Applied Soft Computing Journal 163 (2024) 111912

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Statistical and density-based clustering of geographical flows for crowd movement patterns recognition

Jianbo Tang [a,b,1], Yuxin Zhao [a], Xuexi Yang [a,b,*], Min Deng [a,b], Huimin Liu [a,b], Chen Ding [a], Ju Peng [a], Xiaoming Mei [a]

[a] Department of Geo-informatics, Central South University, Changsha, Hunan 410083, China
[b] Hunan Geospatial Information Engineering and Technology Research Center, Changsha 410083, China

HIGHLIGHTS

- Local spatial autocorrelation analysis is extended for geographical flow data to adaptively determine the paprameters.
- Statisical constraints are used and the significance of each flow cluster is evaluated with Monte Carlo permutation test.
- The proposed method can find natural clusters in geographical flow data, avoiding the false clusters generated by chance.

ARTICLE INFO

Keywords:
Flow clustering
OD flow
Hotspot detection
Statistical significance of clusters
Movement patterns

ABSTRACT

With the rapid development of sensors and communication technologies, it has become easy to collect large-scale and long-term crowd movement positioning data, which brings new opportunities for studying crowd movement patterns. This paper proposes a novel Statistical and Density-Based Clustering algorithm (SDBC) to identify implicit significant spatial aggregation patterns in geographical flow data. Unlike existing flow clustering algorithms, this method identifies the hot spots of origin-destination (OD) flows based on local spatial statistics and density-growing clustering. It also evaluates the significance of the identified geographic flow clusters, effectively reducing the identification of spurious clusters generated by chance in data. In our method, the spatial neighborhood of each flow is first obtained based on spatial proximity, temporal similarity, and directional similarity. Then, the number of flows in the spatial neighborhood of each flow is calculated and used as the density measure. Based on this, high-density flows are automatically detected using local spatial aggregation statistics, and a hierarchical density-based clustering strategy is developed to merge adjacent high-density flows to generate candidate flow clusters. Finally, we perform permutation tests to infer the statistical significance of each flow cluster and eliminate the candidate clusters generated by chance. Experiments on synthetic data and real-world taxi trajectory data were conducted to evaluate the effectiveness of the proposed method. Results show that the proposed method can accurately identify the statistically significant flow clusters of different shapes and densities and performs better than the available state-of-the-art flow clustering algorithms.

1. Introduction

The movement of urban crowds from one geographical region to another form the geographical flows in urban space. For example, the traffic flows are formed by people from their residences to their workplaces and taxis picking up and dropping off passengers from the boarding points [1]. These flows can be represented as a conceptual model containing an origin and a destination and are usually represented as origin-destination (OD) pairs in the literature [2]. Detecting spatiotemporal clusters in OD flow data is of great value in revealing the human mobility patterns in space and can provide decision-making support to applications such as urban planning, traffic management, and intelligent recommendation.

In recent years, with the emergence of various mobile devices and the rapid development and popularization of mobile positioning technology, a large amount of mobile trajectory data has been generated,

* Corresponding author at: Department of Geo-informatics, Central South University, Changsha, Hunan 410083, China.
E-mail address: yangxuexi@csu.edu.cn (X. Yang).
[1] ORCID: https://orcid.org/0000–0002-9780–0439

https://doi.org/10.1016/j.asoc.2024.111912
Received 3 September 2023; Received in revised form 21 May 2024; Accepted 17 June 2024
Available online 28 June 2024

such as mobile signal data, vehicle trajectories, Wi-Fi positioning trajectory data, and Bus IC card records. With the increasing advancement of location-aware devices and technologies, it is becoming easier to obtain a large amount of geo-referenced flow data [3]. These data are often related to the movement of individuals. Although the movement characteristics of different individuals have particular differences, the group formed by individuals may have common behavioral characteristics and show relatively stable movement patterns in space [4]. Therefore, generalizing individual mobility as group mobility can demonstrate macro-geographical connections and identify implicit human mobility patterns [5]. Murray et al.[6] visualized the flow data on maps and explored the spatial interaction patterns between different urban regions. Clustering is essential for detecting implicit groups or aggregation patterns in geographic big data. Scholars have developed some representative spatiotemporal clustering algorithms for detecting clusters in geographic flow data. For example, Zhu and Guo [5]developed a flow hierarchical clustering method to aggregate similar flows for mapping extensive spatial flow data. Tao and Thill [7] extended the classical local K-function to detect the spatial clusters in flow data. Song et al. [8] proposed a spatial scan statistical method based on ant colony optimization to identify arbitrarily shaped clusters in flow data. Similarly, Liu et al. [9] presented a shared nearest-neighbor-based clustering method for detecting OD flow clusters with inhomogeneous densities. Yan et al. [1] recently developed a spatiotemporal flow L-function to detect the flow clusters in geographical flow data by considering spatial attributes and temporal information. The flow clusters related to commuting, entertainment, tourism, and transportation may reflect underlying crowd travel demands and movement patterns on weekdays and weekends [10,11]. The detection of geographical flow clusters has attracted the attention of many scholars, and several geographical flow clustering methods have been developed.

Most existing geographical flow clustering methods are extensions of traditional spatial point clustering algorithms. To identify flow clusters with different shapes and densities, many clustering strategies have been proposed. However, most of these clustering algorithms rely heavily on parameter settings (such as the density threshold), making it challenging to identify natural clusters in real-world flow data without prior knowledge. Based on the spatial distribution characteristics of flow data, some spatial statistics-based methods, such as K-function [7], L-function [1,12], and scan statistics [8] have been proposed to detect the flow clusters, reducing the dependence on manual parameter settings. Although existing methods can effectively identify flow clusters with different shapes and densities, selecting optimal parameters and evaluating the significance of the flow clusters are still worth further investigation. The lack of evaluation of the significance of clusters may lead to the identification of spurious clusters generated by chance in data. Therefore, this paper proposes a statistical and density-based clustering algorithm to identify significant clusters in geographical flow data. Compared with related studies, the main contribution of this study lies in:

(1) Based on the idea of density-growing clustering, the proposed method inherits the advantages of density-based clustering that can find arbitrary-shaped clusters and eliminate the noise. To reduce the difficulty of parameter settings in density-based clustering, such as the thresholds to determine the high-density entities, we extend the local spatial aggregation statistics (i.e., Getis-Ord G_i^* statistic) for geographical flow data to detect the local aggregated high-density flows, with which the density thresholds can be determined adaptively.
(2) The method in this paper constructs some statistical constraints for aggregating high-density flows into candidate clusters and evaluates the statistical significance of the candidate clusters using Monte Carlo permutation tests, which can avoid the identification of some false clustering patterns generated by noises in the flow data and ensure the reliability of clustering results.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the details of the proposed method. Experiments and the comparative analysis are presented in Section 4. Section 5 gives a case study of our method on real-world taxi trajectory data to reveal the crowd movement patterns. Section 6 summarizes this study and outlines the future work.

2. Related work

With the rise of human mobility research, many scholars have conducted in-depth research on detecting clustering patterns in geographical flow data and made remarkable progress. Currently, the research on OD flow clustering can be divided into two main directions according to the flow data used. One is to find the hotspot interactive patterns or flow clusters using the region-aggregated flow data, in which the individual OD flows are first matched to the demarcated geographical units (such as street blocks or partition grids). Then, the number of OD flows between two geographical units is counted. A volume threshold is used to determine which unit pairs are hotspots. In these studies, the flows are aggregated into regions of different scales, which may lead to the modifiable areal unit problem (MAUP). At the same time, some short-distance flows may be ignored. For these reasons, some researchers detect flow clusters directly based on individual flows without aggregating the flow data first, which is the other research direction of geographical flow clustering. The study of this paper also belongs to the latter. Existing clustering methods of geographical flows can be roughly divided into three categories: density-based, hierarchical, and spatial statistics-based flow clustering methods.

2.1. Density-based flow clustering

Density-based flow clustering methods extend concepts such as spatial neighborhood, density, core objects, and density connection for geographical flows. The main idea of density-based clustering methods is to identify flow clusters in data space that are contiguous regions of high-density flows, separated by other low-density regions. The low-density flows that are not in any cluster are considered as noise. In recent years, three of well-known density-based clustering algorithms, namely Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [13], ordering points to identify the clustering structure (OPTICS) [14] and density domain decomposition (DECODE) [15], have been extended to identify flow clusters [7,16,17]. These density-based methods can be used to detect flow clusters of arbitrary shapes. However, due to the sensitivity of clustering parameters (such as density threshold), identifying clusters with heterogeneous densities remains challenging, especially without prior knowledge [17]. Among these methods, density-domain decomposition methods can be used to automatically determine clustering parameters and more accurately identify flow clusters of arbitrary shapes and densities. However, limited by the assumptions of the decomposition algorithm, it is still complicated to distinguish the variable density flows in a cluster into feature and noise. The parameter estimation process of the density-domain decomposition method is time-consuming (traversing all possible parameters), leading to low efficiency for massive geographical flow data. The OPTICS flow clustering algorithm can identify clusters with different densities; however, the quality of clustering results is greatly affected by different parameter settings, such as density and spatial neighborhood thresholds. Although density-based clustering methods are effective for finding flow clusters of arbitrary shapes, these methods still exhibit some limitations. Firstly, the density-based clustering methods are sensitive to parameters, and the parameter selection is usually empirical. Secondly, density connection-based clustering strategies are susceptible to heterogeneous density and weak connectivity, making it easy to merge nearby high-density clusters incorrectly.

2.2. Hierarchical clustering

Hierarchical clustering methods for flow data are usually simple extensions of the traditional spatial hierarchical clustering methods for point data. In these methods, the similarity or distance between two flows is usually first defined based on the spatiotemporal proximity and direction consistency of these flows, and then uses a step-by-step strategy to gradually aggregate two flows with high similarity from the data space to generate a hierarchical clustering structure or conversely, uses a top-down strategy to divide the entire flow data into different subclusters [18]. Zhu and Guo [5] aggregated flows from similar origins to similar destinations into clusters with an extended hierarchical clustering method. Yao et al. [19] propose a stepwise spatiotemporal flow clustering method considering the temporal and geometric characteristics of the OD flows. An advantage of hierarchical clustering methods is that the clustering patterns at different scales are retained, and flow clusters with different densities can be found under different level thresholds. Recently, Tao and Thill [20] proposed a flow clustering method named flow-HDBSCAN, which combines the advantages of density-based clustering and hierarchical clustering and can reveal the internal data structure of flow clusters.

2.3. Spatial statistics-based clustering

Spatial statistics-based clustering methods, which extend from traditional statistical indices for point data, include Moran's I and its local version [21], Getis-Ord G statistic [22], and Ripley's K-function [7]. These methods utilize spatial statistics, such as K-function, to describe global or local clustering characteristics of univariate, bivariate, and network-constrained flows [23,24]. Some classical spatial statistics, such as Moran's I statistic [25], have been applied to measure the spatial autocorrelation of flows through the similarity relationship between flows. Several studies have shown that using Moran's I to measure spatial autocorrelation between residuals of spatial interaction models can improve the prediction performance of flows. Liu et al. [21] extended Moran's I statistic to the flow space to measure the flow's global and local spatial autocorrelation by defining the flow's spatial correlation as a vector dot product. Vectors with high local correlation indices are identified as flow clusters. However, this method cannot determine the location of those identified flow clusters. Similarly, Tao and Thill [7] combined the local K-function with hotspot identification to detect anomalous flow clusters by measuring the spatial similarity between flows. On this basis, Tao et al. [23] further utilized the Cross K-function and Flow K-function to propose a spatial statistics method called Flow Cross K-function to detect global and local bivariate patterns of geographical flows. Recently, a spatial L-function, derived from the K-function, has been developed as a clustering method to identify distinct clusters by estimating cluster scales and grouping the flows inside them. However, this method is limited to weakly separated spatial clusters and is susceptible to noise between clusters [12]. The above methods identify the aggregation patterns in geographical flow data according to the statistical distribution characteristics of the data. These methods are usually insensitive to parameters and even employ nonparametric techniques. However, these methods are based on specific prior statistical distribution assumptions, and their applicability to diverse data needs to be improved. Another representative statistic-based approach is the scanning method, which identifies distinct clusters by locating significant clusters with high likelihood ratios using a moving and size-changing window in space [8,26,27]. Gao et al. [26] extended spatial scan statistics [28] to identify flow clusters at different scales. In addition, Song et al. [8] detect OD flow clusters of arbitrary shapes using ant colony optimization in a spatial scanning statistical method by replacing the initial data model. Further, to detect multiple significant flow clusters of arbitrary shapes, Tao and Thill. [29] applied the multi-directional optimal ecological community-based algorithm (AMOEBA), a data-driven and bottom-up approach for flow clustering, to identify anomalous spatial interaction regions with high-value or low-value in large flow datasets. This method creates a spatial weight matrix of flow data based on the identified flow groups. The algorithm extends AMOEBA to geographical flow data by defining appropriate spatial flow neighborhoods. Inspired by this, Liu et al. [27] proposed an innovative bivariate flow clustering method based on AMOEBA. They defined a bivariate local Getis-Ord statistic in an iterative process to detect irregularly shaped flow clusters. These scanning methods can find statistically significant clusters in flow data. However, there are also some drawbacks to these approaches. Firstly, the predefined window shape may lead to identified clusters with regular shapes. Secondly, since many scanning windows of different sizes, shapes, and directions need to be tested one by one, the efficiency of the scanning method is relatively low when the data volume is large.

In summary, most existing flow clustering methods are extensions of traditional spatial point clustering algorithms. Although spatial flow clusters of different shapes and densities can be identified, the existing methods still have the following shortcomings that need to be improved:

(1) Existing methods based on density clustering can find flow clusters of different densities and shapes; however, they are usually sensitive to parameters. In practical applications, it is challenging to select the optimal parameters without prior knowledge;
(2) The existing research mainly focuses on the discovery of clusters in flow data, but there is often a lack of evaluation of the statistical significance of the identified clusters, which may lead to the identification of some spurious clusters that occurred by chance. Although some scanning methods can test the significance of clusters, most of these methods are based on greedy searching and optimization algorithms, and these methods are usually unsuitable for large-volume flow data because of low efficiency.

Except for the abovementioned research, deep learning-based clustering methods for detecting crowd movement patterns have gained attention recently. For example, researchers have proposed crowd flow detection methods using fully convolutional networks (FCNs) to identify and track the movement of dense crowd centers in drone-captured video sequences, which shows high efficiency and accuracy in managing large crowds and complex scenarios, making it ideal for applications in smart cities, traffic management, and security surveillance [30]. Compared with traditional clustering algorithms, deep learning-based clustering methods have higher adaptability and can automatically represent and learn complex data features, alleviating the problems of traditional clustering algorithms in optimal parameter selection and noise data processing. However, deep learning-based clustering methods also have certain shortcomings, such as the need for a large amount of training data and the poor interpretability of the models. To solve the limitations of existing methods, this paper combines the advantages of multiple clustering strategies and identifies OD flow clusters based on local spatial statistics and density-growing clustering. The proposed novel clustering method improves the adaptability for different datasets and the reliability of the clustering results. Details of the proposed method are as follows.

