International Journal of Geo-Information

MDPI

Article

A Simple Line Clustering Method for Spatial Analysis with Origin-Destination Data and Its Application to Bike-Sharing Movement Data

Biao He [1,2], Yan Zhang [2], Yu Chen [2] and Zhihui Gu [2,3,*]

1 Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources, Shenzhen 518034, China; whu_hebiao@hotmail.com
2 College of Architecture and Urban Planning, Shenzhen University, Shenzhen 518060, China; zhyan@szu.edu.cn (Y.Z.); szuchenyu@szu.edu.cn (Y.C.)
3 Shenzhen Key Laboratory for Optimizing Design of Built Environment, Shenzhen 518060, China
* Correspondence: gzh@szu.edu.cn; Tel.: +86-755-2673-2869

Received: 26 April 2018; Accepted: 27 May 2018; Published: 29 May 2018

Abstract: Clustering methods are popular tools for pattern recognition in spatial databases. Existing clustering methods have mainly focused on the matching and clustering of complex trajectories. Few studies have paid attention to clustering origin-destination (OD) trips and discovering strong spatial linkages via OD lines, which is useful in many areas such as transportation, urban planning, and migration studies. In this paper, we present a new Simple Line Clustering Method (SLCM) that was designed to discover the strongest spatial linkage by searching for neighboring lines for every OD trip within a certain radius. This method adopts entropy theory and the probability distribution function for parameter selection to ensure significant clustering results. We demonstrate this method using bike-sharing location data in a metropolitan city. Results show that (1) the SLCM was significantly effective in discovering clusters at different scales, (2) results with the SLCM analysis confirmed known structures and discovered unknown structures, and (3) this approach can also be applied to other OD data to facilitate pattern extraction and structure understanding.

Keywords: clustering method; spatial linkage; origin-destination trips; bike-sharing movement

1. Introduction

Origin-destination (OD) data are a special type of trajectory data that concern origin and destination locations but ignore the actual trajectory route. OD matrices represent one of the most important sources of information used for strategic planning and the management of transportation networks [1]. Traditionally, urban planning and transportation engineering rely on household questionnaires or census and road surveys that are conducted every 5–10 years to develop methodologies for OD matrix estimations. Recent improvements in Big Data and tracking facilities have made it possible to collect a large amount of travel data for moving objects. However, previous studies of OD matrices, which have been based on point statistics over administrative or traffic spatial units, quickly become illegible as the data size increases due to the massive intersections and overlapping of OD flows [2]. Automatic algorithms that are able to extract useful information from these sources have consequently acquired great interest [3].

There are two basic types of clustering algorithms related to OD data: point clustering of origin or destination points and spatial clustering of actual trajectories. However, few studies have focused on the clustering of OD lines directly. If traffic analysis is the main research subject, the actual

trajectory method becomes highly important, and it would be necessary to track the location of each movement at all times. However, if the spatial linkage of a special interest is the concern, for instance, the strongest linkage between a working place and a residential center, then OD trips are worth studying. The clustering of OD data could help us understand the strongest connections between locations and their spatial characteristics.

In this article, we focused on clustering methods for OD lines. We proposed a new approach for the analysis of OD movements, which can discover spatial linkage patterns, such as location characteristics and spatiotemporal trends in movements. This simple line clustering method (SLCM) aggregates OD lines into small spatial clusters of sufficient size to reveal the spatial characteristics in terms of movement. By adopting the entropy theory and the probability distribution function, this method requires minimal domain knowledge to determine input parameters. The backward clustering process ensures statistically significant clustering. The approach was proven to be effective and scalable for large datasets through a case study using the Mobike bike-sharing OD trip datasets collected in the Nanshan District of Shenzhen, China.

2. Related Research

2.1. Point Clustering Method for OD Data

The point clustering method for OD data was derived from the traditional OD matrix analysis. This method mainly generates OD flow by counting the numbers of origin and destination points in a spatial region of interest (e.g., administration areas or traffic-affected zones), which can be defined subjectively by users or derived from the data [3–8]. For the latter definition, the main option is to use the density-based grouping method to cluster geographically closed points to obtain the region of interest. For example, Guo et al. presented an approach to group spatial points into clusters, derive statistical summaries, and visualize spatiotemporal mobility patterns [9]. In addition, they presented a flow-based density estimation method and a flow selection method to normalize and smooth flows with a controlled neighborhood size and detect high-level patterns in the data [2,10]. Mao et al. presented another novel approach for the spatial clustering of OD pairs based on traffic grid partitioning to discover spatiotemporal patterns in urban commuting and the job–housing balance. This method can create clusters using traffic grids of different sizes to adapt to different density surfaces and identify threshold values from the statistics of OD clusters to extract urban job-housing structures [11].

However, a major limitation of these methods is the fixed boundary constraint for the region of interest. Whether it is defined subjectively by users or derived from the data, once the boundary of the region of interest is fixed, its spatial heterogeneity is ignored. For example, two similar and adjacent OD trips would be regarded as completely different behaviors just because their endpoints are clustered into two different regions. Therefore, we proposed a new clustering method to identify the spatial linkage without fixed boundary.

2.2. Trajectory Clustering Methods

Another research method related to OD line clustering is the trajectory clustering method. Existing works can be classified into two groups: partitioning and hierarchical clustering. For instance, Traclus, a well-known partition-and-group algorithm for clustering trajectories, partitions a trajectory into a set of line segments and groups similar line segments into a cluster [12]. Cao et al. proposed an algorithm to segment a long trajectory into sub segments by user-specified period [13]. Ferrero et al. extended the concept of time series *shapelets* to discover relevant subtrajectories [14]. The method for calculating the similarity or distance of a line segment in these methods can offer valuable insight. For example, Chen et al. defined a distance function with three components: (i) the perpendicular distance, (ii) the parallel distance, and (iii) the angular distance. These components were adapted from similar measures that were used for pattern recognition [15]. Some researchers have adopted this

definition and made some improvements [12,16–19]. However, the physical meaning of this definition is not very clear, especially in the weight setting section.

On the other hand, trying to incorporate the complexity of real trajectories, hierarchical clustering algorithms build models by introducing global or local variables, such as the speed, duration, curvature, and other descriptors of trajectories [20–24]. Mohammad et al. showed the extraction of new point features: bearing rate, the rate of change of the bearing rate and the global and local trajectory features, like medians and percentiles, enables many classifiers to achieve high accuracy (96.5%) and f1 (96.3%) scores [25]. Moreno et al. considered machine learning and context information to enrich trajectory data [26]. These methods are relatively complex and effective when it is necessary to consider the actual trajectory of movement, but for the clustering tasks that are only interested in the OD of trajectory, the method is not applicable. However, if we need to further consider the non-spatial attributes of the OD, such as the characteristics of the moving objects, land-use of OD, etc., the hierarchical clustering method can be of great significance.

