Transportation
https://doi.org/10.1007/s11116-023-10415-y

Deciphering flow clusters from large-scale free-floating bike sharing journey data: a two-stage flow clustering method

Wendong Chen[1] · Xize Liu[1] · Xuewu Chen[1] · Long Cheng[1] · Jingxu Chen[1]

Accepted: 8 August 2023

Abstract

Extracting flow clusters consisting of many similar origin–destination (OD) trips is essential to uncover the spatio-temporal interactions and mobility patterns in the free-floating bike sharing (FFBS) system. However, due to occlusion and display clutter issues, efforts to identify inhomogeneous flow clusters from large journey data have been hampered to some extent. In this study, we present a two-stage flow clustering method, which integrates the Leiden community detection algorithm and the shared nearest-neighbor-based flow (SNN_flow) clustering method to efficiently identify flow clusters with arbitrary shapes and uneven densities. The applicability and performance of the method in detecting flow clusters are investigated empirically using the FFBS system of Nanjing, China as a case study. Some interesting findings can be drawn from the spatio-temporal patterns. For instance, the share of flow clusters used to meet the "first-/last-mile" demand at metro stations is reasonably high, both during the morning (71.85%) and evening (65.79%) peaks. Compared with the "first-/last-mile" flow clusters between metro stations and adjacent workplaces, the solution of the "first-/last-mile" flow clusters between metro stations and adjacent residences is more dependent on the FFBS system. In addition, we explored the shape and density distribution of flow clusters from the perspective of origin and destination points. The endpoint distribution characteristics demonstrate that the shape distribution of metro station point clusters is generally flatter and the spatial points within them are more concentrated than other sorts of point clusters. Our findings could help to better understand human movement patterns and home-work commute, thereby providing more rational and targeted decisions for allocating FFBS infrastructure resources.

Keywords Free-floating bike sharing · Origin–destination flow · Clustering · Commuting trips · Spatio-temporal pattern

✉ Xuewu Chen
chenxuewu@seu.edu.cn

[1] School of Transportation, Southeast University, Dongnandaxue 2, Nanjing 211189, China

Published online: 28 August 2023

Introduction

In recent years, smartphone-operated, non-station-based bike fleets (i.e., free-floating bike sharing, hereafter referred to as FFBS) have witnessed exponential growth worldwide (Cheng et al. 2022b; Hirsch et al. 2019). For instance, in less than two years since its launch in North America, FFBS has rapidly expanded to over 200 systems operating in more than 150 cities (Hirsch et al. 2019). Meanwhile, the FFBS system has ushered in a golden age of expansion in China, with its implementation in over 200 cities and a total of 23 million bikes in just a few years (Gu et al. 2019). By equipping shared bikes with a Global Positioning System (GPS) device, the FFBS system allows users to rent a nearby bike and return it in any suitable place (e.g., on-street corrals and sidewalk racks) via mobile applications (Zhao and Ong 2021). It greatly improves the flexibility and accessibility of journeys by offering "door-to-door" services for local residents (Cheng et al. 2022b).

As FFBS schemes continue to grow in popularity around the world, large FFBS journey data with individual mobile locations and trajectories are readily available (Chen et al. 2022b). In many studies based on journey data, researchers have found that point clusters with high FFBS usage were more concentrated near metro stations, residential neighborhoods, and office buildings (Chen and Ye 2021; Guo and He 2020; Du et al. 2019b). The findings provide valuable insights for understanding spatio-temporal travel characteristics, predicting regional demand, and designing scheduling strategies. However, most of them only considered the origin (O) or destination (D) points of FFBS trips, while the studies from the origin–destination (OD) flow perspective are limited (Chen et al. 2022b; Zhang et al. 2021). Although the discovery of OD flow clusters consisting of many similar trips is essential for unveiling daily human mobility and home-work commuting patterns (Guo et al. 2020; Liu et al. 2022a; Wood et al. 2010), the focus of existing FFBS studies has not yet been extended from O or D points to OD flows.

Due to occlusion and display clutter issues, a substantial amount of trips overlap and intersect each other, making it challenging to discover flow clusters in large flow data (Zhu and Guo 2014). Some studies aggregate trips using predefined spatial areal units (e.g., regular grids, and traffic analysis zones) (Chen et al. 2022b; Wood et al. 2010; Zheng et al. 2021). This aggregation approach is valid in reducing the flow cluttering problem, but it ignores the flow patterns at local scales (Zhu and Guo 2014; Zhu et al. 2019). In recent years, several flow clustering methods have been developed in an attempt to extract flow clusters from large flow data (Guo et al. 2020; Gao et al. 2018; Tao and Thill 2016). These clustering methods mitigate the cluttering and overlapping issues by extracting clusters of similar trips, while maximizing the spatial resolution of the data (Song et al. 2019; Zhu and Guo 2014). Nevertheless, detecting flow clusters with irregular shapes and uneven densities from large flow data is still a huge challenge (Liu et al. 2022a), which we will review in the "Flow clustering methods" section.

In reality, due to the nature of the short-distance trips, shared bikes are likely to stay near their initial assigned location (Zhang and Meng 2019). Inspired by this, we propose a two-stage flow clustering method by integrating the Leiden community detection and the shared nearest-neighbor-based flow (SNN_flow for short) clustering methods. More concretely, in Stage I, the Leiden algorithm is leveraged to partition the entire study area into multiple FFBS activity zones with strong intra-connections, thus decomposing a large flow clustering problem into multiple small sub-problems. In Stage II, the FFBS flow clusters with varying shapes and densities in each activity zone are identified separately using the SNN_flow method, and then the extraction results of all activity zones are merged.

Taking the FFBS system in Nanjing, China as a case study, an empirical investigation is performed on the applicability and performance of the two-stage flow clustering method in identifying flow clusters. This study tackles the following two research questions: (i) What are the typical characteristics of the spatio-temporal patterns of FFBS flow clusters? (ii) What are the similarities and diversities in the shape and density distribution of FFBS flow clusters? This study contributes to the existing literature in two ways. First, it proposes a two-stage flow clustering method that can be leveraged to efficiently detect FFBS flow clusters with arbitrary shapes and inhomogeneous densities from large-scale journey data. Second, it unveils the spatio-temporal patterns and endpoint distribution characteristics of FFBS flow clusters, which could help transportation planners and decision-makers to better understand the heterogeneity of flow clusters and thus take rational measures to make the resource allocation of the FFBS system as balanced as possible.

The remainder of the paper is structured as follows. "Literature review" section provides an overview of FFBS OD flows and flow clustering methods. "Two-stage flow clustering method" section introduces the two-stage flow clustering method in detail. "Study area and data description" section describes the study area and the data used. "Results and discussions" section presents the research findings. Finally, our main conclusions are summarized and policy implications are drawn in "Conclusions and policy implications" section.

Literature review

FFBS OD flows

Numerous existing studies on FFBS data analysis have focused on revealing the mechanisms influencing O or D point usage patterns. These studies have investigated different aspects of this issue, including socio-demographics (Link et al. 2020; Orvin and Fatmi 2021), weather conditions (Peters and MacKenzie 2019), land use (Chen and Ye 2021; Cheng et al. 2020a), built environment (Guo and He 2020; Shen et al. 2018), and access to metro system (Cheng et al. 2022a, 2023; Ma et al. 2019). On balance, FFBS usage is higher in areas with denser populations, comfortable weather conditions, higher land use mix, friendlier cycling environments, and better interchange facilities. These findings are of great importance in many facets such as cycling facility planning (Zhao and Ong 2021), bike scheduling strategy design (Chang et al. 2021), and ridership activity prediction (Xu et al. 2018). However, most of them use isolated models to analyze trip origins and destinations, and few investigate FFBS trips from the perspective of OD flows.

Abstracting flow clusters from large-scale, chaotic journey data is crucial to reveal the spatio-temporal dynamics of human mobility and commuting patterns (Liu et al. 2022a). Currently, only some initial works have looked at OD flows using FFBS journey data. Based on the Shanghai Mobike dataset, Du et al. (2019b) visualized the spatio-temporal distribution of FFBS OD flows by exploiting the ODPFM (O-D Proportion Flow Map) tool. They found that the spatial distribution of FFBS OD flows varied considerably by land use type and period of time. Zheng et al. (2021) constructed an OD spatial network using Beijing Mobike dataset and investigated the unbalance characteristics of the FFBS system. The results suggested that most of the study areas are in a relatively flat stage of supply and demand, while a few areas have large imbalances in resource supply and demand. Drawing on a four-month FFBS OD flow dataset in Singapore, Zhang et al. (2021) identified some activity zones from cycling behaviors by applying a modularity optimization community detection method. They found that the activity zones yielded from the FFBS networks are locally clustered. Furthermore, taking Nanjing, China as an example, Chen et al. (2022b) constructed a spatial interaction network using FFBS OD flows. Based on this, the urban activity zone borders were delineated leveraging the Leiden algorithm. They pointed out that the FFBS activity zone borders overlap more with natural borders (e.g., water bodies and mountains) than with administrative borders.

The aforementioned studies addressing FFBS mobility patterns often focus on the OD flows from one individual area to the other, providing valuable findings for the spatial interactions of the FFBS system. However, since there is no fixed station constraint for FFBS bikes, these studies typically use regular grids to aggregate FFBS usage (Du et al. 2019b; Zhang et al. 2019; Zheng et al. 2021), and few have investigated FFBS flow clusters from a finer spatial resolution. Many questions regarding what are the typical spatio-temporal patterns of FFBS flow clusters and whether they have varying shape and density distributions remain unanswered. Therefore, it is necessary to employ an efficient flow clustering method to detect inhomogeneous flow clusters from large-scale FFBS journey data.

Flow clustering methods

According to the basic principles of flow clustering, the related methods focus on the following categories: hierarchical clustering, statistical-based clustering, and density-based clustering.

In the hierarchical clustering methods, researchers first calculate the similarity between OD trips based on specific metrics (e.g., OD point locations and flow properties), and then use a specified strategy (e.g., agglomerative and divisive) to construct OD flow data into a hierarchical structure to identify flow clusters (Guo et al. 2020). For instance, Zhu and Guo (2014) considered both start and end positions in defining the similarity of two trips and proposed an agglomerative clustering approach to handle large-scale flow data. Yao et al. (2018) developed a new spatial similarity metric based on the angle and length differences between any pair of trips, and a similar agglomerative clustering approach was applied to extract flow clusters. Xiang and Wu (2019) proposed a new hierarchical clustering method (called TOCOFC) to obtain flow clusters from the original, chaotic trips. The method introduces a similarity metric to measure the spatio-temporal similarity between different trips, and then employs a recursive optimum cut-based approach to partition trips. In summary, the hierarchical clustering methods have been widely utilized in small-scale OD flow datasets, but may not be applicable to large OD flow datasets like FFBS journey data because of their high computational complexity (Liu et al. 2022a).

In the statistical-based clustering methods, researchers have extended the traditional spatial statistics in hopes of detecting flow clusters from large OD flow datasets. For instance, Liu et al. (2015) improved the global and local Moran's I statistics to extract flow clusters containing highly spatially correlated trips, and conducted an empirical study using taxi data from Shanghai, China as a case study. Tao and Thill (2016) proposed a *K*-function extension method for OD flow data to upgrade its detection target from point clusters to flow clusters. In addition, Gao et al. (2018) introduced a multidimensional spatial scan statistics approach to identify flow clusters. These statistical-based clustering methods can effectively detect statistically significant flow clusters, but they have obstacles in detecting flow clusters with arbitrary shapes[1] (Song et al. 2019).

Given that some density-based clustering algorithms (e.g., DBSCAN (Ester et al. 1996) and OPTICS (Ankerst et al. 1999)) are well able to identify irregularly shaped point clusters, researchers have successfully upgraded point clustering to flow clustering by improving these traditional algorithms (Gallego et al. 2018; Tao and Thill 2016). Although such density-based clustering approaches have competitive advantages in detecting arbitrarily shaped clusters, they exhibit poor clustering performance when the density of OD flows is unevenly distributed (Reddy and Bindu 2017). Furthermore, these methods are mainly developed based on Euclidean spatial distances (Liu et al. 2022a). However, related studies have shown that Euclidean distance-based clustering methods may introduce a significant systematic bias in the presence of network constraints (Besse et al. 2016; Yamada and Thill 2010). For FFBS journey data, the above methods are clearly not applicable as their OD points are typically strongly constrained by road networks (Hua et al. 2020; Zhang et al. 2019). To handle these issues, Liu et al. (2022a) recently presented a shared nearest-neighbor-based flow (SNN_flow) clustering method, which possesses superior performance in identifying clusters of network-constrained OD flows with irregular shapes and inhomogeneous distributions. However, for the city-level OD flow data, the efficiency of the SNN_flow method does not seem to be ideal as it has a relatively high time complexity.