3. Methodology

This paper proposes a novel statistical and density-based clustering method to identify the significant flow clusters in geographical flow data. The proposed method consists of three steps, including *flow density measure definition* and calculation, *identification of high-density flows*, and flow *clustering with statistic constraint*. The overview of the proposed method is shown in Fig. 1. The method first obtains the spatial neighborhood of each flow according to spatial proximity, temporal similarity, and directional similarity between the flows. Based on this, the number of flows in the spatial neighborhood of a flow is counted and used as the density measure of this flow. Then, a spatial hotspot

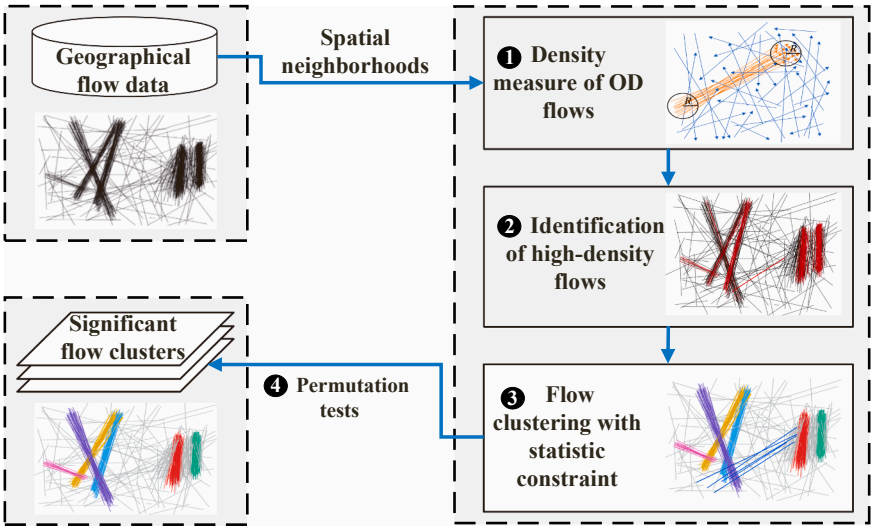


Fig. 1. The overview of the proposed method.

detection algorithm for flow data is applied to find high-density flows (denoted as seeds) using the Getis-Ord's G_i^* statistic. Further, a statistical constrained clustering algorithm is developed to combine the neighboring high-density flows to generate candidate flow clusters. Finally, the statistical significance of each flow cluster is evaluated based on Monte-Carlo permutation tests, and clusters that are not significant are removed.

3.1. Spatial neighborhood and density measure of OD flows

In this paper, a geographical flow is represented as an origin-destination (OD) pair, denoted as $l_i = (O_i, D_i)$, in which O_i is the origin point and D_i is the destination point. Spatial neighbors of a flow (say l_i), are usually defined as flows in the data space whose O points and D points are simultaneously located in the buffer zones of the O point and the D point of flow l_i, as shown in Fig. 2. The sizes of the buffer zones are controlled by two distance parameters, R_O and R_D, and in general, $R_O = R_D$ (hereafter, we use R to represent these two distance parameters). The set of spatial neighbors of a flow is defined as the spatial neighborhood of the flow. In this paper, we obtain the spatial neighborhood of each flow based on spatial proximity, temporal similarity, and directional similarity. For a flow l_i, the flow l_j is the spatial neighbor of l_i if it satisfies the following conditions:

1) *Spatial proximity*: the O point and the D point of flow l_j are in the buffer zones of the O and D points of flow l_i respectively. In the experiment, the buffer size R is set to 300 m by default according to the average size of our interested urban areas, such as the working places, traffic hubs, residential areas, shopping malls, etc.;
2) *Temporal similarity*: there is an intersection between the time periods from the origins to the destinations of the two flows, l_i and l_j;
3) *Directional similarity*: the angle between l_i and l_j is less than the angle threshold θ. The flow can be expressed as a vector from the O point to the D point, and the angle between two flows is calculated by the angle of these two vectors. According to previous studies, the angle threshold θ is set to 30 degree by default.

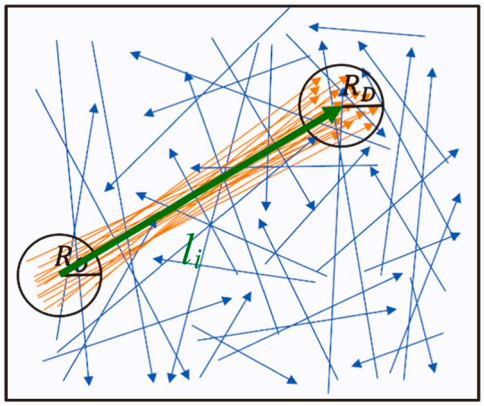


Fig. 2. Illustration of the spatial neighborhood of an OD flow.

Based on the spatial neighborhood of each OD flow, we define the number of flows in the spatial neighborhood of a flow as the density of this flow. The higher the density of a flow, the more aggregated flows around it.

The parameter R is vital for determining the density of the flow, as it sets the local scale for analysis when calculating the flow density. If R is small, the buffer area is limited, leading to the consideration of only those flows that are very close as spatial neighbors. This could result in an underestimation of the flow's spatial density, especially in scenarios where flow distribution is relatively uniform. Conversely, if R is set too large, the spatial density of the flow could be overestimated, as it may include too many irrelevant flows, blurring the distinction between cluster features and noise. In this study, we propose a method for the optimal parameter selection based on spatial proximity density variance, to choose appropriate values for R and θ. This method is based on the following observations.

(1) When R is small, the spatial neighborhood of a flow might only include a few neighboring flows, failing to reflect the overall spatial clustering characteristic of the flow. Moreover, since the calculation of spatial density is based on the number of neighbors within a smaller range, the spatial density of the flow tends to be unstable with high variance, where even minor changes in flow can lead to significant fluctuations in spatial density variance.
(2) As R increases, the variance in spatial density might begin to decrease, as the spatial neighborhood of the flow includes more neighboring flows, making the local density more stable and consistent. However, with a further increase in the value of R, the density variance might increase again, reflecting the increased likelihood of including irrelevant flows or noise.
(3) When R is overly large, almost all flows are included within a broad buffer area, leading to a blurring of the true spatial clustering characteristics of the flow. The variance in spatial density

might drop to very low levels, but it is limited. In such cases, low variance might no longer indicate stability in flow clustering but rather result from an overly broad neighborhood.

In this study, the impact of the parameter R (buffer size) on the spatial density of the flow is analogous to the role of k in the kth-nearest neighbor method for estimating spatial density, with both used to determine the neighborhood of a given object in space. The difference lies in that, in the kth-nearest neighbors method, the size of the neighborhood is defined by the number of neighbors k, whereas in the proposed method, the neighborhood is defined by buffer size R. Pei et al. [31] proposed a nonparametric index defined as the variance of the $(k+1)$-th nearest distance to the k-th nearest distance (RKD) to capture the variance of the kth nearest distance with different k values, which can further help to estimate the appropriate value of k.

Based on the analysis above, we use an index to estimate the optimal values of the spatial proximity parameters R and θ, which can be named the Ratio of spatial proximity Density Variance (RDV). It is achieved by comparing the changes in the spatial neighborhood density variance of flows under different R and θ values. The specific formula is as follows:

$$\mathrm{RDV} = \frac{Var_{R+\Delta R,\theta+\Delta\theta}}{Var_{R,\theta}} \Big/ R_C \tag{1}$$

where $Var_{R,\theta}$ is the spatial neighborhood density variance of flows under the buffer size R and angle threshold θ; $Var_{R+\Delta R,\theta+\Delta\theta}$ is the spatial neighborhood density variance of flows under the buffer size $R+\Delta R$ and angle threshold $\theta+\Delta\theta$; R_C is a constant representing the expected ratio of variance change in the spatial neighborhood density of flows under a certain range of R and θ values, used to help standardize the calculation of RDV values, making it more accurately reflect the changes in spatial neighborhood characteristics of flows under different spatial proximity parameters.

As displayed in Fig. 3, we use a simulated OD flow dataset to illustrate the selection of the optimal R and θ values based on the RDV value. The simulated OD flow data is displayed in Fig. 3(a), and the data contains a large volume of noise. Regarding the average size of our interested urban areas, such as workplaces, traffic hubs, residential areas, shopping malls, etc., we set the buffer size R range between [150,500] meters, with ΔR set to 10 m; based on the definition of

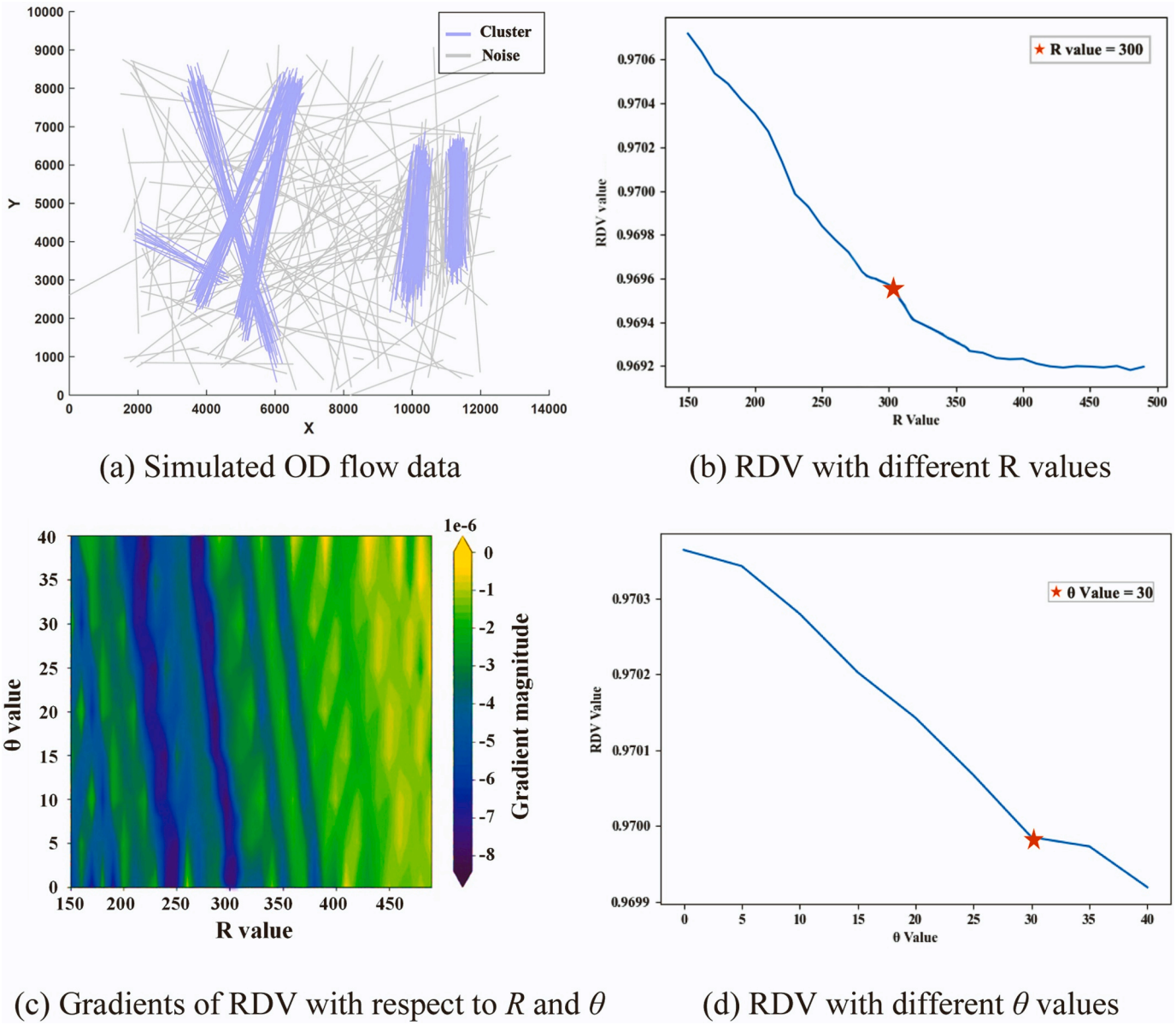


Fig. 3. Estimation of the appropriate R and θ values.

directional similarity, we set the range of θ between [0,45], with $\Delta\theta$ set to 5. In Fig. 3(b), the gradient map illustrates that regions with the least RDV growth align with the darkest stripes, indicative of gradients near zero. Notably, two such vertical stripes are prominent near $R = 250$ m and 300 m, suggesting stabilization of RDV values at these radial distances over the entire θ range. This is further corroborated by Fig. 3(c), where the RDV curve with respect to R flattens near $R = 300$ m, highlighting a potential optimal R value for stable RDV measures. Fig. 3(d) complements this by showing that at R=250 m, an θ of 30 degrees results in a minimal sensitivity of RDV to changes. Therefore, the appropriate R value can be set to 300 m, while the optimal value for θ can be set to 30 degrees. Under these parameter settings, the proposed method can completely and accurately identify the clusters within the simulated data.

3.2. Identification of high-density flows

After the above calculations, we obtain the spatial neighborhood and the density of each flow. Given a geographical flow data set $\boldsymbol{L} = \{l_1, l_2, l_3 \ldots, l_N\}$, and let $\{z_1, z_2, z_3 \ldots, z_N\}$ be the densities of flows in $\boldsymbol{L}$. To find the high-density flows, Getis-Ord's G_i^* statistic is used to identify the local aggregated high-density flows. For a given flow l_i, Getis-Ord's G_i^* statistic is calculated as:

$$G_i^* = \frac{\frac{\sum_{j=1}^{N} w_{i,j}(z_j - \overline{Z})}{S}}{\sqrt{\frac{\left[N\sum_{j=1}^{N} w_{i,j}^2 - \left(\sum_{j=1}^{N} w_{i,j}\right)^2\right]}{N-1}}} \tag{2}$$

where N is the total number of flows in $\boldsymbol{L}$, z_j represents the density of the flow l_j, and w_{ij} is defined as an indicator function that is one if flow l_j exists in the spatial neighborhood of flow l_i and zero otherwise. $\overline{Z}$ and S represent the mean and standard deviation of the densities of all the flows, respectively. The expected value of Getis-Ord's G_i^* statistic is 0, which means that there is not concentrated in high- or low-attribute values. Ord and Getis [32] showed that the G_i^* statistic is approximately normally distributed, and the G_i^* statistic returned for each flow is a z-score value showing where features either high value or low value cluster based on neighboring features. The z-score is a standard deviation of the standard normal distribution and very high z-scores indicating the existence of the clustering pattern associated with very small p-values are found in the right tails of the normal distribution. A larger (positive) value of the G_i^* statistic of a spatial flow, a higher concentration of spatial flows around this flow.

For each flow l_i in $\boldsymbol{L}$, if the G_i^* statistic value (i.e., z-score value) of flow l_i is greater than a threshold value λ, the flow is then identified as a high-density flow (also referred to as a *hotspot*). The identified high-density flows will be further used as *seed flows*. Each of them will be initialized as an initial flow cluster which grows and merges to form the candidate clusters. In applications, typical values of λ are 1.65, 1.95 or 2.58 with confidence levels of 90 %, 95 %, or 99 % respectively. A confidence level of 99 % is the most conservative, indicating that the probability of the observed density value occurred by random chance is very small (less than a one percent probability). For data with a large amount of noise, a more critical λ value could be used with a higher confidence level. In this paper, to balance the accuracy and the computational efficiency, λ is set to 0 according to experiments and previous studies.