3. SLCM Clustering Method

3.1. The Definitions

The key idea of the SLCM is finding the centerlines via sufficient neighboring lines from all OD lines. The following definitions and parameters were used in this method.

Definition: centerline and its neighboring lines. Let L be the database of lines, O be the origin points, D be the destination points, and Li be the line connecting the ith point of O and D. If Lj, a line connecting Oj and Dj, fell within the searching radius of Oi and Di, then Lj is defined as a neighboring line of the centerline Li (Figure 1). A centerline can have more than one neighboring line and the number of its neighboring lines, denoted by $Nls(Li)$, is defined as:

$$Nls(L_i) = \left\{ L_j \in L \mid dist(O_i, O_j) \leq Dr \cap dist(D_i, D_j) \leq Dr \right\} \quad (1)$$

where Dr represents the searching radius, and $dist(Oi, Oj)$ and $dist(Di, Dj)$ represent the Euclidean distances of two endpoints.

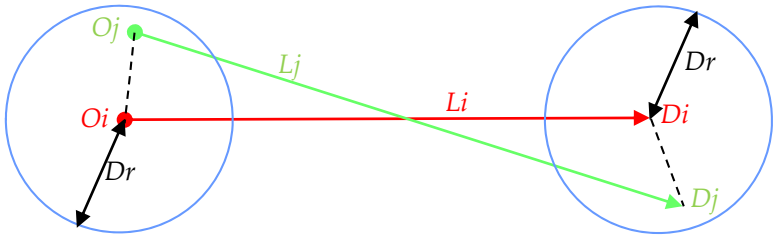

Figure 1. Defining neighboring line of a centerline.

Parameter 1: searching radius Dr. Since the key idea of the SLCM is to find similar lines from the OD database within a given radius, the first parameter that needs to be defined is the searching radius. This parameter will directly determine the shape of the clustering results. In general, the larger the searching radius, the more neighboring lines can be found. However, there is a special case that needs attention. When line Lj is too short (e.g., shorter than 2 times of Dr), meeting definition 1 does not guarantee spatial resemblance because Lj could be in a different or even the opposite direction of Li (i.e., green lines in Figure 2).

Therefore, we added a limitation parameter for Li to be defined as a center line for which the length of Li must be greater than $2Dr/\sin 45^\circ$ ($\approx 2.83Dr$), to ensure that the neighboring lines not only are geographically close to Li, but also in line with the direction of Li. This threshold guarantees an angle less than 45° between a centerline and its neighboring line, and the change in length will not

change by more than 2 times *Dr*. Notice that this definition excludes all OD lines shorter than 0.83*Dr* in the clustering analysis. An angle less than 45° can be used if stricter results are desired or the direction is of greater concern. This definition is elaborated on in Figure 3.

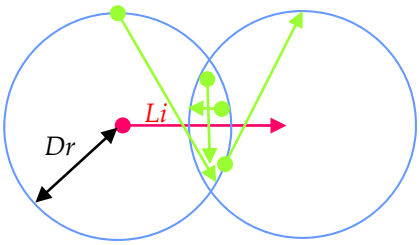

Figure 2. Special cases of definition 1 (the dots represent O, and the arrows represent D).

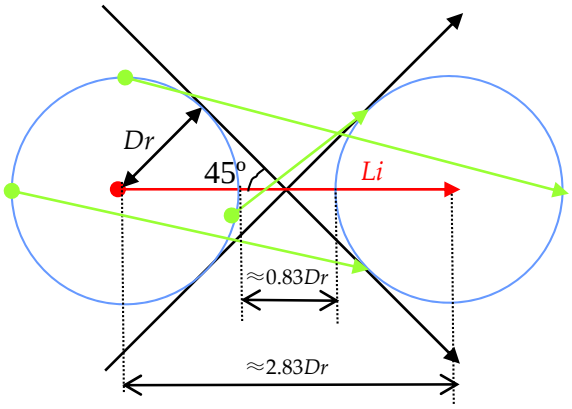

Figure 3. The length constraint of *Li* (the dots represent O, and the arrows represent D).

Parameter 2: the length limitation of centerline *Lm*, which is derived from the searching radius and is approximately equal to 2.83*Dr*. *Li* can be used as a centerline to calculate the neighboring lines in the following step only if the length of *Li* is longer than 2.83*Dr*; otherwise, *Nls(Li)* is set to 1. Obviously, this limitation will not allow lines shorter than 0.83*Dr* to participate in the final clustering. A large search radius will result in more information loss.

Parameter 3: the minimum number of neighboring lines, *Minlines*. In general, a centerline is considered more representative if it has more neighboring lines. This parameter excludes all centerlines with a number of neighboring lines less than the threshold *Minlines* that are not considered representative in cluster analysis.

In summary, the principle of SLCM is not complicated and can be easily applied with only two parameters pre-defined: the searching radius (*Dr*) and the minimum number of neighborhood lines (*Minlines*). Notice that: (1) large *Dr* may lead to information loss, such that OD lines shorter than 0.83*Dr* will be excluded in the clustering analysis, and (2) the definition of *Minlines* lacks statistical meaning. Determining values of the two parameters to avoid these limitations is critical and the following strategy is proposed.

3.2. Determining the Parameters

In general, with sufficient knowledge of the OD data source and a clear study objective, *Dr* and *Minlines* can be directly specified subjectively. For example, if we want to find the strongest spatial connection in the job-house OD data to help design a public bus line and stations with an impacted area of 500 m, we can set the *Dr* as 500 m and the *Minlines* to be the minimum demand of the bus line. However, experience-based parameters may not always be optimal. Therefore, in the absence of

prior knowledge, we recommended the following method developed based on the entropy theory and distribution probability.

In information theory, entropy is related to the amount of uncertainty for an event associated with a given probability distribution [27]. If all the outcomes are equally likely, then the entropy should be maximized. In the worst clustering scenario, *Nls* tends to be uniform for a *Dr* that is too small, and *Nls* becomes 1 for almost all lines; for a *Dr* that is too large, it becomes the total number of lines for almost all lines. Thus, the entropy becomes maximized. In contrast, in a good clustering scenario, the entropy of all *Nls* tends to be skewed. Thus, we used the entropy definition in Formula (2) and (3) and find the optimal value of *Dr* that minimizes *H(L)*.

$$\mathrm{H(L)} = -\sum_{i=1}^{n} p(L_i) * log_2 p(L_i) \tag{2}$$

$$p(L_i) = Nls(L_i) / \sum_{i=1}^{n} Nls(L_i) \tag{3}$$

Certainly, the initial *Dr* begins with a small value and gradually increases to calculate the entropy of all *Nls*. This heuristic method provides a reasonable range where the optimal value is likely to reside.

