International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tgis20

Automatic road network selection method considering functional semantic features of roads with graph convolutional networks

Jianbo Tang, Min Deng, Ju Peng, Huimin Liu, Xuexi Yang & Xueying Chen

To cite this article: Jianbo Tang, Min Deng, Ju Peng, Huimin Liu, Xuexi Yang & Xueying Chen (05 Aug 2024): Automatic road network selection method considering functional semantic features of roads with graph convolutional networks, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2024.2387198

To link to this article: https://doi.org/10.1080/13658816.2024.2387198

Published online: 05 Aug 2024.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2024.2387198

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Check for updates

Automatic road network selection method considering functional semantic features of roads with graph convolutional networks

Jianbo Tang[a,b], Min Deng[a,b], Ju Peng[a], Huimin Liu[a], Xuexi Yang[a,b] and Xueying Chen[a]

[a]Department of Geo-Informatics, Central South University, Changsha, Hunan, China; [b]Hunan Geospatial Information Engineering and Technology Research Center, Changsha, Hunan, China

ABSTRACT

Road network selection plays a key role in map generalization for creating multi-scale road network maps. Existing methods usually determine road importance based on road geometric and topological features, few evaluate road importance from the perspective of road utilization based on human travel data, ignoring the functional values of roads, which leads to a mismatch between the generated results and people's needs. This paper develops two functional semantic features (i.e. travel path selection probability and regional attractiveness) to measure the functional importance of roads and proposes an automatic road network selection method based on graph convolutional networks (GCN), which models road network selection as a binary classification. Firstly, we create a dual graph representing the source road network and extract road features including six graphical and two functional semantic features. Then, we develop an extended GCN model with connectivity loss for generating multi-scale road networks and propose a refinement strategy based on the road continuity principle to ensure road topology. Experiments demonstrate the proposed model with functional features improves the quality of selection results, particularly for large and medium scale maps. The proposed method outperforms state-of-the-art methods and provides a meaningful attempt for artificial intelligence models empowering cartography.

ARTICLE HISTORY
Received 11 July 2023
Accepted 29 July 2024

KEYWORDS
Road network selection; graph convolutional network; functional features; map generalization; POI data

1. Introduction

Road network serves as the fundamental space for human travel activities, and multi-scale urban road network maps form the data foundation of current location-based applications in cities. Road network selection, a crucial generalization operator, is essential for producing smaller-scale road network maps from larger-scale maps, which aims to retain important roads and omit less important ones (Li and Zhou 2012). Two critical challenges in road network selection are evaluating the importance of roads

CONTACT Ju Peng daisy_pj@csu.edu.cn

and determining the optimal number of roads to retain (Foerster *et al.* 2010). Compared to other map elements, such as points and polygons, the road network is presented as polylines with more complex geometries and topological relationships, posing greater difficulties in identifying the required important roads while preserving the overall connectivity of the selected road network. Moreover, the close relationship between the road network and human activities further complicates the road network selection. Therefore, multiple factors that describe the importance of roads from different perspectives should be considered and still require further research and exploration, especially in the era of geospatial big data (Zheng *et al.* 2021).

Compared with rural road networks, urban road networks exhibit higher density and structural complexity, which is the focus of this study. Numerous studies have explored the urban road network selection (Thomson and Richardson 1995, Chen *et al.* 2009, Benz and Weibel 2014, Zheng *et al.* 2021). Existing road network selection methods can be divided into semantic-based methods, graph-based methods, stroke-based methods, mesh density-based methods, and machine learning-based methods. These methods typically construct road network graphs or stokes, evaluating road importance based on geometric, topological, and semantic features. Although these studies have received remarkable achievements for road network selection, some limitations still need further investigation. First, previous studies primarily focus on evaluating road importance using graphical features. The social function values of roads are rarely considered due to the limited acquisition of human travel data, leading to inconsistencies between the road selection results and practical requirements. For example, some less traveled roads are preserved, whereas some frequently used roads may be omitted. Second, existing methods heavily rely on manually setting selection parameters, such as mesh density thresholds and road feature weights, which is challenging without prior knowledge.

In recent years, the advancements in positioning, navigation, and sensor technologies have facilitated the acquisition of human activity-related data, providing new data sources for studying human-road relationships and human travel patterns (Çolak *et al.* 2016, Siła-Nowicka *et al.* 2016, Peng *et al.* 2023, Peng *et al.* 2024). Points of Interest (POIs) surrounding the roads are the motivations that attract humans to travel, while the trajectories are the records of the human travel process. The collaborative consideration of trajectories and POIs can reflect the social attributes (i.e. functional value) and actual utilization of roads from the perspective of human travel. They can help select roads that are more consistent with human cognition. In addition, with the rapid development of artificial intelligence, artificial neural networks have emerged as a promising data-driven approach for learning the experience and knowledge of cartographic experts (Zhou and Li 2014). Some studies have demonstrated the feasibility of deep learning-based methods in improving the automation and intelligence of map generalization (Touya *et al.* 2019, Zheng *et al.* 2021, Ai 2022, Yu and Chen 2022), motivating the development of deep learning methods for road network selection.

Inspired by these, this study introduces two new features to assess the functional importance of roads and proposes a GCN-based method for automatic road network selection. Specifically, the road network selection is modelled as a binary classification with GCNs. Six graphical structure features and two functional semantic features are used to evaluate the importance of the roads. Further, an improved loss function with the connectivity penalty and the refinement process are performed to ensure the connectivity of the selected roads. The main contributions of this paper are summarized as follows:

- Two new features (i.e. *travel path selection probability* and *regional attractiveness*) are proposed to measure the functional importance of the roads besides commonly used graphical structure features.
- A GCN-based road network selection model is developed. The model can automatically learn the number of roads to be retained at different target scales and the weights of different features for road network selection.
- Experiments are conducted to compare the performances of the proposed method and baseline methods, and the influence of different feature combinations on the selection of road networks at different scales is analyzed.
- The transferability and generalization ability of the proposed method are explored with road network data in different cities.

The remainder of this paper is organized as follows. Section 2 reviews the related work for road network selection. Section 3 introduces the proposed method in detail and Section 4 evaluates the performance of the proposed method and analyses the effects of the proposed method with different feature combinations. Section 5 discusses other datasets for evaluating road functional importance, the significance of road features at different scales, and the transferability of the proposed method. Section 6 concludes the study.

2. Related work

For road network selection, Töpfer's radical law is most used to determine how many roads should be selected at the target scale in map generalization (Töpfer and Pillewizer 1966). Jiang (2015) modelled the map generalization as a head/tail breaks process and proposed the head/tail breaks statistics to guide the selection of urban streets. With the recent advances in artificial intelligence, machine learning-based methods such as BP (Zhou and Li 2014), SVM (Zhou and Li 2017), and GNN (Zheng *et al*. 2021, Guo *et al*. 2023) have emerged as powerful new techniques for automatically determine the number of roads retained at the target scale. And to determine which specific roads to retain at the target scale, many methods have been proposed. These methods can be categorized into semantic-based methods, graph theory-based methods, stroke-based methods, mesh density-based methods, and machine learning-based methods. They are reviewed in detail below.

- ***Semantic-based methods***: These methods evaluate the importance of roads based on attribute information, such as road level and road type, and select a certain number of roads in order of their importance. The semantic information of roads is often employed as auxiliary features in conjunction with geometric and topological road features for road network selection.