3.3. Candidate flow clusters detection by clustering with statistic constraint

To discover candidate clusters, a statistical clustering algorithm is proposed. Based on the detected high-density flows (or flow-seed) in the previous step, we initialize each *seed flow* as a flow cluster, represented by $\boldsymbol{C} = \{\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3 \ldots, \boldsymbol{C}_m\}$, where m is the number of detected high-density flows. Here we define several concepts for each initialized cluster.

Definition 1. (Cluster's member): For the initialized cluster $\boldsymbol{C}_i$, the member of the cluster $\boldsymbol{C}_i$ contains only the flow l_i at the beginning. If two clusters $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$ be merged, the members of the merged cluster contain all the members of the cluster $\boldsymbol{C}_i$ and cluster $\boldsymbol{C}_j$.

Definition 2. (Cluster's neighborhood): For the initialized cluster $\boldsymbol{C}_i$, its neighbors are the flows in the spatial neighborhood of its member l_i. If cluster $\boldsymbol{C}_i$ and cluster $\boldsymbol{C}_j$ be merged, the neighbors of the merged cluster $\{\boldsymbol{C}_i, \boldsymbol{C}_j\}$ are the union of the neighbors of $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$. Note that, the neighbors of the merged cluster exclude its members.

Definition 3. (Cluster's density): The number of flows within the neighborhood of a cluster.

The G_i^* statistic of a flow cluster $\boldsymbol{C}$, denoted as $\boldsymbol{C} = \{l_1, l_2, l_3 \ldots, l_k\}$ ($\boldsymbol{C} \subset \boldsymbol{L}$) is calculated as [33]:

$$G_{\boldsymbol{C}}^* = \frac{\sum_{l_i \in \boldsymbol{C}} (z(l_i) - \overline{Z})}{S\sqrt{\frac{(N \times W_{\boldsymbol{C}} - W_{\boldsymbol{C}}^2)}{N-1}}} \tag{3}$$

where N is the total number of flows in $\boldsymbol{L}$, $\overline{Z}$ and S represent the mean and standard deviation of $\{z_1, z_2, z_3 \ldots, z_N\}$ respectively. $z(l_i)$ is the density value of flow l_i, $W_{\boldsymbol{C}}$ ($W_{\boldsymbol{C}}$=k) is the number of flows in cluster $\boldsymbol{C}$. This formula has the same meaning as the above Eq. 1. The $G_{\boldsymbol{C}}^*$ statistic can be regarded as a normalized concentration index. The larger the positive value of the $G_{\boldsymbol{C}}^*$ statistic indicates, the flows in the spatial neighborhood of the cluster $\boldsymbol{C}$ are more concentrated, and the cluster $\boldsymbol{C}$ is more likely to be a significant flow cluster. Therefore, the $G_{\boldsymbol{C}}^*$ statistic is used to measure whether two adjacent flows or clusters can be merged. If two adjacent flows belong to a cluster, then when these flow clusters and the flows of their neighbors are combined into one cluster, the $G_{\boldsymbol{C}}^*$ statistic will become larger, proving that the clusters can be merged to become a larger flow cluster.

In the clustering algorithm, we test all the possible combinations of a *seed flow* with its neighboring adjacent flows to search for the optimal combination of these flows to maximize the $G_{\boldsymbol{C}}^*$ statistic. The proposed Statistical and Density-Based Clustering algorithm (SDBC) for OD flows is briefly described in Table 1.

3.4. Statistical significance test of Candidate clusters

Among the candidate flow clusters, there may be some false clusters that occur by chance. To distinguish the false clusters, we apply Monte Carlo permutation tests to assess the statistical significance of the candidate clusters. For geographical data, spatially random distribution, also called complete spatial randomness (CSR), is usually used as the null model for statistical significance tests of clustering patterns [1,34,35]. Under the null model of the CSR, OD flows are generated by an independent random process, which means the flow has an equal probability of occurring in any location in the study area, and the location of one flow is independent of the location of another flow. Based on this, we propose the null hypothesis H_0: *if there is no flow clustering pattern exists in the data, then the flow data appear to be completely randomly distributed*.

Under the null hypothesis H_0, for each flow cluster (say $\boldsymbol{C}_i$), the number of flows falling within the domain range of the cluster is used as the test statistic, denoted as $N(\boldsymbol{C}_i)$, and Monte Carlo simulation is used for testing the statistical significance. In Fig. 4, it illustrates the domain range of the flow cluster $\boldsymbol{C}_i$. For the cluster $\boldsymbol{C}_i$, the O points and D points of the flows in $\boldsymbol{C}_i$ are severally used to generate the minimum boundary polygons (i.e., convex hulls), and the buffer zones with the width of R_c

Table 1
Statistical and Density-Based Clustering algorithm.

Algorithm 1: Pseudo-code of SDBC

Input: OD flow data, the distance parameter R, the angle threshold θ (θ=30° by default), and the G_i^* statistic value threshold λ (the default value of λ is 0).

Output: A set of flow clusters $\boldsymbol{C}$

$\boldsymbol{L} = [l_1, l_2, l_3 \ldots l_N]$; // Initialized as the OD flows of the study area

$\boldsymbol{R} = [l_1, l_2, l_3 \ldots l_M]$; // The selected high-density flows (represented as seeds)

$\boldsymbol{C} = [C_1, C_2, C_3 \ldots, C_M]$; // The initialized clusters by the seeds

$\boldsymbol{S_i} = [q_1, q_2, q_3 \ldots q_i]$; // Spatial neighborhood of flow l_i $(l_i \in \boldsymbol{L})$

$\boldsymbol{S} = [N \times S_i]$; // Spatial neighborhood of $\boldsymbol{L}$

$q \in \boldsymbol{S}_i$; // One flow in $\boldsymbol{S}_i$, which also has its spatial neighborhood named $\boldsymbol{S}_q$

// Step 1. Identify high-density flows

Generate spatial neighborhood $\boldsymbol{S}$;

Compute $\boldsymbol{G}_i^*$ for all l_i in $\boldsymbol{L}$ $(1 \le i \le n)$ according to Eq. (1);

Filter out high-density flows to satisfy $\boldsymbol{G}_i^* > \boldsymbol{\lambda}$ as seeds $\boldsymbol{R}$ for clustering in the next step;

// Step 2. Clustering with statistic constraint

Initialized seeds $\boldsymbol{R}$ as the initial clusters $\boldsymbol{C}$;

Global $max_{\boldsymbol{G}_C^*}$ = -infinity; // Initialize to negative infinity

for $i \leftarrow 0$ to length($\boldsymbol{C}$) **do**:

  $\boldsymbol{S}_i$ = the spatial neighborhood of the current Cluster $\boldsymbol{C}_i$;

  if (length($\boldsymbol{S}_i$)≠ ∅) **then**

    $max_{\boldsymbol{G}_C^*}$ = -infinity; // Initialize to negative infinity

    for all $q \epsilon \boldsymbol{S}_i$ **do**:

      Compute $G_{\boldsymbol{C}}^*$ of $\{\boldsymbol{C}_i, q\}$ according to Eq. (2);

      if ($G_{\boldsymbol{C}}^* > G_i^*$ && $G_{\boldsymbol{C}}^* > max_{\boldsymbol{G}_C^*}$) **then**

        $max_{\boldsymbol{G}_C^*} \leftarrow G_{\boldsymbol{C}}^*$;

      end if;

    end for;

    if $max_{\boldsymbol{G}_C^*} \le$ Global $max_{\boldsymbol{G}_C^*}$ **then**

      break; // If the $max_{\boldsymbol{G}_C^*}$ no longer grows, jump out of the outermost loop

    end if;

    Global $max_{\boldsymbol{G}_C^*} \leftarrow max_{\boldsymbol{G}_C^*}$; // Update the global maximum $\boldsymbol{G}_{\boldsymbol{C}}^*$

    Merge {Cluster $\boldsymbol{C}_i$, q} combination that makes the Global $max_{\boldsymbol{G}_C^*}$ largest;

    $\boldsymbol{C}_i \leftarrow \{\boldsymbol{C}_i, q\}$; // Update the cluster $\boldsymbol{C}_i$

    $\boldsymbol{S}_i \leftarrow \boldsymbol{S}_i + \boldsymbol{S}_q$; // Update the spatial neighborhood of $\boldsymbol{C}_i$

    Clear $\boldsymbol{S}_q$; // Delete the flows in the neighborhood $\boldsymbol{S}_q$ because q has already merged with $\boldsymbol{C}_i$

  end if

end for

return $\boldsymbol{C} \leftarrow \{\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_K\}$ (K is the number of clusters remaining in $\boldsymbol{C}$)

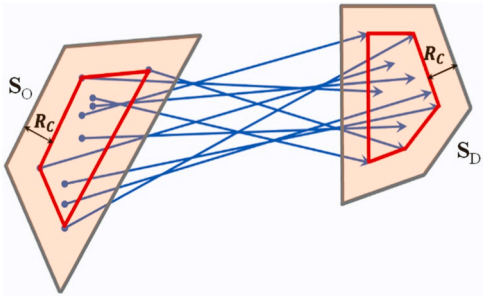


Fig. 4. Illustration of the domain range of the flow cluster.

(R_c=R_o=R_D) are generated outside each of the minimum boundary polygons are denoted as $\mathbf{S}_O$ and $\mathbf{S}_D$, respectively. $\mathbf{S}_O$ and $\mathbf{S}_D$ are defined as the domain range of the cluster $\boldsymbol{C}_i$.

In Monte Carlo simulations, for original OD flows in the study area, the O points and D points of the flows are randomly permutated and then connected in corresponding order to form a new random flow data (with the same number of flows as the original data). For each simulated random flow data, we calculate the value of $N(\boldsymbol{C}_i)$. Repeat the process p times to obtain the values of $N(\boldsymbol{C}_i)$ on the simulated random flow datasets, $\{N(\boldsymbol{C}_i)^1, N(\boldsymbol{C}_i)^2, N(\boldsymbol{C}_i)^3, \ldots, N(\boldsymbol{C}_i)^p\}$($p$=999), which is used to estimate the probability distribution of the test statistic $N(C_i)$ under the null hypothesis. Then, the Monte Carlo *p-value* of $N(\boldsymbol{C}_i)$ is calculated as:

$$\mathrm{p-value}(N(\boldsymbol{C}_i)) = \frac{\sum_{i=1}^{p} I\left(N(\boldsymbol{C}_i)^i \geq N(\boldsymbol{C}_i)^0\right)}{p} \tag{4}$$

where $N(\boldsymbol{C}_i)^0$ is the observed value of the number of flows in the domain range of the cluster $\boldsymbol{C}_i$, and $I(\bullet)$ is an indicator function. If the Monte Carlo *p-value* of $N(\boldsymbol{C}_i)$ is less than the significance level α (α=0.05), the candidate flow cluster $\boldsymbol{C}_i$ is identified as a significant cluster, and the non-significant candidate flow clusters that occurred by chance will be removed.

The statistical significance of each candidate cluster is determined by comparing the observed concentration of flows within the cluster's domain against a distribution of concentrations generated through repeated random permutations of the flow data. Upon applying the Monte Carlo permutation tests, we calculated *p-values* for each candidate cluster, comparing these against a significance level of α=0.05 to determine their statistical significance. A Monte Carlo *p-value* less than the predetermined significance level of α=0.05 indicated that the occurrence of the flow cluster significantly deviated from what would be expected under the CSR model, thereby confirming the presence of a meaningful flow cluster.

3.5. Computational complexity of the proposed method

Given a geographical flow dataset containing N origin-destination (OD) flows, the computational complexity of the proposed method is determined by five components. (1) Reading and constructing flow objects from spatial OD point data has a complexity of $O(N)$, where N is the number of flow data; (2) Identification of spatial neighbors of each flow involves two nested loops, each iterating over all N flows, resulting in time complexity of $O(N^2)$; (3) Calculating the density of each flow, and the complexity is $O(N)$; (4) Statistically constrained clustering process includes identifying high-density flows, initializing clusters, and iteratively merging these clusters. Assuming all spatial flows are identified as high-density flows and without the use of efficient data structures, the clustering step's computational complexity can be as much as $O(N^2)$; (5) Statistical significance testing of candidate clusters is determined by the number of simulations p and the number of flow data N. The most critical part is the processing of each simulated dataset, which has a time complexity of $O(pN)$. Since p is usually a fixed number (such as 999), the overall complexity of this process is primarily influenced by the number of flow data N. Overall, the total computational complexity of the proposed flow clustering algorithm is at most $O(N^2)$ without using efficient data structures.

Among them, the most time-consuming parts of the algorithm are the neighbor identification and cluster merging processes, as both involve quadratic-level iterative computations. Firstly, the time complexity can be reduced with some efficient spatial index structures (such as R-trees or KD-trees) for searching the neighborhoods as used in the DBSCAN algorithm [13]. Furthermore, using big data technology to optimize resource allocation, especially machine learning and predictive analytics methods can be used to intelligently identify neighbors and merge clusters [36,37], thereby reducing the computational complexity of these time-consuming parts. Secondly, for handling large flow datasets in real-world applications, parallel computing techniques can be used to improve the efficiency, which includes data parallelization (dividing data into small batches and processing them in parallel on multiple processors), task parallelization [38] (executing different computational steps in parallel threads or processors) with Hadoop, Apache Spark, etc. Specifically, we can partition the source flow data into different subsets according to the spatial distribution of the origin points or the destination points, and then the flow clustering algorithm can be parallelly executed with different subsets. Since the origin points or the destination points of the flows in the same cluster need to meet the necessary condition of spatial proximity, the data partition based on the origin or destination points will not seriously affect the quality of the clustering result, while could provide computational efficiency by parallelizing the clustering operation on big data.

4. Simulation data and experimental results

4.1. Design of simulated flow datasets

In this paper, we use seven simulated flow data, ***SD***1∼***SD***7, to evaluate the effectiveness of our method, as shown in Fig. 5. The OD flow data ***SD***1 contains three flow clusters of different lengths and densities with similar directions. ***SD***2 contains two flow clusters of different shapes and densities with 509 noise flows. ***SD***3 has three flow clusters of different shapes and different lengths. It consists of a total of 949 flows, of which the number of noise flows is 509 (53.64 %). The flow data ***SD***4 consists of a total of 959 flows, which includes three overlapping flow clusters and 509 noises. ***SD***5 contains three flow clusters of different densities and different directions with neighboring original points, and contains a total of 1034 flows, including 509 noises. ***SD***6 contains two flow clusters with different densities and different lengths, in which the larger cluster covers the small cluster. The data ***SD***7 in Fig. 6 are used to evaluate the effectiveness of the proposed method for detecting different-density clusters with adjacent high-density clusters. There are six clusters in ***SD***7 with high, medium, and low densities, also the six clusters have different directions and sizes. Statistics of each cluster in the simulated data are listed in Table 2.

4.2. Comparison methods and quality evaluation measure

To evaluate the effectiveness of the proposed method, based on the above simulated flow datasets, eight state-of-the-art flow clustering methods, namely, density-based method [16], hierarchical clustering method [5], statistic-based method [7], a stepwise spatio-temporal flow clustering method [19], spatial flow L-function based method (SpatialflowL) [12], flowHDBSCAN [20], Density domain decomposition [17], and the modified SNN_flow [9] are compared.

