BiFlowAMOEBA for the identification of arbitrarily shaped clusters in bivariate flow data

Qiliang Liu, Jie Yang, Min Deng, Wenkai Liu and Rui Xu

Department of Geo-informatics, Central South University, Changsha, Hunan, China; State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, Hubei, China

ABSTRACT
A bivariate flow cluster is a group of two types of spatial flows, where both types of flows have high (or low) values, or one type of flow has a high value while the other has a low value. Identifying bivariate flow clusters aids in understanding the complex interactions between different flow patterns. Detecting bivariate flow clusters remains challenging because statistics for quantitatively assessing bivariate flow clusters are lacking and the shapes and sizes of clusters vary. This study proposes a novel bivariate flow clustering method (BiFlowAMOEBA) by improving a multidirectional optimum ecotope-based algorithm (AMOEBA) which embeds local Getis-Ord statistic in an iterative procedure to detect irregular-shaped clusters. We define a bivariate local Getis-Ord statistic for quantitatively assessing bivariate flow clusters, use a hierarchical clustering strategy to construct clusters, and evaluate the statistical significance of clusters using a Monte Carlo simulation. Experimental results of simulated datasets show that BiFlowAMOEBA can identify bivariate flow clusters of different shapes more accurately and completely, compared with two state-of-the-art methods. Two case studies show that BiFlowAMOEBA helps not only unveil the interactions between public transport and taxi services but also identifies competition patterns between taxis and ride-hailing services.

ARTICLE HISTORY
Received 16 August 2021
Accepted 23 April 2022

KEYWORDS
Bivariate flow data; clustering; spatial association; spatial data mining

1. Introduction
Spatial flows are typically defined as the movement or travel between various locations, such as inter-county migration, daily commutes, and taxi trips (Andrienko et al. 2017, Vrotsou et al. 2017). Spatial flow data can be generally classified into two types: (i) individual flows that represent individual activities, such as taxi rides with origin-destination (OD) locations; (ii) aggregated flows that represent the movements of a group of people or objects, such as taxi OD flows between different traffic analysis regions (Guo and Zhu 2014). This study only focuses on aggregated flow data which...
have become increasingly available in the era of big data. A bivariate flow records the numbers (or values) of two types of flows between the same OD regions, e.g. taxi OD flows and public transit OD flows between the same residential area and workplace. The comparison of spatial distributions of two types of flows is an essential part of spatial association analysis and can provide new insights for understanding complex interactions between different transportation modes (Gao et al. 2018, Tao 2021). The anomalous spatial associations between two types of flows in local regions can be modeled as bivariate flow clusters (Tao and Thill 2020). A bivariate flow cluster is a group of two types of spatial flows, where both types of flows have high (or low) values or one type of flow has a high value while the other has a low value. Suppose there are two types of flows: type-I and type-II flows, four types of bivariate flow clusters can be specified: (i) $H_I - H_{II}$ clusters: both type-I and type-II flows have high values; (ii) $H_I - L_{II}$ clusters: type-I flows have high values and type-II flows have low values; (iii) $L_I - H_{II}$ clusters: type-I flows have low values and type-II flows have high values; (iv) $L_I - L_{II}$ clusters: both type-I and type-II flows have low values. Such clusters are useful for identifying interactions between different transportation modes (e.g. public transport and taxi services), which may aid in optimizing transportation planning (Zhang et al. 2018). These bivariate flow clusters are also useful for identifying competition patterns between different transportation services (e.g. taxi and ride-hailing services), which may be used by governments in formulating appropriate policies to ensure the favorable competition of different transportation services (Tao and Thill 2019b). We draw upon the following two examples to illustrate the practical value of identifying bivariate flow clusters.

(i) Unveiling interactions between public transport and taxi services. In any city, public transport and taxi services are the two primary modes of transportation. Clusters with a high density of taxi OD flows and a low density of public transit OD flows may indicate a lack of convenient public transportation between the origins and destinations of travel. Clusters with a low density of taxi OD flows and a high density of public transport OD flows, may indicate that public transport meets the travel OD travel demand. Such information may assist urban planners in optimizing public transport routes.

(ii) Unveiling competition patterns between taxi and ride-hailing services. Clusters with a high density of taxi OD and ride-hailing OD flow may indicate that traditional taxi businesses are facing serious competition from ride-hailing services (e.g. DiDi and Uber). Clusters with a high density of taxi OD flows and a low density of ride-hailing OD flows may represent the exclusive market territory of taxi services. Clusters with a low density of taxi OD flows and a high density of ride-hailing OD flows may represent the exclusive market territory of ride-hailing services. These bivariate flow clusters may be useful for the establishment of coordinated development of taxi and ride-hailing services.

Although some flow clustering methods have been developed in recent years, these methods are mainly designed for clustering univariate spatial flows (Liu et al. 2022). Local indicators of spatial association of bivariate flow data (e.g. local flow Cross K-function and BiFlowLISA) can be used to evaluate how the value of type-I flows is associated with the value of nearby type-II flows (Tao and Thill 2019b, 2020). However,
these local indicators were not designed for comparing the spatial distributions of two types of flows and identifying bivariate flow clusters of different shapes. To fill this gap, this study developed a novel bivariate flow clustering method, BiFlowAMOEBA, by improving the multidirectional optimum ecotope-based algorithm (AMOEBA) (Aldstadt and Getis 2006). We defined a bivariate local Getis-Ord statistic for bivariate flow data, developed a hierarchical clustering strategy to construct bivariate flow clusters, and finally evaluated the statistical significance of bivariate flow clusters using a Monte Carlo simulation. Experimental results on simulated datasets indicated that, compared with two state-of-the-art methods, BiFlowAMOEBA can identify bivariate flow clusters with increased accuracy. BiFlowAMOEBA was also applied to unveil the interactions between public transport and taxi services in Beijing, China, and to identify competition patterns between taxi and ride-hailing services in Xiamen, China. The bivariate flow clusters discovered by the BiFlowAMOEBA may provide new insights for designing and optimizing transportation systems.

2. Related work

Existing spatial clustering methods have been extended to cluster univariate spatial flows by modifying spatial proximity measures between the two flows. Three types of flow clustering methods are presently available: statistics-based methods that extend local indicators of spatial associations to the flow context [e.g. the local Getis-Ord statistics (Berglund and Karlström 1999, Tao and Thill 2019a), local Moran’s I (Liu et al. 2015), and Ripley’s K function (Tao and Thill 2016)]; hierarchical methods that hierarchically organize each flow based on an agglomerative or divisive strategy [e.g. graph partitioning methods (Adrienko and Adrienko 2011, Xiang and Wu 2019) and average linkage method (Zhu and Guo 2014, Yao et al. 2018)]; and density-based methods that identify high-density subsets of flows [e.g. density domain decomposition (Song et al. 2019a) and shared nearest-neighbor-based clustering methods (Liu et al. 2022)]. These methods were designed to handle univariate flow data, and are unsuitable to cluster bivariate flow data.

Recently, some local indicators of spatial associations have been developed to identify bivariate flow clusters, such as local Flow Cross K-function and local Flow Bivariate Moran’s I (Tao and Thill 2019b, 2020). These local indicators provide useful exploratory spatial data analysis tools for investigating the interactions between different spatial flow patterns (e.g. how the value of type-I flows associates with the value of nearby type-II flows). However, they are not designed for comparing the spatial distributions of a pair of spatial flow datasets (Tao and Thill 2020). In addition, developing bivariate flow clustering methods is necessary to further describe the shapes and extents of bivariate flow clusters. Spatial scan statistics have been modified to cluster flow data (Gao et al. 2018, Song et al. 2019b, Liu et al. 2020). Bivariate flows may be regarded as cases and controls, respectively. The spatial scan statistic was designed to identify clusters with a high ratio between the cases and controls. In a cluster discovered by spatial scan statistics, at least one type of flow exhibits clustering cannot be guaranteed. The density-based two-component clustering method (Pei et al. 2015) is effective for
identifying bivariate point clusters. However, this method cannot be used to detect bivariate flow clusters.

