The Journal of Transport and Land Use jtlu.org
Vol. 17 No. 1 [2024] pp. 143–161

JTLU

Exploring spatial association between residential and commercial urban spaces: A machine learning approach using taxi trajectory data

Lei Zhou
School of Internet of Things
Nanjing University of Posts and Telecommunications
zhoulei@njupt.edu.cn

Chen Wang
School of Internet of Things
Nanjing University of Posts and Telecommunications
1021173514@njupt.edu.cn

Weiye Xiao (corresponding author)
Nanjing Institute of Geography and Limnology
Chinese Academy of Sciences
ywxiao@niglas.ac.cn

Haoran Wang
School of Geographic Science
Nanjing Normal University
191302046@njnu.edu.cn

Abstract: Human mobility datasets, such as traffic flow data, reveal the connections between urban spaces. A novel framework is proposed to explore the spatial association between urban commercial and residential spaces via consumption travel flows in Shanghai. A social network analysis and a community detection method are employed using taxi trajectory data during the daytime to validate the framework. The machine learning-based approach, such as the community detection method, can overcome the limitation regarding spatial uncertainty and spatial effects. The empirical findings suggest that people's commercial activities are sensitive to the power of accessible commercial centers and travel distances. The high-level commercial centers would contribute to the monocentric structure in the outer urban region based on consumption flows. In the central urban region, increasing the number of high-level commercial centers and making the powers of commercial centers hierarchical can contribute to a polycentric mobility pattern of people's consumption. This research contributes to the literature by providing a novel framework to model, analyze and visualize people's mobility based on the trajectory big data, which is promising in future urban research.

Article history:
Received: January 1, 2023
Received in revised form: August 3, 2023
Accepted: January 9, 2024
Available online: February 29, 2024

1 Introduction

Urban spaces have diverse functions, and they are interconnected through human activities such as daily commuting, consumption behavior, and the exchange of goods, capital, and even information within the city (Burger et al., 2014; Gao et al., 2013; Safira & Chikaraishi, 2022; Wu et al., 2021; Zhong et al., 2014). Commercial and residential spaces carry most people's daily activities other than work, which are the most closely

https://doi.org/10.5198/jtlu.2024.1800
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

linked physical spaces in cities (Zhou, Wang, et al., 2023). Therefore, the spatial association between commercial and residential spaces is a critical indicator of urban vitality. In this paper, the "spatial association" between commercial and residential spaces refers to how the consumption behaviors of residents connect commercial and residential spaces, which could reflect distributional dependency, locational proximity, and spatial accessibility (Nicoletti, Sirenko & Verma, 2023; Zhou, Wang, et al., 2023). Much attention has been paid to the spatial association between commercial and residential spaces in urban geography and planning research (Smith et al., 2022; Zhou, Wang, et al., 2023).

In China, the large metropolis has undergone population booming and urban land expansion, with a transformation of urban structure towards polycentricity, a promising strategy to avoid some urban maladies such as traffic congestion and rising inequality. Such an urban transformation also relocates commercial and residential spaces via decentralization (Zhang et al., 2022). Existing literature suggests that decentralization might contribute to the physical disconnection between commercial and residential spaces (Meltzer & Schuetz, 2012; Zhou et al., 2022). Moreover, the rapidly developing e-retail and online consumption has essentially influenced the attributes, formats, and scales of urban commercial space (Zhou et al., 2022), weakening the significance of the physical proximity to commercial space. Therefore, a new urban commercial and residential space nexus is being created, requiring more work on the spatial association pattern between residential and commercial space in China.

Human activities like people's travel flows could reflect the association between urban spaces (Kolodin Ferrari, et al., 2021; Li et al., 2022; Tian et al., 2019; Wu et al., 2021). Current research has paid much attention to residential space, and literature has documented the association between residential space and others including employment space (Xiao, Wei et al., 2021; Zhang et al., 2017), education space (Boussauw et al., 2014; Cheng et al., 2017; Dignum et al., 2022; Sun et al., 2021), healthcare space (Li et al., 2022) and commercial space (Anguera-Torrell & Cerdan, 2021; Fan et al., 2011; Zhou et al., 2022), etc. This research has been carried out in two aspects: one is a static perspective, referring to the spatial correlation of the distribution of morphological elements in a certain region (Sun et al., 2021; Xiao, Wei et al., 2021; Zhou et al., 2022); the other is the functional spatial association between urban spaces via human activities, contributing to a dynamic perspective (Cheng et al., 2017; Zhang et al., 2017). Taking the jobs-housing relationship as an example, scholars traditionally applied the jobs-housing spatial mismatch index to analyze the association between residential and working space (Xiao et al., 2023). However, empirical studies suggest that the dynamic perspective referring to people's mobility can better capture the spatial association between urban spaces (Cheng et al., 2017; Zhang et al., 2017). The studies on urban commuting suggest that commuting mode and time have become a critical concern for quantitatively describing the jobs-housing relationship (Jeddi Yeganeh et al., 2018; Wang & Chai, 2009; Xiao, Wei et al., 2021). The analysis of people's commuting behavior could paint a nuanced and accurate picture of the jobs-housing relationship; however, it requires detailed data sets. Thus, plenty of studies were carried out based on smartcard commuting data and cellular signaling data (Xiong et al., 2021; Zhang et al., 2017; Zheng et al., 2021).

Regarding the spatial association between commercial and residential spaces, current research primarily focuses on mixed land use, such as the ratio between facilities and houses (Meltzer & Schuetz, 2012; Zhou et al., 2022). In urban reality, urban flows such as consumption travel flow could better reflect the association between commercial and residential spaces (Fan et al., 2011). However, the limited data on people's travel flows are the barriers to analyzing the association between commercial and residential spaces at

an intra-urban level, leaving such spatial association and the underlying mechanism a gap in research that remains to be filled.

In the big data era, Location Based Services (LBS) and the proliferation of personal real-time trajectory data provide new channels for exploring human mobility and consuming behavior (Chen & Yeh, 2022). Trajectory data in various forms is increasingly available, such as volunteer positioning data, smart card public transport data (Liu et al., 2016; Xiao & Wei, 2023), public bicycle rental data (Song et al., 2021) and cellular signaling data (Chen & Yeh, 2022). Trajectory data contains location and time which can fully reflect spatial temporal dynamics of human mobility. Trajectory data has been widely used in urban research, such as detecting functional zones and discovering hidden patterns (Siła-Nowicka et al., 2016; Yuan et al., 2014; Zhong et al., 2014). Compared with the traditional data set, the trajectory data could significantly reduce the spatial uncertainty created by the analytical units and depict the accurate spatial association between urban spaces (Dignum et al., 2022; Cheng et al., 2017; Fan et al., 2011; Sun et al., 2021; Xiao, Wei et al., 2021).

In this research, a taxi trajectory data set of Shanghai for one week in 2015 is employed to identify the spatial association between residential and commercial space. We focus on daytime travel between 10:00 am and 4:00 pm to exclude commuting trips. In recent years, the popularity of online taxi booking platforms, such as Didi Taxi, benefits people's daily travel and allows researchers to access residents' travel preferences and consumption behavior. So far, taxi trajectory data has brought ample research outcomes in various fields (Gao et al., 2017; Li et al., 2016; Li et al., 2021; Wu et al., 2020). For example, the frequencies of pick-ups and drop-offs can reveal residents' travel patterns and identify urban mixed land use and functional zones (Hu et al., 2021; Xu et al., 2022). The O-D analysis of taxi trajectory data can help explore the spatial interaction between urban spaces (Liu et al., 2012). Utilizing taxi trajectory data in Shanghai, Li et al. (2021) explored the interaction among absolute space, relative space, and relational space using structural equation modeling. Thus, taxi trajectory data has been proven to be a powerful data source in the research on spatial interaction between urban spaces.