By and large, existing studies have tried many clustering algorithms in the flow cluster extraction problem to provide valuable insights for unveiling human mobility patterns. However, they still have gaps in effectively detecting flow clusters of varying shapes and densities from large-scale OD flow data. To this end, we combine the respective advantages of community detection and SNN_flow methods, and propose a two-stage flow clustering method to extract FFBS flow clusters, trying to provide profound insights for uncovering human mobility patterns and allocating infrastructure resources.

Two-stage flow clustering method

In Stage I, the study area is divided into multiple FFBS activity zones using the Leiden algorithm. Based on this, the FFBS flow clusters within each activity zone are identified separately in Stage II employing the SNN_flow method.

Stage I: activity zone delineation

A three-step identification framework is developed to delineate the FFBS activity zone borders, as depicted in Fig. 1. First, we construct an undirected weighted network ($G=(V, E)$) upon the FFBS trips (Fig. 1a, b), where V is the vertex set of the network G, consisting of the centroids of all spatial units; E is the edge set of the network G, consisting of all links between each pair of centroids; each edge e corresponds to a weight $W(e)$, which represents the size of OD flow (i.e., FFBS trip count).

Second, based on the community detection algorithm, all vertices are divided into multiple vertex groups (Fig. 1c). The basic principle of this algorithm is to form a community

[1] The shape of a flow cluster is determined by the direction and length of the similar trips it contains (Tao and Thill 2016).

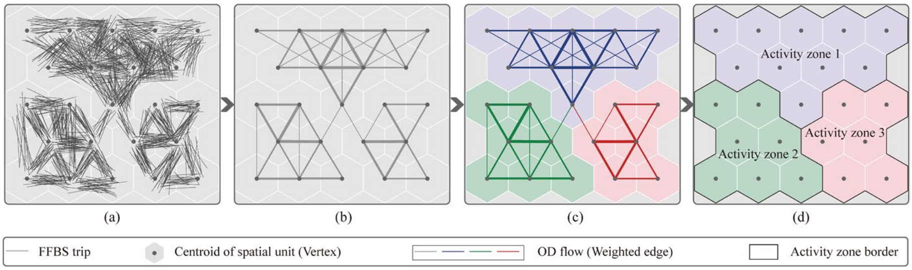


Fig. 1 Identification framework of FFBS activity zones: **a** FFBS trip preparation; **b** undirected weighted network construction; **c** community structure division; and **d** activity zone delineation

structure based on the degree of connection between vertices. That is, vertices located in the same community are relatively closely connected, while vertices located in different communities are very sparsely connected to each other (Girvan and Newman 2002). The Louvain algorithm is a classical community detection algorithm, covering two elementary phases (i.e., node local movement and network aggregation), which provides an efficient solution for vertex grouping (Jin et al. 2021). However, it has been found that the Louvain algorithm may derive some internally poorly connected communities during the community structure partitioning process (Traag et al. 2019). To overcome this defect, Traag et al. (2019) recently extended the Louvain method by adding a partition refinement phase and proposed the so-called Leiden algorithm, which is more computationally efficient and uncovers better partition structures. In this work, the Leiden community detection algorithm is exploited to discover the partition structure of the FFBS system.

In addition, we measure the performance of the communities partitioned by the Leiden algorithm based on the modularity Q. The value of Q lies between zero and one, and the larger the value, the more stable the corresponding community structure (Jin et al. 2021). The modularity Q of the weighted network (Arenas et al. 2008) can be written as:

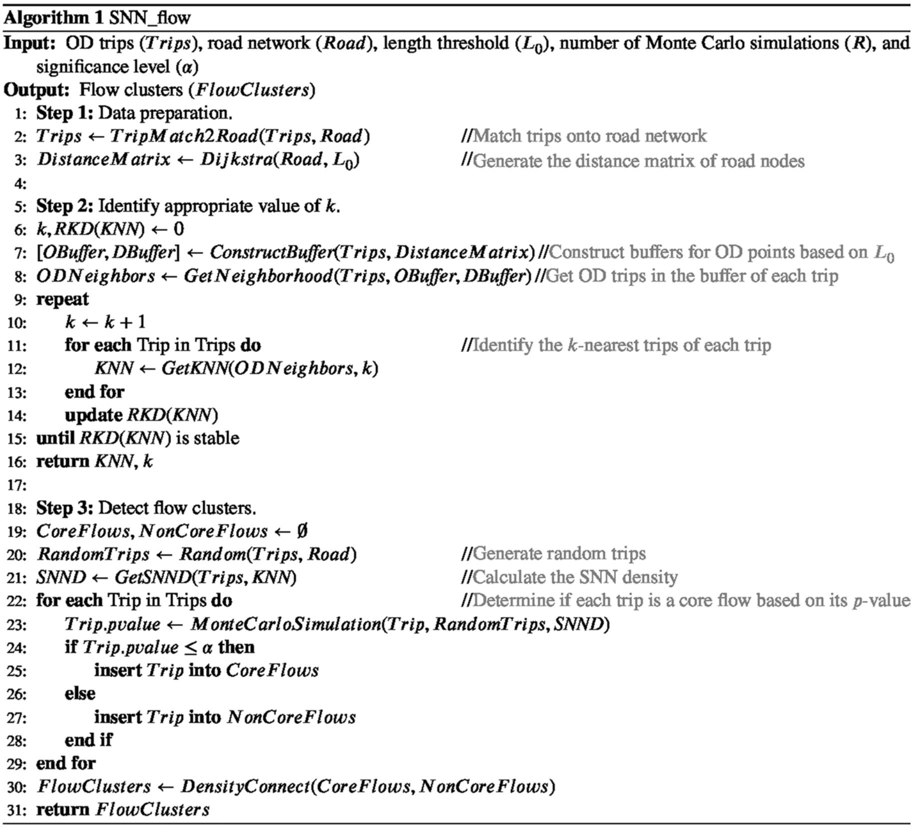
$$Q = \frac{1}{2W}\sum_{ij}\left(W_{ij} - \frac{s_i s_j}{2W}\right)\delta\left(c_i, c_j\right) \tag{1}$$

where W is the weight values of all edges, W_{ij} is the weighted adjacency matrix (i.e., the weight value of the edge between vertices i and j), s_i (s_j) refers to the strength of vertex i (j), c_i (c_j) refers to the community to which vertex i (j) is partitioned, and $\delta(\cdot)$ is an indicator function, if $c_i = c_j$, then $\delta = 1$, else, $\delta = 0$.

Third, the borders of the community structures are automatically identified and highlighted with the help of the ArcGIS Dissolve Boundaries tool (Fig. 1d). These communities have dense FFBS trips within them and can serve as effective proxies for user activity spaces (Chen et al. 2022b). The study area is eventually separated by the borders of the FFBS activity zones.

Stage II: flow cluster identification

For each FFBS activity zone delineated in the previous phase, the SNN_flow method is exploited to identify its respective flow clusters. Specifically, the SNN_flow method consists of three essential steps. First, the FFBS trips and road network datasets are collected and preprocessed as inputs for the subsequent steps ("Data preparation" section). Second, a suitable k value is estimated to determine the clustering scale ("Appropriate k-value estimation" section). Finally, the flow clusters of FFBS are detected based on the SNN density ("Flow cluster detection" section). The pseudo-code of SNN_flow is given in Algorithm 1.



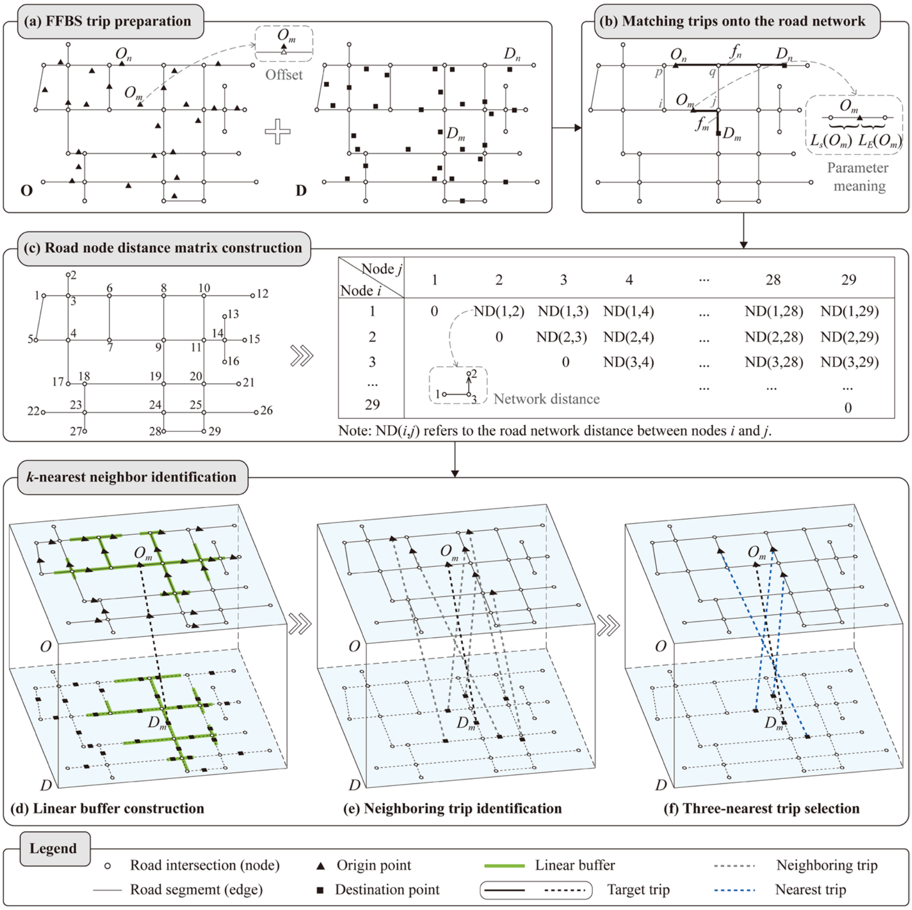
Algorithm 1 SNN_flow

```
Input: OD trips ($Trips$), road network ($Road$), length threshold ($L_0$), number of Monte Carlo simulations ($R$), and
    significance level ($\alpha$)
Output: Flow clusters ($FlowClusters$)
Step 1: Data preparation.
$Trips \leftarrow TripMatch2Road(Trips, Road)$            //Match trips onto road network
$DistanceMatrix \leftarrow Dijkstra(Road, L_0)$           //Generate the distance matrix of road nodes

Step 2: Identify appropriate value of $k$.
$k, RKD(KNN) \leftarrow 0$
$[OBuffer, DBuffer] \leftarrow ConstructBuffer(Trips, DistanceMatrix)$ //Construct buffers for OD points based on $L_0$
$ODNeighbors \leftarrow GetNeighborhood(Trips, OBuffer, DBuffer)$ //Get OD trips in the buffer of each trip
repeat
    $k \leftarrow k + 1$
    for each Trip in Trips do                             //Identify the $k$-nearest trips of each trip
        $KNN \leftarrow GetKNN(ODNeighbors, k)$
    end for
    update $RKD(KNN)$
until $RKD(KNN)$ is stable
return $KNN$, $k$

Step 3: Detect flow clusters.
$CoreFlows, NonCoreFlows \leftarrow \emptyset$
$RandomTrips \leftarrow Random(Trips, Road)$              //Generate random trips
$SNND \leftarrow GetSNND(Trips, KNN)$                     //Calculate the SNN density
for each Trip in Trips do                                 //Determine if each trip is a core flow based on its $p$-value
    $Trip.pvalue \leftarrow MonteCarloSimulation(Trip, RandomTrips, SNND)$
    if $Trip.pvalue \leq \alpha$ then
        insert $Trip$ into $CoreFlows$
    else
        insert $Trip$ into $NonCoreFlows$
    end if
end for
$FlowClusters \leftarrow DensityConnect(CoreFlows, NonCoreFlows)$
return $FlowClusters$
```

Data preparation

With no fixed dock limitation, FFBS bikes can be parked freely near buildings along urban streets (as long as parking is permitted), such that the location of some OD points is somewhat offset from the road segment (e.g., point O_m in Fig. 2a). To address this issue, a map matching approach is used to match each pair of OD points onto their nearest road segment (White et al. 2000). A road network is then constructed based on the existing road dataset to search for the network distance[2] between OD point pairs as a proxy for trip trajectories. Network distance is usually a more accurate reflection of users' actual travel behavior

[2] Network distance refers to the shortest path between two points using a road network, where the shortest path is measured by the travel weight (e.g., travel time or distance) of the network edges (Apparicio et al. 2008).