The density-based clustering method is modified from the OPTICS algorithm by redefining three important parameters, i.e., *Eps* (the spatial distance threshold to construct the neighborhood of a flow), *MinPts* (the density threshold to identify the high-density flows) and *Eps*′ (the

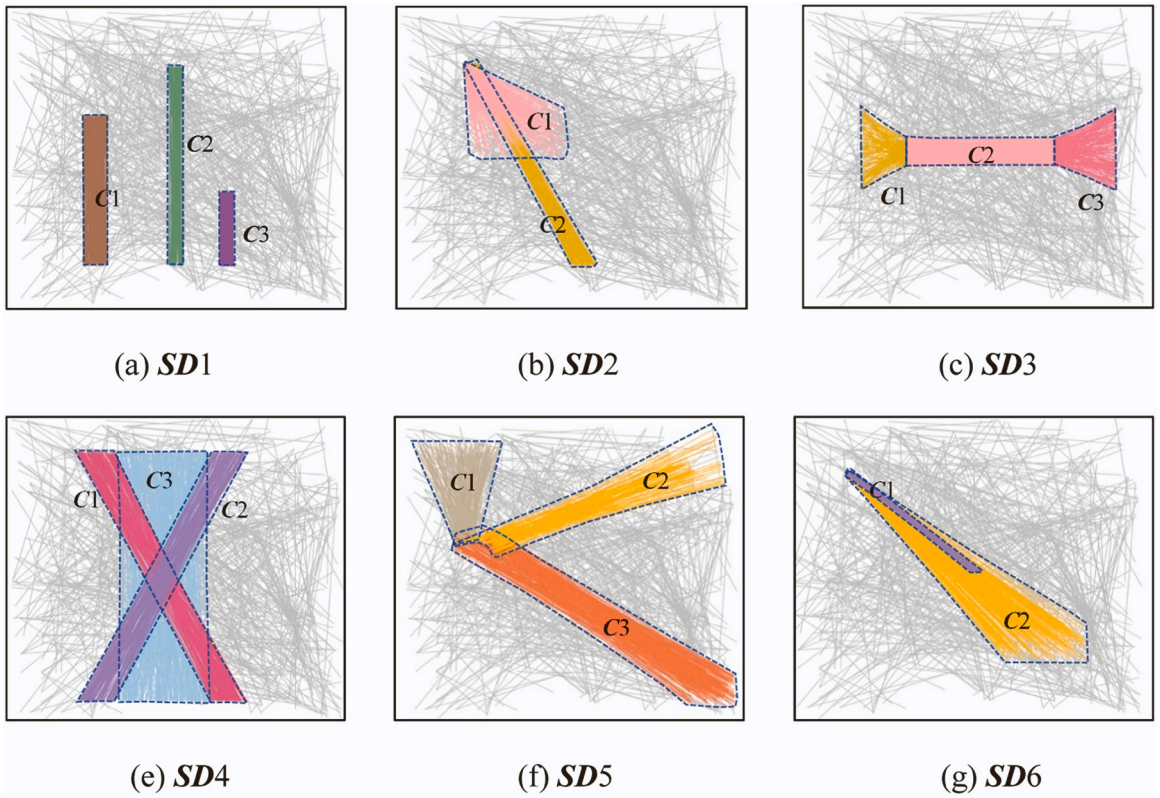


Fig. 5. Simulated OD flow data ***SD***1∼***SD***6.

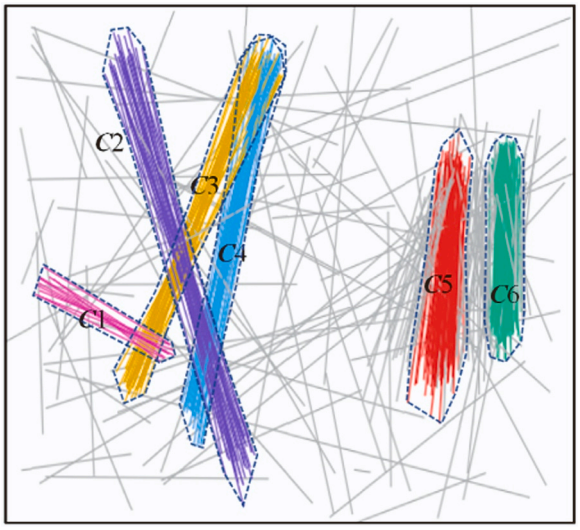


Fig. 6. Simulated OD flow data ***SD***7.

clustering distance). Based on the original research, it is suggested that the values of *Eps* and *MinPts* should be set to sufficiently large values to achieve significant results [14]. The parameter *MinPts* also can be determined empirically as [*ln*(numbers of flows)] [39]. The value of *Eps* can be established by taking the average of the *k*th nearest distances among flows [8]. The selection of the clustering distance *Eps'* (where *Eps'* is equal to or less than *Eps*) for assigning cluster memberships was accomplished through a straightforward process of analyzing the reachability plot [16]. For the flowHDBSCAN method [20], the minimum cluster size (*MinFlows*) needs to be set in advance. In the comparison experiment, we set *MinFlows* as 5, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80 respectively, and we chose the best clustering results among them for comparison. Similar to flowHDBSCAN, the density domain decomposition method [17] was proposed to detect flow clusters of different densities. In this method, the mixed probability distribution of the *k*-th nearest flow distances is first acquired and the flows are decomposed into two components based on the *k*-th nearest flow distances, i.e., dense flows and sparse flows. Sparse flows are then classified as noise, while dense flows are further grouped into flow clusters using the density-connected clustering strategy. The value of *k* (indicating the number of nearest flow distances) should be set manually for the density domain decomposition method. In the experiment, we set *k* as 5, 8, 10, 12, 15, 17, 20, 23, 25, 27, 30, 32, 35, 37, 40 respectively and chose the best clustering results for comparison.

As for the hierarchical clustering method [5], there is only one parameter *k*, which means the number of neighbors of origin or destination points. The parameter *k* used for this comparative method adheres to the principles outlined in previous research and was established to be 30 for this comparison experiment. The stepwise spatial-temporal clustering method [19] integrates temporal and spatial aspects of flow data clustering. This method introduces a novel metric for spatial similarity that considers both the direction and length of flows. Additionally, it establishes a measurement for temporal similarity. Leveraging spatial and temporal similarity metrics, this method applies a clustering framework to systematically aggregate flows, creating a hierarchy of flow clusters.

For the statistic-based clustering method [7], the algorithm first defines vectors with the coordinates of the starting and destination points of OD flows, and calculates the distances or dissimilarities between every two flows, generating a distance matrix. Next, compute the local K-function of all flow events for a series of scales *r*. Evaluate the statistical significance by simulating a set of flows within the study area, calculating the local K-function values for these simulated flows, and repeating this process. By sorting the simulation results, obtain upper and lower significance envelopes. Finally, compare the actual results with these envelopes to determine whether the observed pattern

Table 2
Statistics of each cluster in the simulated data.

Data	Number of flows in each cluster							Total number of flows
***SD*1**	***C*1**		***C*2**		***C*3**		**Noise**	934
	150		75		200		509	
	(16.06 %)		(8.03 %)		(21.41 %)		(54.50 %)	
***SD*2**	***C*1**		***C*2**		**Noise**			809
	100		200		509			
	(12.36 %)		(24.72 %)		(62.92 %)			
***SD*3**	***C*1**		***C*2**		***C*3**		**Noise**	949
	190		100		150		509	
	(20.02 %)		(10.54 %)		(15.80 %)		(53.64 %)	
***SD*4**	***C*1**		***C*2**		***C*3**		**Noise**	959
	100		100		250		509	
	(10.43 %)		(10.43 %)		(26.07 %)		(53.07 %)	
***SD*5**	***C*1**		***C*2**		***C*3**		**Noise**	1034
	150		200		175		509	
	(14.51 %)		(19.34 %)		(16.92 %)		(49.23 %)	
***SD*6**	***C*1**		***C*2**		**Noise**			809
	100		200		509			
	(12.36 %)		(24.72 %)		(62.92 %)			
***SD*7**	***C*1**	***C*2**	***C*3**	***C*4**	***C*5**	***C*6**	**Noise**	384
	10	28	29	38	74	77	128	
	(2.60 %)	(7.29 %)	(7.55 %)	(9.90 %)	(19.27 %)	(20.05 %)	(33.33 %)	

exhibits clustering or dispersion. The key of this algorithm lies in the assessment of proximity based on flow events and the statistical significance evaluation of the local K-function, thereby revealing the clustering patterns of spatial flow events. The significance level (α) and the number of Monte Carlo simulations (m) were set to 0.05 and 999, respectively. Also, the maximum value of the clustering scale r can be determined using the global K function [17]. SpatialflowL is also a statistic-based clustering method to analyze the spatial clustering pattern of geographical flows. Its clustering process involves several steps. Firstly, it performs a significance test to determine the presence of spatial clustering patterns. Next, it estimates the spatial aggregation scale r, which is the only parameter of the SpatialflowL method based on the global L-function. Finally, it identifies the flows with the top 1 % local L-function values at the estimated scales and combines them to form the dominant clusters [12]. Recently, the SNN_flow [9] is proposed to detect the flow clusters in the road network. In this method, the origin and destination points of the flows are first matched onto the road network and the shortest path distances between the origin and destination point pairs are used to define the flow distance. The number of shared nearest neighbors of a flow is defined as the flow density, and the Monte Carlo simulation test for the flow density is performed to identify the high-density flows for density-connected clustering. As our simulated flow data do not consider the road network settings, we used the Euclidean distances between the origin and destination point pairs to define the flow distance instead of the shortest path distances in the experiment. The value of k (the number of nearest neighbors) is set to 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 in turn, the number of Monte Carlo simulations (R) is set to 99, and the significance level (α) is set to 0.05 for the modified SNN_flow method. We finally chose the best clustering results among them.

The parameter settings of these flow clustering methods and the proposed method are listed in Table 3. We implemented the proposed method and the comparison methods in Python (version 3.7) and tested the algorithms on Intel Core i7 CPUs running at 2.0 GHz with 16 GB of RAM. The data and codes are available on Github: https://github.com/csu-mapping/SDBC.

The performances of these clustering methods are evaluated with the adjusted rand index (ARI) [40]. ARI is a widely used index to measure the accuracy of the clustering algorithm. Let $\boldsymbol{D}$ be a geographical flow data with n OD flows (cardinality $|\boldsymbol{D}| = n$), $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\}$ be the clustering result of $\boldsymbol{D}$ generated by the clustering algorithm, $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K$ are non-empty disjoint subsets of $\boldsymbol{D}$ such that their union equals to $\boldsymbol{D}$. Let $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_T\}$ be the benchmark clustering result of $\boldsymbol{D}$. ARI can be calculated as:

$$\mathrm{ARI}(\boldsymbol{C}, \boldsymbol{V}) = \frac{n(n-1)\sum_{i=1}^{K}\sum_{j=1}^{T}\binom{m_{ij}}{2} - 2\sum_{i=1}^{K}\binom{|\boldsymbol{c}_i|}{2}\sum_{j=1}^{T}\binom{|\boldsymbol{v}_j|}{2}}{n(n-1)\left(\sum_{i=1}^{K}\binom{|\boldsymbol{c}_i|}{2} + \sum_{j=1}^{T}\binom{|\boldsymbol{v}_j|}{2}\right)\Big/ 2 - 2\sum_{i=1}^{K}\binom{|\boldsymbol{c}_i|}{2}\sum_{j=1}^{T}\binom{|\boldsymbol{v}_j|}{2}} \tag{5}$$

where $|\boldsymbol{c}_i|$ and $|\boldsymbol{v}_j|$are the number of flows in the cluster $\boldsymbol{c}_i$ and $\boldsymbol{v}_j$, m_{ij} is the number of flows belonging to both $\boldsymbol{c}_i$ and $\boldsymbol{v}_j$. ARI is a number between 0 and 1, and a higher ARI indicates a stronger agreement between the clustering result $\boldsymbol{C}$ and the benchmark $\boldsymbol{V}$.

4.3. Experimental results on simulated data

The clustering results of the proposed method for the simulated data ***SD*1**~***SD*6** are shown in Fig. 7. The proposed method successfully finds all the predefined clusters in the simulated datasets. The ARIs of the results of different clustering algorithms are shown in Table 4. Note that, since the dominant cluster detected by SpatialflowL [12] is extracted as a whole, it cannot directly separate it into individual clusters. Thus, we did not calculate and compare its ARIs in the experiment.

Table 4 presents the ARI values for the clustering results obtained by the comparative methods. For the six simulated datasets, ARI values for the clustering results by the SDBC were generally the highest. For ***SD*2**, the modified SNN_flow method [9] performed best, followed by the density domain decomposition method [17] (with k=23) and the

Table 3
Parameter specification for the methods to be tested.

Category	Methods	Parameters
Density-based clustering	Extension of OPTICS for flow data [16]	*MinPts* (the minimum number of objects in a neighborhood, here it was set to 20.), *Eps* (the spatial distance threshold to construct the neighborhood of a flow), *Eps'* (the clustering distance, *Eps'*≤*Eps*)
	flowHDBSCAN [20]	*MinFlows* (the minimum cluster size)
	Density domain decomposition [17]	*k* (*k*-th nearest flow distances for distinguishing dense flows from sparse flows)
Hierarchical clustering	Hierarchical clustering method [5]	*k* (the number of neighbors of origin or destination points, it was set to 30 here in the experiment)
	A Stepwise Spatial-Temporal clustering method [19]	*k* (the number of nearest flows, which was set to 25 here in the experiment), α (the size coefficient, which was set to 0.55)
Statistic-based clustering	Spatial statistical approach based on local K-function [7]	*r* (spatial scales), α (the significance level, it was set to 0.05), *m* (repeat times of Monte Carlo simulations, it was set to 999)
	Spatial flow L-function (SpatialflowL) [12]	*r* (the spatial aggregation scale, it can be obtained when the global L-function value reaches the maximum)
	SNN_flow with Euclidean flow distance [9]	*k* (the number of nearest neighbors), the number of Monte Carlo simulations (*R*) was set to 99, and the significance level (α) as set to 0.05
	The proposed method (SDBC)	*R* (the distance parameter of buffer size, it was set to 300 here in the experiment), θ (the angle threshold, it was set to 30°)

proposed SDBC. The SDBC performed better than the other flow clustering methods, especially for clusters of different shapes, densities, and sizes that are overlapping or close to each other, such as ***SD***4, ***SD***5 and ***SD***6.

Further to evaluate the effectiveness of the proposed method, the clustering results by eight state-of-the-art clustering methods in Table 2 are shown in Fig. 8. In Fig. 8(a), we can see that the proposed method can successfully find the arbitrarily shaped significant clusters in spatial data with noise, and it performed better than the other methods. Among these methods, the density-based method [16] can identify two high-density flow clusters (***C***5 and ***C***6) and two medium-density flow clusters (***C***3 and ***C***4), and cannot separate them well. This is because density-based clustering methods rely on pre-set density thresholds, making it difficult for them to effectively separate clusters when the transition in density between clusters is smooth. Just like this case where two high-density clusters (***C***5 and ***C***6) with similar densities, are smoothly connected by some noise, it becomes difficult to separate these two clusters. Moreover, due to the presence of clusters with significant differences in density within ***SD***7, finding parameter settings suitable for identifying clusters of all different densities is quite challenging. The parameters selected through sorted *K_distance* and the clustering results by the density-based method are shown in Fig. 8(c)-(d).