We propose a novel framework to explore the spatial association patterns between commercial and residential urban spaces at the township level (Figure 1). First, people's consuming trips are extracted from taxi trajectory data, including the trip origins, destinations, and time. A weighted directional network for people's consumption flow is constructed, whose nodes and association intensity are analyzed using a social network analysis method. Then, the community detection method is adopted to divide the commercial and residential spatial association clusters into communities via spatial interaction. Finally, we will examine how different commercial centers contribute to different spatial association patterns. In this research, we focus on the disparities in the spatial association patterns between commercial and residential urban spaces, which was not clarified in previous studies like Li et al. (2021). The analysis results are expected to help detect urban structures and better understand the impact of commercial centers on commuting based on mobility data in large metropolitans.

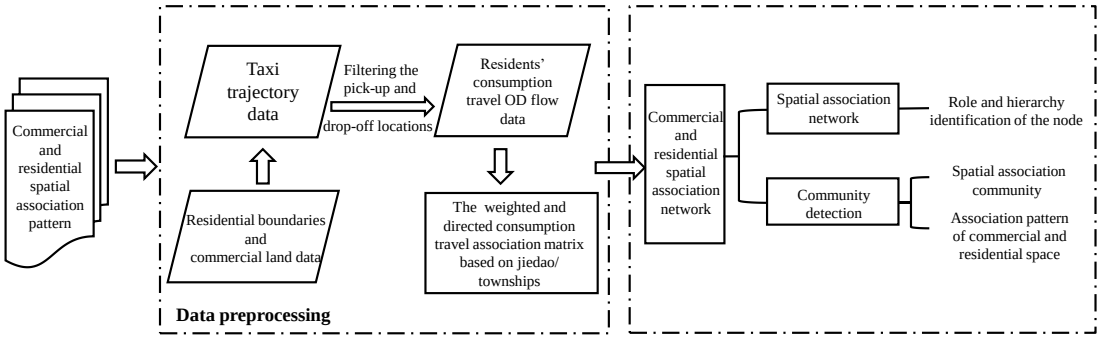


Figure 1. Research framework

2 Methodology

2.1 Study area

Shanghai is an emerging global city in China with various commercial commodities and services. Currently, Shanghai's urban structure is transforming from a monocentric to a polycentric, and the commercial centers are also decentralizing (Xiao, Wang, et al., 2021). There are 205 commercial centers in Shanghai (Figure 2), which are classified into four categories, following Zhou, Xiao et al.'s methodology based on consumption volume (2023). The first-level commercial centers contain all the 21 municipal-level commercial centers in Shanghai. The second-level commercial centers contain all the 48 district-level commercial centers. The third-level commercial centers contain the 70 large and middle scale community-level commercial centers. The fourth-level commercial centers contain the 66 small scale community-level commercial centers in Shanghai.

Meanwhile, the high housing price in Shanghai has intensified urban decentralization, contributing to the physical disconnection between commercial and residential spaces. Transportation flow is critical in connecting residential and commercial spaces, as a taxi is one of the most common ones. According to a Comprehensive Transportation Annual Report of Shanghai in 2016, there were about 50,000 taxis in Shanghai, with taxi trips accounting for about 6% of the total trips. Therefore, taxi trajectory data represents the spatial association between commercial and residential spaces in Shanghai. We focus on taxi trips in 16 administrative districts of Shanghai, including 215 subdistricts (jiedao/township) (Figure 2). The expressway and highway in Shanghai, i.e., the inner ring, middle ring, outer ring, and suburban ring, divide the city into a ring shape spatial structure, including the central urban regions within the inner ring road, the ring road regions between the inner ring road and outer ring road, and the outer urban regions out of the outer ring road.

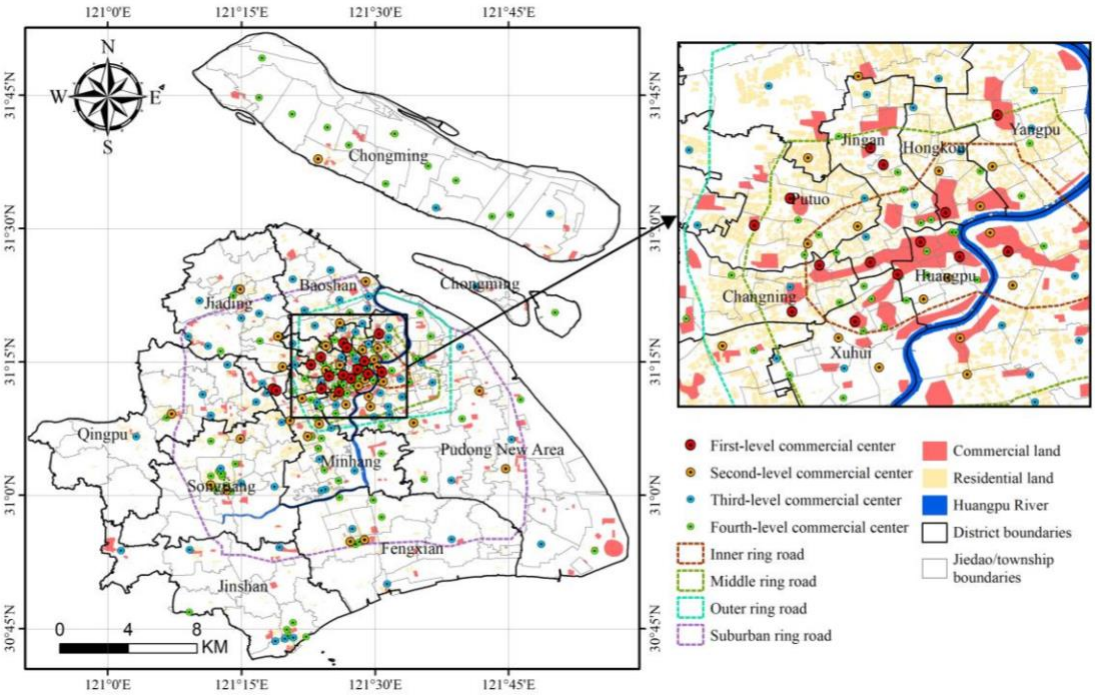


Figure 2. Study area

2.2 Data preprocessing

A dataset containing about 13,000 regular licensed taxis with GPS records collected from the major cab companies in Shanghai was applied. We selected the taxi trajectory data for one week from April 1 to 7, 2015, containing information such as taxi ID, pick-up and drop-off time, coordinates, etc. (Table 1). 2,578,992 valid trips were obtained after eliminating the data with incomplete information, including trip distances greater than 100 km and less than 100 m, trip time longer than 6 hours and shorter than 1 minute, travel speed greater than 33 m/s, and trip end outside of Shanghai. Detailed preprocessing steps of the data follow Liu et al.'s work (2012). Each trip record contains the origins and destinations of each consumption trip. The boundary data of residences (from amap.com) and commercial land-use data (from the land-use map of Shanghai in 2016) were applied to identify the residences and commercial spaces (Figure 2). To avoid the impact of residents' commuting travel, we selected the taxi trips whose origins are less than 70 meters from the residences and whose destinations are within 200 meters of the commercial land from 10:00 to 16:00 as the O-D data of residents' consumption travels. A total number of 141,236 consumption trips were obtained.