Node j / Node i	1	2	3	4	...	28	29
1	0	ND(1,2)	ND(1,3)	ND(1,4)	...	ND(1,28)	ND(1,29)
2		0	ND(2,3)	ND(2,4)	...	ND(2,28)	ND(2,29)
3			0	ND(3,4)	...	ND(3,28)	ND(3,29)
...					...	...	...
29							0

Note: ND(i,j) refers to the road network distance between nodes i and j.

Fig. 2 Identification process of the three-nearest neighbors of a certain trip (f_m): **a** FFBS trip preparation; **b** matching FFBS trips onto the road network; **c** road node distance matrix construction; and **d**–**f** *k*-nearest neighbor identification

than a straight-line path (i.e., Euclidean distance) (Apparicio et al. 2008; Páez et al. 2020). Taking the f_m trip composed of origin point (O_m) and destination point (D_m) as an example, the spatial distribution of its network distance is shown in Fig. 2b. More specifically, the f_m trip can be expressed as (O_m, D_m, $L_S(O_m)$, $L_E(O_m)$, $L_S(D_m)$, $L_E(D_m)$), where $L_S(O_m)$ and $L_E(O_m)$ denote the length of the shortest path between the origin point (O_m) and the start/end node[3] of the road segment where it is located, respectively; similarly, $L_S(D_m)$ and $L_E(D_m)$ are the shortest path lengths from D_m to the start node and end node of its road segment, respectively.

According to the basic principle of the SNN_flow algorithm, it is necessary to further search for the *k*-nearest neighbors corresponding to each trip (Liu et al. 2022a).

[3] To facilitate the distinction, we set the node with the smaller longitude (latitude) value in a road segment as the start node and the other node as the end node.

Hence, we need to calculate the distance (for this study, the network distance is used) between any two trips in advance. Taking f_m and f_n as an example, the network distance between them is calculated by the following formula (Shu et al. 2021):

$$ND(f_m,f_n) = ND(O_m, O_n) + ND(D_m, D_n) \tag{2}$$

where $ND(O_m, O_n)$ refers to the network distance between the origin points of f_m and f_n, and $ND(D_m, D_n)$ refers to the network distance between the destination points of f_m and f_n. For the two origin points O_m and O_n located at the road segments S_{ij} and S_{pq} in Fig. 2b, the network distance $ND(O_m, O_n)$ between O_m and O_n can be chosen as the minimum value from the following two cases: (i) $L_S(O_m)+ND(i, p)+L_S(O_n)$; (ii) $L_E(O_m)+ND(j, q)+L_E(O_n)$. From this, we can see that the network distance between two trips contains two parts, one is an uncertain distance consisting of road nodes and OD points (e.g., $L_S(O_m)$, and the other is a fixed distance consisting of road nodes (e.g., $ND(i, p)$).

In reality, the network distance between road nodes is fixed and does not change with the location of the OD points. Therefore, we can calculate this part of the distance in advance to reduce the workload of calculating the network distance between OD points and improve computational efficiency. Figure 2c displays an example of constructing a distance matrix based on local network nodes.

Appropriate k-value estimation

As mentioned above, in order to obtain the SNN density of a trip, we need to first identify its network-constrained k-nearest neighbors (Fig. 2f). In practice, most trips are within an acceptable distance from their kth nearest neighbors when the value of k is not large (Pei et al. 2012). This means that we only need to compute the network distance of f_m from those trips within a certain range from it. In this study, linear buffers of length L_0 are drawn for the origin and destination points of each trip, respectively. Figure 2d depicts the linear buffers at both ends of f_m. The network distances from the O_m (or D_m) point to the other origins (or destinations) in the buffer are all less than L_0. L_0 depends on the most of trip distances for a given transportation mode. For FFBS, 5 km is typically considered as its longest trip distance (Chen 2021), hence that is the length threshold (L_0) adopted in this study.

After a linear buffer is constructed for each trip, its neighboring trips located within the buffer can be further extracted. Figure 2e illustrates the neighboring trips of f_m within its buffer. Then, combined with the local road node distance matrix, we efficiently calculate the network distance between the target trip f_m and its neighboring trips, and on this basis, identify the k-nearest trips of f_m. Figure 2f shows the three-nearest trips identified from the neighboring trips of f_m (assuming $k=3$).

As we can see, estimating the appropriate value of k is very critical as its magnitude determines the reasonableness of the SNN density distribution. Either too low or too high k-value will affect the normal estimation of SNN density. Many researchers have applied the ratio between the variance of the $(k+1)$th nearest distance and that of the kth nearest distance (*RKD* for short, which refers to the capitalized initials of ratio, $(k+1)k$ and distance, respectively) to estimate the appropriate k value with good performance (Liu et al. 2022a; Pei 2011), and hence that is the index used in this study.

$$RKD = \frac{\mathrm{Var}^*_{k+1}(x)}{\mathrm{Var}^*_k(x)} / R_k \quad (k \geq 1) \tag{3}$$

where $\mathrm{Var}^*_k(x)$ denotes the variance of the kth nearest distance of the trips (i.e., the distance between each trip and its kth nearest trip is first calculated, and then the variance of all distances is calculated); $\mathrm{Var}^*_{k+1}(x)$ has a similar meaning; and R_k is a constant term whose value is equal to the ratio of the expectation value of the above two distances. As the k value increases, the RKD value will gradually level off. We can easily identify the magnitude of the k value when RKD is at the leveling-off change point. For more details, please refer to the study of Pei (2011).

Flow cluster detection

In the previous subsection, we identified k neighboring trips for each trip, and in this subsection, the SNN density of each trip is estimated to finally detect flow clusters. Following Ester et al. (1996) and Liu et al. (2022a), we introduce some important concepts for SNN algorithm:

- **Definition 1** (SNN *similarity*). The number of nearest neighbors shared by the k-nearest trips of any two trips. For trips f_m and f_n, their k-nearest trips can be expressed as $KNN(f_m)$ and $KNN(f_n)$, and the SNN similarity of them can be expressed as:

$$SNN(f_m, f_n) = \left| KNN(f_m) \cap KNN(f_n) \right| \tag{4}$$

- **Definition 2** (*Directly reachable*). If $SNN(f_m, f_n) \geq k/2$, the two trips, f_m and f_n, are directly reachable.
- **Definition 3** (SNN density). It refers to the number of trips that are directly reachable from a particular trip (e.g., f_m).
- **Definition 4** (*Core flow*). For a particular trip, f_m, if p-value$(f_m) \leq \alpha$ (α is the significance level), then it is regarded as a core flow. The p-value of f_m can be written as:

$$p\text{ - value}(f_m) = \frac{\sum_{i=1}^{R} I_i(SNND_r(f_m) \geq SNND_o(f_m))}{1 + R} \tag{5}$$

where $I_i(\cdot)$ is an indicator function, if $SNND_r(f_m) \geq SNND_o(f_m)$, then $I_i = 1$, else, $I_i = 0$. $SNND_r(f_m)$ and $SNND_o(f_m)$ refer to the SNN density of f_m calculated from the random trips and observed trips. R is the number of Monte Carlo simulations. Some researchers confirmed that Monte Carlo simulation can minimize the sampling effort without affecting the overall performance of the model when $\alpha = 0.05$, $R = 99$ (Silva et al. 2009; Liu et al. 2022b), which is used in this study.

- **Definition 5** *(Border flow)*. A trip that is directly reachable from a core flow but is not itself a core flow.
- **Definition 6** (*Noise flow*). A trip that is neither a core flow nor directly reachable from one.

The following steps describe how the SNN algorithm detects flow clusters based on the density-connectivity mechanism. (i) A case (f_m) is randomly selected from the dataset. The case f_m is considered a core flow if its p-value is smaller than or equal to α. Immediately

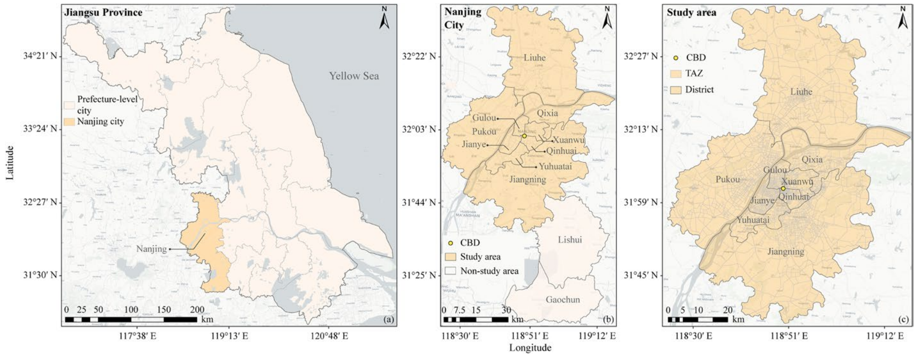


Fig. 3 Spatial distribution of **a** Jiangsu province; **b** Nanjing city; and **c** study area

afterwards, f_m is added to an initial cluster and a cluster ID is assigned to it (e.g., C_k). If the case f_m is not a core flow, the SNN algorithm moves on to another case; (ii) We assume that the algorithm selects a case (f_m) and finds it is a core flow. The algorithm then visits each of the reachable cases that are directly reachable with f_m and repeats the same task: calculate the SNN density. If the reachable case is also identified as a core flow, it is added to the C_k cluster; (iii) If the algorithm finds a reachable case that is directly reachable with f_m but has a p-value greater than α, then this case is considered as a border flow. A border flow can still be added to a C_k cluster as long as it is directly reachable from any core flow in the C_k cluster. The search continues recursively until all reachable cases of f_m are visited; (iv) The algorithm selects a case in the dataset that it has not visited before and starts the process of (i)–(iii) all over again. Those cases that are neither core flows nor directly reachable from one are grouped into the noise flows. Finally, a flow cluster consisting of core and border flows aggregates a certain number of spatially similar trips.

As stated earlier, the SNN_flow method consisting of three essential steps is utilized to identify the flow clusters for each activity zone. It is thus necessary to finally merge the flow clusters of all activity zones for subsequent analysis.

Study area and data description

Study area

Nanjing is the capital of Jiangsu province of China, a megacity and the second largest city in the East China region (Fig. 3a). Nanjing had a total area of 6587 km^2 and a population of 8.33 million as of 2018. There are 11 administrative districts, six of which are urban districts (i.e., Gulou, Jianye, Xuanwu, Qinhuai, Yuhuatai, and Qixia) and the remaining five are suburban districts (i.e., Liuhe, Pukou, Jiangning, Lishui, and Gaochun) (Cheng et al. 2020b), as shown in Fig. 3b.

Since the beginning of 2017, FFBS was first launched in Nanjing and quickly attracted numerous users due to its advantages such as flexible mobility and smart rental process (Hua et al. 2020). FFBS is usually backed by venture capital funding. For profit-making purposes, most bikes are assigned to densely populated areas with high demand (Cheng et al. 2020a; Gu et al. 2019). Nanjing is no exception, and citizens in its peripheral districts (i.e., Lishui and Gaochun) have no FFBS bikes to use. Therefore, the remaining nine administrative districts of Nanjing are selected as the study area (see Fig. 3c). Note that the traffic analysis zone (TAZ) within the study area was adopted as the spatial unit for delineating the FFBS activity zones (see "Stage I: Activity zone delineation" section).

Data description

The FFBS journey data were provided by Mobike, which at the time had the largest share of the FFBS fleet in Nanjing (Cheng et al. 2022b). The dataset records journey information of users, including fields such as user ID, bike ID, unlock time, lock time, coordinates of origins and destinations. We focus on the mobility pattern of the FFBS system on weekdays. In this study, data for only three consecutive weekdays (from 12 (Tuesday) to 14 (Thursday) September 2017) are used due to data availability. Nevertheless, they could still serve as a valid sample to validate the applicability of the method and unravel the daily patterns of FFBS trips (Guo and He 2020). During this period, the average temperature in Nanjing was between 20 °C and 28 °C with no rainfall, which was suitable for outdoor activities such as cycling. To mitigate the interference of abnormal data, we removed FFBS journeys with travel times less than 2 min or longer than 120 min (Chen et al. 2022b; Zhao et al. 2015). Nearly 1.9 million trips made by a total of 190,008 bikes were eventually recorded.