Another parameter, *Minlines*, is calculated from probability distribution functions in order to ensure its statistical significance. We first tested if *Nls* follows a normal distribution, which can be used to obtain all spatial clustering with higher confidence. As shown in Figure 4, any data with distribution similar to normal distribution can be transferred into a standard normal distribution with z-scores and *p*-value [28]. Z-scores are standard deviations. The *p*-value is the probability that the observed spatial pattern was created from a random process. A small *p*-value means it is very unlikely (with a small probability) that the observed spatial pattern is a result of a random processes. With $p < 0.01$ as the significance level, *Minlines* can be calculated as:

$$Minlines = \mathrm{average}(Nls) + 2.58 * \mathrm{SD}(Nls) \tag{4}$$

where SD is the standard deviation.

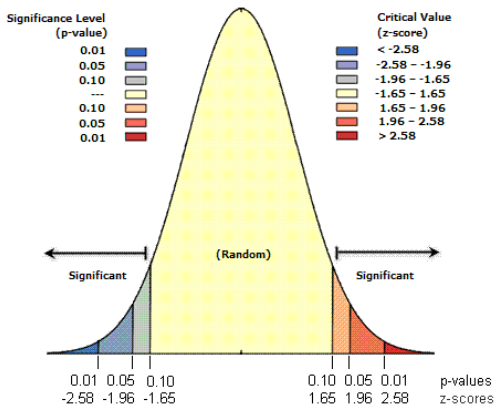

Figure 4. The *p*-value and z-scores of Standard normal distribution (from ArcGIS help [28]).

However, studies have also shown that many human activities follow a power law distribution instead of a normal distribution. For example, the distributions of a wide variety of physical, biological, and man-made phenomena were found approximately following a power law over a wide range of magnitudes [29]. Therefore, if the OD data does not meet the normal distribution, a power law distribution test is recommended. In general, most of the OD lines in space are discrete, and only a

few are clustered together. Thus, we suggest testing the OD dataset with Pareto (type-I) distribution, and to calculate the cumulative distribution probability (CDP) using Formula (5) [30].

$$p(x > x_m) = 1 - (\frac{x_m}{x})^{\alpha} \quad (5)$$

where x_m represents the minimum possible value of x, and α is a positive parameter which can be obtained from the power law mode:

$$f(x) = c * x^{-(\alpha+1)} \quad (6)$$

where c and α are regression coefficients. The Pareto type-I distribution is characterized by scale parameter x_m and shape parameter α, which is known as the tail index (Figure 5).

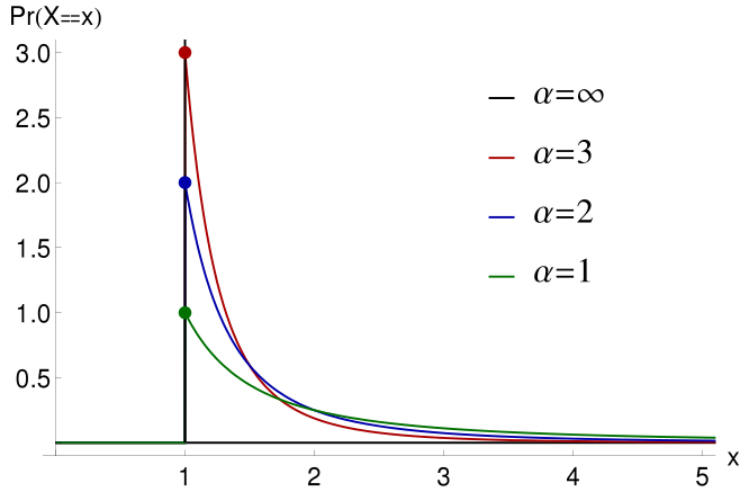

Figure 5. Pareto type-I probability density functions for various α with x_m = 1 (from Wikipedia [31]).

To find the strong spatial connections that are statistically significant, we set *Minlines* as 95% and 99% of CDPs (i.e., $p < 0.05$ and $p < 0.01$). In general, when the radius is very small, *Nls* values would approach 1 in most cases, and α would be a large value. With increasing radius, the number of *Nls* with a value of 1 decreases, and α decreases. However, when α is less than 1, the expected value of a random variable following the Pareto distribution is ∞, where the tail of the distribution has an infinite area, and the probability density function becomes meaningless [32]. Therefore, when α is less than 1, *Minlines* is set as the null, and no centerline will be considered.

Of course, other distribution tests can be performed. The key is finding the suitable probability distribution to help extract spatial clustering with high significance level.

3.3. Clustering Process Flowchart

Overall, our classification method is simple in principle and the parameter initialization is adaptable. Once the optimal values of *Dr* and *Minlines* are set either subjectively or derived from data, the clustering process can be implemented following the flowchart in Figure 6.

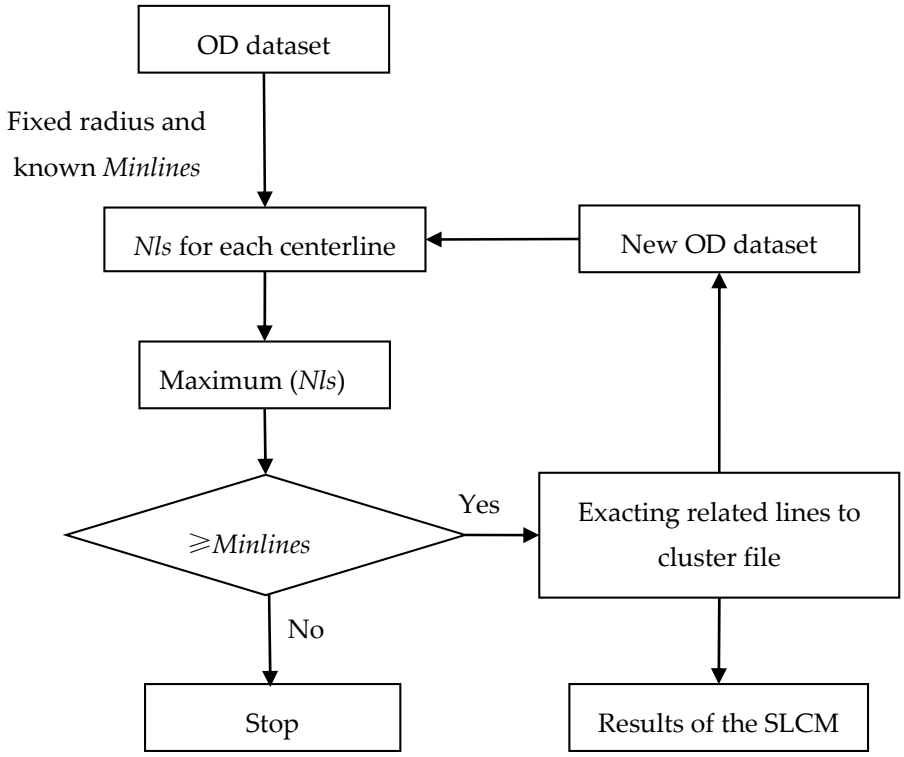

Figure 6. The overall process of the SLCM with fixed radius.