Table 1. Data format and examples of the pick-up and drop-off labels for taxi

Taxi ID	Pick-up time	Pick-up coordinates	Drop-off time	Drop-off coordinates
13	2015/4/4 13:22	121.3752°E,31.3244°N	2015/4/4 13:49	121.37761°E,31.3128°N
281	2015/4/4 13:57	121.4519°E,31.3300°N	2015/4/4 14:17	121.4619°E,31.2990°N
12958	2015/4/4 13:51	121.3937°E,31.1728°N	2015/4/4 13:57	121.3887° E,31.1982° N
27588	2015/4/4 14:08	121.4710° E,31.2347° N	2015/4/4 14:25	121.4632° E,31.2755° N
…	…	…	…	…

2.3 Methodology

2.3.1 Social network analysis

The consumption trips contribute to a spatial association network, representing numerous nodes and connections separately. Social network analysis is applied to analyze the association density and direction of commercial and residential space, as it can effectively identify the spatial organization and the hierarchy of the spatial association network (Song et al., 2021). A weighted and directed network that can be denoted as $G \coloneqq (N, L, W)$ was constructed (Liu et al., 2015). $N \coloneqq (n_1, n_2, n_3, \dots, n_n)$ is a node set of the network, including the spatial units (i.e., jiedao/ townships) that trip origins are residences and destinations are consuming places. $L \coloneqq (l_{(1,1)}, l_{(1,2)}, l_{(1,3)}, \dots, l_{(i,j)})$ is a link set formed by the consumption travel flows between origins and destinations. The 141,236 consumption travel OD flow data was converted into a directional residential, and commercial spatial incidence matrix with the number of consumption travels as the link weight, which can be represented as $W \coloneqq (w_{(1,1)}, w_{(1,2)}, w_{(1,3)}, \dots, w_{(i,j)})$. It should be noted that, in the weighted directed network, the weight-set W corresponds to the link-set L one-to-one. If the number of consumption travel flows from one origin to one destination is greater than 0, it is believed that there is a commercial and residential spatial association between the two spatial units. The strength of spatial association between two spatial units is determined by the number of consumption trips between two urban spaces. Then, an i×i consumption travel association matrix of i spatial units is established.

Node centrality is the association degree of one node with the other nodes within the network. It is an important index to measure the node hierarchy within the network, which reflects whether the node is the core one. Node centrality is applied to analyze the role and hierarchy of each spatial unit node in the commercial and residential spatial association network. Within a weighted and directed network, node centrality includes out-degree and in-degree. In this study, the out-degree of a spatial unit is the number of residents' consumption travel flow across the spatial unit that they are living in, while the in-degree spatial unit is the number of residents' consumption travel flow in from other spatial units.

$$C(n_i) = \sum_{j=1}^{k} (C_{in}(n_i), C_{out}(n_i)) \quad (1)$$

$$C_{in}(n_i) = \sum_{j=1}^{k} r_{ij,in} \quad (2)$$

$$C_{out}(n_i) = \sum_{j=1}^{k} r_{ij,out} \quad (3)$$

Where, $C(n_i)$ is the centrality of node i; $C_{in}(n_i)$ is the in-degree of node i; k is the number of nodes within the network; $r_{ij,in}$ is the number of consumption travel flows from j to i; $C_{out}(n_i)$ is the out-degree of node i; $r_{ij,out}$ is the number of consumption travel flows from i to j.

Combined with the function of each township, the ratio between in- and out-degree is applied to analyze the role and hierarchy of each spatial unit node in the commercial and residential spatial association network. The role of the townships are classified into three categories, commercial-dominated, residential-dominated, and commercial and residential mixed townships. The critical values of the ratio between in- and out-degree classifying the role of townships are determined based on the distribution of the data.

2.3.2 Community detection

Community detection is a network analysis classification based on topological linkages and attributes. (Daepp et al., 2022; Song et al., 2021). Specifically, Rosvall and Bergstrom (2007) introduced the Infomap algorithm on the basis of information theory, which can be applied to reveal the structure of communities in weighted directed networks. Thus, we adopt the Infomap algorithm to identify the cluster structure of the commercial and residential spatial association network, which can further contribute to extracting the spatial association patterns. The basic idea of this algorithm is to quantify node connection based on information flow across the network. Information flows substantially more between intra-community nodes than between inter-community nodes. The Infomap algorithm takes the coding length of the trajectory of a random walker as the objective function of optimization and transforms the network partition problem into an information compression coding problem that minimizes the description length (Xu et al., 2017; Zhong et al., 2014). This method is applied based on the weighted and directed network from residents' consumption flow.

The Infomap algorithm is based on information compression coding. There are three popular coding methods: equal-length coding, Huffman coding, and double-layer Huffman coding. In this study, the double-layer Huffman coding is applied given its high code reusability and computational speed. A double-layer Huffman coding method is adopted to encode the network nodes with their adjacent connected edges. Then, using random walking, it seeks the trajectory of the best ergodic network. The coding length is determined by Formula 4.

$$L(M) = q_{out}H(Q) + \sum_{i=1}^{m} p^i H(p^i) \tag{4}$$

$$q_{out} = \sum_{i=1}^{m} q_{out}^i \tag{5}$$

$$H(Q) = -\sum_{i=1}^{m} \frac{q_{out}^i}{\sum_{j=1}^{m} q_{out}^j} \log\left(\frac{q_{out}^i}{\sum_{j=1}^{m} q_{out}^j}\right) \tag{6}$$

$$p^i = \sum_{\alpha \in i} p_\alpha + q_{out}^i \tag{7}$$

Where the q_{out} is the proportion of all codes that represent group names in the code; q_{out}^i is the probability of occurrence of the name of group i; $H(Q)$ is the average length in bytes required to encode group names; p^i is the proportion of codes in the code for all nodes belonging to group i; $H(P^i)$ is the average length in bytes required to encode all nodes in group i.

Figure 3 depicts the workflows of the Infomap algorithm, whereas Figure 3a depicts the traditional approach in which each node is considered as a separate community with its own code. Figure 3b depicts the unique codes of nodes inside communities, whereas nodes in different communities may be identical. Figure 3c depicts the Community code.

Node weights, connected edge weights, and connection directions are important attributes for mining the community structure of networks with actual flow interactions. However, most of the traditional algorithms still cannot effectively take into account the weighted and directed network. Most of them perform network segmentation without information on network weights and connection directions. The performance of the algorithms varies, and the results are often deviated from the real world (Xu et al., 2017). It has been demonstrated that the Infomap algorithm is efficient in community detection studies, which can fully consider topological attributes such as node weights, connected edge weights, and connection directions. Thus, this method provides great adaptability and robust performance for the community segmentation of real-world network. Lancichinetti and Fortunato (2009) conducted a comparative examination of some

algorithms for community detection on diverse graphs: the Girvan and Newman benchmark, Lancichinetti-Fortunato-Radicchi benchmark, and random graphs. They concluded that Rosvall and Bergstrom's (2007) Infomap method outperforms the others on the benchmarks they examined. With excellent performances of low computational complexity and rapid calculation speed, this method can also be used with weighted and directed networks, which allows one to examine huge network sizes. The publications by Rosvall and Bergstrom (2007) contain additional information regarding this algorithm.

