International Journal of
Geo-Information

Article

An OD Flow Clustering Method Based on Vector Constraints: A Case Study for Beijing Taxi Origin-Destination Data

Xiaogang Guo [1,†], Zhijie Xu [2,†], Jianqin Zhang [1,*], Jian Lu [1] and Hao Zhang [1]

[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 102616, China; 2108160317006@stu.bucea.edu.cn (X.G.); 2108521518001@stu.bucea.edu.cn (J.L.); 2108521518022@stu.bucea.edu.cn (H.Z.)

[2] School of Science, Beijing University of Civil Engineering and Architecture, Beijing 102616, China; xuzhijie@bucea.edu.cn

* Correspondence: zhangjianqin@bucea.edu.cn

† Joint first authors with equal contribution.

Received: 16 January 2020; Accepted: 20 February 2020; Published: 22 February 2020

Abstract: Origin-destination (OD) flow pattern mining is an important research method of urban dynamics, in which OD flow clustering analysis discovers the activity patterns of urban residents and mine the coupling relationship of urban subspace and dynamic causes. The existing flow clustering methods are limited by the spatial constraints of OD points, rely on the spatial similarity of geographical points, and lack in-depth analysis of high-dimensional flow characteristics, and therefore it is difficult to find irregular flow clusters. In this paper, we propose an OD flow clustering method based on vector constraints (ODFCVC), which defines OD flow event point and OD flow vector to express the spatial location relationship and geometric flow behavior characteristics of OD flow. First, the OD flow vector coordinate system is normalized by the Euclidean distance-based OD flow event point spatial clustering, and then the OD flow clusters with similar flow patterns are mined using adjusted cosine similarity-based OD flow vector feature clustering. The transformation of OD data from point set space to vector space is realized by constraining the vector coordinate system and vector similarity through two-step clustering, which simplifies the calculation of high-dimensional similarity of OD flow and helps mining representative OD flow clusters in flow space. Due to the OD flow cluster property, the k-means algorithm is selected as the basic clustering logic in the two-step clustering method, and a sum of squared error perceptually important points algorithm considering silhouette coefficients (SSEPIP) is adopted to automatically extract the optimal cluster number without defining any parameters. Tested by origin-destination flow data in Beijing, China, new traffic flow communities based on traffic hubs are obtained by using the ODFCVC method, and irregular traffic flow clusters (including cluster mode, divergence mode, and convergence mode) with representative travel trends are found.

Keywords: origin-destination (OD) flow clustering; vector constraints; OD flow event point; OD flow vector

1. Introduction

Origin-destination (OD) flow is the semantic recognition and feature extraction of complex trajectory data. It clearly expresses the geographic information of the origin and destination points of real trajectory, the implicit trajectory flow direction and distance, as well as specific thematic attributes (such as population migration, logistics and freight flow, traffic flow, etc.) [1]. However, with the popularization of GPS positioning and the increase of Internet of Things sensors, massive mobile trajectory data has also emerged. How to find the flow pattern and explore the human–earth interaction in dense OD trajectory data is an important issue in mobile trajectory data mining [2,3].

Some scholars use visual analysis methods such as edge bundling, OD point clustering to solve the phenomenon of overlapping and displaying confusion of edges [1,4–6], thus highlighting the larger flow of OD clusters. Some scholars also use spatial clustering for pattern recognition by O-point clustering, D-point clustering, OD-point clustering, and OD flow (edge) clustering for different application scenarios [7–11]. In terms of research ideas and methods of OD flow clustering, most researchers regard OD flow data as a set of O and D points. According to the spatial characteristics of OD points, the point clustering algorithm is used to realize OD flow clustering through double-iteration [12–15]. These OD flow clustering algorithms are easily constrained by the spatial distribution of OD points and the setting of search radius or internal connectivity parameters. They do not have the ability to discover irregular flow clusters actively.

The existing clustering methods of geographic OD flow rely on the characteristics of geographical units and functional areas, and too closely link the inherent land use data of the origin and destination points with the dynamic geographic flow behavior. In the model driven flow clustering method, the discovered flow patterns are more classic and inferred from traditional location factors and dynamic factors, and they do not fully mine the intrinsic value of OD trajectory data. How to take full advantage of "every valuable and real data" with the idea of data-driven methods, and how to use the data representation of OD flow to mine the geographical flow pattern to verify, update, supplement, and modify the geographical unit, land-use type, and functional division, are important research contents for studying the interaction pattern of geospatial subspace and describing the geographical flow.

Facing the problems of time delay and non-dynamics caused by the dependence of inherent geographic units in geographic flow pattern mining, this paper proposes an OD flow clustering method based on vector constraints. We plan to mine the dynamic interaction mode of flow space through the data characteristics of geographic flow and provide new tools for the mining of complex and irregular flow patterns.

In this study, the spatial and geometric (behavior) attributes of OD flow are expressed by defining OD flow event point and OD flow vector, then, OD flow vector coordinates are normalized by event point spatial clustering. On this basis, OD flow clusters with similar flow patterns are found by OD flow vector clustering. Finally, taking Beijing taxi OD data as an example, this study uses the method to find taxi traffic flow communities and irregular shape clusters with the same flow pattern.

2. Related Research

2.1. Visualization of OD Flow

The visualization methods of OD flow data mainly include the flow map [16–21], the OD matrix, and the OD map [22–25]. Among them, the flow map better reflects the spatial characteristics of OD data. But there are also some problems in that method, such as the visual cluttering problem, the modifiable area unit problem, the normalization problem, the salience bias, and so on [1,8,11]. In order to solve these problems, scholars have proposed edge rerouting [26], edge bundling [27,28], matrices of multiple maps [24,29] and other means to reduce clutter, but it does lead to the loss of spatial information of OD points, and the relationship between OD flows is difficult to be perceived. The location of OD points is aggregated by means of spatial clustering and graph partitioning [17,30], but arbitrary clustering results in loss of spatial resolution and the meaninglessness of clustering patterns. By using default geographic units or multiscale clustering [31], scale differences and significant deviation can be solved. However, OD flows between different scales cannot be quantitatively compared with each other when the scale, scope, and sampling accuracy of datasets are significantly different, and the patterns in flow graphs tend to be controlled by flows with longer geographic distances. In order to solve these problems comprehensively, some scholars have proposed a new method for flow map generalization [1] which can be used to deal with different scale flow datasets. The idea of this method is to solve the visualization problem of OD flow through the point set density distribution of OD points.

2.2. Clustering of OD Flow

Clustering is an important tool for pattern discovery. After long-term development, a variety of clustering algorithms have been created and optimized, such as hierarchical clustering, density-based clustering, model-based clustering, partition-based clustering, and grid-based clustering [32] (pp. 2–19). New clustering algorithms include semi-supervised clustering [32] (pp. 136–155), spectral clustering, and clustering based on non-negative matrix factorization [32] (pp. 157–213), as well as high-dimensional data clustering [33], graph clustering [34,35], uncertain data clustering, and multisource related data clustering for complex problems [36,37]. However, no matter how clustering algorithm develops under data-driven or algorithm-driven, similarity is the core issue of clustering algorithm. Through a literature review, the design ideas of OD flow clustering algorithm are classified into two categories, point-based OD flow clustering and line-based OD flow clustering.

Point-based OD flow clustering defines the similarity index as similarity measure based on OD points. There are many similarity measures, such as Euclidean distance, Manhattan distance, Chebyshev distance, and so on. In the process of OD flow clustering, it is usually through the OD nested point clustering algorithm. Some scholars have proposed a simple line clustering algorithm to find the closest spatial relationship by searching the adjacent lines of OD travel within a certain radius [15]. However, in the specific algorithm, the search radius based on OD points is adopted, which is dependent on the similarity of points to iterate. Some scholars proposed a spatial scan statistical method based on ant colony optimization to detect OD clusters of arbitrary shape [38]. The definition of OD clusters also depends on the internal connectivity of the OD points. Therefore, in this paper, this kind of algorithm is regarded as OD flow clustering based on point idea.

Line-based OD flow clustering is a clustering algorithm that is based on line (trajectory) similarity. In map synthesis, time series clustering, and trajectory clustering, there are many measurement parameters for similarity of measurement lines. For complex trajectories (multi-segment lines), there are DTW distance, minimum outsourcing rectangle distance, longest common subsequence distance, editing distance, Frechet distance, and so on [39]. For line group synthesis, it is macroscopically defined as similar geometric features, similar spatial relations, and similar attributes (semantics) [40], specifically, including topological similarity, direction mean, circular variance, average length, tortuosity coefficient, line group density, etc. When researching the similarity of simple line type (primary OD flow) and sub-trajectory, the similarity index of geometric constraints is considered [41,42]. For example, the similarity distance of sub-trajectory in TR-OPTICS trajectory clustering algorithm is measured as horizontal distance, vertical distance, and angle distance [43].

OD flow is a simple linear form in geometry, which consists of O and D points. Therefore, when studying OD flow clustering, we can use OD points to iterate clustering, and use low-dimensional point similarity to restrict the similarity of OD flow. Although high-dimensional data can be processed by specific dimension reduction algorithms, its inherent characteristics are weakened or ignored [32] (pp. 216–220). The flow direction, flow distance, and flow space of OD flow, which researchers are more concerned about, are not intuitively reflected. The difficulty of the line-based flow clustering algorithm is to define the similarity function suitable for OD flow clustering according to the spatial and attribute characteristics of OD flow.

3. OD Flow Clustering Method Based on Vector Constraints

This section introduces the OD flow clustering method based on vector constraints. The method is described in the following three aspects: the definition of related concepts, model parameters (clustering number and distance function), and the details of clustering process. The components of the method are shown in Figure 1.

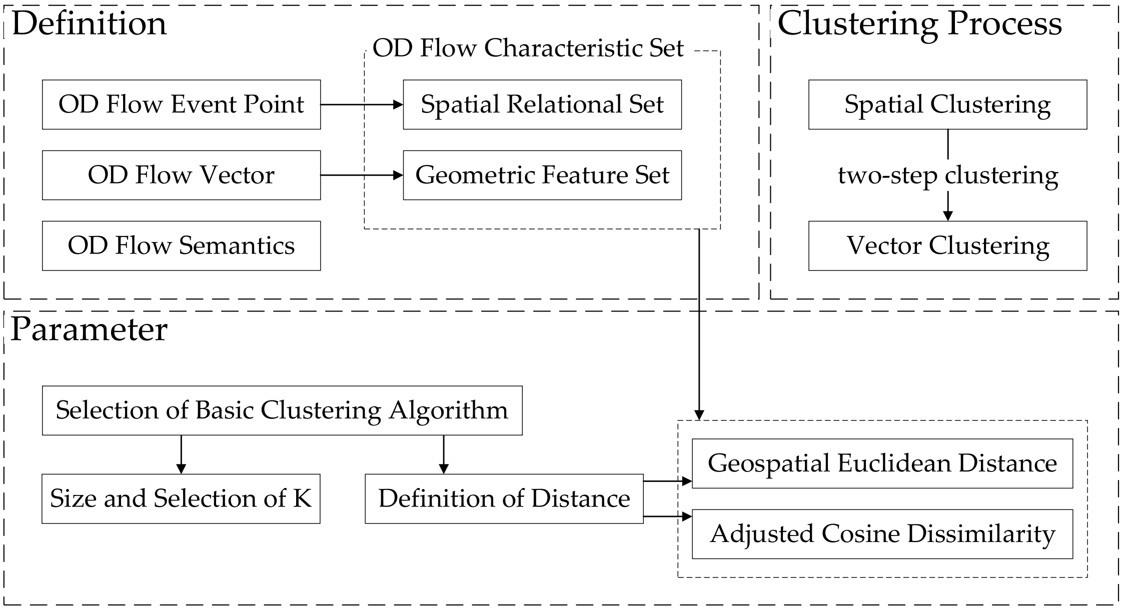


Figure 1. Component diagram of method related concepts.