However, this simple clustering method still cannot solve all the problems mentioned above. Lines shorter than the specified value (i.e., 0.83*Dr*) will be excluded in clustering, which results in loss of information. Therefore, we propose another more flexible clustering procedure to avoid this drawback. The complex version of the SLCM clustering process consists of two parts: determining the parameters forward and searching the clusters backward, as shown in Figure 7. In this complex scenario, the optimal search radius is not used as the only search radius, but as the maximum search radius in the following step. Clustering backward means that clustering is started with the optimal search radius and ends with the minimum radius. With the optimal radius, we first search for the centerline of the maximum *Nls*. If the maximum *Nls* is greater than the *Minlines*, we extract the centerline and their neighboring lines into the cluster file and mark them. The remaining lines are then recalculated to determine the next centerline that satisfies the conditions. When no qualifying centerline appears, we moved on to the next smaller search radius and repeated this clustering process until the minimum search radius was reached. This backward clustering process can preserve spatially connected clustering with as much significance as possible under different search radii.

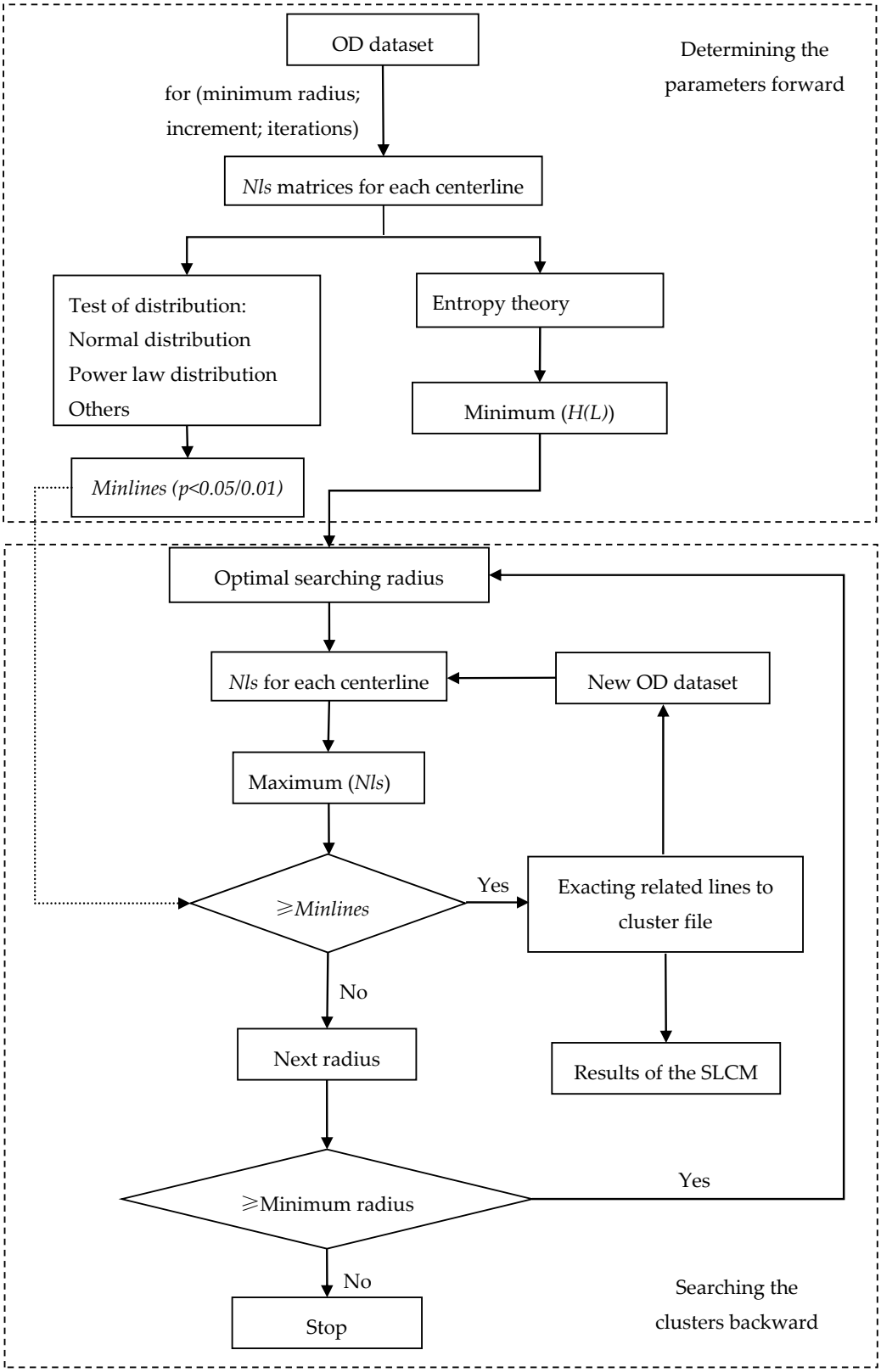

Figure 7. The overall process of the SLCM with flexible radius.

4. Case Study

To evaluate the effectiveness of the proposed approach, a case study with Mobike bike-sharing movement data was carried out in this paper. Our study area is the Nanshan District of Shenzhen, China. Shenzhen, the youngest megacity in China, was founded only 40 years ago. By the end of 2016, the city had 11.9 million people in an area of 1997.27 km^2. As one of the main downtown areas, Nanshan District has a permanent population of 1.87 million people and a density of 7235 person/km^2 [33].

Founded in 2014, Mobike is one of the fastest-growing bike-sharing companies in China. With services available in 13 cities, it is known for its distinct orange-hued and GPS-equipped bikes. Mobike provides an easy-to-use application and an efficient geo-local-station system that attracts a large number of users. By the end of 2016, the number of China's monthly active users (MAU) for bike-sharing had reached 4.32 million. Mobike achieved 72.5% of the market share, accounting for more than 3.13 million MAU [34]. These data also provide the possibility of collecting a large amount of information on citizens' mobility. Using high-frequency scanning, the positions of available bikes can be obtained at every minute and the bike-sharing OD lines can be rebuilt by tracking the locations of every bike.