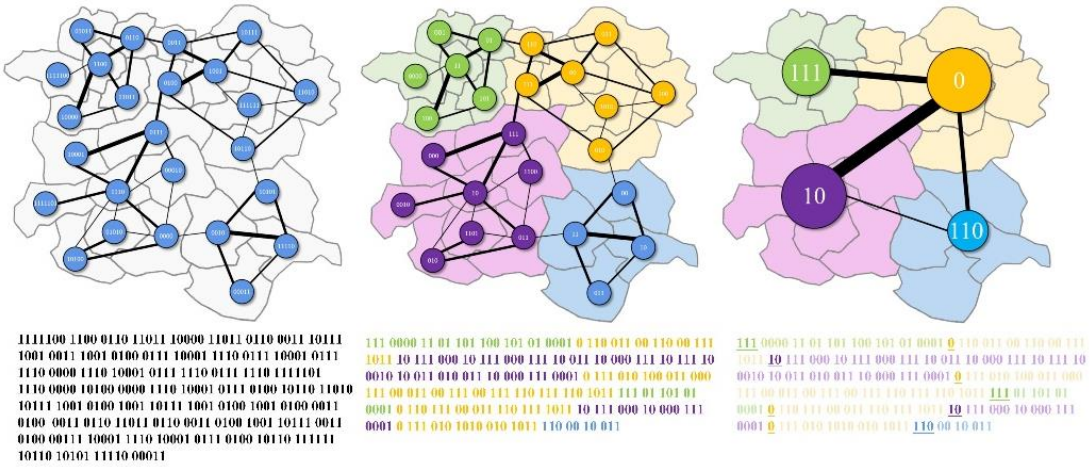


Figure 3. Infomap encoding and clustering process

(a) Huffman coding construction based on random walk probability; (b) hierarchical coding; (c) category coding in hierarchical coding. The matching coding sequences are presented at the bottom, and hierarchical coding sequences are shorter.

3 Spatial distribution of commercial spaces and residential spaces

According to the ratio between in- and out-degree from the taxi trajectory data, we classify the townships in Shanghai into three categories, commercial-dominated (the ratio is smaller than 0.6), residential-dominated (the ratio is larger than 1.7), and others (commercial and residential mixed). 72 townships with limited in- and out-degree cannot be identified as any category because there are only a few commercial trips.

(1) Commercial-dominated townships

Commercial-dominated townships are the main inflow areas of residents' consumption travels, whose in-degree is usually much higher than out-degree. Spatially, most of the commercial-dominated townships are identified in the central urban regions, and there are also some in the outer urban regions (Figure 4). According to the node in-degree, the commercial-dominated townships are classified into three categories: high-, medium- and low-level. There are many high-level commercial-dominated townships within the inner ring. Medium-level commercial-dominated townships generally have high-grade commercial centers (Figure 1). With a relatively large quantity, low-level commercial-dominated townships are mainly distributed around the outer ring and suburban ring. They emerged in recent years, and most are not of high grade and do not have high-end commercial formats. These results validate the effectiveness of the trajectory data in detecting commercial-dominated regions at the township level. However, there is no significant relationship between the in-degree and the hierarchy of the commercial centers, which might be because they have different hinterlands. It

manifests the drawback of the network analysis focusing on the nodes, which ignores the spatial association between nodes.

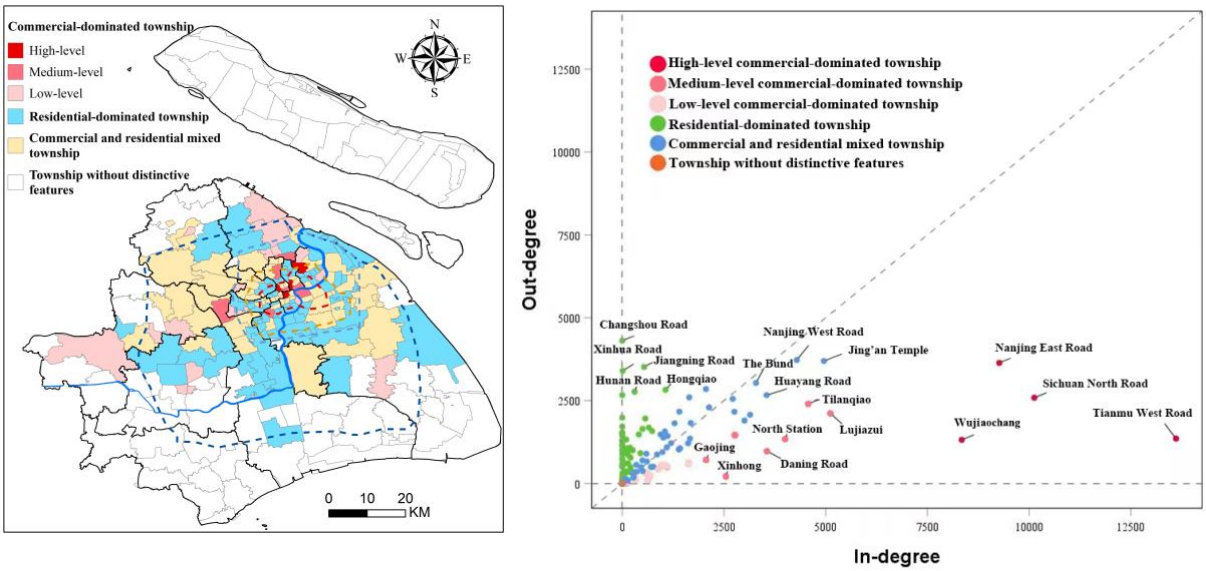


Figure 4. Role identification of *jiedao* or township

(2) Residential-dominated townships

The residential-dominated townships are the main outflow areas of residents' consumption travels. Most of the residential-dominated townships have no commercial centers or only a few low-level commercial centers. A total number of 72 townships are identified as residential-dominated, which are widely distributed from the edge of the inner ring to the suburban ring, with a concentrated and continuous distribution in the central urban regions and an outward extension in the outer urban regions. Most residential-dominated townships in Shanghai still adopt the traditional "residential neighborhood" development model. The lag behind commercial development makes the residents consume in other commercially developed townships. It raises an issue regarding the analytical scale when scholars are trying to estimate the relationship between residences and the space for commercial activities. In the central urban region, a small analytical area is enough because people's residences and commercial activities are highly concentrated. In the outer urban regions, the residences cover a large area, making some of them disconnected from the commercial centers.

(3) Commercial and residential mixed townships

The balanced outflow and inflow of residents' consumption travel in the townships suggest a high mixture of commercial and residential functions. 54 townships are identified as the commercial and residential mixed type, with node centrality degrees at medium or low levels. They are mainly clustered in the northwest of Shanghai. Among them, the node centrality of Jing'an Temple (a first-level commercial center) and Nanjing West Road (a first-level commercial center) within the inner ring is relatively high. These two regions are the two most famous commercial centers in Shanghai, whose commercial functions are much better than some commercial-dominated townships. However, these two regions might have spillover effects, raising the number of commercial trips to the surrounding regions. The township-level analysis cannot fully consider such spillover effects because it treats every township individually and ignore the spatial effects.

Regarding the background above, there are many disadvantages of the township-level analysis regarding spatial heterogeneity created by different analytical scales between inner urban regions and outer urban regions and spatial auto-correlation effects from spatial spillover. Also, the current analysis cannot tell the mechanism underlying the relationship between commercial trip generation and the level of commercial centers. Thus, a community detection method is employed for further investigating the spatial association between commercial and residential urban spaces.

4 Spatial association between residential and commercial spaces

Figure 5 presents the seven communities detected based on the spatial association network between commercial and residential spaces. We find that the Huangpu River is the natural boundary and contributes to the eastern part of Shanghai, a separated community. At the initial stages of urban development, trips between residential and commercial spaces cannot go across natural barriers. Thus, a strong connection has been established between commercial and residential spaces, creating the confined or semi-confined commercial and residential spatial association. Despite significant improvements in public transportation, existing links between commercial and residential spaces remain difficult to break due to path dependency.

