

HEProOE: A Hyperedge Enhanced Probabilistic Optimal Estimation method for detecting Spatial Fuzzy Communities

Xiao He¹, Zhongan Tang^{2,3}, Baoju Liu^{1,3,*}, Jia Duan^{2,3}, and Min Deng^{1,3}

¹ Department of Geo-informatics, Central South University, Changsha, 410083, China

² The Third Surveying and Mapping Institute of Hunan Province, Changsha 410083, China

³ Hunan Geospatial Information Engineering and Technology Research Center, Changsha 410083, China

* baojuliu@csu.edu.cn

Abstract

Identifying spatial communities with human mobility data has emerged as a key approach to understanding urban spatial structure. However, relying solely on human mobility data to partition spatial communities ignores the semantic information and may fragment large, semantic consistent Indivisible Regions (IRs) such as college campuses. Furthermore, individual spatial units often belong simultaneously to multiple IRs, creating membership uncertainty, while the spatial stochasticity of human movements inherently introduces ambiguity to the boundaries of spatial fuzzy communities. To address these challenges, we proposed the Hyperedge-Enhanced Probabilistic Optimal Estimation method (HEProOE) that integrated the hyperedge into spatial fuzzy community detection, representing IRs as semantic consistent regions. First, IRs were represented as hyperedges, where each spatial unit holds a probabilistic community membership. Second, a novel distance-weighted Jensen-Shannon (JS) divergence metric was introduced to measure the semantic consistency within each hyperedge. Finally, this metric was converted into a new likelihood component and seamlessly integrated with the mobility-based ProOE model, yielding a unified framework that simultaneously optimizes for both mobility patterns and semantic consistency. Experimental results demonstrated that HEProOE uncovers spatial fuzzy communities with significantly higher semantic consistency, providing an effective tool for a more authentic understanding of urban spatial structures.

Introduction

Urban spatial structures fundamentally shape human mobility patterns, as individuals traverse different zones to access diverse activity destinations^{1,2}. These movement trajectories aggregate into macroscopic spatial communities at the urban scale³. Consequently, detecting these communities from human mobility data has become a critical analytical strategy for deciphering urban spatial structure^{3,4}. The resulting community structures can be used to analyze the dynamic interactions between urban functional areas⁵, identify important areas of human activity⁶. Furthermore, the identified boundaries can be used to help differentiate urban policies and facility site selection^{7,8}.

However, methods that rely solely on mobility data often ignore the semantic consistency between spatial units, resulting in inconsistent semantics of detected communities. IRs are spaces that have been given distinct character through human experience, function, and identity^{9,10}. For example, a university campus, a financial district, or a residential neighborhood are not merely collections of coordinates, they are artificially defined priori regional boundaries according to their semantic attributes. Integrating IRs is therefore essential to ensure semantic consistency in the detected spatial communities.

This recognition has driven a divergence in research on spatial community detection^{11–13}. Current algorithms can be categorized into two paradigms based on whether they incorporate semantic attributes or not: (1) Methods based solely on human mobility, and (2) Hybrid methods, which incorporate both mobility and semantic attributes.

Methods based solely on human mobility

Methods based solely on human mobility are among the most widely used for identifying spatial communities, with the aim of deeply reflecting the urban spatial structure from the perspective of human mobility. Based on the structure of mined spatial communities, this category can be further divided into non-overlapping, overlapping, and fuzzy communities¹⁴.

Non-overlapping community detection methods are well-established in non-geographic contexts. Initially, many algorithms were developed for this context, such as those based on modularity optimization (e.g., Louvain method), information theory (e.g., Infomap)^{15–17} or graph neural networks¹⁸, were directly applied to spatial networks. However, these non-geographic methods often produce geographically fragmented and scattered communities, as they do not account for spatial proximity and continuity constraints. To address this, researchers have proposed various spatial adaptations. Some studies modified the modularity function by integrating a gravity model, which penalizes connections between distant spatial units, thereby promoting geographically continuous communities^{19,20}, or by using inverse distance weighting to prioritize connections between neighboring spatial units²¹. Other approaches enforced spatial contiguity more explicitly through specialized algorithms. For instance, the Spatial Tabu Optimization for Community Structure (STOCS) method employs a continuity-constrained Tabu optimization search to enhance robustness against data noise²². Similarly, the Density and Adjacency Expansion-Based Spatial Structural Community Detection Algorithm (DASSCAN) identifies communities by expanding from dense core nodes to adjacent ones based on structural similarity²³, while ant colony optimization-based spatial scan statistic (ACOScan) leverages contiguity-constrained ant colony optimization to group interconnected road segments²⁴. These techniques have significantly improved the geographical contiguity of non-overlapping spatial communities. Despite these advancements, the non-overlapping paradigm, which assigns each spatial unit to a single community, struggles to capture the complex reality of urban spaces where a location can serve multiple functions and belong to different activity zones.

To better represent this functional multiplicity, overlapping community detection methods were introduced, allowing spatial units to belong to more than one community^{25–27}. Recent approaches have refined this concept. For example, the Spatial Modularity and Edge Similarity (SMES) method identifies spatially cohesive communities by combining spatial modularity with edge similarity²⁸, while the link-based method integrates geographical weighting (GWCD) to effectively capture spatially overlapping regions²⁹.

While offering a more realistic depiction, a key limitation persists in this paradigm: most overlapping methods provide only a binary (member or non-member) assignment, failing to quantify the varying degrees of association. Recognizing that the stochasticity of human mobility and the transitional nature of urban functions create inherently ambiguous boundaries, the concept of fuzzy communities has emerged. Fuzzy methods assign each spatial unit a probability distribution of membership across all communities, thereby capturing the uncertainty and strength of its affiliation. To date, this area remains relatively underexplored in spatial contexts. The Probabilistic Optimal Estimation (ProOE) model is a notable exception, formally quantifying this fuzziness within a probabilistic framework³⁰. Nevertheless, even these advanced models primarily rely on pairwise interactions, potentially overlooking the semantic consistency of larger functional entities.

Hybrid methods

While methods based solely on human mobility excel at revealing the structure of human flows, they inherently lack semantic attributes, making it difficult to interpret the underlying functions of the detected communities^{13,31}. To bridge this gap, hybrid methods have been developed, which integrate mobility patterns with auxiliary semantic data, most commonly Points of Interest (POIs), to produce communities that are both structurally sound and functionally coherent³².

Early approaches sought to explain or label mobility-derived communities in a post-hoc manner. For example, researchers first detected communities from mobility flows and then employed statistical models like logistic regression to quantify the explanatory power of different POI categories in shaping these communities³¹ or simply used POI profiles to assign functional labels such as "commercial" or "residential" to the identified zones³². Subsequently, the paradigm shifted from post-hoc explanation towards a more integrated approach where semantic consistency actively guides the community detection process. This evolution involved moving beyond static POI counts to analyzing the semantics of entire trajectories. Methods like TODMIS (Trajectory cOmmunity Discovery using Multiple Information Sources) were developed to model individual trips as semantic sequences, enabling the discovery of communities based on behavioral similarity rather than mere spatial proximity³³. Other studies constructed composite networks where edge weights were a function of both human mobility and semantic similarity, calculated using techniques like ontology, thereby embedding functional context directly into the network structure³⁴. More recent studies have advanced this paradigm further by explicitly modeling complex urban semantics and their temporal dynamics^{35,36}.

Despite these advances, a critical limitation persists in how current methods model the functional and semantic structure of cities. The prevailing approaches rely on complex, bottom-up procedures—such as aggregating discrete POI data or applying topic models—to infer semantic meaning for small, arbitrary spatial units. This process not only introduces methodological complexity but, more importantly, it overlooks the powerful, top-down structural information inherent in the urban landscape.

Large-scale entities like university campuses, financial districts, or major parks are not mere statistical aggregates of micro-level data; they are established, human-defined functional wholes that directly shape mobility. By treating them as collections of independent cells, existing models disregard their intrinsic coherence. In network science, these multi-unit, higher-order relationships are naturally represented by hyperedges³⁷⁻³⁹. While hypergraph-based methods like Hypergraph-MT exist for community detection, they are fundamentally non-spatial and ill-suited for geographic analysis⁴⁰.