- ***Graph theory-based methods***: These methods abstract the road network into a graph (Thomson and Richardson 1995) and are quantitatively evaluated using complex network theory indicators (e.g. shortest path, centrality, minimum spanning tree) to assess road importance (Mackaness and Beard 1993, Jiang and Claramunt 2004, Porta *et al*. 2006, Gülgen and Gökgöz 2011). A certain number of roads that occupy important positions in the road network are then selected. These methods primarily focus on the graphical characteristics (i.e. connectivity and topology) of the road network (Richardson and Thomson 1996), forming the foundation for many other subsequent road network selection methods.
- ***Stroke-based methods***: The concept of strokes was first proposed by Thomson and Richardson (1999) based on the 'good continuation' grouping principle of Gestalt theory as the basic unit for road network selection. The construction of strokes usually relies on geometric criteria (Liu and Li 2019), road names, or road levels with strategies like every-best-fit, self-best-fit, and self-fit (Jiang *et al*. 2008, Zhou and Li 2012). Then, stroke length, average stroke density, centrality indicators, and stroke level are often employed to calculate the importance of strokes (Liu *et al*. 2010, Xu *et al*. 2012, Yang *et al*. 2013, Weiss and Weibel 2014). Recently, Yu *et al*. (2020) extended the classical stroke-based selection method with traffic flow patterns based on the thought that strokes with a close relationship in the traffic flow system should be selected simultaneously, which enriches the topological connectivity relationships between strokes. In their method, four common features of stroke length, degree centrality, closeness centrality, and betweenness centrality were used to determine the relative importance of strokes. While known for maintaining good road continuity at target scales, stroke-based methods heavily depend on specific parameters such as road deflection angle threshold and the weights of different features, potentially reducing their applicability.
- ***Mesh density-based methods***: This type of method was proposed by Hu *et al*. (2007) and Chen *et al*. (2009), aiming to maintain the relative density of the road network before and after selection. This approach first divides the road network space into sub-regions (meshes). Then, the relative importance of bounding road segments for different meshes is calculated and the least important road segments are eliminated. Benz and Weibel (2014) utilized an extended stroke-mesh algorithm for road network selection at a medium scale.
- ***Machine learning-based methods***: These methods aim to apply intelligence models to improve the automation and intelligence of road network selection. Traditional machine learning-based methods such as self-organizing neural networks (Jiang and Harrie 2004), genetic algorithms (Van Nimwegen *et al*. 1999, Mathew and Isaac 2014), BP neural networks (Zhou and Li 2014), support vector machines (SVM) (Zhou and Li 2017), decision tree models (Karsznia *et al*. 2022) have been used for road network selection, which reduces the difficulty of feature integration compared to the above methods. With the remarkable progress of deep neural networks in feature extraction and classification, deep learning methods also exhibit great potential in automatic and intelligent map generalization (Touya *et al*. 2019, Ai 2022, Courtial *et al*. 2023). For example, Zheng *et al*. (2021) explored the applicability of three graph neural networks (GNNs) for road network selection, demonstrating the

feasibility and effectiveness of GNNs in map generalization. However, their work only considered statics graphical features and focused on small-scale maps. The effectiveness of GNNs, considering more semantic features at medium or large-scale maps in more cities, requires further exploration.

Although existing research has achieved significant improvements in road network selection, some limitations still remain and are worth further exploration. First, the importance of roads is usually determined based on the graphical features of roads, such as road level, length, connectivity, and centrality, overlooking the social attributes and functional values of roads. As a result, the generated road network map may not be consistent with actual human travel demands. Second, existing road network selection methods rely on manually setting selection parameters (e.g. the number of roads to be selected and the weights of different road features), resulting in low automation and intelligence. To overcome these limitations, this study proposes two new functional features besides six commonly used graphical features to evaluate the importance of roads. A CGN-based method is proposed for automatic road network selection, which models connectivity loss and integrates the topology refinement process. The framework of the proposed method is shown in Figure 1, which comprises four critical steps: road network abstraction representation, road feature extraction, road network selection model construction, and road selection result evaluation.

3. Methodology

3.1. Dual graph abstraction of the road network

The primitive graph and dual graph are two commonly used ways to represent the road network (Porta *et al.* 2006). The primitive graph abstracts the intersections or road endpoints as nodes and road segments as edges, as shown in Figure 2(b), which

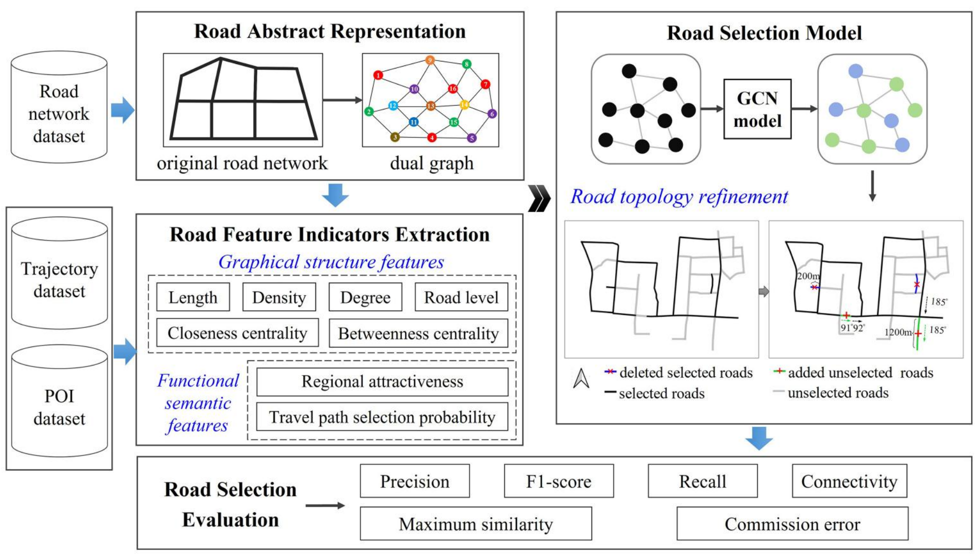


Figure 1. The framework of the proposed method.

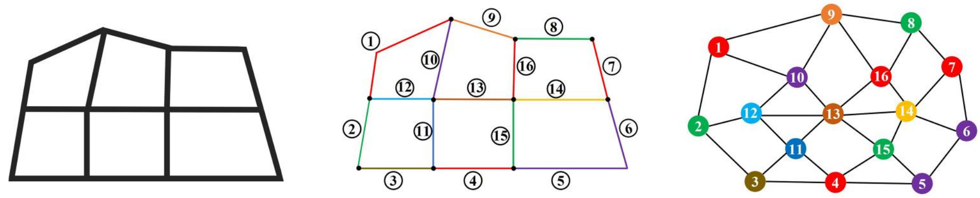


Figure 2. Illustration of the road network abstraction.

directly reveals the spatial structure of the road network. However, the implicit expression of topological relationships between different roads is not conducive to road segment-based geographical analysis. In contrast, the dual graph abstracts road segments as graph nodes and connections between road segments as edges, as shown in Figure 2(c), which directly reveals topological relationships between different road segments. Therefore, the dual graph is better suitable for road network selection, which can be naturally transformed into the binary classification of nodes in the dual graph. Here, we give a formal description of the binary classification model for road network selection using the dual graph.

Let N be the number of road segments in the road network and K be the number of road features. The road network can be represented as an undirected graph, denoted as $\mathbf{G}=(\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{X}, \boldsymbol{Y})$, where:

- $\boldsymbol{V}= \{v_i|0 \leq i \leq N\}$ is the set of graph nodes representing road segments in the origin road network.
- $\boldsymbol{E}= \{e_{ij}=(v_i, v_j) \mid 0 \leq i \leq N, 0 \leq j \leq N\}$ denotes the set of graph edges denoting the connectivity between road segments. e_{ij} exists if and only if $i \neq j$ and node v_i is directly connected to node v_j.
- $\boldsymbol{X}= \{x_{ik}|0 \leq i \leq N, \ 0 \leq k \leq K\}$ is a K-dimensional feature matrix used to calculate road importance.
- $\boldsymbol{Y}= \{y_i|0 \leq i \leq N\}$ denotes the classification label of road segment v_i at the target scale with 1 for selected and 0 for unselected.