3.1. Definition

3.1.1. OD Flow Event Point

$$P_{od} = \left\{X_{P_{od}}, Y_{P_{od}}\right\},\ X_{P_{od}} = (X_O + X_D)/2,\ Y_{P_{od}} = (Y_O + Y_D)/2 \tag{1}$$

where X_O and Y_O are geographic coordinates of origin point (O-point), and X_D and Y_D are geographic coordinates of destination point (D-point). According to Equation (1), P_{od} is the midpoint of the OD flow geometric line. Taking the taxi OD trajectory data as an example, OD flow is the taxi trajectory data which contains the semantic information of passengers boarding and disembarking positions. The generation of an OD flow represents a passenger's travel behavior by taxi. Some scholars have proposed a spatiotemporal point process model, which regards the starting point and ending point of taxi as two different point processes [44,45]. If judging by the semantic information of OD flow nodes, OD flow is a point process with two different properties. If judging from the source of data collection, that is, taxi GPS data with passenger travel events (attributes), OD flow is an event with passenger semantics in the process of taxi operation, which is regarded as a point process. In this study, line can be abstracted as a point based on map generalization in small and medium scale, and further interpreted from the perspective of spatial and temporal point process. OD flow is regarded as an event of urban crowd activity and abstracted as a point process. Its spatial attributes are represented by the midpoint of geometric line of OD flow. Therefore, we define P_{od} as the event point of OD flow, which has the spatial attribute of OD flow. It should be emphasized that the original intention of using point coordinates to represent the spatial location of OD flow is to treat OD flow as a whole and as a line object and, then, use OD flow event points to represent the overall spatial location attributes of OD flow.

3.1.2. OD Flow Vector

$$\vec{OD} = (\Delta X, \Delta Y) = (X_D - X_O, Y_D - Y_O) \tag{2}$$

Equation (2) shows that $\vec{OD}$ is the geometric vector of OD flow. Taking taxi OD trajectory data as an example, O-point is taxi GPS location when passenger boarding incident occurs, and D-point is taxi GPS location when passenger alighting incident occurs. OD flow is a directed line segment. As a semantic extraction of complex trajectory data, OD flow has no entity meaning in geographic space, but it represents passenger flow in geographic space in semantic space. Although there is no real track based on road network in OD flow, there are clear directions of crowd activity and spatial and temporal distances between OD. In this study, OD flows are considered as geometric vectors. The size and direction of OD flows are expressed by the modulus and direction of OD flows.

3.1.3. OD Flow Semantics

The semantics of OD flow can be regarded as events of urban crowd activities, which are usually inferred from the semantics of OD points [46] (pp. 130–158). For example, from residential to office area is regarded as commuting, from residential to business circle is considered as shopping. Accordingly, the semantic information of OD flows depends heavily on the accuracy and granularity of point of interest (POI) data. In this study, the semantic information of OD points is not extracted and aggregated in advance based on urban functional areas and urban travel rules. There is no clustering trend for high-dimensional data in the whole space. Semantic space and geographic space do not necessarily have good similarity in clustering of OD flows. Therefore, it is not concerned with similar clustering of specific semantic features of OD flows in this paper. It is hoped that spatial clustering of OD flows can be used to mine the flow rules and potential patterns of OD flows.

3.1.4. OD Flow Characteristic Set

$$C_{OD} = \left\{ \begin{array}{c} Spatial\ Relational\ Set \big| (OD\ Flow\ Event\ Point\ Spatial\ Relations), \\ Geometric\ Feature\ Set \big| (Vector\ Size\ Relation, Vector\ Direction\ Relation) \end{array} \right\} \tag{3}$$

Without considering the semantic similarity, some scholars have proposed that the target features of spatial line group can be summarized as a set of spatial relations (spatial topological relations, spatial direction relations, and spatial distance relations) and a set of geometric features (line length and average length, tortuous coefficient, and line group density) [40]. Owing to the particularity of OD line structure, it is unnecessary to pay attention to the topological relationship and tortuous coefficient of OD [40,43]. The direction of OD flow is not calculated by the direction angle but expressed by the geometric vector feature. The spatial distribution and distance of OD flow are replaced by the distribution density and distance of the event points of OD flow. Hence, we define the data structure of OD flow as:

$$C_{OD} = \left\{ P_{od}, \vec{OD} \right\} \tag{4}$$

3.2. Parameter

3.2.1. Selection of Basic Clustering Algorithm

In the research of spatial pattern recognition of OD flow clustering, there are three main methods to identify spatial distribution [38]. Among them, there are two main types of improved classical clustering algorithm, hierarchical clustering algorithm for OD flow and density clustering algorithm based on origin and destination points. The advantage of the hierarchical clustering algorithm is that the structure of clustering results is tree-like, and it can be expressed by multiscale clustering. The advantage of density-based clustering algorithm is that it has the potential to mine spatial clusters of OD points with arbitrary shapes by connecting high-density spatial entities with continuous spaces into clusters. The third method focuses on extending the traditional spatial statistical method to identify OD flow clustering anomalies by defining new OD flow similarity. The limitations of existing hierarchical clustering and density clustering algorithms are as follows: First, the definition of distance and the size of value are uncertain and secondly, the constraint of discovering clusters of arbitrary shape due to the splitting of OD points in OD flows. The limitation of the modification based on spatial statistical algorithm lies in the loss of OD flow information caused by dimensionality reduction of OD flow similarity definition.

This research adopts the K-means clustering algorithm, the main reason is the definition and selection of clusters [47,48]. Different clustering algorithms have different definitions and mining abilities for clusters because of their different logic. The expected clustering result of this algorithm is OD flow with close spatial relationship and similar geometric shape within clusters, so it is more suitable for partition-based clustering method, which is based on central cluster definition.

3.2.2. Size and Selection of K

The selection of seed points is an important step in K-means clustering. The size of K determines the number of clusters. The selection of K affects the efficiency of iteration. In previous studies, elbow method and silhouette coefficient are classic indexes to evaluate clustering effect [49,50]. By traversing K, sum of squared error (SSE) and silhouette coefficients under different K values are calculated, and the elbow nodes of SSE curve and the corresponding larger silhouette coefficients are found. In previous studies, naked eye judgment was often used. In the algorithm design of this study, perceptually important points (PIP's) is adopted to automatically identify SSE elbow points [51], and the silhouette coefficient is used to check.

By calculating SSE, the elbow method is used to find the relationship between K value and the real clustering number.

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2 \tag{5}$$

where C_i is the ith cluster, p is the sample point in C_i, and m_i is the mean of all samples in C_i. When K is less than the real clustering number, the increase of K leads to a significant decrease in SSE, and when K reaches the real clustering number, the gain of clustering effect decreases rapidly with an increase of K. Therefore, the K value corresponding to the elbow inflection point of the SSE curve is the real clustering number of data.

Silhouette coefficient is a clustering evaluation method which combines cohesion and separation. For any vector i, its silhouette coefficient is:

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{6}$$

where $a(i)$ is the average distance from vector i to all other points in the cluster to which it belongs, and $b(i)$ is the minimum distance from vector i to all points in the cluster that it does not belong. The average of all the silhouette coefficients is the total silhouette coefficients of the clustering results. The closer the silhouette coefficient approaches to 1, the better the cohesion and separation are. But the silhouette coefficient is a relative evaluation index. The silhouette coefficient fluctuates with the change of K. It is a non-convex curve. There are many local optimum solutions. Usually, the elbow method is needed to assist, and the K value corresponding to the local maximum of the silhouette coefficient is chosen as the optimal clustering number.

When determining the size of K, the elbow method is affected by subjective factors, and there are multiple local maximum values using silhouette coefficient. Therefore, the SSE perceptually important points algorithm considering silhouette coefficients (SSEPIP) is adopted to automatically extract the optimal clustering number. The process of SSEPIP is as follows.

The SSE curve is defined as sequence P, where the first two PIP's are the first and last points of P, and the next PIP is the point in P with maximum distance to the first two PIP's. The distance is defined as the vertical distance between the test point P_3 and the straight line connecting two adjacent PIP's (Figure 2):

$$VD(p_3, p_c) = \left| y_c - y_3 \right| = \left| \left(y_1 + (y_2 - y_1) \cdot \frac{x_c - x_1}{x_2 - x_1} \right) - y_3 \right| \tag{7}$$

where $x_c = x_3$. PIP's algorithm is often used for data compression, so the number of PIP's changes with the experimental requirements. In this experiment, sequence P (SSE) is a monotone curve, and there are inflection points around the third PIP, and therefore only once PIP recognition is needed, as shown in Figure 3.

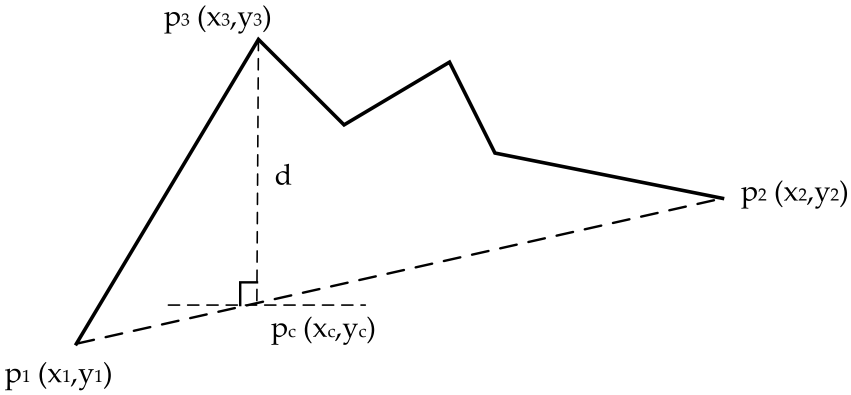


Figure 2. Capturing sequence fluctuations by measuring vertical distance.