The road dataset was obtained from Amap (https://ditu.amap.com/), one of the most popular mapping service providers in China. In order to calculate the network distance between adjacent FFBS trips (see "Data preparation" section), a road network needs to be constructed on the basis of the original road dataset with the help of ArcGIS Network Analyst Extension. In addition, we applied a solution recently developed by Xu (2022) to further refine the connectivity of the road network by checking and modifying its topology (https://github.com/xuxinkun0591/gaode2/).

Another dataset we adopted is the land use map provided by the Nanjing Planning Bureau. The land use map consists of many polygons with different shapes, and each

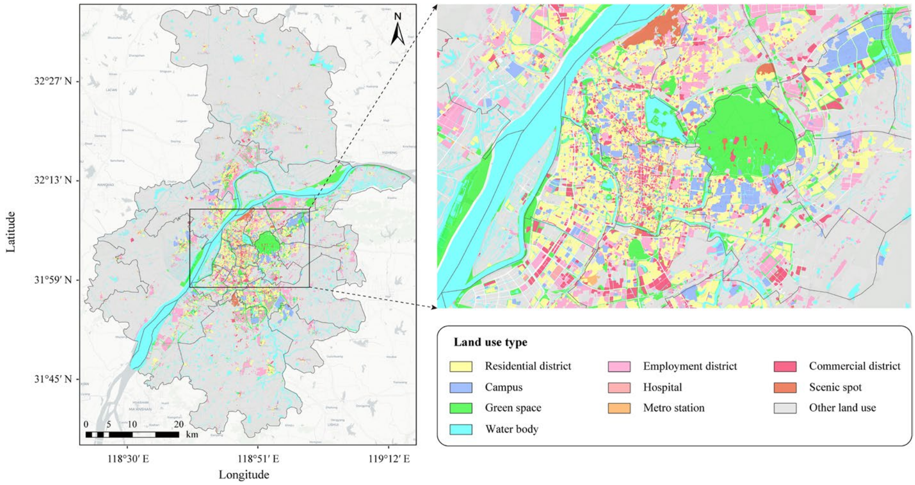


Fig. 4 Spatial distribution of different land use types in the study area

polygon has a corresponding land use type attribute. In line with a related study (Pan et al. 2012), a variety of land use types are divided, including campus, hospital, scenic spot, metro station, employment district, residential district, commercial district, green space, water body, and other land use. The spatial distribution of different land use types in the study area is shown in Fig. 4. Based on the land use information, we can initially infer the travel purpose of FFBS trips in the subsequent analysis (Lei et al. 2020).

Results and discussions

Flow clusters identification

Application of the two-stage flow clustering method

In a recently published work by Chen et al. (2022b), the partitioning of FFBS activity zones has been examined employing the Leiden algorithm using the same data source. The community structure of the study area is obtained in just a few seconds due to the low time complexity of the Leiden algorithm, which proves that this algorithm is very efficient. The study pointed out that the most robust community structure was yielded when the entire study area was divided into 22 FFBS activity zones (Fig. 5b). It can be seen from Fig. 5a, b that the FFBS activity zone borders coincide with the established administrative borders in a small percentage.

To validate the rationality of the FFBS activity zone delineation, the proportional distribution of FFBS trips within and between regions is investigated for three weekdays (September 12 to 14, 2017), as shown in Fig. 5c, d. First, as we can see, while most FFBS trips are distributed within the same administrative district (87.89%), there is still a certain share of FFBS trips used to connect different administrative districts (12.11%). This distribution characteristic is more prominent in urban districts (e.g., Gulou, Qinhuai, and Xuanwu districts). By contrast, the activity zones delineated by the Leiden algorithm have stronger intra-zone connections (92.34%). While the number of activity zones is increasing, FFBS trips between them show the opposite trend (i.e., inter-zone trips, 7.66%). It means that activity zone borders could portray FFBS user travel behavior and urban spatial structure in a more reasonable way. Therefore, by dividing the study area into multiple activity zones, a complex network can be decomposed into multiple sub-networks. This process is expected to significantly improve the computational efficiency of the SNN_flow method while minimizing the effect of inter-zone connections.

Then, the flow clusters within each activity zone are detected separately using the SNN_flow method, and the flow cluster detection results are further merged for all activity zones. It is noteworthy that morning peak (7:00–9:00, referred to as AM) and evening peak (17:00–19:00, referred to as PM) are considered the focus of flow clusters analysis, as FFBS usage is higher and more time-concentrated during these periods. In addition, we extract flow clusters with the number of similar trips greater than 30 from the daily AM and PM peaks to ensure that the number of flow clusters is within a reasonable range (Liu et al. 2022a). Taking the activity zone 14 as a case study, the details of flow clusters identification are illustrated in Appendix 1.

Table 1 depicts a summary of the flow clusters identified for all activity zones during the AM and PM peaks. On the whole, the number of similar trips and flow clusters stabilized

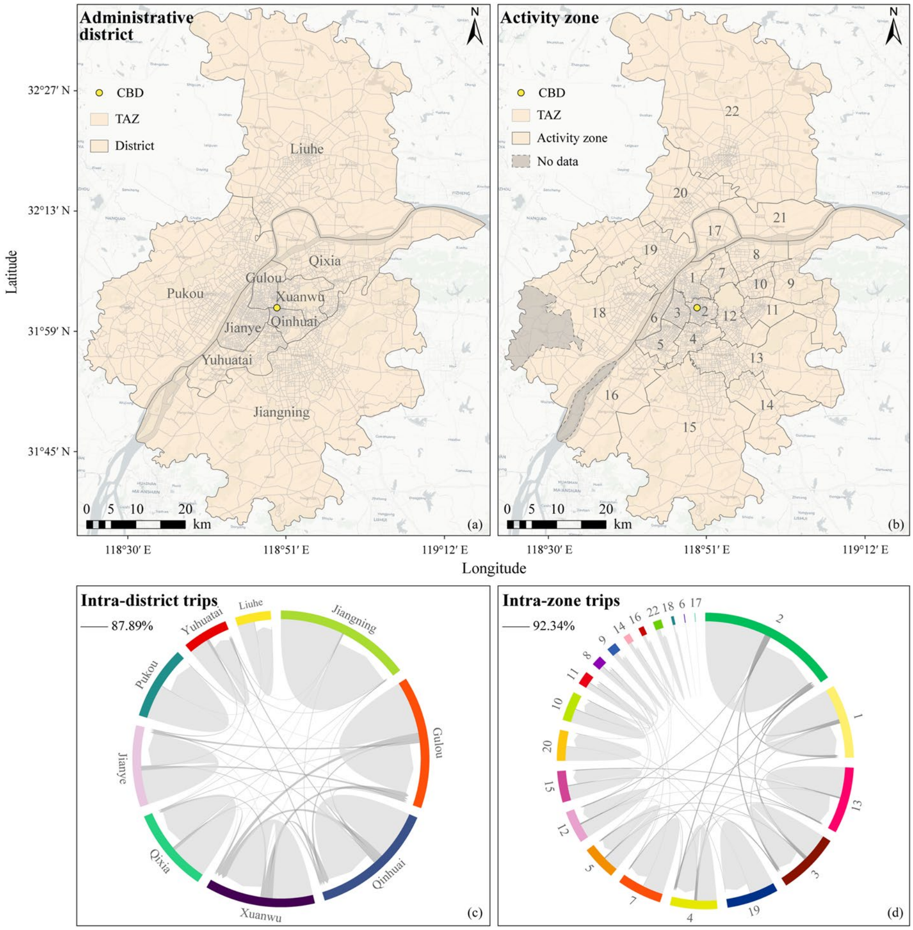


Fig. 5 Spatial distribution of **a** administrative districts and **b** activity zones in the study area; and proportional distribution of FFBS trips from September 12 to 14, 2017 of **c** administrative districts and **d** activity zones in the study area. *Note* spatial distribution of activity zones **b** is adapted from Chen et al. (2022b). Those connections with less than 100 FFBS trips are not displayed in this figure **c**–**d** to avoid display clutter issues

at an equilibrium level during three different weekdays. For instance, during the AM peak, the number of flow clusters remained between 375 and 404, corresponding to a percentage of similar trips located between 16.92% and 17.57%. Nevertheless, we found salient differences in the number and size of flow clusters identified between the AM and PM peaks. Taking September 12, 2017 (Tuesday) as an example, while the number of raw trips during the PM peak (117,386) was larger than that during the AM peak (108,387), the number of flow clusters extracted during these two time periods showed an opposite trend (328 for PM vs. 375 for AM), and the corresponding number and proportion of similar trips also followed this trend. This implies that more flow clusters are identified and the size of the flow clusters is usually larger during the AM peak compared to the PM peak (see the mean values in Table 1). A plausible explanation is that commuters tend to have less stringent time constraints for returning home during the PM peak, during which they may complete

Table 1 Summary results of flow cluster detection from September 12 to 14, 2017

Date	Period	No. of raw trips	No. of similar trips (percentage)	No. of flow clusters	Mean (unit: similar trips / flow cluster)[a]	Std. Dev. (unit: similar trips /flow cluster)[b]
Sept. 12, 2017 (Tuesday)	AM	108,387	18,443 (17.02%)	375	49.2	30.2
	PM	117,386	15,889 (13.54%)	328	48.4	28.8
Sept. 13, 2017 (Wednesday)	AM	112,232	19,716 (17.57%)	404	48.8	30.6
	PM	118,320	15,657 (13.23%)	337	46.5	27.6
Sept. 14, 2017 (Thursday)	AM	111,751	18,907 (16.92%)	390	48.5	30.8
	PM	117,217	15,126 (12.90%)	329	46.0	27.4
Total	AM	332,370	57,066 (17.17%)	1169	48.8	31.1
	PM	352,923	46,672 (13.22%)	994	47.0	28.3

[a] It refers to the average number of similar trips contained in each flow cluster

[b] It refers to the standard deviation of the number of similar trips contained in each flow cluster

Table 2 Running time of SNN_flow and the two-stage flow clustering methods for estimating appropriate *k* values

No. of raw trips	SNN_flow method		Two-stage flow clustering method	
	Time (s)	No. of *k* values	Time (s)	No. of *k* values
10,000	138	1[a]	31	22[b]
20,000	436	1	130	22
30,000	1323	1	286	22
40,000	2859	1	557	22
50,000	6957	1	934	22

[a] The entire study area is taken as input and a unique *k* value is output

[b] For each activity zone, one corresponding *k* value is output separately. The study area is divided into 22 activity zones, so there are 22 *k* values in total (Fig. 11a–c shows the case of determining the appropriate *k* value for activity zone 14)

some discretionary activities (e.g., shopping, eating, and entertainment) (Chen et al. 2022c; Ji et al. 2017). This leads to a reduction in the share of commuting demand that concentrates a large number of similar trips.

Efficiency comparison of flow clustering methods

In this subsection, we focus on comparing the efficiency of SNN_flow method and the two-stage flow clustering method (Leiden & SNN_flow) in extracting flow clusters. The largest difference between the two methods in the process of identifying flow clusters is the input. More specifically, the former takes the dataset of the entire study area as input, while the latter first partitions the study area into 22 activity zones, and then takes the dataset of each activity zone as input separately. Both methods take a little time in the data preparation step (see "Data preparation" section), but the former method has difficulty in obtaining results within a limited time in the flow cluster detection step (see "Flow cluster detection" section). Under this circumstance, the running time of the appropriate *k*-value estimation step (see "Appropriate k-value estimation" section) was selected as a proxy to compare the efficiency of these two methods in this study. It is noteworthy that both methods were implemented in Python 3.8.11. All computational experiments were conducted on a desktop with a 2.90 GHz computer processing unit and 64 GB memory.

