IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 9, 2019, accepted December 26, 2019, date of publication December 30, 2019, date of current version January 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963107

Identifying Flow Clusters Based on Density Domain Decomposition

CI SONG, TAO PEI, AND HUA SHU

State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 101408, China

Corresponding author: Tao Pei (peit@lreis.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 41525004, Grant 41421001, and Grant 41601430, and in part by the Key Research Program of Frontier Science, Chinese Academy of Sciences under Grant QYZDY-SSW-DQC007.

ABSTRACT Flow clustering is one of the most important data mining methods for the analysis of origin-destination (OD) flow data, and it may reveal the underlying mechanisms responsible for the spatial distributions and temporal dynamics of geographical phenomena. Existing flow clustering approaches are based mainly on the extension of traditional clustering methods to points by redefining basic concepts or some spatial association indictors of flows and the implementation of classic clustering processes, such as aggregating, collecting or searching. However, current techniques still suffer from two main problems: poor identification accuracy and complicated parameter selection processes. To resolve these problems, a new clustering method is proposed in this study for arbitrarily shaped flow clusters based on the density domain decomposition of flows. Simulation experiments based on our method and existing methods show that our method outperforms the three most commonly used methods in terms of the overall identification rate and almost all F1 measures, and it does not require any manual adjustments during the parameter selection process. Finally, a case study is conducted on taxi trip data from Beijing. Several flow clusters are identified to represent different types of residents' travel behaviors, including daily commuting, return travel, tourism and behaviors on special days.

INDEX TERMS Origin-destination (OD) flow, flow space, flow clustering, density domain decomposition, point process.

I. INTRODUCTION

The movement of a geographical object between two locations (e.g., the daily commute from dwelling to workplace [1], immigration between states [2] or delivery services in a city [3]) can be presented as an origin-destination (OD) flow [4], [5]. Such flows, which include human, commodity, information, capital and relationship flows, can be thought of as interactions or relationships between two georeferenced places and may reflect the underlying mechanisms responsible for the spatial distributions and temporal dynamics of geographical phenomena [6]–[8].

In recent years, with the emergence of different kinds of mobile devices, large amounts of OD flow data have emerged, including mobile phone signal data [9], GPS trajectories [10], footprints [11], WiFi positioning trace data [12], logistics records [13], and transaction records [14]. Therefore, many spatial analysis and geodata mining methods, such as abnormal flow detection [15], flow cluster identification [6], [16], [17], and flow estimation or prediction techniques [18]–[20], have been developed to discover patterns in these OD flow data. Among these methods, flow clustering is commonly used to discover the distribution characteristics of flows. In early years, it was mainly used to evaluate the global roles of cities in transportation [21] or improve our understanding of the geographical patterns in residents' mobility [22], [23]. However, in recent years, this method has been extended to detect malaria hotspots in the epidemiology field [24], to analyze saving propensities and wealth distributions [25], to understand the interdisciplinary nature of knowledge absorption [26], and many other fields [27], [28].

The associate editor coordinating the review of this manuscript and approving it for publication was Corrado Mencar.

The main principle of generating a flow clustering algorithm is to extend the traditional point clustering algorithms by redefining some basic concepts (such as distance, density and reachability) or some spatial association indictors. On this basis, current flow clustering methods can be grouped into three categories: hierarchical-based clustering, density-based clustering and statistics-based clustering. In hierarchical clustering methods for flow data, the distance of an OD flow should be defined according to the OD locations [29], [30] and, sometimes, the attributes of the flows [6], [31]; furthermore, an agglomerative or divisive strategy should be used to organize each flow into a hierarchy [32], [33]. These methods can identify flow clusters at different spatial scales, and they are usually proposed to solve problems associated with flow cluster identification, generalization and visualization [34]–[37]. In density-based clustering methods for flow data, the main aspiration is to find high-density subsets of flow data [38]–[40]; accordingly, the definition of the local density of an OD flow should also be defined based on the quantity or reachability of each flow. Then, a traditional algorithm such as DBSCAN [41] or OPTICS [42] can be extended to identify flow clusters based on a density-connected process. Moreover, these methods are insensitive to outlier flows because unnecessary noise can be eliminated since not all flow data need to be clustered [43]. Finally, in statistics-based clustering methods, the definitions of spatial association indictors or other statistic measures, such as Moran's I [29], [44], Getis-Ord's G [8], Ripley's K-function [6], and the log-likelihood ratio [30], can be extended to describe the local aggregative characteristics of the flow subsets. The objective of these methods is usually to find the subset of flows with the optimal values of statistical measures. These methods can usually create a significant description of spatial homogeneity, thereby providing a standard measure for comparing flow clusters [45]–[47].

Although many flow clustering methods have been proposed in recent years, some problems remain unsolved. First and foremost is the identification rate for arbitrarily shaped flow clusters. Most of the above methods are very useful at identifying certain types of flow clusters. However, when encountering some irregularly shaped flow clusters, these methods may not be as effective [30]. Second, in most cases, the parameters need to be manually selected, which is often difficult. Determination of an unknown number of flow clusters is always a troublesome problem, and some other scale parameters, such as distance threshold and neighborhood range, are hard to set in clustering algorithms.

To solve these problems, in this study, we propose a clustering method for arbitrarily shaped flow clusters based on the density domain decomposition of flows. In our method, a flow dataset is assumed to be composed of clusters with high-density flows and noise with low-density flows. A mixed probability density model of k-th nearest neighbor distances is used to separate clusters and noise. The model parameters can be estimated through an EM algorithm, and thus, flow clusters can be determined. This method is believed to automatically identify arbitrarily shaped flow clusters with high accuracy.

The remainder of this paper is arranged as follows. Section II introduces some basic concepts about flows. Section III describes the proposed method in detail. Section IV presents the simulation experiment and compares our method with popular methods presented in previous studies. Section V presents a case study involving taxi trip data in Beijing. Finally, Section VI provides the conclusions and future work.