This reveals a critical gap in the literature, which is divided between two incomplete perspectives. On one hand, spatial models (like ProOE) operate on pairwise interactions but lack a mechanism to enforce the semantic integrity of known urban districts, often fracturing them illogically. On the other hand, generic hypergraph models can represent higher-order groups but fail in urban contexts because they ignore fundamental geographic principles like distance decay and spatial proximity.

To address these gaps, this paper introduced a novel framework that conceptualizes IRs as hyperedges within a spatial network. In this formulation, an IR is defined as a set of constituent spatial units. This allows us to enforce semantic consistency by ensuring that spatial units within the same IR share a similar community membership. We design a distance-weighted JS divergence metric to measure this semantic consistency within the hyperedge (IR), which is then transformed into a likelihood function. This function is seamlessly integrated with an existing mobility-based probabilistic model, creating a unified framework that balances mobility flows with semantic consistency. This approach yields the following key contributions:

(1). A Unified Probabilistic Framework for Hybrid Community Detection: We developed a principled hybrid method to integrate the structure of IRs into a fuzzy, probabilistic community detection framework. This framework explicitly considers both the interaction structure of human mobility and the semantic consistency between spatial units. By formulating the semantic consistency as a likelihood component, our model elegantly balances the influence of human-defined IRs with the organic, fuzzy community structures revealed by mobility data, leading to the identification of communities with significantly clearer, more stable, and more realistic boundaries.

(2). A Novel Hypergraph Representation for IRs: We introduced a new paradigm for encoding semantic consistency by modeling IRs as hyperedges. This direct and interpretable method surpasses bottom-up topic modeling and naturally represents complex spatial realities like overlapping and nested IRs, offering a fundamentally new perspective on integrating semantics into spatial network analysis and enabling the framework to produce partitions that are more interpretable and better aligned with human-perceived urban geography.

Methods

To capture the complex, multi-unit nature of IRs such as university campuses or financial districts, this study models them as hyperedges. An IR's functional indivisibility stems from the strong semantic consistency among its constituent spatial units, driven by factors like information sharing or capital flow. Traditional graphs, which rely on binary edges, are limited to representing pairwise relationships and are thus less suited for capturing this group-level coherence. By employing hyperedges—a concept from hypergraph theory that allows a single edge to connect multiple nodes—the collective semantic integrity of an IR can be effectively represented. This representation is crucial for addressing a common issue where conventional community detection algorithms might fragment these functionally unified regions into smaller, disconnected communities.

The framework for delineating urban spatial communities is founded on two core principles that link human mobility with semantic regions. The first principle is that trips are more likely to occur between spatial units that belong to the same spatial community. The second is that spatial units within the same

IR should exhibit highly similar community membership distributions due to the IR's semantic influence. These principles are translated into two core computational objectives: 1) transforming the probabilistic community membership of spatial units into an estimation of trip volumes between them, and 2) enforcing semantic consistency by maximizing the similarity of community membership distributions for all units within a single IR. By jointly optimizing these two objectives, the model is designed to generate a more accurate and meaningful delineation of spatial communities.

As illustrated in Figure 1, the methodology begins by representing the underlying community structure of urban space through these two primary components: human mobility and semantic consistency. A Spatial Trip and Semantic Graph (STSG) is then constructed, where spatial units are represented as nodes, human mobility is modeled using binary edges, and IRs are modeled using hyperedges. Central to this framework is a probabilistic model where each spatial unit's affiliation to various communities is represented as a community membership distribution. For the human mobility component, the modeling process follows the procedure of the Probabilistic Optimal Estimation (ProOE) model³⁰. ProOE estimates trip volumes by adjusting the community membership distribution of each spatial unit, an approach that effectively captures the ambiguity in the interplay between spatial units and communities. Analogously, for the semantic component, the consistency within an IR is quantified by adjusting the community membership distributions of all units within its corresponding hyperedge to maximize their similarity.

Finally, the model seeks a joint solution by optimizing the community membership distributions to satisfy both mobility and semantic constraints. The first objective, following the ProOE framework, is to adjust these distributions until the estimated trip volumes closely approximate the real, observed trip volumes. Concurrently, the second objective is to maximize the similarity of membership distributions for all spatial units within each hyperedge, thereby reinforcing the semantic consistency of the IRs. To manage the trade-off between these two objectives, a semantic weight parameter is introduced. This parameter allows for balancing the influence of trip data against the structural constraint of semantic coherence, enabling a flexible and robust delineation of urban spatial communities.

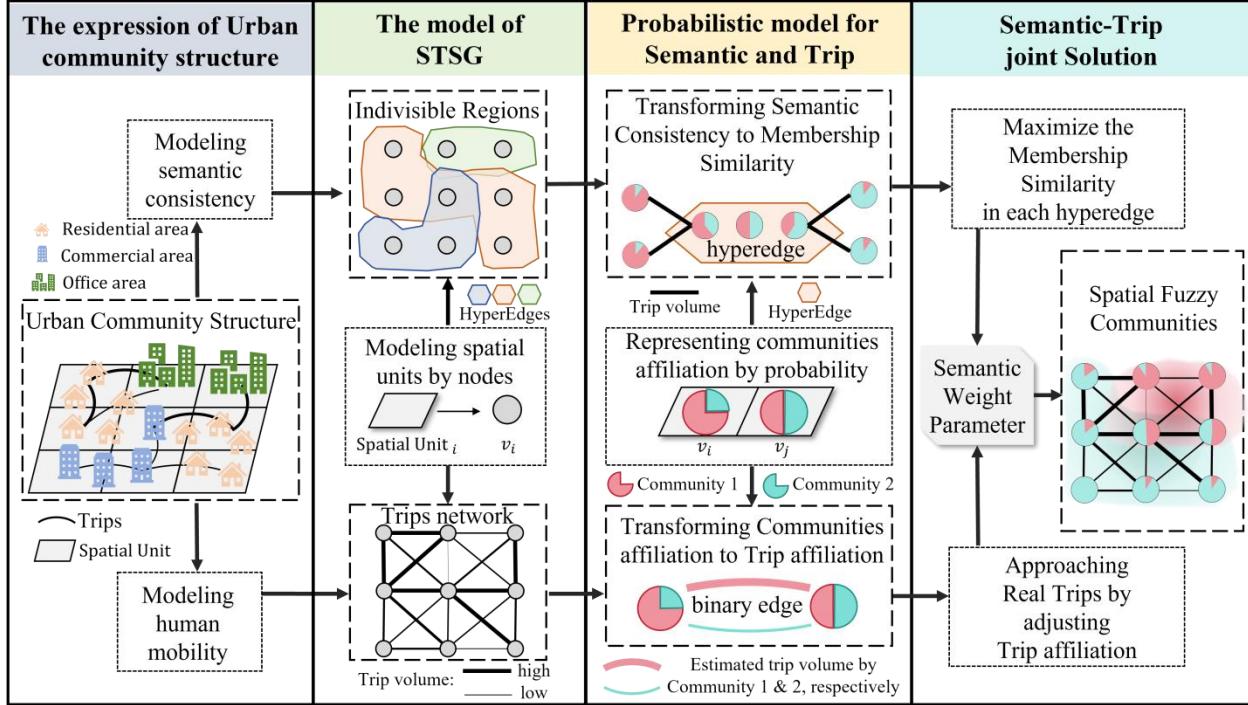


Figure 1. Illustration of the proposed Hyperedge-Enhanced Probabilistic Optimal Estimation method. The pie chart shows the spatial units' membership to different communities. The colors represent different communities.

Modeling Semantic Consistency with Hyperedges

The HEProOE framework was built upon a novel representation of urban space that explicitly models the semantic consistency of IRs. STSG was defined as $G=(V, E, H, R)$, where V is the vertex set, containing all spatial units; E is the edge set, containing all trips; H is the hyperedge set, containing all IRs; R is the weight set, containing all trip volumes. Our method is defined by four core components:

Spatial Units are modeled as Nodes: The city is partitioned into a set of N discrete spatial units (e.g., TAZs), which are represented as nodes $v_i (i \in N, v_i \in V)$ in a network.

Trips are modeled as Edges: an OD pair (origin and destination pair for a trip) from vertex v_i to vertex v_j is regarded as edge $e_{ij} (e_{ij} \in E)$.

IRs are modeled as Hyperedges: An IR (e.g., university campus) is modeled as a hyperedge h . A hyperedge is a generalization of a standard edge that can connect any number of nodes. Here, it groups all the spatial units $\{v_1, \dots, v_i, \dots\} (v_i \in h)$ that fall within the boundaries of that IR, formally representing it as a single, cohesive entity.

