International Journal of Geo-Information

MDPI

Article

Delineating Source and Sink Zones of Trip Journeys in the Road Network Space

Yan Shi [1], Bingrong Chen [1], Jincai Huang [2,*], Da Wang [1], Huimin Liu [1] and Min Deng [1]

[1] Department of Geo-Informatics, Central South University, Changsha 410083, China; csu_shiy@csu.edu.cn (Y.S.); brchen@csu.edu.cn (B.C.); 215001023@csu.edu.cn (D.W.); lhmgis@csu.edu.cn (H.L.); dengmin@csu.edu.cn (M.D.)
[2] Big Data Institute, Central South University, Changsha 410083, China
* Correspondence: huangjincaicsu@csu.edu.cn

Abstract: Source–sink zones refer to aggregated adjacent origins/destinations with homogeneous trip flow characteristics. Current relevant studies mostly detect source–sink zones based on outflow/inflow volumes without considering trip routes. Nevertheless, trip routes detail individuals' journeys on road networks and give rise to relationships among human activities, road network structures, and land-use types. Therefore, this study developed a novel approach to delineate source–sink zones based on trip route aggregation on road networks. We first represented original trajectories using road segment sequences and applied the Latent Dirichlet Allocation (LDA) model to associate trajectories with route semantics. We then ran a hierarchical clustering operation to aggregate trajectories with similar route semantics. Finally, we adopted an adaptive multi-variable agglomeration strategy to associate the trajectory clusters with each traffic analysis zone to delineating source and sink zones, with a trajectory topic entropy defined as an indicator to analyze the dynamic impact between the road network and source–sink zones. We used taxi trajectories in Xiamen, China, to verify the effectiveness of the proposed method.

Keywords: origin–destination flow; source–sink zone; topic-level aggregation; spatial-constrained SOM network

Citation: Shi, Y.; Chen, B.; Huang, J.; Wang, D.; Liu, H.; Deng, M. Delineating Source and Sink Zones of Trip Journeys in the Road Network Space. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 150. https://doi.org/10.3390/ijgi13050150

Academic Editors: Wolfgang Kainz and Hartwig H. Hochmair

Received: 23 February 2024
Revised: 25 April 2024
Accepted: 26 April 2024
Published: 30 April 2024

1. Introduction

With the accelerating urbanization of the modern world, the spatial discretization of urban areas based on distinct and purposeful functions leads to frequencies and rhythms of cross-regional human mobility, powered by diversified transportation means [1]. With the associated social need to connect physically separated spaces with publicly accessible road and transportation networks, the large-scale flows of humans along them have led to multiple urban problems, including traffic congestion, environmental pollution, and emergency situations [2,3]. Consequently, there is a compelling need to accurately predict the spatiotemporal patterns and flows of traffic based on a solid understanding of the supply and demand formation mechanisms between urban facilities and human residences. Mastering this capability would enable urban planners to optimize the spatial structures of cities [4]. Today, the wide use of vehicle-mounted Global Positioning System (GPS) and other location-based services has led to the mass generation of geospatial trajectory data, from which the accurate delineation of homogeneous source–sink (i.e., origin–destination (OD)) zones is expected to provide new and powerful methods of characterizing and predicting large-scale spatiotemporal human mobility trends.

Most extant OD zone identification studies study either homogeneous volumes of inflows and outflows using spatial clustering operations [5] or the communities resident in the flow space, which are depicted by directional OD pairs. However, apart from OD densities and volumes, the specific trajectory information of individual trips provides critical route-sequence knowledge that directly impacts the efficacy and utility of road

networks [6]. Given their meaningful and influential results, these studies lack insight into the key semantic features that reflect human intentions. We contend that this type of metric is the most important of all measures when it comes to urban planning and prediction.

Nevertheless, eight case examples can be identified regarding the global similarities between distinct OD trip trajectories and their multifactor discrepancies. Figure 1a,b illustrate the first two cases, in which similar and discrepant route sequences are indicated for spatially adjacent ODs. Figure 1c,d show two other cases for spatially distant ODs. The other four are shown in Figure 1e–h, which differentiate the conditions of originating or departing from spatially adjacent areas via the given trajectories. The heterogeneity of these case examples clearly reflects the need to understand the purposes of the trips. Because detailed feature modeling is needed in this respect, this study provides a novel urban OD zone delineation approach based on topic modeling and the aggregation of trip routes in the road network space.

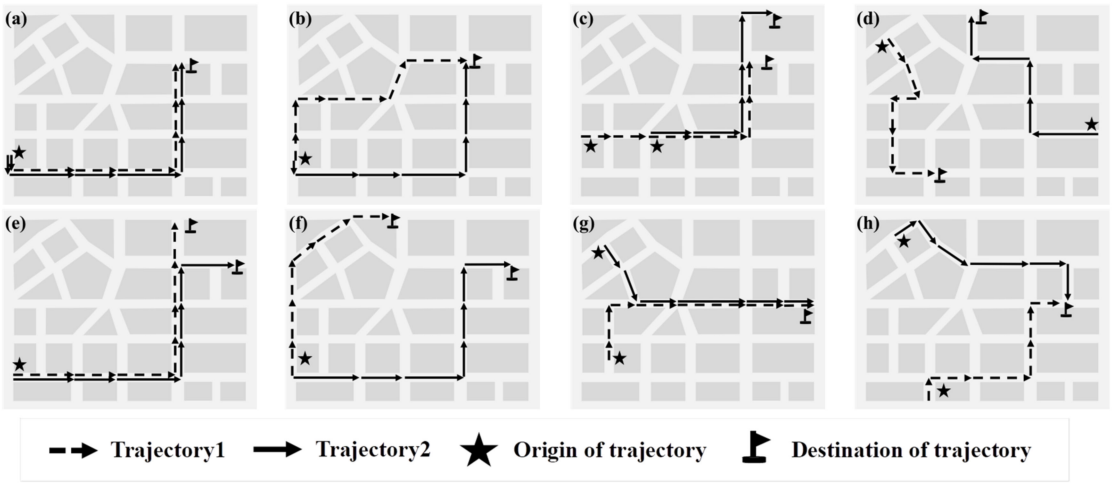


Figure 1. Eight typical cases regarding the global similarities of two distinct trip trajectories. (**a**,**b**) denote similar/discrepant route sequences with spatially adjacent origins and destinations; (**c**,**d**) denote similar/discrepant route sequences with spatially separated origins and destinations; (**e**–**h**) denote similar/discrepant route sequences under the condition of originating from or leaving for spatially adjacent areas by the two trip trajectories.

The remainder of this article is organized as follows: Section 2 reviews existing related studies to illustrate the research gap. Section 3 elaborates on the proposed approach to semantic OD zone delineation. In Section 4, experimental analyses using real-world taxi trajectory data are conducted to verify the performance of the proposed method. Section 5 summarizes the conclusions and points out future study directions.

2. Related Works

There are two types of studies on the delineation of OD zones based on individual trip trajectories in urban road network spaces. The first uses spatial clustering analysis to identify homogeneous areas separately from starting and stopping locations, and a few of these further consider the directionality of trip flows. The second uses binary variables determined by spatial OD pairs and recognizes spatial interaction types using either flow clustering or community detection operations.

2.1. Detecting Spatially Homogeneous Areas Based on OD Points

An urban OD zone comprises spatially continuous subareas of significant homogeneity based on aggregated points of origin and destination. Most studies construct spatially random temporal variables based on inflow/outflow volumes. Thus, spatial clusters and hot spots can be detected.

For example, Lee [7] adopted the classical k-means algorithm to cluster the boarding points of passengers extracted from taxi trajectories in order to find trip source zones of diverse popularity. Yue et al. [8] utilized the single-linkage-based hierarchical clustering method to mine time-dependent attractive regions and cross-regional movement patterns from taxi trajectory data. Liu et al. [9] partitioned an entire region into OD areas by modeling the temporal frequency discrepancies between pick-ups and drop-offs. Yang et al. [10] combined kernel density estimation with natural breaks to extract OD zones with crowd convergence–divergence features, whereas Fang et al. [11] conducted a spatiotemporal analysis of convergence and divergence patterns to study human mobility behaviors. Liu et al. [12] explored the relationship between functional urban polycentricity and human mobility through a multi-view analysis in the Tokyo Metropolitan Area. Huang et al. [13] proposed an approach for estimating urban functional distributions using semantics-preserved Point of Interest (POI) embedding to understand urban functionalities and their spatial distribution.

Other studies aggregate spatially adjacent areas of significant departure and arrival frequencies. For instance, Scholz et al. [4] employed a novel spatial local autocorrelation model to detect human activity hotspots and revealed dynamic patterns based on San Francisco taxi trajectories. Zhao et al. [14] developed a density-based trajectory clustering method using individual positions recorded by mobile phone base stations over time and used data field theory to extract urban traffic hotspot areas in terms of pick-up and drop-off behaviors. Deng et al. [15] examined the constraints of urban road networks on human mobility and combined spatial core point statistical identification with cluster expansion to discover OD hotspots adaptively in road network spaces. Several studies have developed grid-based spatial expansion or network-constrained bivariate spatial scanning statistics to detect urban black holes and volcanoes formed by significant aggregations of inflows and outflows. Yang et al. [16] used a constraint-based approach for identifying the urban–rural fringe of polycentric cities, delineating transitional areas between urban and rural environments. Shi et al. [17] developed a new framework for capturing urban recreational hotspots from GPS data, showcasing the use of spatial heterogeneity to identify areas of leisure and recreation in urban environments.

2.2. Detecting Spatial Interaction Areas Based on OD Pairs

Regional interactions depicted by large-scale trip flows can be used to characterize spatiotemporal cross-regional travel patterns in terms of classifying 'purpose' based on the functions of the visited facilities [18]. Considering the directional interaction features of OD pairs, many studies have performed semantic clustering and community characteristic analysis and detection.

Spatial point clustering has been applied to individual flows to aggregate OD zones [19]. For example, Zhu et al. [20] proposed a hierarchical flow clustering method to visualize large-scale inflow and outflow patterns by measuring OD flow similarities. Liu et al. [21] constructed a spatial autocorrelation index for vectors formed by directional OD taxi trips and identified hot-zone pairs using kernel density estimation. Tao et al. [22] improved the local K function to quantify spatial proximity relationships between distinct OD pairs and proposed the flowAMOEBA multidirectional optimum ecotope-based algorithm to identify hot flow clusters. Yao et al. [23] considered the similarities of OD flows in terms of overlapping geometric and temporal features and employed a k-nearest-neighbor strategy to extract spatiotemporal flow clusters. Gao et al. [24] connected origins with destinations to form binary variables and proposed a statistical spatial scanning approach to recognize significant aggregated OD flows. Song et al. [25] designed an OD flow cluster detection method that combines spatial scan statistics with ant colony optimization. Xing et al. [26] introduced "Flow Trace" as a novel representation of intra-urban movement dynamics. Wang et al. [27] proposed a classification-based multifractal analysis method for identifying urban multifractal structures, taking geographic mapping into consideration.

Several community detection studies have constructed spatial interaction graphs to analyze OD flows for zone partitioning via sub-graphs of strong internal connectivity. For example, Zhang et al. [28] designed a modular community detection algorithm that combines regional function similarities using point-of-interest (POI) spatial distributions and OD interaction intensities using bus travel smart-card data. Jia et al. [29] grouped influential nodes determined by betweenness centrality to construct an interaction network whose spatial kernel density was leveraged to recognize OD hotspots in distinct time slices. Kang et al. [30] measured hub locations in time-evolving spatial interaction networks, highlighting the role of spatiotemporal coupling and group centrality in identifying key nodes within urban networks. Sobolevsky et al. [31] constructed a spatial interaction network according to the individual trip information gleaned from mobile phone records to create a series of optimized modular sub-communities. Zhong et al. [32] extracted graph-related features and used the Infomap community detection algorithm to identify the spatial layouts of urban hubs, spokes, and rims based on boarding–alighting locations. Cao et al. [33] introduced a method for constructing multi-level urban clusters, focusing on population distributions and interactions and offering insights into how communities form and interact within cities.

2.3. Critical Analysis of Existing Studies

From the extant studies reviewed in this article, we can divide source–sink zone delineation methods into OD point- and flow-based types. Point-based methods generally make independent assumptions on pick-up and drop-off behaviors while separately identifying source and sink zones based on spatial clustering in homogeneous partitions. These methods are advantageous for the discovery of significant gathering and dispersal patterns. However, the independence assumption ignores important sequential relationships between the origins and destinations of individual trips. Hence, they are greatly limited in correlating sources and sinks with origins and destinations.

Flow-based methods further impose pairwise connection constraints on OD neighborhoods using spatial interaction graphs. Hence, OD pairs are treated as dependent binary systems that use links to aggregate binary trip rules by extending unitary points [33]. However, the necessary rigid constraints typically break the homogeneity into disconnected fragments, making it difficult to investigate large-scale human mobility patterns.

As mentioned in Section 1, for any urban trip, the origin and destination must be spatially associated through a route sequence on a road network instead of being directly connected in a Euclidean space. With this in mind, relevant studies depend on route sequences to characterize individual trip flows [34]. However, there is still a need to learn the latent semantic features of the route sequences and OD points to build a cross-regional model that reflects the purpose. Therefore, this study provides a semantic aggregation approach for delineating cross-regional urban source–sink zones considering ODs and trip routes in the road network space.

3. Methodology

In this paper, we propose a novel method for delineating urban OD zones based on road network structure. The method includes several key steps as shown in Figure 2: high-dimensional modeling of trajectories with road segments, aggregation of travel routes, region clustering, and analysis of trajectory topic entropy. First, original high-frequency trajectory points are subjected to map matching to associate them with the most suitable road segments. Subsequently, the LDA model is employed to model the relationship between road segments and travel trajectories. Second, the Wasserstein distance calculation is incorporated into clustering analysis to achieve a topic-level aggregation of trip routes, with a focus on those expressed by the modeled probability vectors of segments. In the third step, traffic analysis zones (TAZs) related to OD points are considered, and a set of multi-dimensional vectors is constructed by concatenating travel route accounts, which are subsequently classified into different clusters. In this way, a spatially constrained

self-organizing network is utilized to detect source–sink areas through multi-dimensional clustering. In the last step, trajectory topic entropy is defined as an indicator to analyze the dynamic impact between the road segments and source–sink zones. This process is described in detail in the following subsections.