According to a report, the number of bikes in the Mobike bike-sharing program in Shenzhen exceeded 200,000 in April 2017 [35]. We found 42,901 bikes in the Nanshan District and 68,883 OD trips on 9 June 2017, with an average straight-line length of 1.25 km and a minimum length of 300 m. At the morning peak time (MPT; 6–9 am), there were 15,458 OD trips, with an average length of 1.26 km and an average travel time of 22 min. The study area and the Mobike OD trips at MPT are shown in Figure 8.

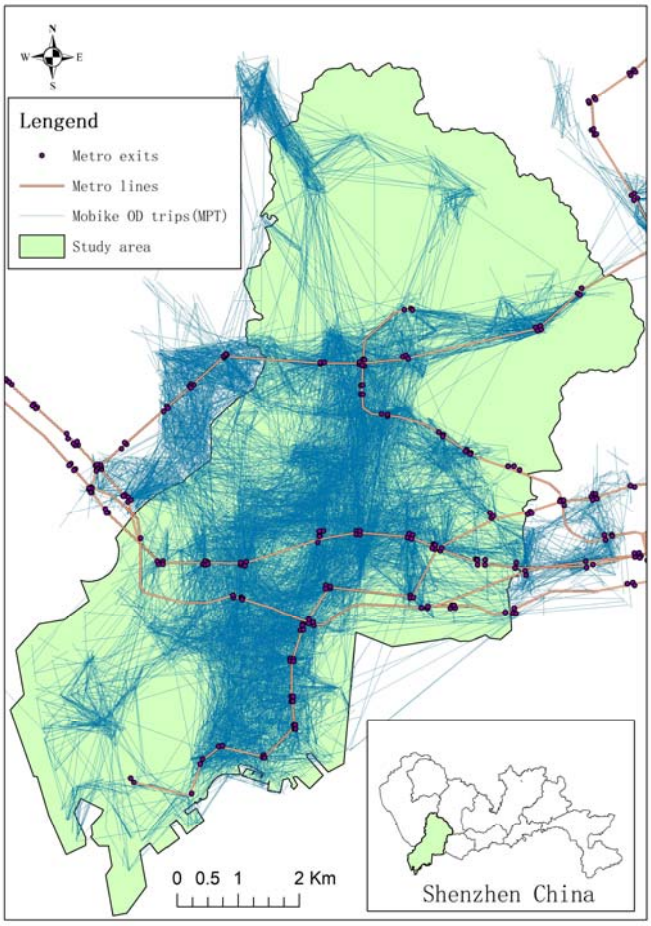

Figure 8. Study area and Mobike OD trips at MPT.

OD data at MPT was used in this study to reveal the spatial patterns of bike movement. We set the initial *Dr* value as 50 m and increased it by 50 m each time. We ran the process for 50 rounds to discover the optimal parameters of *Dr* and *Minlines* and obtain the clustering results for the optimal parameters. The analysis progress and results are shown in the following section.

4.1. The Parameter Determination

First, we calculated the entropy of all *Nls* values with different *Dr* values. We noted that with increasing *Dr*, the entropy of *Nls* decreased first and then increased. This result is consistent with our initial prediction. The real minimum *H(L)* is achieved at round 30 (*Dr* = 1500 m). However, at round 5, with *Dr* of 250 m, the entropy of all *Nls* values began steady as shown in Figure 9.

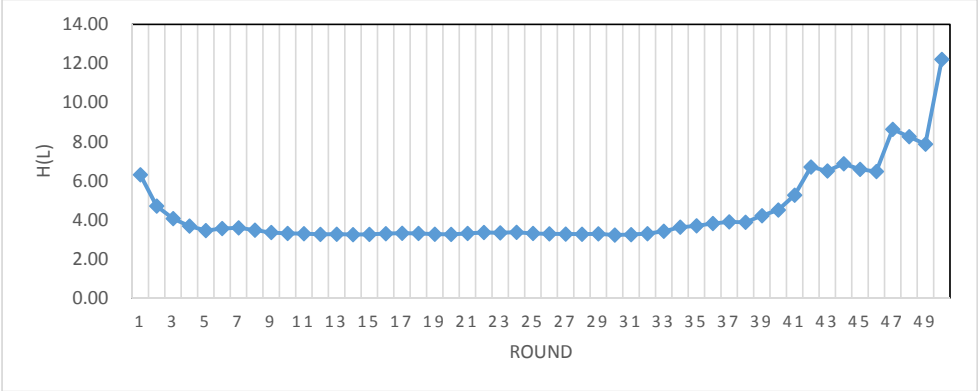

Figure 9. Entropy of all *Nls* values with different radiuses from round 1 to round 50.

Furthermore, we tested the distribution of all *Nls* values with normal distribution and power law distribution for each round. As shown in Figure 10, the goodness of fit with power law mode is very high, which means that Pareto probability distribution function would be more suitable for determining the value of *Minlines*. Then, we calculated the CDPs for each round, which are listed in Table 1. After round 13, the values of α were less than 1, and the distribution probability of *Minlines* became meaningless. Combining the calculation of entropy values, the optimal *Dr* of 600 m and the corresponding *Minlines* were found at round 12.

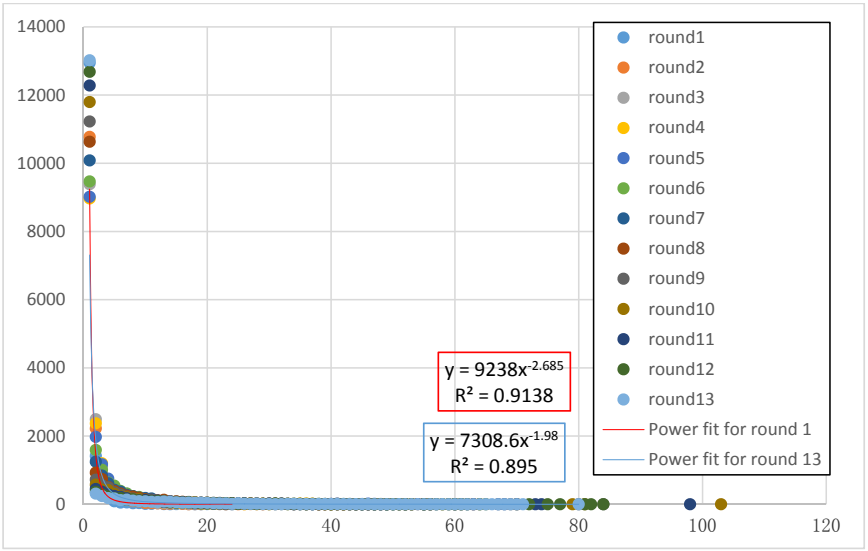

Figure 10. Distribution of all *Nls* values with different radius and the power law fitting lines.