In recent years, increasing attention has been paid to analyze the relationships between different transportation modes, e.g. the interactions between public transport and taxi services and the competition patterns between taxis and ride-hailing services considered in the case studies of this work. For example, odds ratio analysis has been used to measure the relative balance of travel demand (the number of origins or destination points) of public transit and taxis in each sub-region (e.g. traffic analysis region), and the relative balance of travel demand of taxis and ride-hailing services (Zhang et al. 2018, Wang et al. 2021); Regression analysis methods (e.g. multinomial linear regression and geographically weighted regression) were used to estimate the effects of ride-hailing services on taxicab ridership (Contreras and Paz 2018, Tang et al. 2021). Most existing work treats the origin points and destination points of flows separately; therefore, internal spatial dependence between OD points was ruined and the analysis results cannot reveal the information about human mobility (Tao and Thill 2019b).

Based on the above analysis, one can find that the bivariate flow clusters considered in this study may provide new insights for analyzing the relationships between different transportation modes. However, identifying bivariate flow clusters of different shapes continues to be a challenge. On that account, this study aims to develop a novel bivariate flow clustering method by upgrading the AMOEBA algorithm.

3. The BiFlowAMOEBA method

In this study, AMOEBA was improved from two aspects: (i) we extended Getis-Ord local statistic $G_C_i$ to a bivariate flow context, and defined a novel bivariate local Getis-Ord statistic, $BG_C_i$; and (ii) we developed a hierarchical clustering strategy with $BG_C_i$ as the objective function to identify irregular-shaped and statistically significant bivariate flow clusters. As the theoretical basis of BiFlowAMOEBA, we first introduce the basic idea of AMOEBA.

3.1. The AMOEBA algorithm

AMOEBA was developed to identify spatial clusters and construct a spatial weight matrix for areal data (Aldstadt and Getis 2006). For a dataset with $N$ number of areas, AMOEBA considers each area and iteratively adds neighboring areas to that seed area, until the magnitude of the local Getis-Ord statistic $G_C_i$ (Ord and Getis 2010) is maximized. $G_C_i$ can be used to determine whether areas with high values or areas with low values tend to cluster in a local region. When a row-standardized spatial weights matrix (the spatial weights matrix is used to identify the neighbourhood for each area) is applied, $G_C_i$ for a given area $i$ can be simplified as

$$G_C_i = \sqrt{\frac{N-1}{N/n_i - 1}} \bar{z}_x$$  \hfill (1)

where $x_i$ is the attribute value at cell $i$; $n_i$ is the number of neighbors (including $i$ itself) of $i$; $\bar{z}_x$ is the average z-score value of neighboring areas of $i$. The detailed derivation of Equation (1) can be found in the Supplement Document.
AMOeba first selects a seed area \( i \) and calculates its \( G_i^* \) value (denoted as \( G_i^*(0) \)). Then, AMOeba calculates the \( G_i^* \) statistic for each region that includes \( i \) and all combinations of its neighboring areas. The combination that maximizes absolute \( G_i^* \) value is recorded. If this \( G_i^* \) value is larger than \( G_i^*(0) \), then this combination will be identified as a new high- \( (G_i^* > 0) \) or low-value \( (G_i^* < 0) \) cluster. The newly-built cluster will be regarded as a new seed to perform the region-growing procedure. The final cluster is constructed until it fails to increase the absolute \( G_i^* \) value by adding any set of neighboring areas. Each area will be selected as the initial seed to construct a cluster. AMOeba finally retains the non-overlapping clusters with the highest \( G_i^* \) values. A Monte Carlo simulation was used to evaluate the statistical significance of these clusters. AMOeba is more suitable for detecting arbitrarily shaped clusters. However, AMOeba is time-consuming because it requires an exhaustive procedure to optimize \( G_i^* \) to identify each cluster. To overcome this limitation, Duque et al. (2011) proposed an alternative formulation for AMOeba that avoids exhaustive procedures, without losing optimality. Widener et al. (2012) introduced a parallel computing method to improve the computational efficiency of AMOeba. Additionally, AMOeba was modified to identify clusters from univariate flow data. For example, a flowAMOeba method was developed by redefining the spatial relationship between the flows (Tao and Thill 2019a).

### 3.2. Bivariate local Getis-Ord local statistic for bivariate flows

To extend the Getis-Ord local statistic \( G_i^* \) to a bivariate flow context, it is vital to properly define the spatial weight matrix first. In this study, each bivariate flow was considered the basic unit for clustering. Compared to areal data, two unique characteristics of flow data should be well-handled: sparse OD matrix and the dyadic nature (Tao 2021).

Flows in a region are commonly stored by an OD matrix, where the rows represent the origins, the columns represent the destinations, and each element represents the number of a flow between two regions. In practice, there are usually a large number of null OD pairs in an OD matrix. For example, there are 1451 traffic analysis regions within the Sixth Ring Road in Beijing; therefore, the size of the OD matrix is 2,105,401. On 9 May 2016, only 91,258 OD pairs recorded taxi trips or smart card transactions. Treating these null OD pairs as zero-value flows will lead to a significant decrease in the mean and variance of a flow dataset. As in the above example, if we treat null OD pairs as zero-value flows, both the mean and variance of this dataset are close to zero. To avoid bias in calculation and reduce the memory requirements, we only kept the OD pairs that exist in the bivariate flow data and defined a bivariate flow model without using a sparse OD matrix. The bivariate flow can be represented as \( Bfi(ROi, RDi, A_i, A_ii, Ni) \), where \( ROi \) and \( RDi \) are origin and destination regions, respectively; \( A_i \) is the total number of type-I flows (e.g. taxi OD flows) from \( ROi \) to \( RDi \); \( A_ii \) is the total number of type-II flows (e.g. public transit OD flows) from \( ROi \) to \( RDi \), and \( Ni \) is the spatial neighborhood of \( Bfi \).

The dyadic nature of flow should be considered to define the spatial neighborhood for each bivariate flow. In this study, \( Ni \) of a bivariate flow \( Bfi \) is defined as a set of
bivariate flows, where the origin polygon of each flow shares at least one edge with \( R_i^O \) and the destination polygon of each flow shares at least one edge with \( R_i^D \). Figure 1(a) shows the spatial neighborhood of a bivariate flow whose origin and destination regions are different. There are some bivariate flows whose origin and destination regions are the same. For example, flows only occur in a traffic analysis region or flows that return to their origins. In this study, these two kinds of flows are not distinguished if they occurred in the same time period. Figure 1(b) shows the spatial neighborhood of a bivariate flow whose origin and destination regions are the same.

After spatial neighborhoods are constructed, the spatial weight matrix for bivariate flows can be calculated. For two bivariate flows \( Bfi \) and \( Bfj \), if \( Bfj \) is a neighbor of \( Bfi \), \( w_{ij} = 1 \); else, \( w_{ij} = 0 \):

\[
w_{ij} = \begin{cases} 1 & \text{if } Bfj \text{ is a neighbor of } Bfi \\ 0 & \text{otherwise} \end{cases}
\]

When a row-standardized spatial weights matrix is applied, according to Equation (1), \( G_i^I \) for the \( i \)th type-I and type-II flows can be represented as:

\[
G_i^I = \frac{1}{\sqrt{N_i^I - 1 \sum_{j=1}^{n_i} w_{ij}^2}} \frac{1}{\sqrt{N_i^I - 1 \sum_{j=1}^{n_i} w_{ij}^2}}
\]

\[
G_i^{II} = \frac{1}{\sqrt{N_i^{II} - 1 \sum_{j=1}^{n_i} w_{ij}^2}} \frac{1}{\sqrt{N_i^{II} - 1 \sum_{j=1}^{n_i} w_{ij}^2}}
\]

where \( N_i^I \) is the number of bivariate flows, \( n_i^I \) is the number of neighbors of flow \( i \). \( \tilde{z}_{AIi} \) and \( \tilde{z}_{AIIi} \) are the average z-score value of neighbors of the \( i \)th type-I and type-II flows, respectively.