3.2. Road features extraction for road selection

In road network selection, road feature extraction for road importance evaluation is a critical step that determines the quality of the road network selection results. This study considers not only the graphical structures of roads but also their social functional values for human travel when assessing the importance of roads. Six graphical structure features and two proposed functional features of roads are extracted and assigned to the feature matrix $\boldsymbol{X}$, which will be used as inputs for the proposed GCN-based road network selection model.

3.2.1. Graphical structure features extraction for road selection

Six graphical structure features of roads include two geometric features (i.e. road length, road density), three topological features (i.e. degree, closeness centrality, betweenness centrality), and one semantic feature (i.e. road type). The definitions of each feature are as follows.

- *Road length* (*L*): Length is a fundamental and commonly used feature to describe the spatial domain of a road. It reflects the range of the road and is positively correlated with its importance.
- *Road density* (*RD*): Maintaining the relative consistency of road density within subregions before and after selection is crucial for evaluating the rationality of the selected road network structure (Liu *et al.* 2009). The calculation of road density follows the method outlined by Tian *et al.* (2016).
- *Degree* (*D*): In the dual graph, the degree of a road v_i is the number of its neighbouring roads directly connected to v_i (Brintrup *et al.* 2016). It reflects the accessibility of the road within the road network, with higher degrees corresponding to greater influence.
- *Closeness Centrality* (*CC*): It is defined as the reciprocal of the total shortest paths from a single road to all other roads (Yi *et al.* 2018). Closeness centrality reflects the proximity of the road to its non-directly connected roads in the entire road network with higher closeness centrality values indicating stronger associations between the given road and other roads.
- *Betweenness Centrality* (*BC*): Betweenness centrality is widely used to measure the ability of a node that allows information to be transmitted from one node to another in a graph. Its calculation is shown in Equation (1), where $o(s, t|\ i)$ denotes the number of shortest paths between any pair of nodes (say node s and node t) that pass through the target node i, and $o(s, t)$ denotes the total number of shortest paths between any pair of nodes (Brandes 2008).

$$BC_i = \sum_{s,t \in V, s,t \neq i} \frac{o(s,t|i)}{o(s,t)} \tag{1}$$

- *Road level* (*RL*): Road level, also called road type, reflects the designed traffic-carrying capacity of a road. In general, the importance of a road is positively correlated with the road level.

3.2.2. Functional semantic features extraction for road selection

The importance of a road is determined not only by its graphical structure features but also by the functional values it carries. For example, if a wide and long road is abandoned for a long time with little traffic flow, it is deemed less important. Conversely, a short and low-level road carrying a large traffic flow can be more crucial. Therefore, two new features, namely *regional attractiveness* and *travel path selection probability* are proposed to evaluate the social functional values of roads. They are combined with six graphical structure features to comprehensively evaluate the importance of roads within the road network.

1. ***Regional attractiveness***: Various POIs are the main destinations that motivate humans to travel, such as malls for shopping, schools for education, and hospitals for healthcare. Roads with more POIs around them are likely to attract more

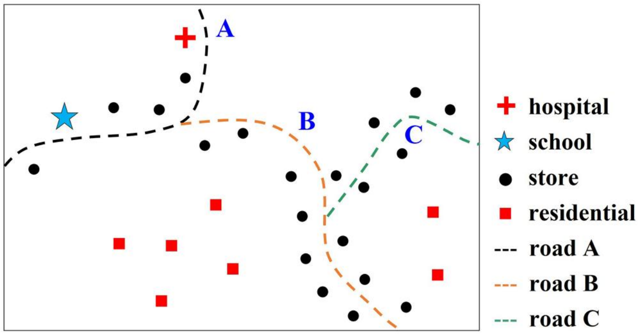


Figure 3. An example of spatially heterogeneous distribution of different types of POIs.

crowds and thus be more important. Current studies indicate that POIs are helpful for road network selection (Xu *et al.* 2018, Han *et al.* 2020). However, due to the diverse functional types of POIs, the impacts of different POIs on the importance of roads vary. Simply considering the number of POIs around the road ignores the heterogeneous distribution of different types of POIs. For example, in Figure 3, roads A, B, and C are surrounded by a school, a hospital, residential buildings, and many shopping stores. If only considering the number of POIs, road B with more shopping stores nearby might be more important. However, the actual factors attracting people to this area are more likely the school and hospital near road A. Simply relying on the number of POIs can lead to misjudging the importance of roads.

To address this issue, we propose an enhanced indicator, *regional attractiveness*, to measure the importance of roads by considering the attractiveness of POIs from the perspective of human travel purposes. If more appealing and weighted POIs are near a road, it is considered to be more important and has a higher probability of being retained. The differences in the number of different types of POIs around the road are considered. The *regional attractiveness* of road i is calculated as:

$$RAN_i = \sum_{m=1}^{M} \frac{w_m * PN_{im}}{L_i} \qquad w_m = \frac{PN}{PN_m} \tag{2}$$

where M is the number of POI types (e.g. residential, commercial, industrial, green spaces), w_m is the weight of POIs of type m, PN_{im} denotes the number of POIs of type m in the road buffer zone, L_i is the length of road i, PN is the total number of all POI types, and PN_m is the total number of POIs of type m in the study area.

The *regional attractiveness* indicator calculates the ratio of the total number of POIs belonging to type m to the total number of POIs of all types in the study area and uses this ratio as a coefficient to adjust the weight of POIs of type m, which can reduce the differences in the number of different POI types. We also consider the line density of POIs by dividing by the road length since the denser the surrounding POIs, the more important the road. In addition, the road buffer zone size should not be too large or too small to exactly represent the surrounding environment influence of the road. In the experiments, we first match the POIs to the source road network using a geometric-based approach (Yang *et al*. 2014), and finally set the buffer size to 300 m according to previous studies (Schipperijn *et al*. 2010, Xu *et al*. 2018) and the maximum road width in the study area.

2. ***Travel path selection probability***: While POIs are the attractors for human travel, traffic flows (calculated from vehicle trajectories) can directly reveal the actual usage of roads, reflecting the importance of roads to some extent. In general, roads with larger traffic flows in the road network are more likely to be selected for travel, indicating their greater importance, which is also consistent with human cognition. Therefore, the functional importance of roads can also be measured by their traffic volume. Traffic volume refers to the number of vehicles passing through a specific location during a given period. As illustrated in Figure 4, considering that there is variability and randomness of traffic volume at small-time scales (e.g. per minute, per hour), the long-term traffic volume (e.g. per week), which exhibits stable patterns, is employed to model the importance of roads. Here, another road functional semantic feature based on the traffic volume, namely, *travel path selection probability*, is defined for the road network selection. *Travel path selection probability* refers to the probability that a road is selected among many roads when people travel, which can reflect the relative usage intensity of the roads. The travel path selection probability of road i is calculated as follows:

$$PT_i = \frac{WT_i}{\sum_{i=1}^{N} WT_i} \tag{3}$$

where WT_i denotes the weekly traffic volume of road i, and N is the number of roads in the study area.