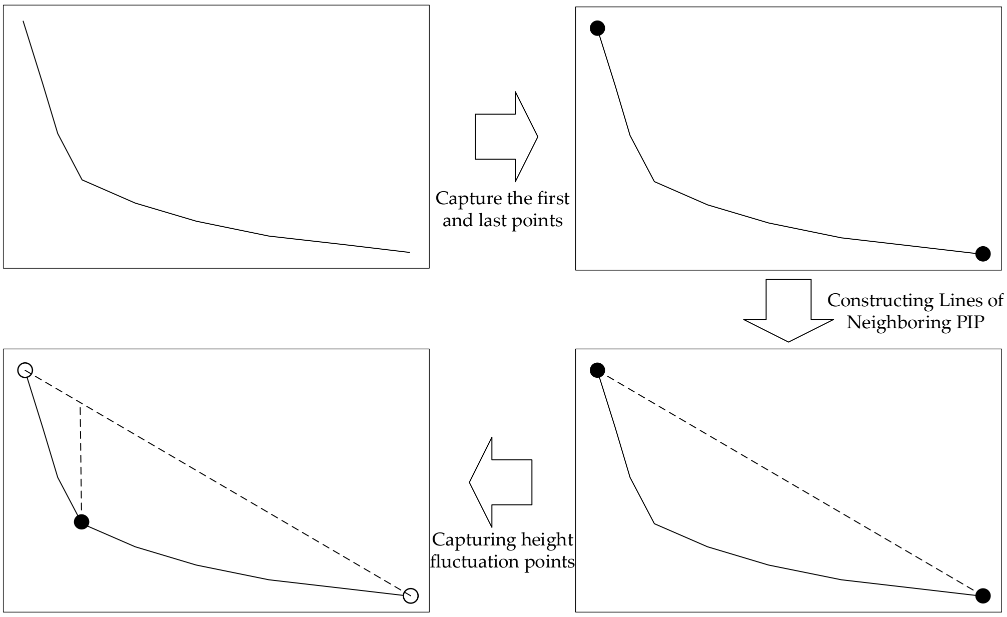

Capture the first and last points
Constructing Lines of Neighboring PIP
Capturing height fluctuation points

Figure 3. Sum of squared error (SSE) inflection point recognition process.

The PIP's algorithm is generally used to compress static data, and it cannot solve the sequence with variable length stably. The third PIP oscillates slightly around the inflection point as the tail point changes, therefore, the local maximum point of silhouette coefficient is used as the constraint condition to help select the best K value. In order to better illustrate the recognition process, the SSE sequence and the silhouette coefficient sequence are standardized. Figure 4 shows the stepwise results of using silhouette coefficient to assist in identifying SSE inflection points. The black line represents SSE sequence and the blue line represents silhouette coefficient sequence.

In the traversal process of K value, each new position needs to be selected. In order to optimize the efficiency of clustering algorithm in each cycle, we evaluate the average silhouette coefficients of each spatial cluster and generate new seed points in the range of the spatial cluster with the smallest silhouette coefficients for the next round of calculation. The specific process is shown in Figure 5.

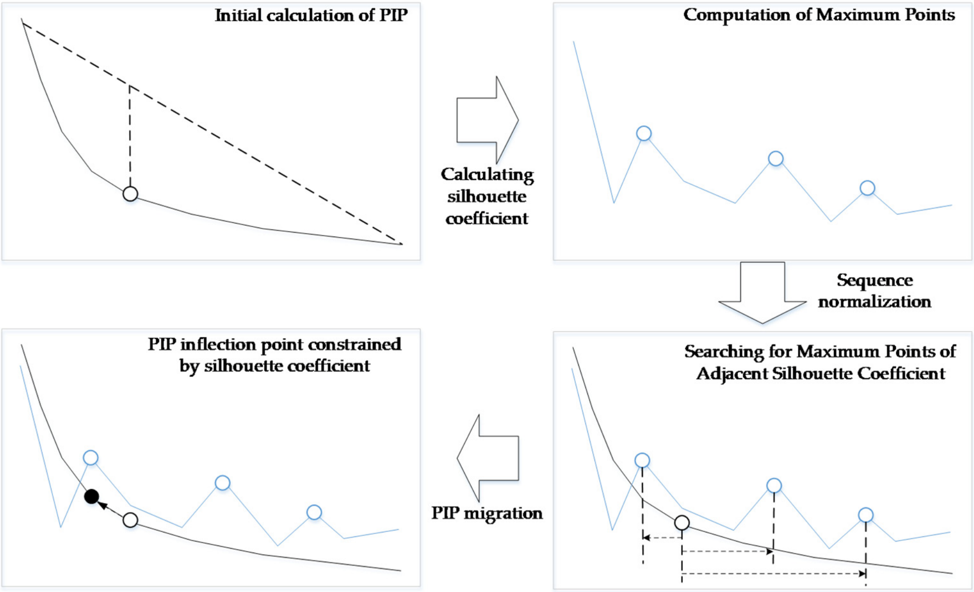

Initial calculation of PIP
Calculating silhouette coefficient
Computation of Maximum Points
Sequence normalization
Searching for Maximum Points of Adjacent Silhouette Coefficient
PIP migration
PIP inflection point constrained by silhouette coefficient

Figure 4. SSE inflection point recognition considering silhouette coefficient.

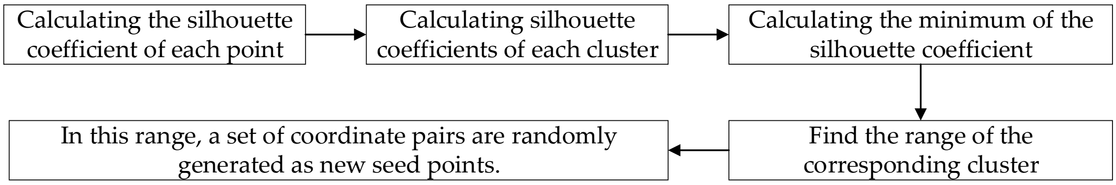

Calculating the silhouette coefficient of each point
Calculating silhouette coefficients of each cluster
Calculating the minimum of the silhouette coefficient
Find the range of the corresponding cluster
In this range, a set of coordinate pairs are randomly generated as new seed points.

Figure 5. New seed point selection process.

3.2.3. Definition of Distance

In the research of spatial statistics methods of OD flow, some scholars obtained the distance of OD flow by weighted summation of O-point distance and D-point distance [52]. Some scholars obtained the distance of OD flow by weighted summation of vector coordinates and attribute variables of OD flow [53]. In this study, we take OD flow as a whole object, and try to construct the corresponding distance function through the spatial and geometric attributes of OD flow. However, in high-dimensional data analysis, data from different dimensions cannot be directly compared and calculated. When constructing distance function, weight allocation has strong subjectivity. In this study, two-step clustering based on spatial dimension similarity and geometric feature similarity is carried out, while Euclidean distance and adjusted cosine similarity are used as spatial distance function and geometric feature distance function.

$$D_{spatial}(i,j) = D_{EUC}\left(P_i, P_j\right) = \sqrt{\left(X_{p_i} - X_{p_j}\right)^2 + \left(Y_{p_i} - Y_{p_j}\right)^2} \tag{8}$$

The spatial characteristic distance of OD flow is defined as the geospatial Euclidean distance of OD flow event points.

$$\begin{aligned} D_{vector}(i,j) = 1 & -Sim_{AdjCos}\left(OD_i, OD_j\right) \\ & = 1 - \frac{(\Delta X_i - R_x)\cdot\left(\Delta X_j - R_x\right) + \left(\Delta Y_i - R_y\right)\cdot\left(\Delta Y_j - R_y\right)}{\sqrt{(\Delta X_i - R_x)^2 + \left(\Delta Y_i - R_y\right)^2}\cdot\sqrt{\left(\Delta X_j - R_x\right)^2 + \left(\Delta Y_j - R_y\right)^2}} \end{aligned} \tag{9}$$

The geometric characteristic distance of OD flow is defined as the adjusted cosine dissimilarity. R is the intra-cluster mean in a given dimension. The adjusted cosine similarity normalizes different dimensions according to the difference of vector angles, indirectly considers the influence factors of vector modulus, and synthetically measures the similarity of vector size and direction [54]. Because the range of the similarity is [−1,1], the distance function is the dissimilarity calculated by the difference.

3.3. Clustering Process

Compared with traditional clustering, the difficulty of line clustering and even high-dimensional clustering is how to deal with high-dimensional information. From a differential point of view, a straight line is constructed from innumerable points, so the process point model is suitable for describing flow space data. Vector is the best descriptive features of line shape. The size and direction of vectors describe the length and angle of line, thus expressing the distance and direction of flow. But vector features cannot describe spatial dimension information. Therefore, we use event points to express the spatial information of the flow. Any high-dimensional object can be mapped into a point object in two-dimensional space. In this study, the spatial characteristics of the OD flow event points are used to reflect the spatial properties of the flow. In this way, we express any flow object through the event points and vectors of the flow, and cluster from the spatial dimension and the geometric dimension in two steps.

The clustering logic is shown in Figure 6.

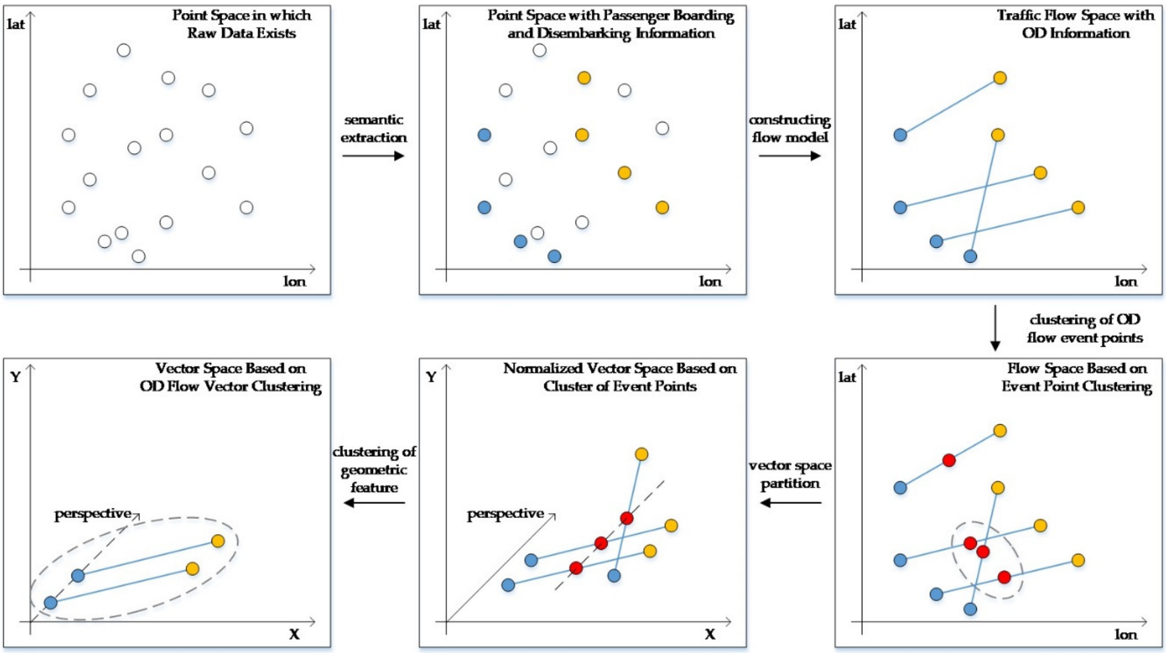

lat
Point Space in which Raw Data Exists
lon
semantic extraction
lat
Point Space with Passenger Boarding and Disembarking Information
lon
constructing flow model
lat
Traffic Flow Space with OD Information
lon
clustering of OD flow event points
lat
Flow Space Based on Event Point Clustering
lon
vector space partition
Y
Normalized Vector Space Based on Cluster of Event Points
perspective
X
clustering of geometric feature
Y
Vector Space Based on OD Flow Vector Clustering
perspective
X

Figure 6. Logic diagram of clustering algorithms.