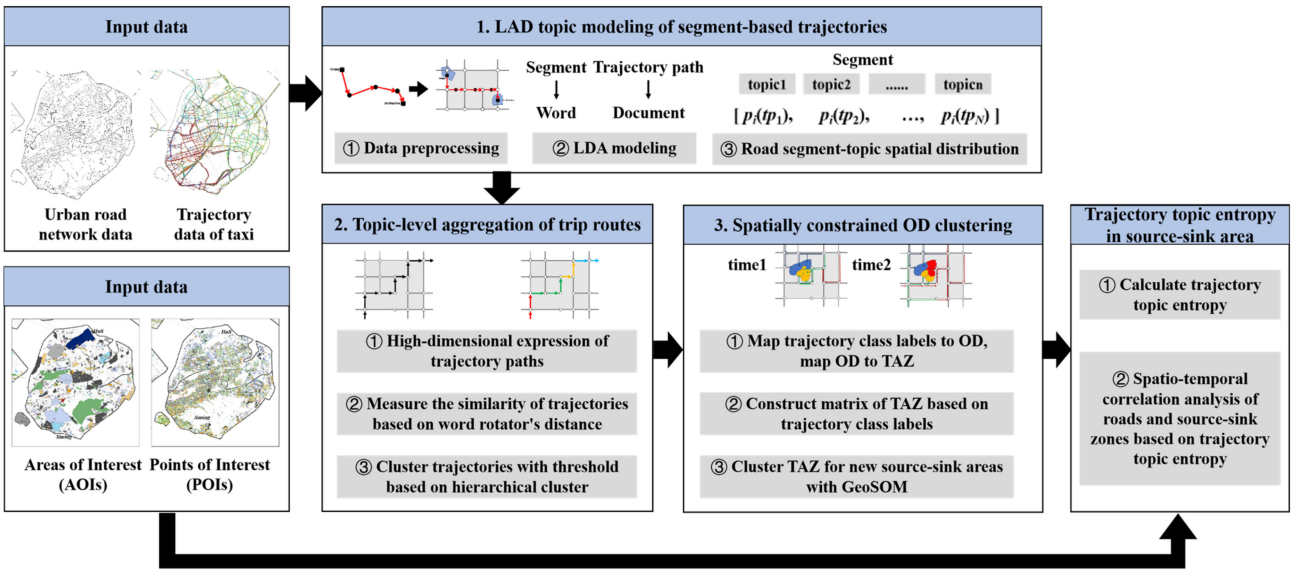


Figure 2. Framework for delineating the source and sink zones of trip journeys in road network space.

In line with the research strategy outlined in this paper, the algorithm pseudocode is provided in Appendix A. This inclusion serves to facilitate a thorough understanding and replication of the methods employed.

3.1. Topic Modeling of Segment-Based Trajectories

The complete trajectory of any trip consists of three elements: the origin area, the moving route, and the destination area. Generally, original trip trajectories are recorded in the form of sequential, discrete points with specific time intervals. The first and last points indicate geographical origin and destination areas, and the remaining points are treated as collected samples of moving routes in the road network space. In this case, we project the origin and destination points into their nearest basic geographical areas (i.e., traffic analysis zones (TAZs)). For the other trajectory points, map-matching operations in the road network are conducted to represent the moving routes by road-segment sequences. Thus, any individual trip trajectory can be denoted as follows:

$$Tr_i = (O\_taz_i, s_{i1}, s_{i2}, \ldots, s_{in}, D\_taz_i), \quad (1)$$

where Tr_i represents the complete trip, i, including its origin point mapping in the TAZ, O_taz_i, and destination point mapping, D_taz_i, along n sequential road segments ($s_{i1}, s_{i2}, \ldots, s_{in}$). s_{in} is the nth segment passed during trip i. This study further generalizes fragmented road segments by considering their geometric and travel semantics to eliminate position frequency uncertainties.

3.1.1. Segment-Based Representation of Trip Trajectories

Most urban trip trajectories use positioning devices that sample at low frequencies. Furthermore, considering the high velocities of moving vehicles, the trajectory points of individual trips have sparse distributions on the road network in most cases. This fact can give rise to the differentiation of trip trajectories, even with consistently moving routes. To address this problem, we first aggregate the road segments into a set of segments based on the principle of geometric continuity. Given any two distinct road segments, s_i and s_j,

the requirement for geometric continuity is that nodes are shared on the road network. Additionally, the two segments can be merged into a single segment if they meet one of the following conditions:

The angle between s_i and s_j is larger than a specific threshold. 120° and 180° are used in this study [35].

Their intersection has no connected segments other than s_i and s_j.

Based on the obtained road segments, any segment-based trip route with n segments can be represented as $trip_i = (stk_1, stk_2, \ldots, stk_n)$.

3.1.2. Topic-Based Representation of Road Segments

In reality, individual trips with coincident travel process preferences usually do not have consistent segment sequences based on the uncertainties of route selection at intersections. This study considers each trip trajectory to be a document that describes a potential topic determined by the specific combination of segments. Thus, diverse travel preferences can be uncovered using natural language processing. In this study, we adopt the latent Dirichlet allocation (LDA) model to aggregate trip routes with similar underlying travel semantics via conditional probability estimations [36]. We treat road segments as words and trajectories as documents, incorporating them into the LDA model. Specifically, for any route, rt_i, the joint probability of its included segment sequence can be estimated by the LDA model as follows:

$$T[(stki_1, stki_2, \ldots, stk_{in})|rt_i] = \prod_{j=1}^{n} \sum_{k=1}^{K} p(tp_k \mid rt_i) \times p(stk_{ij} \mid tp_k), \quad (2)$$

where $p(tp_k \mid rt_i) = p(\theta_i \mid \alpha) \times p(tp_k \mid \theta_i)$, $p(stk_{ij} \mid tp_k) = p(\Phi_k \mid \beta) \times p(stk_{ij} \mid \Phi_k)$. Here, K denotes the number of topics, and tpk indicates the kth latent topic. Additionally, $p(tp_k \mid \theta_i)$ represents the probability that $trip_i$ generates tp_k, where the probability distribution parameter, θ_i, can be determined by the Dirichlet distribution with hyperparameter α. The probability of generating stk_{ij} by tp_k (i.e., $p(seg_{ij} \mid \Phi_k)$) is estimated similarly by leveraging parameter Φ_k and hyperparameter β. The LDA model can be trained through parameter estimation with the help of Gibbs sampling by setting the number of topics with coherence. Thus, any segment, seg_i, on the road network can be assigned a topic with a specific probability and expressed as a K-dimensional probabilistic vector, $stk_i = [p_i(tp_1), p_i(tp_2), \ldots, p_i(tp_K)]$.

3.2. Topic-Level Aggregation of Trip Routes

Each road segment is decomposed into a K-dimensional vector based on the topic modeling of trip trajectories by constructing membership relationships with travel-related semantic topics. To further characterize the preferences of travel processes, a trip route is depicted as a sequence of topic distributions using a two-dimensional (2D) tensor of size $n \times K$, as follows:

$$Trip_i^{n \times K} = \begin{bmatrix} seg_{i1} \\ seg_{i2} \\ \vdots \\ seg_{in} \end{bmatrix} = \begin{bmatrix} p_{i1}(tp_1) & p_{i1}(tp_2) & \cdots & p_{i1}(tp_K) \\ p_{i2}(tp_1) & p_{i2}(tp_2) & \cdots & p_{i2}(tp_K) \\ \vdots & \vdots & \ddots & \vdots \\ p_{in}(tp_1) & p_{in}(tp_2) & \cdots & p_{in}(tp_K) \end{bmatrix} \quad (3)$$

Based on this representation, this study performs topic-level trip route aggregation by measuring the similarities between topic–vector sequence pairs. Among the various distance measurements, the Wasserstein distance can quantify the similarities in the probability space by minimizing the transference costs between variable probability distributions [37]. The Wasserstein distance has advantages in analyzing discrete probability distributions with an unequal number of observations compared with other probability distance metrics (e.g., Kullback–Leibler and Jensen–Shanon). To adapt the observations to multidimensional vectors, we employ a modified Wasserstein distance (i.e., word rotator distance (WRD))

designed for word–vector sequences and measure the topical similarities between distinct trip routes, which do not necessarily pass through the same number of segments.

Given the topic–vector sequences of any two trip routes, Rt_i and Rt_j, where $Rt_i = [p_{i1}(\bullet), p_{i2}(\bullet), \ldots, p_{im}(\bullet)]$ and $Rt_j = [p_{j1}(\bullet), p_{j2}(\bullet), \ldots, p_{jn}(\bullet)]$, their WRD can be expressed as follows:

$$\begin{gathered} WRD(Rt_i, Rt_j) = \min_{\gamma(Trip_{ix}, Trip_{jy}) \geq 0} \sum_{x=1}^{m} \sum_{y=1}^{n} \gamma\left(Rt_{ix}, Rt_{jy}\right) \times d(Rt_{ix}, Rt_{jy}), \\ s.t \sum_{y=1}^{n} \gamma(Rt_{ix}, Rt_{jy}) = \frac{1}{m}, \sum_{x=1}^{m} \gamma(Rt_{ix}, Rt_{jy}) = \frac{1}{n}, \end{gathered} \tag{4}$$

where $d(Rt_{ix}, Rt_{jy}) = 1 - \frac{Rt_{ix} \cdot Rt_{jy}}{||Rt|| \times ||Rt_{jy}||}$. Here, $||.||$ denotes the modulus of vectors. We employed a hierarchical clustering strategy to extract homogeneous route agglomerations based on the distances obtained between trip route pairs. These include two phases of clustering (i.e., aggregation and partition). The purpose of aggregation operations is to generate a hierarchical tree by iteratively merging the two sub-clusters with the smallest Ward's distance, which is calculated as follows:

$$Ward\_Dis(SCi, SCj) = SSD(SCij) - SSD(SCi) - SSD(SCj), \tag{5}$$

where $SSD(SC\bullet) = \sum_{m=1}^{|SC_\bullet|} \sum_{n=1}^{|SC_\bullet|} \left[\frac{WRD(Trip_m, Trip_n)}{|SC_\bullet|} \right]^2$. Inversely, the partition process aims to divide the hierarchical tree layer-by-layer by cutting the edge that connects the two sub-clusters with the largest Ward's distance. The partition operations are terminated until the largest Ward's distance is smaller than a given threshold similarity distance, Sd. Finally, route agglomerations are recognized by collecting trip routes within the same connected sub-trees.

3.3. Spatially Constrained OD Clustering

We defined OD zones as collections of spatially adjacent OD areas associated with trip route agglomerations. The key to OD zone delineation is to link the aggregated topic-level trip routes to the corresponding OD areas. Accordingly, this study constructs a multidimensional vector for each OD area to quantitatively represent travel process preferences. We then designed a spatially constrained self-organizing map neural network to determine the homogeneous source–sink zones using high-dimension clustering.

3.3.1. OD Zone Characterization by Embedding Route Agglomerations

Influenced by TAZ area imbalance, large discrepancies exist between spatially adjacent zones in terms of the volumes of urban inter-regional trips caused by significant areal differences. This areal heterogeneity problem leads to unreasonable outliers in the clustering results and negatively affects the reliable delineation of OD zones linked to characteristic travel processes [38]. Therefore, this study constructs area-weighted vectors by embedding the clustered agglomerations of trip routes into the OD zones.

For each zone, the associated outward/inward trip routes are directly determined based on their sequential relationships with boarding/alighting points located within the zone. Linked to the obtained route agglomerations in Section 3.2, the class labels of the outward/inward trip routes are used to construct a high-dimension vector, by which the underlying preference information of travel processes can be attached to the OD zone. Mathematically, given any traffic analysis zone, taz_i, the OD characteristic vectors are, respectively, represented as follows:

$$oz_i = [N_i(op1), N_i(op2), \ldots, N_i(opj), \ldots, N_i(opk)]/Area_i, \tag{6}$$

$$dz_i = [N_i(ip1), N_i(ip2), \ldots, N_i(ipj), \ldots, N_i(ipk)]/Area_i, \tag{7}$$

where $N_i(opj)$ and $N_i(ipj)$ represent the numbers of associated outward and inward trip routes with taz_i, respectively, belonging to the route agglomeration with class label j. $Area_i$

denotes the area of the taz_i. Hence, we can characterize all OD zones in the study area with n TAZs as follows:

$$OZ = \begin{bmatrix} N_1(op1) & N_2(op1) & \cdots & N_n(op1) \\ N_1(op2) & N_2(op2) & \cdots & N_n(op2) \\ \vdots & \vdots & \ddots & \vdots \\ N_1(opk) & N_2(opk) & \ldots & N_n(opk) \end{bmatrix} \begin{bmatrix} 1/Area_1 & & & \\ & 1/Area_2 & & \\ & & \ddots & \\ & & & 1/Area_n \end{bmatrix}, \quad (8)$$

$$DZ = \begin{bmatrix} N_1(ip1) & N_2(ip1) & \cdots & N_n(ip1) \\ N_1(ip2) & N_2(ip2) & \cdots & N_n(ip2) \\ \vdots & \vdots & \ddots & \vdots \\ N_1(ipk) & N_2(ipk) & \ldots & N_n(ipk) \end{bmatrix} \begin{bmatrix} 1/Area_1 & & & \\ & 1/Area_2 & & \\ & & \ddots & \\ & & & 1/Area_n \end{bmatrix}, \quad (9)$$

3.3.2. OD Zone Clustering Based on a Spatially Constrained Self-Organized Map (SOM) Network

A SOM network is an unsupervised artificial neural network that uses a competitive learning mechanism among neurons for network optimization [39]. Many studies have confirmed that SOMs have unique advantages in extracting dominant low-dimensional patterns implied in high-dimension sequence data by capturing complicated nonlinear inter-object relationships. Moreover, it preserves the intrinsic topological structure for modeling spatial dependencies in low-dimension spaces. By leveraging these specialties, this study imposes spatial continuity constraints to detect OD zones from high-dimensional vectors characterized by aggregated outward/inward trip routes.