3.3. Road network selection model based on graph convolutional network

In the road network selection process, there are only two states of a road: selected or unselected, and its importance is determined not only by its own features but also by

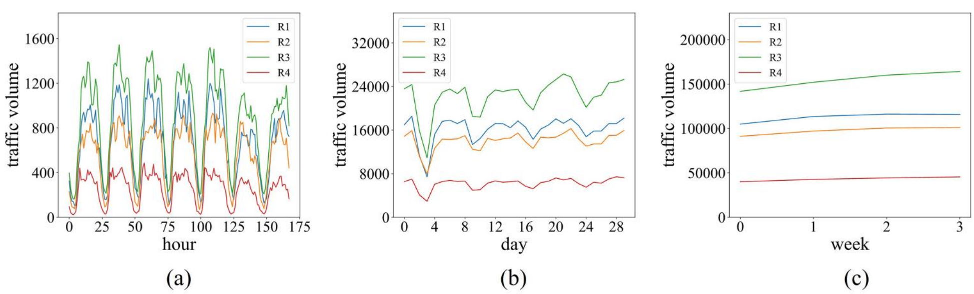


Figure 4. Comparison of traffic volumes of four roads (R1, R2, R3, and R4) in Beijing at different time scales in November 2019 of (a) at the hour scale, (b) at the day scale, and (c) at the week scale.

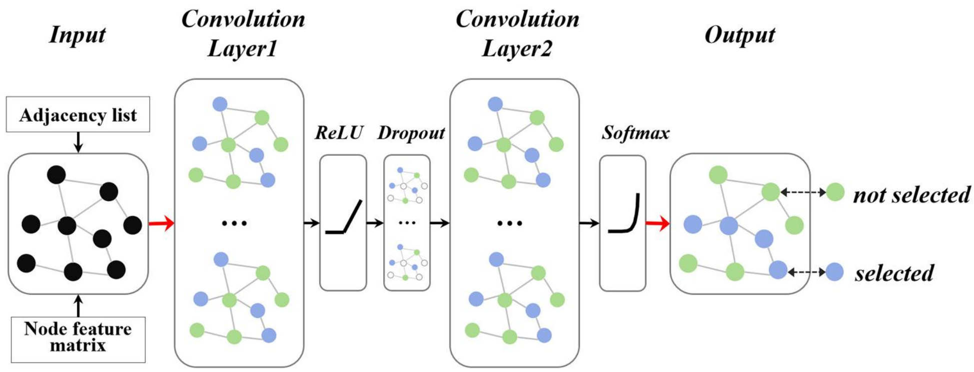


Figure 5. Road network selection model with GCNs.

the characteristics of its neighbouring roads. Therefore, the road network selection problem can be naturally modelled as a binary classification task, where selected roads are marked as 1, and unselected roads as 0. We propose an automatic road network selection model based on the graph convolutional network (GCN) owing to its powerful ability of feature extraction and information aggregation. As illustrated in Figure 5, it includes an input layer, two convolutional layers, a dropout layer, and an output layer. The input layer receives an $N*K$ feature matrix, where N denotes the number of road nodes and K is the number of road features. The dropout layer randomly discards certain neurons to alleviate the overfitting problem. The convolutional layer is employed to extract salient feature information from road nodes and their adjacent road nodes, and it is described as:

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}X^{(l)}W^{(l)}) \tag{4}$$

where σ represents a nonlinear activation function, $\sim D$ denotes the degree matrix, $\sim A$ is the adjacency matrix, $X^{(l)}$ is the output matrix of the last layer l, and $W^{(l)}$ denotes the connection weight matrix of layers l and $l+1$. More detailed information can be found in Kipf's work (Kipf and Welling 2016).

In the last output layer, the *Softmax* function is used to perform node classification for road network selection. The output is an $N*1$ matrix, and each value indicates the selection result of the corresponding road with selected equal to 1 and unselected equal to 0. The proposed GCN-based road network model can extract the implicit features of roads, capture the influence of adjacent roads, and estimate the optimal weights of different features.

For training classification models, the cross-entropy loss function is commonly used (Shore and Johnson 1980). To preserve road connectivity at the target scale map, we introduce a connectivity loss term to the cross-entropy loss function. First, we define an indicator function $\boldsymbol{I}(.)$ to measure road connectivity. If both ends of road i are connected to other roads, the value of $\boldsymbol{I}(i)$ equals 0. If only one end of road i is connected to other roads (e.g. hanging roads), the value of $\boldsymbol{I}(i)$ equals 1. Otherwise (road i is isolated), the value of $\boldsymbol{I}(i)$ equals 2. The proposed indicator function differs from the concept of degree. As shown in Figure 6, there are four roads A, B, C, and D with the

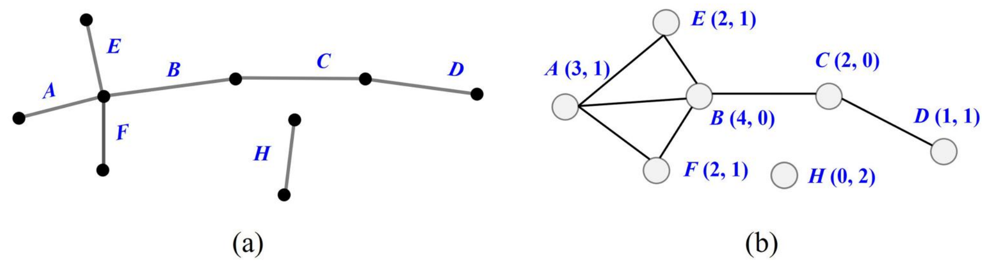


Figure 6. The illustration of *I*(.) of (a) the original road network, and (b) the dual graph of the original road network. The first and second elements of the tuple in (b) are the degree and *I*(.) values of roads, respectively.

same level, two hanging roads *E*, and *F*, and one isolated road *H*. The degrees of *A*, *B*, *C*, and *D* are 3, 4, 2, and 1, respectively, while their indicator values are 1, 0, 0, and 1, respectively. Although road *A* has a higher degree than road *C*, the ***I***(.) value of road *C* is smaller than road *A*, indicating that road *C* is more likely to be selected in terms of road connectivity. For the hanging roads *E* and F, the degree values are both 3, while the ***I***(.) values are both 1. For the isolated road *H*, the degree is 0, and the ***I***(.) value is 2. In this case, the selection model with connectivity loss tends to omit isolated and hanging roads rather than well-connected roads.

Then, the connectivity loss is defined as the ratio of the sum of ***I***(.) values to the number of selected roads. In addition, a regularization item is incorporated to mitigate overfitting in the road network selection model. It is defined as the ratio of the number of selected roads to the total number of roads at the source scale. Using the extended loss function, the road selection model aims to maintain the maximum similarity between the selected and original roads while preserving good connectivity among the selected roads. The complete proposed loss function is formulated in Equation (5), where the first item is the cross-entropy loss, the second item is the connectivity loss, and the last item denotes the regularization item.

$$\begin{aligned} L = &(-(y\log(\hat{y}) + (1-y)\log(1-\hat{y}))) + (\lambda_1 * sum(I(\hat{y}_{\text{label}=1}))/|\hat{y}_{\text{label}=1}|)+) \\ &+ (\lambda_2 * |\hat{y}_{\text{label} = 1}|/|\hat{y}|) \end{aligned} \tag{5}$$

where y represents the road selection labels of the ground truth, $\hat{y}$ denotes the road selection labels predicted by the model, $|\hat{y}|$ is the number of roads in the source scale map, $\hat{y}_{\text{label} = 1}$ represents the set of roads selected by the model from the source scale map, and $|\hat{y}_{\text{label} = 1}|$ is the number of the selected roads by the model, λ_1 and λ_2 are two hyperparameters to adjust the loss function.

Figure 7 gives an example to illustrate the effectiveness of the proposed connectivity loss using the road network in Figure 6. There are three road selection results *a*, *b*, and *c*. The values of the proposed connectivity loss of the road selection results in Figure 7 are 1, 3/4, and 1/2, respectively. Obviously, the road selection result *c* is better than *b*, and result *b* is better than *a* in terms of good connectivity. The connectivity loss can guide the model to select roads with good connectivity when selecting the same number of roads.