Table 1. Statistical results for different radius from round 1 to round 15.

Round	*Dr (m)*	R^2	*α*	*Minlines (95% of CDP)*	*Minlines (99% of CDP)*	*Entropy*	*Max (Nls)*
1	50	0.914	1.685	6	16	6.30	24
2	100	0.893	1.397	9	28	4.70	31
3	150	0.919	1.175	13	51	4.06	33
4	200	0.911	1.124	15	61	3.68	42
5	250	0.886	1.172	15	51	3.45	51
6	300	0.952	1.238	13	42	3.55	55
7	350	0.930	1.359	10	30	3.59	55
8	400	0.902	1.296	11	35	3.47	63
9	450	0.917	1.152	14	55	3.35	84
10	500	0.923	1.066	17	76	3.30	103
11	550	0.918	1.046	18	82	3.29	98
12	600	0.909	1.005	20	98	3.25	84
13	650	0.895	0.980	-		3.26	80
14	700	0.892	0.863	-		3.24	88
15	750	0.899	0.801	-		3.25	92

4.2. Clustering Process with a Fixed Radius

To verify the efficiency of the selected parameters, we performed the clustering process for each specified *Dr* from round 1 to round 12. Table 2 summarizes the basic information of all clustering results. The minimum CDP is set as 95% ($p < 0.05$). The results are as follows: (1) when the searching radius is small, more gathering lines with short lengths are observed; (2) when the entropy is small, the number of clusters is low, and the total number of lines is relatively high. The average number of lines in one cluster is largest in round 12; (3) when the searching radius is large, the number of clusters with statistical significance is low.

Table 2. Features of the clusters with different fixed radii.

	R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**	**R11**	**R12**
Number of clusters	37	37	29	35	45	76	150	139	95	69	58	55
Number of total clustered lines	733	492	560	793	1101	1594	2566	2607	2267	1937	1787	1784
Average number of clustered lines	20	13	19	23	24	21	17	19	24	28	31	32
Length of centerline	0.34	0.70	0.69	0.74	0.88	1.05	1.22	1.37	1.47	1.62	1.76	1.91

Figure 11 shows the visualization of the cluster analysis from round 1 to round 12. We discovered some interesting features. (1) When the search radius was small, the clustering results indicated spatial relationships mainly to the metro stations, with very few exceptions. This is consistent with Mobike's important role in connecting the public transit in "the last mile" [36]. The search radius determined the clustering results. (2) As the search radius increases, the number of extracted clustering lines with a 95% CDP increases, and the spatial aggregation features are not obvious. (3) Compared with the 95% CDP, clusters with a 99% CDP have a clearer spatial feature, especially when the search radius is large. However, due to the high requirements of CDP, no suitable clusters could be identified under several radii.

Thus, we recommend the complex version of SLCM clustering method with a 99% CDP. The advantage of this method is that significant aggregation areas can be preserved as much as possible.

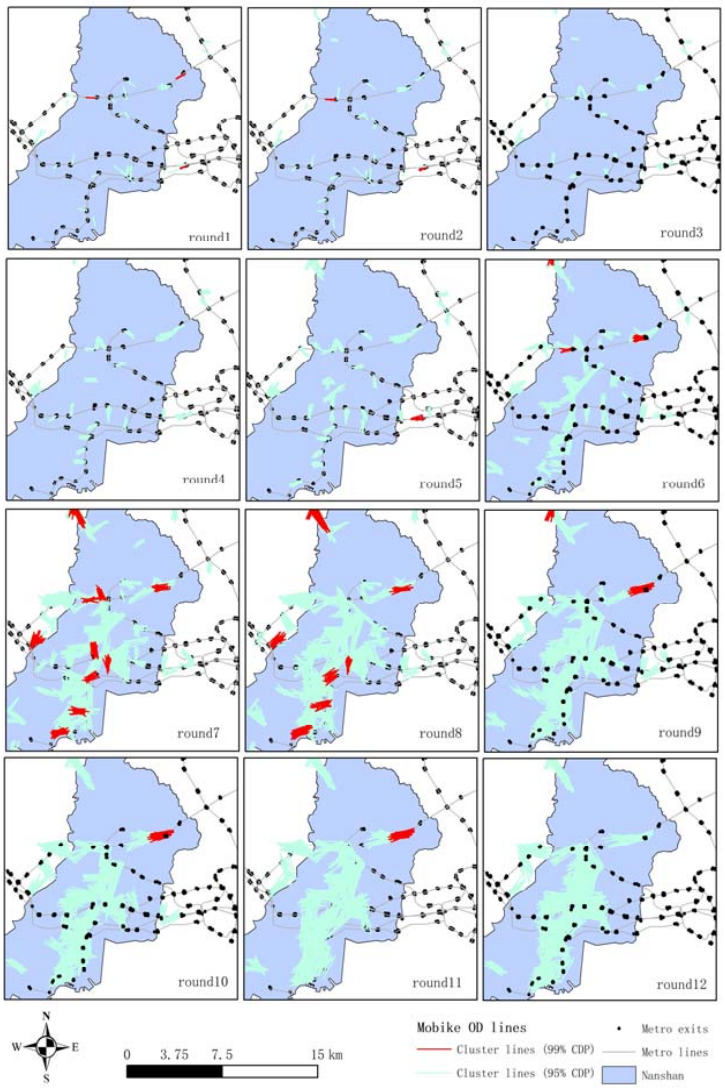

Figure 11. Clustering results for fixed radiuses from round 1 to round 12.

4.3. Clustering Results with Flexible Radius

By using the backward clustering process of the SLCM, the most significant line clusters with 99% of CDP can be preserved. Finally, we found 15 clusters with strong spatial linkages for different search radii (Table 3). These clusters are essentially superimposed from round 1 to round 12 of the specified radius clustering results. It is worth noting that C13 was the exact cluster we mentioned with a low significant level under a large radius. The clustering method with flexible radius helped confirm it. Compared with clustering analysis of the fixed radius, this analysis can fully identify the areas with strong linkages at all spatial scales, with the characteristics of different search radii and centerline lengths.

Table 3. Features of the centerline with flexible radius.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
Radius (m)	550	450	400	400	400	400	400	400	350	350	350	350	250	200	150
Length of centerline (km)	1.6	1.3	1.1	1.7	1.2	1.1	1.3	1.2	1.0	1.0	1.0	1.0	0.7	0.5	0.7
Number of clustered lines	98	60	63	46	42	39	37	35	55	36	34	30	51	29	17

In addition, we found some interesting connections between the origin and destination areas. (1) Most of these connected areas were near metro entrances. However, not all bike trips were related to the metro. In Figure 12c, some destinations were closer to the entrance of a nearby mountain park. (2) Comparing with ordinary residential areas, the residents of the city-village in Shenzhen had a stronger demand and showed a higher frequency of Mobike application. In Figure 12b, more origins of the city-village than ordinary residential houses were identified, and (3) in general, the demand for a connection with public transportation was relatively high, whether it was from a place of residence to a metro/bus station or from a metro/bus station to a work place (Figure 12a,d).