II. BASIC CONCEPTS ABOUT FLOWS

Before we describe the details of our method, several basic concepts about OD flows must be introduced.

Definition 1 (OD Flow and Flow Space): An OD flow is defined as two tuples consisting of a 2-D origin point and a 2-D destination point in flow space. An OD flow can be expressed as $f =< \mathbf{x}^O, \mathbf{x}^D >$, where $\mathbf{x}^O = (x^O, y^O)$ and $\mathbf{x}^D = (x^D, y^D)$ denote the coordinates of the origin point and destination point, respectively. Therefore, the flow space is a metric space that is expressed as the Cartesian product of two 2-D planes ($\mathbf{R}^2 \times \mathbf{R}^2$), and each flow can be seen as a 4-D point in this flow space.

Definition 2 (Flow Distance): The flow distance is the fundamental measurement ρ in the flow space. Two types of distance measurements in flow space are defined as follows:

Chebyshev Distance:

$$\rho_c(f_i, f_j) = \max(\rho^O, \rho^D) = \max(|\mathbf{x}_i^O - \mathbf{x}_j^O|, |\mathbf{x}_i^D - \mathbf{x}_j^D|) \quad (1)$$

Manhattan Distance:

$$\rho_m(f_i, f_j) = \rho^O + \rho^D = |\mathbf{x}_i^O - \mathbf{x}_j^O| + |\mathbf{x}_i^D - \mathbf{x}_j^D| \quad (2)$$

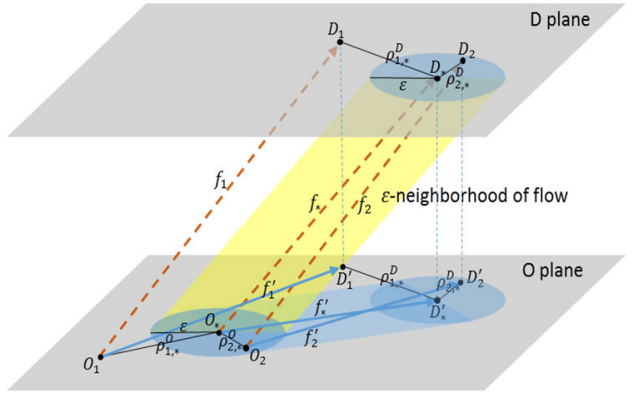
where ρ^O (or ρ^D) represents distance between origin points (or destination points).

Definition 3 (ε-Neighborhood of Flow): The ε-neighborhood of flow, denoted $N_\varepsilon(f)$, is defined by $N_\varepsilon(f) = \{f' | \rho(f, f') \leq \varepsilon\}$. This ε-neighborhood can be expressed as a flow sphere with center f and radius ε in the flow space. The volume of this sphere, denoted $V_{N_\varepsilon(f)}$, can be calculated as follows:

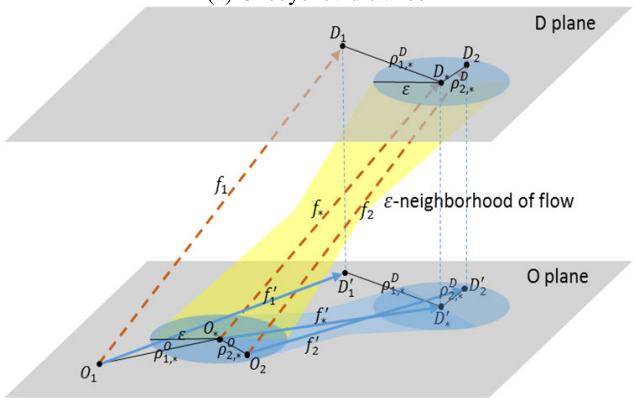
$$V_{N_\varepsilon(f)} = \int\int_{N_\varepsilon(f)} \rho^O \rho^D \mathbf{d}\rho^O \mathbf{d}\rho^D \iint \mathbf{d}\theta^O \mathbf{d}\theta^D \quad (3)$$

where (ρ^O, θ^O) and (ρ^D, θ^D) are the polar coordinates of the 2-D origin plane and the 2-D destination plane, respectively. Based on the above two definitions of the flow distance, the volumes are $\pi^2\varepsilon^4$ for the Chebyshev distance and $\pi^2\varepsilon^4/6$ for the Manhattan distance. Figure 1 shows the ε-neighborhood of flow based on both types of distances.

Definition 4 (Flow Density): The flow density, denoted λ, is the number of flows per unit volume. For a flow zone (a subset of the flow space) $\mathbf{z} \subset \mathbf{R}^2 \times \mathbf{R}^2$, the flow density $\lambda(\mathbf{z})$ can be calculated as $\lambda(\mathbf{z}) = n_\mathbf{z}/V_\mathbf{z}$, where $n_\mathbf{z} = |\{f | f \in \mathbf{z}\}|$ is the number of flows in $\mathbf{z}$ and $V_\mathbf{z}$ is the volume of $\mathbf{z}$. For a



(a) Chebyshev distance



(b) Manhattan distance

FIGURE 1. ε-Neighborhood of flow f_* based on two types of distance (the yellow part). The blue areas in O plane are the projected ε-Neighborhood of flow f_* in 2D-plane. The red dash arrows (f_*, f_1, f_2) are flows in flow space and the blue solid arrows (f'_*, f'_1, f'_2) are the projected flows in 2D-plane. ρ^O (or ρ^D) represents distance between origin points (or destination points). In this figure, $\rho^O_{1,*}$ and $\rho^D_{1,*}$ are both larger than ε, and therefore, f_1 is not in the ε-neighborhood of flow f_*. $\rho^O_{2,*}$ and $\rho^D_{2,*}$ are both smaller than ε, and therefore, based on the Chebyshev distance, f_2 is in the ε-neighborhood of flow f_*. However, based on the Manhattan distance, f_2 is not in the ε-neighborhood of flow $f*$ for $\rho^O_{2,*} + \rho^D_{2,*} > \varepsilon$.