Figure 6 shows the process of transforming OD flow data from point space to flow space, and then to vector space in clustering algorithm logic. The white dots represent the original taxi GPS data, the blue dots represent the passenger boarding position, the yellow dots represent the passenger getting off position, and the red dots represent the OD flow event point. The proposed algorithm can be divided into the following two steps: The first step is to constrain the vector coordinate system by spatial clustering of OD flow event points, and the second step is to constrain the OD flow vector characteristics by clustering the similarity of geometric vectors. The original OD data exists in discrete GPS trajectory point space. In previous studies, paired OD point set data was obtained by semantic extraction. In this paper, we construct OD flow dataset by calculating OD flow event points and OD flow vector of OD point pairs and transform OD point set space into OD flow space. In OD flow space, we describe the expression of OD flow data as two dimensions, i.e., spatial dimension and geometric feature dimension, and define that the elements in OD flow cluster should satisfy both spatial dimension similarity and geometric feature dimension similarity. In this way, the OD flow clustering process is realized by the two-step clustering method. The first is the process of "space division". In small-scale or multiscale analysis, the geospatial location attributes of OD flow are expressed by OD flow event points. Therefore, the OD flow is divided into several spatial clusters in the flow space by using the OD flow event point clustering. The size and direction of OD flow are different in each spatial cluster, while the spatial location relationship between OD flows is relatively close. Then, it is the "vector clustering" process. In this process, only considering the geometric characteristics of OD flows in each spatial cluster, adjusted cosine similarity is calculated by OD flow vectors and clustered. The implicit premise is that OD flows in each spatial cluster are translated before geometric feature clustering, ignoring the spatial location differences of OD clusters in the same spatial cluster, and then the OD flow vector coordinate system is unified. Therefore, after "space partitioning", each spatial cluster in OD flow space is transformed into an independent vector space, and all vector clustering processes run in parallel.

In this study, the spatial distance and morphological distance are not integrated into a composite flow distance function. The reason is that the fusion of spatial distance and morphological distance is very complex, and the two features depend on each other and influence each other. Some studies tried to use weighted distance function to express the flow distance, but the problem of multiscale expression and global normalization cannot be well solved [52,53]. From the global point of view, the scale differences caused by different lengths, angles, and spatial locations cannot be well solved. From the local point of view, the density distribution of different dimensions has a significant impact on clustering results, so the global optimal solution cannot be obtained. Therefore, this study attempts to solve the global distribution through dimension segmentation, first through spatial clustering, in order to obtain a cluster set with close spatial relations within the cluster and assume that each spatial cluster exists in a separate and unified vector space coordinate system. Then, clustering is carried out by the geometric feature distance in the spatial cluster to solve the problem of uneven local feature density, and representative vector clusters in different spatial clusters are obtained respectively. The specific steps are as follows:

Step 1 At present, most OD flow data storage forms are O-point coordinates (X_O, Y_O), D-point coordinates (X_D, Y_D), and thematic attributes. Therefore, it is necessary to extract OD flow event points and calculate flow vectors to obtain OD flow feature set.

Step 2 K-means clustering based on the spatial distance of event points. The K_S value increases from 2 to the optimal number of spatial clusters solved by SSEPIP.

Step 3 For each spatial cluster (N in number), K-means clustering is carried out based on the geometric feature distance of OD flow vectors. The K_{VN} value increases from 2 to the optimal number of vector clusters solved by SSEPIP.

Step 4 By calculating the average of OD flow event points and OD flow vectors in clusters, we can get representative flows of clusters, and visualize them (expressing the direction of OD flows by moving points).

The origin-destination flow clustering vector constraints (ODFCVC) method is suitable for distributed computing environment, especially for the third step of the algorithm, each spatial cluster performs the geometric feature clustering operation independently.

The flow chart of clustering method is shown in Figure 7.

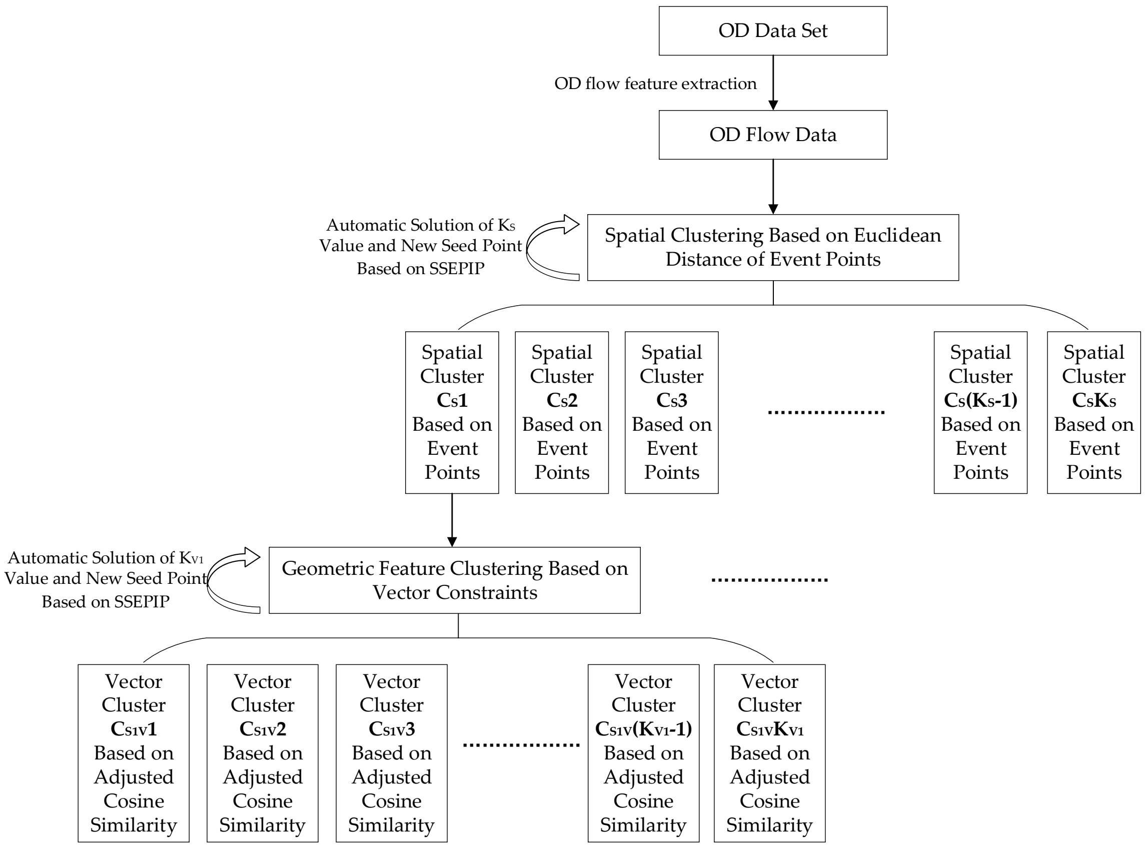

OD Data Set
OD flow feature extraction
OD Flow Data
Automatic Solution of Ks Value and New Seed Point Based on SSEPIP
Spatial Clustering Based on Euclidean Distance of Event Points
Spatial Cluster Cs1 Based on Event Points
Spatial Cluster Cs2 Based on Event Points
Spatial Cluster Cs3 Based on Event Points
....
Spatial Cluster Cs(Ks-1) Based on Event Points
Spatial Cluster CsKs Based on Event Points
Automatic Solution of Kv1 Value and New Seed Point Based on SSEPIP
Geometric Feature Clustering Based on Vector Constraints
....
Vector Cluster Cs1v1 Based on Adjusted Cosine Similarity
Vector Cluster Cs1v2 Based on Adjusted Cosine Similarity
Vector Cluster Cs1v3 Based on Adjusted Cosine Similarity
....
Vector Cluster Cs1v(Kv1-1) Based on Adjusted Cosine Similarity
Vector Cluster Cs1vKv1 Based on Adjusted Cosine Similarity

Figure 7. Origin-destination flow clustering vector constraints (ODFCVC) method flow chart.

C_SN is the N-th spatial cluster based on OD flow event points, $N = 1, 2, \ldots\ldots, K_S$. K_S is the optimal solution of global spatial cluster K value. $C_{SNV}M$ is the M-th vector cluster contained in the N-th spatial cluster, $M = 1, 2, \ldots\ldots, K_{VN}$. K_{VN} is the optimal solution of vector cluster K value based on the adjusted cosine similarity in the N-th spatial cluster.

Assuming that the distance matrix is a symmetric matrix. A $2 * n * n$ matrix with a diagonal of 0 is constructed by clustering O and D points for n OD flows in previous flow clustering methods. However, in the ODFCVC method, the distance matrix size is:

$$C = n^2 + \sum_{i=1}^{k_s} {n_i}^2 \left(\sum_{i=1}^{k_s} n_i = n \right) \tag{10}$$

When constructing the distance of composite flow, the full dimension feature matrix generates unnecessary redundancy, because when the difference between one dimension is too large, there is no need to consider the similarity of the other dimensions. Therefore, through the gradual clustering of different dimensions, first, we cluster on the spatial feature dimension, grouping OD flows based on spatial clusters and unification of vector coordinate systems. Then, we cluster on geometric feature dimension, and extract representative vector features through adjusted cosine similarity in each spatial cluster. This method also improves the normalization error and local feature loss caused by solidifying vector features into four or eight directions in previous studies [55,56].

4. Experiments and Analysis

Section 4 introduces an example of traffic flow pattern mining using the ODFCVC method. This section contains three experiments. The first experiment is to analyze the OD data of taxis in Beijing by using the ODFCVC method. The second and third experiments are to analyze the spatial cluster and vector cluster generated by clustering.

4.1. Taxi OD Flow Clustering Based on ODFCVC

Taxi OD flow is a kind of trajectory with the location information of taxi passengers getting on and off by semantics extraction from taxi trajectory data generated by GPS positioning. Compared with the complex real trajectory, OD flow does not depend entirely on the real road network data and can directly reflect the characteristics of urban residents' travel. It is an important data source for mining the spatial and temporal activities of urban population [46] (pp. 60–61). The data we used in the experiment are some taxi GPS trajectory data (more than 12 150 pieces) from 6 a.m. to 9 a.m. on 11 January 2008 in Beijing. The data format is the raw taxi GPS trajectory data structure, including the taxi encryption number, GPS feedback time, real-time longitude and latitude, riding status, riding events, speed, direction angle, and other fields [46] (pp. 76–79). In order to facilitate the visualization of clustering results, we use JavaScript + HTML + CSS front-end web development technology to carry out all the experiments and use JavaScript language to write clustering and visualization methods. The visualization of taxi OD flow without clustering analysis is shown in Figure 8.