To further define a bivariate \( G_i^I \) for bivariate flows, the bivariate spatial dependence between bivariate flows should be quantitatively modeled by considering both the univariate spatial associations of two types of flows and their mutual spatial relationship should be considered (Wartenberg 2010, Anselin 2019). In this study, a bivariate spatial association measure (\( L \) statistic) (Lee 2001) was employed to measure the bivariate spatial dependence.

\[
L = \sqrt{SSS_X} \cdot \sqrt{SSS_Y} \cdot r_{X,Y}
\]

\[
= \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \cdot \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}} \cdot \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}
\]

where \( SSS_X \) and \( SSS_Y \) are the spatial smoothing scalars of variable \( X \) and \( Y \), respectively. \( r_{X,Y} \) is the Pearson’s correlation coefficient between the spatial lags of two variables. \( x_i (\sum_j w_{ij} x_j) \) and \( y_i (\sum_j w_{ij} y_j) \) are the spatial lags of variable \( X \) and \( Y \) of area \( i \), respectively. \( \bar{x} \) and \( \bar{y} \) are the means of variable \( X \) and \( Y \), respectively. \( -\bar{x} \) and \( -\bar{y} \) are the means of the spatial lags of variables \( X \) and \( Y \), respectively. \( SSS_X \) and \( SSS_Y \) can be...
used to represent the univariate spatial association of each variable. If variable X (or Y) is spatially clustered, \( \text{SSS}_X \) (or \( \text{SSS}_Y \)) will be large because the spatial lags of X (or Y) will be similar to the original values of X (or Y) (Lee 2001). The Pearson’s correlation coefficient between spatial lags of two variables provides a mutual spatial relationship between two variables. Detailed mathematical justification can be found in the Supplement Document. A local bivariate spatial association \( L_i^* \) can be regarded as the relative contribution each area makes to \( L \). When a row-standardized spatial weights matrix is applied, \( L_i^* \) can be simplified as:

\[
L_i^* = \tilde{z}_{xi} \tilde{z}_{yi}
\]  

(3)

The detailed derivation of Equation (3) can be found in the Supplement Document. From Equation (3), we can find that the local bivariate spatial association between two variables can be measured by the product of the local average z-scores for these two variables (\( \tilde{z}_{xi} \) and \( \tilde{z}_{yi} \)). From Equation (1), we can find that \( G_i^* \) can be regarded as the product of local average z-scores (\( \tilde{z}_{Ai} \) and \( \tilde{z}_{AII} \)) for a variable and a scalar \( \left( \sqrt{\frac{N-1}{\nu_i}} \right) \) related to the number of local units (for two local regions with the same value of \( \tilde{z}_{Ai} \), the region with more units indicates a stronger local spatial dependence). Motivated by \( G_i^* \) and \( L_i^* \), a bivariate \( G_i^* \) should consist of two parts: (i) local bivariate spatial association between two variables; (ii) a scalar for considering the number of local units. To incorporate local bivariate spatial association in \( G_i^* \), the bivariate \( G_i^* \) can be defined as the product of \( L_i^* \) and \( G_i^* \) (Lee and Cho 2015).

When sparse OD matrix issue and the dyadic nature of bivariate flows were considered to construct spatial neighborhoods and spatial weight matrix, the bivariate \( G_i^* \) for bivariate flows can be defined as:

\[
BG_i^* = \sqrt{\frac{N_i^s - 1}{n_i^s - 1} \tilde{z}_{Ai}^s \tilde{z}_{AII}^s}
\]  

(4)

The higher the magnitude of \( BG_i^* \), the more anomalous the spatial associations between two types of flows in local regions. When \( \tilde{z}_{AII}^s \) and \( \tilde{z}_{Ai}^s \) both have large positive values or small negative values, \( \tilde{z}_{Ai}^s \tilde{z}_{AII}^s \) will be large and can reveal high positive bivariate spatial dependence between type-I and type-II flows. For \( \tilde{z}_{AII}^s \) and \( \tilde{z}_{Ai}^s \), if one has a large positive value and the other has a small negative value, \( \tilde{z}_{Ai}^s \tilde{z}_{AII}^s \) will be small and can reveal high negative bivariate spatial dependence between type-I flow and type-II flows. As the size of the local region increases \( \left( \sqrt{\frac{N_i^s - 1}{n_i^s - 1}} \right) \) increases), local bivariate spatial association usually becomes weak \( \left( \tilde{z}_{Ai}^s \tilde{z}_{AII}^s \right) \) decreases). Similar to \( G_i^* \), when an increase in \( \sqrt{\frac{N_i^s - 1}{n_i^s - 1}} \) cannot compensate for a decrease in \( \tilde{z}_{Ai}^s \tilde{z}_{AII}^s \), \( BG_i^* \) will decrease.

### 3.3. Hierarchical clustering strategy for identifying bivariate flow clusters

\( BG_i^* \) was used as an objective function for searching bivariate flow clusters. In each cluster, each two bivariate flow \( Bf_i \) and \( Bf_j \) should be spatially reachable: there is a path \( Bf_1, Bf_2, \ldots, Bf_n \) with \( Bf_1 = Bf_i \) and \( Bf_n = Bf_j \), where each flow \( Bf_k+1 \) is the
neighbor of $Bf_k$, $k = 1, 2, \ldots, n-1$. The exhaustive strategy used by AMOEBA (Aldstadt and Getis 2006) is time-consuming in identifying bivariate flow clusters; therefore, it is not applicable to sizeable flow datasets. The constructive AMOEBA (Duque et al. 2011) was developed based on a proposition of $G^*_i$: the first $n^*_i$ neighboring areas with the highest attribute values that can maximize $G^*_i$ for a local region. However, this proposition is not applicable to $BG^*_i$; thus, we design a hierarchical clustering strategy to identify bivariate flow clusters that avoids exhaustive searching.

At the beginning of hierarchical clustering, both $A^l_i$ and $A^l_j$ were standardized (the mean was zero and the variance was one). For each type of cluster, we select candidate flows for clustering according to the values of $\tilde{z}^*_A$ and $\tilde{z}^*_A$ for example, only the flows with positive values of $\tilde{z}^*_A$ and $\tilde{z}^*_A$ will be clustered to identify $H_l - H_{ll}$ clusters. If we want to search $L_l - L_{ll}$ clusters, only the flows with negative values of $\tilde{z}^*_A$ and $\tilde{z}^*_A$ should be selected as candidate flows. We initialize each candidate flow as a bivariate flow cluster, i.e., $BFC_s = \{BFC_i, 0 < i < n_c\}$, where $n_c$ is the number of candidate flows. The original $BG^*_i$ value for each initial cluster was calculated as $\tilde{z}^*_A \tilde{z}^*_A$.

To decide which two neighbouring clusters $BFC_i$ and $BFC_j$ ($\exists Bf_i \in BFC_i$ and $\exists Bf_j \in BFC_j$ are spatial neighbors) should be merged first, the objective function ($BG^*_i(BFC_i \cup BFC_j)$) and the gain of the objective function ($Gain(BFC_i \cup BFC_j)$) can be defined as:

$$BG^*_i(BFC_i \cup BFC_j) = \sqrt{\frac{N^* - 1}{n_{ki}^* - 1} \tilde{z}^*_A \tilde{z}^*_A}$$

$$Gain(BFC_i \cup BFC_j) = |BG^*_i(BFC_i \cup BFC_j)| - \max\{|BG^*_i|, |BG^*_j|\}$$

Where $n^*_{ki,j}$ is the number of bivariate flows in $BFC_i \cup BFC_j$, $\tilde{z}^*_A$ and $\tilde{z}^*_A$ are the average z-score values of type-I and type-II flows in $BFC_i \cup BFC_j$, respectively. We calculated $BG^*_i(BFC_i \cup BFC_j)$ and $Gain(BFC_i \cup BFC_j)$ for every two neighbouring clusters. For two clusters $BFC_i$ and $BFC_j$, if $|BG^*_i(BFC_i \cup BFC_j)|$ is the maximum and $Gain(BFC_i \cup BFC_j) > 0$, then these two clusters are combined to form a new cluster. Four types of bivariate flow clusters can be identified using $|BG^*_i(BFC_i \cup BFC_j)|$ and $Gain(BFC_i \cup BFC_j)$.

We iteratively merged the two neighbouring clusters until the gain of the objective function is less than zero by combining any two clusters. Unlike AMOEBA, our method is not a region growing method; therefore, the selection of seed flows is not required and the discovered clusters are not overlapped.