single flow, the local flow density $\lambda(f)$ can be calculated as follows:

$$\lambda(f) = \lim_{\varepsilon \to 0} \frac{n_{N_\varepsilon(f)}}{V_{N_\varepsilon(f)}} \tag{4}$$

III. METHOD

Based on the concepts defined above, a density domain decomposition model of flows is proposed as an extension of the point process decomposition model [48] to identify flow clusters. This method can be divided into four steps, as shown in Figure 2. First, we determine whether the flow set is homogeneous by using several quantitative indices proposed in our previous work (such as NLH*, A-w) [47]. If the flow set is not homogeneous, we proceed to the second step. Otherwise, the flow set is considered homogeneous and cannot be decomposed. Second, a mixed probability density function (pdf) of the k-th nearest distances of flows is generated to describe the density domain model of the flow set. Third, the parameters of this pdf are evaluated by an

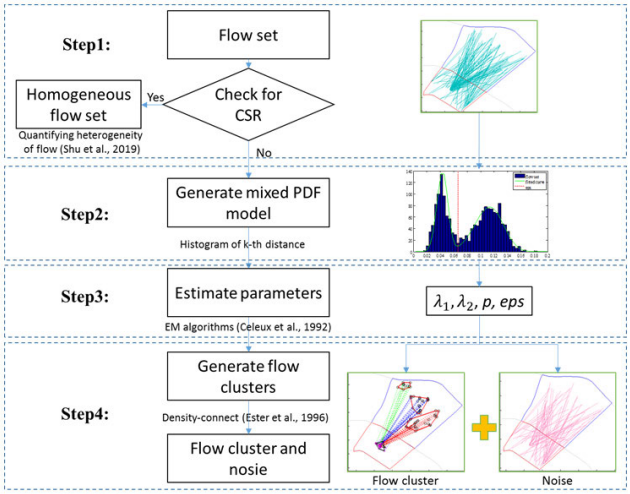


FIGURE 2. Density domain decomposition model of flows.

expectation-maximization (EM) algorithm, and all the flows are decomposed into two components with different densities that correspond either to dense flows or sparse flows. Each sparse flow is seen as noise, while each dense flow can be generated into flow clusters based on the density-connected clustering concept [36] in the final step. Since our previous work [47] provides the details of the first step, we introduce only the remaining three steps.

A. MIXED PDF OF THE K-TH DISTANCES OF FLOWS

For one homogeneous flow set, where the flow density is $\lambda(f) \equiv \lambda$, the probability distribution of the k-th nearest flow distances F_{D_k} can be acquired by traversing the pdf that includes 0, 1, 2, …, k-1 flows within the D_k-neighborhood:

$$F_{D_k}(\varepsilon) = 1 - P(D_k \geq \varepsilon) = 1 - P(n_{N_\varepsilon} \leq k) \tag{5}$$

where k is the ordinal number of nearest neighbors and n_{N_ε} is the number of flows in the ε-neighborhood. In this case, $P(n_{N_\varepsilon} = k)$ follows a Poisson distribution and is expressed as $P(n_{N_\varepsilon} = k) = e^{-\lambda V_{N_\varepsilon}} (\lambda V_{N_\varepsilon})^k / k!$, and the pdf $f_{D_k}(\varepsilon)$ is the derivative of $F_{D_k}(\varepsilon)$:

$$f_{D_k}(\varepsilon) = \frac{\mathbf{d}F_{D_k}(\varepsilon)}{\mathbf{d}\varepsilon} = \lambda \frac{e^{-\lambda V_{N_\varepsilon}} (\lambda V_{N_\varepsilon})^{k-1}}{(k-1)!} \frac{dV_{N_\varepsilon}}{d\varepsilon} \tag{6}$$

Therefore, the mixed pdf of the k-th nearest flow distances of the two homogeneous flow sets with different densities, for example, λ_1 and λ_2 ($\lambda_1 > \lambda_2$), can be expressed as follows:

$$D_k(\varepsilon) = p f_{D_k}(\varepsilon | k, \lambda_1 \pi) + (1-p) f_{D_k}(\varepsilon | k, \lambda_2 \pi) \tag{7}$$

where p is the proportion rate of flow clusters, λ_1 is the density of the flow cluster and λ_2 is the density of noise. Since a flow set can be seen as a mixture of flow clusters and noise, equation (7) can be used to describe the density domain model of flows.

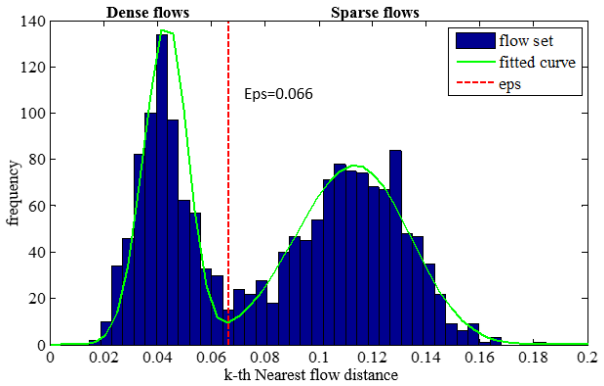


FIGURE 3. Histogram of the k-th nearest flow distances for distinguishing dense flows from sparse flows. The yellow curve represents the fitted curve of the histogram. The red line is the boundary that divides dense flows from spare flows, where the probability $\hat{\delta}_i^{t+1}$ equals 0.5.