The contemporary SOM network generates a specified number of rectangular or hexagonal grids to organize the initialized neurons with random weight vectors in the low-dimensional space. In this study, as shown in Figure 3, we utilize hexagonal grids in a 2D space to construct one-to-one mappings between TAZs and neurons while preserving spatial continuity. For each neuron, the weight vector is initialized as the OD characteristic vector of the corresponding TAZ. Given any taz_i, we first determine the winner neuron as the one with the smallest cosine distance with input vector xi and iteratively update the winner neuron with weight vector $wv_{j\times}$, as follows:

$$Wv_j \times (t+1) = wv_j \times (t) + \alpha_0 \cdot [x_i - wv_j \times (t)], \quad (10)$$

where t denotes the number of iterations, and the learning rate, α_0, is initialized at 0.5. Furthermore, by considering the spatial dependencies of adjacent neurons, those that are spatial neighbors of tazi are updated as follows:

$$wv_k(t+1) = wv_k(t) + \alpha(sde_{j\times,k}) \cdot [x_i - wv_k(t)], \text{ if } taz_k \in SN_i, \quad (11)$$

where $sde_{j\times,k} = \exp[-\frac{Dis^2[wv_{j\times}(t),\, wv_k(t)]}{2\sigma^2}]$, $\alpha(t) = \alpha_0 \cdot sde_{j\times,k}$. Here, SN_i collects the TAZ within the first- and second-order spatial neighborhoods of taz_i. $Dis[wv_{j\times}(t),\, wv_k(t)]$ denotes the cosine distance between the weight vectors of the winner neuron, j, and the k^{th} neuron. Iterative updates are terminated when all neuron weight factors converge. We then utilize the U matrix to visualize the high-dimensional mapping results of the spatially constrained SOM network in the original high-dimensional spaces. These OD zones are thus highlighted by an amalgamation of adjacent neurons with similar matrix values.

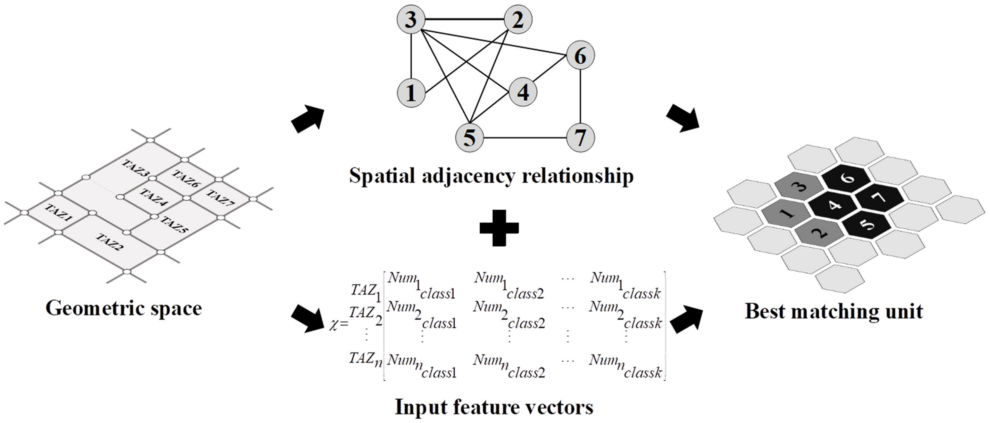


Figure 3. Illustration of a spatially constrained SOM network.

3.3.3. Trajectory Topic Entropy in Traffic Zones

Track topic entropy is an index to measure the uncertainty or complexity of the topic distribution in a traffic area. Within a traffic zone, there are N trajectories that encompass M distinct topics. For the ith trajectory, its topic number is T_i. Define n_j as the number of times topic j appears within the traffic zone:

$$n\_j = \sum_{i=1}^{n} \delta(T\_i, j), \tag{12}$$

Let $\delta(x, y)$ be the Kronecker Delta function, where $\delta(x, y) = 1$ if x = y and if $x \neq y$, $\delta(x, y) = 0$. Next, compute the frequency p_j of topic j within the traffic zone.

Using the Shannon entropy formula, calculate the trajectory topic entropy. Let H(T) denote the trajectory topic entropy in the traffic zone, then:

$$H(T) = -\sum_{j=1}^{M} (p\_j \times \log(p\_j)), \tag{13}$$

Here, the logarithm is the natural logarithm (base e), and H(T) represents the uncertainty in the distribution of road segment topics within the traffic zone. A larger value of H(T) indicates greater uncertainty in the topic distribution within the traffic zone. A higher entropy means that the trajectory topic in the region is more complex or diverse, while a lower entropy may indicate that the trajectory topic is more single or concentrated. Therefore, this can be seen as an indicator of the complexity of travel behavior within a traffic community. Reveal regional functional characteristics: Different types of regions (such as business districts, residential districts, school districts, etc.) may have different track topic distributions, resulting in different track topic entropy. For example, a commercial area may have more track topics because people visit for various reasons, while a residential area may have fewer topics because people's travel may be primarily to get home. Therefore, trajectory topic entropy can provide a new method for understanding and identifying regional functions.

4. Experimental Results and Analyses

The superiority and practicality of the proposed method were verified by performing comparative experiments on the collected taxi trajectory datasets. Section 4.1 elaborates on the real-world data taken from the study area. Section 4.2 evaluates the proposed method using quantitative comparisons with similar methods. Section 4.3 then provide the spatiotemporal analyses and detected results, respectively.

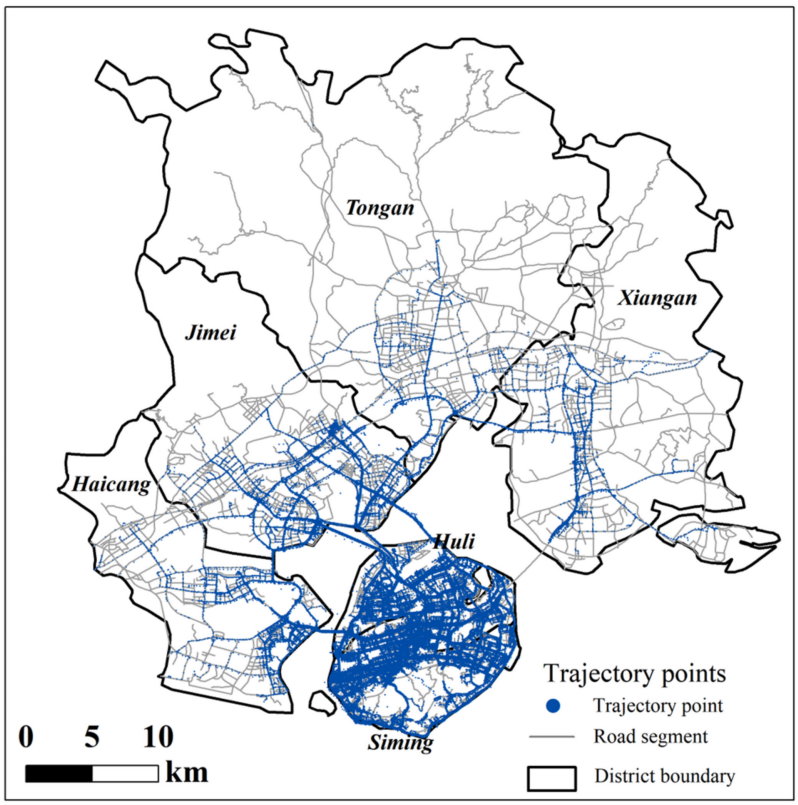
4.1. Real-Life Dataset Description

This study takes the urban area of Xiamen City as the study area. Xiamen was designated one of the first special economic zones by the China State Council. As such, its economy has grown rapidly in recent years, and taxi vehicles have become a popular mode

of public transportation. Currently, very many taxi trajectories have been recorded by GPS receivers installed in vehicles. Hence, we used that data.

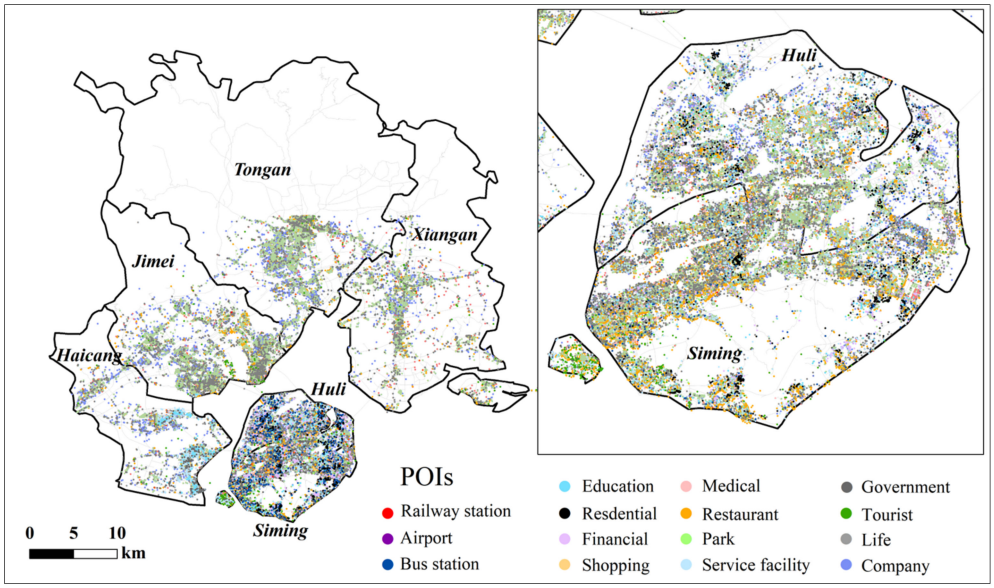
Figure 4a illustrates the spatial distribution of taxi trajectories on the Xiamen road network, which has 7573 segments, including expressways, primary, secondary, urban arterial, and secondary arterial roads. The trajectories were obtained from the digital China Innovation Contest, DCIC2020 (https://data.xm.gov.cn/opendata-competition/#/contest_explain) (accessed on 28 April 2024). Trajectory points were recorded approximately every 15 s from 31 May to 7 June 2019. Each trajectory point is mapped to the spatially adjacent road segment based on its (longitude, latitude, and direction) information. This process extracts the sequence of road segments traversed by the travel trajectory. The LDA model is fed with the processed trajectory data (now in the form of documents), identifying latent topics within the dataset. Each topic is characterized by a distribution over the "words" (road segments), indicating the likelihood of each segment being part of a particular topic. A dictionary is constructed from our "documents," mapping each unique segment to a unique integer ID. This step is crucial for the LDA algorithm, which operates on numerical representations. Each trajectory document is then converted into a bag-of-words (BoW) format. In BoW, a document is represented as a list of (wordID, frequency) tuples, where "wordID" corresponds to a road segment and "frequency" to its occurrence within the document. The LDA model is trained on this corpus, extracting topics that are essentially clusters of frequently occurring segments.

After preprocessing, 1,982,977 trajectories were selected. Figure 4b,c spatially visualize the POIs and areas of interest (AOIs) of Xiamen. The POI data contain 141,816 points labeled by 15 classes, and the AOI data include 2851 multi-polygons labeled by 12 classes. They are mostly concentrated in the Siming and Huli districts.

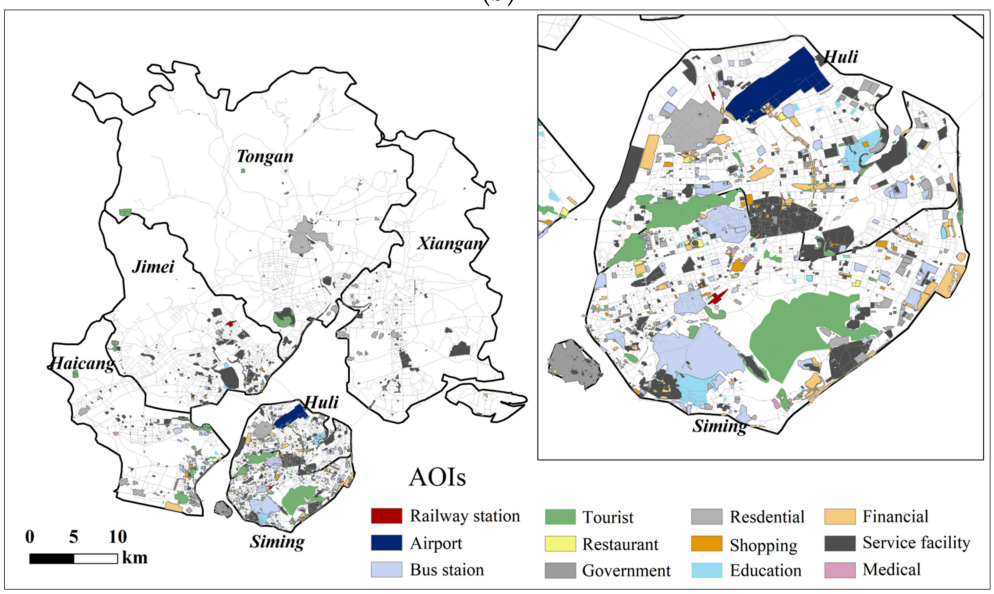


(**a**)

Figure 4. *Cont.*



(**b**)



(**c**)

Figure 4. Study area and real-life datasets. (**a**) Collected trajectories and road map data in the study area. (**b**) POIs in the study area. (**c**) AOIs in the study area.

4.2. Spatiotemporal Analysis of the Results of the Proposed Method

4.2.1. Parameter-Setting Analysis

The number of topics is a crucial LDA parameter used to aggregate trajectories with similar route semantics; thus, topic coherence is an effective evaluation index for assessing topic quality [40]. A higher coherence value indicates better topic classification. The number of topics was initially set to five in this study, and we optimized it based on the value of the coherence index at its peak. Figure 5 shows the coherence indices for different numbers of topics in terms of evening peak trajectories. These optimized topics are listed in Table 1. In an empirical analysis, it was discovered that the selection of Latent Dirichlet Allocation (LDA) algorithm parameters for trajectories under varying time periods is crucial to achieving significant spatial clustering in visualization. When the number of topics corresponding to the maximum coherence value is not chosen, the visualization of the most relevant topic space for road segments does not exhibit prominent spatial aggregation. However, when the number of topics associated with the highest coherence

value is selected, the visualization results display a markedly pronounced spatial clustering effect. This highlights the importance of carefully selecting LDA algorithm parameters to attain meaningful and effective spatial representations of data.