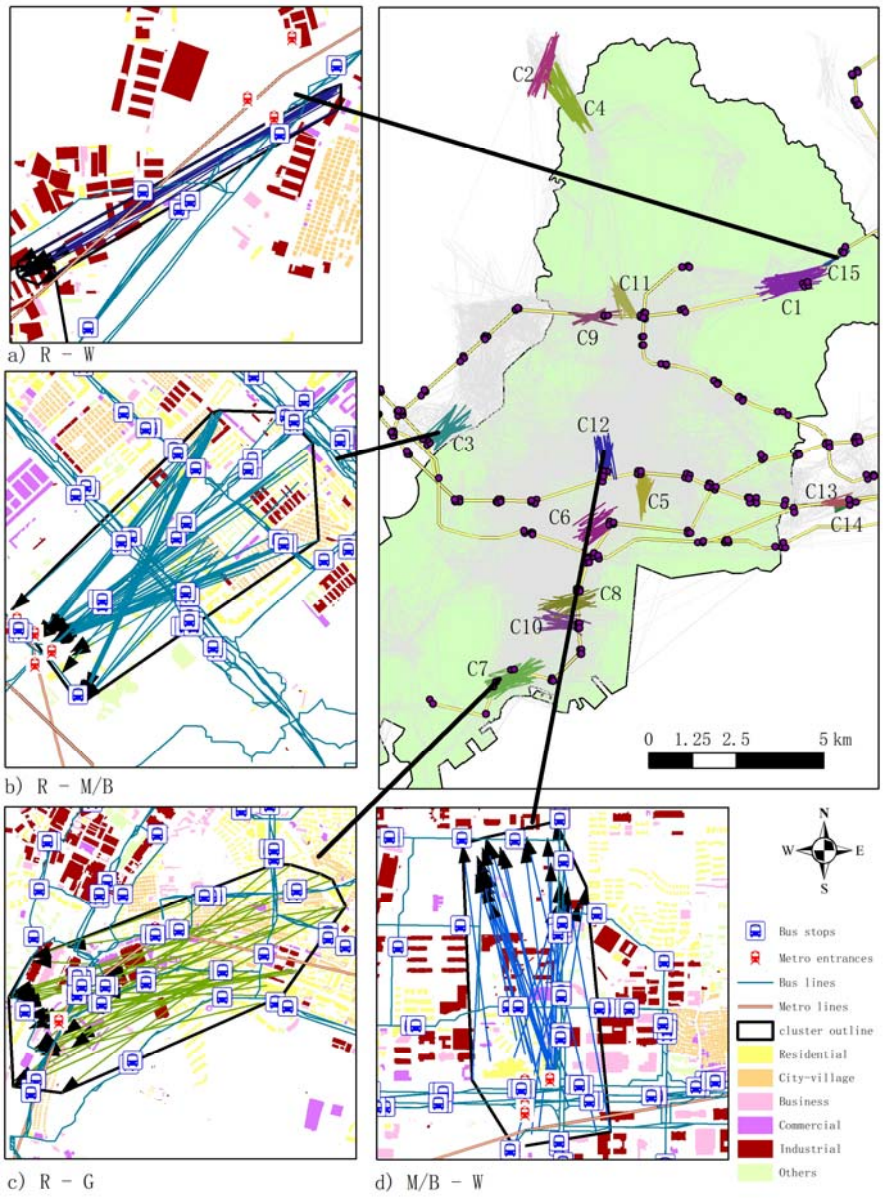

Figure 12. The results of the SCLM and typical clusters with different trip purposes.

5. Discussion and Conclusions

In this paper, we proposed a new method for extracting spatial linkage information from large OD data (i.e., the SLCM). This method is more intuitive and simple for clustering OD lines. The key concept is to identify neighboring lines by searching for the endpoints of OD lines within a certain radius (*Dr*). When the number of neighboring lines exceeds a threshold (*Minlines*), the cluster is considered to be a

significant cluster. This method adopts entropy theory and the probability distribution function to determine the parameters, by which we can effectively obtain reliable parameters. We can directly use empirical values to conduct the clustering method for a specific radius to find the strongest contact areas. While in the absence of prior knowledge, the backward clustering process can help us find all significant clusters using a flexible radius.

To demonstrate the effectiveness of the SLCM clustering method, we have conducted a case study using the Mobike OD data at MPT in the Nanshan District of Shenzhen, China. A comparison of the clustering process showed that the complex version of SLCM can be used to effectively identify spatial clustering. The results showed that this method can be used to synthesize massive flows, find major patterns, and confirm known structures (e.g., typical travel purposes of bike-sharing for the 'last kilometer' problem) and discover unknowns (i.e., special places with more attractions). This type of clustering analysis is also applicable for other commonly encountered OD data, including migration, phone calls, world trade, etc., especially floating location data.

Overall, we believe that combined with other data sources, the new clustering method for OD lines proposed in this paper can be used to obtain new insights into movement data. This work is the start of a series of analyses; currently, we are undertaking in-depth studies on spatial pattern recognition to further explore impact factors as citizen intra-city trips are often influenced by resource allocations and functional areas, such as educational, entertainment, business, and residential areas.

Author Contributions: B.H. and Y.C. contributed materials and analysis tools; Y.Z. contributed to the analysis framework, and Z.G. wrote the paper.

Acknowledgments: This project was supported by the National Natural Science Foundation of China (grant No. 51778366) and the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation at the Ministry of Land and Resources (KF-2016-02-13). The authors would like to thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the final version of the paper.

Conflicts of Interest: The authors declare no conflicts of interest. The founding sponsors had no role in the design of the study, the collection, analyses, or interpretation of data, the writing of the manuscript, or in the decision to publish the results.