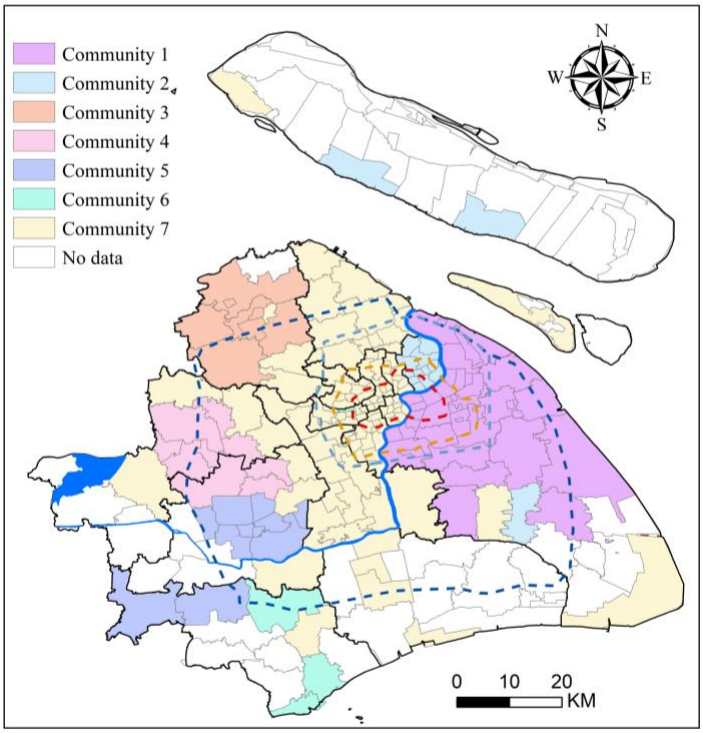


Figure 5. The distribution of spatial association community

Also, there is a significant urban-rural segmentation in the western Shanghai, as the outer urban regions contribute to self-contained communities with few connections with the urban center. For instance, communities 3, 4, 5, 6 are the commercial and residential spaces in the outer urban regions, with a few connections with central urban regions. Thus, there is a distance limitation for people's commercial activities. Although the high-

level commercial centers are located in the urban center, people prefer the low-level commercial centers within their consuming distance. People would choose the commercial center with the highest level within their consuming distance rather than the ones out of their consuming distance.

Another notable point is that there are some communities spatially disconnected, such as communities 2, 5, and 6. It is difficult to explain why people do not consume nearby and tend to travel long for consumption, particularly for community 2, whose townships are highly disconnected. It is possible that the commercial centers are super powerful in the regions, which dominate the commercial activities and generate many long-distance consuming trips. This result indicates that a spatial mismatch existed in commercial and residential spaces.

Figure 6 further visualizes the consumption within the communities to describe how different commercial centers contribute to different spatial associations between commercial and residential spaces. Communities 1, 2, and 7 have first-level commercial centers and strong internal connections. These cores even attract some long-distance consumption trips because of lacking powerful commercial centers in the surrounding area. Communities 3, 4, 5, and 6 do not have first-level commercial centers and only have weak internal connections. Based on the above results, we try to figure out the relationship between the location and level of commercial centers and their spatial association pattern with residential spaces. Indeed, the resulting spatial association patterns reveal some interesting characteristics. The organization of the communities could be classified in the following four ways.

(1) Monocentric spatial association

The first-level commercial centers at the periphery of the central urban regions can contribute to a monocentric spatial association pattern with a small hinterland, i.e., Community 2 (Figure 6a). The total consumption flows within the community is 4,970, of which 70% flows to the first-level commercial center. For the periphery of the outer urban regions, the second-level commercial sites are clustered as a center and have a strong attraction to the surrounding residents with a significant polarization effect, forming the monocentric spatial association pattern, such as Community 3 & 5 (Figure 6b & 6c). The total consumption flow within the community is only 58 and 65, respectively.

(2) Dicentric spatial association

The middle-level commercial centers of different functions located separated by a certain distance at the periphery of the outer urban regions contribute to dicentric spatial association patterns, i.e., Community 4 (Figure 6d). With a second- and third-level commercial center as the dura-core, a low-level spatial association pattern is formed with only 51 consumption flows.

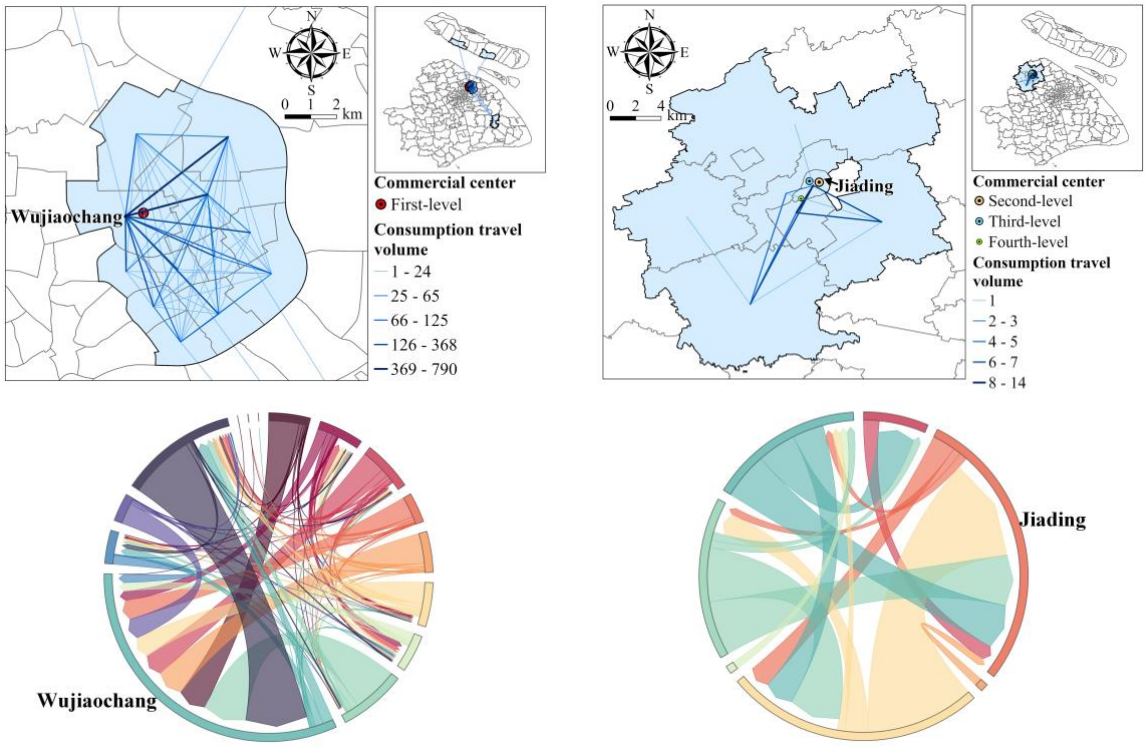
(3) Polycentric spatial association

The first-level commercial centers clustered in the central urban regions provide the ploycentric spatial association patterns with a large hinterland. However, due to the natural barrier of the Huangpu River between the districts in the central urban regions and the Pudong New Area, two polycentric spatial association patterns are formed, i.e., Community 1 & 7 (Figure 6d & 6f). Community 1 presents a clear polycentric and networked association pattern with significant spatial dependency and hierarchical characteristics (Figure 6e). Its total consumption flow reaches 8,606. Compared with the commercial center in Community 2, the ones in Community 1 follow the hierarchical structure, as there are second-, third-, and fourth-level commercial centers. As the distance would decrease residents' preference to consume, it is difficult for people in Community 1 but far from Lujiazui to consume in this first-level commercial center. That explains why communities 1 & 2 share a similar structure of commercial centers but have a different spatial association between residential and commercial space.