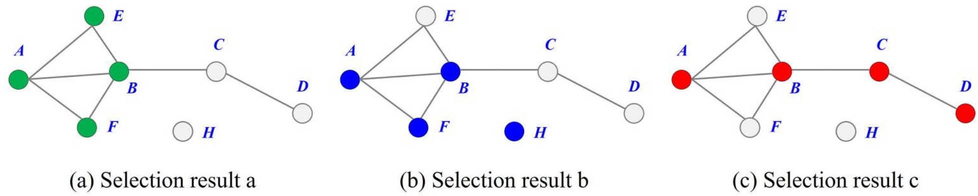


Figure 7. Illustration of the connectivity loss to preserve roads with good connectivity.

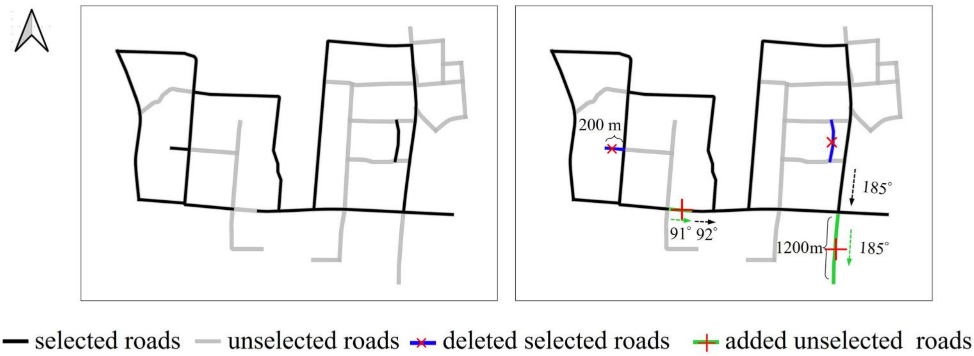


Figure 8. The illustration of the refinement process of the selected road network selection.

3.4. Refinement of the selected roads using the principle of road continuity

Maintaining road connectivity is crucial for reliable navigation applications. However, the road selection results obtained by the above GCN-based selection model may contain disconnected or dangling roads, which would disrupt the topological integrity of the road network. This issue is particularly pronounced when utilizing functional road features, where shorter and lower-level roads carrying large traffic volumes and distributed with important POIs may be selected. To address this issue, we further propose a refinement strategy based on the principle of stroke construction to ensure the connectivity of the selected roads. Let E_s be the set of the selected roads, and E_u be the set of the unselected roads, the refinement process comprises three key steps and is illustrated in Figure 8:

Step 1: Remove isolated roads and hanging roads with a very short length (e.g. less than 50 m).

Step 2: Fill gaps and connect adjacent roads in similar directions to ensure the selected road network maintains proper connectivity. Let E_i denote the neighbors of the selected road e_i. For each neighbor $e_{ik} \in E_i$, if the following conditions are met:

a. e_{ik} is not currently in the selected road set.
b. The direction difference between e_{ik} and e_i is less than a predefined angular θ, where θ is determined according to the stroke construction principles outlined by Thomson and Richardson (1999).
c. e_i is not a hanging road, or its length exceeds a given threshold (as specified in Step 3).

Then e_{ik} is added to the selected road set E_s and removed from the unselected road set E_u. This step ensures the continuity of the selected roads.

Step 3: Remove hanging roads that are shorter than the specified length threshold to avoid unnecessary fragments. The rationale behind this step stems from the concept of the smallest visual object (SVO) for line generation (Li and Openshaw 1992). Roads that are too short to be visually perceptible on the target scale map may introduce unnecessary clutter. Specifically, if the length of a hanging road is less than d mm on the target scale map (i.e. the ground length of the road is less than d mm $\times$ M, where M is the denominator of the target scale), the road will be removed. The parameter d can be determined based on specific application requirements and cartographic conventions. Generally, it should be greater than 0.5 mm to ensure the legibility of the selected roads on the target scale map.

After the above steps, the remaining roads are the final result of the proposed method.

4. Experiments

4.1. Study area and data processing

To evaluate the effectiveness of the proposed method, experiments were conducted on real road network data within the Fifth Ring Road in Beijing, China, covering an area of 667 square kilometers. This area was chosen for its dense and complex road network, as well as the abundant floating car trajectory data and POIs. In the experiments, road segments were used as basic units for road network selection. The raw road network data in 2019 at the source scale of 1:10,000 in the study area was downloaded from OpenStreetMap (https://www.openstreetmap.org). Road centreline extraction, topology check, and correction were performed to obtain the source road network, which finally contains 11,727 road edges and 7,829 intersections or endpoints. Road networks at target scales of 1:50,000 and 1:200,000 in 2019 were obtained from the National Mapping Agency (NMA) (https://www.tianditu.gov.cn) and aligned with the road network at the scale of 1:10,000. The three handled road networks are shown in Figure 9. It should be noted that road network maps from NMA are generally produced by cartographers, where road structural characteristics, functional values, cultural factors, and practical application demands are considered. As many scholars have done in previous studies, in the experiments, the road network data obtained from NMA were used as the benchmark data to evaluate the performance of the road network selection method. Hereafter, the road network at the source scale is referred to as the source road network, the corresponding benchmark road network at the target scale is referred to as the target road network, and the road network selected by the generalization algorithm is referred to as the selected road network.

For this study, floating car trajectory data from November 1 to November 30, 2019, and POIs in 2019 were collected to derive the functional semantic features of the roads, i.e. travel path selection probability and regional attractiveness. The floating car trajectory data and POIs are shown in Figure 10. Noise trajectories such as redundant, stranded, and abnormal trajectory points were first eliminated manually. Then, the ST-matching algorithm (Lou *et al.* 2009) was implemented to match trajectories to the

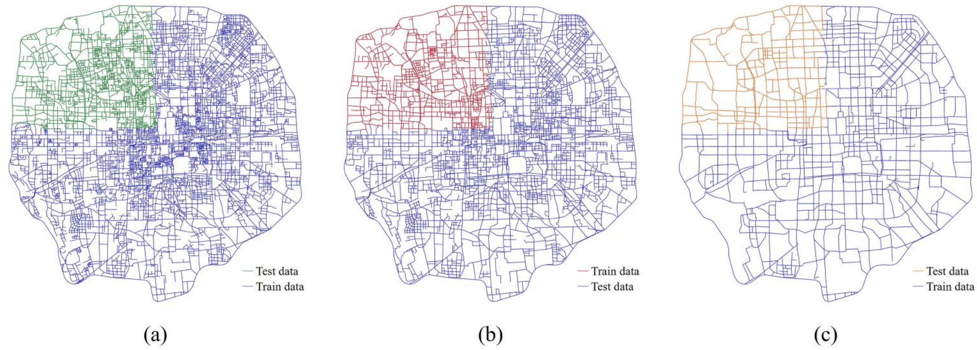


Figure 9. Study area and road network data. (a) The source road network data at 1:10,000; (b) the target road network data at scale 1:50,000; (c) the target road network data at scale 1:200,000.