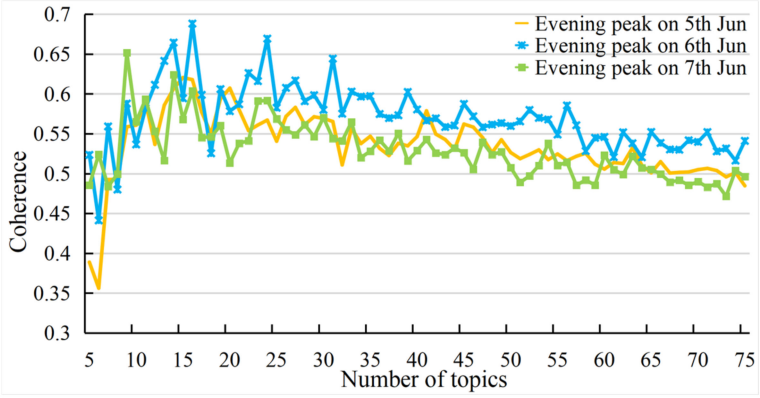


Figure 5. Coherence indexes under different numbers of topics regarding the trajectories in the evening peaks.

Table 1. Optimal number of topics in different time periods.

Dates	Time Period	Number of Topics	Date	Time Period	Number of Topics
31 May	Morning peak Evening peak	26 18	1 June	Morning peak Evening peak	20 19
2 June	Morning peak Evening peak	75 14	3 June	Morning peak Evening peak	12 16
4 June	Morning peak Evening peak	69 15	5 June	Morning peak Evening peak	45 15
6 June	Morning peak Evening peak	24 16	7 June	Morning peak Evening peak	15 9

Sensitivity analysis is employed to evaluate the performance of a model under various initial conditions. In the context of the GeoSOM clustering algorithm, sensitivity can be assessed by running the model multiple times while altering the initialization weights. The following presents a simplified Python implementation for analyzing the sensitivity of GeoSOM clustering results. The MiniBatchKMeans algorithm is utilized as the fundamental clustering algorithm for GeoSOM; however, the GeoSOM class can be modified to accommodate other clustering algorithms according to specific requirements.

Sensitivity analysis is conducted by executing GeoSOM multiple times with varying random initialization weights and calculating the silhouette coefficients for all results. The stability of the clustering outcomes can be gauged by computing the mean and standard deviation of the silhouette coefficients. A smaller standard deviation suggests that the clustering results are less influenced by the initial state and are more stable, whereas a larger standard deviation indicates that the clustering results are more sensitive to the initial state and are less stable.

Another crucial parameter is the hierarchical cluster number of trip routes. We divided the hierarchical tree under the threshold, Sd, which was mentioned in Section 3.2. Figure 6 shows the relationship between Sd and the cluster numbers. The cluster number tends to be stable when Sd is greater than the inflection point, based on which the cluster number was set. Table 2 lists the optimized cluster numbers.

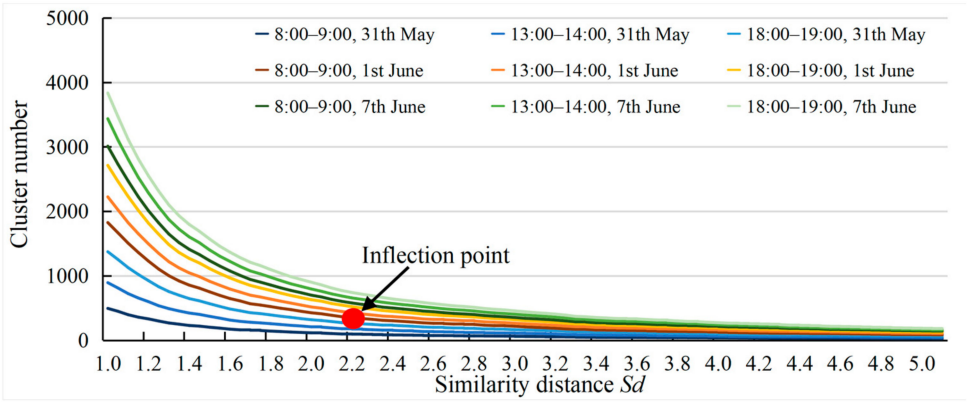


Figure 6. Trip route cluster numbers under different Sd values.

Table 2. Optimal Sd values under hierarchical clustering.

Date	Time	Sd	Date	Time	Sd	Date	Time	Sd
31 May	8:00–9:00	2.60	1 June	8:00–9:00	2.70	2 June	8:00–9:00	2.70
	13:00–14:00	2.80		13:00–14:00	2.50		13:00–14:00	2.55
	18:00–19:00	2.40		18:00–19:00	2.55		18:00–19:00	2.45
3 June	8:00–9:00	2.60	4 June	8:00–9:00	2.40	5 June	8:00–9:00	2.75
	13:00–14:00	2.65		13:00–14:00	2.60		13:00–14:00	2.70
	18:00–19:00	2.50		18:00–19:00	2.35		18:00–19:00	2.55
6 June	8:00–9:00	2.75	7 June	8:00–9:00	2.25	8 June	8:00–9:00	2.70
	13:00–14:00	2.90		13:00–14:00	2.55		13:00–14:00	2.50
	18:00–19:00	2.60		18:00–19:00	2.90		18:00–19:00	2.55

4.2.2. Spatio-Temporal Variation Patterns of Road Segment Topic

Through the analysis of the interaction characteristics of the trajectories between ODs, it is found that the spatial distribution of the same topic road segments presents interesting change rules. As shown in the Figure 7 below, the spatial distribution of road sections in three time periods of weekdays, ordinary weekends, and Dragon Boat Festival is shown, and the pattern change index analysis and spatial distribution pattern analysis of the same topic road section are shown as follows:

The R3 topic and the R2 topic are stable on weekdays and ordinary weekends, and the spatial coverage of the topic is R2 covering Lujiang Street and R3 covering the street. The coverage of R1 and R7 shows a dynamic trend and is highly correlated with the dynamic mobility trend of the working population and the residential population. Spatial density analysis based on line elements with similar topic labels reveals significant variations in connectivity strength between major roads and surrounding areas. R1, representing Chenggong Avenue, a major arterial in Xiamen, shows substantial connectivity differences with adjacent streets across various times. R7, or Jiahe Road, closely connects with Jialian Street during peak periods on weekends, weekdays, and holidays, but it associates strongly with Jiaotong Street during weekend peaks and with Jialian Street and Yundang Subdistrict during holiday peaks, highlighting dynamic spatial relationships between main and neighborhood roads.

Figure 7 shows the topic-embedded road segments. Several popular road segments (e.g., Chenggong Avenue, Xiahe Road, and Yunding North Road) were identified by the trip route topics. Weekdays and weekends shared approximately 43.6% of the topic-embedded segments, which were well-connected topologically and reflected the most intra-urban trips. Notably, the spatial distributions of the topics presented apparent time-varying characteristics. Through the analysis of interactive features between origin–destination (OD) trajectories, it was observed that the spatial distribution of road segments with the same topic exhibits intriguing changing patterns. The figure below illustrates the

thematic spatial distribution of road segments during three time periods: weekdays, regular weekends, and the Dragon Boat Festival. Morphological change index analysis and spatial distribution pattern analysis were conducted on the road segments on the same topic.

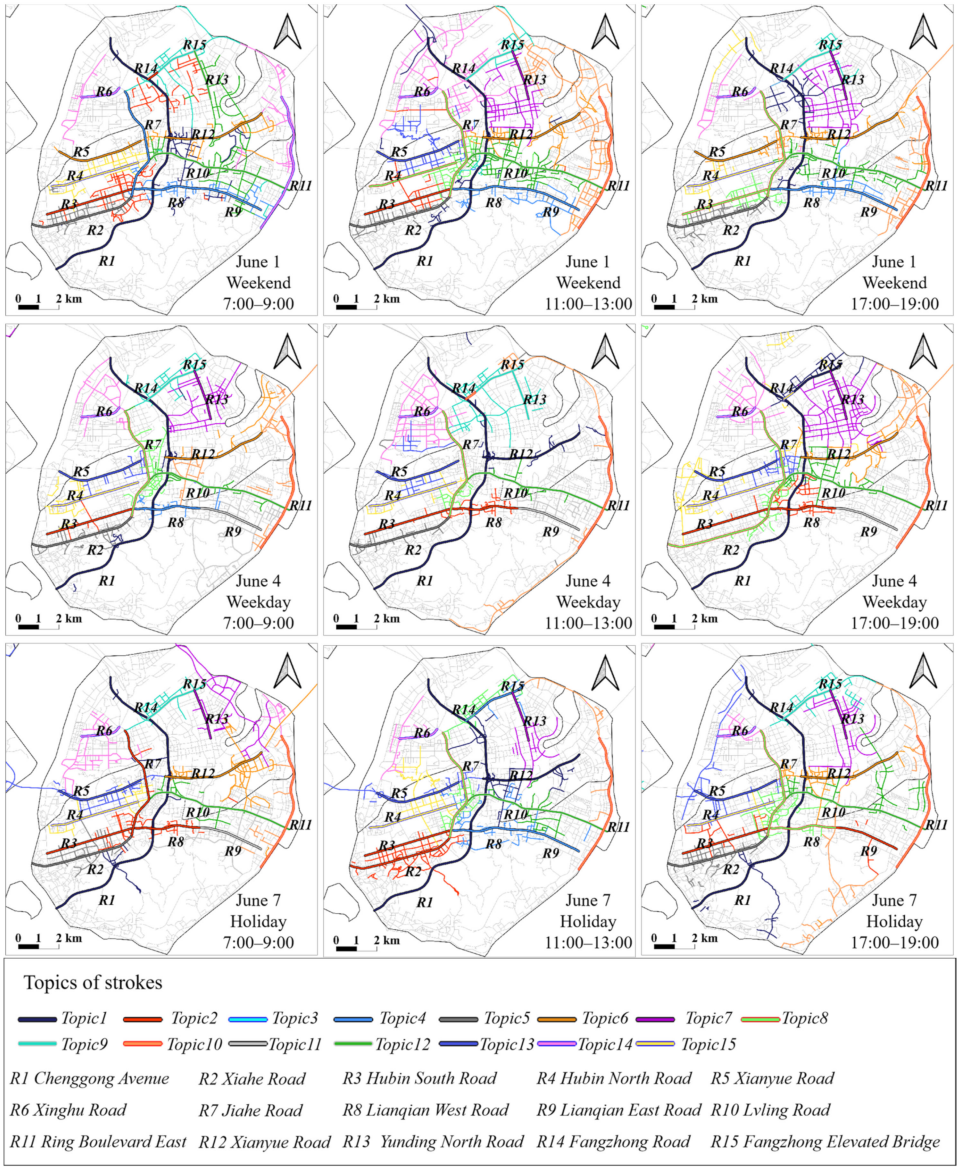


Figure 7. Spatial distribution of road segment topics in different time periods.

In this context, the R3 and R2 topics demonstrated stable structures during weekdays and regular weekends, with the spatial coverage of the R2 topic encompassing the Lujiang Street area and the R3 topic covering the Yuandang Street area. The coverage areas of the R1 and R7 topics exhibited dynamic changing trends, which were highly correlated with the dynamic flow patterns of the working and residential populations. Furthermore, the study considered the diffusion and contraction effects of different thematic road segment collections occurring during various time periods. This comprehensive analysis provides valuable insights into the underlying patterns and dynamics of spatial distributions in relation to distinct topics and time frames.

Analysis of eight trajectory patterns associated with road section topics in Figure 8.

Most of the trajectories in mode a appear on short-distance trips connecting popular areas, and the trajectories in mode a are aggregated into the same category. a There are also trajectories in mode a that form most of the similarities and only a small part of the

differences in the middle, which do not have an impact on the clustering results because the difference sections driven in the middle belong to the same spatial topic.

In Figure 8, the judgment of whether the trajectories appearing in modes c and e will be aggregated into the same category is based on the percentage of overlap and the similarity of the topic of the road section in the non-overlap place. If both indicators have high values, then the trajectories are aggregated into one category; otherwise, they are in different categories. Mode c mainly depends on the size of the share of the same path, and the large share is clustered, while the small share is influenced by the topic of the road sections around the starting and ending points of different trajectories, and the similarity is low. And the two modes f and h are mostly classified as different categories despite the existence of the same starting point or end point, due to path differences, in the case that the route section topics are not similar.

Different patterns of trajectories show special characteristics for regional clustering results: the higher the proportion of trajectory type combinations of d, f, and h, the more dispersed the regional clustering results, and the higher the proportion of trajectory types of a, c, g, and e, the larger the range of regional aggregation.

The trajectories of different modes show special characteristics for the regional clustering results: the higher the proportion of trajectory type combinations of d, f, and h, the more dispersed the regional clustering results, the higher the proportion of a, c, g, and e trajectory types, and the larger the range of regional aggregation.