The statistical significance of all the discovered clusters was evaluated using a Monte Carlo simulation. A conditional permutation approach was used to generate random bivariate flow datasets. In this study, the null OD pairs were not regarded as zero-value flows. Therefore, we do not permutate the observed $A^l_i$ and $A^l_j$ values at null OD pairs. The endpoint locations and direction of each non-null bivariate flow were unchanged. Each random bivariate flow dataset was obtained by randomly permutating the observed $A^l_i$ and $A^l_j$ values at observed $RO - RO$ pairs. The conditional permutation approach does not generate new $RO - RO$ pairs and new $A^l_i$ and $A^l_j$ values. Therefore, the statistical distribution of the observed bivariate flows can be retained in random bivariate flows. The $p$-value of each bivariate flow cluster $BFC_i$ is
calculated as:

\[
p - \text{value}(BFC_i) = \frac{\sum_{j=1}^{R} I_j \left( BG_i^\text{ran}(BFC_i) \right)}{1 + R}
\]

where \( BG_i^\text{ran}(BFC_i) \) and \( BG_i^\text{obs}(BFC_i) \) are the \( BG_i \) values for \( BFC_i \), calculated based on random and observed flow data, respectively. \( I(\cdot) \) is an indicator function, if \( BFC_i^\text{ran} \) and \( BFC_i^\text{obs} \) belong to the same type of bivariate flow cluster and \( |BG_i^\text{ran}(BFC_i)| \geq |BG_i^\text{obs}(BFC_i)| \), then \( I = 1 \); else, \( I = 0 \). If \( p - \text{value}(BFC_i) \leq \alpha \) (\( \alpha \) is the significance level), then \( BFC_i \) is a true cluster.

### 3.4. Implementing BiFlowAMOEBA

BiFlowAMOEBA requires three parameters: target cluster type \((T_c)\), significance level \((\alpha)\), and number of Monte Carlo simulations \((R)\). The pseudo-code of the BiFlowAMOEBA is performed as follows.

**Algorithm: BiFlowAMOEBA** (Input: Bivariate flows, \( T_c \), \( \alpha \), \( R \))

1. **StandardizedFlows = StdAttributeValue(Bivariate flows)** //Standardize \( A^I \) and \( A^II \) values for each bivariate flow
2. **Candidateflows = SelectCandidateFlows (StandardizedFlows, \( T_c \))** //Select candidate flows
3. **BFCs = InitializeBFCs(Candidateflows)** //Initialize bivariate flow clusters
4. **Max_Gain = SearchLargestIncrementValue(BFCs)** //Search the largest gain value
5. **while Max_Gain > 0**
6. **BFCs = MergeCluster(BFCs)** //Merge the two clusters with the largest gain value
7. **Max_Gain = SearchLargestIncrementValue(BFCs)** //Search the largest gain value
8. **end while**
9. **for each BFC in BFCs do**
10. **BFC.pvalue = MonteCarloSimulation (StandardizedFlows, BFC, R)** //Calculate p-value for each cluster
11. **if BFC.pvalue \leq \alpha**
12. **insert BFC into TrueClusters**
13. **end for**
14. **return TrueClusters**

The time and space complexities of BiFlowAMOEBA mainly depend on two elements: the identification of bivariate flow clusters using the hierarchical clustering strategy, the time complexity does not exceed \( O(n_c^2) \) and the space complexity is approximately \( O(n_c) \); the identification of significant bivariate flow clusters using a Monte Carlo simulation, where the time complexity does not exceed \( O(Rn_cN) \), and the space complexity is approximately \( O(N) \). The total time complexity of BiFlowAMOEBA is approximately \( O(Rn_cN) \), and the total space complexity of BiFlowAMOEBA is approximately \( O(N) \).

### 4. Experiments

BiFlowAMOEBA was compared with two state-of-the-art methods [AntScan_flow (Song et al. 2019b) and BiFlowLISA (Tao and Thill 2020)]. For the three methods, the
significance level, $\alpha$ was set to 0.01, and the number of Monte Carlo simulations $R$ was set to 999. For AntScan_flow, the iteration times were set to 100, the ant number was set to 200, the elite ant number was set to 30, and the evaporation coefficient was set to 0.99. All three methods were implemented in Python 3.8.1, on a workstation with a 2.20 GHz computer processing unit and 32 GB memory running the Microsoft Windows 10 operating system. In the two case studies, we mainly aim to understand how the two transportation modes complement or compete with each other. For the two transportation modes, if the travel demand between different regions is low, it is hard to determine whether these two transportation modes are complementary or competitive. Therefore, $L_i - L_{ii}$ clusters are not considered in the two case studies.

4.1. Simulation experiment

The simulated data sets were generated using a procedure similar to the methods developed by Tao and Thill (2019a) and Song et al. (2019b). We designed two groups of simulated data in a predefined region with 250 polygons (Figure 2). Four types of bivariate flow clusters (i.e. $H_i - H_{ii}$, $H_i - L_{ii}$, $L_i - H_{ii}$, and $L_i - L_{ii}$ clusters) were predefined in the simulated datasets. Group I contains four data sets and each data set contains a single bivariate flow cluster. Group II contains one data set including four bivariate flow clusters of different types and shapes. For each data set, we first generated $R^O - R^D$ pairs for each cluster, and then assigned two anomalous values (high or low) for each $R^O - R^D$ pairs as $A^i$ and $A^{ii}$ values, to form a bivariate flow cluster. Finally, to generate noise flows, we randomly selected 1000 $R^O - R^D$ pairs and assigned two medium values as $A^i$ and $A^{ii}$ values. $R^O - R^D$ pairs of clusters and noise flows are not duplicated in each data set (the origin or destination regions of noise flows can locate inside clusters). We assume that the medium value follows a normal distribution.

![Figure 2](image-url)

Figure 2. Two groups of simulated data. Group I: (a–d); Group II: (e).
distribution $N(\mu, \sigma)$. At a significance level of 0.01, $(\mu - 3\sigma)$ and $(\mu + 3\sigma)$ can be used as the thresholds of the low value and high values. We also assume that high and low values follow normal distributions $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$ ($\mu_1 > (\mu + 3\sigma)$ and $\mu_2 < (\mu - 3\sigma)$), respectively. Therefore, $A^l$ values were randomly extracted from normal distributions as $N(150, 5)$ (high-value), $N(100, 5)$ (medium-value), and $N(50, 5)$ (low-value); $A^u$ values were randomly extracted from normal distributions as $N(150, 5)$ (high-value), $N(120, 5)$ (medium-value), and $N(90, 5)$ (low-value).

The performance of the three methods is shown in Table 1. Precision and recall (Tan et al. 2006) were used to evaluate the performance of the three methods:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$

Where $TP$, $FP$, and $FN$ are the values of true positive, false positive (also called type-I error), and false negative (also called type-II error), respectively. High values of precision and recall correspond, respectively to low type I and type II error rates.

From Table 1, we can find that BiFlowAMOEBA can detect predefined clusters completely and accurately: the values of precision and recall usually exceed 0.9. Although BiFlowLISA usually detects predefined clusters accurately, some clusters cannot be completely identified by BiFlowLISA. The clusters identified by AntScan_flow are more complete than that identified by BiFlowLISA. However, AntScan_flow wrongly combined some noise flows into clusters. We can conclude that BiFlowAMOEBA performs better than the two state-of-the-art methods on the four simulated datasets.

### 4.2. Case study of Beijing smart card transaction and taxi trajectory data

BiFlowAMOEBA was first applied to identify interactions between public transit and taxi services in Beijing. The study area is located within the Sixth Ring Road of Beijing (Figure 3). Eleven places of interest were highlighted in the study area, to analyze the detected clusters. Table 2 provides a detailed description of these places. The taxi GPS trajectories were generated by nearly 30,000 taxis on 9 May 2016. Each trajectory records the taxi ID, sampling time, location, and status (occupied or not). Smart card transactions were generated by nearly 25,000 buses on 9 May 2016. Each transaction records the boarding time, alighting time, boarding station, and alighting station of a passenger. The OD flows of smart card transactions and taxi trajectories were matched.