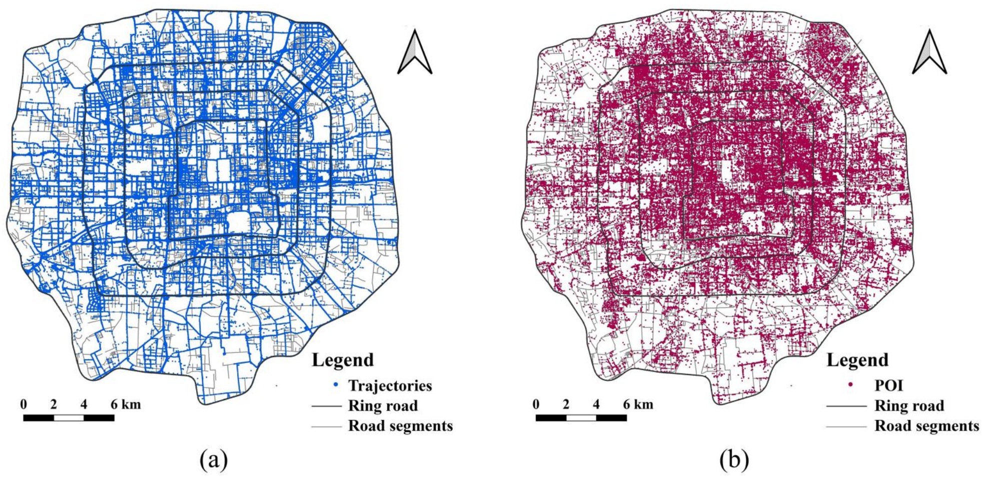
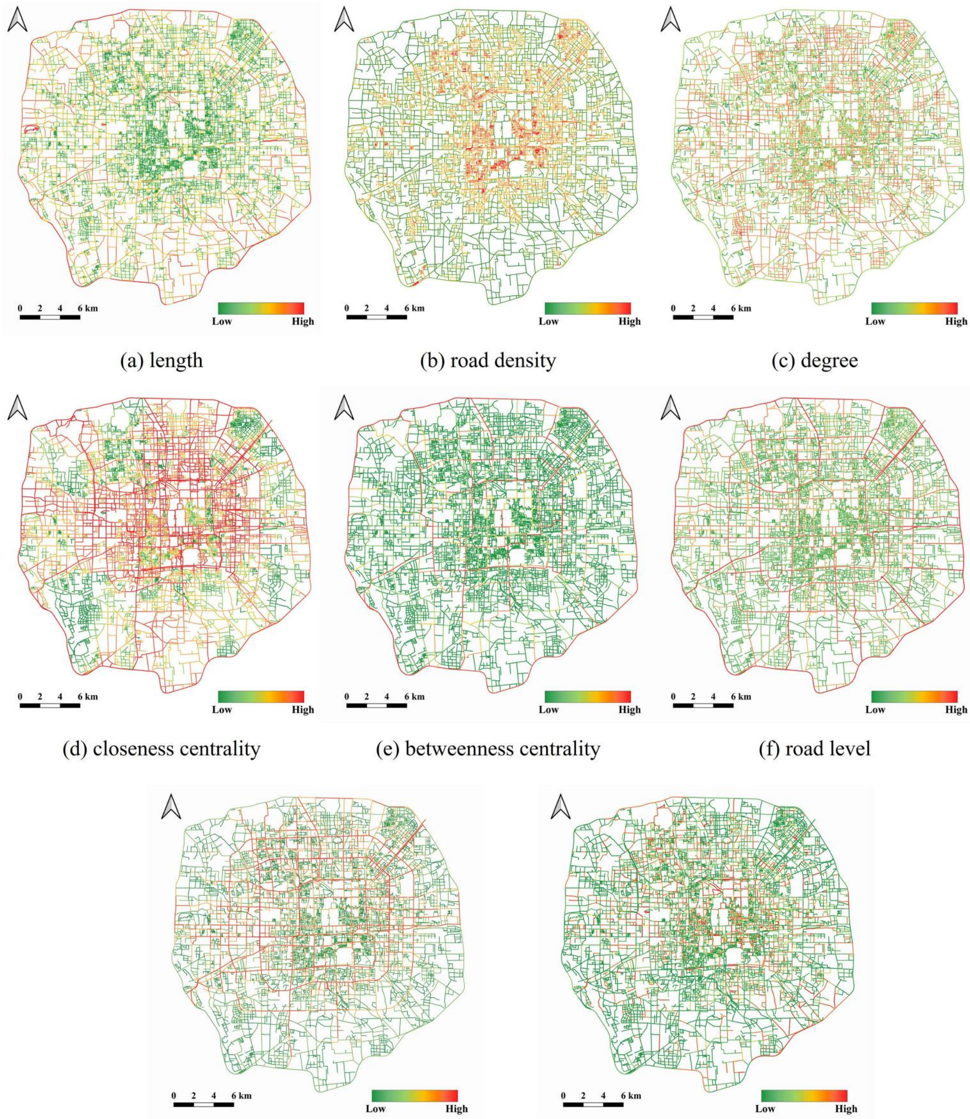


Figure 10. Floating car trajectory data and POIs. (a) Trajectory data on 2019-11-01 from 16:00 to 17:00 in the study area; (b) POIs within the Fifth Ring Road of Beijing.

road segments. Notably, in the trajectories and road segments matching process, the merged two-way road still has two directions and can match the trajectories passing through its left or right lanes, and we summed the traffic flows in two directions as the traffic volume of the merged two-way road for calculating the travel path selection probability. To calculate the regional attractiveness of each road in the road network, a coordinate transformation was first performed on the POIs obtained from Amap (https://lbs.amap.com) to address the coordinate offset issue due to the map encryption policy. Then we matched the POIs to the source road network using the geometric-based approach from Yang *et al*. (2014) and extracted the POIs within the buffer zones of the roads.

After that, the dual graphs were constructed and eight road features including six graphical structure features and two functional sematic features were extracted, as shown in Figure 11. The Pearson coefficient of 0.218 between travel path selection probability and regional attractiveness indicates that there is little correlation between the two proposed features, which can also be observed from Figure 11(g–h).



(a) length (b) road density (c) degree

(d) closeness centrality (e) betweenness centrality (f) road level

(g) travel path selection probability (h) regional attractiveness

Figure 11. The visualization of the road importance features.

4.2. Experimental design

4.2.1. Baselines and evaluation indicators

Four representative baseline methods were used to evaluate the effectiveness of the proposed method: the stroke-based method (Liu *et al.* 2010), RBF neural network-based method (Jeatrakul and Wong 2009) (hereafter abbreviated as RBF), Backward Propagation Network-based method (Zhou and Li 2014) (abbreviated as BP), and Support Vector Machine-based method (Zhou and Li 2017) (abbreviated as SVM). Additionally, the recent road network selection method based on GCNs by Zheng *et al.* (2021) was compared in ablation experiments. Six commonly used indicators, including *Precision*, *Recall*, *F*1-*score*, *Similarity* (Zhou and Li 2011), *Commission error* (Tian *et al*. 2019), and *Connectivity* (Li and Zhou 2012), were used to evaluate the performance of different methods.

1. *Precision*: It is defined as the proportion of correctly selected roads (positive samples) in the selected road network, which evaluates the algorithm's capability to discriminate against negative samples (i.e. roads that should not be selected).
2. *Recall*: It is defined as the proportion of roads in the target road network that are correctly selected by the algorithm, which measures the algorithm's ability to find all positive samples.
3. *F*1-*score*: It is the harmonic mean of precision and recall, which comprehensively measures the algorithm's ability to identify positive samples and is calculated as:

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

4. *Similarity*: If the selected road network retains the main roads of the target road network, the similarity value will be larger, indicating more consistency of the selected road network to the target road network (Zhou and Li 2011). The value of *Similarity* ranges from 0 to 1 and is calculated as:

$$Similarity = \frac{A \cap B}{A + B - A \cap B} \tag{7}$$

where A and B denote the total length of roads in the selected and target road networks, respectively, and $A \cap B$ represents the total length of common roads in the selected and target road networks.