We randomly sampled from the dataset and generated five datasets with different numbers of raw trips. The running times of these two methods in estimating *k* values for these five datasets are displayed in Table 2. It is found that when the number of raw trips increases to a certain threshold, the time spent by the SNN_flow method is incredibly high. For example, when the number of raw trips increased to 50,000, its running time is nearly 7000 s. In contrast, the running time of the two-stage flow clustering method is in an acceptable range. On the other hand, FFBS trips have the distinct characteristics of short distance and local aggregation (Chen et al. 2022b; Zhang et al. 2021), and thus it seems more reasonable to extract the corresponding *k* value for each activity zone than to extract a unique *k* value from the entire study area. In summary, for the FFBS system, the two-stage flow clustering method that divides the study area into multiple activity zones and

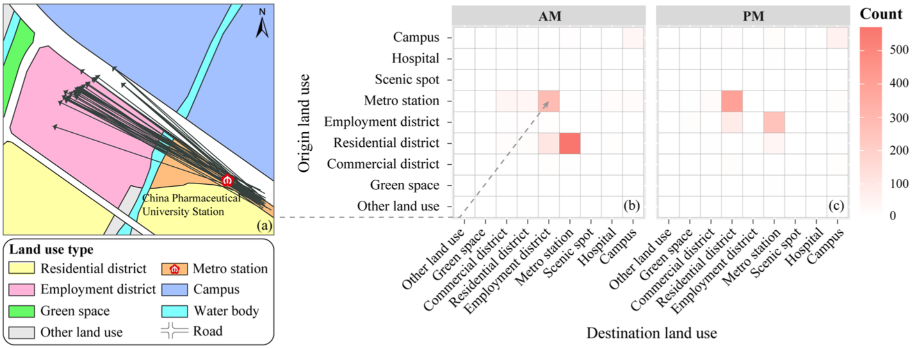


Fig. 6 Matching results of origin–destination land use types of flow clusters. (**a**) a matching case of "metro station → employment district" type flow cluster; matrix of origin–destination land use types for all flow clusters during (**b**) the AM peak and (**c**) the AM peak from September 12 to 14, 2017

then treats them separately is more efficient and reasonable than the SNN_flow method that directly treats the entire study area.

Spatio-temporal patterns of flow clusters

Inference of potential travel purpose

In this subsection, we focus on the spatio-temporal patterns of the flow clusters identified in "Application of the two-stage flow clustering method" section. First of all, the travel purpose of the flow clusters is inferred by combining the land use information (see Fig. 6). More concretely, if the proportion of the origin (destination) points of a flow cluster that falls into a certain land parcel exceeds 50%, we assign the land use type of this parcel to the head (end) of this flow cluster. Note that for an origin (destination) point that does not fall into any of the parcels, we group it into the parcel nearest to it. As shown in Fig. 6a, in the case of this identified flow cluster, most of its origins and destinations fall into parcels of the metro station type and employment district type, respectively. Therefore, it is reasonable to assume that this is a flow cluster for addressing the "last-mile" demand between a metro station and a workplace.

Figure 6b, and c illustrates the matching results of origin–destination land use types for the AM peak and PM peak flow clusters from September 12 to 14, 2017. For the AM peak (as shown in Fig. 6b), those OD flow clusters of the "residential district → metro station" type have the highest share (47.73%). This is followed by the flow clusters of the "metro station → employment district" type (24.12%). Those flow clusters that span directly from residential districts to employment districts also have a share, coming in third (8.98%). The remaining types of flow clusters (e.g., "metro station → commercial district", "campus → campus") are fewer in number during the AM peak, together accounting for less than 20% of the total. Similar to the AM peak, the OD points of the flow clusters during the PM peak are primarily concentrated in three land use types: metro station, residential district, and employment district, but their trip chain order is the opposite of that of the AM peak (see Fig. 6c). To put it another way, the flow clusters during the PM peak are dominated by return-home trips, including "metro station → residential district" (40.14%),

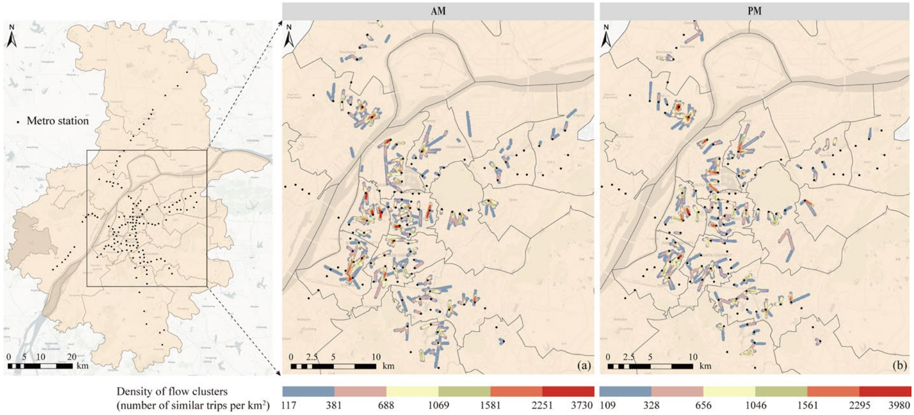


Fig. 7 Spatial distribution of FFBS flow cluster density during **a** the AM peak and **b** the PM peak from September 12 to 14, 2017. Note: the numerical intervals in the legend are divided by the Jenks Natural Breaks Classification tool

"employment district → metro station" (25.65%), and "employment district → residential district" (8.15%).

Overall, the percentage of flow clusters used to meet "first-/last-mile" demand between metro stations and adjacent residences/workplaces is considerable, both during the AM (71.85%) and PM (65.79%) peaks. This implies that FFBS commuting trips with similar spatio-temporal characteristics mostly occur near metro stations. Another interesting finding is that the proportion of flow clusters addressing the "first-/last-mile" between metro stations and adjacent residences (47.73% for AM, 40.14% for PM) was considerably higher than those addressing the "first-/last-mile" between metro stations and adjacent workplaces (24.12% for AM, 25.65% for PM). One reason is that many companies provide commuter shuttles for their employees as an optional way to address the "first-/last-mile" needs of the metro system (Johnson et al. 2015; Kou et al. 2022). The other reason may be that many workplaces (e.g., industrial parks, government agencies) rarely allow FFBS bikes parking inside for management purposes, and the parking spaces available for commuters near the gates are usually limited (Chen and Ye 2021). This somewhat reduces the possibility of choosing FFBS as the connection mode of the metro system.

Analysis of spatio-temporal distribution characteristics

The spatial distribution of FFBS flow clusters during the peak hours from September 12 to 14, 2017 was depicted with the help of the Line Density tool in ArcGIS (see Fig. 7). The length and direction of the flow clusters are characterized by the centerline extracted from the similar trips, and the size of the flow clusters is weighted by the number of similar trips. As shown in Fig. 7, the redder the color of the grid, the higher the number of similar trips occurring at that location. As expected, the density of flow clusters during the AM peak is generally larger than that of flow clusters during the PM peak.

As shown in Fig. 7, metro stations perform a considerable role in the formation of FFBS flow clusters. In order to provide nuanced and appropriate guidance to relevant policies, it is necessary to investigate from which metro stations these flow clusters converge and diverge. First, four types of flow clusters related to metro stations are labeled according to peak hours and trip chain order, namely AM "first-mile" clusters (i.e., "residential district → metro station"), AM "last-mile" clusters (i.e., "metro station → employment district"), PM "first-mile" clusters (i.e., "employment district → metro station"), and PM "last-mile" clusters (i.e., "metro station → residential district"). Then, the number of similar trips from September 12 to 14, 2017 corresponding to these four types of flow clusters is aggregated to each metro station, as shown in Fig. 8.

Some interesting findings can be drawn from Fig. 8. For instance, for the AM "first-mile" clusters, those metro stations that converge a large number of similar trips from residences are principally located outside the city center (Fig. 8a). As for the AM "last-mile" clusters, those metro stations that diverge plenty of similar trips to workplaces are mostly concentrated in the core city (see Fig. 8c). This coincides with the work of Gan et al. (2020) that the residential-oriented metro stations are located in more remote areas than the employment-oriented metro stations concentrated in urban cores. They argued that a major reason is that these relatively remote areas often grew out of under-functioning urban villages, lacking companies and enterprises that can provide a substantial number of job opportunities.

During the PM peak, the spatial distribution of similar trips arriving at the metro stations (Fig. 8b) is similar to that of similar trips departing from the metro stations during the AM peak (Fig. 8c). Figure 8a and d also follow the same trend. Nevertheless, we find a significant difference in the AM "first-mile" clusters (Fig. 8a) and PM "last-mile" clusters (Fig. 8d). Specifically, the residential-based metro stations are located in more remote peripheral areas during the PM peak compared to the AM peak. This may be due to the fact that many commuters will have discretionary activities (e.g., shopping, eating, and entertainment) in their return-home journeys during the PM peak, and inner areas with more commercial land uses appear to be better able to meet these flexible needs (Chen et al. 2022c).

Endpoint distribution characteristics of flow clusters

In this section, two classical tools in spatial analysis, namely standard deviational ellipse (SDE)[4] and calculate distance band from neighbor count (CDBFNC),[5] are adopted to portray the shape and density distribution of flow clusters (Zhu et al. 2016).

We focus on three types of work-related flow clusters during the AM peak (i.e., "residential district → metro station", "metro station → employment distric", and "residential district → employment district") and three types of return-home-related flow clusters during the PM peak (i.e., "employment district → metro station", "metro station → residential district", and "employment district → residential district"), all of which have a high share (see "Inference of potential travel purpose" section for details). It is worth noting that we

[4] SDE is widely utilized to examine the directionality and shape of the spatial points. one, two and three standard deviation(s) (input parameter) indicate that the ellipse cover 68%, 95% and 99% of all points, respectively. The latest help document from ArcGIS 10.8 describes how the tool works. https://desktop.arcgis.com/zh-cn/arcmap/latest/tools/spatial-statistics-toolbox/h-how-directional-distribution-standard-deviationa.htm

[5] CDBFNC reflects the degree of aggregation of spatial points by calculating the average distance from a set of points to the specified *n*th nearest neighbor (*n* is an input parameter). For more details on this tool, please refer to the latest help documentation from ArcGIS 10.8. https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-statistics-toolbox/calculate-distance-band-from-neighbor-count.htm

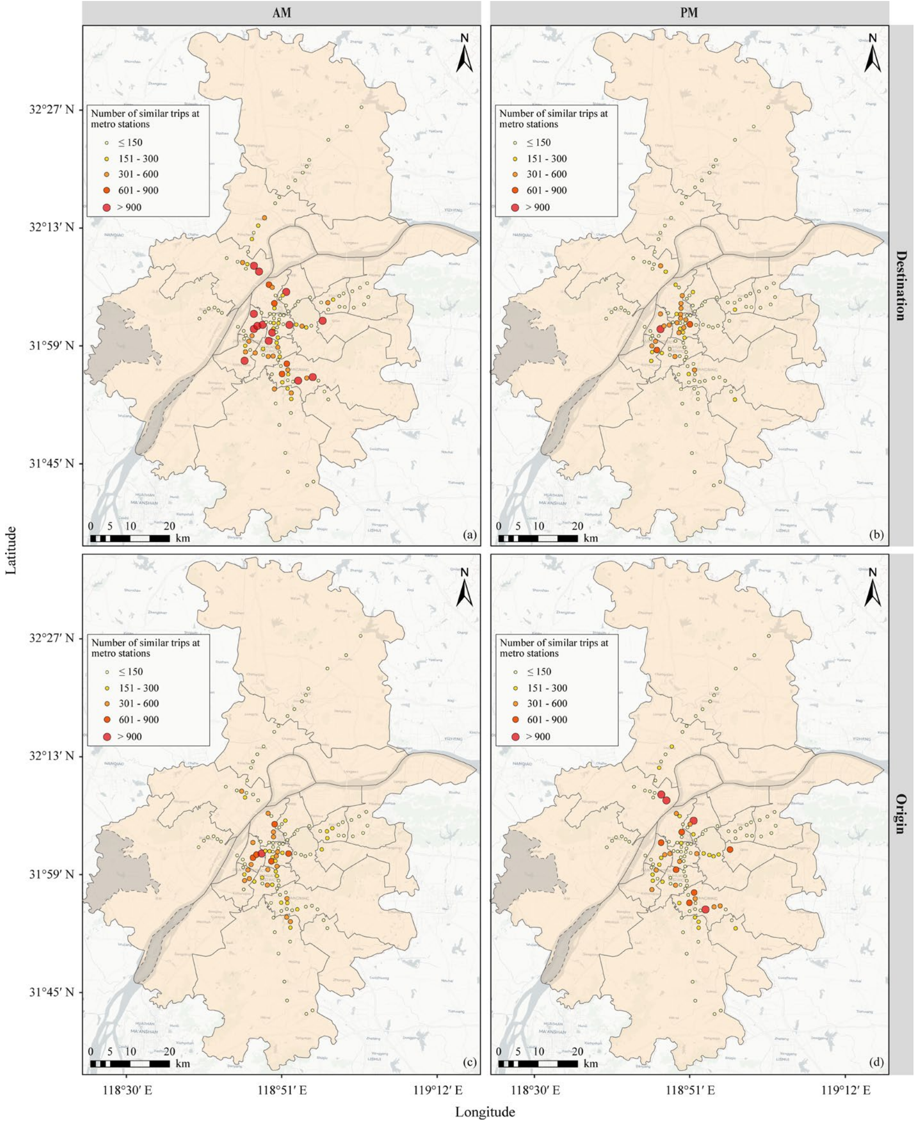


Fig. 8 Aggregation results of the number of similar trips from September 12 to 14, 2017 corresponding to the four types of flow clusters at the metro stations. **a** "first-mile" clusters (residential district → metro station) arriving at metro stations during the AM peak; **b** "first-mile" clusters (employment clusters → metro station) arriving at metro stations during the PM peak; **c** "last-mile" clusters (metro station → employment clusters) departing from metro stations during the AM peak; and **d** "last-mile" clusters (metro station → residential district) departing from metro stations during the PM peak

need to extract the endpoints of these flow clusters as the input of the two tools, as both of them are limited to processing point data (Zhu et al. 2016). For the AM peak from September 12 to 14, 2017, the total number of flow clusters in terms of three work-related types is 945, corresponding to 1890 point clusters (945 × 2, i.e., a flow cluster contains one origin point cluster and one destination point cluster). According to the point cluster land use

