center with a precise boundary and 2) is more flexible with cities of different structures, particularly cities that are likely to be polycentric. One the other hand, since cities are complex and distinct, feasibility of the approach in a specific city is influenced by the knowledge of this city. More knowledge about socio-economic characteristics and spatial range of city helps yielding a better identification result. First of all, the identification relies much on spatial scale. Similar to other identification methods (e.g., kernel density estimation), the three methods in this study are influenced by the parameters that control the range of spatial relationships between venues. An appropriate range of spatial relationship between venues needs some knowledge of city center (e.g., an approximate spatial size of city center). Secondly, knowledge of socio-economic characteristics (e.g., economic structure, industrial distribution, history and culture, etc.) is necessarily required to select appropriate candidate landmarks used to be further matched with typical clusters. For instance, in a historical or tourism city, city center is likely to be located in the inner city or ‘old city’, and thus central plaza or city hall are more considered to be candidate landmarks; whereas in an industrial or financial city, city center is likely to be located in the CBD, and thus skyscrapers are more considered to be candidate landmarks. Moreover, empirical result demonstrates the validity of using LBSN data for identifying city center in spite of data bias. This suggests that data bias does not large influence the validity of using LBSN for city center identification. Thus, empirical result demonstrates that LBSN data has a large potential and usefulness for identifying city center.

Finally, we further compare the three methods. Table 4 lists some characteristics of three methods. Apart from the type, other characteristics are also discussed here. On the one hand, DBSCAN needs two parameters whereas other two methods need only one. Thus, DBSCAN is relatively difficult to choose appropriate parameters since there are two parameters required being determined simultaneously. On the other hand, boundaries delineated by DBSCAN are more geometrically regular than those by the other two methods (see Figure 5). Moreover, some hints might be useful to city center identification in the future. Compared to DBSCAN and GN, LGOG seems to be likely to identify few city centers (see Table 3). In other words, seemingly LGOG is sensitive to the first largest city center whereas DBSCAN and GN are sensitive to the second largest city center. This suggests that intuitively LGOG might be suitable for city center identification in a monocentric city whereas DBSCAN and GN might be in a polycentric city.

Table 4: Comparisons of the three identification methods

Method	LGOG	DBSCAN	GN
Method type	Clustering method for either high values or low values	Density-based clustering method	Connectivity-based method
Number of parameters	One	Two	One
Geometric regularity of boundary	Compact	Very compact	Disaggregated
Suggested urban structure	Monocentric	Polycentric	Polycentric

5 Conclusion and future work

The experiments in this work show that city centers and their boundaries can be detected with high accuracy by using LBSN data. Three different methods have been used and compared with each other regarding identification of city centers. In overall, they have different advantages and disadvantages, and therefore seem to be suitable for cities of different urban structure. DBSCAN method can delineate more geometrically regular boundaries but have more parameters required being determined than LGOG and GN methods. And LGOG might be suitable for city center identification in a monocentric city whereas DBSCAN and GN might be in a polycentric city.

Compared to other mobility data sources (such as traffic flows and mobile phone records), LBSN check-in data have some advantages. First, user-generated check-ins are available for free in terms of some downloading tools (e.g., API). Second, check-in data has a larger spatial scale than traffic flows and mobile phone records, since the position of a check-in can be represented by the position of a venue. For example, the position of geo-referenced check-in is at the street level, whereas the position of traffic flow data is more likely to be at the census tract level.

Check-in data has some limitations. There is a representativeness issue, e.g., the bias of age group and bias of place category when check-in data is used to represent human mobility. Normally, young users contribute the vast majority of check-ins. In addition, users check in more at some categories of venues (e.g., airport, restaurant, shop, railway station, etc.) than other categories of venues. Compared to the categories of venues like restaurant, shop or workplace, home venues attract relatively few check-ins, because the majority of users do not often check in at home venues (e.g., private houses and apartments). The heterogeneity in the popularities of LBSN venue categories is somewhat consistent with the heterogeneity in the contributions of venue categories to the identification of city center, seemingly reducing the influence of LBSN data bias on the city center identification. However, the bias of place category still has a negative effect on delineating a precise city center boundary close to the actual one. Such negative effect could somewhat decrease with an increasing volume of check-ins and active users. Apart from the representative issue, there is a ‘data sparseness’ problem with check-in data in sparsely populated areas since the time period of the data set in this study is 2009-2010 when the social media is not as popular as it is currently. Even at present, the data density is still not high in some countries or regions (e.g., developing countries). However, the ‘data sparseness’ issue could not large influence city center identification since city center is normally located in densely populated areas, but could influence some other studies, such as an investigation of the commuting patterns of inhabitants.

Regarding validation of the identification result, firstly, there is not much ground truth data of city center with a precise boundary since urban planners and economic geographers normally delineate a city center with a vague boundary or just report an approximate central location of city center. For urban planners and economic geographers, a city center with a relatively precise boundary is more useful than that with only an approximate location. Secondly, the majority of the approaches rely much on spatial scale. In other words, the results of identification methods (e.g., kernel density methods, clustering methods, etc.) are much influenced by their parameters that control the range of spatial relationship. Thus, how to find out the optimum parameters when there is little knowledge of the study cite is a vital issue.

Some other aspects should be considered in future research. In order to get a better understanding of urban structure and urban dynamics, further analysis of interactions between different urban centers in a city should be included, once a large volume data set of LBSN is acquired. Moreover, it might be more interesting if the category of the venue (e.g., restaurant, office or shop) was to be taken into consideration.

References:

Alonso, W. (1960). A theory of the urban land market. *Papers and Proceedings Regional Science Association*, 6, 149–157.

Alonso, W. (1964). *Location and Land Use: Toward a General Theory of Land Rent*. Cambridge, MA, USA, Harvard University Press.