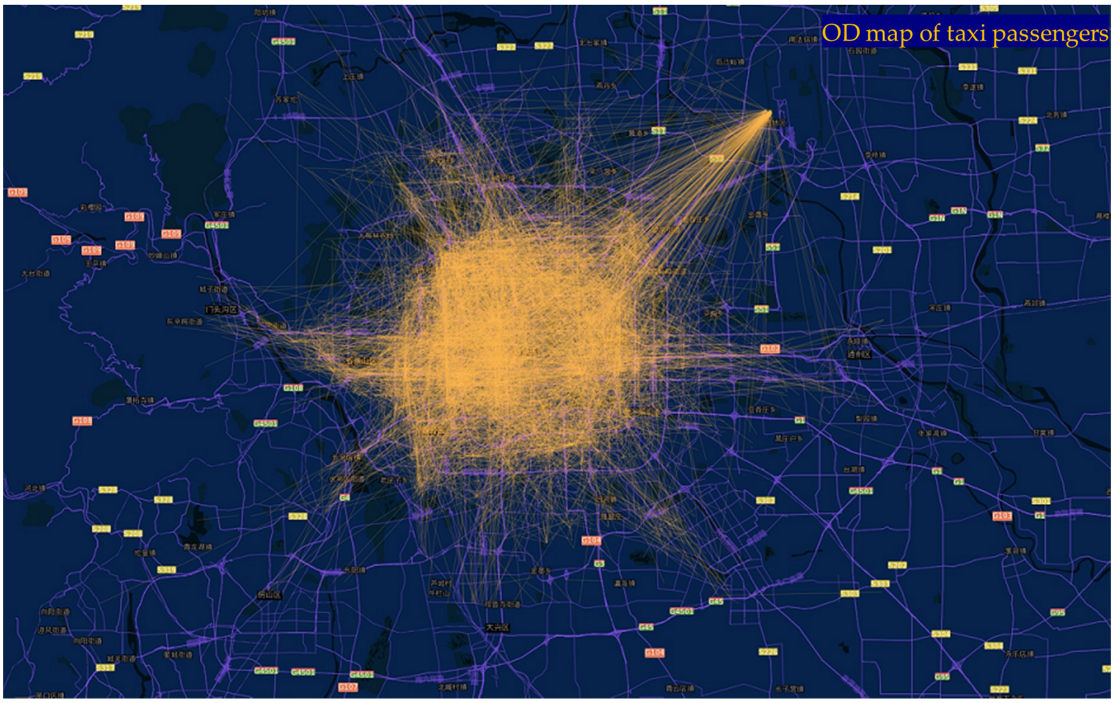


Figure 8. Unprocessed taxi data origin-destination (OD) flow map.

Through the automatic determination and test of the optimal K value, the K value of the spatial cluster based on OD flow event points is 4, and the K value of the vector cluster contained in each spatial cluster is 4, 4, 5, and 4, respectively.

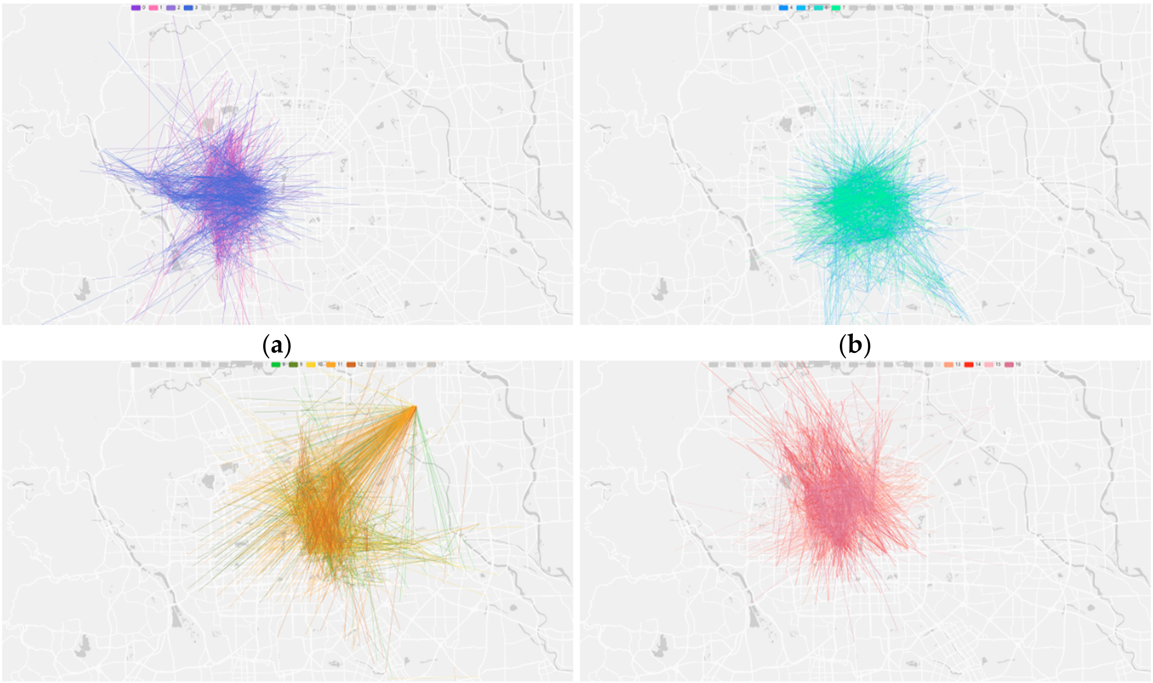
In order to verify the significance of the number k of clustering results solved by SSEPIP, we use the stability of clustering as the evaluation criteria [57]. The method of draw a random subsample of the original data set without replacement is used to generate perturbed versions (p1, p2, p3) of the dataset, and the sampling rate is 0.8. The distance function uses the minimal matching distance. The experimental results of clustering stability in the process of "space division" and "vector clustering" are shown in Table 1. The stability index shows the effectiveness of SSEPIP to solve the optimal clustering number automatically.

Table 1. Experimental results of clustering stability based on minimal matching distance.

k	First Step Clustering			Second Step Clustering [1]											
				C_S1			C_S2			C_S3			C_S4		
	p1	p2	p3	p1	p2	p3	p1	p2	p3	p1	p2	p3	p1	p2	p3
2	26	26	26	85	85	85	113	113	113	1	1	1	49	49	49
3	2782	2781	2781	168	168	168	732	661	784	6	6	6	944	944	50
4	47	47	47	148	154	148	134	134	134	6	64	6	108	108	118
5	51	51	51	129	129	132	370	370	370	74	38	38	246	246	246
6	231	231	48	471	471	471	322	179	409	87	87	87	613	613	613
7	67	67	56	693	721	420	157	159	159	536	536	536	342	405	652
8	2193	2193	1346	595	208	599	356	356	356	115	106	106	343	544	341
9	342	342	342	176	767	569	230	229	358	575	478	593	356	612	630
10	153	153	774	188	185	344	263	279	267	682	692	702	372	372	452

[1] In the first step clustering, the best cluster number K_S is 4.

The OD flow spatial cluster generated in the clustering process is shown in Figure 9, and the final clustering result is shown in Figure 10. In order to observe the overall flow trend of 17 clusters, we calculate the OD flow event point mean and OD flow vector mean of all kinds of clusters as representative visualization descriptive indicators. The visualization results are shown in the nested view in Figure 10.


(**a**) (**b**)

(**c**) (**d**)

Figure 9. Four taxi OD flow spatial clusters and communities based on OD flow event point clustering process. Spatial clusters (**a**,**b**,**d**) contain four vector clusters respectively, and spatial cluster (**c**) contains five vector clusters.

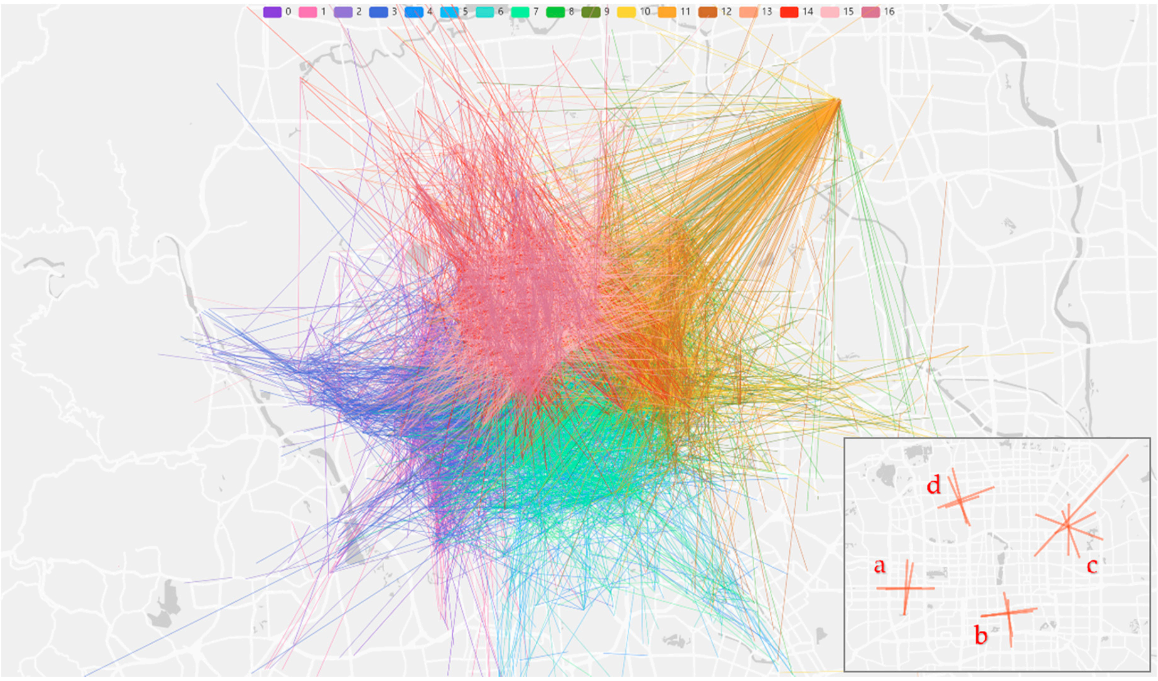


Figure 10. Clustering results of ODFCVC.

4.2. Spatial Cluster Analysis of Taxi OD Flow Clustering

Figure 9 is the visualization result of OD flow event point spatial clustering based on spatial partition index extraction after the ODFCVC method. By comparing with Figure 10, we see that the cluster product of the first step clustering "spatial partition" is the OD flow vector coordinate system constraint, and the OD flow cluster is obviously divided into four spatial clusters, each of which constrains different vector clusters. This compound flow model is mainly influenced by the calculation of OD flow event points. OD flow is not a real trajectory, and there is no midpoint of the trajectory. By defining the OD flow event point, the midpoint of the OD flow is regarded as the spatial abstraction of the OD flow. Therefore, the midpoint of OD flow has certain physical significance in clustering analysis and pattern recognition.

The original intention of spatial clustering based on OD flow event points is, on the one hand, to satisfy the similarity conditions in OD flow clusters in spatial dimension, and on the other hand, to simplify the amount of data when calculating geometric similarity and enhance the expression of local feature difference. However, whether OD flow spatial cluster has physical significance is worth our in-depth consideration. Therefore, we use community discovery algorithm in network analysis to realize OD flow communities based on different geographical units and try to understand the physical significance of taxi OD flow spatial cluster through comparative analysis.

We find communities in the graph constructed by OD flow using Clauset–Newman–Moore (CNM) greedy modularity maximization [58,59]. The corresponding process is shown in Figure 11. The network analysis function is realized by NetworkX software package, and the community visualization is realized by ArcMap. The taxi OD flow network is constructed by taking Beijing traffic zone and Beijing street unit as nodes [60]. Through overlay analysis of different geographic units and OD flow, the OD interactive graph of taxi trip based on geographic units is obtained. Then, we use the most classical module-based community discovery algorithm CNM to get the OD flow community without geographical space constraints, and visually display on a map. Figure 12 shows the community distribution of taxi OD flow network with traffic zone and block (street unit) as nodes, respectively. We annotate the six main communities (the number of nodes in the community is more than 10% of the total nodes), which is represented by ①–⑥.