In Fig. 8(b), the hierarchical-based clustering method [5] successfully identified two high-density clusters ***C***5 and ***C***6 and two medium-density clusters ***C***3 and ***C***4, but the two low-density clusters ***C***1 and ***C***2 could not be separated, and there was still a small amount of

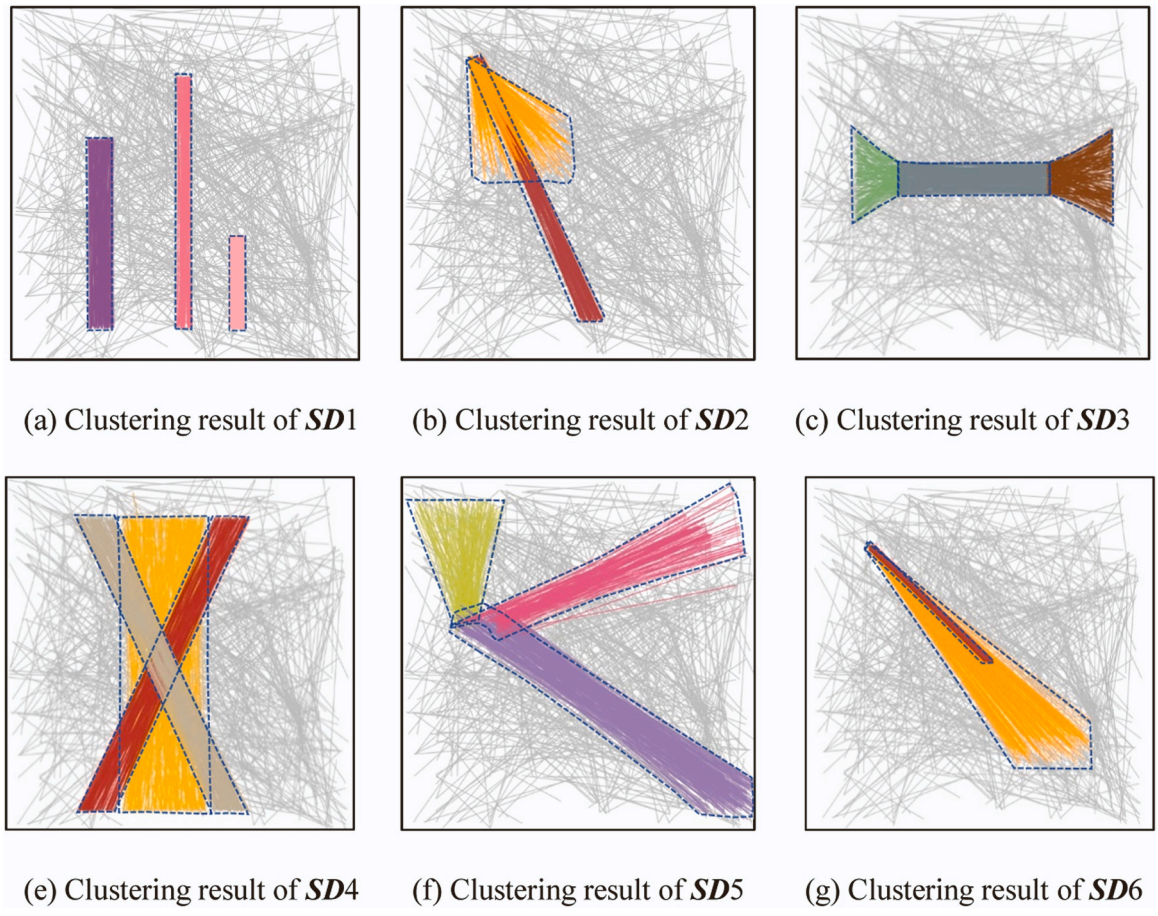


Fig. 7. Clustering results of ***SD***1∼***SD***6 by the proposed method (indicates the contours of the ground truth clusters in ***SD***1∼***SD***6).

Table 4
ARIs of the clustering results by different clustering algorithms*.

Data set	*SD1*	*SD2*	*SD3*	*SD4*	*SD5*	*SD6*	*SD7*
Density-based method [16]	0.9802	0.9740	0.9607	0.8238	0.8523	0.8773	0.7324
Hierarchical clustering method [5]	0.7638	0.8949	0.7280	0.8743	0.8491	0.7681	0.7986
Statistic-based method [7]	0.9668	0.8497	0.7803	0.9686	0.9588	0.9080	0.8601
A stepwise spatial-temporal flow clustering [19]	0.9764	0.9433	0.9537	0.9827	0.9658	0.9722	0.8314
flowHDBSCAN [20]	0.9737	0.7964	0.7013	0.7859	0.4693	0.9323	0.8236
	(*mf*=80)	(*mf*=60)	(*mf*=60)	(*mf*=3)	(*mf*=15)	(*mf*=80)	(*mf*=10)
Density domain decomposition [17]	0.9448	0.9913	0.9742	0.8831	0.7867	0.9781	0.7350
	(*k*=35)	(*k*=23)	(*k*=35)	(*k*=40)	(*k*=32)	(*k*=5)	(*k*=15)
SNN_flow with Euclidean flow distance [9]	0.9975	**0.9925**	0.8906	0.8481	0.9923	0.8924	0.7193
	(*k*=30)	(*k*=35)	(*k*=30)	(*k*=15)	(*k*=30)	(*k*=35)	(*k*=20)
The proposed method (SDBC)	**0.9982**	0.9882	**0.9803**	**0.9948**	**0.9953**	**0.9925**	**0.9457**

*Note: *mf* refers to *MinFlows* (the minimum cluster size)

noise in the identification results. It can identify clusters of different shapes and sizes, but when the density of the clusters varies greatly, the algorithm may misjudge the noises as members of clusters. The measurement of distance between flows is crucial for the formation of clusters in the hierarchical-based clustering method. If the distance metric used, such as nearest neighbor or single linkage, allows clusters to be connected by flows that are relatively close together, low-density clusters might merge even if they are in different directions. As for the results of stepwise spatial-temporal flow clustering method [19] shown in Fig. 8(i), the two low-density clusters ***C***1 and ***C***2 and two medium-density clusters ***C***3 and ***C***4 were successfully identified, but the two high-density clusters ***C***5 and ***C***6 could not be separated. This is because the distance measurement in this clustering method is effective in distinguishing low-density and medium-density clusters in ***SD***7 dataset, because the spatial distance between these clusters is large, and the difference in flow direction is also obvious. However, in high-density regions, the flows of multiple clusters may be spatially very close and flow in similar directions, making it difficult to distinguish between these distance measures.

As for the statistical method [7], we can see that when the scale *r* is relatively small, the lowest-density cluster cannot be found, but medium and high-density clusters can be identified. This is because, at a smaller scale, the characteristics of low-density clusters are not distinct enough to differentiate them from background noise, leading to their non-recognition. With the expansion of the scale *r*, low-density clusters are found, but high-density clusters are linked together and cannot be separated. This is because at a larger spatial scale, low-density clusters become recognizable due to their relative aggregation within the cluster. However, high-density clusters, due to their proximity and overlapping areas, may be mistakenly identified as a single large cluster. These results indicate that the statistical clustering method is highly sensitive to the choice of spatial scale. Different scale settings directly impact the accuracy of clustering results, especially when dealing with complex datasets containing clusters of varying densities. The results of clustering under different parameter selections of the statistical method are shown in Fig. 8(e)-(f).

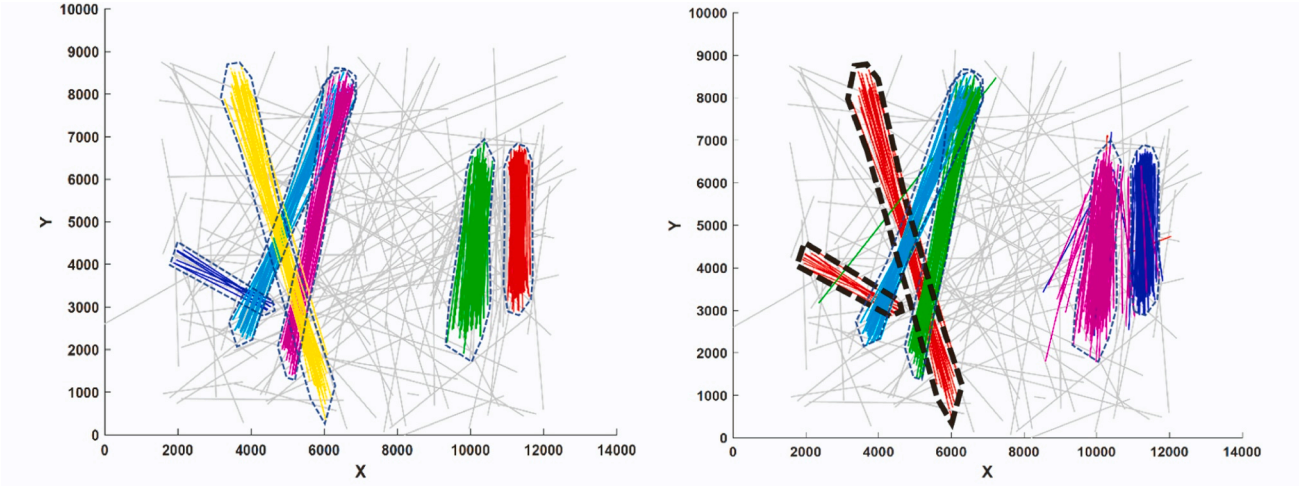
As for the SpatialflowL method [12] which excels at identifying prominent, high-density clusters, such as C5 and C6 were firstly identified due to its emphasis on aggregation scale. Then, the firstly extracted high-density clusters ***C***5 and ***C***6 are eliminated from ***SD***7 dataset, and for the remaining data, the SpatialflowL function is executed to obtain new clusters. Repeat the above process until no clusters can be found in the remaining data. Ultimately, after three iterations of this process, the final clustering results are as shown in the Fig. 8(h). The SpatialflowL method focus may limit its ability to distinguish closely situated clusters or to detect low-density clusters. Its reliance on specific scale parameters might lead to the identification of only the most dominant high-density clusters, potentially overlooking smaller, less dense, but equally significant clusters. The aggregation scale selected through Global SpatialflowL function (L(*r*)) and the results of clustering of the SpatialflowL method are shown in Fig. 8(g)-(h). The flowHDBSCAN, density domain decomposition, and SNN_flow methods can find all the high-density flow clusters in ***SD***7. However, due to the flow distance measurement used in the SNN flow does not consider the orientation consistency, some noise flows nearby the clusters were wrongly merged. The density domain decomposition method missed the low-density flow clusters at the left bottom corner. Noting that the clustering results of these three algorithms are selected from the results under multiple parameters.

The clustering results show that the proposed method is effective and it performed better than the other five clustering methods for detecting flow clusters of different densities and different shapes in data with noise. First, the SDBC algorithm can deal with spatial flow clusters of arbitrary shapes and different densities, as well as noise interference. The statistic-based clustering method [7] may not be able to separate multiple clusters due to the fixed size of the scan window. The density-based clustering method [12] may encounter challenges when working with flow datasets that include clusters of varying densities due to difficulties in selecting appropriate parameters. Second, our method can be considered as an almost parameter-adaptive algorithm. In the OPTICS algorithm, three key parameters should be determined manually often leads to difficulties in identifying all flow clusters. Third, our method considers the significance of spatial flow clustering and avoids identifying some false clusters that may occurred by chance, while other methods do not take this into account. Table 5 shows the comparisons of the abovementioned clustering methods.

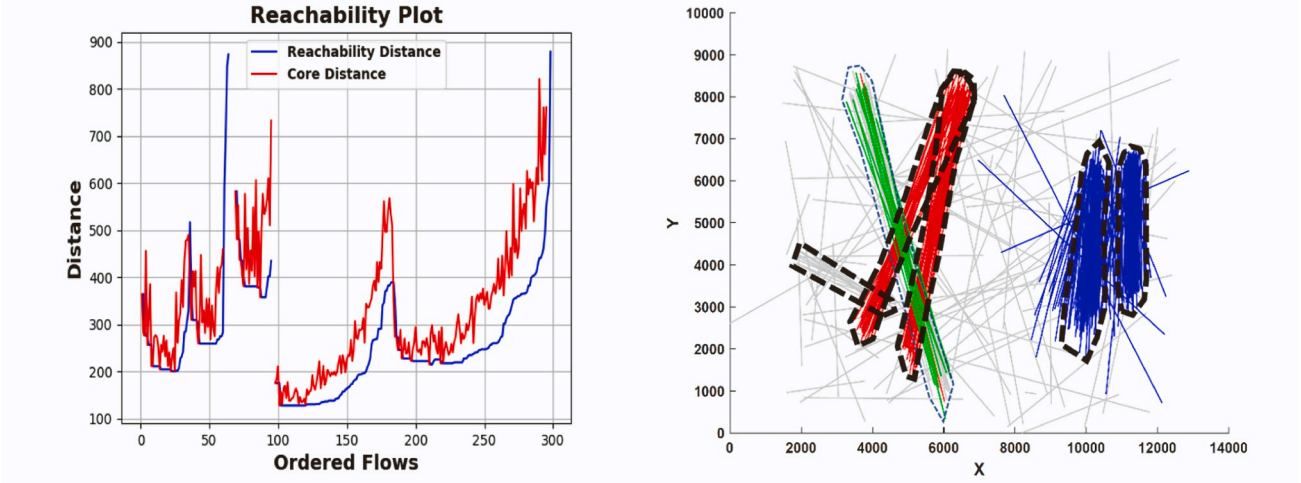
5. Case study with Wuhan taxi OD flow data

5.1. Study area and datasets

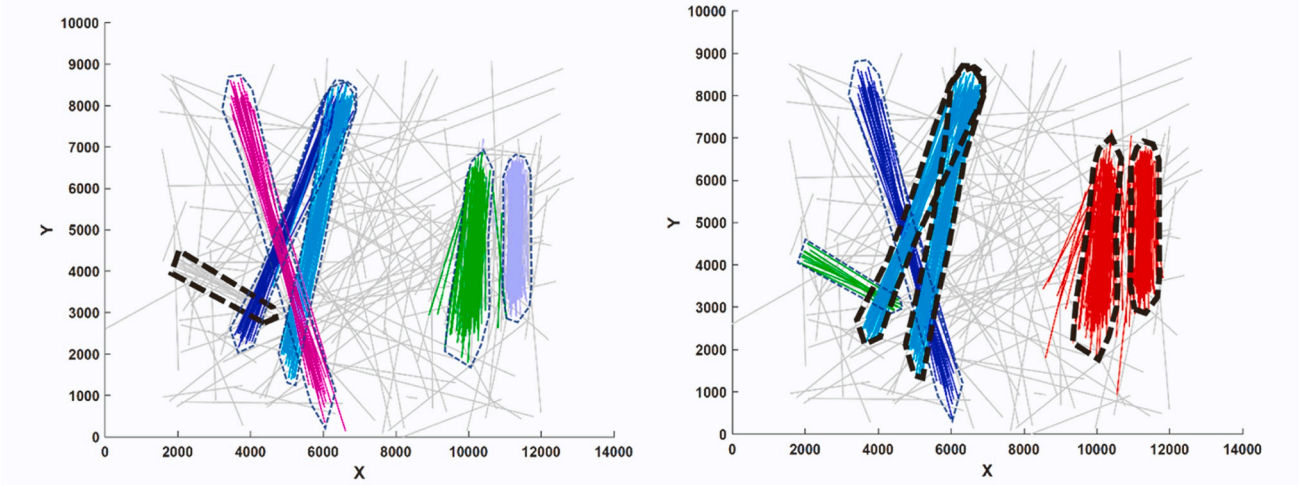
Taxi trajectory data accounts for a large proportion of the traffic flow in the urban road network, and is often used to reveal the mobility patterns of crowds in urban space. In this paper, we performed a case study with taxi trajectory data in Wuhan, China, to further validate the applicability of our method on real-world flow data. The study area is located within the Fifth Ring Road in Wuhan, China, as shown in Fig. 9 (a). Due to the occlusion of GPS devices and other problems, there are some noisy points in the vehicle trajectory data. We first extracted the trajectory data inside the range of the study area. The noise or error sampling points in the trajectory data were eliminated according to the distance and time interval between adjacent trajectory points, and maximum speed limit. If the distance or time interval between two adjacent trajectory points exceeds a given threshold (e.g., 500 m or 120 s), the trajectory is interrupted at the trajectory point; If the speed of the vehicle at one trajectory point exceeds a given threshold (e.g., 120 km/h), the trajectory point is removed and the trajectory was divided into two sub-trajectory at this point. After these preprocessing steps, we extracted the pick-up and drop-off point pairs to form the OD flow dataset from the remaining 223,354 taxi trajectories in Wuhan on