| Dataset | Cluster type | BiFlowAMOEBA | BiFlowLISA | AntScan_flow |
|---------|--------------|--------------|------------|--------------|
|         | Cluster type | Precision    | Recall     | Precision    | Recall     | Precision | Recall     |
| Group I | $H_I - H_U$  | 0.96         | 0.96       | 0.92         | 0.44       | 0.52      | 0.46       |
|         | $H_I - L_U$  | 0.96         | 0.93       | 0.89         | 0.51       | 0.59      | 0.52       |
|         | $L_I - H_U$  | 1.00         | 0.91       | 1.00         | 0.52       | 0.69      | 0.55       |
|         | $L_I - L_U$  | 0.97         | 1.00       | 0.72         | 0.35       | 0.52      | 0.61       |
| Group II| $H_I - H_U$  | 0.92         | 0.92       | 0.80         | 0.67       | 0.63      | 0.83       |
|         | $H_I - L_U$  | 0.91         | 0.92       | 0.76         | 0.67       | 0.67      | 0.83       |
|         | $L_I - H_U$  | 0.90         | 1.00       | 0.70         | 0.78       | 0.61      | 0.81       |
|         | $L_I - L_U$  | 1.00         | 0.90       | 0.42         | 0.45       | 0.28      | 0.64       |

| Group I | $H_I - H_U$  | 0.96         | 0.96       | 0.92         | 0.44       | 0.52      | 0.46       |
|         | $H_I - L_U$  | 0.96         | 0.93       | 0.89         | 0.51       | 0.59      | 0.52       |
|         | $L_I - H_U$  | 1.00         | 0.91       | 1.00         | 0.52       | 0.69      | 0.55       |
|         | $L_I - L_U$  | 0.97         | 1.00       | 0.72         | 0.35       | 0.52      | 0.61       |
| Group II| $H_I - H_U$  | 0.92         | 0.92       | 0.80         | 0.67       | 0.63      | 0.83       |
|         | $H_I - L_U$  | 0.91         | 0.92       | 0.76         | 0.67       | 0.67      | 0.83       |
|         | $L_I - H_U$  | 0.90         | 1.00       | 0.70         | 0.78       | 0.61      | 0.81       |
|         | $L_I - L_U$  | 1.00         | 0.90       | 0.42         | 0.45       | 0.28      | 0.64       |

The OD flows of smart card transactions and taxi trajectories were matched.
to 1,451 traffic analysis regions using the ray intersection method (Taylor 1994). In this case study, we regarded taxi and public transit OD flows as types-I and -II flows, respectively. Bivariate flow clusters were detected in two periods: morning rush hours (7:00–9:00) and afternoon rush hours (16:00–19:00).

The bivariate flow clusters identified by BiFlowAMOEBA, BiFlowLISA, and AntScan_flow during morning rush hours are displayed in Figures 4–6 (red, green, and yellow polygons represent the origin, destination, and overlapping origin and destination locations, respectively). Due to space limitations, the bivariate flow clusters identified during afternoon rush hours are shown in the Supplement Document. To present clear and readable figures for each type of cluster, we only present a maximum of ten clusters with the highest absolute values of the statistic $BG_i^2$ for BiFlowAMOEBA, a maximum of ten clusters with the largest number of flows for BiFlowLISA and a maximum of ten clusters with the highest values of log-likelihood ratio for AntScan_flow.

4.2.1. $H_I$–$H_{II}$ clusters

The $H_I$–$H_{II}$ clusters identified by BiFlowAMOEBA (Figure 4(a)) demonstrate that taxi and public transport are popular modes of travel around or between some regions during morning rush hours. Travel distances were usually short in these clusters. These
Clusters I, II, III, and IX were around commercial areas, e.g., Wangjing, Sanlitun, and Xinjiekou. Cluster IV, V, and VI were around technology hubs, e.g., Fengtai Science and Technology Park, Zhongguancun and Zhongguancun Software Park. Additionally, traffic was high in some residential areas and transportation hubs. For example, Cluster VII and VIII were around residential areas near Fourth Ring Road and Fifth Ring Road, Cluster X was around Sihui Coach Station.

Table 2. Descriptions of interested places in Figure 3.

| No. | Name of subarea                          | Description                                                                                                                                 |
|-----|------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| 1   | Beijing Capital International Airport   | One of the two main international airports serving Beijing                                                                               |
| 2   | Zhongguancun Software Park               | Home to a large number of tech-related companies as well as a spacious multi-purpose events space suitable for large outdoor events |
| 3   | Wangjing                                 | A major residential, technological and business area in the northeast of Beijing                                                          |
| 4   | Asian Games Village                      | The site of the 1990 Asian Games and a major residential area and a subdistrict of the Chaoyang District of Beijing                          |
| 5   | Chinese Academy of Sciences              | The national academy for the natural sciences of the People’s Republic of China and it functions as the national scientific think tank and academic governing body |
| 6   | Sanlitun                                 | An important commercial area containing many popular bar streets and international stores; a popular destination for shopping, dining and entertainment |
| 7   | Xinjiekou                                | A famous business district in Beijing. It has a long history and numerous cultural relics and scenic spots                                   |
| 8   | Tongzhou West Railway Station            | Tongzhou West Railway Station is a third-class station on the Beijing-Chengde Railway                                                     |
| 9   | Sihui Coach Station                      | A modern and intelligent long-distance highway passenger terminal                                                                        |
| 10  | Beijing Central Business District        | The primary area for finance, media, and business services in Beijing                                                                       |
| 11  | Fengtai Science and Technology Park      | A technology hub located in the southwest of Beijing                                                                                       |

Figure 4. $H_I - H_{II}$ clusters identified during morning rush hours: (a) clusters identified by BiFlowAMOEBA; (b) clusters identified by BiFlowLISA.
The $H_1 - H_{II}$ clusters identified by BiFlowLISA are shown in Figure 4(b). AntScan_flow does not find $H_1 - H_{II}$ clusters. According to the definition of $H_1 - H_{II}$ cluster, a high-quality cluster should have as high $A^I$ value as possible and as high $A^{II}$ value as possible. The means of $A^I$ and $A^{II}$ for clusters identified by BiFlowAMOEBA are 6.71 and 315.56. The means of $A^I$ and $A^{II}$ for clusters identified by BiFlowLISA are 1.21 and 50.84. The quality of the clustering results obtained by BiFlowAMOEBA is higher than that obtained by BiFlowLISA. By comparing the clustering results of BiFlowAMOEBA and BiFlowLISA, we found that BiFlowLISA identify some clusters that were not identified by BiFlowAMOEBA, e.g. Cluster II, Cluster VI, and Cluster IX in Figure 5(b). The means of $A^I$ and $A^{II}$ for Cluster II, VI, and IX are 1.05 and 55.86, 1.06 and 57.13, and 1.11 and 43.44, respectively. For BiFlowAMOEBA, the means of $A^I$ and $A^{II}$ for the cluster with the lowest absolute value of $BG^I_i$ are 4.5 and 336. We can find that the $H_1 - H$ clusters identified by BiFlowAMOEBA are more anomalous than that identified by BiFlowLISA.

4.2.2. $H_1 - L_{II}$ clusters

In Figure 5(a), $H_1 - L_{II}$ clusters identified by BiFlowAMOEBA, including I–III, V, VI, and VIII, were mainly discovered far from downtown. This indicates that people residing in these areas prefer to choose taxi for commuting during morning rush hours. Clusters IV and VII represent the imbalanced traffic flows between Beijing Capital International Airport and the regions within the Fifth Ring Road. We can infer that people in commercial areas and science/education areas prefer to take a taxi to the airport during morning rush hours.

The $H_1 - L_{II}$ clusters identified by BiFlowLISA and AntScan_flow are shown in Figures 5(b,c). According to the definition of $H_1 - L_{II}$ cluster, a high-quality cluster should have as high $A^I$ value as possible and as low $A^{II}$ value as possible. The means of $A^I$ and $A^{II}$ for clusters identified by BiFlowAMOEBA, BiFlowLISA, and AntScan_flow are 7.24 and 0.32, 1.69 and 3.84, and 1.37 and 7.63, respectively. The quality of the clustering results obtained by BiFlowAMOEBA is higher than that obtained by the other two methods. BiFlowLISA and AntScan_flow found some clusters which were not identified by BiFlowAMOEBA, e.g. Cluster I, Cluster V, Cluster IX and Cluster X in
Figure 5(b) and Cluster VI, Cluster VIII and Cluster IX in Figure 5(c). In Figure 5(b), the means of $A^I$ and $A^II$ for Cluster I, V, IX, and X are 1.42 and 4, 1.35 and 5.07, 1.08 and 0.75, and 1.57 and 3.21, respectively. In Figure 5(c), the means of $A^I$ and $A^II$ for Cluster VI, VIII, and IX are 1.45 and 8.66, 1 and 0, and 1.23 and 20.85. For BiFlowAMOEBA, the means of $A^I$ and $A^II$ for the cluster with the lowest absolute value of $B_{G_i}/C_{3i}$ are 5 and 0.45. We can find that the $H_I - L_{II}$ clusters identified by BiFlowAMOEBA are usually more anomalous than that identified by the other two methods.