Fuzzy Memberships are modeled as Node Attributes: Following the principles of fuzzy community detection, each spatial unit i is assigned a probabilistic membership vector $u_i = \{u_{i1}, u_{i2}, \dots, u_{ik}, \dots, u_{iK}\}$, where u_{ik} is the probability that unit i belongs to community k , and $\sum_K u_{ik} = 1$.

To capture the semantic consistency within IRs, this study measured the similarity between the community membership probability distributions of its constituent spatial units. A critical requirement for this measure is symmetry, as the consistency between two units should be independent of the order of comparison. We therefore employ the Jensen-Shannon (JS) divergence. As a symmetric and bounded variant of the Kullback-Leibler (KL) divergence, JS divergence provides a principled and numerically stable method for this task. It allows our model to effectively penalize pairs of spatial units within the same Identified Region (IR) that exhibit highly divergent community memberships, thereby enforcing semantic coherence. For spatial units i and j , their JS divergence can be defined as:

$$JS(i,j) = \sum_k^K \frac{u_{ik} \ln \frac{u_{ik}}{0.5(u_{ik}+u_{jk})} + u_{jk} \ln \frac{u_{jk}}{0.5(u_{ik}+u_{jk})}}{2} \quad (1)$$

To express the assumption that the JS divergence between spatial units within the entire hyperedge should be as small as possible, the average JS divergence among all spatial units in the IR is calculated. To model the decay of node influence within the hyperedge with spatial distance, an inverse distance function is incorporated into the model as a distance decay factor. Finally, the above results are transformed into probabilities, ultimately obtaining the likelihood probabilities for all hyperedges, with the results as follows:

$$P(H|u) = \prod_h^H e^{-\frac{\sum_{i \in h} \sum_{j \in h, j \neq i} \frac{JS(i,j)}{d_{ij}}}{\frac{D_h(D_h-1)}{2}}} \quad (2)$$

Where d_{ij} is the distance between spatial units v_i and v_j .

Hyperedge-Enhanced Probabilistic Optimal Estimation model

Based on the same framework to express probabilistic membership between spatial units and spatial communities in ProOE, it models the membership to community for a certain trip by $u_{ik}u_{jk}$. Furthermore, to more accurately model human mobility in cities, ProOE introduces an intensity parameter (I) and a probability density function (PDF), to model the effect of spatial heterogeneity (the difference in the intensity of the internal trips of different communities) and trip distance volume distributions. Putting these elements together, we can model the estimated interaction volume:

$$\lambda_{ij} = \sum_k^K u_{ik}u_{jk}I_k PDF(d_{ij}) \quad (3)$$

Regarding the real trip volume R_{ij} , as a random event under the current community, it can be given as $E(R_{ij}) = \lambda_{ij}$. Thus, assumed that the real trip volume R_{ij} and the estimated trip volume λ_{ij} , follow the Poisson distribution, i.e., $R_{ij} \sim Poisson(\lambda_{ij})$. The likelihood of this distribution can then be defined as follows:

$$P(R|u,I,PDF) = \prod_{i,j}^N e^{-\lambda_{ij}} \frac{\lambda_{ij}^{R_{ij}}}{R_{ij}!} \quad (4)$$

where $R_{ij}!$ represents the factorial of R_{ij} . Combing likelihood of trips and IRs together, we can obtain the total likelihood as follows:

$$P(R,H|u,I,PDF) = P(R|u,I,PDF) + \alpha \times P(H|u) \quad (5)$$

where α is a semantic weight parameter used to balance trip and IR likelihood. By maximizing the above formula, we can obtain the optimal membership.

Model Optimization using the EM Algorithm

To solve this model, we utilized an expanded EM algorithm introduced in ProOE. To simplify the computation, we first convert formula (5) to log-likelihood:

$$\ln P(R,H|u,I,PDF) = \ln P(R|u,I,PDF) + \alpha \times \ln P(H|u) \quad (6)$$

To maintain computational and formal simplicity, α is still extracted as a separate parameter. $\ln P(R|u,I,PDF)$ and $\ln P(H|u)$ are respectively:

$$\ln P(R|u,I,PDF) = \sum_{i,j}^N R_{ij} \ln \lambda_{ij} - \sum_{i,j}^N \lambda_{ij} - \sum_{i,j}^N R_{ij}! \quad (7)$$

$$\ln P(H|u) = \sum_h^H - \frac{\sum_{i \in h} \sum_{j \in h, j \neq i} \frac{\sum_k^K u_{ik} \ln \frac{2 \times u_{ik}}{(u_{ik} + u_{jk})} + u_{jk} \ln \frac{2 \times u_{jk}}{(u_{ik} + u_{jk})}}{d_{ij}}}{2 \times D_h (D_h - 1)} \quad (8)$$

The Joint Jensen's inequality is as follows:

$$\ln \sum_k^K X_{ijk} \geq \sum_k^K q_{ijk} \ln \frac{X_k}{q_{ijk}} \quad (9)$$

where q_{ijk} is the probability of satisfying $\sum_k^K q_{ijk} = 1$. By combining the above equations and calculating the partial derivatives of the optimal parameters following the original ProOE³⁰. We arrived at the iterative update formulas for the EM algorithm.

(1) E-Step: Calculate the expected value of the hidden variables:

$$q_{ijk} = \frac{u_{ik} u_{jk} I_k PDF(d_{ij})}{\sum_k^K u_{ik} u_{jk} I_k PDF(d_{ij})} \quad (10)$$

where q_{ijk} represents the probability that trip e_{ij} is generated by community k .

(2) M-Step: Maximize the log-likelihood function and update the model parameters:

$$I_k = \frac{\sum_{i,j}^N R_{ij} q_{ijk}}{\sum_{i,j}^N u_{ik} u_{jk} PDF(d_{ij})} \quad (11)$$

Differentiating with respect to u_{ik} yields the update formula for community membership that considers semantic consistency. Depending on the membership relationship between node i and the hyperedge, it can be divided into three cases:

When node i does not belong to any hyperedge:

$$u_{ik} = \frac{\sum_{i,j}^N R_{ij} q_{ijk}}{I_k \sum_k^K u_{ik} PDF(d_{ij})} \quad (12)$$

When node i belongs to one or some hyperedges and has no interaction with the outside:

$$u_{ik} = \frac{1}{2} e^{-\frac{\sum_{h \in H, i \in h} \frac{1}{2 \times D_h(D_h-1)} \sum_{j \in h, j \neq i} \frac{\ln(u_{ik} + u_{jk})}{d_{ij}}}{\sum_{h \in H, i \in h} \frac{1}{2 \times D_h(D_h-1)} \sum_{j \in h, j \neq i} \frac{1}{d_{ij}}}} \quad (13)$$

When node i belongs to one or some hyperedges and has interaction with the outside:

$$u_{ik} = \frac{\sum_{i,j}^{ij \in E} R_{ij} q_{ijk}}{\sum_k^K I_k u_{jk} PDF(d_{ij}) + \alpha \times \sum_{h \in H, i \in h} \frac{1}{2 \times D_h(D_h-1)} \sum_{j \in h, j \neq i} \frac{1}{d_{ij}} \ln \frac{2 \times u_{ik}}{(u_{ik} + u_{jk})}} \quad (14)$$

Thus, the solution to the problem in formula (5) is converted to iterative updates in formula (10)–formula (14). Repeating this process allowed us to obtain optimal estimates for parameters u , q , and I .

Study area and datasets

New York City (NYC) was selected as the study area, representing a quintessential global metropolis with a dense population and a complex network of social and economic interactions, which has fostered the development of numerous well-established communities, each with distinctive characteristics in different areas (Figure 2a). Manhattan is rich with such well-defined IRs. For instance, the Financial District in Lower Manhattan constitutes a finance-dominated community, while Midtown Manhattan emerges as a commercial-cultural community (Figure 2c). This clear organization of IRs provides a strong basis for evaluating the model's ability to capture semantically meaningful urban structures.