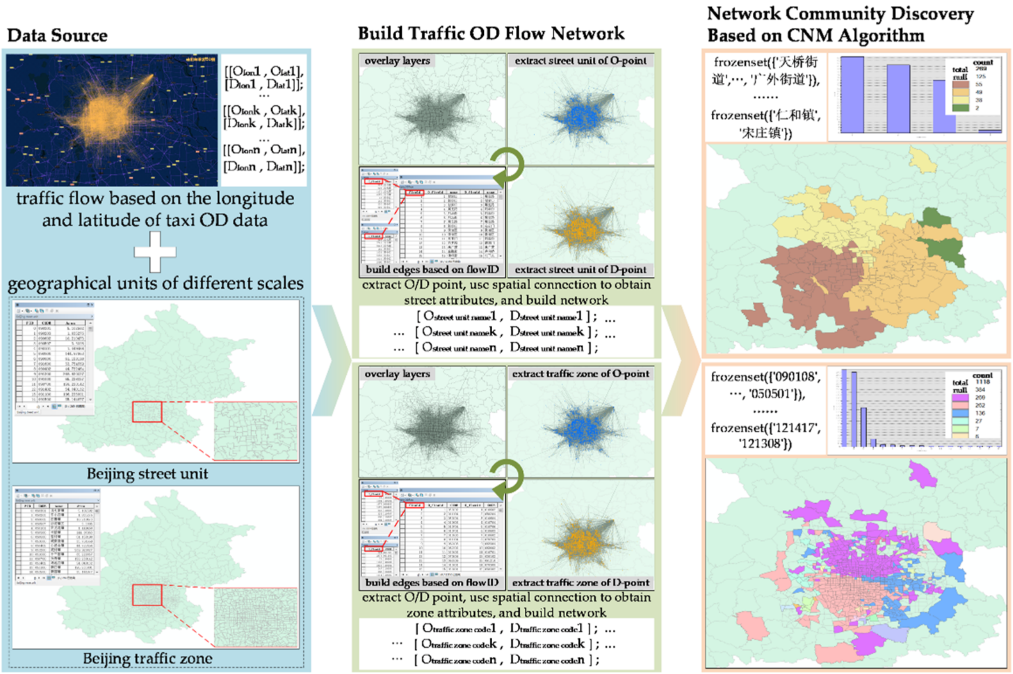
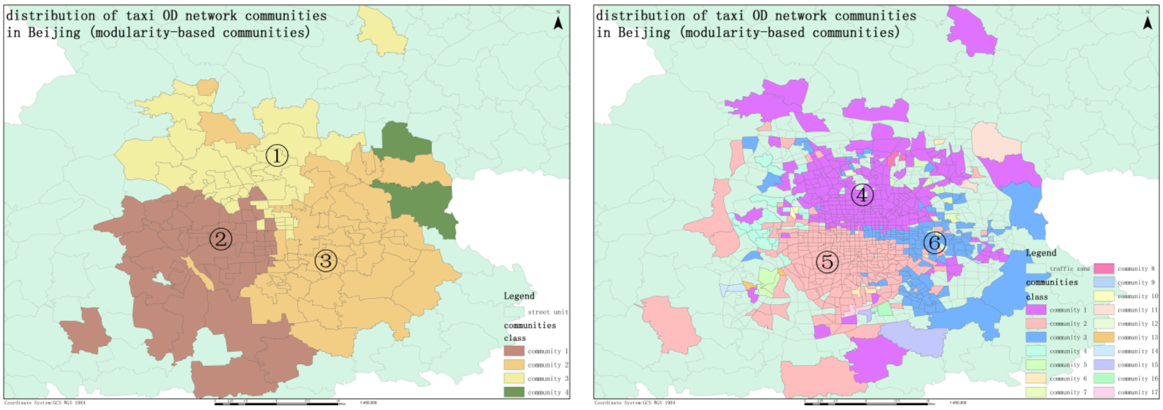

Data Source
traffic flow based on the longitude and latitude of taxi OD data
geographical units of different scales
Beijing street unit
Beijing traffic zone
Build Traffic OD Flow Network
overlay layers
extract street unit of O-point
build edges based on flowID
extract street unit of D-point
extract O/D point, use spatial connection to obtain street attributes, and build network
extract traffic zone of O-point
extract traffic zone of D-point
extract O/D point, use spatial connection to obtain zone attributes, and build network
Network Community Discovery Based on CNM Algorithm
frozenset({'天桥街道',⋯,'广外街道'}),
……
frozenset({'仁和镇', '宋庄镇'})
frozenset({'090108', ⋯, '050501'}),
……
frozenset({'121417', '121308'})
count
total
null

Figure 11. Discover communities using the Clauset–Newman–Moore (CNM) algorithm and visualize geographically.


distribution of taxi OD network communities in Beijing (modularity-based communities)
①
②
③
Legend
street unit
communities class
community 1
community 2
community 3
community 4
distribution of taxi OD network communities in Beijing (modularity-based communities)
④
⑤
⑥
Legend
traffic zone
communities class

(**a**) (**b**)

Figure 12. Taxi OD flow network communities based on (**a**) street unit and (**b**) traffic zone.

By comparing Figures 9 and 12, we analyze the relationship and difference between the OD flow spatial cluster extracted by the ODFCVC method and the OD flow network community obtained by classic network community algorithm. First of all, comparing the network communities mined by the CNM algorithm on different granularity of geographical units, there are some similarities and differences between them.

In terms of similarities, there are three main traffic flow communities, covering the central area of Beijing. The communities are located in the north, southwest, and southeast of the central urban area of Beijing. Each community has a certain degree of geographical spatial connectivity. In terms of differences, due to the more detailed division of traffic zones, the community formed is also more fragmented, and there are geographical gaps within the community. There are community boundary contradictions in the west, northeast, and south of Beijing central area.

Then, compared with Figures 9 and 12, it can be found that the spatial partition results based on the first step clustering of the ODFCVC method have strong similarity with the network community mining, and some interesting phenomena are found. The number of spatial clusters obtained by clustering is 4. Because the clustering results are based on the K-means algorithm, the spatial clusters have global characteristics and ignore local anomalies, therefore, the spatial clusters obtained by Euclidean distance clustering of OD flow event points have internal continuity. Spatial cluster (a) corresponds to the north of network community ②, spatial cluster (b) corresponds to the east of network community ⑤ and the south of community ② and ③, spatial cluster (c) corresponds to the intersection of network community ⑥ and community ③ and community ④, and spatial cluster (d) corresponds to network community ①. It can be seen that the OD flow spatial cluster obtained by clustering has certain practical significance. In the south of Beijing, where the network community is controversial, the new spatial cluster (b) obtained by the ODFCVC method is beneficial to explain and unify the network community obtained by different geographical units.

Using the OD flow midpoint as the event point is easier to find the flow clustered mode. The physical significance of taxi OD flow cluster is due to the attraction of urban functional areas and the social interaction of transportation hubs, resulting in traffic flow cluster with obvious spatial division. Because the formation of traffic flow community is dependent on the urban traffic hub, we compare the cluster result with the spatial distribution of Beijing traffic hub [61] and find that the two have obvious spatial correlation (Transport Center (a) includes Beijing West Railway Station and Liuliqiao Passenger Transport Hub. Transport Center (b) includes Beijing South Railway Station, Songjiazhuang Transport Hub, and Nanyuan Airport. Transport Center (c) includes Beijing Railway Station, Sihui Public Transport Hub, and Beijing Capital International Airport. Transport Center (d) includes Xiyuan Transport Hub). Through the spontaneous behavior of urban people's travel activities, urban traffic centers can be identified without the influence of functional areas of origin-destination points. It can also be found that the transport hub not only serves as a passenger flow distribution center, but also attracts the traffic interaction around it. Therefore, the ODFCVC method can better identify OD flow communities and discover OD flow clusters with potential spatial connections.

4.3. Vector Cluster Analysis of Taxi OD Flow Clustering

On the basis of the recognition of OD flow event point spatial cluster, this clustering method can find representative geometric feature clusters with arbitrary shape. In previous studies, the similarity of OD points is often constrained by defining regular search space, or additional geographic unit partition, or uniformly continuous density space, so as to obtain rule clusters with similar geometry, or irregular clusters with similar semantic features or uniform density of OD points. The morphological structure of these clusters depends on the parameter definitions of hierarchical and density-based clustering algorithms. However, due to the solidification of search radius, intra-cluster connectivity, and other parameters, existing aggregation algorithms cannot deal very well with the global density of non-uniform line sets.

Traditional methods pay more attention to "geographic attraction" to "flow behavior" [46] (pp. 111–114), but the ODFCVC method is the opposite. For taxi OD trajectory analysis, the spatial location and thematic attributes of OD points are static geoproblems, and the spatiotemporal trajectory generated by taxi activities is a problem of urban dynamics. In the past, people's dynamic spatiotemporal modeling and analysis focus more on the rigorous causal inference. Because the characteristics of OD point's functional area meet the law of urban activity, it will produce corresponding travel behavior. Therefore, when travel behavior has similar OD points, this kind of travel behavior is the same mode. This kind of analysis method which first has "geographic attraction" and then "flow behavior" has a great dependence on the accuracy of OD data, and is limited by the factors of urban functional areas and POI data updating, does not have the premise of initiative discovery of new patterns. In the data-driven way, the proposed method tries to find the geographic similarity attracting such behavior by analyzing the flow behavior and excavate the urban subspace interaction under the new urban science paradigm.

The whole process of the algorithm does not need any preset parameters. It only needs to calculate the optimal K value of the spatial cluster and the geometric cluster, respectively, by the silhouette coefficient, the sum of squares of errors, and other indicators. The first step of spatial clustering and adjusted cosine similarity solves the feature loss caused by global data normalization as far as possible.

Moreover, because of the geometric constraints, the method does not use the conventional two-point constraints and is only affected by the K value of geometric cluster based on spatial cluster optimization and adjusted cosine similarity, therefore, OD flows find clusters with irregular distribution. Therefore, the ODFCVC method recognizes not only clusters of similar patterns with regular shapes, but also clusters with convergence and divergence patterns. As shown in Figure 10c, the origin points of convergence pattern and destination points of evacuation pattern affected by traffic center do not have the characteristics of point similarity in traditional clustering method, that is, they do not have uniform connectivity density or similar spatial distance. However, using the ODFCVC method, we find the main traffic flow divergence and convergence modes based on the influence of Capital International Airport.

Since the ODFCVC method can find the convergence and divergence mode of the flow clusters, we also try to explain it by the point set density distribution in vector space. Figures 13 and 14 show the distribution of kernel density of OD point data in geographic space and vector space. The kernel density analysis and result visualization are realized by ArcMap.