B. PARAMETER EVALUATION FOR DECOMPOSING FLOWS OF DIFFERENT DENSITIES

Once the density domain model of flows is generated, the next step is to evaluate the parameters and decomposing flows of different densities. In this process, an EM algorithm [49], [50] is applied to evaluate the parameters $(\lambda_1, \lambda_2, p)$ based on the histogram of the observed k-th nearest flow distances (Fig. 3). A summary of the algorithm can be seen as follows:

E-Step:

$$E(\hat{\delta}_i^{t+1}) = \frac{\hat{p}^t f_{D_k}(\varepsilon_i|k, \hat{\lambda}_1^t)}{\hat{p}^t f_{D_k}(\varepsilon_i|k, \hat{\lambda}_1^t) + (1-\hat{p}^t) f_{D_k}(\varepsilon_i|k, \hat{\lambda}_2^t)} \tag{8}$$

M-Step:

$$\hat{\lambda}_1^{t+1} = \frac{k \sum_{i=1}^n \hat{\delta}_i^{t+1}}{\pi^2 \sum_{i=1}^n \varepsilon_i^4 \hat{\delta}_i^{t+1}},$$

$$\hat{\lambda}_2^{t+1} = \frac{k \sum_{i=1}^n (1-\hat{\delta}_i^{t+1})}{\pi^2 \sum_{i=1}^n \varepsilon_i^4 (1-\hat{\delta}_i^{t+1})}, \quad p^{t+1} = \frac{\sum_{i=1}^n \hat{\delta}_i^{t+1}}{n} \tag{9}$$

where n is the number of flows and t is the iteration time. $\hat{\delta}_i^{t+1}$ is the probability that flow f_i belongs to a dense flow. If $\hat{\delta}_i^{t+1} \geq 0.5$, flow f_i can be marked as a dense flow; otherwise, it is marked as a sparse flow. In this step, parameters λ_1, λ_2 and p can be estimated by the iteration process, and parameter k can be selected by the fitting accuracy of $D_k(\varepsilon)$. Therefore, the density domain model of flows can be determined, and flows can be decomposed into a dense part and a sparse part.

C. IDENTIFYING FLOW CLUSTERS BASED ON DENSITY -CONNECTED CLUSTERING

After decomposing process, we can filter out noise (sparse flows) from the flow set and obtain candidate features (dense flows) that can be collected into flow clusters. In this process, the classic DBSCAN algorithm can be improved for identification of flow clusters by modifying the **density-connected** concept of flows [41]. Here, we present the definition that flow f_p is **density-connected** to flow f_q with respect to (wrt) **Eps** and **MinPts** if there is a chain of flows $f_{p_1}, f_{p_2}, \ldots, f_{p_n}, p_1 = q, p_n = p$ such that $|N_{eps}(f_{p_i})| \geq \text{MinPts}(i = 2, 3, \ldots, n-1)$ and $f_{p_{i-1}} \in N_{Eps}(f_{p_i})$, $(i = 2, 3, \ldots, n)$. On this basis, a flow cluster can be identified as a set in which all features are density-connected to each other. Parameter **MinPts** is equal to k, and parameter **Eps** can be estimated by the following formula:

$$pf_{D_k}(eps|k, \lambda_1\pi) = (1-p) f_{D_k}(eps|k, \lambda_2\pi) \tag{10}$$

Then:

$$eps = \left(\frac{\ln \frac{1-p}{p} + k \ln \frac{\lambda_2}{\lambda_1}}{\pi^2(\lambda_2 - \lambda_1)} \right)^{1/4} \tag{11}$$

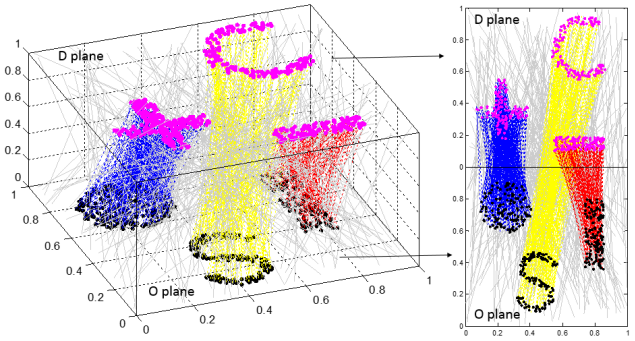


FIGURE 4. Simulated dataset consisting of noise and three flow clusters. The flow space is described as a "box" in the left figure, in which the bottom plane stands for the "O plane" and the top plane stands for the "D plane". The origin points (black points) and destination points (magenta points) of the flows are all limited to $[0,1]^2$ 2-D planes. Noise is presented as gray lines, and C-1 (138 flows), C-2 (163 flows), and C-3 (376 flows) are presented as red lines, yellow lines and blue lines, respectively. The density of noise is approximately 10^3, and the density of clusters is approximately 10^5.

IV. SIMULATION EXPERIMENTS

In this section, we design and analyze a Monte Carlo simulation experiment with 100 sets of simulated flow data to validate our method. Each dataset is composed of noise and three clusters with different shapes (i.e., bar-strip, "S"-"C" and "O"-"+"), coded as C-1, C-2 and C-3, respectively. The density of each flow cluster is much higher than that of the noise (100 times greater). An example of the simulated dataset is displayed in Figure 4. In this experiment, each dataset is processed by our method, and a significance test is designed to validate all the identified flow clusters. Through the test, the A-w statistic of flow clusters [47] is used to test whether the density of flow clusters is different from that of noise clusters.

These simulated datasets are also evaluated by three other flow clustering methods for comparison: a hierarchical clustering method, a density-based clustering method and a spatial statistics-based clustering method. For the hierarchical clustering method, we use the algorithm proposed by Guo in 2014 [7]. This algorithm iteratively merges flows to form

TABLE 1. Experimental settings for three methods.