5. *Commission Error*: It measures the difference between the selected road network and the target road network with a value ranging from 0 to 1 (Tian *et al*. 2019). A lower commission error value indicates a smaller difference between the selected and target road networks. It is calculated as:

$$Commission\ \ Error = \frac{A \cap \sim B}{A} \tag{8}$$

where A and B are defined as in the *Similarity* metric, and $A \cap \sim B$ represents the total length of roads that are in the selected road network but not in the target road network.

6. *Connectivity*: It measures the connectivity of the road network (Li and Zhou 2012) and ranges from 0 to 1. The larger the connectivity value, the better the road network topology. If there exists a path between any two roads in the road network, the connectivity value is 1. It is calculated as:

$$Connectivity = \frac{\sum_{i \in N} \sum_{j \in N: j \neq i} a_{ij}}{N(N-1)} \tag{9}$$

where N is the total number of roads in the road network, a_{nd} indicates the connection status of roads i and j. If there exists a path from roads i to j, a_{ij} equals 1; otherwise, a_{ij} equals 0.

4.2.2. Experimental settings

In the experiments, a quarter of the source road network data in the upper left corner was selected as the testing data for the machine learning-based method and the proposed method, and the rest was used for the training of these methods. The target scales for the experiments are 1:50,000 and 1:200,000. The parameter settings of the proposed method and the baseline methods are shown in Table 1. The grid search algorithm was used to find the optimal hyperparameter combinations for the proposed method. For the stroke-based method, the number of roads to be selected is determined according to Töpfer's principles of selection (Töpfer and Pillewizer 1966). The hidden layer sizes of RBF, BP, SVM, and GCN models were all set to 128. The Adam optimizer (Kingma and Ba 2014) was employed to calculate and update the weights and the initial learning rate was set to 0.01. The cross-entropy function (Shore and Johnson 1980) was used as the loss function for RBF-based and BP-based methods and the proposed method utilized the improved loss function presented in Equation (5).

4.3. Results and comparison analysis

The road network selection results of the proposed method are visualized at target scales of 1: 50,000 and 1:200,000. As shown in Figure 12, it indicates that the proposed method effectively retains the primary roads and preserves good spatial connectivity and topological structure of the road network at both target scales. Although some hanging roads exist in the road network selection results, they are mostly located near boundaries and in sparsely populated areas, which is acceptable. In addition, we find that the proposed method performs better for road network selection at the medium scale (1:50,000) compared to the small scale (1:200,000). The possible reason is that the functional semantic features may play a more important role in road network selection at medium or large scales, whereas graphical structure features (such as road length and road level) have a greater impact at small scales.

Table 1. Parameter settings of the proposed methods and the baseline methods.

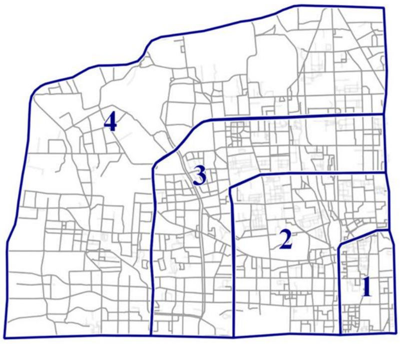
No.	Methods	Parameters
1	Stroke-based method (Liu *et al.* 2010)	The number of roads to be selected at the target scale is determined according to Töpfer's principles of selection.
2	RBF-based method (Jeatrakul and Wong 2009)	Feature number = 8, hidden size = 128, epochs = 1000, optimizer = Adam, loss function = cross entropy function.
3	BP-based method (Zhou and Li 2014)	Feature number = 8, epochs = 1000, optimizer = Adam, loss function = cross entropy function.
4	SVM-based method (Zhou and Li 2017)	Feature number = 8, kernel = linear, C = 1.0, gamma = auto.
5	The proposed method	Feature number = 8, convolutional layer number = 2, hidden size = 128, epochs = 1000, learning rate = 0.01, optimizer = Adam, $\lambda_1 = 10$, $\lambda_2 = 1$, loss function = the proposed loss function (see Equation (5)).

(a) (b)

Figure 12. Road network selection results by the proposed method of (a) at the target scale of 1:50,000, and (b) at the target scale of 1:200,000.

As shown in Figure 12(a), the road density distribution is uneven, with some areas having denser roads. For example, the selected roads in the circled areas *A* and *B* are denser than in other areas. This is because these two areas contain plentiful POIs and attract more crowds there, thus bringing large traffic volume. Area *A*, called Zhongguancun Science Park(Z-park), gathers the largest number of high-tech enterprises in China, while area *B*, the government activity concentration area of Xicheng District, gathers a large number of administrative departments. Despite some roads in the two areas being short and low-level, they are frequently chosen routes and play significant roles in the road network at this scale. By considering multiple road features comprehensively, the proposed method adaptively selects roads with varying densities and ensures the suitability of their display at different target scales. This can also be observed in Figure 12(b), where the selected roads in regions *A* and *B* are less densely distributed at the smaller scale. Further, we checked the official map at the scale of 1:50,000 and found that roads in area *A* and area *B* are similar to the selection results of our method. There may be densely distributed roads in the large-scale maps (e.g. 1:50,000). In this case, the concern about the visual conflict problem of the map is unnecessary, because there is sufficient space for the representation of the selected roads in these target scale maps. In smaller-scale maps, visual conflict issues need to be considered due to the limitation of space on a map. However, in the selection results of our method on smaller-scale maps (e.g. 200,000), the selected roads are sparsely distributed (see Figure 12(b)), which also meets the needs of map representation. In the proposed method, we apply the refinement steps described in Section 3.4 to remove the short hanging roads that could not be displayed at the target scale map from the selection results, which can further reduce the occurrence of visual conflicts.

Furthermore, we evaluated the density change of four partitions (see Figure 13(a)) at the source scale and two target scales. The result shown in Figure 13(b) demonstrates that the selected roads largely preserve their relative density before and after



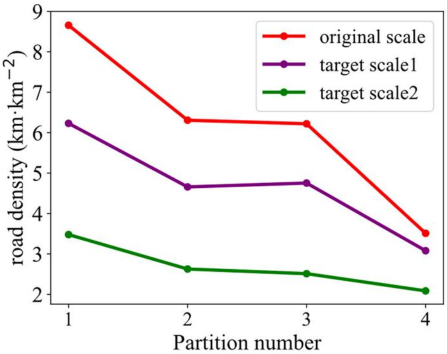


Figure 13. The road density contrasts before and after the selection of (a) partitions in the test area, and (b) road density in different partitions at different scales.

Table 2. Evaluation of different road network selection results at the target scale of 1:50,000.

Methods	Precision	Recall	F1-score	Similarity	Commission error	Connectivity
Stroke-based method (Liu *et al*. 2010)	**0.8789**	0.7559	0.8128	0.7220	**0.0899**	0.9777
RBF (Jeatrakul and Wong 2009)	0.8303	0.9098	0.8682	0.7988	0.1658	0.9688
RBF with all features	0.8432	0.9104	0.8755	0.8158	0.1454	0.9815
BP (Zhou and Li 2014)	0.8415	0.9104	0.8746	0.8056	0.1567	0.9684
BP with all features	0.8540	0.9215	0.8864	0.8298	0.1355	0.9774
SVM (Zhou and Li 2017)	0.7912	0.9410	0.8596	0.7899	0.1838	0.9795
SVM with all features	0.8217	0.9393	0.8766	0.8141	0.1562	0.9903
The proposed method	0.8764	**0.9448**	**0.90923**	**0.8570**	0.1179	**1.0000**

selection. Overall, the results demonstrate that the proposed road network selection method can automatically determine the number of roads to be selected at different target scales, maintain road connectivity, and preserve the relative density of roads in different areas.