Two primary datasets were utilized. First, IRs (hyperedges) were sourced from OpenStreetMap (OSM). We extracted all polygon features in NYC tagged with the 'place' key, which OSM uses to denote named geographical entities. From these, 18 prominent polygons with semantically explicit names, such as "Upper West Side" or "Lenox Hill," were selected to serve as the predefined IRs in our model (Figure 2c). Second, human mobility data were obtained from the New York City Taxi and Limousine Commission (TLC). We aggregated 8,328,114 trips from public yellow and green taxi records from June to August 2023. To construct the mobility network, these trips were georeferenced and aggregated between NYC's

262 official spatial units (Traffic Analysis Zones, TAZs) (Figure 2a), forming a weighted, undirected network where nodes are spatial units and edge weights represent the total trip volume. Figure 2b visualizes the overall trip in NYC.

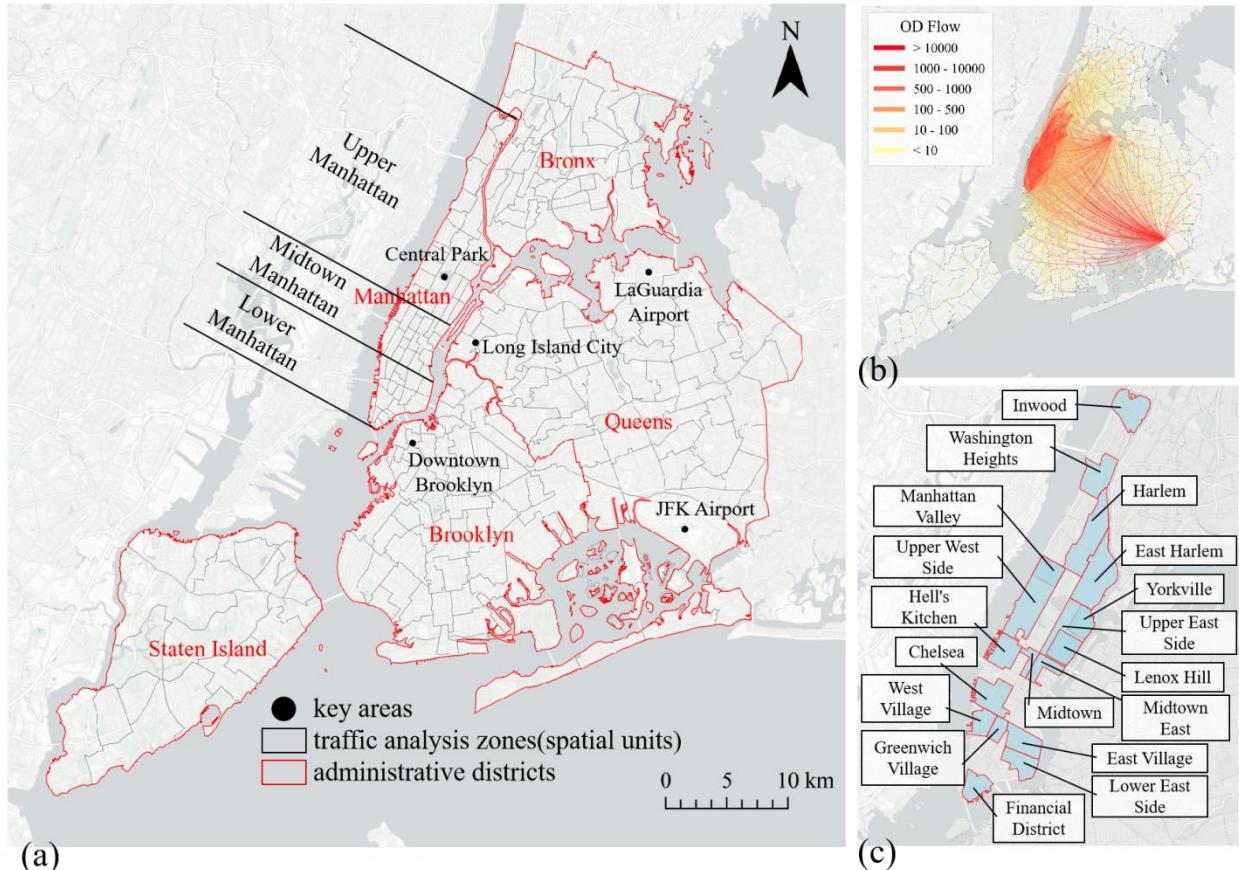


Figure 2. Overview of study area and data. (a) Overview of the research area. (b) Trip data for New York City. (c) IRs obtained from OpenStreetMap [Data © OpenStreetMap contributors; licensed under ODbL 1.0; opendatacommons.org/licenses/odbl/].

Results

The HEProOE model incorporates two crucial parameters: α , which balances the weighting between the mobility-based interaction component and the semantic consistency component, while K is the number of communities to detect. The influence of these two parameters is systematically investigated in the Section: Parameter Sensitivity Analysis. To mitigate sensitivity to initialization, all experiments were conducted with 5 independent runs using randomized starting points. The solution achieving the maximum log-likelihood was retained as the optimal result.

Spatial fuzzy community detection results

Figure 3a presents the spatial fuzzy communities across NYC as detected by HEProOE. A mixed-color shading technique was employed to represent the fuzzy nature of community affiliations, with each spatial unit reflecting its membership in different communities³⁰. A key finding is the model's performance in Manhattan. Notably, Manhattan is partitioned into several distinct major IRs, including Lower Manhattan (Comm_4), Midtown Manhattan (Comm_2, Comm_3), and Upper Manhattan (Comm_1, Comm_5). This division accurately reflects the integrity of major IRs, demonstrating that HEProOE effectively integrates semantic consistency to produce meaningful communities. Beyond Manhattan, the model also identifies other significant urban communities, such as those centered around Downtown Brooklyn and Long Island City, which are characterized by high population density and intense human mobility.

To quantitatively assess the quality of the detected communities, we employed two metrics from the ProOE model³⁰: the Confidence Index (ConI), which measures the internal coherence of a whole community, and the Certainty Index (CerI), which assesses the membership certainty of each individual spatial unit. As illustrated in Figure 3b, communities are ordered by their ConI scores. The Manhattan-based communities (Comm_1–Comm_5) exhibit the highest ConI values, underscoring Manhattan's central role in the city's mobility network. Specifically, Comm_1 (Upper East side) and Comm_2 (Midtown) achieve the highest ConI scores, reflecting their status as dominant residential, cultural, and commercial communities. The spatial pattern of CerI (Figure 3c) reveals high assignment certainty in the cores of local communities, such as Lower Manhattan and Downtown Brooklyn, reinforcing their status as stable anchors. Conversely, areas with low certainty—notably Midtown Manhattan, the Upper East Side, and major transportation hubs like LaGuardia and JFK Airports—are characterized by diverse, city-wide travel patterns. Their ambiguous membership appropriately reflects their real-world function as connectors that serve the entire city rather than belonging to a single, localized community. This ambiguity also explains their sensitivity to parameter changes, as explored in Section: Parameter Sensitivity Analysis.