Method	Distance Measurement	Clustering Strategy	Parameter Setting
Hierarchical-based	Shared-Nearest-Neighbor (**SNN**)*	Merge nearest clusters iteratively	$k = 50$
Density-based	Chebyshev distance	OPTICS	$\min pts = \lceil \log(n) \rceil$ $eps = 0.12$
Statistics-based	Chebyshev distance	Merge neighbors with maxL	$\max L = 0.16$

$$^{*}\mathbf{SNN}(f_p, f_q) = 1 - \frac{|KNN(\mathbf{x}_p^O, k) \cap KNN(\mathbf{x}_q^O, k)|}{k} \times \frac{|KNN(\mathbf{x}_p^D, k) \cap KNN(\mathbf{x}_q^D, k)|}{k}$$

a hierarchy of flow clusters and is believed to be effective at aggregating spatial flows and simplifying flow sets into groups [37]. The only parameter k is set to ensure that, on average, each flow has five flow neighbors. For the density-based clustering method, we apply the classic trajectory clustering technique proposed in 2006 [39]. This method takes each flow as a trajectory with only two points and adopts an improved OPTICS algorithm to identify the flow clustering structure. Parameter **MinPts** is empirically set as and parameter **Eps** is set as average k-th distance of flows. For the statistics-based clustering method, we use a local version of the L-function (the L-function is the normalization of Ripley's K-function) to identify flow clusters within the simulated dataset. This algorithm first identifies the aggregation scale parameter (maxL) based on the global L-function and then calculates the local L values using this scale. Finally, flows with top 1% local L values are merged as the dominant cluster. This approach has been shown to be useful for detecting spatial clustering patterns in flow data [6], [31]. These experimental settings are listed in TABLE I.

Since all methods can successfully identify significant flow clusters, we analyze the average identification rates shown in TABLE 2 for comparison. From these results, we can see that our method outperforms the other methods in terms of the overall identification rate and almost all F1 measures. The other three methods have good recall rates, but, in general, their precisions are unsatisfactory (less than 90%). It is worth noting that the precision of our method is much higher than those of other methods, especially for the "S"-"C"-shaped flow cluster, and the shape of the identified flow cluster is reconstructed very well (Fig. 5). To identify irregularly shaped flow clusters, higher precision means that the original shape of the flow cluster can be better maintained. Thus, despite a few defects in the recall rate, we believe that our method is generally superior at identifying irregularly shaped flow clusters.

V. CASE STUDY

A. DATA DESCRIPTION

We apply our flow clustering method to taxi trip data in Beijing to identify different flow cluster patterns of daily traffic. Our dataset contains records of GPS trajectories from more than 25,000 taxis (more than 1/3 of all taxis in Beijing). Each record is described by five fields: <taxi ID, current time, longitude, latitude, status>. Thus, the OD flows from each taxi can be extracted according to changes in status.

TABLE 2. Comparison among the flow cluster results from our method and other methods.

Method	Cluster	Nflow	Precision	Recall	F1
Our method	bar - strip	138	**95.77%**	99.28%	**97.49%**
	"S" - "C"	163	**93.42%**	83.63%	88.25%
	"O" - "+"	376	**96.32%**	99.49%	**97.88%**
	All	677	**95.56%**	**95.63%**	**95.59%**
Hierarchical-based	bar - strip	138	86.60%	**99.99%**	92.82%
	"S" - "C"	163	80.35%	93.08%	86.25%
	"O" - "+"	376	94.81%	99.43%	97.06%
	All	677	89.26%	98.01%	93.43%
Density-based	bar - strip	138	86.67%	99.89%	92.81%
	"S" - "C"	163	83.68%	**93.94%**	**88.52%**
	"O" - "+"	376	91.61%	99.97%	95.61%
	All	677	87.94%	98.52%	92.93%
Statistics-based	bar - strip	138	84.91%	99.97%	91.83%
	"S" - "C"	163	83.88%	91.88%	87.70%
	"O" - "+"	376	89.75%	**100.00%**	94.60%
	All	677	87.27%	98.04%	92.34%

Bold numbers denote the highest value in comparison with the counterparts.

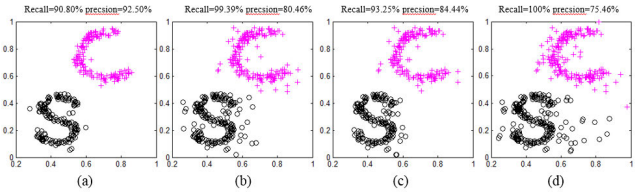


FIGURE 5. Examples of different methods for the identification of "S"-"C"-shaped flow clusters. The black circles imply the origin points of the identified flows, and the magenta crosses imply the destination points of the identified flows. (a) Our method; (b) hierarchical-based method; (c) density-based method; (d) statistics-based method.

Here, we choose six region pairs as study areas within Beijing during different periods to discover different flow patterns of residents' travel behaviors in Beijing (Figure 6). The first two region pairs, A and B, are two main commuter flows with short distances distributed between the east and west districts of Beijing, respectively. Next, two region pairs, C and D, are return flows from certain transportation hubs, e.g., an airport (C) and Beijing South Railway Station (D). The last two region pairs, E and F, are traffic flows on two special days, e.g., National Day (E) and Qingming Festival (F). TABLE 3 shows the study areas and descriptions of the data used in our case study.