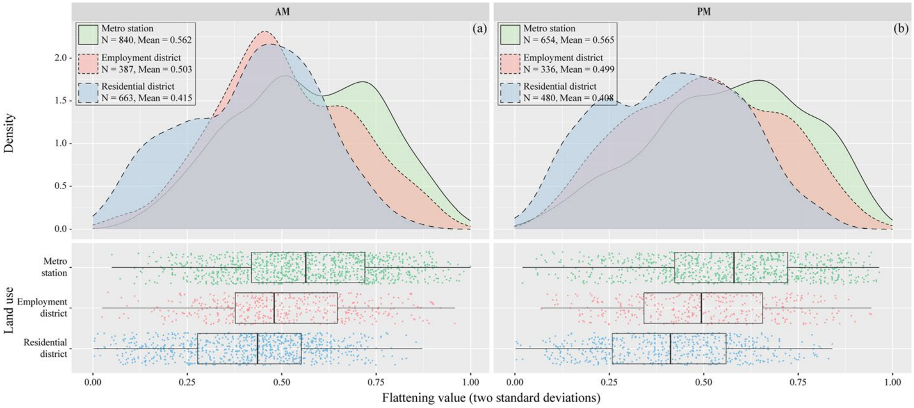


Fig. 9 Distribution of flattening values at the origins and destinations of flow clusters during **a** the AM peak and **b** the PM peak from September 12 to 14, 2017

type, we further divide the 1890 observations into three categories (840 for metro station, 663 for residential district, and 387 for employment district). By analogy, there are 1470 observations in the PM peak from September 12 to 14, 2017 (654 for metro station, 480 for residential district, and 336 for employment district).

We set two standard deviations as the input parameter for SDE, so that the ellipse covers as many points in the point cluster as possible (95%) with less influence from outliers. The tool finally outputs the long and short semi-axes of each ellipse. To be more intuitive, we use the flattening value to depict the shape of the ellipse (point cluster). The flattening value is equal to the ratio of the difference between the long and short semi-axes to the long semi-axes. Its value spans from zero to one, and the closer the value is to one, the flatter the shape of the point cluster. Figure 9 shows the distribution of flattening values for the 1890 (1470) point clusters during the AM (PM) peak of the three weekdays in the form of kernel density and box plots. On the whole, the highest flattening values are found for the metro station point clusters during the AM peak (mean = 0.562), followed by the employment district point clusters (mean = 0.503), and the lowest flattening values for the residential district point clusters (mean = 0.415) (see Fig. 9a). This implies that the shape distributions of employment district and metro station point clusters are inclined to be flatter than that of residential district point clusters. This is perhaps due to the fact that during the peak-hour periods, parking spaces near the entrances of office buildings, especially metro stations, are often in short supply (Zhao and Ong 2021), resulting in many travelers having to park their FFBS bikes along the surrounding sidewalks. In contrast, the shape distribution of residential district point clusters tends to be more circular. The major reason for this may be the existence of many non-gated residential communities in Nanjing (Xinhua Daily 2022), which allows travelers scattered there to park FFBS bikes closer to their exact destination. The distribution of flattening values during the PM peak is basically the same as that during the AM peak, except that it is more uniform (see Fig. 9b). The potential rationale is that, as we discussed in "Application of the two-stage flow clustering method" section, the transaction time and location of return-home trips during the PM peak tend to be less concentrated compared to those of work-related trips during the AM peak (Chen et al. 2022c; Ji et al. 2017).

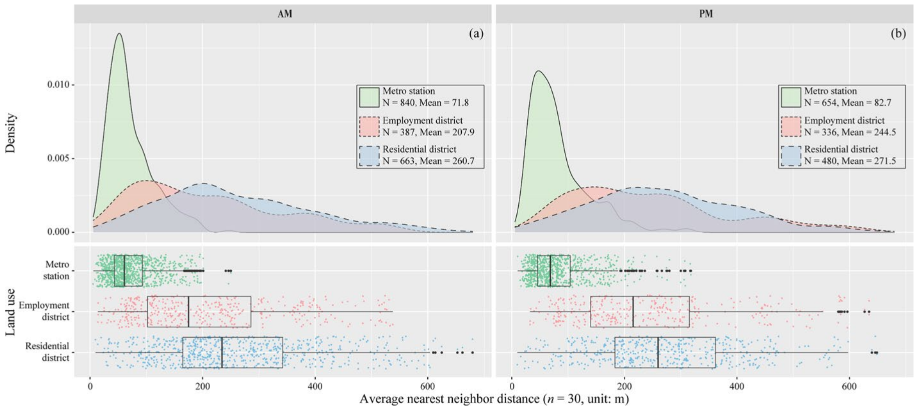


Fig. 10 Distribution of average nearest neighbor distances at the origins and destinations of flow clusters during **a** the AM peak and **b** the PM peak from September 12 to 14, 2017

Since the number of spatial points falling into each point cluster is more than 30 (the recognition threshold for flow clusters is 30, see "Flow clusters identification" section for details), we set the input parameter (n) of CDBFNC tool to 30. The tool finally returns the average distance of all spatial points of the point cluster to the 30th nearest neighbor. Figure 10 illustrates the distribution of average nearest neighbor distance for the 1890 (1470) point clusters during the AM (PM) peak of the three weekdays. As shown in Fig. 10a, the overall distribution of the average nearest neighbor distance of the metro station point clusters is significantly shorter (mean = 71.8 m). This result further indicates that metro stations with relatively limited parking resources often need to carry the operational management pressure incurred by the rapid convergence and divergence of FFBS bikes during peak hours. For the residential district point clusters, we find that their average nearest neighbor distances are longer in general (mean = 260.7 m). In other words, the spatial distribution of points within the residential district point clusters is more dispersed than the other two types of point clusters. One possible reason is that, compared with the compact office buildings and metro stations, many large residential neighborhoods in Nanjing are located in less-developed peripheral areas (Cheng et al. 2022b; Gan et al. 2020), where there is relatively sufficient space for bike parking. The average nearest neighbor distance exhibits essentially the same overall distribution during the AM and PM peaks, except that its distribution is somewhat more uniform during the PM peak (see Fig. 10b). This finding is similar to the distribution of flattening values during the peak hours (Fig. 9), further demonstrating that FFBS trips in the morning are more intensively concentrated.

Conclusions and policy implications

Discovering FFBS similar trips is of great importance for understanding spatio-temporal interactions and human mobility patterns. However, extracting flow clusters consisting of similar trips from large-scale, chaotic journey data remains under-researched. To deal with this issue, this study presents a two-stage flow clustering method, which integrates the Leiden community detection algorithm and the SNN_flow clustering method to efficiently identify flow clusters with arbitrary shapes and inhomogeneous densities. Taking the Nanjing FFBS system as a case study, we demonstrate that the methodological framework helps to significantly improve the efficiency of flow cluster identification.

The results of flow cluster detection (see Table 1) show that although the number of raw trips is higher during the PM peak, the number of flow clusters and corresponding similar trips identified during this period are notably less than those during the AM peak. From the perspective of spatio-temporal patterns, some interesting findings can also be drawn. First, the share of flow clusters used to meet the "first-/last-mile" demand between metro stations and adjacent residences/workplaces is quite high during both the AM (71.85%) and PM (65.79%) peaks. Second, the share of the "first-/last-mile" flow clusters between metro stations and adjacent residences (47.73% for AM, 40.14% for PM) is markedly higher than that of the "first-/last-mile" flow clusters between metro stations and adjacent workplaces (24.12% for AM, 25.65% for PM). Third, the residential-based metro stations in the "first-/last-mile" flow clusters are principally located out of the city center, while the employment-based metro stations in the "first-/last-mile" flow clusters are mostly concentrated in the core city, which is more pronounced during the PM peak. We also investigate the shape and density distribution of the flow clusters. The endpoint distribution results show that metro station point clusters typically have a flatter, linear-like shape distribution than residential point clusters. In addition, we find that spatial points in metro station point clusters are more concentrated, and their density distribution is generally higher than that of other sorts of point clusters.

The spatio-temporal patterns of flow clusters could assist transportation planners and decision-makers in establishing effective policies and regulations to facilitate the rational use of FFBS infrastructure resources. First, extracting flow clusters that concentrate a large number of similar trips could provide nuanced guidance for FFBS operators to allocate resources more efficiently. For instance, during the epidemic prevention and control period, knowing the spatio-temporal dynamics of similar trips could help enhance the efficiency of staff in cleaning and disinfecting bikes (Teixeira and Lopes 2020). Second, metro stations, as the primary departure/arrival places of FFBS similar trips, play a crucial role in addressing the "first-/last-mile" commuting demand of local residents. However, around metro stations, there is often an obstacle in addressing the operational management pressure incurred by the rapid convergence and divergence of FFBS bikes, and the tidal phenomenon of "no bikes to rent or no parking spaces to return" often occurs during the peak hours (Chen et al. 2022a). An effective solution is to predict the similar trips in certain areas in advance according to the past spatio-temporal distribution of flow clusters, thereby reserving a certain amount of bikes and parking spaces for users. Third, compared with the "first-/last-mile" between metro stations and adjacent workplaces, the solution of the "first-/last-mile" between metro stations and adjacent residences is more dependent on the FFBS system. Although FFBS has attracted many users due to its convenience of payment and parking, it is clearly vulnerable to extreme weather such as heavy rain and low temperatures (Shen et al. 2018). In contrast, microcirculation bus – a recently emerging public transportation mode – can provide short-distance travelers with a safer and more comfortable travel service (Du et al. 2019a). Therefore, microcirculation bus service is expected to be the preferred mode of connection to meet the "first-/last-mile" demand between metro stations and adjacent residential neighborhoods under severe meteorological conditions. Fourth, jobs-housing imbalance leads to different FFBS-metro usage patterns during the AM and PM peaks. The differences are critical for designing FFBS fleet rebalancing strategies. For instance, during the AM peak, many metro stations outside the city center may be piled up with a great deal of returned shared bikes, and staff will need to clean them in a timely manner. During the PM peak, the provision of shared bikes near these metro stations becomes insufficient, and it is necessary to allocate more bikes there in advance to address the return-home demand.

The endpoint distribution of flow clusters also provides scholars and decision-makers with some valuable insights and policy implications. Specifically, we inferred from Fig. 9 that the narrow space near metro station entrances results in many users having to park their FFBS bikes along sidewalk racks. This implies that the catchment area with a certain radius size (i.e., acceptable walking distance, e.g., 300 m) generated in the center of a metro station may not be able to accurately capture the FFBS-metro integrated use (Cheng et al. 2022b; Xu et al. 2019). Therefore, it seems more reasonable to construct the catchment area in terms of network walking distance rather than in a straight line. In addition, geo-fenced parking spaces have been put into use in many cities around the world to tackle the disorderly parking of shared bikes (Zhang et al. 2019; Cheng et al. 2022b). It is found that differences in land use types (e.g., metro station, residential district, and employment district) and time of day (e.g., morning peak and evening peak) can bring varying distributions of shape and density in FFBS parking areas. To improve the efficiency of parking utilization, transportation planners may consider flexibility in the size and shape of geo-fenced areas to meet parking needs.