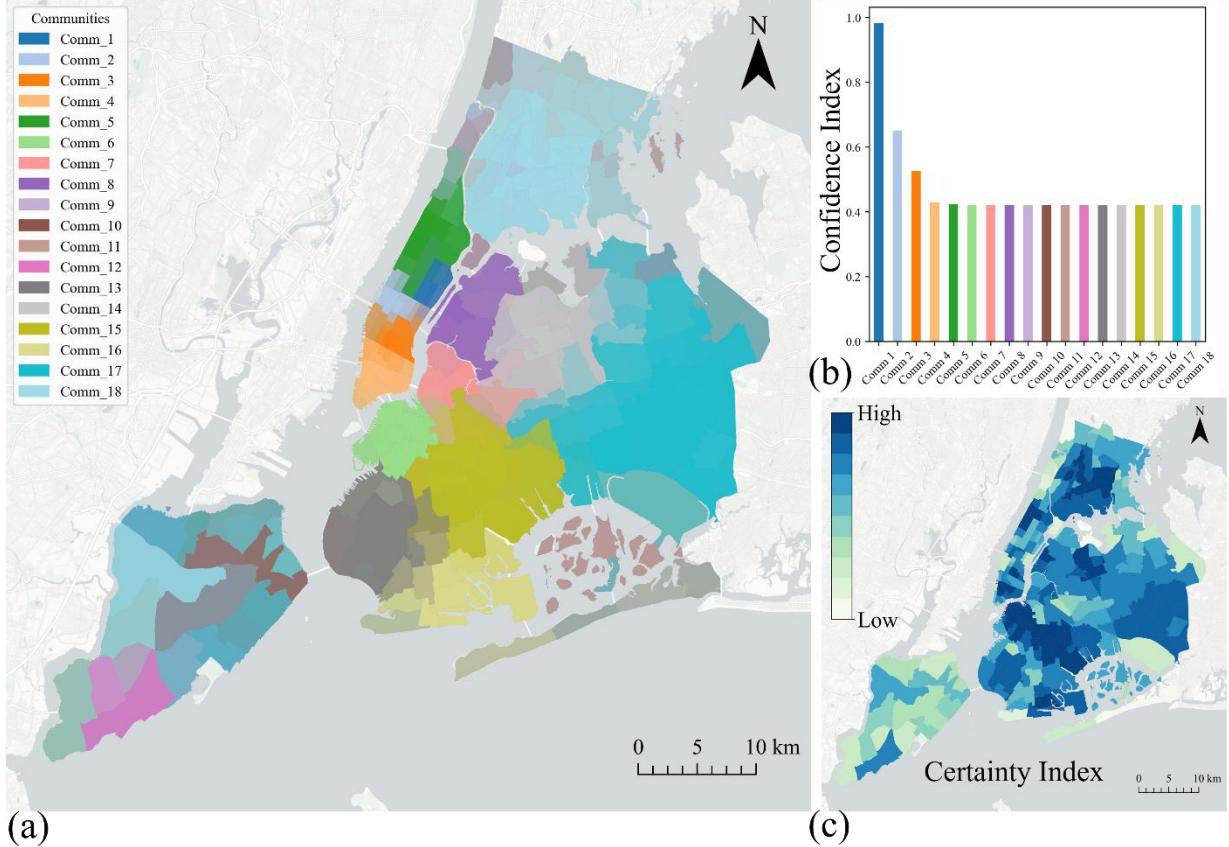


Figure 3. Detection result based on HEProOE. (a) Spatial fuzzy community. (b) Confidence Index (ConI). (c) Certainty Index (CerI).

Comparative experiment

To evaluate the effectiveness of our proposed semantic consistency component, we compared our HEProOE model against two baselines: its base model, ProOE, and Hypergraph-MT, a state-of-the-art method for higher-order community detection⁴⁰. The results clearly demonstrate the limitations of models that overlook either semantic coherence or spatial context.

The first baseline, ProOE, relies exclusively on pairwise mobility interactions and consequently fails to preserve the integrity of large, semantically coherent IRs. As shown in Figure 4b, this leads to the fragmentation of well-known districts like the Upper East Side and Hell's Kitchen, which are either fractured into smaller communities or assigned to ambiguous, low-certainty zones. This issue is fundamental to its design: without a mechanism to recognize IRs as unified entities, ProOE cannot distinguish between internal and external mobility flows, causing it to split cohesive neighborhoods that interact with multiple external areas.

The second baseline, Hypergraph-MT, also produces unsatisfactory results, albeit for different reasons. While it can model higher-order interactions, it lacks mechanisms to account for spatial effects like spatial heterogeneity and distance decay. When applied to urban data, it disproportionately focuses on areas with

extreme interaction volumes. This results in the over-partitioning of Manhattan into a dense cluster of small, indistinguishable communities, as seen in Figure 4c.

In stark contrast, HEProOE successfully integrates both mobility patterns and semantic information. By representing IRs as hyperedges, it enforces semantic consistency, ensuring that spatial units within a predefined region (such as the Upper East Side) are grouped into the same community. The resulting partition, depicted in Figure 4a, features sharp, meaningful boundaries that preserve the structural integrity of key districts. This outcome aligns far more closely with the city's established functional and perceived geography. This comparison validates our approach, showing that by integrating semantic consistency, HEProOE overcomes critical limitations inherent in models that rely on mobility data or non-spatial methods alone.

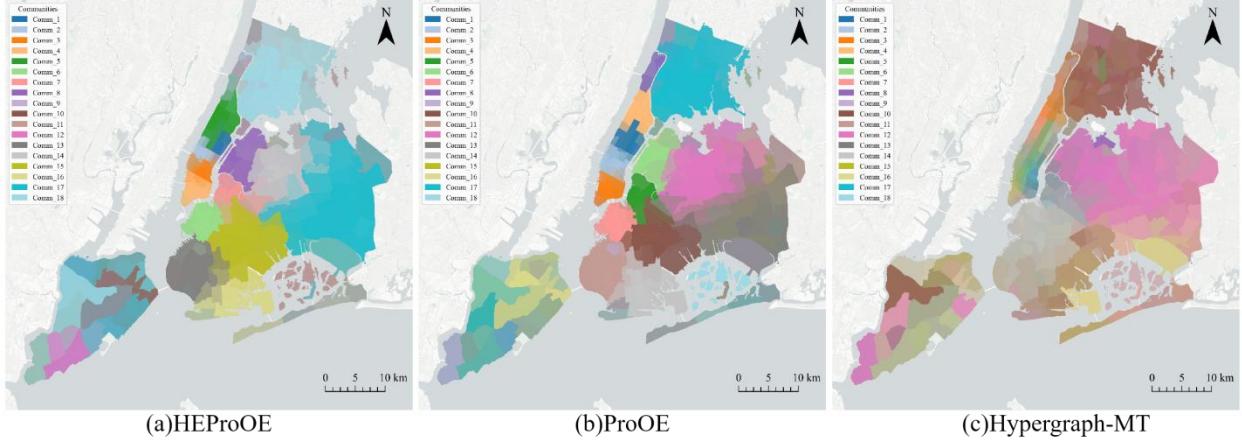


Figure 4. Comparison between community detection results from (a) HEProOE, (b) ProOE, and (c) Hypergraph-MT

To provide a quantitative basis for our comparison, we measured the alignment between each model's community partitions and the predefined Identified Regions (IRs) using Fuzzy Normalized Mutual Information (FNMI)³⁰. The results are telling: HEProOE achieved the highest FNMI score (0.440), surpassing both ProOE (0.410) and Hypergraph-MT (0.312). This score confirms that our model's partitions correspond most closely to the city's established semantic geography, a finding corroborated by visual analysis.

Parameter Sensitivity Analysis

The parameter α governs the trade-off between the log-likelihood of the interaction component and the semantic consistency component. Given that the number of trips is orders of magnitude larger than the number of semantic hyperedges, α must be sufficiently large to prevent the semantic consistency from being overshadowed. To establish a theoretically grounded starting point for α , we note that the interaction log-likelihood is dominated by the term its first term, approximating $\frac{N(N-1)}{2} \times \bar{R} \times \ln \bar{R}$ (\bar{R} is the mean value of all trip values), while the semantic consistency log-likelihood scales with the number of hyperedges, $|H|$. This suggests a theoretical relationship: $\alpha \approx \frac{N(N-1)}{2} \times \bar{R} \times \ln \bar{R} / |H|$.

However, the actual values of trip volumes (R_{ij}) can vary significantly, causing large fluctuations in the log-likelihood terms. Therefore, we introduced a scaling coefficient, β , to fine-tune the balance more effectively. The final weighting parameter is thus defined as:

$$\alpha = \beta \times \frac{N(N-1)}{2} \times \bar{R} \times \ln \bar{R} \div |H| \quad (15)$$

In our experiments, we tune β to explore the sensitivity of the model to the relative importance of semantic consistency.

The number of communities, K , is another user-defined parameter that influences the granularity of the results. Since K and β are not independent, we conducted a dual-parameter sensitivity analysis by varying both simultaneously. Figure 5 shows the model's log-likelihood across different combinations of K and β . The curves on the left and bottom of the plot show the maximum log-likelihood achieved for each value of K and β , respectively.

As the number of communities K increases, the log-likelihood gradually increases, indicating that a larger number of communities enhances the model's ability to fit the original interaction flow and IR data, which aligns with the conclusions of the original ProOE model. However, the relationship between β and log-likelihood is not simply positive correlation. As β increases, the model's log-likelihood first rises gradually and then declines. This occurs because when β is very small, the interaction component has a dominant influence on the model, weakening its ability to capture semantic consistency. As β gradually increases, the model can better capture semantic consistency during training. However, an excessively large β may cause the model to overly focus on the influence of semantic consistency, and the complex relationships—such as overlap and inclusion—between different IRs make it difficult to solve effectively. Based on the inflection points of the curves for K and β versus log-likelihood, we selected $K=18$ and $\beta=1$ as benchmarks. We further present results for different β values when $K=18$ and for different community counts K when $\beta=1$ to further illustrate the conclusions in the following sections.