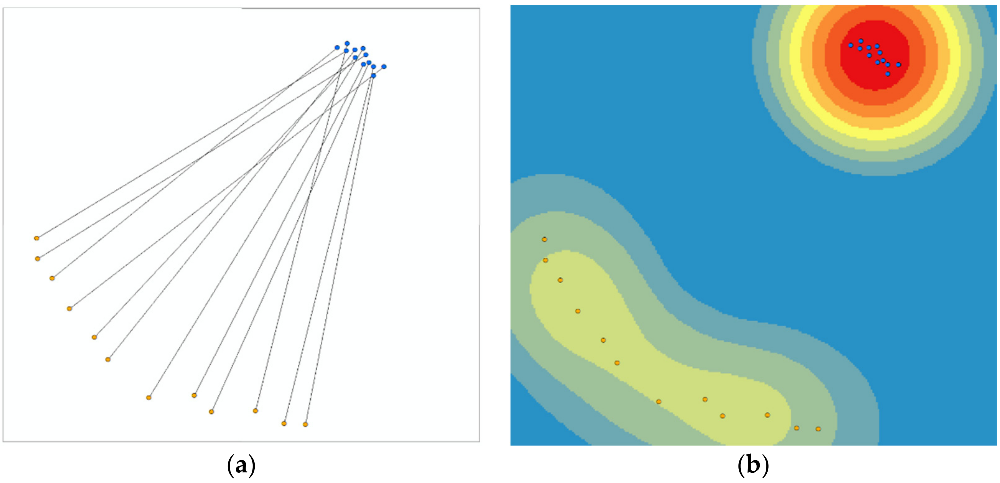


Figure 13. (**a**) OD flows and (**b**) kernel density distribution of OD points in geographical space.

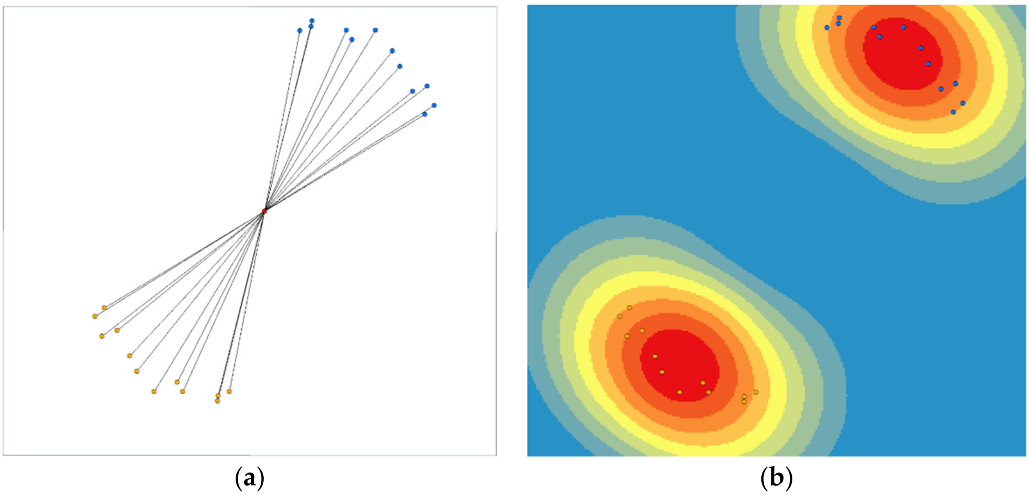


Figure 14. (**a**) OD flows and (**b**) kernel density distribution of OD points in vector space.

Density consistency is an important criterion for evaluating clustering results. In density-based clustering, OD flows with OD points in the same density space tend to be regarded as the same pattern. In geographical space, the O-point of divergence mode traffic flow and the D-point of convergence mode traffic flow are in high-density space, while the D-point of divergence mode traffic flow and the O-point of convergence mode traffic flow are in low-density space. Figure 13b shows the heterogeneity of the density distribution of OD points in geospatial space. Therefore, it is difficult to find these patterns automatically by using the traditional clustering methods of density and simple connectivity index.

In the algorithm flow proposed in this paper, OD flow is mapped to vector space, and the relative geographic location of OD points changes with the aggregation of OD flow event points. Because only the direction and size of vector are considered in the vector coordinate system, we can set the OD flow event point as the intersection point and move the relative position of OD flow vector. It can be found that the density distribution of OD points is relatively homogeneous. Figure 14b shows the homogeneity of the density distribution of OD points in vector space.

5. Conclusions and Discussion

In this paper, a two-step clustering method for OD data is proposed. The pattern characteristics of OD flow are represented by OD flow event points and OD flow vector, and OD data is mapped from massive point set space to independent vector feature space. This method simplifies the complexity of OD flow similarity calculation and pays more attention to the overall spatial distribution and movement trend of OD flow. Compared with the previous studies, the proposed method breaks away from the line clustering idea based on two-point clustering, pays more attention to the overall (high-dimensional) similarity of OD flows, optimizes the dimension of feature matrix in the clustering process, and achieves the automatic optimal clustering number calculation without any parameters. The ODFCVC method can mine arbitrary shape OD flow clusters with representative characteristics and find OD flow communities, which is conducive to optimizing traffic zone planning and analyzing OD flow dynamics problems.

The ODFCVC method can be combined with the existing research [15,38]. On the one hand, it can be compared with the results of previous algorithms to evaluate the dynamic functional area attributes of geographical units. On the other hand, it can be combined with density-based algorithm and partition-based algorithm to strengthen the recognition ability of density domain intensity and construct multiscale clustering results while maintaining the flow characteristics.

Through the experiment of Beijing Taxi OD data, the method mines out the important traffic centers and traffic flow communities affected by traffic hubs in Beijing without relying on geographical units, which makes up for the shortage of traditional traffic engineering and urban planning by using transport capacity, passenger flow, construction scale, spatial accessibility, and other indicators to evaluate the importance of traffic hubs. In addition, this method breaks through the limitations of the "parallel line" experience mode in the previous pattern mining, and finds the cluster mode (OD points are similar respectively), divergence mode (O points are similar), and convergence mode (D points are similar) of traffic flow at the same time, which is more suitable for the real traffic flow.

Through the kernel density analysis of OD points in geographic space and vector space, we found that the ODFCVC method can map the OD flow of irregular shape pattern to the vector space with homogeneous OD point density by using the method of geometric constraints, which meets the conditions of traditional OD flow density clustering. Therefore, the method based on density clustering can also be applied to pattern mining of OD flow in vector space.

The original intention of traffic flow pattern and OD flow pattern mining is not simply from data to pattern, different measurement functions and different indicators can produce a variety of flow patterns. However, how to use the patterns found by clustering algorithm reasonably and apply them to traffic planning, urban planning, and other fields is the value of research. Previous clustering algorithms rely heavily on the idea that "OD points are similar, so OD flows are similar", and deeply study different measurement indexes such as spatial similarity, thematic similarity, and land use type similarity of OD points [8,39]. The ODFCVC method does not rely on the similarity index of OD points, and can excavate representative flow patterns. It excavates the internal connection between OD points of nongeographic spatial similarity in the same pattern, and update and iterate from "observing the flow of human activities" to "observing the points of land use types". It also provides new means for the research of urban land use renewal, urban dynamic function area mining, urban internal space interaction, multiscale traffic district planning, and so on.

Information extraction is an important obstacle to traffic information services and applications. The division of traffic zones (communities) is an important component of traffic surveys, travel demand forecasting, trip generation, and trip distribution [62]. The traditional traffic zone division method cannot reflect the latest or real-time traffic patterns and the consistent characteristics within a traffic zone and ignores the mobility and community characteristics of traffic behavior [62,63]. Our research applies the ODFCVC method to traffic OD flow, which can identify traffic flow communities with frequent internal interactions and regional interaction behaviors with typical travel patterns. It provides a new means for dividing traffic zones and revealing the spatial structure characteristics.

The method also has strong expansibility. First, the similarity function, that is, different distance functions can be replaced according to the research requirements when measuring the similarity of spatial relations and geometric features of flows. Secondly, the basic clustering algorithm, that is, as long as the logic of multistep clustering based on dimension deconstruction is concerned, each step can be replaced by a clustering algorithm that relies on different clustering centers to meet the needs of researchers. Thirdly, the research object and analysis dimension, that is, for OD flow data, this paper only analyses the spatial dimension and dynamic feature dimension but does not consider the influence of time dimension. It can expand multidimensional data analysis by step clustering. For any high-dimensional geometric form such as area data and volume data, the geometric center and high-dimensional geometric vectors can be used to express spatial and morphological features. Finally, the computing environment, that is, because the whole operation process of the method presents a tree-like diffusion pattern, clustering analysis of each dimension can be computed distributed on the basis of the results of the previous clustering step.

However, the method still has some shortcomings in multilevel structure expression and similarity judgment. The ODFCVC method considers the complexity of OD flow clustering from the perspective of spatial location and geometric vectors, but there is no multiscale partition and pattern mining for any single dimension space, which is mainly due to the limitations of K-means algorithm. In the aspect of similarity threshold, we adopt the traditional silhouette coefficient and SSE to automatically obtain the optimal K value. However, these parameters are the optimal solution of spatial clustering for the whole sample and the optimal solution of vector space geometric feature clustering for the spatial cluster, which lack a certain prior constraint for similarity. Therefore, in future research, considering that the dataset is affected by the change of the field of view and scale, we plan to optimize the ODFCVC method to have the ability of multiscale analysis, and therefore mine multiscale compound flow patterns and complex flow patterns.

Author Contributions: Conceptualization, Xiaogang Guo and Zhijie Xu; methodology, Xiaogang Guo; software, Xiaogang Guo; validation, Xiaogang Guo; formal analysis, Zhijie Xu and Xiaogang Guo; investigation, Xiaogang Guo and Jianqin Zhang; resources, Jianqin Zhang and Zhijie Xu; data curation, Hao Zhang, Xiaogang Guo, and Jian Lu; writing—original draft preparation, Xiaogang Guo; writing—review and editing, Xiaogang Guo, Jian Lu, Jianqin Zhang, and Zhijie Xu; visualization, Xiaogang Guo; supervision, Jianqin Zhang; project administration, Jianqin Zhang and Xiaogang Guo; funding acquisition, Jianqin Zhang, Zhijie Xu, and Xiaogang Guo. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant No. 41771413, grant No. 41701473), the Beijing Natural Science Foundation (grant No. 8202013), the China Scholarship Council (grant No. 201909960002), and the BUCEA Post Graduate Innovation Project (grant No. PG2019007).

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments.