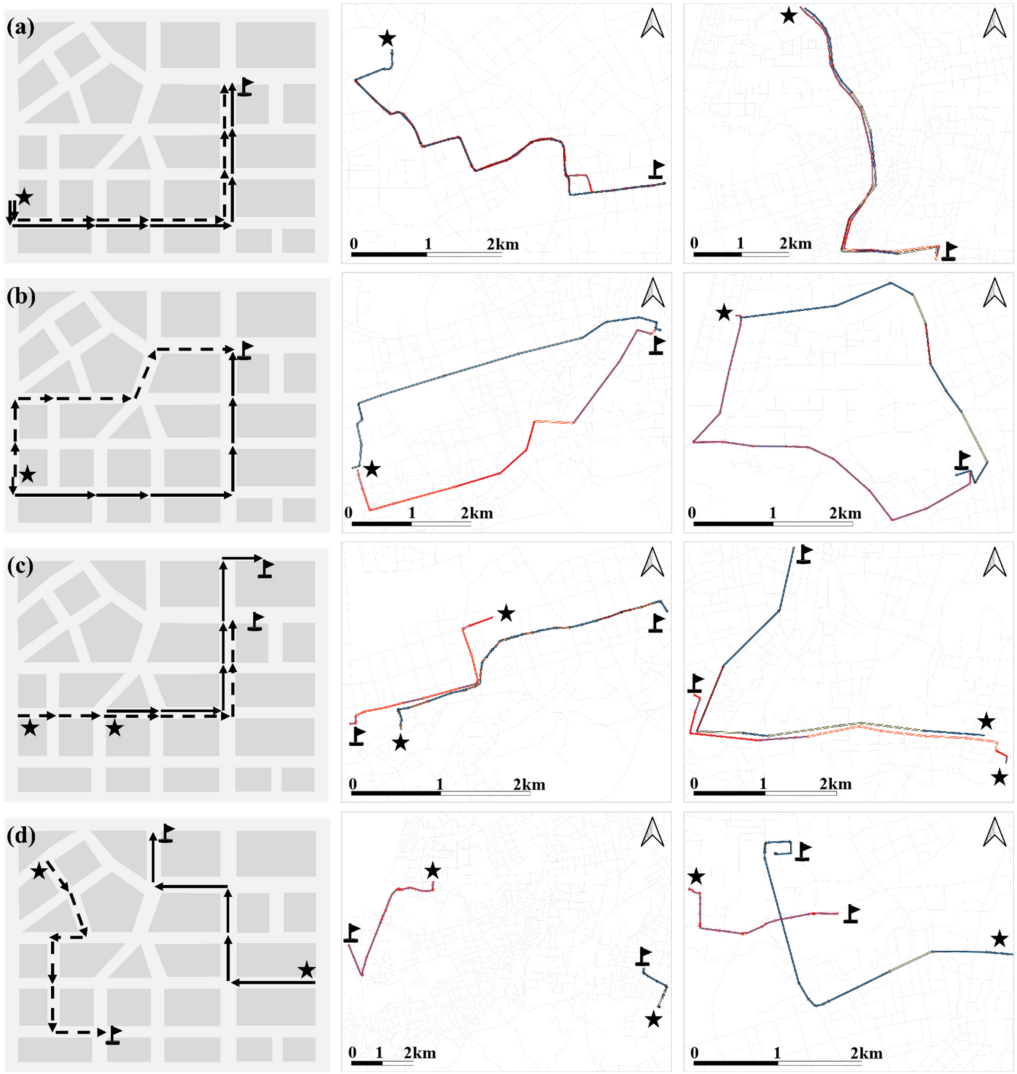


Figure 8. *Cont.*

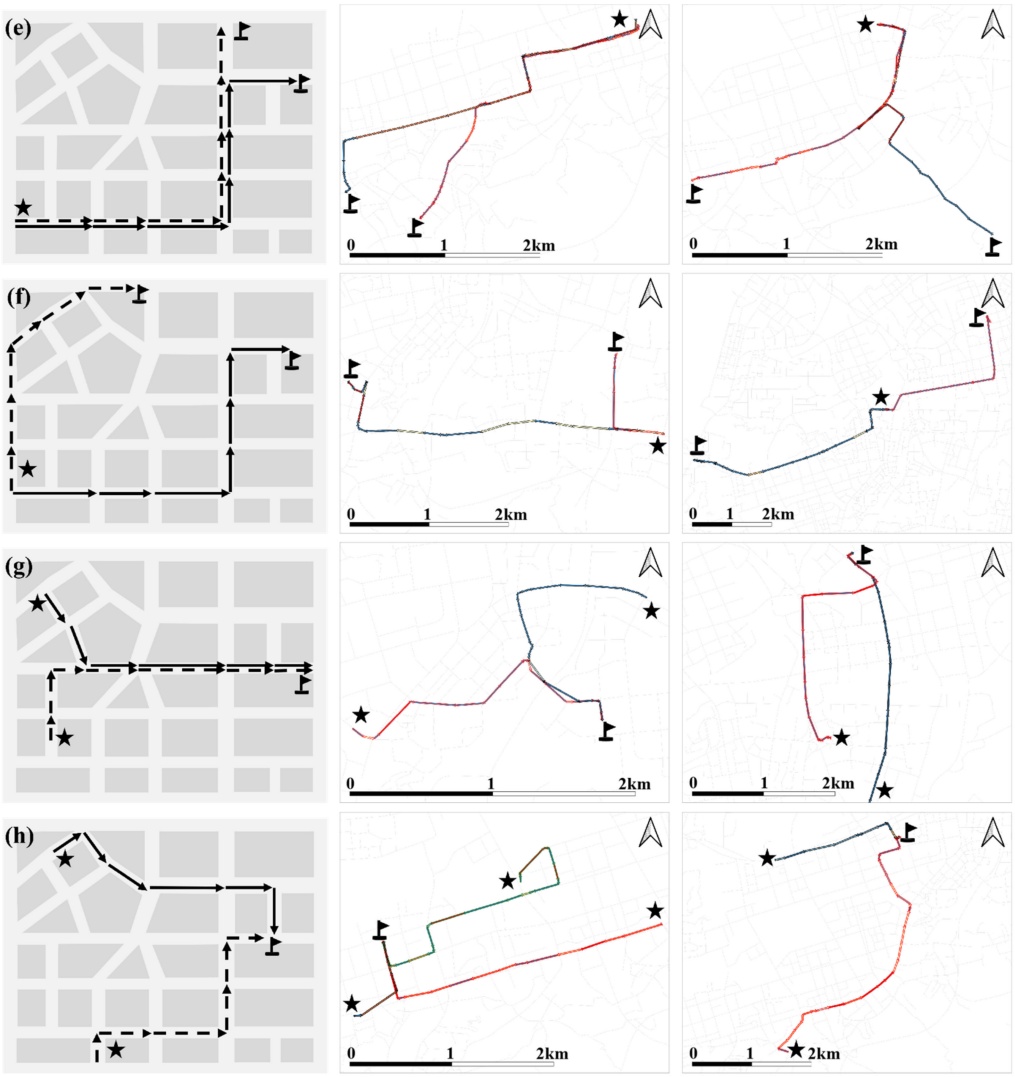


Figure 8. (**a**–**h**) Typical cases with experimental results.

From the spatial clustering results obtained for different time periods, the spatial topic entropy of trajectory data was calculated, and violin plots were generated as shown in Figure 9. Trajectory entropy is an indicator used to describe the randomness of trajectory distributions, typically employed for analyzing the complexity and uncertainty of trajectory data. The calculation method involves converting the trajectory sequences into a probability distribution and subsequently computing the information entropy of that distribution. The value range for trajectory entropy generally lies between zero and one. Higher values indicate greater unevenness, uncertainty, and complexity in the trajectory data distribution, while lower values imply a more uniform, certain, and simplified distribution. In trajectory analysis, trajectory entropy is commonly utilized to describe the diversity and changing trends of trajectories. For instance, in taxi trajectory analysis, trajectory entropy can be employed to compare the driving patterns and route selection diversity of taxis during different time periods. It was observed from Figure 9 that the distribution patterns of the global spatial topic entropy remained highly similar across different time periods. This suggests that despite potential variations in traffic flow and activity patterns within regions at different time periods, these changes did not significantly impact the overall spatial topic entropy of trajectory data. The relative stability of the global spatial topic entropy across different time periods indicates that the semantic complexity of trajectory data remains consistent over time. On a case-by-case basis, in Section 4.2.4, we will illustrate

how different types of trajectories are classified by combining topics and influencing the final regional clustering results.

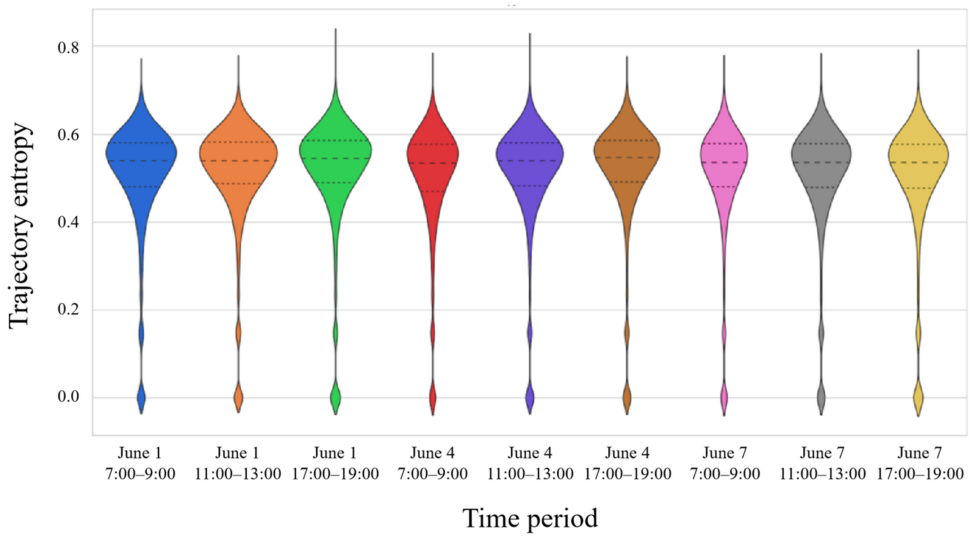


Figure 9. Spatial entropy distribution of global trajectories.

4.2.3. Analysis of the Dynamic Aggregation Situation of the Region and the Change in Trajectory Dynamic Characteristics

The global spatial distributions are shown, and it is found that the global zonal aggregation effect is not very significant. From the spatial global aggregation trend, it is difficult to extract the area with strong aggregation based on the trajectory features extracted in this paper to divide the similar area, and through the analysis, it is found that the parcels with adjacent spatial location will lead to different trajectories of cab driving because of the difference in the function of the parcels and the direction of the passenger drop-off point, which can be mainly explained by the fact that passengers generally tend to get off near the destination location and hope the driver can park on the right, which leads to the form direction of the trajectory on the road section, and despite the high similarity of the starting part of the trajectory, it will still be disturbed by the driving direction on different road sections.

Figure 10 shows the OD zones detected by the proposed method of origin zones in different time periods. At the morning peak, commuting individuals flowed mostly from residential areas to workplaces, and the opposite trip flows occurred at the evening peak. In summary, the proposed method accurately identified residential areas and workplaces as source and sink zones, respectively, at the morning peak and residential areas and workplaces as sink and source zones, respectively, at the evening peak on the weekday. Combined with the detected sink zones, metro stations (e.g., Qianpu Junction) that are strongly connected by tidal trip flows were detected as OD areas at the morning peaks of weekdays.

Figure 11 shows the detected results of destination zones in different time periods. Compared with the results of weekday peaks, intra-urban travel at the weekend peaks was concentrated at cultural and recreational facilities. Moreover, short-distance travel to cultural facilities and parks mostly originated from nearby residential areas, whereas long-distance travel to these areas was by individuals from administrative areas, airports, and railway stations. On weekdays, it was unsurprising that people usually travel to offices and companies, whereas restaurants, entertainment facilities, and parks are more attractive on weekends. Like the weekday results, residential areas were found to play a critical role in originating and terminating trip flows at morning and evening peaks, respectively.

4.2.4. Spatial–Temporal Distribution Pattern Analysis of Local Topic Entropy

Xiamen is a famous tourist city, and travel patterns were intuitively found to differ from weekdays and weekends. The main distinguishing characteristic of a holiday is that large amounts of long-distance travel appear on arterial roads. Large numbers of individuals are attracted to popular scenic spots, such as Zeng Cuo An Village, especially on holidays. Heping Pier is the main pier for travel to Gulangyu Island; therefore, on major holidays, it becomes a significant sink zone at morning peaks and afternoon periods, and it absorbs crowd flows mainly from hotels, residences, etc. By contrast, this pier is also a prominent source zone that carries crowds dispersing into areas with hotels, restaurants, and transportation facilities during evening peaks. Thus, three typical areas were selected in Figure 12 for local display and trajectory distribution analysis.

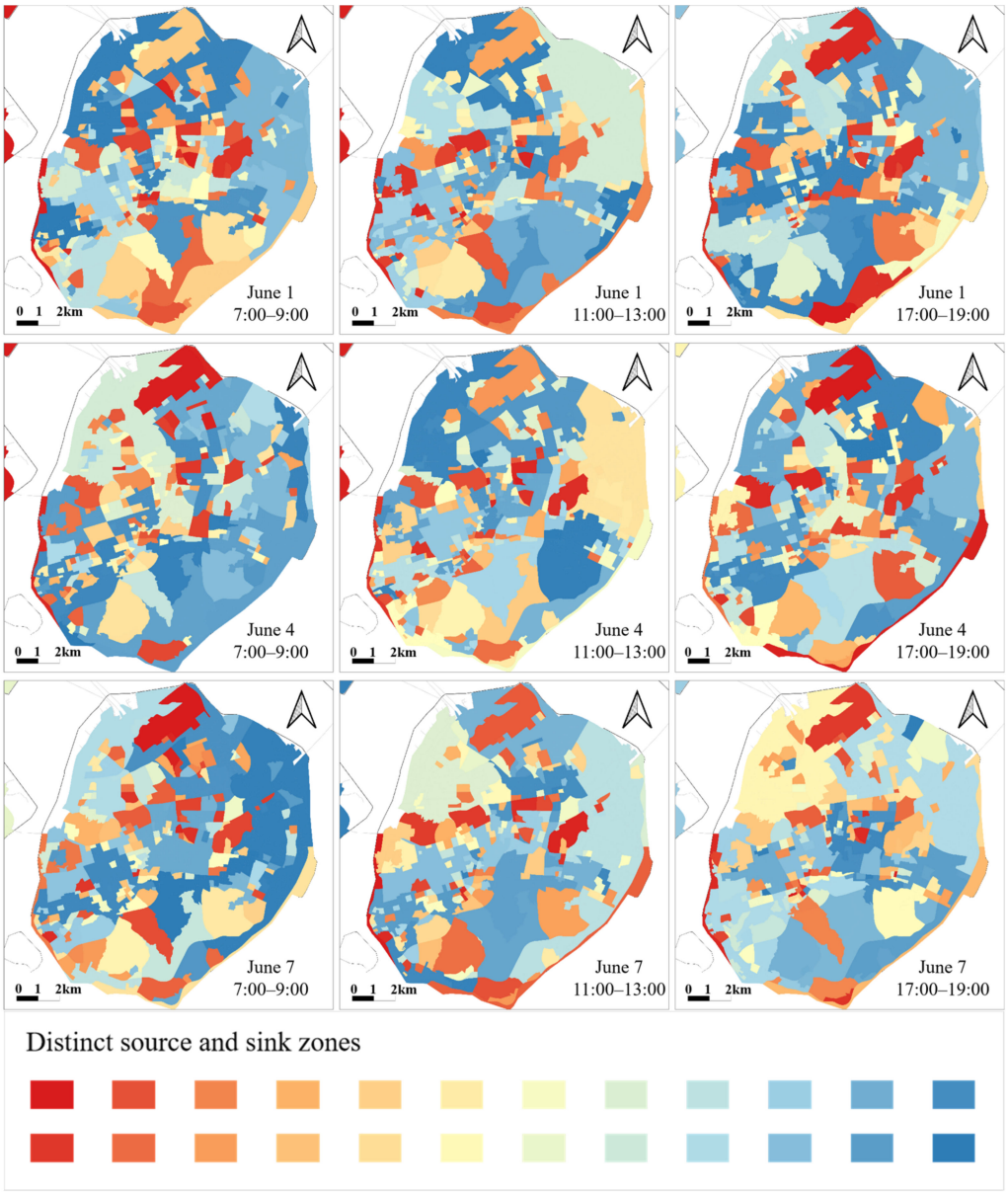


Figure 10. Global spatial distribution of origin zones in different time periods.