(a) Results of the proposed method (R=300, θ=30) (b) Results of hierarchical clustering method (k_value=30)

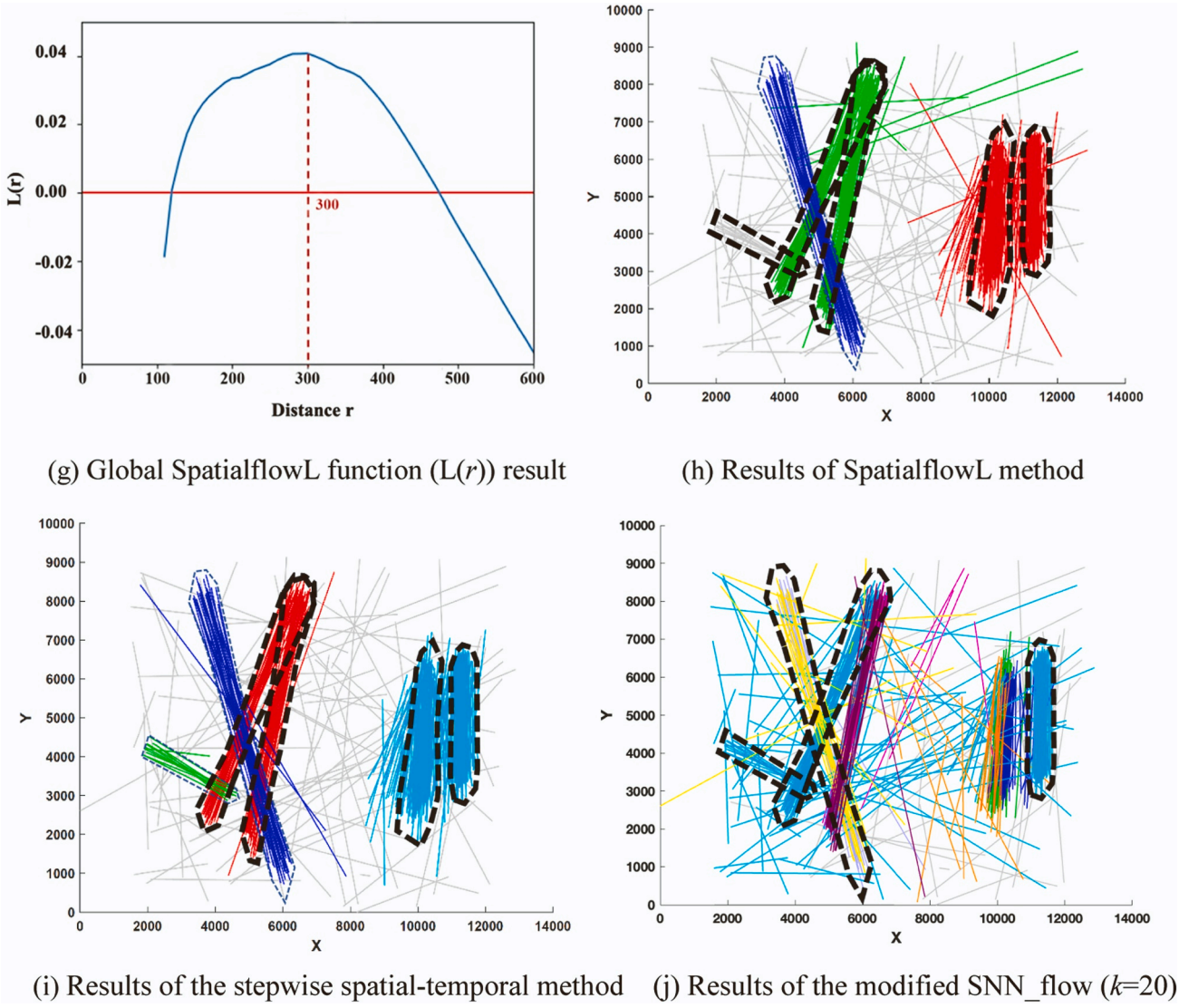


(c) Sorted $K\_distance$ graph (d) Results of Density-based method (Eps=900, $MinPts$=20, Eps'=600)



(e) Results of Statistical-based method (r=400) (f) Results of Statistical-based method (r=1500)

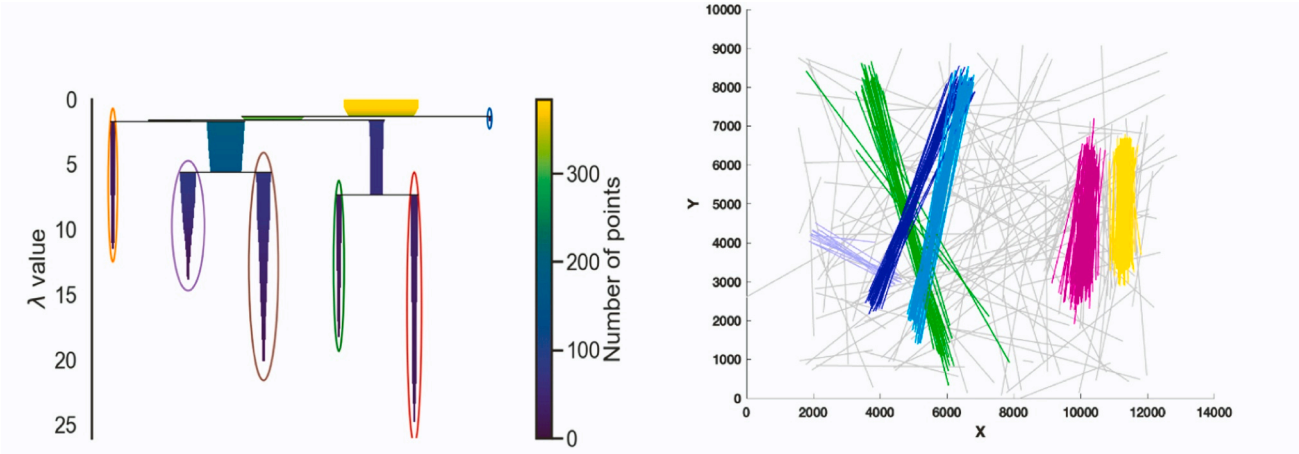
Fig. 8. Clustering results of ***SD7*** by different clustering methods (indicates the contours of the ground truth clusters in ***SD7*** and indicates the wrongly detected clusters).



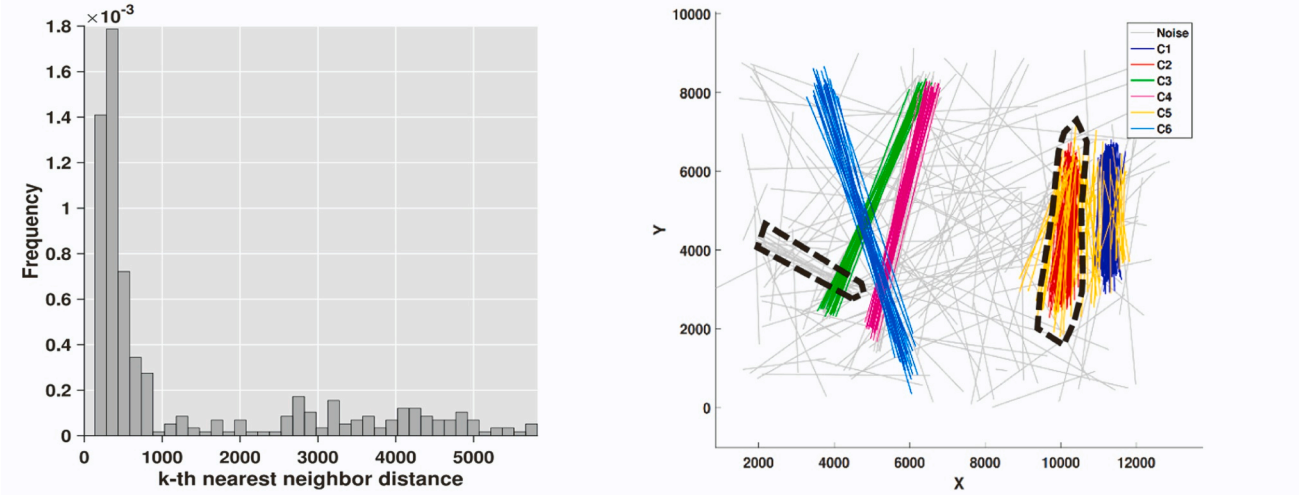
(g) Global SpatialflowL function (L(*r*)) result (h) Results of SpatialflowL method

(i) Results of the stepwise spatial-temporal method (j) Results of the modified SNN_flow (*k*=20)

Fig. 8. (*continued*).



(k) Nearest neighbor hierarchical clustering tree (left) and the simplified hierarchical clustering tree (right) of the result by the flowHDBSCAN (*Minflows*=10)



(l) Distribution of *k*-th nearest flow distances (left) and the best clustering result (right) by the density domain decomposition method (*k*=15)

Fig. 8. (*continued*).

May 1, 2014. The OD flow is a pick-up and drop-off point pair in an individual trajectory, which represents the flow vector of the movement of the vehicle in space from the pick-up point to the corresponding drop-off point. In this case study, a total of 232,353 OD flows were used as the input data. The heatmap of the OD points distribution is shown in Fig. 9 (b). The records of each OD flow include the flow ID, location of the pick-up point, sampling time of the pick-up point, location of the drop-off point, and sampling time of the drop-off point.

Based on the above extracted OD flows in Wuhan on May 1, 2014, to reveal the spatiotemporal movement patterns of crowds, we detected the OD flow clusters during three special periods, namely the morning rush hours (8:00–10:00 am), the afternoon rush hours (16:00–19:00), and the nighttime (19:00–23:00).

From the clustering results, we selected 12 popular sites in Wuhan where the identified flow clusters are located for evaluating the identified clusters, see Table 6 for a specific representation of these sites.

The identified popular sites are primarily distributed around transportation hubs, commercial centers, densely populated residential areas, and university campuses. The local areas or stations that have an important influence on crowd traveling and mobility in the city were well identified by the proposed flow clustering method, which provides useful guidance for revealing the crowd mobility patterns in the urban space and applications such as store location selection, and transportation planning.

5.2. Flow clusters identified during the morning rush hour

The OD flow clusters identified by the proposed method during the morning peak are presented in Fig. 10. To emphasize the prominent OD flow cluster interaction patterns, we excluded small clusters with fewer than 30 OD flows. It is evident that during the morning rush hours, a significant concentration of OD flow clusters occurs at crucial transportation hubs like Hankou Railway Station, Wuhan Railway Station, and subway stations. Among the detected clusters, we chose 6 to delve into the analysis of OD interaction patterns during the morning rush hours.

Among them, the originating and terminating points of unidirectional Cluster I are predominantly located near Hankou Station and

Table 5
Comparisons of different clustering methods.

Category	Method	Distance Measurement	Discovery of clusters with uneven density	Identify noise	Test statistical significance of clusters	Computational complexity (*N* is the number of flows)
Density-based clustering	Extension of OPTICS for flow data [16]	Flow distance	×	√	×	$O(N\log N)$ with efficient spatial access methods such as R*-tree
	flowHDBSCAN [20]	Flow distance	√	√	×	$O(N\log N)$ with efficient spatial access methods such as R*-tree
	Density domain decomposition [17]	Chebyshev distance	√	√	×	$O(N^2+M)$ M means the number of the Markov Chain Monte Carlo simulations)
Hierarchical clustering	Hierarchical clustering method [5]	Chebyshev distance	×	×	×	$O(kMN\log(MN))$ k means k nearest points; M is the average number of flow neighbors
	A Stepwise Spatial-Temporal clustering method [19]	Flow distance	√	×	×	$O(kN\log N)$ k is the number of nearest flows
Statistic-based clustering	Spatial statistical approach based on local K-function [7]	Flow distance	×	√	√	$O(RN^2)$ R is the number of Monte Carlo replications
	Spatial flow L-function (SpatialflowL) [12]	Flow distance	×	√	√	$O(MN^2\log N)$ M is the number of distances (scales)
	SNN_flow [9]	Flow distance	√	√	×	$O(N^2\log N+RN^2k)$ k means k nearest neighbors; R is the number of Monte Carlo simulations
	The proposed method	Flow distance	√	√	√	$O(N^2)$ without using any spatial access method such as R*-tree in neighbor queries

Note: "√" means the method supports the function, "×" means the method does not support the process

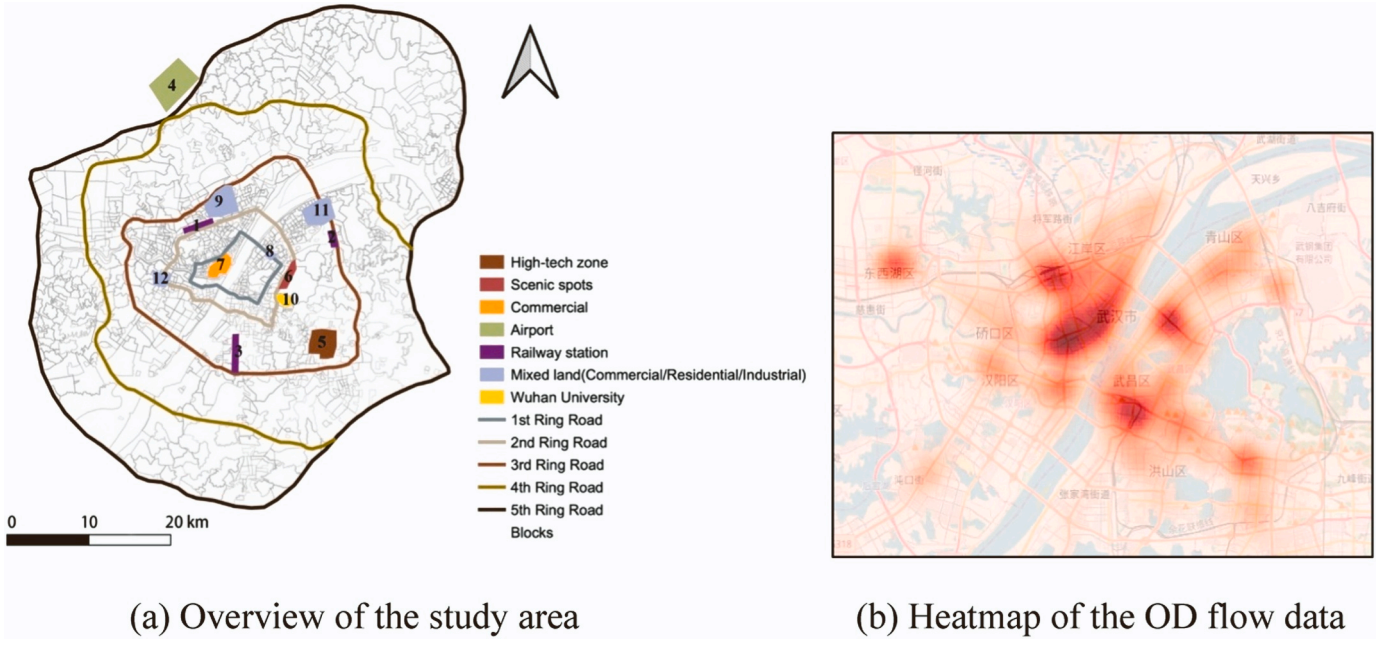


Fig. 9. The study area and the used OD flow data.