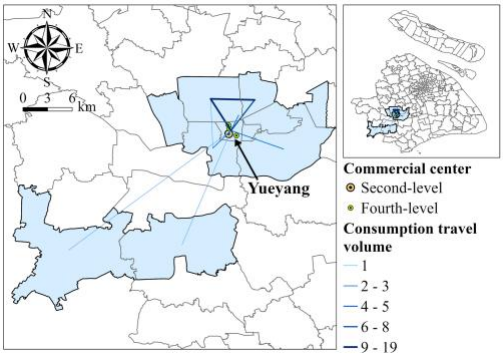
Community 7 is a typical polycentric spatial association pattern. It has a total consumption flows of 100,419, with six first-level commercial centers as its cores, making it a spatially compact and closely connected spatial association (Figure 6f). In addition, the association intensity tends to decay quickly with the increase in the distance from the core commercial center. About 90% of the inter-township consumption flow in the peripheral townships is less than 50, indicating weak association intensity within these areas. Thus, the regions with many strong commercial centers would contribute polycentric urban structures and include most commercial activities.

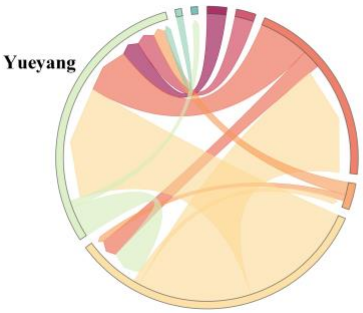
(4) Homogeneous low-level spatial association

The low-level commercial centers with similar functions at the periphery of the outer urban areas tend to form the homogeneous low-level spatial association pattern with a small hinterland of their nearby neighborhood, i.e., Community 6 (Figure 6g). The total number of consumption flows is only 5, which is relatively balanced distributed. The inter-township association network shows a relatively low density of balanced connections.

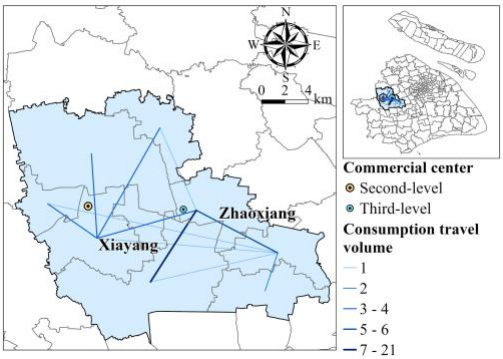


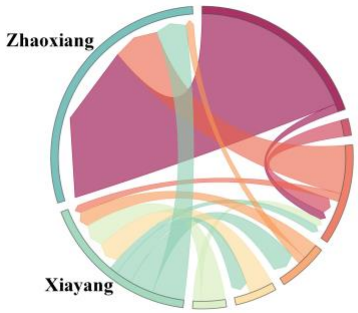
a. Monocentric spatial association (Community 2) **b.** Monocentric spatial association (Community 3)



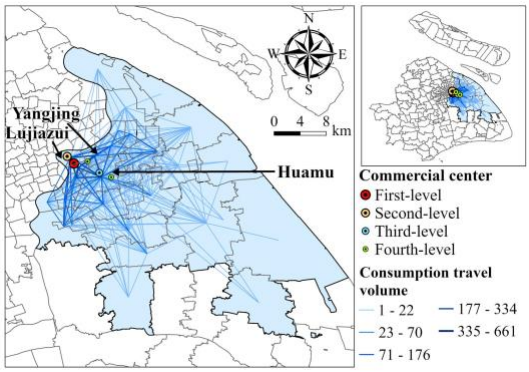


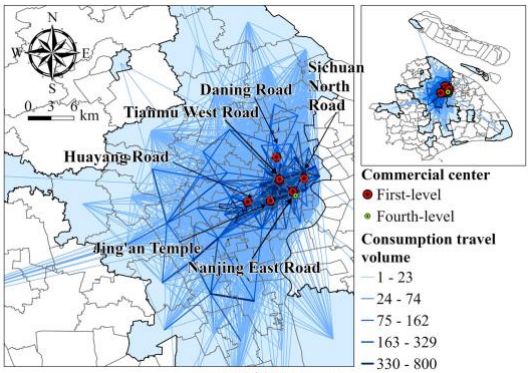
c. Monocentric spatial association (Community 5)

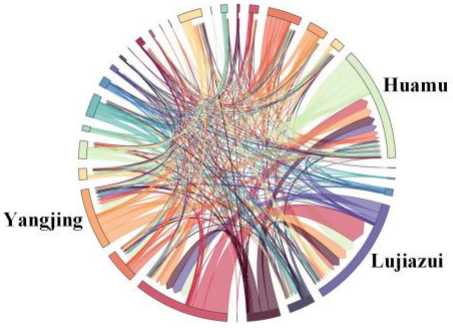


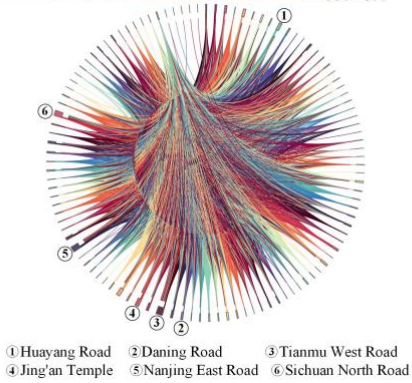


d. Dicentric spatial association (Community 4)

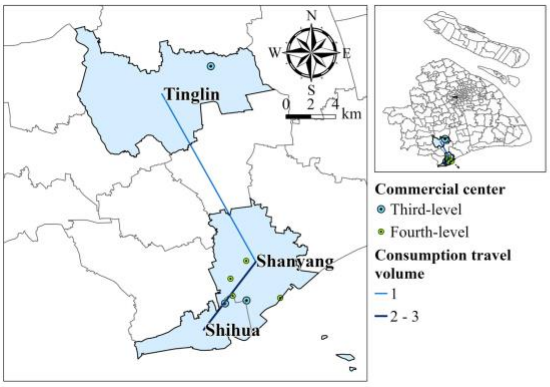


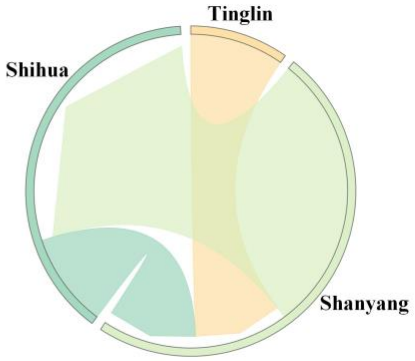






e. Polycentric spatial association (Community 1) **f.** Polycentric spatial association (Community 7)





g. Homogeneous low-level spatial association (Community 6)

Figure 6. The spatial association network and pattern of each community

Moreover, we further investigate the average consumption travel time of each community (Figure 7). The average travel time of intra-community consumption travel flows are longer than those of inter-community consumption travel flows. The polycentric spatial association pattern in community 1 & 7 has contributed to the longest inter-community average travel time between commercial and residential spaces, with the average travel time reaching 10.9 and 14.9 minutes, which confirms that the high-level commercial centers have a larger hinterland. However, the average travel time of other

communities with fewer low-level commercial centers is less than 10 minutes, which indicates that low-level commercial centers are more localized for the nearby neighborhood.