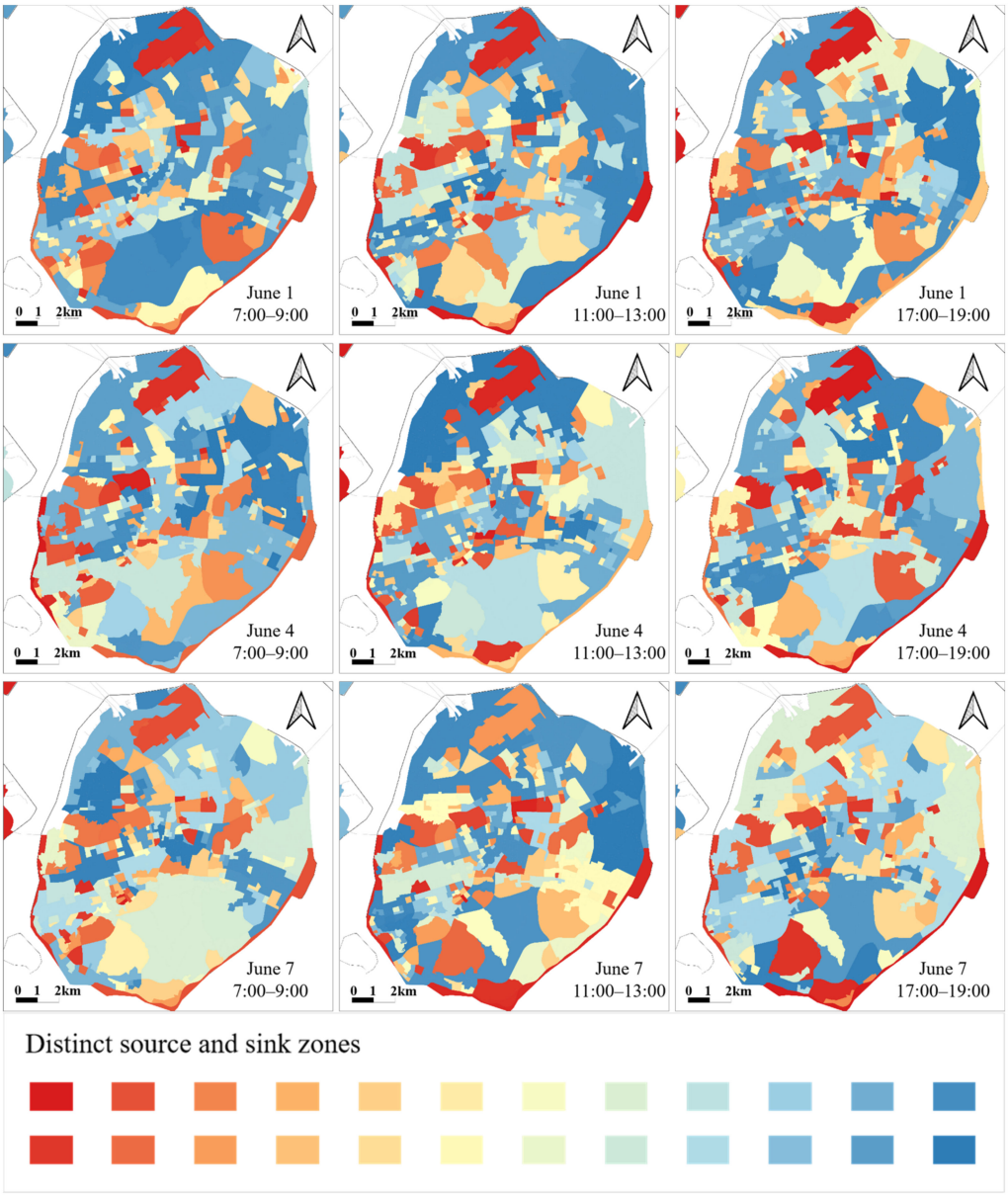


Figure 11. Global spatial distribution of destination zones in different time periods.

Using the metric of global spatial distribution, we found some interesting phenomena in the results. Meanwhile, we named an indicator called the spatial POI complexity score that measures the complexity of the types of POIs within a spatial area, which, based on the quantity and variety of POIs within a region, can be calculated as follows:

$$\text{spatial POI complexity score} = \frac{Number\ of\ POI\ Categorie \times Total\ Number\ of\ POI}{max(Number\ of\ POI\ Categorie \times Total\ Number\ of\ POI)} \quad (14)$$

As shown in Figure 13, the Xiamen Lujiang Hotel is the first four-star hotel in Fujian Province, located at 54 Lujiang Road, on the bank of the Lujiang River, in the heart of Xiamen's bustling downtown area, adjacent to customs, foreign trade, commerce, and an economic center. Vientiane City in Xiamen is located in a very advantageous location, in the core of Xiamen Island. From the subway line 1, the "Lotus Intersection" station can be found in this mall, and the traffic is very convenient. Compared with the results of weekday peaks, intra-urban travel at the weekend peaks was concentrated at cultural and recreational facilities. Moreover, short-distance travel to cultural facilities and parks mostly originated from nearby residential areas, whereas long-distance travel to these areas was

by individuals from administrative areas, airports, and railway stations. On weekdays, it was unsurprising that people usually travel to offices and companies, whereas restaurants, entertainment facilities, and parks are more attractive on weekends. The structure of Lujiang Hotel changes during the evening peaks on weekdays, weekends, and holidays. The complexity of the combined structure in the surrounding area is positively correlated with the complexity of the trajectory type. Road section clusters gather together, expressing various traffic patterns. Road section clusters represent travel in road network space and are related to functional areas, or POIs, in geographic space.

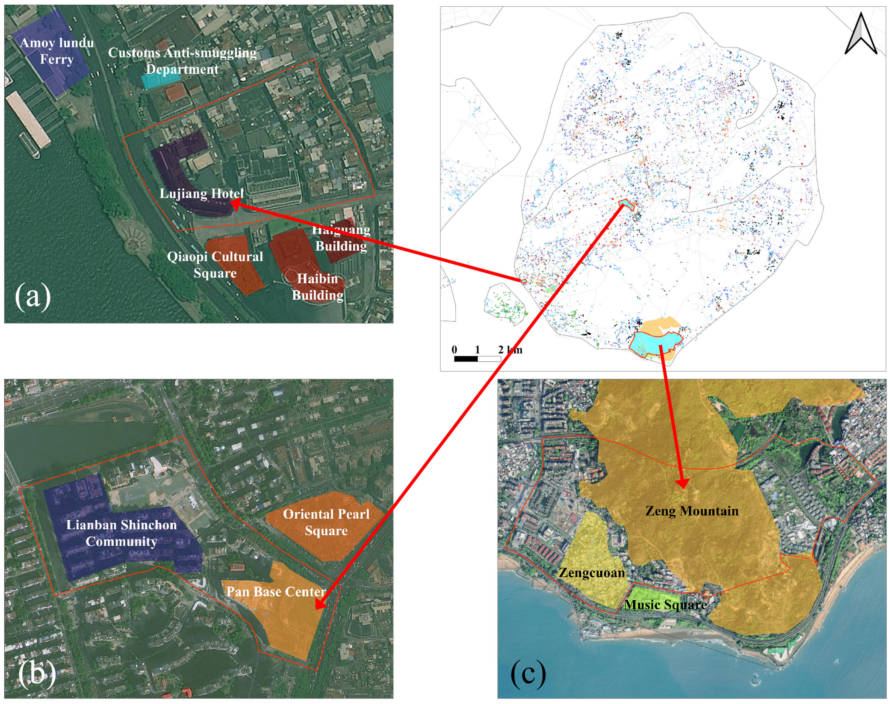


Figure 12. Results of regional clustering of three typical regions. (**a**) Lujiang Hotel; (**b**) Pan Base Center; (**c**) Zeng Cuo An.

The Panji Commereial Center in Xiamen as shown in Figure 14, located in the bustling city center, is a large-scale integrated mall that encompasses shopping, dining, and entertainment. Leveraging its unique geographical position and high-quality consumer experience, Panji Commereial Center has emerged as the preferred shopping destination for both local residents and tourists. Proximity to several major transportation routes and adjacency to numerous commercial and cultural landmarks provide residents with significant convenience and a wide array of shopping options. During weekdays, the flow of people and the interaction routes within the Panji Commereial Center are stable, with fewer visitors primarily consisting of individuals engaged in work or routine shopping. On weekends, the similarity in the volume of people and associated roads indicates an algorithmic result of area merging, reflecting the center's profound attraction to tourists. The diverse interaction routes of the Panji Commereial Center and its surrounding roads on holidays, which significantly differ from the neighboring entertainment and cultural areas, highlight the center's strong appeal to visitors.

Focusing on Figure 15, the sink zones at the morning and evening peaks were mainly detected at entertainment hubs (e.g., Zeng Cuo An Village). Like the weekday results, residential areas were found to play a critical role in originating and terminating trip flows at morning and evening peaks, respectively. Through the spatial distribution of the aggregation pattern and the trajectory of reaching the region in these three typical regions in different time periods, we find that human travel has a significant force on the dynamic changes of regional functions, and the analysis using the trajectory entropy and the contour

coefficient of the region shows that there is a certain correlation between the two. Xiamen is a famous tourist city, and travel patterns were intuitively found to differ from weekdays and weekends. The main distinguishing characteristic of a holiday is that large amounts of long-distance travel appear on arterial roads. Large numbers of individuals are attracted to popular scenic spots, such as Zeng Cuo An Village, especially on holidays. Heping Pier is the main pier for travel to Gulangyu Island; therefore, on major holidays, it becomes a significant sink zone at morning peaks and afternoon periods, and it absorbs crowd flows mainly from hotels, residences, etc. By contrast, this pier is also a prominent source zone that carries crowds dispersing into areas with hotels, restaurants, and transportation facilities during evening peaks.

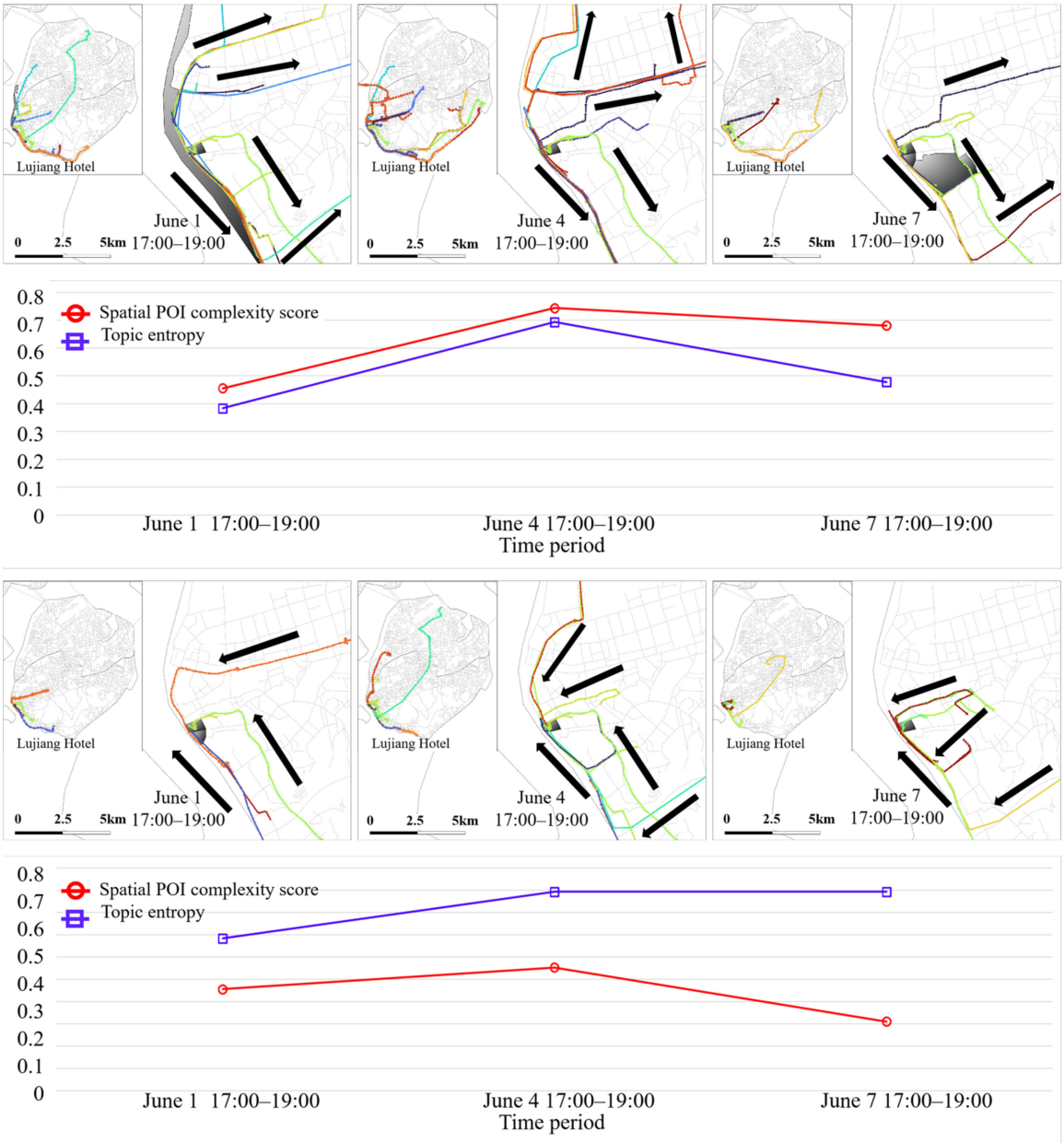


Figure 13. Results of regional clustering of Lujiang Hotel and visualization of the impact of different types of trajectories on Lujiang Hotel.