Wuhan Railway Station denoted as the path from Hankou Station to Wuhan Railway Station. This cluster likely signifies the spatial interaction pattern between these two transport hubs during the morning rush hour. The endpoints of converging Clusters II and III are situated at Hankou Station and Wuchang Station respectively, illustrating that individuals from distinct regions travel to Hankou Station and Wuchang Station during the morning rush hour. Additionally, it's notable that the service coverage of Hankou Station is notably broader than that of Wuhan Railway Station. The starting point of the OD flow in Cluster IV is mainly distributed near the Xudong subway station, and the end point is near the East Lake tourist area, including the Hubei Provincial Museum, Wuhan East Lake Ocean Paradise, and other tourist attractions. According to the identification results of this cluster, we can find that people are more inclined to take taxis to their destinations near subway stations, which is consistent with our perception. Cluster V originates near Wuhan University and Central China Normal University, both prominent universities in Wuhan, with a concluding point at Wuchang Station. Given the data's collection time during the May Day holiday period, this cluster's identification likely depicts the commuting patterns of students traveling to or returning to their hometowns when heading to the transportation hub. Cluster VI consists of long-distance flows, starting from the commercial center and ending at Wuhan Tianhe Airport, perhaps symbolizing commuters' movement from the city center to the transportation hub.

Table 6
Some popular sites in Wuhan where the identified flow clusters are located.

No.	Name of the site	Description
1	Hankou Railway Station	One of the major railway stations in Wuhan.
2	Wuhan Railway Station	One of the major railway stations in Wuhan.
3	Wuchang Railway Station	One of the major railway stations in Wuhan.
4	Tianhe airport	The first 4 F civil international airport in central China and one of the eight regional hub airports in China.
5	Guanggu	Generally, refers to Wuhan East Lake New Technology Development Zone
6	Donghu Lake	A well-known eco-tourism scenic spot and the first batch of national-level scenic spots.
7	Jianghan Road、Hanzheng Street	Famous century-old commercial streets in Wuhan.
8	Xudong	One of the important business districts in the center of Wuhan.
9	Houhu	It generally refers to Houhu Street under the jurisdiction of Jiang, an District of Wuhan City.
10	Wuhan University	One of the most famous university in China.
11	Honggangcheng	Located in Qingshan District, there are many factories gathered here, Wuhan Iron and Steel Group, etc.
12	Wangjiawan	Located at the intersection of Hanyang Avenue and Longyang Avenue as the core, diverges to the surrounding areas, and is an emerging commercial circle.

5.3. Flow clusters identified during the afternoon rush hour

The OD flow clusters detected by the proposed method during the afternoon peak hours are shown in Fig. 11. Among the 6 identified flow clusters we selected, Cluster I may represent the commuter flow from Wuhan Railway Station to Hankou Station during the afternoon rush hour. The starting points of Clusters II and III are located at Hankou Station and Wuhan Station, respectively, and the end areas are situated near Hanzheng Street and Jianghan Road Pedestrian Street in Wuhan's Business District, as well as in the business district along the Yangtze River. This may signify the influx of tourists from the transportation hub into Wuhan's business district. Cluster IV may represent the commuter flow from the commercial area near Jianghan Road Pedestrian Street to the residential area of Houhu. Cluster V mainly represents the commuter flow from the Wuhan Hanzheng Street business district to the residential area of Jiangteng Community. Cluster VI represents a set of commuter flows across the Yangtze River, originating from the commercial district near the Wuhan Municipal Government to the residential district of Xudong Community on the other side of the river. From the above analysis, we can observe that during the afternoon rush hour, the prominent commuting patterns include flows from the commercial area to residential areas and between transportation hub stations.

5.4. Flow clusters identified during the nighttime

Fig. 12 illustrates the OD flow clusters detected using the method introduced in this paper during the nighttime period. From these, we have identified six distinct flow clusters. Cluster I possibly signifies the commuting flow from the nearby commercial district of Wuhan to the residential vicinity of Wangjiawan. For cluster II, the destinations are predominantly dispersed around Hankou Station, suggesting a probable commuter flow from the riverside commercial district to the transport hub. In the case of Cluster III, the initiation point is close to Wuhan Station, while the final destination aligns with the East Lake Development Zone of Optics Valley. This might indicate a commuter flow from transportation hubs to high-tech development zones. Clusters IV and V are characterized by their opposite starting and ending points, suggesting commutes between the riverside commercial area and the residential region near the Second Ring Road during the nighttime. This pattern might reflect the interactive mode engendered by nocturnal entertainment activities. Lastly, Cluster VI has its starting point near the Red Steel City, concluding at a residential area beyond the Second Ring Road. The concentration of factories around Honggang City on Heping Avenue potentially signifies the nighttime commuting flow of workers returning home after their night shifts.

The clustering results of the proposed method on synthetic and real-

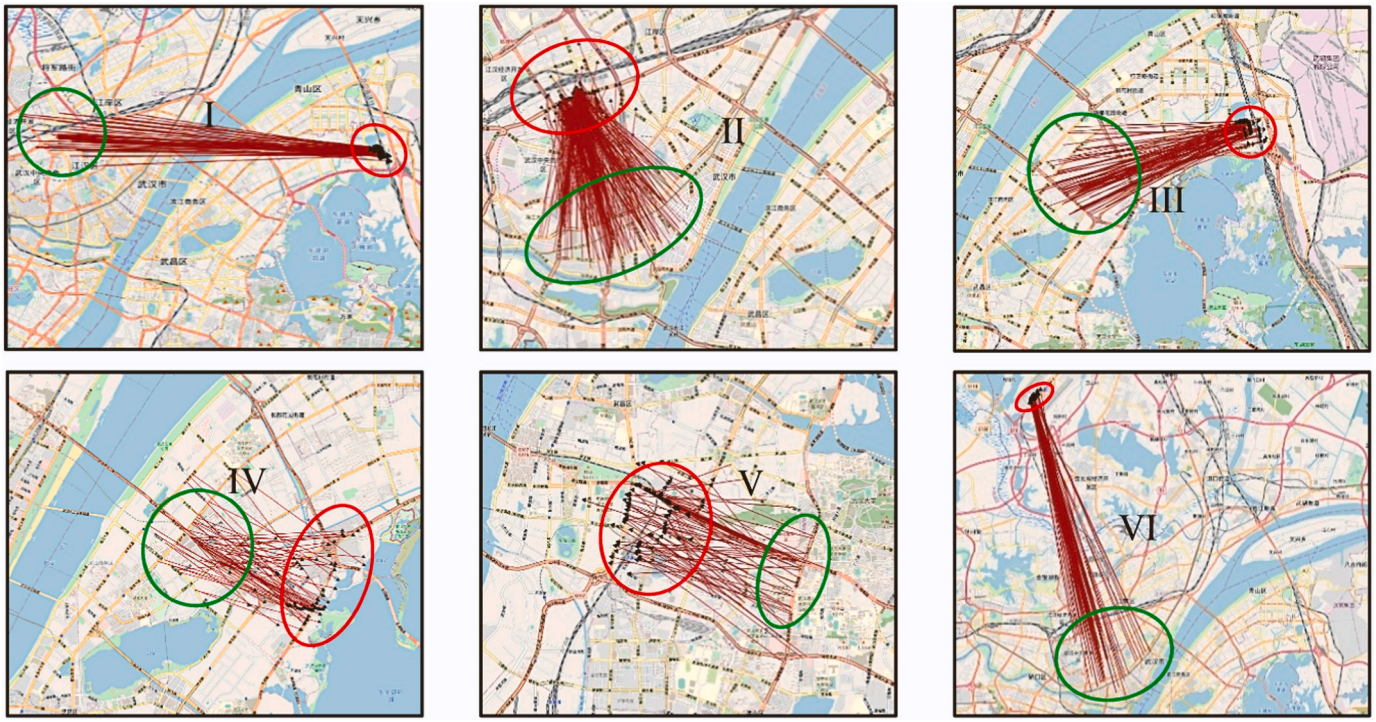


Fig. 10. OD flow clusters detected during the morning peak hours (8:00–10:00 am).

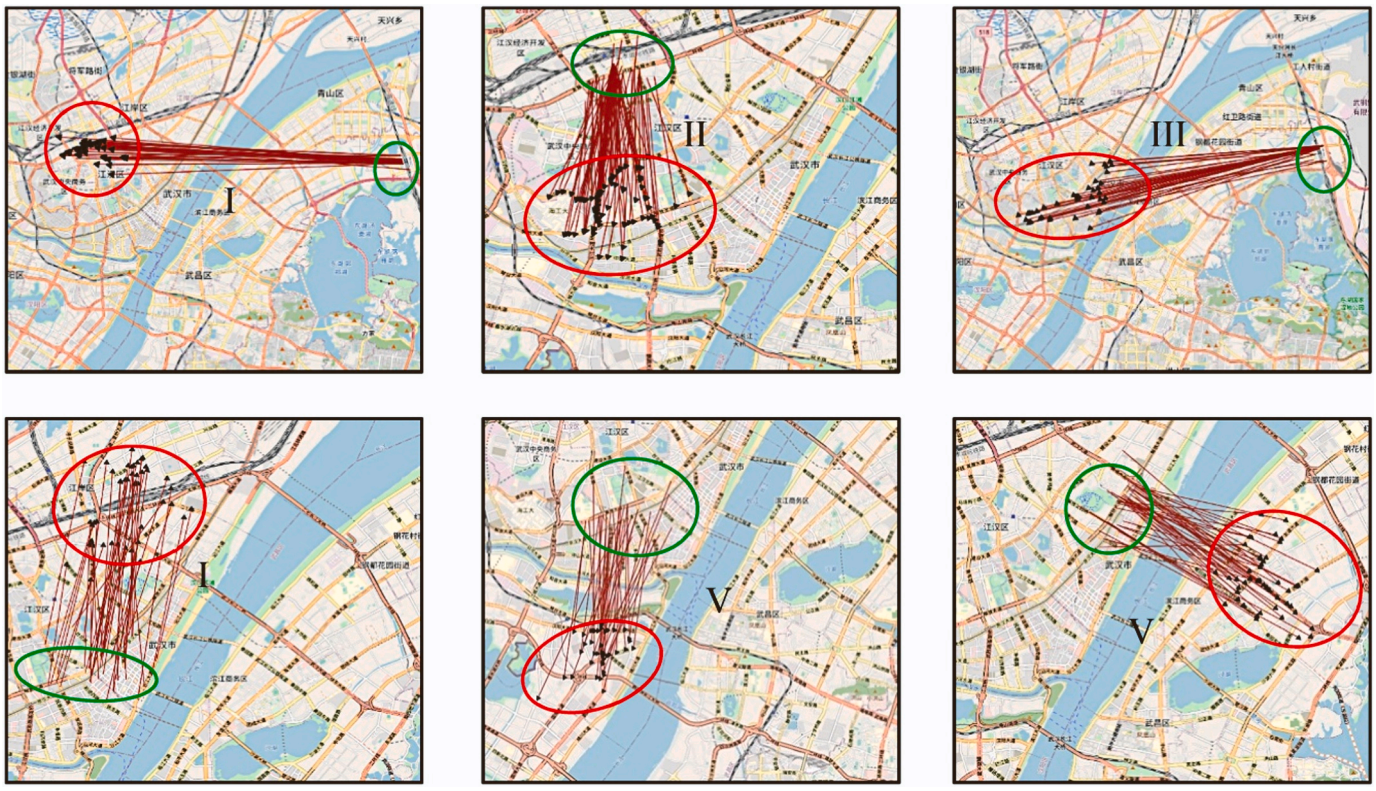


Fig. 11. OD flow clusters detected during the afternoon peak hours (16:00–19:00).

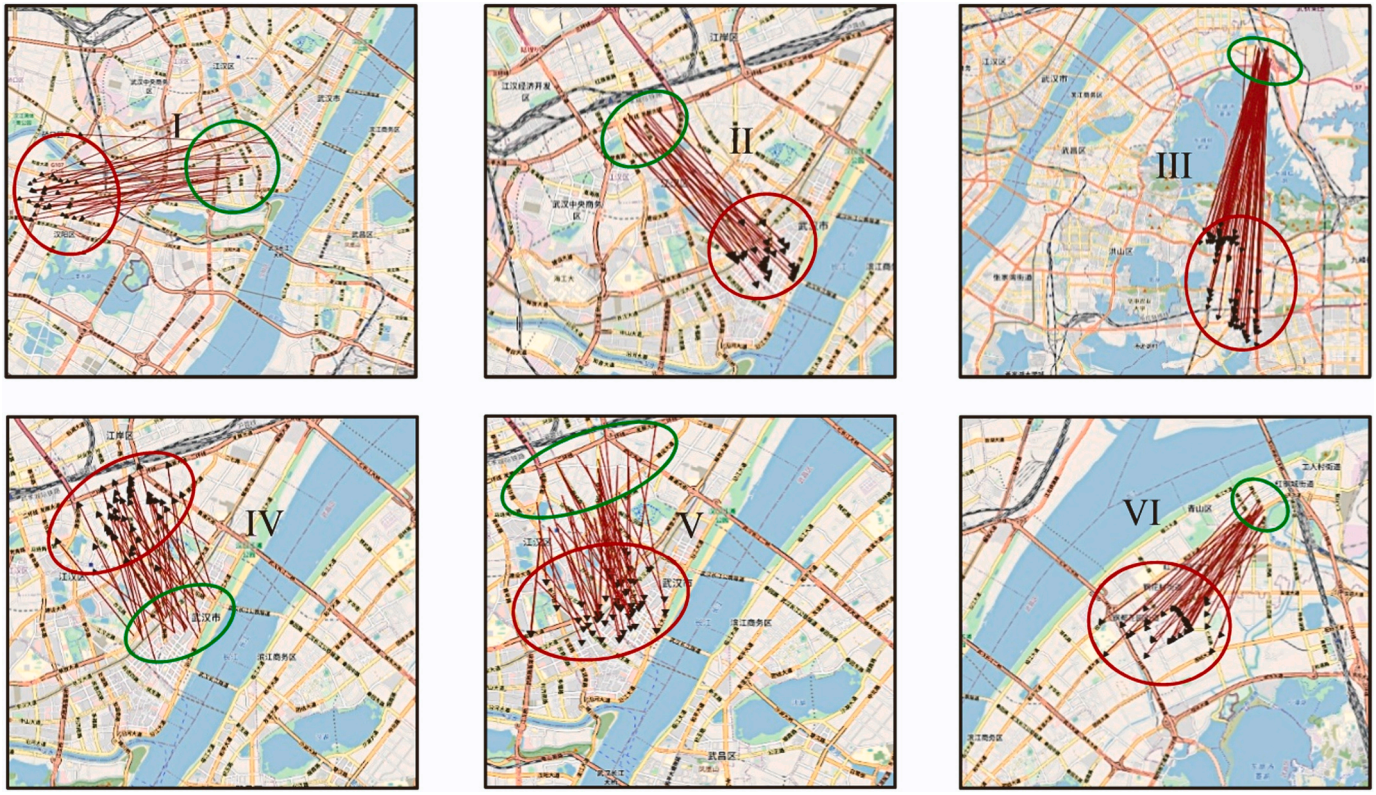


Fig. 12. OD flow clusters detected during the nighttime (19:00–23:00).