In addition, the journey data of travel modes such as taxis and buses usually have a larger order of magnitude compared to those of FFBS. Many studies have pointed out that traditional flow clustering methods may have some hindrance in efficiently extracting flow clusters from the above travel modes (Liu et al. 2022a; Song et al. 2019). The two-stage flow clustering method proposed in this study may be an effective solution. To be specific, the community detection algorithm is utilized to first divide the entire study area into multiple activity zones with strong intra-connections, thus decomposing a large flow clustering problem into multiple small sub-problems.

Admittedly, there are several limitations to this study. First, we conducted an empirical analysis based on cross-sectional data (three-weekday FFBS journey data), which makes it difficult to trace the evolutionary mechanism of FFBS flow clusters over time. This study will be extended by performing a longitudinal analysis if a longer period of journey data becomes available in the future. Second, this study did not focus on individual-level mobility patterns, which are also important for understanding home-work commuting. Therefore, exploring the similarities and diversities of FFBS flow clusters among different user groups is also a worthwhile research topic. Furthermore, only the flow clusters within each activity zone were extracted in this study. Although the proportion of FFBS trips between different activity zones is small (7.66%), the identification and analysis of their flow clusters can be further taken into account, which is worth of on-going study. Nevertheless, as a first attempt to extract FFBS flow clusters and investigate their spatio-temporal patterns, our findings could provide further insights into human movement patterns and home-work commuting behavior.

Appendix 1

Taking the activity zone 14 as a case study,[6] its flow clusters detected in the AM peak hours from September 12 (Tuesday) to 14 (Thursday), 2017 are displayed in Fig. 11. Figure 11a–c shows the *RKD* plots of journey data on different weekdays. According

[6] All datasets that support the findings of this case study are available on “figshare.com”, with the identifier at the private link: https://doi.org/10.6084/m9.figshare.21324837.v1

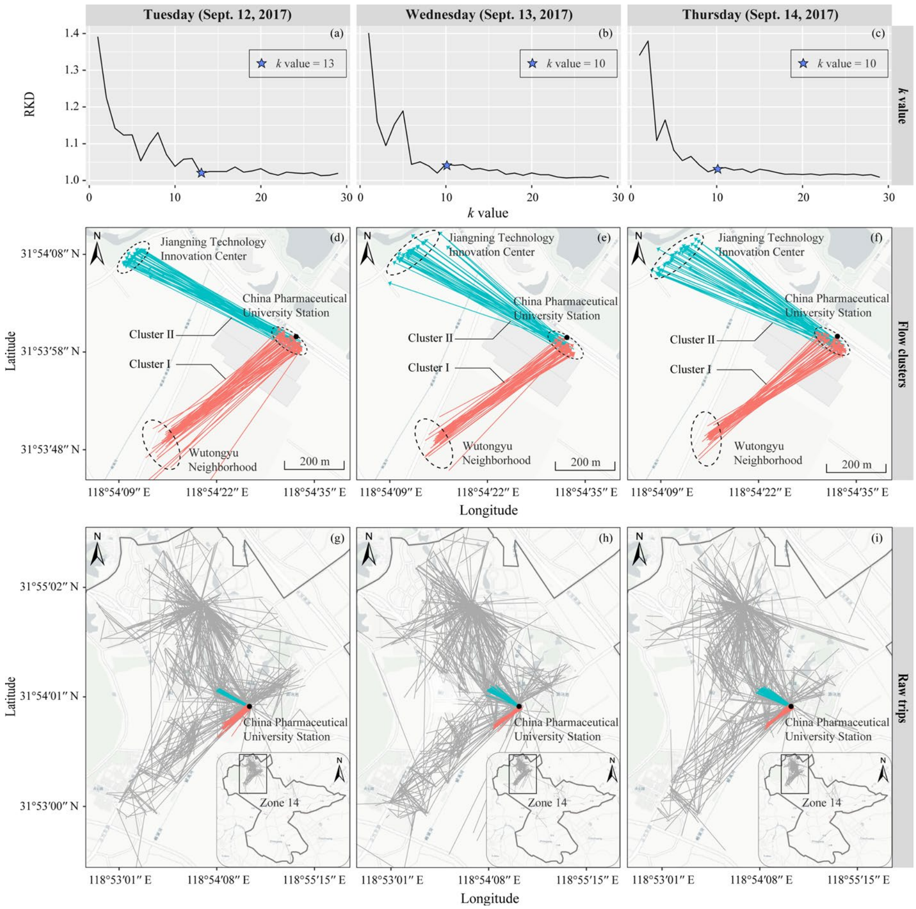


Fig. 11 Flow clusters identification results for activity zone 14 during the AM peak from September 12 to 14, 2017. **a**–**c** *RKD* plots; **d**–**f** spatial distribution of the identified flow clusters; and **g**–**i** spatial distribution of raw trips

to the identification rule provided in "Stage II: Flow cluster identification" section, the values of *k* for these three weekdays were set to 13, 10, and 10, respectively. Using the determined *k* values as the input parameters of SNN_flow, the spatial distribution of the identified flow clusters is depicted in Fig. 11d–f.

During the AM peak on three different weekdays, we identified two flow clusters in activity zone 14 with similar spatial distributions, albeit slightly different in the number of similar trips. Cluster I is composed of many similar trips from Wutongyu Neighborhood (a residential area) to China Pharmaceutical University Station (the terminal metro station of Line 1). Cluster I represents the AM "first-mile" commute pattern between these two locations (see "Analysis of spatio-temporal distribution characteristics" section for details). The spatial distribution of similar trips contained in Cluster II is mainly from China Pharmaceutical University Station to Jiangning Technology Innovation Center (an office building). Cluster II represents the AM "last-mile" commute pattern between these two locations (see "Analysis of spatio-temporal distribution characteristics" section for details). In short, it can be seen from this case that the two-stage flow clustering method allows us to efficiently abstract flow clusters with irregular shapes and uneven densities from large-scale, chaotic FFBS trips (see Fig. 11g–i).

Acknowledgements This research is funded by the Joint Funds of the National Natural Science Foundation of China (Grant No. 52172316, 52372301 and 71901059). The authors wish to thank Beijing Mobike Technology Co., Ltd. for providing the data used in this study.

Author contributions The authors confirm contribution to the paper as follows: Conceptualization: WC, XL, XC; Methodology: WC; Formal analysis and investigation: WC, XL, XC; Writing—original draft preparation: WC; Writing—review and editing: WC, LC, JC. All authors reviewed the results and approved the final version of the manuscript.

Data availability statements The data that support the findings of this study are available from the corresponding author upon request.

Declarations

Competing interests The authors declare no competing interests.

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

References

Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: OPTICS: ordering points to identify the clustering structure. ACM SIGMOD Rec. **28**, 49–60 (1999). https://doi.org/10.1145/304181.304187

Apparicio, P., Abdelmajid, M., Riva, M., Shearmur, R.: Comparing alternative approaches to measuring the geographical accessibility of urban health services: distance types and aggregation-error issues. Int. J Health Geogr. **7**, 1–14 (2008). https://doi.org/10.1186/1476-072X-7-7

Arenas, A., Fernandez, A., Gomez, S.: Analysis of the structure of complex networks at different resolution levels. New J. Phys. **10**(5), 053039 (2008). https://doi.org/10.1088/1367-2630/10/5/053039

Besse, P.C., Guillouet, B., Loubes, J.M., Royer, F.: Review and perspective for distance-based clustering of vehicle trajectories. IEEE Trans. Intell. Transp. Syst. **17**, 3306–3317 (2016). https://doi.org/10.1109/TITS.2016.2547641

Chang, X., Wu, J., Sun, H., de Almeida Correia, G.H., Chen, J.: Relocating operational and damaged bikes in free-floating systems: a data-driven modeling framework for level of service enhancement. Transp. Res. Part A Policy Pract. **153**, 235–260 (2021). https://doi.org/10.1016/j.tra.2021.09.010

Chen, D.: Free-floating bike-sharing green relocation problem considering greenhouse gas emissions. Transp. Saf. Environ. **3**, 132–151 (2021). https://doi.org/10.1093/tse/tdab001

Chen, E., Ye, Z.: Identifying the nonlinear relationship between free-floating bike sharing usage and built environment. J. Clean. Prod. **280**, 124281 (2021). https://doi.org/10.1016/j.jclepro.2020.124281

Chen, W., Chen, X., Chen, J., Cheng, L.: What factors influence ridership of station-based bike sharing and free-floating bike sharing at rail transit stations? Int. J. Sustain. Transp. **16**, 357–373 (2022a). https://doi.org/10.1080/15568318.2021.1872121

Chen, W., Chen, X., Cheng, L., Liu, X., Chen, J.: Delineating borders of urban activity zones with free-floating bike sharing spatial interaction network. J. Transp. Geogr. **104**, 103442 (2022b). https://doi.org/10.1016/j.jtrangeo.2022.103442

Chen, W., Liu, X., Chen, X., Cheng, L., Wang, K., Chen, J.: Exploring year-to-year changes in station-based bike sharing commuter behaviors with smart card data. Travel Behav. Soc. **28**, 75–89 (2022c). https://doi.org/10.1016/j.tbs.2022.02.005

Cheng, L., Yang, J., Chen, X., Cao, M., Zhou, H., Sun, Y.: How could the station-based bike sharing system and the free-floating bike sharing system be coordinated? J. Transp. Geogr. **89**, 102896 (2020a). https://doi.org/10.1016/j.jtrangeo.2020.102896

Cheng, L., Yang, M., De Vos, J., Witlox, F.: Examining geographical accessibility to multi-tier hospital care services for the elderly: a focus on spatial equity. J. Transp. Health. **19**, 100926 (2020b). https://doi.org/10.1016/j.jth.2020.100926

Cheng, L., Jin, T., Wang, K., Lee, Y., Witlox, F.: Promoting the integrated use of bikeshare and metro: a focus on the nonlinearity of built environment effects. Multimodal Transp. **1**(1), 100004 (2022a). https://doi.org/10.1016/j.multra.2022.100004

Cheng, L., Wang, K., De Vos, J., Huang, J., Witlox, F.: Exploring non-linear built environment effects on the integration of free-floating bike-share and urban rail transport: a quantile regression approach. Transp. Res. Part A Policy Pract. **162**, 175–187 (2022b). https://doi.org/10.1016/j.tra.2022.05.022

Cheng, L., Huang, J., Jin, T., Chen, W., Li, A., Witlox, F.: Comparison of station-based and free-floating bikeshare systems as feeder modes to the metro. J. Transp. Geogr. **107**, 103545 (2023). https://doi.org/10.1016/j.jtrangeo.2023.103545

Du, B., Qiao, Y., Zhao, J., Sun, L., Lv, W., Huang, R.: Urban micro-circulation bus planning based on temporal and spatial travel demand, in: 2019a IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 981–988. IEEE (2019a). https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00193

Du, Y., Deng, F., Liao, F.: A model framework for discovering the spatio-temporal usage patterns of public free-floating bike-sharing system. Transp. Res. Part C Emerg. Technol. **103**, 39–55 (2019b). https://doi.org/10.1016/j.trc.2019.04.006

Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, pp. 226–231 (1996)

Gallego, C.E.V., Comendador, V.F.G., Nieto, F.J.S., Martinez, M.G.: Discussion on density-based clustering methods applied for automated identification of airspace flows. In: 2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), pp. 1–10. IEEE (2018). https://doi.org/10.1109/DASC.2018.8569219

Gan, Z., Yang, M., Feng, T., Timmermans, H.: Understanding urban mobility patterns from a spatiotemporal perspective: daily ridership profiles of metro stations. Transportation **47**, 315–336 (2020). https://doi.org/10.1007/s11116-018-9885-4

Gao, Y., Li, T., Wang, S., Jeong, M.H., Soltani, K.: A multidimensional spatial scan statistics approach to movement pattern comparison. Int. J. Geogr. Inf. Sci. **32**(7), 1304–1325 (2018). https://doi.org/10.1080/13658816.2018.1426859

Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. u.s.a. **99**, 7821–7826 (2002). https://doi.org/10.1073/pnas.122653799

Gu, T., Kim, I., Currie, G.: To be or not to be dockless: Empirical analysis of dockless bikeshare development in China. Transp. Res. Part A Policy Pract. **119**, 122–147 (2019). https://doi.org/10.1016/j.tra.2018.11.007

Guo, Y., He, S.Y.: Built environment effects on the integration of dockless bike-sharing and the metro. Transp. Res. D Transp. Environ. **83**, 102335 (2020). https://doi.org/10.1016/j.trd.2020.102335