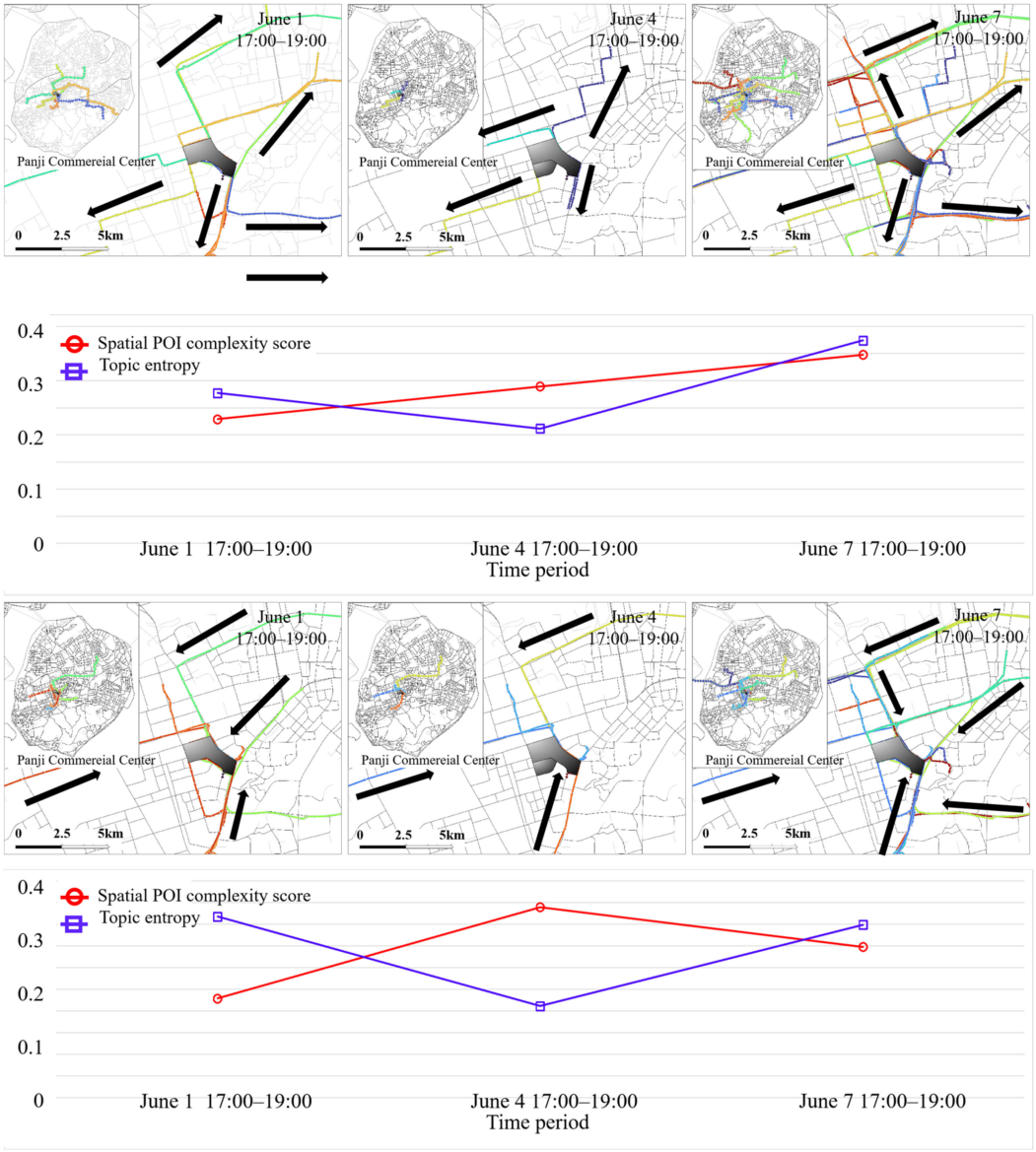


Figure 14. Results of regional clustering of the Panji Commereial Center and visualization of the impact of different types of trajectories on the Panji Commereial Center.

4.2.5. Experimental Comparisons and Evaluations

Existing studies on OD zone extraction mostly focus on source and sink data. In this study, we selected four typical methods for comparative experiments, including Liu's method [9], Zhu's method [20], Fang's method [11], and Jia's method [28]. We found that the proposed method can extract source and sink areas for critical functional public places. Liu's method, which focuses on the time series of inflow and outflow differences, obtained several zones with significant aggregation and dissipation patterns in their central areas. The OD zones were spatially discrete, and Zhu's method uses hierarchical clustering to aggregate OD flows. The extracted zones had large spatial sizes owing to the influence of the OD pair with the largest number of flows. Fang's method considers the stability of the inflow and outflow time series for regionalization. However, we found that it could not identify small areas with attractive functions in the suburbs (e.g., snack-gathering places). Jia's method uses a spatial interaction network for community detection, which leverages multiple parameters and results in the generation of more isolated zones, especially in the suburbs.

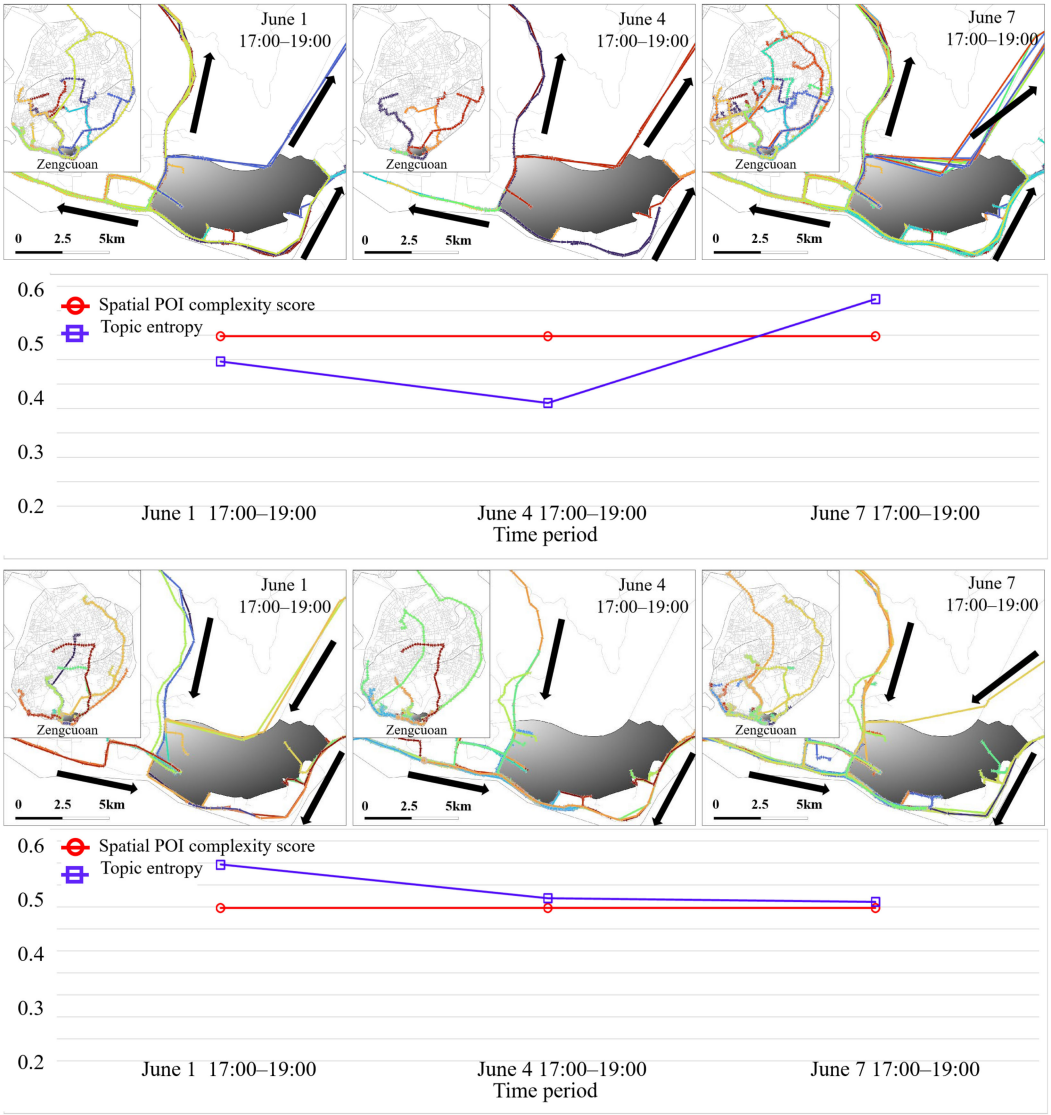


Figure 15. Results of regional clustering of Zengcuoan and visualization of the impact of different types of trajectories on Zengcuoan.

Five validation indices (i.e., the Duun index [41], Silhouette index [42], Davis–Bouldin index [43], SD index [44], and S_Dbw index [45] were selected to quantitatively evaluate the performance of the five methods. Per Section 3.1.2, we first vectorized the topic-embedded trajectory clusters for each basic spatial unit and calculated index values based on the distances between distinct units. For the Duun and Silhouette indices, larger values indicate better clustering performance, and the other three indicate the opposite.

Table 3 lists the quantitative evaluation results of the five indexes. We obtained the optimal results of the four baseline methods using parameter testing. The proposed method outperformed the other four in all five evaluation indices for the six groups of data across different time periods. This confirms that the proposed method is superior in depicting homogeneous zones with high similarities in terms of trip flow volume and sequence.

Table 3. Quantitative evaluations of different methods using clustering validation indexes.

Date	Time Period	Method	Quantitative Evaluation Indexes				
			Dunn	**Sil**	**DB**	**SD**	**S_Dbw**
Weekend (1 June)	Morning	Liu's	0.016	0.490	0.653	0.326	1.101
		Zhu's	0.734	0.743	0.687	0.196	0.456
		Fang's	0.580	0.825	0.340	0.075	0.159
		Jia's	0.741	0.803	0.622	0.145	1.389
		Proposed	0.880	0.954	0.257	0.045	0.108
	Evening	Liu's	0.592	0.754	0.963	1.002	2.269
		Zhu's	0.847	0.636	0.829	0.628	0.662
		Fang's	0.438	0.817	0.446	0.251	0.195
		Jia's	0.122	0.596	0.759	0.432	8.207
		Proposed	1.578	0.919	0.423	0.202	0.179
Workday (4 June)	Morning	Liu's	0.645	0.815	0.836	0.413	0.804
		Zhu's	0.537	0.734	0.764	0.723	2.478
		Fang's	0.802	0.657	0.805	0.564	6.574
		Jia's	0.315	0.810	0.973	0.654	1.978
		Proposed	1.286	0.927	0.631	0.275	0.167
	Evening	Liu's	0.582	0.815	0.934	0.497	2.547
		Zhu's	0.704	0.938	0.749	0.305	0.957
		Fang's	0.679	0.804	0.631	0.482	1.367
		Jia's	0.457	0.733	0.834	0.592	4.578
		Proposed	0.834	1.174	0.627	0.257	0.844
Holiday (7 June)	Morning	Liu's	0.572	0.658	0.869	0.361	0.705
		Zhu's	0.540	0.803	0.939	0.738	3.379
		Fang's	0.791	0.584	0.756	0.431	8.317
		Jia's	0.232	0.781	0.932	0.542	1.289
		Proposed	1.688	1.029	0.533	0.312	0.059
	Evening	Liu's	0.527	0.933	0.899	0.341	0.369
		Zhu's	0.642	1.041	0.738	0.291	0.424
		Fang's	0.595	0.959	0.338	0.324	0.375
		Jia's	0.306	0.785	0.943	0.616	8.391
		Proposed	0.924	1.246	0.547	0.273	0.214

4.3. Discussion

The main contribution of this paper is to map the dynamic topic information of the road network under the constraint of trajectory interaction to the traffic zones, resulting in the dynamic aggregation of traffic zones and thus assisting in the analysis of the dynamic changes of semantic functions in the regions. Based on the spatiotemporal detection results of Xiamen, we have obtained some interesting findings:

Firstly, based on the LDA model, the spatial distribution of road segment topics is obtained, and it is found that the spatial distribution of road segment topics exhibits significant aggregation in some areas. Road segments with the same topic and significant aggregation are mostly distributed around the main roads of the city, reflecting the interactive radiation range of the main roads to the surrounding road segments. Although the specific semantic information of the spatial distribution of these road segment topics cannot be directly analyzed, it can be found that these aggregation results reflect the travel preferences and travel intensity of the trajectories on the road segments.

Section 4.2.3 shows the mapping of trajectory clustering labels to the regions where the starting points are located and analyzes the spatial distribution characteristics of the regions based on the similarity of trajectory clustering results. Although the experimental results show a low overall aggregation of regional divisions, the changes in regional aggregation patterns over different time periods can assist in extracting interesting places. The more complex and diverse the spatial combinations of trajectories within the arriving area, the

greater the amplitude of aggregation pattern changes, the larger the semantic functional differences, and the stronger the comprehensiveness. Conversely, the more stable the spatial distribution of trajectories in the arriving area, the more stable the changes in the spatial patterns of the regions, and the more stable the semantic functions.

Finally, the trajectory topic entropy in the traffic zones where the starting points of the trajectories are located is analyzed. It is found that areas with higher trajectory topic entropy, such as transportation hubs (airports and train stations), and areas with comprehensive functions and diverse dynamic changes in semantic functions, form a distinct contrast with areas with lower trajectory topic entropy and stable, single-function characteristics. When the trajectory topic entropy of a region is significantly different from that of the surrounding regions, it is easier to identify the prominent feature areas in Section 4.2.3 based on the aggregation results of trajectory clustering labels. At the same time, the experimental results also found that the frequency distribution of a certain topic in a region has a high similarity with the spatial distribution of the corresponding road segment topic, but the complex interaction of trajectories makes the combination of road segment topics in the traffic zones more complicated.