Table 7
Statistics of the simulated data and ARIs of the clustering results.

Noise levels	Number of flows		Total number of flows	ARI
Low level	Flows in clusters 256 (75.0 %)	Noise 64 (25.0 %)	320	0.9985
Medium level	Flows in clusters 256 (66.7 %)	Noise 128 (33.3 %)	384	0.9891
High level	Flows in clusters 256 (50 %)	Noise 256 (50 %)	512	0.8988
	Flows in clusters 256 (40 %)	Noise 384 (60 %)	640	0.8713

world taxi trajectory data demonstrate its superiority in identifying significant flow clusters of different shapes and densities compared to existing algorithms. Based on the clustering results of the taxi trajectory data in Wuhan city, we can find the hotspots of spatial interactions between one area to others in the city, which has important guiding significance for us to further explore and analyze people's travel purpose, and in deeper to analysis the distribution and accessibility of urban facilities and resources and further reveal the formation mechanism of traffic flows to provide help for urban planning and traffic control. With the movement data of crowds in the urban cities, we can also apply the proposed clustering method to detect the groups of crowds sharing the common movement patterns, which provides reference for better understanding human mobility patterns and helps to improve facilities allocation and location planning, traffic control, and epidemic prevention, etc.

6. Discussions and conclusions

6.1. Discussions and sensitivity analysis

The applicability of a clustering algorithm to datasets with different noise levels is an important criterion to evaluate the robustness and adaptability of an algorithm. To test the robustness of the proposed method to noise, we used flow datasets containing noise of different levels for the experimental analysis. Datasets containing one preset flow cluster and random noises with a noise ratio lower than 30 %, a noise ratio between 30 % and 50 %, and a noise ratio exceeding 50 %, were used as the input data respectively. The statistics of the simulated data with noise and the ARIs of the clustering results by the proposed method are shown in Table 7.

From the evaluation results in Table 7, it can be seen that, for the data with a lot of random noise, the cluster recognition accuracy of the proposed method is still high. As the noise level increases to 30 % or higher, the clustering accuracy of the proposed method decreases slightly, and the random noise flows near the edge of the cluster are easy to recognize as the members of the cluster. In a high-level noise environment (noise ratio exceeds 50 %), the accuracy of clustering is decreased, but the cluster recognition accuracy is still above 80 %.

Due to the complexity of the geographical environment, there may be a lot of noise in the geographical flow data in the practical applications. Improving the robustness of the SDBC algorithm in noisy environments can be achieved through data preprocessing such as noise filtering or data smoothing to reduce the impact of noise. Moreover, by adjusting the density threshold λ using the adaptive parameter estimation as described in Section 3.1, the robustness of the proposed method to noise could be improved. The process described in [32] also can be performed to estimate the value of the density threshold λ. In general, the density threshold λ is set to 0 by default.

6.2. Conclusions

In this paper, we propose a novel clustering algorithm (SDBC) to identify significant clusters of arbitrary shapes and different densities in geographical flow data with noise. The method identifies the high-density flows as seeds for density-growing clustering based on local spatial statistics and permutation tests. We defined the distance between the flows and extended the Getis-Ord G_i^* statistic for the geographical flows. This extension enables us to identify regions of high-density flow clusters. The clustering results on simulated datasets and real-world taxi trajectory data demonstrated that the proposed method can effectively identify spatial flow clusters of different shapes and densities. Moreover, the proposed method exhibits superior performance for detecting significant flow clusters from data with noise compared to the available state-of-the-art spatial flow clustering methods.

Nevertheless, some limitations of the proposed method also exist. Firstly, considering the computational efficiency and time complexity of the proposed method, future work can further reduce the computational complexity and realize the parallelization of the proposed method for geographical flow big data. For instance, employing spatial index structures like R-tree or kd-tree can enhance the efficiency of neighbor querying. Although parallel processing can significantly increase computational speed, data consistency and synchronization issues deserve further investigation. In distributed systems, improper data handling can affect the accuracy of the final results. Therefore, in practical applications, it's necessary to optimize and improve the algorithm according to the specific needs of the scenario, achieving a balance between the performance, accuracy, and implementation complexity. Secondly, although the current method for selecting parameters R and θ is effective, it may not be universally optimal across different types of flow data, especially in diverse urban environments with varying flow characteristics. Therefore, further work could focus on developing a method for more adaptive parameter determination. This may involve integrating with various features of the flows, such as the flow length and direction changes, and using machine learning models to automatically adjust parameters to obtain more accurate flow patterns in different data scenarios. Additionally, the dynamic neighborhood updating mechanisms based on real-time data analysis to adapt to immediate changes in urban travel flows, enhancing the timeliness and accuracy of clustering results. Moreover, defining variable spatial neighborhoods for different flows according to the variance of density may also help to better identify flow clusters of variable density. The study of dynamic neighborhood updating mechanisms for spatiotemporal flow data could promote the proposed method to adapt to real-time change patterns recognition of the flow clusters, which is our future research work. Lastly, considering urban flow patterns at different levels or scales, based on the proposed method, the development of multi-scale flow data clustering algorithms is also worth studying, to not only capture the large-scale (macro) but also the local (micro) flow clustering patterns, which can help to improve the understanding of human multi-scale movement patterns in space.

CRediT authorship contribution statement

Huimin Liu: Writing – review & editing, Conceptualization. **Chen Ding:** Visualization, Software, Data curation. **Ju Peng:** Visualization, Software, Data curation. **Xiaoming Mei:** Writing – review & editing, Conceptualization. **Yuxin Zhao:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Xuexi Yang:** Writing – review & editing, Writing – original draft, Visualization, Resources, Methodology, Data curation, Conceptualization. **Min Deng:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Jianbo Tang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

Acknowledgements

This study was supported by the Funds of National Key Research and Development Program of China (No. 2022YFB3904203), National Science Foundation of China (Nos. 42271462, 42271485, 42171459, 42171441), the Hunan Provincial Natural Science Foundation of China (No. 2024JJ1009, 2022JJ40585, 2022JJ30703, 2024JJ8343), the Hunan Province Natural Resources Science and Technology Project (20230121XX), and the Frontier Cross Research Project of Central South University (2023QYJC002).

References

[1] X. Yan, T. Pei, C. Song, X. Liu, Estimating spatiotemporal aggregation scales by revisiting the spatiotemporal L-function, Trans. GIS 27 (2) (2023) 592–604, https://doi.org/10.1111/tgis.13034.
[2] G. Andrienko, N. Andrienko, G. Fuchs, J. Wood, Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data, IEEE Trans. Vis. Comput. Graph. 23 (9) (2016) 2120–2136.
[3] J. Cuenca-Jara, F. Terroso-Sáenz, M. Valdés-Vela, A. Skarmeta, Classification of spatio-temporal trajectories from volunteer geographic information through fuzzy rules, Appl. Soft Comput. 86 (2020) 105916, https://doi.org/10.1016/j.asoc.2019.105916.
[4] S. Dutta, A. Das, B. Patra, CLUSTMOSA: Clustering for GPS trajectory data based on multi-objective simulated annealing to develop mobility application, Appl. Soft Comput. 130 (2022) 109655, https://doi.org/10.1016/j.asoc.2022.109655.
[5] X. Zhu, D. Guo, Mapping large spatial flow data with hierarchical clustering, Trans. GIS 18 (3) (2014) 421–435.
[6] A.T. Murray, Y. Liu, S.J. Rey, L. Anselin, Exploring movement object patterns, Ann. Reg. Sci. 49 (2011) 471–484, https://doi.org/10.1007/s00168-011-0459-z.
[7] R. Tao, J.C. Thill, Spatial cluster detection in spatial flow data, Geogr. Anal. 48 (4) (2016) 355–372.
[8] C. Song, T. Pei, T. Ma, Y. Du, H. Shu, S. Guo, Z. Fan, Detecting arbitrarily shaped clusters in origin-destination flows using ant colony optimization, Int. J. Geogr. Inf. Sci. 33 (1) (2019) 134–154. ⟨http://ir.igsnrr.ac.cn/handle/311030/51359⟩.
[9] Q. Liu, J. Yang, M. Deng, C. Song, W. Liu, SNN_flow: a shared nearest-neighbor-based clustering method for inhomogeneous origin-destination flows, Int. J. Geogr. Inf. Sci. 36 (2) (2021) 253–279, https://doi.org/10.1080/13658816.2021.1899184.
[10] Z. Fang, X. Yang, Y. Xu, S.-L. Shaw, L. Yin, Spatiotemporal model for assessing the stability of urban human convergence and divergence patterns, Int. J. Geogr. Inf. Sci. 31 (11) (2017) 2119–2141, https://doi.org/10.1080/13658816.2017.1346256.
[11] J. Peng, H. Liu, J. Tang, C. Peng, X. Yang, M. Deng, Y. Xu, Exploring crowd travel demands based on the characteristics of spatiotemporal interaction between urban functional zones, ISPRS Int. J. Geo-Inf. 12 (2023) 225, https://doi.org/10.3390/ijgi12060225.
[12] H. Shu, T. Pei, C. Song, X. Chen, S. Guo, Y. Liu, J. Chen, X. Wang, C. Zhou, L-function of geographical flows, Int. J. Geogr. Inf. Sci. 35 (4) (2021) 689–716, https://doi.org/10.1080/13658816.2020.1749277.
[13] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, 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, 2–4, August, Portland, OR (1996) 226–231.
[14] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, OPTICS: Ordering Points To Identify the Clustering Structure. Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, 1999, pp. 49–60.
[15] T. Pei, A. Jasra, D.J. Hand, A.-X. Zhu, C. Zhou, DECODE: a new method for discovering clusters of different densities in spatial data, Data Min. Knowl. Discov. 18 (2009) 337–369, https://doi.org/10.1007/s10618-008-0120-3.
[16] M. Nanni, D. Pedreschi, Time-focused clustering of trajectories of moving objects, J. Intell. Inf. Syst. 27 (3) (2006) 267–289.
[17] C. Song, T. Pei, H. Shu, Identifying flow clusters based on density domain decomposition, IEEE Access 8 (2019) 5236–5243.
[18] X. Guo, Z. Xu, J. Zhang, J. Lu, H. Zhang, An OD flow clustering method based on vector constraints: a case study for Beijing taxi origin-destination data, ISPRS Int. J. Geo-Inf. 9 (2) (2020) 128.
[19] X. Yao, D. Zhu, Y. Gao, L. Wu, P. Zhang, Y. Liu, A stepwise spatio-temporal flow clustering method for discovering mobility trends, IEEE Access 6 (2018) 44666–44675, https://doi.org/10.1109/ACCESS.2018.2864662.
[20] R. Tao, J.-C. Thill, C. Depken, M. Kashiha, flowHDBSCAN: a hierarchical and density-based spatial flow clustering method, UrbanGIS17 (2017) 1–8.
[21] Y. Liu, D. Tong, X. Liu, Measuring spatial autocorrelation of vectors, Geogr. Anal. 47 (3) (2015) 300–319.
[22] S. Berglund, A. Karlström, Identifying local spatial association in flow data, J. Geogr. Syst. 1 (3) (1999) 219–236.
[23] R. Tao, J.-C. Thill, Flow cross K-function: a bivariate flow analytical method, Int. J. Geogr. Inf. Sci. 33 (10) (2019) 2055–2071.
[24] Z. Kan, M.-P. Kwan, L. Tang, Ripley's K-function for network-constrained flow data, Geogr. Anal. 54 (4) (2021) 769–788.
[25] P. Moran, Notes on continuous stochastic phenomena, Biometrika 37 (1) (1950) 17–23.
[26] Y. Gao, T. Li, S. Wang, M.-H. Jeong, K. Soltani, A multidimensional spatial scan statistics approach to movement pattern comparison, Int. J. Geogr. Inf. Sci. 32 (7) (2018) 1304–1325.
[27] Q. Liu, J. Yang, M. Deng, W. Liu, R. Xu, BiFlowAMOEBA for the identification of arbitrarily shaped clusters in bivariate flow data, Int. J. Geogr. Inf. Sci. 36 (9) (2022) 1784–1808, https://doi.org/10.1080/13658816.2022.2072850.
[28] M. Kulldorff, A spatial scan statistic, Commun. Stat. - Theory Methods 26 (6) (1997) 1481–1496.
[29] R. Tao, J.-C. Thill, flowAMOEBA: identifying regions of anomalous spatial interactions, Geogr. Anal. 51 (1) (2019) 111–130, https://doi.org/10.1111/gean.12161.
[30] G. Castellano, E. Cotardo, C. Mencar, G. Vessio, Density-based clustering with fully-convolutional networks for crowd flow detection from drones, Neurocomputing 526 (2023) 169–179.
[31] T. Pei, A nonparametric index for determining the numbers of events in clusters, Math. Geosci. 43 (3) (2011) 345–362, https://doi.org/10.1007/s11004-011-9325-x.
[32] J.K. Ord, A. Getis, Local spatial autocorrelation statistics: distributional issues and an application, Geogr. Anal. 27 (4) (1995) 286–306.
[33] J.C. Duque, J. Aldstadt, E. Velasquez, J.L. Franco, A. Betancourt, A computationally efficient method for delineating irregularly shaped spatial clusters, J. Geogr. Syst. 13 (4) (2011) 355–372.
[34] J.C. Huang, J.B. Tang, Discovery of arbitrarily shaped significant clusters in spatial point data with noise, Appl. Soft Comput. 108 (2021) 107452, https://doi.org/10.1016/j.asoc.2021.107452.
[35] R. Karne, S. Tk, Clustering algorithms and comparisons in vehicular ad hoc networks, Mesop. J. Comput. Sci. (2023) 121–129, https://doi.org/10.58496/MJCSC/2023/014.
[36] M.A. Mohammed, N. Ţăpuş, A novel approach of reducing energy consumption by utilizing big data analysis in mobile cloud computing, Mesop. J. Big Data (2023) 110–117, https://doi.org/10.58496/MJBD/2023/015.
[37] R. Zaib, O. Ourabah, Large scale data using K-means, Mesop. J. Big Data (2023) 36–45, https://doi.org/10.58496/MJBD/2023/006.
[38] D. Han, A. Agrawal, W.K. Liao, A. Choudhary, Parallel DBSCAN Algorithm Using a Data Partitioning Strategy with Spark Implementation. 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 305–312, https://doi.org/10.1109/BigData.2018.8622258, 2018.
[39] D. Birant, A. Kut, ST-DBSCAN: an algorithm for clustering spatial-temporal data, Data Knowl. Eng. 60 (1) (2007) 208–221.
[40] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1) (1985) 193–218.