B. FLOW CLUSTER RESULTS

The flow cluster results are shown in Figure 7. In the morning peak commuter flows of Wangjing, flow clusters from three different residential communities to the Taiyanggong subway station are identified (Fig. 7a). The origins of these flow clusters include school district communities

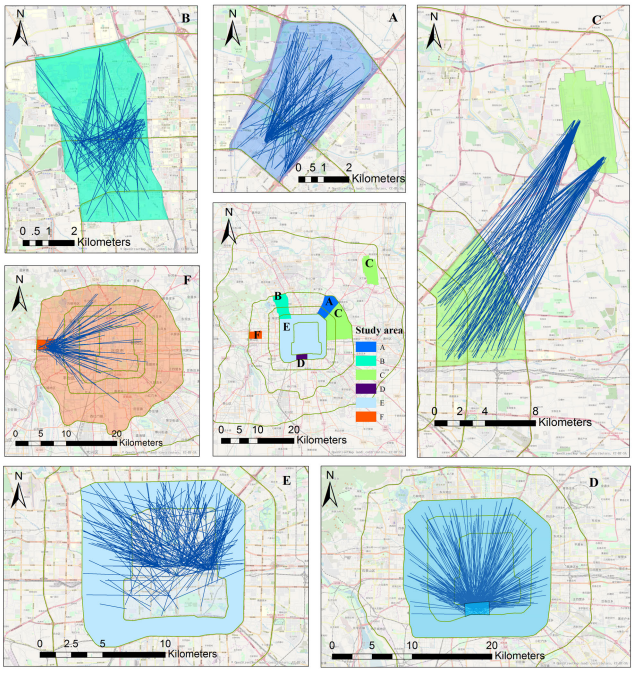


FIGURE 6. Study areas and OD flows.

TABLE 3. Study area 2 and data description.

ID	OD region	Period	Restriction
A	From Wangjing to Taiyanggong	Weekdays 7 am-9 am	East to West
B	From university area to Zhongguancun	Weekdays 7 am-12 am	North to South
C	From airport to areas outside the East 3rd Ring Road	Weekdays 9 pm-12 pm	North to South
D	From Beijing South Railway Station to 4th Ring Road	Weekdays 9 pm-12 pm	South to North
E	From the 3rd Ring Road to the 2nd Ring Road	National Day 4 am-6 am	North to South
F	From the 5th Ring Road to Babaoshan	Qingming Festival 4 am-10 am	South to North

Bold numbers denote the highest value in comparison with the counterparts

along the Wangjing North Road (Wangxin Garden and Shangjing New Route, blue flow clusters in Fig. 7a), residential communities in Wangjingxiyuayan (yellow flow clusters in Fig. 7a) and business-living buildings in the Huajiadi community. The Taiyanggong subway station (subway line 10) is commonly used by residents in the Wangjing area. For commuter flows from the university area to Zhongguancun, only one flow cluster is identified, from Wudaokou to Zhongguancun (Fig. 7b). The origins of this flow cluster are mainly distributed throughout the Dongsheng Science and Technology Park and Wudaokou commercial district, whereas the destinations are relatively dispersed, mainly distributed in the commercial center and residential communities in Zhongguancun. These results reflect the daily commuting behaviors in northwestern and northeastern Beijing.

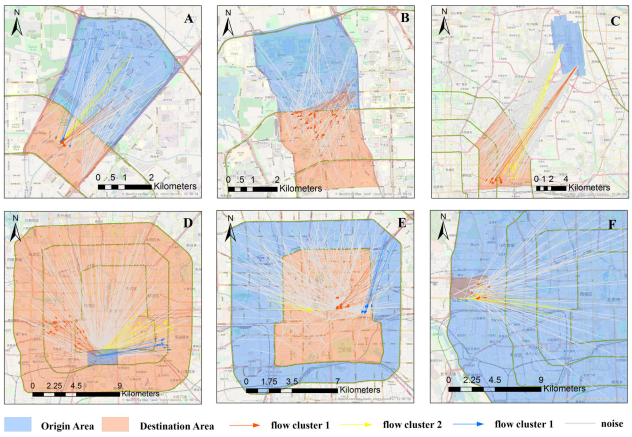


FIGURE 7. OD flow clusters in six different study areas of Beijing.

Figures 7c and 7d show the return flows from two transportation hubs at night on the weekdays. Two flow clusters are identified from the airport to areas outside the East 3rd Ring Road (Fig. 7c). The destinations are distributed in the residential communities near the Sihui Bridge (red arrows in Fig. 7c) and Chaoyang Joy City, including the Yuanyangguoji community, Ciyunli community and Pearl Rome Jiayuan community. For the return flows from Beijing South Railway Station, three flow clusters are identified; the destinations of these flow clusters are mainly distributed in Beijing West Railway Station, the communities around the Guangqumen Bridge and the Panjiayuan community. From these results, we can see that the return flows from the airport are more concentrated than the flows from a railway transportation hub.

Figures 7e and 7f show the flow clusters on two special days. Flows from the 3rd Ring Road to the 2nd Ring Road are composed of three flow clusters in Figure 7e. The red flow cluster and yellow flow cluster are morning tourist flows from residential communities along the East and West 3rd Ring Roads to Tiananmen, respectively, whereas the blue flow cluster may represent travel flows from the Sanyuan Bridge to Beijing Railway Station. The flows in Figure 7f mainly represent trips to Babaoshan for those seeking to sweep graves during the Qingming Festival. The origins of these flow clusters are mainly distributed in several communities near Wukesong and Beijing West Railway Station, including the Mingrijiayuan, Jiujiefang, and Muxidinanli communities. These results reflect purposeful travel behaviors on special days.

VI. CONCLUSION AND FUTURE WORK

In this study, we propose a method for flow clustering based on the density domain decomposition of flows. This method identifies arbitrarily shaped flow clusters with high accuracy and does not need any parameters in the clustering process. Simulation experiments show that our method outperforms three commonly used methods in terms of the overall identification rate and almost all F1 measures. The proposed method is applied to different taxi data in Beijing as a case study, and

it can also be easily extended to other OD flow data such as resident travel paths and migration and logistics data, which may help to provide information for urban management and region planning.

However, this method has some limitations. First, the parameter estimation process is very time consuming because we must traverse all possible parameters and then choose the optimal one for decomposing the flow set. Second, in our method, the flow set is assumed to be composed of two components, features and noise. Thus, some OD flows with different densities may not be separated because they are all considered features. Future research will focus on the analysis of superposed flow sets, which may contain more than two homogeneous flow sets with different densities, and on improving the calculation efficiency.