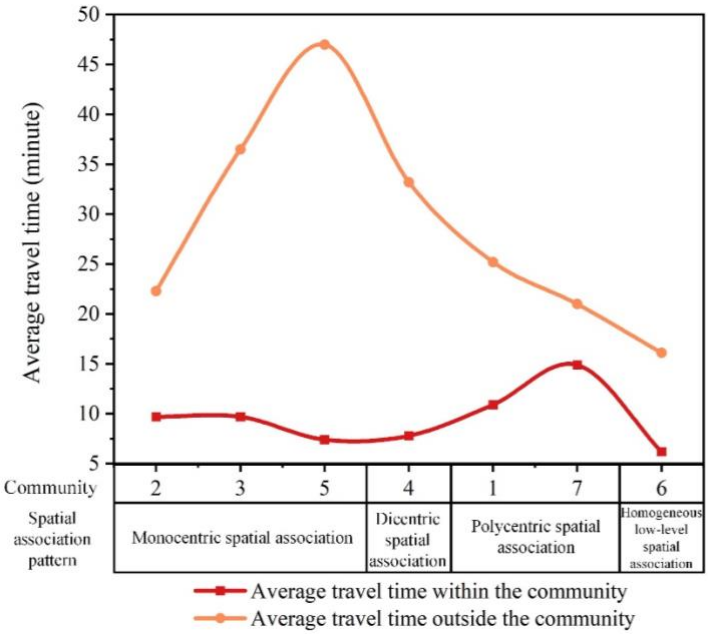


Figure 7. The average travel time of inter-community and intra-community consumption travel flow

5 Conclusion and discussions

Commercial and residential spaces are the most closely linked physical spaces in cities. The spatial association between commercial and residential spaces, i.e., the distributional dependency, locational proximity, and spatial accessibility of the two spaces with the connection of the consumption behaviors of residents, is a critical indicator of urban vitality (Nicoletti, Sirenko & Verma, 2023; Zhou, Wang et al., 2023). This study aims to provide a novel approach to exploring the spatial association between residential and commercial urban spaces drawing upon the trajectory of big data. The traditional methodologies using a fixed analytical scale, such as township in section 3, are disadvantaged in identifying the functional relationship. Specifically, there is a spatial heterogeneity between central and suburban regions at the township level, and the spillover effects from the powerful commercial centers contribute to spatial auto-correlation effects. Our research outcomes suggest that the machine learning-based approach, such as the community detection method, is less sensitive to analytical units and spatial effects. The workflow of this research provides methodological implications for future urban research using trajectory data to reduce the impact of the spatial uncertainty caused by analytical scales. It can also be applied to studies of the spatial association of other functional spaces, e.g., jobs and housing.

Empirically, the community detection results reveal that there is already a segmentation in Shanghai, as outer urban regions have self-contained commercial activities. The primary reason is that residents' commercial behavior is determined by both the level of the commercial centers and the travel distance. There might be a dynamic trade-off between the power of the accessible commercial centers and the travel distance, as some people in the outer urban regions prefer the lower-level commercial centers close to them. This research cannot support further exploration of such a trade-off

between travel distance and the power of commercial centers, which is expected to be revealed in future research.

Moreover, the high-level commercial centers are powerful in attracting commercial trips, no matter in the central urban region or the outer urban regions. However, the power of high-level commercial center in the outer urban regions is much higher than the ones in the central urban regions and contribute to a monocentric pattern. Our empirical results suggest two ways to generate a polycentric structure: (1) the regions have several high-level commercial centers; (2) there are a few commercial centers that are hierarchical. Polycentric urban development has been promoted to help reduce traffic congestion, reduce inequality, and promote sustainable urban development (Liu et al., 2016). The evidence provided in this research can support urban planning for a more polycentric urban structure. In recent years, the government has promoted polycentric strategies such as building new towns to contribute to the suburbanization of the population. However, due to the lack of self-sufficient commercial centers at the initial stage, the consumption travel distance of residents in the peripheral areas is much greater. The community detection results can help assess current outcomes of urban practices concerning the sub-centers and their hinterland, which could help government optimize the commercial and residential facilities during suburbanization.

Overall, the machine learning-based network analysis using big data is a good opportunity to deepen our understanding of the urban network. It provides tools for modeling, analysis, and visualization of urban flow data. For example, Yu et al. (2020) used a community detection model and commuting data based on mobile phone to identify commuting demand. Now network analysis has gained currency to identify the changes in people's mobility via different travel modes while the data sources are getting ample and diverse (Gibbs et al., 2020; Yang et al., 2019). The advances in methodologies are also expected to improve urban theories about people's urban travel in the age of data explosion.

Regarding the datasets, the study utilizes taxi trajectory data to identify residents' consumption flows, which might be biased in estimating consumption travels. We only considered the taxi trips between 10:00 am and 4:00 pm to exclude commute trips. Given the scarcity of real data on residents' consumption flows, the taxi trajectory data is a parametric substitution for the consumption flows. As one type of public transportation, taxi targets a certain population group, and some consuming behavior might not take a taxi like some low-income migrant workers would take the bus or subway rather than taxi. Future studies can combine with multisource data to explore the spatial association pattern between urban spaces from different perspectives.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant number: No. 42071212, No. 42201231, No. 41701185), the Qinglan Project in Jiangsu Province.

Data availability

The data that support the findings of this study are not publicly available as they are part of an ongoing dissertation. They are, however, available from the author upon reasonable request.

References

Anguera-Torrell, O., & Cerdan, A. (2021). Which commercial sectors co-agglomerate with the accommodation industry? Evidence from Barcelona. *Cities, 112*, 103112.

Boussauw, K., Van Meeteren, M., & Witlox, F. (2014). Short trips and central places: The homeschool distances in the Flemish primary education system (Belgium). *Applied Geography, 53*, 311–322.

Burger, M. J., van der Knaap, B., & Wall, R. S. (2014). Polycentricity and the multiplexity of urban networks. *European Planning Studies, 22*(4), 816–840.

Chen, Z., & Yeh, A. G. (2022). Delineating functional urban areas in Chinese mega city regions using fine-grained population data and cellphone location data: A case of Pearl River Delta. *Computers, Environment and Urban Systems, 93*, 101771.

Cheng, L., Chen, C., & Xiu, C. (2017). Excess kindergarten travel in Changchun, Northeast China: A measure of residence-kindergarten spatial mismatch. *Journal of Transport Geography, 60*, 208–216.

Daepp, M. I. (2022). Small-area moving ratios and the spatial connectivity of neighborhoods: Insights from consumer credit data. *Environment and Planning B: Urban Analytics and City Science, 49*(3), 1129–1146.

Dignum, E., Athieniti, E., Boterman, W., Flache, A., & Lees, M. (2022). Mechanisms for increased school segregation relative to residential segregation: A model-based analysis. *Computers, Environment and Urban Systems, 93*, 101772.

Fan, Y., Khattak, A., & Rodríguez, D. (2011). Household excess travel and neighborhood characteristics: Associations and trade-offs. *Urban Studies, 48*(6), 1235–1253.

Gao, S., Wang, Y., Gao, Y., & Liu, Y. (2013). Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality. *Environment and Planning B: Planning and Design, 40*(1), 135–153.

Gao, S., Janowicz, K., & Couclelis, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS, 21*(3), 446–467.

Gibbs, H., Liu, Y., Pearson, C. A., Jarvis, C. I., Grundy, C., Quilty, B. J., ... & Eggo, R. M. (2020). Changing travel patterns in China during the early stages of the COVID-19 pandemic. *Nature Communications, 11*(1), 5012.

Hu, S., Gao, S., Wu, L., Xu, Y., Zhang, Z., Cui, H., & Gong, X. (2021). Urban function classification at road segment level using taxi trajectory data: A graph convolutional neural network approach. *Computers, Environment and Urban Systems, 87*, 1010619.