5. Conclusions and Future Work

This study proposed a novel OD zone delineation approach based on trip route topic modeling and trajectory aggregations in the road network space. Trip routes were first reconstructed using road segment sequences covered by trajectories, and the LDA model was employed to learn distinct topics hidden in a series of trip routes. A hierarchical clustering operation was then carried out to aggregate the topic-embedded trip routes by introducing the WRD. Finally, the trajectory clusters for each basic spatial unit were vectorized, and a spatially constrained SOM network was adopted to detect the source and sink zones. Comparative experiments on taxi trajectories in Xiamen demonstrated the superior efficacy of the proposed method. Additionally, significant differences in travel characteristics among weekdays, weekends, and holidays were uncovered using a time-dependent analysis of the detected zones.

Future research should focus on three directions. The first is to deeply explore the spatiotemporal characteristics of intra-urban travel behaviors by integrating trip purpose information based on multiple transportation methods. The second is to investigate the influence of various factors (e.g., land use, weather conditions, and infrastructure) on the formation of OD zones. The third is to develop methods for dynamically predicting destinations, purposes, and routes of intra-urban travel on the road network.

Author Contributions: Conceptualization, Yan Shi and Jincai Huang; methodology, Yan Shi and Bingrong Chen; validation, Da Wang and Huimin Liu; formal analysis, Bingrong Chen and Huimin Liu; investigation, Deng Min; resources, Deng Min; data curation, Jincai Huang and Bingrong Chen; writing—original draft preparation, Bingrong Chen and Jincai Huang; writing—review and editing, Yan Shi and Jincai Huang; visualization, Bingrong Chen; supervision, Yan Shi and Jincai Huang; funding acquisition, Yan Shi and Min Deng. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (No. 2021YFB3900904); the National Key Research and Development Program of China, (No. 2018YFB1004603); National Natural Science Foundation of China (No. 42071452, 42171459, 42371477), the Hunan Provincial Natural Science Foundation of China (No. 2022JJ20059), the science and technology innovation Program of Hunan Province (No. 2023RC3032), the National Key R&D Program of China (No. 2021YFB3900904), Central South University Innovation-Driven Research Programme (No. 2023CXQD013); and the Guangdong Science and Technology Strategic Innovation Fund (the Guangdong-Hong Kong-Macau Joint Laboratory Program; No. 2020B1212030009).

Data Availability Statement: The data and codes supporting the findings of this study are available at figshare.com: "https://figshare.com/s/29d65fb0311bbd43093b (accessed on 28 April 2024)".

Acknowledgments: The authors thank the editor and reviewers for their useful comments and suggestions regarding the improvement of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

In line with the research strategy outlined in this paper, we provide Algorithm A1 pseudocode to facilitate a thorough understanding and replication of the methods employed.

Algorithm A1: Discovering Source and Sink Zones in Trip Routes

Input: Set of trajectories, Road segment sequences
Output: Source and Sink Zones

```
Reconstruct trajectories using road segment sequences
  FOR each trajectory in Set of Trajectories
    Road Network Binding: Associating trajectory points with adjacent road segments based on (lon, lat, direction)
    Reconstruct trip route based on road segment sequence
Apply LDA model to learn distinct topics in trip routes
  Initialize an empty list, Trajectory Documents
  FOR each trajectory in Set of Trajectories
    Create a document representation of the trajectory
    Document = Convert trajectory into a sequence of road segment identifiers
    Add Document to Trajectory Documents
  Preprocess Trajectory Documents
    Tokenize each Document in Trajectory Documents
    Remove rare and common tokens, if necessary
    Create a dictionary of all unique tokens across Trajectory Documents
Convert Trajectory Documents into a Bag-of-Words (BoW) format
  FOR each Document in Trajectory Documents
    Convert Document into BoW using the dictionary
  Store the result in Corpus
Train LDA model
  Specify the number of topics, N
  LDA_Model = Train LDA using Corpus, Dictionary, and N
Analyze the result
RETURN LDA_Model
Apply hierarchical clustering on topic-embedded trip routes
  FOR each topic in Topics
    Calculate Word Mover's Distance (WRD) between trip routes
  Clusters = Hierarchical Clustering(Topic-embedded trip routes, WRD)
  Vectorize trajectory clusters for each basic spatial unit
  Vectorize cluster
Detect source and sink zones through a spatially constrained GeoSOM network
  GeoSOM_Input = Vectorized trajectory clusters
  GeoSOM_Network = Train GeoSOM(GeoSOM_Input)
  Source_Sink_Zones = Identify zones(GeoSOM_Network)
RETURN Source_Sink_Zones
```

References

1. Long, J.A.; Weibel, R.; Dodge, S.; Laube, P. Moving Ahead with Computational Movement Analysis. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 1275–1281. [CrossRef]
2. Moreira-Matias, L.; Fernandes, R.; Gama, J.; Ferreira, M.; Mendes Moreira, J.; Damas, L. On Recommending Urban Hotspots to Find Our next Passenger. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013.
3. Huang, W.; Cui, L.; Chen, M.; Zhang, D.; Yao, Y. EstiEstimating Urban Functional Distributions with Semantics Preserved POI Embedding. *Int. J. Geogr. Inf. Sci.* **2022**, *36*, 1905–1930. [CrossRef]

4. Scholz, R.W.; Lu, Y. Detection of Dynamic Activity Patterns at a Collective Level from Large-Volume Trajectory Data. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 946–963. [CrossRef]
5. Yuan, J.; Yu, Z.; Xing, X. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 186–194.
6. McKenzie, G.; Janowicz, K.; Gao, S.; Gong, L. How Where Is When? On the Regional Variability and Resolution of Geosocial Temporal Signatures for Points of Interest. *Comput. Environ. Urban Syst.* **2015**, *54*, 336–346. [CrossRef]
7. Lee, J.; Inhye, S.; Gyung-Leen, P. Analysis of the Passenger Pick-Up Pattern for Taxi Location Recommendation. In Proceedings of the 2008 Fourth International Conference on Networked Computing and Advanced Information Management, Gyeongju, Republic of Korea, 2–4 September 2008; IEEE: Gyeongju, Republic of Korea, 2008; pp. 199–204.
8. Yue, Y.; Zhuang, Y.; Li, Q.; Mao, Q. Mining Time-Dependent Attractive Areas and Movement Patterns from Taxi Trajectory Data. In Proceedings of the 2009 17th International Conference on Geoinformatics, Fairfax, VA, USA, 12–14 August 2009.
9. Liu, Y.; Wang, F.; Xiao, Y.; Gao, S. Urban Land Uses and Traffic 'Source-Sink Areas': Evidence from GPS-Enabled Taxi Data in Shanghai. *Landsc. Urban Plan.* **2012**, *106*, 73–87. [CrossRef]
10. Yang, X.; Zhao, Z.; Lu, S. Exploring Spatial-Temporal Patterns of Urban Human Mobility Hotspots. *Sustainability* **2016**, *8*, 674. [CrossRef]
11. Fang, Z.; Yang, X.; Xu, Y.; Shaw, S.-L.; Yin, L. Spatiotemporal Model for Assessing the Stability of Urban Human Convergence and Divergence Patterns. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 2119–2141. [CrossRef]
12. Liu, K.; Murayama, Y.; Ichinose, T. Exploring the Relationship between Functional Urban Polycentricity and the Regional Characteristics of Human Mobility: A Multi-View Analysis in the Tokyo Metropolitan Area. *Cities* **2021**, *111*, 103109. [CrossRef]
13. Huang, J.; Tang, J. Discovery of arbitrarily shaped significant clusters in spatial point data with noise. *Appl. Soft Comput.* **2021**, *108*, 107452. [CrossRef]
14. Zhao, P.; Qin, K.; Ye, X.; Wang, Y.; Chen, Y. A Trajectory Clustering Approach Based on Decision Graph and Data Field for Detecting Hotspots. *Int. J. Geogr. Inf. Sci.* **2016**, *31*, 1–27. [CrossRef]
15. Deng, M.; Yang, X.; Shi, Y.; Gong, J.; Liu, Y.; Liu, H. A Density-Based Approach for Detecting Network-Constrained Clusters in Spatial Point Events. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 466–488. [CrossRef]
16. Yang, J.; Dong, J.; Sun, Y.; Zhu, J.; Huang, Y.; Yang, S. A Constraint-Based Approach for Identifying the Urban–Rural Fringe of Polycentric Cities Using Multi-Sourced Data. *Int. J. Geogr. Inf. Sci.* **2022**, *36*, 114–136. [CrossRef]
17. Shi, H.; Huang, H.; Ma, D.; Chen, L.; Zhao, M. Capturing Urban Recreational Hotspots from GPS Data: A New Framework in the Lens of Spatial Heterogeneity. *Comput. Environ. Urban Syst.* **2023**, *103*, 101972. [CrossRef]
18. Wu, W.; Zheng, Y.; Cao, N.; Zeng, H.; Ni, B.; Qu, H.; Ni, L.M. MobiSeg: Interactive region segmentation using heterogeneous mobility data. In Proceedings of the 2017 IEEE Pacific Visualization Symposium (PacificVis), Seoul, Republic of Korea, 18–21 April 2017; pp. 91–100.
19. Chawla, S.; Yu, Z.; Hum, J. Inferring the Root Cause in Road Traffic Anomalies. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, 10–13 December 2012; IEEE: Brussels, Belgium, 2012; pp. 141–150.
20. Zhu, X.; Guo, D. Mapping Large Spatial Flow Data with Hierarchical Clustering. *Trans. GIS* **2014**, *18*, 421–435. [CrossRef]
21. Liu, Y.; Tong, D.; Liu, X. Measuring Spatial Autocorrelation of Vectors. *Geogr. Anal.* **2015**, *47*, 300–319. [CrossRef]
22. Tao, R.; Thill, J.C. BiFlowLISA: Measuring spatial association for bivariate flow data. *Comput. Environ. Urban Syst.* **2020**, *83*, 101519. [CrossRef]
23. Yao, Y.; Li, X.; Liu, X.; Liu, P.; Liang, Z.; Zhang, J.; Mai, K. Sensing Spatial Distribution of Urban Land Use by Integrating Points-of-Interest and Google Word2Vec Model. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 825–848. [CrossRef]
24. Gao, S.; Janowicz, K.; Couclelis, H. Extracting Urban Functional Regions from Points of Interest and Human Activities on Location-based Social Networks. *Trans. GIS* **2017**, *21*, 446–467. [CrossRef]
25. 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.* **2019**, *33*, 134–154. [CrossRef]
26. Xing, X.; Yuan, Y.; Huang, Z.; Peng, X.; Zhao, P.; Liu, Y. Flow Trace: A Novel Representation of Intra-Urban Movement Dynamics. *Comput. Environ. Urban Syst.* **2022**, *96*, 101832. [CrossRef]
27. Wang, J.; Lu, F.; Liu, S. A Classification-Based Multifractal Analysis Method for Identifying Urban Multifractal Structures Considering Geographic Mapping. *Comput. Environ. Urban Syst.* **2023**, *101*, 101952. [CrossRef]
28. Zhang, T.; Duan, X.; Li, Y. Unveiling Transit Mobility Structure towards Sustainable Cities: An Integrated Graph Embedding Approach. *Sustain. Cities Soc.* **2021**, *72*, 103027. [CrossRef]
29. Jia, T.; Yu, X.; Li, X.; Qin, K. Identification and Analysis of Urban Influential Regions Using Spatial Interaction Networks. *Trans. GIS* **2021**, *25*, 2821–2839. [CrossRef]
30. Kang, C.; Jiang, Z.; Liu, Y. Measuring Hub Locations in Time-Evolving Spatial Interaction Networks Based on Explicit Spatiotemporal Coupling and Group Centrality. *Int. J. Geogr. Inf. Sci.* **2022**, *36*, 360–381. [CrossRef]
31. Sobolevsky, S.; Szell, M.; Campari, R.; Couronné, T.; Smoreda, Z.; Ratti, C. Delineating Geographical Regions with Networks of Human Interactions in an Extensive Set of Countries. *PLoS ONE* **2013**, *8*, e81707. [CrossRef]
32. Zhong, C.; Arisona, S.M.; Huang, X.; Batty, M.; Schmitt, G. Detecting the dynamics of urban structure through spatial network analysis. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 2178–2199. [CrossRef]

33. Cao, W.; Dong, L.; Cheng, Y.; Wu, L.; Guo, Q.; Liu, Y. Constructing Multi-Level Urban Clusters Based on Population Distributions and Interactions. *Comput. Environ. Urban Syst.* **2023**, *99*, 101897. [CrossRef]
34. McKenzie, G.; Daniel, R. Measuring Urban Regional Similarity through Mobility Signatures. *Comput. Environ. Urban Syst.* **2021**, *89*, 101684. [CrossRef]
35. Zhou, M.; Yue, Y.; Li, Q.; Wang, D. Portraying Temporal Dynamics of Urban Spatial Divisions with Mobile Phone Positioning Data: A Complex Network Approach. *ISPRS Int. J. Geo-Inf.* 2016; 5, 240.
36. Blei, D.M.; Andrew, Y.N.; Michael, I.J. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
37. Rubner, Y. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121. [CrossRef]
38. Cao, J.; Xia, T.; Li, J.; Zhang, Y.; Tang, S. A Density-Based Method for Adaptive LDA Model Selection. *Neurocomputing* **2009**, *72*, 1775–1781. [CrossRef]
39. Kohonen, T. Essentials of the Self-Organizing Map. *Neural Netw.* **2013**, *37*, 52–65. [CrossRef] [PubMed]
40. Mimno, D.; Wallach, H.; Talley, E.; Leenders, M.; McCallum, A. Optimizing semantic coherence in topic models. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 262–272.
41. Dunn, J.C. Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **1974**, *4*, 95–104. [CrossRef]
42. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
43. Davies David, L.; Donald, W. Bouldin. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* 1979; PAMI-1, 224–227.
44. Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. On Clustering Validation Techniques. *J. Intell. Inf. Syst.* **2001**, *17*, 107–145. [CrossRef]
45. Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. Understanding of Internal Clustering Validation Measures. In Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Sydney, Australia, 13 December 2010; pp. 911–916.