Anas, A., Arnott, R. and Small, K. (1998). Urban spatial structure. *Journal of Economic Literature*, 36, 1426–1464.

Balcan, D., Hu, H., Goncalves, B., Bajaradi, P., Poletto, C., Ramasco, J.J., Paolotti, J. J. D., Perra, N., Tizzoni, M., Van den Broeck, W., Colizza, V. and Vespignani, A. (2009). Seasonal transmission potential and activity peaks of the new influenza a (h1n1): a monte carlo likelihood analysis based on human mobility. *BMC Medicine*, 7, 45.

Bao, J., Zheng, Y., and Mokbel., M.F. (2012). Location-based and Preference-Aware Recommendation Using Sparse Geo-Social Networking Data. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, Redondo Beach, California, USA, November 6-9, 2012, pp. 199-208.

Borruso, G. and Porceddu, A., (2009). A tale of two cities: Density analysis of CBD on two midsize urban areas in northeastern Italy. *In*: Murgante B., Borruso G. and Lapucci A. (Eds.) *“Geocomputation and Urban Planning” - Series: Studies in Computational Intelligence*, 176, Berlin Heidelberg, Springer-Verlag, pp. 37-56.

Clauset, A., Shalizi, C.R. and Newman, M.E.J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661–703.

Carol, H. (1960). The hierarchy of central functions within the city. *Annals of the Association of American Geographers*, 50, 419-438.

Cheng, Z., Caverlee, J., Lee, K. and Sui, D. (2011). Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, Barcelona, Spain, July 17-21, 2011, pp. 81-88.

Cho, E., Myers, S. A. and Leskovec, J. (2011). 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, August21-24, 2011, pp. 1082-1090.

Chowell, G., Hyman, J.M., Eubank, S. and Castillo-Chavez, C. (2003). Scaling laws for the movement of people between locations in a large city. *Physical Review E*, 68(6), 066102.

Clark, C. (1951). Urban Population Densities. *Journal of the Royal Statistical Society. Series A*, 114, 490–496.

Eubank, S, Guclu, H, Kumar, V.S.A., Marathe, M., Srinivasan, A., Toroczkai, Z. and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429, 180–184.

Getis, A., and Ord. J.K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24 (3), 189–206.

González, M.C., Hidalgo, C. and Barabási, A.L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196), 779–782.

Gornig, M. and Häussermann, H. (2002). Berlin: Economic and Spatial Change. *European Urban and Regional Studies*, 9 (4), 331–41.

Hollenstein, L. and Purves, R. S. (2010). Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, 1(1), 21-48.

Jiang, S., Ferreira J. and Gonzalez, M. (2012). Discovering urban spatial-temporal structure from human activity patterns. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, August 12, 2012, Beijing, China, pp.95-102.

Kloosterman, R. and Musterd, S. (2001). The polycentric urban region: towards a research agenda. *Urban Studies*, 38: 623–633.

Liang, X, Zhao, J, Dong, L, and Xu, K (2013) Unraveling the origin of exponential law in intra-urban human mobility. *Scientific Reports*, 3, 2983.

Liu, Y., Sui, Z., Kang, C. and Gao, Y. (2014). Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data. *PLoS ONE*, 9(1): e86026.

Lüscher, P. And Weibel, R. (2012). Exploiting empirical knowledge for automatic delineation of city centers from large-scale topographic databases. *Computers, Environment and Urban Systems*, 37, 18-34.

Montello, D.R., Goodchild, M.F., Gottsegen, J. and Fohl, P. (2003). Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. *Spatial Cognition & Computation: An Interdisciplinary Journal*, 3:2-3, 185-204.

Murphy, R.E. and Vance, J.E. (1954). Delimiting the CBD. *Economic Geography*, 30, 189-222

Murphy R.E. (1972). *The Central Business District: A Study in Urban Geography*. London, Longman.

Noulas, A., Scellato, S., Mascolo, C. and Pontil, M. (2011). An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proceedings of Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, July 17–21,pp. 570-573.

Ord, J.K. and Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis,* 27(4), 286 – 306.

O'Sullivan, D. and Unwin, D.J. (2010). *Geographic Information Analysis (2nd Edition)*. John Wiley & Sons.

Ratti, C., Pulselli, R.M., Williams, S. and Frenchman, D. (2006). Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5), 727 – 748.

Roth, C., Kang, S.M., Batty, M. and Barthélemy, M. (2011). Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. *PLoS ONE*, 6(1): e15923.

Scellato, S., Noulas, A. and Mascolo, C. (2011). Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Diego, CA, USA, August21-24, 2011, pp. 1046-1054.

Taubenboeck, H., Klotz, M., Wurm, M., Schmieder, J., Wagner, B., Wooster, M., Esch, T. and Dech, S. (2013). Delineation of Central Business Districts in mega city regions using remotely sensed data. *Remote Sensing of Environment*, 136, 386-401.

Thurstain-Goodwin, M. and Unwin, D. (2000). Defining and Delineating the Central Areas of Towns for Statistical Monitoring using Continuous Surface Representations. *Transactions in GIS*, 4 (4), 305-318.

Van Criekingen, M., Bachmann, M. and Lennert, C.G.M. (2007). Towards polycentric cities. An investigation into the restructuration of intra-metropolitan spatial configurations in Europe. *Belgeo*, 1, 15-30.

Wang, P., González, M., Hidalgo, C. and Barabási, A.L. (2009). Understanding the spreading patterns of mobile phone viruses. *Science,* 324, 1071–1075.

Wei, L.Y., Zheng, Y., and Peng, W.C. (2012). Constructing popular routes from uncertain trajectories. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Beijing, China, August 12-16, 2012, pp. 195-203.