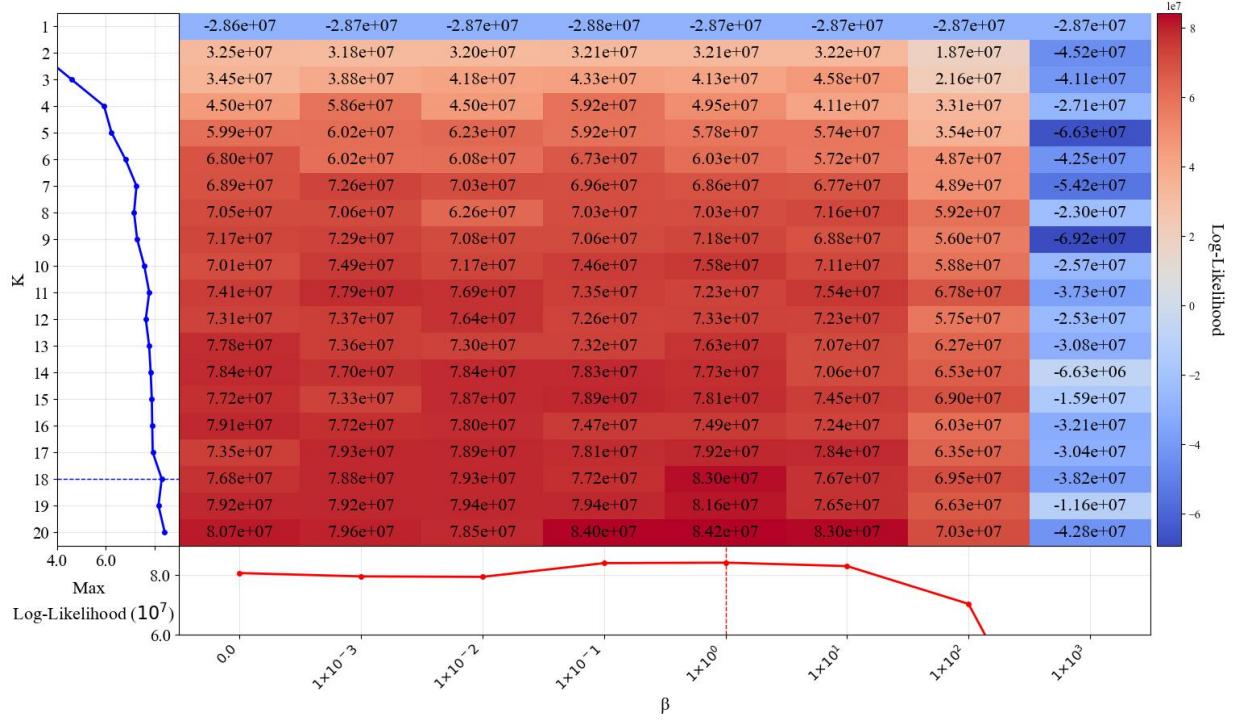


Figure 5. Fitted log-likelihood under different parameters. The color represents different log-likelihood values, while the curves on the left and bottom respectively indicate the maximum log-likelihood values when varying K and β .

Impact of semantic weight parameter

To better illustrate the conclusion, only Manhattan Island and its adjacent areas were visualized. As Figure 6 shows, at low β values, the community structure closely resembles that of the original ProOE, where boundaries are fluid and primarily dictated by mobility intensity. Due to the lack of consideration for semantic consistency within IRs, the results here divide a large number of IRs into two or more communities, such as Upper East Side and Upper West Side. As β increases, the influence of semantic consistency becomes more pronounced. The detected community boundaries become sharper and align more strictly with the predefined IRs, progressively reducing their fragmentation. For excessively high α values, the communities risk becoming simple aggregations of the input IRs, potentially ignoring strong interaction patterns that contradict the predefined semantics. The value of β used for the main results was chosen because it offered an optimal balance, maximizing the overall log-likelihood while producing a partition that is both empirically supported by mobility data and consistent with established urban IRs.

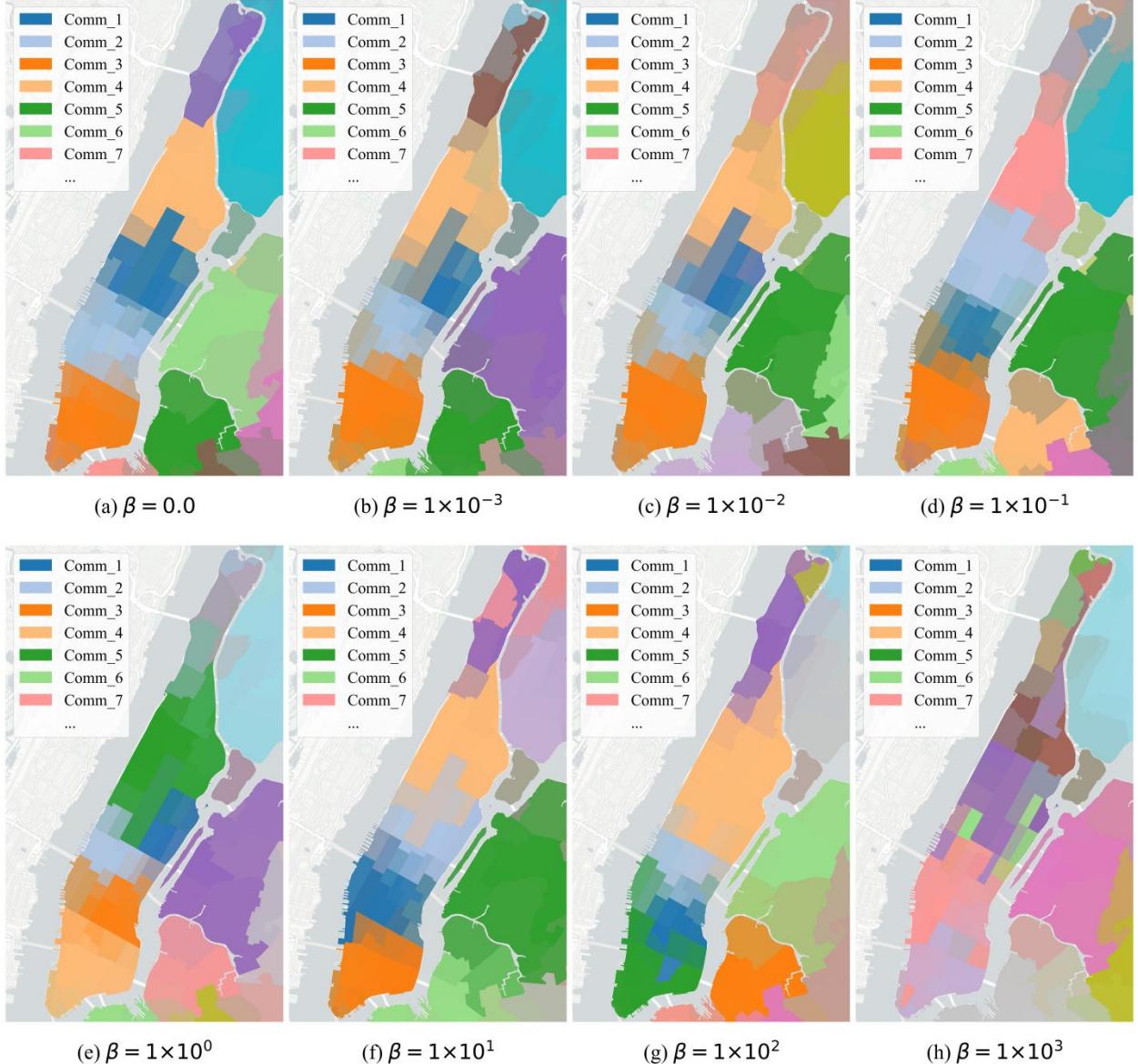


Figure 6. Comparison of community division results in the Manhattan area under the same number of communities but different β values (an additional scaling coefficient β derived from α). (a) $\beta=0.0$, (b) $\beta=1\times10^{-3}$, (c) $\beta=1\times10^{-2}$, (d) $\beta=1\times10^{-1}$, (e) $\beta=1\times10^0$, (f) $\beta=1\times10^1$, (g) $\beta=1\times10^2$ and (h) $\beta=1\times10^3$.