Guo, X., Xu, Z., Zhang, J., Lu, J., Zhang, H.: An OD flow clustering method based on vector constraints: a case study for Beijing taxi origin-destination data. ISPRS Int. J. Geo-Inf. **9**, 128 (2020). https://doi.org/10.3390/ijgi9020128

Hirsch, J.A., Stratton-Rayner, J., Winters, M., Stehlin, J., Hosford, K., Mooney, S.J.: Roadmap for free-floating bikeshare research and practice in North America. Transp. Rev. **39**, 706–732 (2019). https://doi.org/10.1080/01441647.2019.1649318

Hua, M., Chen, X., Zheng, S., Cheng, L., Chen, J.: Estimating the parking demand of free-floating bike sharing: A journey-data-based study of Nanjing, China. J. Clean. Prod. **244**, 118764 (2020). https://doi.org/10.1016/j.jclepro.2019.118764

Ji, Y., Fan, Y., Ermagun, A., Cao, X., Wang, W., Das, K.: Public bicycle as a feeder mode to rail transit in China: the role of gender, age, income, trip purpose, and bicycle theft experience. Int. J. Sustain. Transp. **11**, 308–317 (2017). https://doi.org/10.1080/15568318.2016.1253802

Jin, M., Gong, L., Cao, Y., Zhang, P., Gong, Y., Liu, Y.: Identifying borders of activity spaces and quantifying border effects on intra-urban travel through spatial interaction network. Comput. Environ. Urban Syst. **87**, 101625 (2021). https://doi.org/10.1016/j.compenvurbsys.2021.101625

Johnson, G., Scher, H., Wittmann, T.: Designing shuttle connections to commuter rail with census origin and destination data. Transp. Res. Rec. **2534**, 84–91 (2015). https://doi.org/10.3141/2534

Kou, W., Wang, J., Liu, Y., Li, C.: Last-mile shuttle planning for improving bus commuters' travel time reliability: a case study of Beijing. J. Adv. Transp. **2022**, 5117488 (2022). https://doi.org/10.1155/2022/5117488

Lei, D., Chen, X., Cheng, L., Zhang, L., Ukkusuri, S.V., Witlox, F.: Inferring temporal motifs for travel pattern analysis using large scale smart card data. Transp. Res. Part C Emerg. Technol. **120**, 102810 (2020). https://doi.org/10.1016/j.trc.2020.102810

Link, C., Strasser, C., Hinterreiter, M.: Free-floating bikesharing in Vienna–A user behaviour analysis. Transp. Res. Part A Policy Pract. **135**, 168–182 (2020). https://doi.org/10.1016/j.tra.2020.02.020

Liu, Q., Yang, J., Deng, M., Song, C., Liu, W.: SNN_flow: a shared nearest-neighbor-based clustering method for inhomogeneous origin-destination flows. Int. J. Geogr. Inf. Sci. **36**(2), 253–279 (2022a). https://doi.org/10.1080/13658816.2021.1899184

Liu, Y., Tong, D., Liu, X.: Measuring spatial autocorrelation of vectors. Geogr. Anal. **47**, 300–319 (2015). https://doi.org/10.1111/gean.12069

Liu, W., Liu, Q., Yang, J., Deng, M.: A network-constrained clustering method for bivariate origin-destination movement data. Int. J. Geogr. Inf. Sci. **37**(4), 767–787 (2022b). https://doi.org/10.1080/13658816.2022.2137879

Ma, X., Zhang, X., Li, X., Wang, X., Zhao, X.: Impacts of free-floating bikesharing system on public transit ridership. Transp. Res. D Transp. Environ. **76**, 100–110 (2019). https://doi.org/10.1016/j.trd.2019.09.014

Orvin, M.M., Fatmi, M.R.: Why individuals choose dockless bike sharing services? Travel Behav. Soc. **22**, 199–206 (2021). https://doi.org/10.1016/j.tbs.2020.10.001

Páez, A., Anjum, Z., Dickson-Anderson, S.E., Schuster-Wallace, C.J., Ramos, B.M., Higgins, C.D.: Comparing distance, time, and metabolic energy cost functions for walking accessibility in infrastructure-poor regions. J. Transp. Geogr. **82**, 102564 (2020). https://doi.org/10.1016/j.jtrangeo.2019.102564

Pan, G., Qi, G., Wu, Z., Zhang, D., Li, S.: Land-use classification using taxi GPS traces. IEEE Trans. Intell. Transp. Syst. **14**, 113–123 (2012). https://doi.org/10.1109/TITS.2012.2209201

Pei, T.: A nonparametric index for determining the numbers of events in clusters. Math. Geosci. **43**, 345–362 (2011). https://doi.org/10.1007/s11004-011-9325-x

Pei, T., Gao, J., Ma, T., Zhou, C.: Multi-scale decomposition of point process data. GeoInformatica **16**, 625–652 (2012). https://doi.org/10.1007/s10707-012-0165-8

Peters, L., MacKenzie, D.: The death and rebirth of bikesharing in Seattle: Implications for policy and system design. Transp. Res. Part A Policy Pract. **130**, 208–226 (2019). https://doi.org/10.1016/j.tra.2019.09.012

Reddy, K.S.S., Bindu, C.S.: A review on density-based clustering algorithms for big data analysis. In: 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 123–130, IEEE (2017). https://doi.org/10.1109/I-SMAC.2017.8058322

Shen, Y., Zhang, X., Zhao, J.: Understanding the usage of dockless bike sharing in Singapore. Int. J. Sustain. Transp. **12**, 686–700 (2018). https://doi.org/10.1080/15568318.2018.1429696

Shu, H., Pei, T., Song, C., Chen, X., Guo, S., Liu, Y., Chen, J., Wang, X., Zhou, C.: L-function of geographical flows. Int. J. Geogr. Inf. Sci. **35**, 689–716 (2021). https://doi.org/10.1080/13658816.2020.1749277

Silva, I., Assunçao, R., Costa, M.: Power of the sequential Monte Carlo test. Seq. Anal. **28**, 163–174 (2009). https://doi.org/10.1080/07474940902816601

Song, C., Pei, T., Ma, T., Du, Y., Shu, H., Guo, S., Fan, Z.: Detecting arbitrarily shaped clusters in origin-destination flows using ant colony optimization. Int. J. Geogr. Inf. Sci. **33**, 134–154 (2019). https://doi.org/10.1080/13658816.2018.1516287

Tao, R., Thill, J.C.: Spatial cluster detection in spatial flow data. Geogr. Anal. **48**, 355–372 (2016). https://doi.org/10.1111/gean.12100

Teixeira, J.F., Lopes, M.: The link between bike sharing and subway use during the COVID-19 pandemic: the case-study of New York's Citi bike. Transp. Res. Interdiscip. Perspect. **6**, 100166 (2020). https://doi.org/10.1016/j.trip.2020.100166

Traag, V.A., Waltman, L., Van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. **9**, 1–12 (2019). https://doi.org/10.1038/s41598-019-41695-z

White, C.E., Bernstein, D., Kornhauser, A.L.: Some map matching algorithms for personal navigation assistants. Transp. Res. Part C Emerg. Technol. **8**, 91–108 (2000). https://doi.org/10.1016/S0968-090X(00)00026-7

Wood, J., Dykes, J., Slingsby, A.: Visualisation of origins, destinations and flows with OD maps. Cartogr. J. **47**, 117–129 (2010). https://doi.org/10.1179/000870410X12658023467367

Xiang, Q., Wu, Q.: Tree-based and optimum cut-based origin-destination flow clustering. ISPRS Int. J. Geo-Inf. **8**, 477 (2019). https://doi.org/10.3390/ijgi8110477

Xinhua Daily: More than 2300 non-gated residential neighborhoods in Nanjing have achieved full coverage of high-standard basic management. http://house.china.com.cn/2115807.htm (2022). Accessed 30 Sept 2022

Xu, X.: The road network data obtained from this processing can be directly used in traffic models. https://www.sohu.com/a/397982966_650480 (2020). Accessed 25 Aug 2022

Xu, C., Ji, J., Liu, P.: The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. Transp. Res. Part C Emerg. Technol. **95**, 47–60 (2018). https://doi.org/10.1016/j.trc.2018.07.013

Xu, D., Bian, Y., Rong, J., Wang, J., Yin, B.: Study on clustering of free-floating bike-sharing parking time series in Beijing subway stations. Sustainability **11**, 5439 (2019). https://doi.org/10.3390/su11195439

Yamada, I., Thill, J.C.: Local indicators of network-constrained clusters in spatial patterns represented by a link attribute. Ann. Assoc. Am. Geogr. **100**, 269–285 (2010). https://doi.org/10.1080/00045600903550337

Yao, X., Zhu, D., Gao, Y., Wu, L., Zhang, P., Liu, Y.: A stepwise spatio-temporal flow clustering method for discovering mobility trends. IEEE Access. **6**, 44666–44675 (2018). https://doi.org/10.1109/ACCESS.2018.2864662

Zhang, J., Meng, M.: Bike allocation strategies in a competitive dockless bike sharing market. J. Cleaner Prod. **233**, 869–879 (2019). https://doi.org/10.1016/j.jclepro.2019.06.070

Zhang, Y., Lin, D., Mi, Z.: Electric fence planning for dockless bike-sharing services. J. Cleaner Prod. **206**, 383–393 (2019). https://doi.org/10.1016/j.jclepro.2018.09.215

Zhang, X., Shen, Y., Zhao, J.: The mobility pattern of dockless bike sharing: a four-month study in Singapore. Transp. Res. D Transp. Environ. **98**, 102961 (2021). https://doi.org/10.1016/j.trd.2021.102961

Zhao, D., Ong, G.P.: Geo-fenced parking spaces identification for free-floating bicycle sharing system. Transp. Res. Part A Policy Pract. **148**, 49–63 (2021). https://doi.org/10.1016/j.tra.2021.03.007

Zhao, J., Wang, J., Deng, W.: Exploring bikesharing travel time and trip chain by gender and day of the week. Transp. Res. Part C Emerg. Technol. **58**, 251–264 (2015). https://doi.org/10.1016/j.trc.2015.01.030

Zheng, Z., Chen, Y., Zhu, D., Sun, H., Wu, J., Pan, X., Li, D.: Extreme unbalanced mobility network in bike sharing system. Physica a. **563**, 125444 (2021). https://doi.org/10.1016/j.physa.2020.125444

Zhu, X., Guo, D.: Mapping large spatial flow data with hierarchical clustering. Trans. GIS. **18**, 421–435 (2014). https://doi.org/10.1111/tgis.12100

Zhu, R., Hu, Y., Janowicz, K., McKenzie, G.: Spatial signatures for geographic feature types: examining gazetteer ontologies using spatial statistics. Trans. GIS **20**, 333–355 (2016). https://doi.org/10.1111/tgis.12232

Zhu, X., Guo, D., Koylu, C., Chen, C.: Density-based multi-scale flow mapping and generalization. Comput. Environ. Urban Syst. **77**, 101359 (2019). https://doi.org/10.1016/j.compenvurbsys.2019.101359

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Wendong Chen is a Ph.D. candidate in the School of Transportation, Southeast University. He received his master's degree from the School of Transportation of Southeast University and his bachelor's degree from the School of Highway of Chang'an University. His research interests are in shared mobility, travel behaviour analysis, and transport and land use integration.

Xize Liu is a Ph.D. candidate in the School of Transportation, Southeast University. He received his master's degree from the School of Transportation of Southeast University and his bachelor's degree from the School of Transportation Science and Engineering of Harbin Institute of Technology. His research interests are in travel behaviour analysis and multimodal transport.

Xuewu Chen received the Ph.D. degree in transportation engineering from Southeast University, China. She is currently a professor with the School of Transportation, Southeast University. Her research interests include urban transportation, multimodal transportation, shared mobility, and travel behaviour analysis.

Long Cheng received the B.S. degree in transport & traffic from Southeast University, Nanjing, China in 2011 and the Ph.D. degree in transport engineering from Southeast University, Nanjing, China in 2016. He is currently an associate professor at the School of Transportation of Southeast University. His research interests include multimodal transport, shared mobility, travel behaviour analysis, and transport and land use integration.

Jingxu Chen received the Ph.D. degree in transportation engineering from Southeast University, China, in 2018. He is currently an associate professor with the School of Transportation, Southeast University. His research interests include multimodal transport, transportation system optimization, and modeling and simulation of real-time transit control systems.