REFERENCES

[1] Y. Liu, F. Wang, Y. Xiao, and S. Gao, "Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai," *Landscape Urban Planning*, vol. 106, no. 1, pp. 73–87, May 2012.

[2] Y. Chun and D. A. Griffith, "Modeling network autocorrelation in space–time migration flow data: An eigenvector spatial filtering approach," *Ann. Assoc. Amer. Geograph.*, vol. 101, no. 3, pp. 523–536, Apr. 2011.

[3] R. Ducret, B. Lemarié, and A. Roset, "Cluster analysis and spatial modeling for urban freight. Identifying homogeneous urban zones based on urban form and logistics characteristics," *Transp. Res. Procedia*, vol. 12, pp. 301–313, Jan. 2016.

[4] M. Castells, "Grassrooting the space of flows," *Urban Geogr.*, vol. 20, no. 4, pp. 294–302, May 1999.

[5] M. F. Goodchild, M. Yuan, and T. J. Cova, "Towards a general theory of geographic representation in GIS," *Int. J. Geograph. Inf. Sci.*, vol. 21, no. 3, pp. 239–260, Mar. 2007.

[6] R. Tao and J.-C. Thill, "Spatial cluster detection in spatial flow data," *Geograph. Anal.*, vol. 48, no. 4, pp. 355–372, Nov. 2016.

[7] X. Zhu and D. Guo, "Mapping large spatial flow data with hierarchical clustering," *Trans. GIS*, vol. 18, no. 3, pp. 421–435, Jun. 2014.

[8] S. Berglund and A. Karlström, "Identifying local spatial association in flow data," *J. Geograph. Syst.*, vol. 1, no. 3, pp. 219–236, Oct. 1999.

[9] Z. Liu, T. Ma, Y. Du, T. Pei, J. Yi, and H. Peng, "Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records," *Trans. GIS*, vol. 22, no. 2, pp. 494–513, Apr. 2018.

[10] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Towards mobile intelligence: Learning from GPS history data for collaborative recommendation," *Artif. Intell.*, vols. 184–185, pp. 17–37, Jun. 2012.

[11] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services," in *Proc. 5th Int. Conf. Weblogs Social Media*, Barcelona, Spain, 2011, pp. 1–8.

[12] Y. Liu, T. Pei, C. Song, H. Shu, S. Guo, and X. Wang, "Indoor mobility interaction model: Insights into the customer flow in shopping malls," *IEEE Access*, vol. 7, pp. 138353–138363, 2019.

[13] P. Peng, J. P. Poon, Y. Yang, F. Lu, and S. Cheng, "Global oil traffic network and diffusion of influence among ports using real time data," *Energy*, vol. 172, pp. 333–342, Apr. 2019.

[14] D. Kondor, M. Pósfai, I. Csabai, and G. Vattay, "Do the rich get richer? An empirical analysis of the bitcoin transaction network," *PLoS ONE*, vol. 9, no. 2, Feb. 2014, Art. no. e86197.

[15] C. Chen, D. Zhang, P. Samuel Castro, N. Li, L. Sun, and S. Li, "Real-time detection of anomalous taxi trajectories from GPS traces," in *Mobile and Ubiquitous Systems: Computing, Networking and Services*, vol. 104, A. Puiatti and T. Gu, Eds. Berlin, Germany: Springer, 2012, pp. 63–74.

[16] Y. Gao, T. Li, S. Wang, M.-H. Jeong, and K. Soltani, "A multidimensional spatial scan statistics approach to movement pattern comparison," *Int. J. Geograph. Inf. Sci.*, vol. 32, no. 7, pp. 1304–1325, Jul. 2018.

[17] D. Guo, "Flow mapping and multivariate visualization of large spatial interaction data," *IEEE Trans. Vis. Comput. Graphics*, vol. 15, no. 6, pp. 1041–1048, Nov. 2009.

[18] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 186–194.

[19] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, and D. Zhang, "Prediction of urban human mobility using large-scale taxi traces and its applications," *Frontiers Comput. Sci.*, vol. 6, no. 1, pp. 111–121, 2012.

[20] M. M. Fischer and D. A. Griffith, "Modeling spatial autocorrelation in spatial interaction data: An application to patent citation data in the European union," *J. Regional Sci.*, vol. 48, no. 5, pp. 969–989, Dec. 2008.

[21] R. Guimera, S. Mossa, A. Turtschi, and L. A. N. Amaral, "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 22, pp. 7794–7799, May 2005.

[22] T. A. S. Nielsen and H. H. Hovgesen, "Exploratory mapping of commuter flows in England and Wales," *J. Transp. Geography*, vol. 16, no. 2, pp. 90–99, Mar. 2008.

[23] A. Rae, "From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census," *Comput., Environ. Urban Syst.*, vol. 33, no. 3, pp. 161–178, May 2009.

[24] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee, "Quantifying the impact of human mobility on Malaria," *Science*, vol. 338, no. 6104, pp. 267–270, Oct. 2012.

[25] M. Zanin, D. Papo, M. Romance, R. Criado, and S. Moral, "The topology of card transaction money flows," *Phys. A, Stat. Mech. Appl.*, vol. 462, pp. 134–140, Nov. 2016.

[26] M. Liu, X. Hu, and J. Li, "Knowledge flow in China's humanities and social sciences," *Quality and Quantity*, vol. 52, pp. 607–626, 2018.

[27] C. Andris, X. Liu, and J. Ferreira, "Challenges for social flows," *Comput., Environ. Urban Syst.*, vol. 70, pp. 197–207, Jul. 2018.