Jeddi Yeganeh, A., Hall, R., Pearce, A., & Hankey, S. (2018). A social equity analysis of the U.S. public transportation system based on job accessibility. *Journal of Transport and Land Use, 11*(1), 1039–1056.

Kolodin Ferrari, T., da Fonseca Feitosa, F., Bogado Tomasiello, D., & Vieira Monteiro, A. M. (2021). Household structure and urban opportunities: Evaluating differences in the accessibility to jobs, education and leisure in São Paulo. *Journal of Transport and Land Use, 14*(1), 841–862.

Lancichinetti A., & Fortunato S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E, 80*(5), 056117.

Li, X., Ma, X., & Wilson, B. (2021). Beyond absolute space: An exploration of relative and relational space in Shanghai using taxi trajectory data. *Journal of Transport Geography, 93*, 103076.

Li, J., Zhang, Y., Wang, X., Qin, Q., Wei, Z., & Li, J. (2016). Application of GPS trajectory data for investigating the interaction between human activity and landscape pattern: A case study of the Lijiang River basin, China. *ISPRS International Journal of Geo-Information, 5*(7), 104.

Li, M., Wang, F., Kwan, M. P., Chen, J., & Wang, J. (2022). Equalizing the spatial accessibility of emergency medical services in Shanghai: A trade-off perspective. *Computers, Environment and Urban Systems, 92,* 101745.

Liu, X., Derudder, B., & Wu, K. (2016). Measuring polycentric urban development in China: An intercity transportation network perspective. *Regional Studies, 50*(8), 1302–1315.

Liu, X., Gong, L., Gong, Y., & Liu, Y. (2015). Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography, 43,* 78–90.

Liu, Y., Wang, F., Xiao, Y., Gao, S. (2012). Urban land uses and traffic source-sink areas: Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning, 106*(1), 73–87.

Meltzer, R., & Schuetz, J. (2012). Bodegas or bagel Shops? Neighborhood differences in retail and household services. *Economic Development Quarterly, 26,* 73–94.

Nicoletti, L., Sirenko, M., & Verma, T. (2023). Disadvantaged communities have lower access to urban infrastructure. *Environment and Planning B: Urban Analytics and City Science, 50*(3), 831–849.

Rosvall, M., & Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences, 104*(18), 7327–7331.

Safira, M., & Chikaraishi, M. (2022). On the empirical association between spatial agglomeration of commercial facilities and transportation systems in Japan: A nationwide analysis. *Journal of Transport and Land Use, 15*(1), 463–480.

Smith, L. G., Ma, M. Y., Widener, M. J. & Farber, S. (2022). Geographies of grocery shopping in major Canadian cities: Evidence from large-scale mobile app data. *Environment and Planning B: Urban Analytics and City Science, 50*(3), 723–739.

Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science, 30*(5), 881–906.

Song, J., Zhang, L., Qin, Z., & Ramli, M. A. (2021). A spatiotemporal dynamic analyses approach for dockless bike-share system. *Computer, Environment and Urban Systems, 85,* 101566.

Sun, W., Jin H., Chen Y., Hu, X., Li Z., Kidd, A., & Liu, C. (2021). Spatial mismatch analyses of school land in China using a spatial statistical approach. *Land Use Policy, 108*,105543.

Tian, Y., Winter, S., & Wang, J. (2019). Identifying residential and workplace locations from transit smart card data. *Journal of Transport and Land Use, 12*(1), 375–394.

Wang, D., & Chai, Y. (2009). The jobs–housing relationship and commuting in Beijing, China: The legacy of Danwei. *Journal of Transport Geography, 17*(1), 30–38.

Wu, C., Smith, D., & Wang, M. (2021). Simulating the urban spatial structure with spatial interaction: A case study of urban polycentricity under different scenarios. *Computers, Environment and Urban Systems, 89,* 101677.

Wu, L., Cheng, X., Kang, C., Zhu, D., Huang, Z., & Liu, Y. (2020). A framework for mixed use decomposition based on temporal activity signatures extracted from big geodata. *International Journal of Digital Earth, 13*(6), 708–726.

Xiao, W., Wei, D., & Li, H. (2021). Understanding jobs-housing imbalance in urban China: A case study of Shanghai. *Journal of Transport and Land Use, 14*(1), 389–415.

Xiao, W., Wei, Y. D., & Chen, W. (2023). Skills mismatch, jobs-housing relationship and urban commuting. *Travel Behavior and Society, 33,* 100610.

Xiao, W., & Wei, Y. D. (2023). Assess the non-linear relationship between built environment and active travel around light-rail transit stations. *Applied Geography, 151,* 102862.

Xiao, Y., Wang, Y., Miao, S., & Niu, X. (2021). Assessing polycentric urban development in Shanghai, China, with detailed passive mobile phone data. *Environment and Planning B: Urban Analytics and City Science, 48*(9), 2656–2674.

Xiong, Q., Liu, Y., Xie, P., Wang, Y., & Liu, Y. (2021). Revealing correlation patterns of individual location activity motifs between workdays and day-offs using massive mobile phone data. *Computer, Environment and Urban Systems, 89,* 101682.

Xu, J., Li, A., Li, D., Liu, Y., Du, Y., Pei, T., … & Zhou, C. (2017). Difference of urban development in China from the perspective of passenger transport around Spring Festival. *Applied Geography, 87,* 85–96.

Xu, J., Liu, J., Xu, Y., Lv, Y., Pei, T., Du, Y., & Zhou, C. (2022). Identification of spatial and functional interactions in Beijing based on trajectory data. *Applied Geography, 145,* 102744.

Yang, Y., Heppenstall, A., Turner, A., & Comber, A. (2019). A spatiotemporal and graph-based analysis of dockless bike sharing patterns to understand urban flows over the last mile. *Computers, Environment and Urban Systems, 77,* 101361.

Yu, Q., Li, W., Yang, D., & Zhang, H. (2020). Mobile phone data in urban commuting: A network community detection-based framework to unveil the spatial structure of commuting demand. *Journal of Advanced Transportation, 2020,* 1–15.

Yuan, N. J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., & Xiong, H. (2014). Discovering urban functional zones using latent activity trajectories. IEEE Transactions on Knowledge and Data Engineering, 27(3), 712–725.

Zhang, L., Zhu, L., Shi, D., & Hui, E. C. (2022). Urban residential space differentiation and the influence of accessibility in Hangzhou, China. *Habitat International, 124,* 102556.

Zhang, P., Zhou, J., & Zhang, T. (2017). Quantifying and visualizing jobs-housing balance with big data: A case study of Shanghai. *Cities, 66,* 10–22.

Zheng, Z., Zhou, S., & Deng, X. (2021). Exploring both home-based and work-based jobs-housing balance by distance decay effect. *Journal of Transport Geography, 93,* 103043.

Zhong, C., Arisona, S. M., Huang, X., Batty, M., & Schmitt, G. (2014). Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science, 28*(11), 2178–2199.

Zhou, L., Liu, M., Zheng, Z., & Wang, W. (2022). Quantification of spatial association between commercial and residential spaces in Beijing using urban big data. *ISPRS International Journal of Geo-Information, 11,* 249.

Zhou, L., Xiao, W., Zheng, Z., & Zhang, H. (2023). Commercial dynamics in urban China during the COVID-19 recession: Vulnerability and short-term adaptation of commercial centers in Shanghai. *Applied Geography, 152,* 102889.

Zhou, L., Wang, C., Zhen, F. (2023). Exploring and evaluating the spatial association between commercial and residential spaces using Baidu trajectory data. *Cities, 141,* 104514.