Impact of Community Number (K)

The number of communities K affects the granularity of model fitting. As the number of communities increases, some large communities are decomposed into smaller ones with stronger semantic consistency. For example, Comm_2 in Figure 7a, which includes Central Park as well as the Upper East Side and Upper West Side, is gradually segmented. Among them, Comm_1, being the densest and most well-known residential area in New York City, is clearly separated. Meanwhile, some regions are located in the fuzzy boundary zones of communities at a coarse granularity (e.g., the overlapping areas of Comm_1,

Comm_2, and Comm_3 in Midtown Manhattan in Figure 7a are gradually merged into independent communities (e.g., Comm_2 and Comm_3 in Figure 7e and Figure 7f). This demonstrates that the varying number of community divisions serves as an important scale variable for understanding urban spatial structure.

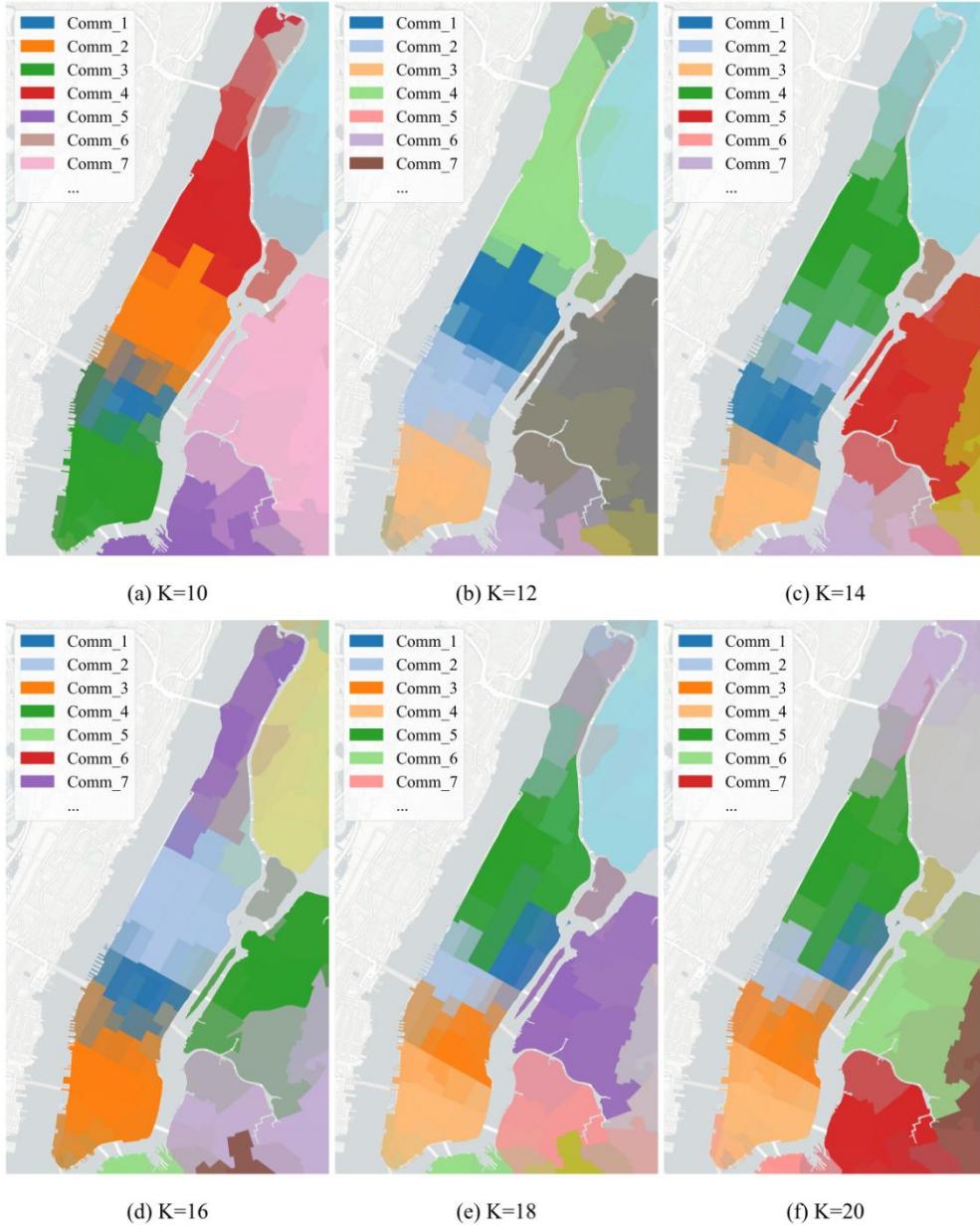


Figure 7. Comparison of community division results in the Manhattan area under the same β values but different number of communities. (a) K=10, (b) K=12, (c) K=14, (d) K=16, (e) K=18 and (f) K=20.

Instability as an Analytical Tool

The differences in community detection results caused by parameter variations indicate the instability of the community. For instance, if a community exhibits different forms under varying β parameters, it suggests potential conflicts between mobilities and semantics within the community. The semantics of certain regions are inherently local, yet they engage in strong interactions with numerous external areas, making it difficult to clearly delineate them as a single community. The model struggles to simultaneously accommodate both aspects (e.g., the fluctuations around Midtown Manhattan in Figure 6). The number of communities, K , essentially represents a scale or granularity of community division. If a community maintains a largely consistent structure across multiple scales, it signifies that the community plays a pivotal role in interactions and semantics within the local area (e.g., regions like Lower Manhattan and Long Island City in Figure 7). Therefore, sensitivity analysis of parameters can serve as a tool for urban spatial structure analysis, as the morphological changes of communities under different parameters may provide more critical information than static communities.

Discussion

The method proposed in this study effectively extends current spatial fuzzy community detection approaches by embedding the semantic consistency in IRs into the community detection framework, enabling the model to simultaneously capture both semantic consistency and human mobility patterns.

A case study based on New York City taxi data demonstrated that this model effectively captures semantically distinct communities within the city, accurately identifying IRs such as the Financial District and Upper East Side. The probabilistic framework employed by the model enables precise modeling and fine-grained exploration of local zones near community boundaries, delivering clearer and more defined boundaries compared to the original ProOE model. Additionally, the incorporation of semantic consistency enhances detection stability. As shown in Figure 3, the communities identified by this method can be further utilized to assess the confidence of community and the certainty of spatial units, laying a foundation for subsequent practical applications. By leveraging semantic enhancement, the proposed model balances semantic and interaction information, uncovering communities with tight interactions and consistent semantic coherence. Adjusting different weighting parameters β allows the mining results to further analyze the alignment between semantic and interaction information within certain communities, as well as the confidence of communities across varying scales.

Although our model effectively detects communities constrained by semantic consistency, it has limitations that point toward promising future directions. Currently, the model relies on predefined IRs and applies a single global weight to their influence. A more sophisticated approach would incorporate richer data sources, such as Points of Interest (POIs) and land use, and allow the strength of semantic consistency to vary across different regions.

A powerful way to address these challenges would be to integrate our probabilistic framework with deep learning architectures. The core principles of our model—such as fuzzy membership and enforcing semantic coherence through higher-order structures—are highly compatible with graph-based deep learning models. Such a hybrid approach could leverage the representation power of deep learning to fuse multi-source data (like POIs and land use) in an end-to-end manner. This would allow the model to

automatically learn the varying semantic strengths of different areas, leading to a more adaptive and powerful framework for discovering semantically rich communities in complex urban environments.

Conclusion

This paper proposed a hyperedge-enhanced probabilistic optimal estimation method for detecting spatial fuzzy communities. By modeling the existing regional partition IRs in urban spaces as hyperedges and designing an intra-hyperedge semantic consistency metric using geographically weighted modified JS divergence, the method models semantic consistency and interaction within a unified probabilistic framework. By formulating both components as likelihoods, it jointly guides the partitioning of spatial fuzzy communities. The results show that, compared to the original model, the hyperedge-enhanced model can better capture the semantic consistency within communities and further analyze the consistency between semantic and interaction information in community detection results, thereby providing an important basis for in-depth analysis of urban spatial structures.

Data availability

The data and code can be requested from the corresponding author or the first author.