4.2.3. $L_I - H_{II}$ clusters

In Figure 6(a), $L_I - H_{II}$ clusters identified by BiFlowAMOEBA mainly represent traffic flows from different residential areas around the Sixth Ring Road to technology hubs.
and commercial areas. The flows in these clusters may indicate urban commuting behavior during the morning rush hours. Clusters I, II, III, and V were from residential areas to Zhongguancun Software Park, Cluster IV was from a residential area to Fengtai Science and Technology Park, and Cluster VI, VIII, and IX were from residential areas to the Chinese Academy of Sciences. The other three $L_I - H_{II}$ clusters were from residential areas to commercial areas (e.g. Xinjiekou and Wangjing). We can infer that people in these residential areas prefer to choose public transit for long-distance commuting during morning rush hours from the discovered clusters.

The $L_I - H_{II}$ clusters identified by BiFlowLISA and AntScan_flow are shown in Figures 6(b,c). According to the definition of $L_I - H_{II}$ cluster, a high-quality cluster should have as low $A^I$ value as possible and as high $A^{II}$ value as possible. The means of $A^I$ and $A^{II}$ for clusters identified by BiFlowAMOEBA, BiFlowLISA, and AntScan_flow are 0 and 889.72, 0 and 43.88, and 0.05 and 91.52, respectively. We can infer that the quality of the clustering results obtained by BiFlowAMOEBA is higher than that obtained by the other two methods.

4.3. Case study of Xiamen taxi and ride-hailing data

BiFlowAMOEBA was applied to identify competition patterns between taxi and ride-hailing services in Xiamen, Fujian Province, China. Figure 7 shows the study area. Seven important sites were highlighted to analyze the detected clusters. Table 3 provides a detailed description of these places. Taxi and ride-hailing orders were collected

Figure 7. The study area.
on 31 May 2019. Each order records the vehicle ID, boarding time, boarding location, alighting time, and alighting location. We chose the nearest road segment as the matched road segment for each OD point (White et al. 2000). OD flows of taxi and ride-hailing were matched to 2,355 road segments. We regarded taxi OD flows as type-I flows and ride-hailing OD flows as type-II flows. Bivariate flow clusters were detected in two periods: morning rush hours (7:00–9:00) and afternoon rush hours (16:00–19:00).

The bivariate flow clusters identified by BiFlowAMOEBA, BiFlowLISA, and AntScan_flow during morning rush hours are displayed in Figures 8–10 (red, green, and yellow polygons represent the origin, destination, and overlapping origin and destination locations, respectively). Due to space limitations, the bivariate flow clusters identified during afternoon rush hours are shown in the Supplement Document. To present clear and readable figures for each type of cluster, we only presented a maximum of ten clusters with the highest absolute values of the statistic $BG_i$ for BiFlowAMOEBA, a maximum of ten clusters with the largest number of flows for BiFlowLISA and a maximum of ten clusters with the highest values of log-likelihood ratio for AntScan_flow.

4.3.1. $H_1 - H_2$ clusters
The $H_1 - H_2$ clusters discovered by BiFlowAMOEBA (Figure 8(a)) indicate that taxi and ride-hailing are popular modes of travel around or between certain regions during morning rush hours. Cluster I–III, V–VIII represent the morning commuting behavior between residential areas and Guanyin Mountain (a famous commercial area). Additionally, increased traffic revealed that the demand for taxi and ride-hailing were high around some residential areas and transportation hubs. Cluster IV and Cluster IX were found around Xiamen Railway Station. Cluster X was around a residential area near Huandao South Road.

The $H_1 - H_2$ clusters identified by BiFlowLISA are shown in Figure 8(b). AntScan_flow does not find $H_1 - H_2$ clusters. The means of $A^1$ and $A^2$ for clusters

| No. | Name of subarea                | Description                                                                 |
|-----|--------------------------------|-----------------------------------------------------------------------------|
| 1   | Xiamen North Railway station  | A large-scale transportation hub integrating the functions of railway, urban road, long-distance bus, rail transit transfer and commercial service |
| 2   | Jimei University              | A key public university in Fujian Province                                 |
| 3   | Xiamen Horticultural Expo Garden | The world's largest aquatic horticulture expo garden with the most complete garden genres, which was built for the 6th China International Garden and Flowers Exposition |
| 4   | Xiamen Gaoqi International Airport | An important regional aviation hub in the southeast coast of China, it is one of the 12 major trunk airports in China |
| 5   | Guanyin Mountain             | An important commercial area containing important international business centers and international conference center |
| 6   | Xiamen Railway Station       | A railway station that combines both ‘traditional’ passenger trains and high-speed (D-series) trains |
| 7   | Xiamen University            | The university is perennially ranked as one of the top academic institutions in Southern China, with strengths in economics and management, fine arts, law, chemistry, journalism, communication, and mathematics. |
identified by BiFlowAMOEBA are 8.30 and 39.30. The means of $A^I$ and $A^II$ for clusters identified by BiFlowLISA are 1.86 and 4.35. Therefore, the quality of the clustering results obtained by BiFlowAMOEBA is higher than that obtained by BiFlowLISA. For Cluster II, IV, and VIII in Figure 8(b) that were not discovered by BiFlowAMOEBA, the maximum means of $A^I$ and $A^II$ are 2.25 and 3.5. For BiFlowAMOEBA, the means of $A^I$ and $A^II$ for the cluster with the lowest absolute value of $BG^I_J$ are 7.5 and 12. It indicates that the $H_I - H_{II}$ clusters identified by BiFlowAMOEBA are more anomalous than that identified by BiFlowLISA.

Figure 8. $H_I - H_{II}$ clusters identified during morning rush hours: (a) clusters identified by BiFlowAMOEBA; (b) clusters identified by BiFlowLISA.
4.3.2. HI\textsubscript{2}LII\textsubscript{1} clusters

In Figure 9(a), HI\textsubscript{2}/C0\textsubscript{1}LII\textsubscript{1} clusters identified by BiFlowAMOEBA mainly represent traffic flows between or around different transportation hubs. Clusters I, II, and X described travel from Xiamen Gaoqi International Airport to Xiamen North Railway station or vice versa, Cluster IV was located from Xiamen Gaoqi International Airport to Xiamen Railway station, Cluster V and VI were located around Xiamen Gaoqi International Airport. These clusters may indicate that taxi is a more popular mode of travel than ride-hailing between hubs or around this transportation during morning rush hours. Other clusters were discovered between residential areas and commercial areas or transportation hubs. Cluster III and VIII showed travel flows from residential areas to Xiamen North Railway station or vice versa. Cluster VII and Cluster IX were located in residential to commercial areas.

The HI\textsubscript{2}/C0\textsubscript{1}LII\textsubscript{1} clusters identified by BiFlowLISA and AntScan_flow are shown in Figures 9(b,c). The means of A\textsubscript{I} and A\textsubscript{II} for clusters identified by BiFlowAMOEBA, BiFlowLISA, and AntScan_flow are 9.6 and 0, 1.70 and 0.12, and 1.62 and 0.34, respectively. It indicates that the quality of the clustering results obtained by BiFlowAMOEBA is higher than that obtained by the other two methods. For the clusters in Figures 9(b,c) which were not identified by BiFlowAMOEBA, the maximum mean of A\textsubscript{I} is 2.57 and the minimum mean of A\textsubscript{II} is 0. For BiFlowAMOEBA, the means of A\textsubscript{I} and A\textsubscript{II} for the cluster with the lowest absolute value of BG\textsubscript{1} are 7 and 0. We can infer that

Figure 9. HI\textsubscript{2}–LII clusters identified during morning rush hours: (a) clusters identified by BiFlowAMOEBA; (b) clusters identified by BiFlowLISA; (c) clusters identified by AntScan_flow.

4.3.2. HI\textsubscript{2}–LII\textsubscript{1} clusters

In Figure 9(a), HI\textsubscript{2}–LII\textsubscript{1} clusters identified by BiFlowAMOEBA mainly represent traffic flows between or around different transportation hubs. Clusters I, II, and X described travel from Xiamen Gaoqi International Airport to Xiamen North Railway station or vice versa, Cluster IV was located from Xiamen Gaoqi International Airport to Xiamen Railway station, Cluster V and VI were located around Xiamen Gaoqi International Airport. These clusters may indicate that taxi is a more popular mode of travel than ride-hailing between hubs or around this transportation during morning rush hours. Other clusters were discovered between residential areas and commercial areas or transportation hubs. Cluster III and VIII showed travel flows from residential areas to Xiamen North Railway station or vice versa. Cluster VII and Cluster IX were located in residential to commercial areas.