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

References

1. Guo, D.; Zhu, X. Origin-destination flow data smoothing and mapping. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 2043–2052. [CrossRef] [PubMed]
2. Liu, Y.; Wang, F.; Xiao, Y.; Gao, S. Urban land uses and traffic "source-sink areas": Evidence from GPS-enabled taxi data in Shanghai. *Landsc. Urban Plan.* **2012**, *106*, 73–87. [CrossRef]

3. Liu, X.; Gong, L.; Gong, Y.; Liu, Y. Revealing travel patterns and city structure with taxi trip data. *J. Transp. Geogr.* **2015**, *43*, 78–90. [CrossRef]
4. Stephen, D.M.; Jenny, B. Automated layout of origin–destination flow maps: U.S. county-to-county migration 2009–2013. *J. Maps* **2017**, *13*, 46–55. [CrossRef]
5. Jenny, B.; Stephen, D.M.; Muehlenhaus, I.; Marston, B.E.; Sharma, R.; Zhang, E.; Jenny, H. Force-directed layout of origin-destination flow maps. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1521–1540. [CrossRef]
6. Graser, A.; Schmidt, J.; Roth, F.; Brändle, N. Untangling origin-destination flows in geographic information systems. *Inf. Vis.* **2017**. [CrossRef]
7. Andrienko, G.; Andrienko, N. A General Framework for Using Aggregation in Visual Exploration of Movement Data. *Cartogr. J.* **2010**, *47*, 22–40. [CrossRef]
8. Wang, Z.; Yuan, X. Visual Analysis of Trajectory Data. *J. Comput.-Aided Des. Comput. Graph.* **2015**, *27*, 9–25.
9. Xin, R.; Ai, T.; Yang, W.; Feng, T. A New Network Voronoi Diagram Considering the OD Point Density of Taxi and Visual Analysis of OD Flow. *J. Geo-Inf. Sci.* **2015**, *17*, 1187–1195. [CrossRef]
10. Chen, Y.; Huang, Z.; Pei, T.; Liu, Y. HiSpatialCluster: A novel high-performance software tool for clustering massive spatial points. *Trans. GIS* **2018**, *22*, 1275–1298. [CrossRef]
11. Wang, S.; Du, Y.; Jia, C.; Bian, M.; Fei, T. Integrating algebraic multigrid method in spatial aggregation of massive trajectory data. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 1–20. [CrossRef]
12. Guo, D.; Zhu, X.; Jin, H.; Gao, P.; Andris, C. Discovering Spatial Patterns in Origin-Destination Mobility Data. *Trans. GIS* **2012**, *16*, 411–429. [CrossRef]
13. Zhu, X.; Guo, D. Mapping Large Spatial Flow Data with Hierarchical Clustering. *Trans. GIS* **2014**, *18*, 421–435. [CrossRef]
14. Pei, T.; Wang, W.; Zhang, H.; Ma, T.; Du, Y.; Zhou, C. Density-based clustering for data containing two types of points. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 175–193. [CrossRef]
15. He, B.; Zhang, Y.; Chen, Y.; Gu, Z. A Simple Line Clustering Method for Spatial Analysis with Origin-Destination Data and Its Application to Bike-Sharing Movement Data. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 203. [CrossRef]
16. Tobler, W. Experiments in migration mapping by computer. *Cartogr. Geogr. Inf. Sci.* **1987**, *14*, 155–163. [CrossRef]
17. Guo, D. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Trans. Vis. Comput. Graph.* **2009**, *15*, 1041–1048. [CrossRef]
18. Selassie, D.; Heller, B.; Heer, J. Divided edge bundling for directional network data. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 2354–2363. [CrossRef]
19. Verbeek, K.; Buchin, K.; Speckmann, B. Flow map layout via spiral trees. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 2536–2544. [CrossRef]
20. Nagel, T.; Maitan, M.; Duval, E.; Moere, A.V.; Klerkx, J.; Kloeckl, K.; Ratti, C. Touching transport a case study on visualizing metropolitan public transit on interactive tabletops. In *Proceedings of International Working Conference on Advanced Visual Interfaces*; ACM Press: New York, NY, USA, 2014; pp. 281–288. [CrossRef]
21. Boyandin, I.; Bertini, E.; Bak, P.; Lalanne, D. Flowstrates: An approach for visual exploration of temporal origin-destination data. *Comput. Graph. Forum* **2011**, *30*, 971–980. [CrossRef]
22. 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; pp. 51–58. [CrossRef]
23. Henry, N.; Fekete, J. Matrixexplorer: A dual-representation system to explore social networks. *IEEE Trans. Vis. Comput. Graph.* **2006**, *12*, 677–684. [CrossRef] [PubMed]
24. Wood, J.; Dykes, J.; Slingsby, A. Visualization of origins, destinations and flows with OD maps. *Cartogr. J.* **2010**, *47*, 117–129. [CrossRef]
25. Wood, J.; Slingsby, A.; Dykes, J. Visualizing the Dynamics of London's Bicycle-Hire Scheme. *Cartogr. Int. J. Geogr. Inf. Geovis.* **2011**, *46*, 239–251. [CrossRef]
26. Phan, D.; Xiao, L.; Yeh, R.; Hanrahan, P. Flow map layout. *IEEE Symp. Inf. Vis.* **2005**, 219–224. [CrossRef]
27. Holten, D. Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. *IEEE Trans. Vis. Comput. Graph.* **2006**, *12*, 741–748. [CrossRef]
28. Tao, F. Visual Analysis of Resident Trip Mode Based on Taxi OD Data. Master's Thesis, Wuhan University, Wuhan, China, 2017.

29. Slingsby, A.; Kelly, M.; Dykes, J.; Wood, J. OD Maps for Studying Historical Internal Migration in Ireland. In Proceedings of the IEEE Conference on Information Visualization (InfoVis), Seattle, DC, USA, 14–19 October 2012.
30. Adrienko, N.; Adrienko, G. Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 205–219. [CrossRef]
31. Wu, J.; Zhu, Y.; Ku, T.; Wang, L. Hot routes detection algorithm based on grid clustering. *J. Jilin Univ.* **2015**, *45*, 274–282. [CrossRef]
32. Zhang, X. *Data Clustering*, 1st ed.; Science Press: Beijing, China, 2018.
33. Kriegel, H.-P.; Kröger, P.; Zimek, A. Clustering high-dimensional data. *ACM Trans. Knowl. Discov. Data* **2009**, *3*, 1–58. [CrossRef]
34. Kernighan, B.W.; Lin, S. An Efficient Heuristic Procedure for Partitioning Graphs. *Bell Syst. Tech. J.* **1970**, *49*, 291–307. [CrossRef]
35. Barnes, E. An algorithm for partitioning the nodes of a graph. In Proceedings of the 1981 20th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes, San Diego, CA, USA, 16–18 December 1981; pp. 303–304. [CrossRef]
36. Aggarwal, C.C.; Yu, P.S. A Survey of Uncertain Data Algorithms and Applications. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 609–623. [CrossRef]
37. Zhang, X.; Zhang, X.; Liu, H.; Liu, X. Multi-Task Multi-View Clustering. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 3324–3338. [CrossRef]
38. Song, C.; Pei, T.; Ma, T.; Du, Y.; Shu, H.; Guo, S.; Fan, Z. Detecting arbitrarily shaped clusters in origin-destination flows using ant colony optimization. *Int. J. Geogr. Inf. Sci.* **2018**, *33*, 1–21. [CrossRef]
39. Gong, X.; Pei, T.; Sun, J.; Luo, M. Review of the Research Progresses in Trajectory Clustering Methods. *Prog. Geogr.* **2011**, *30*, 522–534. [CrossRef]
40. Liu, T.; Du, Q.; Mao, H. Spatial Similarity Assessment Model and Its Application in Line Groups. *Geomat. Inf. Sci. Wuhan Univ.* **2012**, *37*, 992–995. [CrossRef]
41. Zhou, X.; Miao, F.; Ma, H.; Zhang, H.; Gong, H. A Trajectory Regression Clustering Technique Combining a Novel Fuzzy C-Means Clustering Algorithm with the Least Squares Method. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 164. [CrossRef]
42. Lee, J.-G.; Han, J.; Li, X.; Gonzalez, H. Traclass: Trajectory Classification Using Hierarchical Region-Based and Trajectory-Based Clustering. *Proc. VLDB Endow.* **2008**, *1*, 1081–1094. [CrossRef]
43. Yang, S.; Bi, S.; Athanase, N.; Huang, T.; Wan, L. Spatial clustering method for taxi passenger trajectory. *Comput. Eng. Appl.* **2018**, *54*, 249–255. [CrossRef]
44. Pei, T.; Li, T.; Zhou, C. Spatiotemporal Point Process: A New Data Model, Analysis Methodology and Viewpoint for Geoscientific Problem. *J. Geo-Inf. Sci.* **2013**, *15*, 793–800. [CrossRef]
45. Pei, T.; Gong, X.; Shaw, S.-L.; Ma, T.; Zhou, C. Clustering of temporal event processes. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 484–510. [CrossRef]
46. Xiao, S.; Fang, Z.; Chen, B.; Yin, L.; Chen, J.; Yang, X. *Space-Time GIS Analysis of Urban Crowd Activities*, 1st ed.; Science Press: Beijing, China, 2018.
47. Li, Z.; Liu, Q.; Tang, J. Towards a Scale-driven Theory for Spatial Clustering. *Acta Geod. Cartogr. Sin.* **2017**, *46*, 1534–1548. [CrossRef]
48. Tang, P.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*; Addison Wesley Press: Boston, MA, USA, 2006.
49. Wu, G.; Zhang, J.; Yuan, D. Automatically Obtaining K Value Based on K-means Elbow Method. *Comput. Eng. Softw.* **2019**, *40*, 167–170. [CrossRef]
50. Zhang, Z.; Guo, X.; Zhang, K. Clustering Center Selection on K-means Clustering Algorithm. *J. Jilin Univ.* **2019**, *37*, 437–441. [CrossRef]
51. Fu, T.; Chung, F.; Luk, R.; Ng, V. Pattern discovery from stock time series using self-organizing maps. In *Workshop Notes of KDD2001 Workshop on Temporal Data Mining*; Springer: New York, NY, USA, 2001; pp. 26–29.
52. Tao, R.; Thill, J.-C. Spatial Cluster Detection in Spatial Flow Data. *Geogr. Anal.* **2016**, *48*, 355–372. [CrossRef]
53. Liu, Y.; Tong, D.; Liu, X. Measuring Spatial Autocorrelation of Vectors. *Geogr. Anal.* **2014**, *47*, 300–319. [CrossRef]
54. Yang, J.; Li, Y.; Cheng, W.; Liu, Y.; Liu, C. EKF–GPR-Based Fingerprint Renovation for Subset-Based Indoor Localization with Adjusted Cosine Similarity. *Sensors* **2018**, *18*, 318. [CrossRef] [PubMed]

55. Rey, S.J.; Murray, A.T.; Anselin, L. Visualizing regional income distribution dynamics. *Lett. Spat. Resour. Sci.* **2011**, *4*, 81–90. [CrossRef]
56. Andrienko, G.; Andrienko, N.; Fuchs, G.; Wood, J. Revealing Patterns and Trends of Mass Mobility Through Spatial and Temporal Abstraction of Origin-Destination Movement Data. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 2120–2136. [CrossRef]
57. Ulrike, V.L. Clustering Stability: An Overview. *Found. Trends Mach. Learn.* **2010**, *2*, 235–274.
58. Mark, N. *Networks: An Introduction*; Oxford University Press: Oxford, UK, 2011; p. 224.
59. Clauset, A.; Newman, M.E.J.; Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **2004**, *70*. [CrossRef]
60. Li, X.; Yang, X.; Chen, H. Study on traffic zone division based on spatial clustering analysis. *Comput. Eng. Appl.* **2009**, *45*, 19–22. [CrossRef]
61. Yang, G.; Song, C.; Pei, T.; Zhou, C.; Shu, H.; Zhang, J. Passengers' OD temporal-spatial distribution characteristics of the external traffic hubs in Beijing. *J. Geoinf. Sci.* **2016**, *18*, 1374–1383. [CrossRef]
62. Dong, H.; Wu, M.; Ding, X.; Chu, L.; Jia, L.; Qin, Y.; Zhou, X. Traffic zone division based on big data from mobile phone base stations. *Transp. Res. Part C Emerg. Technol.* **2015**, *58*, 278–291. [CrossRef]
63. Yildirimoglu, M.; Kim, J. Identification of communities in urban mobility networks using multi-layer graphs of network traffic. *Transp. Res. Part C Emerg. Technol.* **2018**, *89*, 254–267. [CrossRef]