References

1. Calabrese, F.; Lorenzo, G.D.; Liu, L.; Ratti, C. Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Comput.* **2011**, *10*, 36–44. [CrossRef]
2. Guo, D.; Zhu, X. Origin-Destination Flow Data Smoothing and Mapping. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 2043–2052. [CrossRef] [PubMed]
3. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; pp. 226–231.
4. Adrienko, N.; Adrienko, G. Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 205–219. [CrossRef] [PubMed]
5. Andrienko, G.; Andrienko, N. Spatio-temporal aggregation for visual analysis of movements. In Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, Columbus, OH, USA, 19–24 October 2008; Ebert, D., Ertl, T., Eds.; IEEE: Washington, DC, USA, 2008; pp. 51–58.
6. Cui, W.; Zhou, H.; Qu, H.; Wong, P.C.; Li, X. Geometry-Based Edge Clustering for Graph Visualization. *IEEE Trans. Vis. Comput. Graph.* **2008**, *14*, 1277–1284. [CrossRef] [PubMed]
7. Buchin, K.; Speckmann, B.; Verbeek, K. Flow Map Layout via Spiral Trees. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 2536–2544. [CrossRef] [PubMed]
8. Holten, D.; Van Wijk, J.J. Force-Directed Edge Bundling for Graph Visualization. *Comput. Graph. Forum* **2009**, *28*, 983–990. [CrossRef]
9. Guo, D.S.; Zhu, X.; Jin, H.; Gao, P.; Andris, C. Discovering spatial patterns in origin-destination mobility data. *Trans. GIS* **2012**, *16*, 411–429. [CrossRef]
10. Zhu, X.; Guo, D. Mapping Large Spatial Flow Data with Hierarchical Clustering. *Trans. GIS* **2014**, *18*, 421–435. [CrossRef]

11. Mao, F.; Ji, M.; Liu, T. Mining spatiotemporal patterns of urban dwellers from taxi trajectory data. *Front. Earth Sci.* **2016**, *10*, 205–221. [CrossRef]
12. Lee, J.G.; Han, J.; Whang, K.Y. Trajectory clustering: A partition-and-group framework. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China, 11–13 June 2007; pp. 593–604.
13. Cao, H.; Mamoulis, N.; Cheung, D.W. Discovery of periodic patterns in spatiotemporal sequences. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 453–467. [CrossRef]
14. Ferrero, C.A.; Alvares, L.O.; Zalewski, W.; Bogorny, V. Movelets: Exploring Relevant Subtrajectories for Robust Trajectory Classification. In Proceedings of the 33rd ACM/SIGAPP Symposium on Applied Computing, Pau, France, 9–13 April 2018.
15. Chen, J.; Leung, M.K.H.; Gao, Y. Noisy Logo Recognition Using Line Segment Hausdorff Distance. *Pattern Recognit.* **2003**, *36*, 943–955. [CrossRef]
16. Lee, J.G.; Han, J.; Li, X.; Gonzalez, H. *Traclass*: Trajectory classification using hierarchical region-based and trajectory-based clustering. Proceedings of VLDB Endowment, Auckland, New Zealand, 23–28 August 2008; pp. 1081–1094.
17. Yuan, G.; Xia, S.; Zhang, L.; Zhou, Y.; Ji, C. An efficient trajectory-clustering algorithm based on an index tree. *Trans. Inst. Meas. Control.* **2011**, *34*, 850–861. [CrossRef]
18. Zhang, D.; Lee, K.; Lee, I.; Zhang, D.; Lee, K.; Lee, I. Hierarchical trajectory clustering for spatio-temporal periodic pattern mining. *Expert Syst. Appl.* **2018**, *92*, 1–11. [CrossRef]
19. Wang, Y.; Qin, K.; Chen, Y.; Zhao, P. Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi GPS data. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 25. [CrossRef]
20. Zheng, Y.; Chen, Y.; Li, Q.; Xie, X.; Ma, W.Y. Understanding transportation modes based on GPS data for web applications. *ACM Trans. Web* **2010**, *4*, 1–36. [CrossRef]
21. Xiao, Z.; Wang, Y.; Fu, K.; Wu, F. Identifying different transportation modes from trajectory data using tree-based ensemble classifiers. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 57. [CrossRef]
22. Campello, R.J.G.B.; Moulavi, D.; Sander, J. Density-based clustering based on hierarchical density estimates. In *PAKDD 2013: Advances in Knowledge Discovery and Data Mining*; Lecture Notes in Computer Science: Volume 7819; Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 160–172.
23. Dodge, S.; Weibel, R.; Forootan, E. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Comput. Environ. Urban Syst.* **2009**, *33*, 419–434. [CrossRef]
24. Tiakas, E.; Papadopoulos, A.N.; Nanopoulos, A.; Manolopoulos, Y.; Stojancivic, D.; DjordjevicKajan, S. Searching for similar trajectories in spatial networks. *J. Syst. Softw.* **2009**, *82*, 772–788. [CrossRef]
25. Etemad, M.; Soares Júnior, A.; Matwin, S. Predicting Transportation Modes of GPS Trajectories Using Feature Engineering and Noise Removal. *Adv. Artif. Intell.* **2018**, 259–264.
26. Moreno, B.; Júnior, A.S.; Times, V.; Tedesco, P.; Matwin, S. Weka-SAT: A Hierarchical Context-Based Inference Engine to Enrich Trajectories with Semantics. In *Advances in Artificial Intelligence. AI 2014*; Lecture Notes in Computer Science, Volume 8436; Sokolova, M., van Beek, P., Eds.; Springer: Cham, Switzerland, 2014.
27. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. [CrossRef]
28. ArcGIS Help. Available online: http://resources.arcgis.com/en/help/main/10.2/index.html#//005p00000006000000 (accessed on 30 July 2013).
29. Yaneer, B.Y.; Concepts: Power Law. New England Complex Systems Institute. Available online: http://www.necsi.edu/guide/concepts/powerlaw.html (accessed on 18 August 2015).
30. Barry, C.A. *Pareto Distributions*; International Co-Operative Publishing House: Fairland, MD, USA, 1983; ISBN 0-89974-012-X.
31. Pareto Distribution. Available online: https://en.m.wikipedia.org/wiki/Pareto_distribution (accessed on 15 January 2018).
32. Clauset, A.; Shalizi, C.R.; Newman, M.E.J. Power-Law Distributions in Empirical Data. *SIAM Rev.* **2009**, *51*, 661–703. [CrossRef]
33. Shenzhen Statistical Yearbook 2017. Available online: http://www.sztj.gov.cn/xxgk/tjsj/tjnj/201712/W020171219625244452877.pdf (accessed on 19 December 2017). (In Chinese and English)

34. Mobike Stays Ahead in Chinese Bike-Sharing Market, Analysis Says. 2017. Available online: http://www.chinadaily.com.cn/business/tech/2017-02/10/content_28163187.htm (accessed on 10 February 2017).
35. 2017 Sharing Bike and Urban Development White Paper. 2017. Available online: http://news.cssn.cn/zx/bwyc/201704/t20170412_3484389.shtml (accessed on 12 April 2017). (In Chinese)
36. Deng, L.F.; Xie, Y.F.; Huang, D.X. Bicycle-sharing facility planning based on riding spatiotemporal data. *Planners* **2017**, *10*, 82–88.