The HI\textsubscript{2}–LII\textsubscript{1} clusters identified by BiFlowLISA and AntScan_flow are shown in Figures 9(b,c). The means of A\textsubscript{I} and A\textsubscript{II} for clusters identified by BiFlowAMOEBA, BiFlowLISA, and AntScan_flow are 9.6 and 0, 1.70 and 0.12, and 1.62 and 0.34, respectively. It indicates that the quality of the clustering results obtained by BiFlowAMOEBA is higher than that obtained by the other two methods. For the clusters in Figures 9(b,c) which were not identified by BiFlowAMOEBA, the maximum mean of A\textsubscript{I} is 2.57 and the minimum mean of A\textsubscript{II} is 0. For BiFlowAMOEBA, the means of A\textsubscript{I} and A\textsubscript{II} for the cluster with the lowest absolute value of BG\textsubscript{1} are 7 and 0. We can infer that
BiFlowAMOEBA identified more anomalous $H_I - L_{II}$ clusters than that identified by the other two methods.

### 4.3.3. $L_I - H_{II}$ clusters

In Figure 10(a), $L_I - H_{II}$ clusters identified by BiFlowAMOEBA mainly represent the urban commuting behavior between commercial and residential areas during morning rush hours. Clusters I–IV and VII–IX were located from residential areas to Guanyin Mountain. Clusters V and X were found from residential areas to commercial areas around Xiamen Horticultural Expo Garden. Cluster VI was found in the Siming district. These clusters may indicate that people residing in these residential areas prefer ride-hailing for commuting during morning rush hours.
The $L_{i} - H_{II}$ clusters identified by BiFlowLISA and AntScan_flow are shown in Figures 10(b,c). The means of $A_{I}$ and $A_{II}$ for clusters identified by BiFlowAMOEBA, BiFlowLISA, and AntScan_flow are 0 and 13.6, 0 and 2.59, and 0.16 and 3.59, respectively. There are some clusters in Figure 10(b) (Cluster VI and IX) and Figure 10(c) (Cluster III, IV, VI, VIII, and X) which were not identified by BiFlowAMOEBA. The minimum mean of $A_{I}$ is 0 and the maximum mean of $A_{II}$ is 4.34 for these clusters. For BiFlowAMOEBA, the means of $A_{I}$ and $A_{II}$ for the cluster with the lowest absolute value of $BG_{i}^*$ are 0 and 6.36. We can infer that BiFlowAMOEBA identified more anomalous $H_{I} - L_{II}$ clusters than that identified by BiFlowLISA and AntScan_flow.

5. Discussion

The experimental results on simulated datasets show that BiFlowAMOEBA performs better than BiFlowLISA and AntScan_flow. For the two real-world datasets, the bivariate flow clusters identified by the three methods are usually different (the clusters identified by BiFlowAMOEBA are more anomalous than that identified by the other two methods). The primary reasons for this are as follows.

(i) BiFlowLISA is useful for evaluating how the value of type-I flows is associated with the value of nearby type-II flows. BiFlowLISA only evaluates the similarity between the $A_{I}$ (or $A_{II}$) value of a target flow and $A_{II}$ (or $A_{I}$) values of its neighboring flows, and does not evaluate the similarity between the $A_{I}$ (or $A_{II}$) value of a target flow and $A_{I}$ (or $A_{II}$) values of its neighboring flows. When BiFlowLISA is applied to identify the bivariate flow clusters which are interested in this study, some spurious clusters may be identified because the $A_{I}$ (or $A_{II}$) value of a target flow and $A_{II}$ (or $A_{I}$) values of its neighboring flows may be different.

(ii) AntScan_flow is a powerful tool for identifying the local regions where the spatial distributions of two kinds of flows differ the most. Although AntScan_flow can identify regions with anomalous ratios between the values of two types of flows (cases and controls), these regions may not indicate true bivariate flow clusters considered in this study and vice versa. AntScan_flow was not designed to identify the $H_{I} - H_{II}$ clusters considered in this study. In addition, when the number of controls is very small, the ratio between cases and controls also will be very high. However, the cases may not exhibit clustering.

(iii) For BiFlowAMOEBA, $BG_{i}^*$ can consider both the univariate spatial associations of two types of flows and their mutual spatial relationship. Moreover, the hierarchical clustering strategy with $BG_{i}^*$ as the objective function can identify irregular-shaped clusters. Therefore, BiFlowAMOEBA is more suitable for detecting the bivariate flow clusters considered in this study. BiFlowAMOEBA is an important complement to existing bivariate flow analysis methods.

The interactions between public transport and taxi services in Beijing and the competition patterns between taxi and ride-hailing services in Xiamen may have potential application value in designing and optimizing public transportation systems.

5.1. Interactions between public transit and taxi services in Beijing

$H_{I} - H_{II}$ clusters (Figure 4(a)) primarily represent short-distance travel flows around commercial areas, technology hubs, and residential areas. These clusters are useful for
identifying regions with high taxi and public transit usage loads. This indicates that taxi services account for public transportation services around these regions during rush hours. Therefore, moderately increasing the frequency of public transportation may help alleviate traffic congestion around these regions during rush hours. \(H_1 - L_{\|}\) clusters (Figure 5(a)) primarily represent travel flows between Beijing Capital International Airport and different regions within the Fifth Ring Road, and travel flows between the Sixth Ring Road and the Fifth Ring Road. These clusters can be used to identify traffic flows with high taxi and low public transit usage loads. This may be attributable to the lack of convenient public transportation between these regions. Therefore, public transportation planning between these regions may need to be improved. \(L_1 - H_{\|}\) clusters (Figure 6(a)) primarily represent the urban commuting behavior between residential areas and commercial areas, and between residential areas and technique hubs. These clusters indicate that public transit services between these regions are convenient. Therefore, public transportation between these regions may be retained in further transportation system planning.

\[5.2. \text{ Competition patterns between taxi and ride-hailing services in Xiamen} \]

\(H_1 - H_{\|}\) clusters (Figure 8(a)) primarily represent travel flows between commercial areas and residential areas, and commercial areas and transportation hubs in Huli and Siming Districts, respectively. These clusters can represent the mutual market territory of taxi and ride-hailing services. This may indicate that taxi and ride-hailing services in these regions compete with each other. \(H_1 - L_{\|}\) clusters (Figure 9(a)) primarily represent travel flows between different transportation hubs or between commercial areas and transportation hubs. These clusters represent the exclusive market of taxi services. This may indicate that taxi services mainly dominate the market between transportation hubs, and between transportation hubs and commercial areas. \(L_1 - H_{\|}\) clusters (Figure 10(a)) primarily represent urban commuting behavior between commercial and residential areas that nearly cover the entire city of Xiamen. These clusters represent the exclusive market for ride-hailing services. This may indicate that ride-hailing service is a more popular commuting tool because of various preferential measures. These competition patterns may provide a reference for the government in formulating appropriate policies to ensure the favorable competition between these two transportation services in a city.

\[6. \text{ Conclusion} \]

In this study, we upgraded AMOEBA from area data to bivariate flow data and developed a bivariate flow clustering method, known as BiFlowAMOEBA. BiFlowAMOEBA can effectively identify irregular-shaped and statistically significant bivariate flow clusters. We extended the local Getis-Ord statistic \(G'_{ij}\) to a bivariate flow context by considering two unique characteristics of flow data (i.e. sparse OD matrix and the dyadic nature) and bivariate spatial dependence between bivariate flows. The newly built bivariate \(BG'_{ij}\) can assess bivariate flow clusters quantitatively. To avoid the time-consuming enumeration of possible clustering combinations, we proposed a hierarchical
clustering strategy with $BG^*$ as the objective function for constructing clusters. Bivariate flow clusters with irregular shapes can be effectively identified without a brute-force search. The statistical significance of bivariate flow clusters was evaluated using a Monte Carlo simulation. The experimental results on simulated datasets showed that BiFlowAMOEBA can identify bivariate flow clusters of different shapes more accurately and completely than the two state-of-the-art methods. Two case studies indicate that BiFlowAMOEBA may provide new insights for designing and optimizing transportation systems. The interaction patterns between public transit and taxi services identified in Beijing may aid in public transportation planning. The competition patterns between taxi and ride-hailing services identified in Xiamen may aid in the establishment of coordinated development among different transportation services.