References

1. Ma, Z. & Zhu, D. Collective flow-evolutionary patterns reveal the mesoscopic structure between snapshots of spatial network. *Int. J. Geogr. Inf. Sci.* 1–32 (2024) doi:10.1080/13658816.2024.2395953.
2. Zhang, H. *et al.* An approach for exploring spatial associations in multi-layer networks based on convergent and divergent flow structures. *Int. J. Digital Earth* **17**, 2436478 (2024).
3. Liu, J. & Yuan, Y. Exploring dynamic urban mobility patterns from traffic flow data using community detection. *Annals of GIS* **30**, 1–20 (2024).
4. Hong, Y. & Yao, Y. Hierarchical community detection and functional area identification with OSM roads and complex graph theory. *Int. J. Geogr. Inf. Sci.* **33**, 1569–1587 (2019).
5. Jia, T. *et al.* Dynamical community detection and spatiotemporal analysis in multilayer spatial interaction networks using trajectory data. *Int. J. Geogr. Inf. Sci.* **36**, 1719–1740 (2022).
6. Liu, H. *et al.* Revealing Urban Spatial Interaction Characteristics and Crowd Travel Patterns from Trajectory Data. *Annals of the American Association of Geographers* 1–19 (2025) doi:10.1080/24694452.2024.2440409.
7. Feng, Z., Zeng, X., Li, W., Tan, Z. & Liu, Y. Revealing emission patterns of urban traffic flows: A complex network theory perspective. *Atmosphere* **16**, 594 (2025).

8. Fang, C. *et al.* Exploring spatial complexity: Overlapping communities in south China's megaregion with big geospatial data. *Comput. Environ. Urban Syst.* **112**, 102143 (2024).
9. Tuan, Y.-F. Space and place: humanistic perspective. in *Philosophy in Geography* (eds Gale, S. & Olsson, G.) 387–427 (Springer Netherlands, Dordrecht, 1979). doi:10.1007/978-94-009-9394-5_19.
10. Maiorani, A. Space and Place as Human Coordinates: Rethinking Dimensions across Disciplines. (Cambridge Scholars Publishing, 2021).
11. Balaska, V. *et al.* Semantic communities from graph-inspired visual representations of cityscapes. *Automation* **4**, 110–122 (2023).
12. Liu, S. & Wang, S. Trajectory community discovery and recommendation by multi-source diffusion modeling. *IEEE Trans. Knowl. Data Eng.* **29**, 898–911 (2017).
13. Huang, L., Yang, Y., Gao, H., Zhao, X. & Du, Z. Comparing community detection algorithms in transport networks via points of interest. *IEEE Access* **6**, 29729–29738 (2018).
14. Azaouzi, M., Rhouma, D. & Ben Romdhane, L. Community detection in large-scale social networks: state-of-the-art and future directions. *Social Network Anal. Min.* **9**, 23 (2019).
15. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 26113 (2004).
16. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech: Theory Exp.* **2008**, P10008 (2008).
17. Li, J. *et al.* A comprehensive review of community detection in graphs. *Neurocomputing* **600**, 128169 (2024).
18. Wang, X. *et al.* Self-supervised graph autoencoder with redundancy reduction for community detection. *Neurocomputing* **590**, 127703 (2024).
19. Expert, P., Evans, T. S., Blondel, V. D. & Lambiotte, R. Uncovering space-independent communities in spatial networks. *Proc. Natl. Acad. Sci.* **108**, 7663–7668 (2011).
20. Gao, S., Liu, Y., Wang, Y. & Ma, X. Discovering spatial interaction communities from mobile phone data. *Trans. GIS.* **17**, 463–481 (2013).
21. Chen, Y., Xu, J. & Xu, M. Finding community structure in spatially constrained complex networks. *Int. J. Geogr. Inf. Sci.* **29**, 889–911 (2015).
22. Guo, D., Jin, H., Gao, P. & Zhu, X. Detecting spatial community structure in movements. *Int. J. Geogr. Inf. Sci.* **32**, 1326–1347 (2018).
23. Wan, Y. & Liu, Y. DASSCAN: a density and adjacency expansion-based spatial structural community detection algorithm for networks. *ISPRS Int. J. Geo-Inf.* **7**, 159 (2018).
24. Liu, Q., Zhu, S., Deng, M., Liu, W. & Wu, Z. A spatial scan statistic to detect spatial communities of vehicle movements on urban road networks. *Geogr. Anal.* **54**, 124–148 (2022).

25. Xie, J., Kelley, S. & Szymanski, B. K. Overlapping community detection in networks: the state-of-the-art and comparative study. *ACM Comput. Surv.* **45**, 1–35 (2013).
26. Gupta, S. K., Singh, D. P. & Choudhary, J. A review of clique-based overlapping community detection algorithms. *Knowl. Inf. Syst.* **64**, 2023–2058 (2022).
27. Xie, H. *et al.* Redundancy-aware masked graph autoencoder for overlapping community detection in attributed networks. *Eng. Appl. Artif. Intell.* **162**, 112821 (2025).
28. Ni, L. *et al.* Local overlapping spatial-aware community detection. *ACM Trans. Knowl. Discovery Data* **18**, 59 (2024).
29. Sekulić, S., Long, J. & Demšar, U. A spatially aware method for mapping movement-based and place-based regions from spatial flow networks. *Trans. GIS.* **25**, 2104–2124 (2021).
30. He, X. *et al.* A probabilistic optimal estimation method for detecting spatial fuzzy communities. *International Journal of Geographical Information Science* **0**, 1–33 (2025).
31. Huang, L., Yang, Y., Zhao, X., Gao, H. & Yu, L. Mining the relationship between spatial mobility patterns and POIs. *Wireless Communications and Mobile Computing* **2018**, 4392524 (2018).
32. Yu, B., Wang, Z., Mu, H., Sun, L. & Hu, F. Identification of urban functional regions based on floating car track data and POI data. *Sustainability* **11**, 6541 (2019).
33. Liu, S., Wang, S., Jayarajah, K., Misra, A. & Krishnan, R. TODMIS: Mining communities from trajectories. in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* 2109–2118 (Association for Computing Machinery, New York, NY, USA, 2013). doi:10.1145/2505515.2505552.
34. Liu, C. & Guo, C. STCCD: Semantic trajectory clustering based on community detection in networks. *Expert Syst. Appl.* **162**, 113689 (2020).
35. Cheng, S., Yang, S., Cheng, X., Li, K. & Zheng, Y. Semantic overlapping community detection with embedding multi-dimensional relationships and spatial context. *Social Network Anal. Min.* **14**, 14 (2023).
36. Dandekar, A., Bressan, S., Abdessalem, T., Wu, H. & Ng, W. S. Detecting communities of commuters: Graph based techniques versus generative models. in *On the Move to Meaningful Internet Systems: OTM 2016 Conferences* (eds Debruyne, C. *et al.*) 485–502 (Springer International Publishing, Cham, 2016). doi:10.1007/978-3-319-48472-3_29.
37. Zhao, Z. *et al.* Graph-based clustering: High-order bipartite graph for proximity learning. *IEEE Trans. Knowl. Data Eng.* **37**, 4649–4663 (2025).
38. Zhao, Z. *et al.* Fuzzy clustering via orthogonal tensor decomposition on high-order anchor graphs. *IEEE Trans. Fuzzy Syst.* **33**, 2987–3000 (2025).
39. Zhao, Z. *et al.* Enhancing clustering performance with tensorized high-order bipartite graphs: a structured graph learning approach. *IEEE Trans. Circuits Syst. Video Technol.* **35**, 2616–2631 (2025).

40. Contisciani, M., Battiston, F. & De Bacco, C. Inference of hyperedges and overlapping communities in hypergraphs. *Nat. Commun.* **13**, 7229 (2022).

Competing interests

The authors declare no competing interests.

Acknowledgements

Funding

This project was supported by the National Natural Science Foundation of China under Grant 42471506; the Provincial Natural Science Foundation of Hunan under Grant 2025JJ40034; the Changsha Distinguished Young Science and Technology Talent Program kq2506011; the Scientific Research Fund of Hunan Provincial Education Department under Grant 23B0013 and the Open Topic of Hunan Geospatial Information Engineering and Technology Research Center under Grant HNGIET2024004.

Author contributions

X.H.: Research framework, literature review, research methods, data processing and paper writing. Z.T.: resources, formal analysis and validation. B.L.: Formal analysis, conceptualization, logical organization and editing. J.D.: resources, formal analysis and validation. M.D.: conceptualization, resources, supervision and funding acquisition. All authors reviewed the manuscript.