[28] P. Peng, Y. Yang, S. Cheng, F. Lu, and Z. Yuan, "Hub-and-spoke structure: Characterizing the global crude oil transport network with mass vessel trajectories," *Energy*, vol. 168, pp. 966–974, Feb. 2019.

[29] Y. Liu, D. Tong, and X. Liu, "Measuring spatial autocorrelation of vectors," *Geogr. Anal.*, vol. 47, no. 3, pp. 300–319, Jul. 2015.

[30] C. Song, T. Pei, T. Ma, Y. Du, H. Shu, S. Guo, and Z. Fan, "Detecting arbitrarily shaped clusters in origin-destination flows using ant colony optimization," *Int. J. Geograph. Inf. Sci.*, vol. 33, no. 1, pp. 134–154, Jan. 2019.

[31] R. Tao and J. C. Thill, "A density-based spatial flow cluster detection method," in *Proc. Int. Conf. Gisci. Short Paper*, Berkeley, CA, USA, 2016, pp. 288–291.

[32] N. Adrienko and G. Adrienko, "Spatial generalization and aggregation of massive movement data," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 2, pp. 205–219, Feb. 2011.

[33] D. Guo, X. Zhu, H. Jin, P. Gao, and C. Andris, "Discovering spatial patterns in origin–destination mobility data," *Trans. GIS*, vol. 16, no. 3, pp. 411–429, Jun. 2012.

[34] D. Phan, L. Xiao, R. Yeh, P. Hanrahan, and T. Winograd, "Flow map layout," in *Proc. IEEE Symp. Inf. Vis. (INFOVIS)*, Minneapolis, MN, USA, Jan. 2006, pp. 29–34.

[35] I. Boyandin, E. Bertini, and D. Lalanne, "Using flow maps to explore migrations over time," in *Proc. 13th AGILE Int. Conf. Geograph. Inf. Sci. Geospatial Vis. Anal. Workshop*, Guimaraes, Portugal, 2010.

[36] J. Wood, J. Dykes, and A. Slingsby, "Visualisation of origins, destinations and flows with OD maps," *Cartograph. J.*, vol. 47, no. 2, pp. 117–129, May 2010.

[37] G. Andrienko, N. Andrienko, G. Fuchs, and J. Wood, "Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin–destination movement data," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 9, pp. 2120–2136, Sep. 2017.

[38] B. Zhu and Q. Huang, "Urban population migration pattern mining based on taxi trajectories," in *Proc. 3rd Int. Workshop Mobile Sens., Future, Brought Big Sensor Data*, Philadelphia, PA, USA, 2013.

[39] M. Nanni and D. Pedreschi, "Time-focused clustering of trajectories of moving objects," *J. Intell. Inf. Syst.*, vol. 27, no. 3, pp. 267–289, Nov. 2006.

[40] T. Pei, W. Wang, H. Zhang, T. Ma, Y. Du, and C. Zhou, "Density-based clustering for data containing two types of points," *Int. J. Geograph. Inf. Sci.*, vol. 29, no. 2, pp. 175–193, Feb. 2015.

[41] M. Ester, H. P. Kriegel, J. Sander, and X. Xiaowei, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, Portland, OR, USA, 1996, pp. 226–231.

[42] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. R. Sander, "OPTICS: Ordering points to identify the clustering structure," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Philadelphia, PA, USA, 1999, pp. 49–60.
[43] T. Pei, A. Jasra, D. J. Hand, A.-X. Zhu, and C. Zhou, "DECODE: A new method for discovering clusters of different densities in spatial data," *Data Mining Knowl. Discovery*, vol. 18, no. 3, pp. 337–369, Jun. 2009.
[44] W. R. Black, "Network autocorrelation in transport network and flow systems," *Geograph. Anal.*, vol. 24, no. 3, pp. 207–222, Sep. 2010.
[45] Y. Lu and J.-C. Thill, "Assessing the cluster correspondence between paired point locations," *Geograph. Anal.*, vol. 35, no. 4, pp. 290–309, Oct. 2003.
[46] A. T. Murray, Y. Liu, S. J. Rey, and L. Anselin, "Exploring movement object patterns," *Ann. Regional Sci.*, vol. 49, no. 2, pp. 471–484, Oct. 2012.
[47] H. Shu, T. Pei, C. Song, T. Ma, Y. Du, Z. Fan, and S. Guo, "Quantifying the spatial heterogeneity of points," *Int. J. Geograph. Inf. Sci.*, vol. 33, no. 7, pp. 1355–1376, Jul. 2019.
[48] T. Pei, A. Zhu, C. Zhou, B. Li, and C. Qin, "A new approach to the nearest-neighbour method to discover cluster features in overlaid spatial point processes," *Int. J. Geograph. Inf. Sci.*, vol. 20, no. 2, pp. 153–168, Feb. 2006.
[49] G. Celeux and G. Govaert, "A classification EM algorithm for clustering and two stochastic versions," *Comput. Statist. Data Anal.*, vol. 14, no. 3, pp. 315–332, Oct. 1992.
[50] S. Byers and A. E. Raftery, "Nearest–neighbor clutter removal for estimating features in spatial point processes," *J. Amer. Stat. Assoc.*, vol. 93, no. 442, pp. 577–584, Jun. 1998.

CI SONG received the B.S. and M.S. degrees from Wuhan University, in 2007 and 2009, respectively, and the Ph.D. degree from the Institute of Geographical Science and Natural Resource Research. He is currently an Assistant Professor with the Institute of Geographical Science and Natural Resource Research. His research interests include spatial data mining, spatial analysis, and geographic information science.

TAO PEI received the Ph.D. degree from the China University of Geosciences, in 1998. He is currently a Professor with the State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research. His research interests include spatial big data mining and geostatistics.

HUA SHU received the B.S. degree from Shaanxi Normal University, in 2013. He is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences. His research interests include spatiotemporal big data mining, mobile computing, and geographic information science.