BiFlowAMOEBA also has two limitations. First, it is time-consuming to evaluate the significance of bivariate flow clusters using Monte Carlo simulation. In the future, high-performance computing techniques (e.g. parallel computing) should be used to improve the computational efficiency of BiFlowAMOEBA. Second, BiFlowAMOEBA can only be used for aggregated bivariate flow data. The modifiable areal unit problem cannot be neglected. In the future, we would like to extend BiFlowAMOEBA for clustering individual OD flows.

**Acknowledgements**

The authors gratefully acknowledge the comments from the editor and the reviewers. The authors would like to thank Dr. Zhou Huang at Peking University for providing the traffic analysis zone data in Beijing.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Author contributions**

Qiliang Liu, Min Deng, and Jie Yang conceived and designed the presented idea. Jie Yang implemented the experiments and analyzed the results. Qiliang Liu and Jie Yang wrote the manuscript. Min Deng reviewed the manuscript and provided comments. Wenkai Liu and Rui Xu collected the research data and checked the code of the proposed method.

**Funding**

This study was funded through support from the National Natural Science Foundation of China (NSFC) (Nos. 41730105 and 41971353) and the Natural Science Foundation of Hunan Province (Nos. 2021JJ20058 and 2020JJ4695).

**Data and codes availability statement**

The data and codes that support the findings of this study are available on ‘figshare.com’, with the identifier at the private link: [https://doi.org/10.6084/m9.figshare.19640511](https://doi.org/10.6084/m9.figshare.19640511)
Notes on contributors

Qiliang Liu is currently an associate professor at Central South University, Hunan, China. His research interests focus on multi-scale spatio-temporal data mining and spatiotemporal statistics. He has published more than 30 peer-reviewed journal articles in these areas.

Jie Yang is currently a Ph.D. candidate at Central South University and his research interests focus on spatio-temporal clustering.

Min Deng is currently a professor at Central South University and the associate dean of the School of Geosciences and info-physics. His research interests are map generalization, spatio-temporal data analysis, and mining.

Wenkai Liu is currently a Ph.D. candidate at Central South University and his research interests focus on spatio-temporal clustering and association rule mining.

Rui Xu is currently pursuing a Ph.D. degree at the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing at Wuhan University, Wuhan, China. His major field of study is the application of GIS in water conservancy and hydropower projects.

References

Adrienko, N., and Adrienko, G., 2011. Spatial generalization and aggregation of massive movement data. *IEEE Transactions on Visualization and Computer Graphics*, 17 (2), 205–219.

Aldstadt, J. and Getis, A., 2006. Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*, 38 (4), 327–343.

Andrienko, G., et al., 2017. Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data. *IEEE Transactions on Visualization and Computer Graphics*, 23 (9), 2120–2136.

Anselin, L., 2019. A local indicator of multivariate spatial association: extending Geary’s C. *Geographical Analysis*, 51 (2), 133–150.

Berglund, S. and Karlström, A., 1999. Identifying local spatial association in flow data. *Journal of Geographical Systems*, 1 (3), 219–236.

Contreras, S. and Paz, A., 2018. The effects of ride-hailing companies on the taxicab industry in Las Vegas, Nevada. *Transportation Research Part A: Policy and Practice*, 115, 63–70.

Duque, J.C., et al., 2011. A computationally efficient method for delineating irregularly shaped spatial clusters. *Journal of Geographical Systems*, 13 (4), 355–372.

Gao, Y., et al., 2018. A multidimensional spatial scan statistic approach to movement pattern comparison. *International Journal of Geographical Information Science*, 32 (7), 1304–1325.

Guo, D.S. and Zhu, X., 2014. Origin-Destination flow data smoothing and mapping. *IEEE Transactions on Visualization and Computer Graphics*, 20 (12), 2043–2052.

Lee, S.I. and Cho, D., 2015. Developing a bivariate AMOEBA technique. In *GeoComputation 2015 Conference*, 20–23 May.

Lee, S.I., 2001. Developing a bivariate spatial association measure: an integration of Pearson’s r and Moran’s I. *Journal of Geographical Systems*, 3 (4), 369–385.

Liu, Q., et al., 2020. Network-constrained bivariate clustering method for detecting urban black holes and volcanoes. *International Journal of Geographical Information Science*, 34 (10), 1903–1929.

Liu, Q., et al., 2022. SNN_flow: a shared nearest-neighbor-based clustering method for inhomogeneous origin-destination flows. *International Journal of Geographical Information Science*, 36 (2), 253–227.

Liu, Y., Tong, D., and Liu, X., 2015. Measuring spatial autocorrelation of vectors. *Geographical Analysis*, 47 (3), 300–319.

Ord, J.K. and Getis, A., 2010. Local spatial autocorrelation statistic: distributional issues and an application. *Geographical Analysis*, 27 (4), 286–306.
Pei, T., et al., 2015. Density-based clustering for data containing two types of points. *International Journal of Geographical Information Science*, 29 (2), 175–193.

Song, C., et al., 2019b. Detecting arbitrarily shaped clusters in origin-destination flows using ant colony optimization. *International Journal of Geographical Information Science*, 33 (1), 134–154.

Song, C., Pei, T., and Shu, H., 2019a. Identifying flow clusters based on density domain decomposition. *IEEE Access*, 8, 5236–5243.

Tan, P.N., Steinbach, M., and Kumar, V., 2006. *Introduction to data mining*. 1st ed. Boston, MA: Addison-Wesley Longman Publishing, Co., Inc.

Tang, J.J., et al., 2021. Uncovering the spatially heterogeneous effects of shared mobility on public transit and taxi. *Journal of Transport Geography*, 95, 103134.

Tao, R., 2021. Big spatial flow data analytics. In: M. Werner and Y.Y. Chiang, eds. *Handbook of big geospatial data*. Switzerland: Springer, 163–183.

Tao, R. and Thill, J.C., 2016. Spatial cluster detection in spatial flow data. *Geographical Analysis*, 48 (4), 355–372.

Tao, R. and Thill, J.C., 2019a. FlowAMOeba: identifying regions of anomalous spatial interactions. *Geographical Analysis*, 51 (1), 111–130.

Tao, R. and Thill, J.C., 2019b. Flow cross K-function: a bivariate flow analytical method. *International Journal of Geographical Information Science*, 33 (10), 2055–2071.

Tao, R. and Thill, J.C., 2020. BiFlowLISA: measuring spatial association for bivariate flow data. *Computers, Environment and Urban Systems*, 83, 101519.

Taylor, G., 1994. Point in polygon test. *Survey Review*, 32 (254), 479–484.

Vrotsou, K., et al., 2017. An interactive approach for exploration of flows through direction-based filtering. *Journal of Geovisualization and Spatial Analysis*, 1 (1–2), 1–15.

Wang, D., Miwa, T., and Morikawa, T., 2021. Comparative analysis of spatial–temporal distribution between traditional taxi service and emerging ride-hailing. *ISPRS International Journal of Geo-Information*, 10 (10), 690.

Wartenberg, D., 2010. Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis*, 17 (4), 263–283.

White, C.E., Bernstein, D., and Kornhauser, A.L., 2000. Some map matching algorithms for personal navigation assistants. *Transportation Research Part C: Emerging Technologies*, 8 (1–6), 91–108.

Widener, M.J., et al., 2012. Developing a parallel computational implementation of AMOeba. *International Journal of Geographical Information Science*, 26 (9), 1707–1723.

Xiang, Q. and Wu, Q., 2019. Tree-based and optimum cut-based origin-destination flow clustering. *ISPRS International Journal of Geo-Information*, 8 (11), 477.

Yao, X., et al., 2018. A stepwise spatio-temporal flow clustering method for discovering mobility trends. *IEEE Access*, 6, 44666–44675.

Zhang, X., et al., 2018. Do different datasets tell the same story about urban mobility — A comparative study of public transit and taxi usage. *Journal of Transport Geography*, 70, 78–90.

Zhu, X. and Guo, D., 2014. Mapping large spatial flow data with hierarchical clustering. *Transactions in GIS*, 18 (3), 421–435.