To further evaluate the effectiveness of the proposed method, a quantitative comparison with four representative baseline methods (Stroke-based method, RBF-based method, BP-based method, and SVM-based method) was conducted at target scales of 1:50,000 and 1:200,000 using the above-introduced evaluation indicators. The RBF, BP, and SVM methods were evaluated in two experiments: using features in the reference paper and using all features in this paper). The evaluation results of different road network selection methods at target scales are listed in Table 2 and Table 3, respectively, with the bolded numbers indicating the best performance for that metric (the subsequent Tables have the same meaning of bold values).

As shown in Table 2 and Table 3, the proposed method outperformed baseline methods across almost all metrics, demonstrating its effectiveness in preserving road topology, overall structure, and shape at the target scales of 1:50, 000, and 1:200, 000. The proposed method achieved a connectivity value of 1, which indicates that the roads selected by our approach are completely connected. In contrast, the road networks selected by baseline methods contain disconnected roads. This disparity underscores the superior ability of the proposed method to ensure the continuity of the

Table 3. Evaluation of different road network selection results at the target scale of 1:200,000.

Methods	Precision	Recall	F1-score	Similarity	Commission error	Connectivity
Stroke-based method (Liu *et al.* 2010)	0.7467	0.8467	0.7935	0.6875	0.2283	0.9984
RBF (Jeatrakul and Wong 2009)	0.7948	0.8394	0.8165	0.7114	0.2048	0.9877
RBF with all features	0.7946	0.8381	0.8158	0.7272	**0.1961**	0.9951
BP (Zhou and Li 2014)	0.7792	0.8525	0.8142	0.7029	0.2223	0.9810
BP with all features	0.7938	0.8342	0.8135	0.7078	0.2094	0.9827
SVM (Zhou and Li 2017)	0.7956	0.8433	0.8188	0.7060	0.2157	0.9902
SVM with all features	0.7850	0.8486	0.8156	0.7124	0.2234	0.9928
The proposed method	**0.7984**	**0.9204**	**0.8550**	**0.7666**	0.2054	**1.0000**

selected road network, which is important for urban planning and navigation-oriented map applications.

At the target scale of 1:50,000, the machine learning-based methods including our proposed method achieved significant improvements over the stroke-based methods in the *recall*, *F1-score*, and *similarity* indicators, indicating their superior capability in avoiding the omission of important small roads at a large or medium scale. Because the stoke-based method prioritizes continuous and primary roads, it excelled in *precision* and exhibited lower *commission error*, whereas it may have omitted some important short roads, resulting in the lowest *recall* value at this scale. Among the machine learning-based baseline methods (RBF, BP, and SVM), they performed better when utilizing all road features compared to solely relying on features from the reference paper. At the target scale of 1:200,000, the stroke-based method performed worst on most evaluation metrics, while the machine learning methods performed better. The machine learning-based baseline methods had comparable performance when using all features and only features from the reference paper, which is different from the results at the scale of 1:50,000. The proposed method had significant improvements in *recall* and *similarity*, suggesting its effectiveness in learning cartographic rules and knowledge implicitly and better adapting to the road network selection task at the target scale than baselines. Moreover, the overall results at the scale of 1:50,000 are better than those at the scale of 1:200,000. It indicates that the road network selection results obtained by current methods have relatively large differences with the official road networks at a smaller scale which needs further exploration.

Next, we visualized the selection results of the machine learning-based methods at the target scale of 1:200,000. As shown in Figure 14, the proposed method is more effective in preserving the spatial shape, topological relationship, and continuity of the selected roads compared to baseline methods. It can be observed that road segments circled in red were better selected by the proposed method, but partially erroneously removed by BP, RBF, and SVM methods, leading to a disruption of the connectivity and incorrect topology relationship of the road network. This further illustrates the effectiveness and applicability of the proposed method for road network selection.

4.4. Ablation experiments

To investigate the effects of the proposed functional features of regional attractiveness and travel path selection probability on the road network selection results, the

Figure 14. Road network selection results of four machine learning methods on test data at 1:200,000: (a) RBF-based method; (b) SVM-based method; (c) BP-based method; (d) The proposed method.

ablation experiments were conducted and the performance of the proposed road selection model with different road feature combination strategies at the target scales of 1:50,000 and 1:200,000 were evaluated. It should be mentioned that in the ablation experiments, in order to directly evaluate the influence of different features on road network selection, the selection results under different feature combinations were not post-processed by the refinement process.

Table 4 presents the evaluation results for road network selection at the target scale of 1:50,000. It demonstrates that the GCN model incorporating all road features enhances the road network selection accuracy compared to models utilizing only graphical structure features. The combinations of graphical structure features and regional attractiveness alone yielded minimal improvement. Regional attractiveness requires further incorporation of travel path selection probability to mitigate the selection of disconnected roads because it may increase the probability of selecting some short roads that are distributed with important POIs. At the target scale of 1:200,000

Table 4. Evaluation of different feature combinations at the target scale of 1:50,000.

Feature combination strategy	Precision	Recall	F1-score	Similarity	Commission Error	Connectivity
Graphical structure features (Zheng *et al.* 2021)	0.8451	0.9115	0.8770	0.8169	0.1424	0.9866
Graphical structure features with regional attractiveness	0.8485	0.9104	0.8783	0.8179	0.1411	0.9855
Graphical structure features with travel path selection probability	0.8632	0.9136	0.8877	0.8358	0.1240	**0.9916**
All road features	**0.8751**	**0.9209**	**0.8974**	**0.8480**	**0.1171**	0.9905

Table 5. Evaluation of different feature combinations at the target scale of 1:200,000.

Feature combination strategy	Precision	Recall	F1-score	Similarity	Commission Error	Connectivity
Graphical structure features (Zheng *et al.* 2021)	0.7827	0.8512	0.8155	0.7154	0.2235	1.0000
Graphical structure features with regional attractiveness feature	0.7929	0.8498	0.8204	0.7254	0.2131	1.0000
Graphical structure features with travel path selection probability	0.7939	0.8551	0.8233	0.7290	0.2123	1.0000
All road features	**0.8007**	**0.8551**	**0.8270**	**0.7339**	**0.2078**	**1.0000**

(Table 5), the GCN model with all features still achieved the best results. However, its performance was not significantly different from the GCN model combining graphical structure features and travel path selection probability, which achieved the highest *Recall* value. Additionally, the first two feature combination strategies exhibited similar *Recall* values. These indicate that the use of regional attractiveness did not improve the road network selection quality and the travel path selection probability has a positive impact on road network selection at a smaller scale.

The above analysis reveals that the proposed two functional semantic features play an important role in road network selection at large or medium scales (e.g. 1:50,000). However, at smaller scales (e.g. 1:200,000), their importance differs, with travel path selection probability contributing more than regional attractiveness. Therefore, assigning different weights of road features at different target scales is essential for practical applications, which is challenging for traditional road network selection methods (e.g. the stroke-based method), while the proposed GCN-based method and other machine learning methods can automatically learn the optimal weights of different road features.

4.5. Hyperparameter sensitivity

The proposed road network selection model contains five crucial hyperparameters: (1) the number of neurons in the hidden layer, (2) the number of training epochs, (3) the learning rate, (4) λ_1 in the connectivity loss, and (5) regularization weight λ_2. To demonstrate model sensitivity, we present a hyperparameter analysis at a 1:200,000 scale, evaluating performance via the F1-score and similarity. Figure 15 displays the model performance attained by systematically varying each hyperparameter, holding all other model configuration details constant. We assess the number of neurons number from 2-2048 (powers of 2), training epochs from 5-1000, learning rates from 0.001-0.2, and λ_1 and λ_2 from 0.01-10.

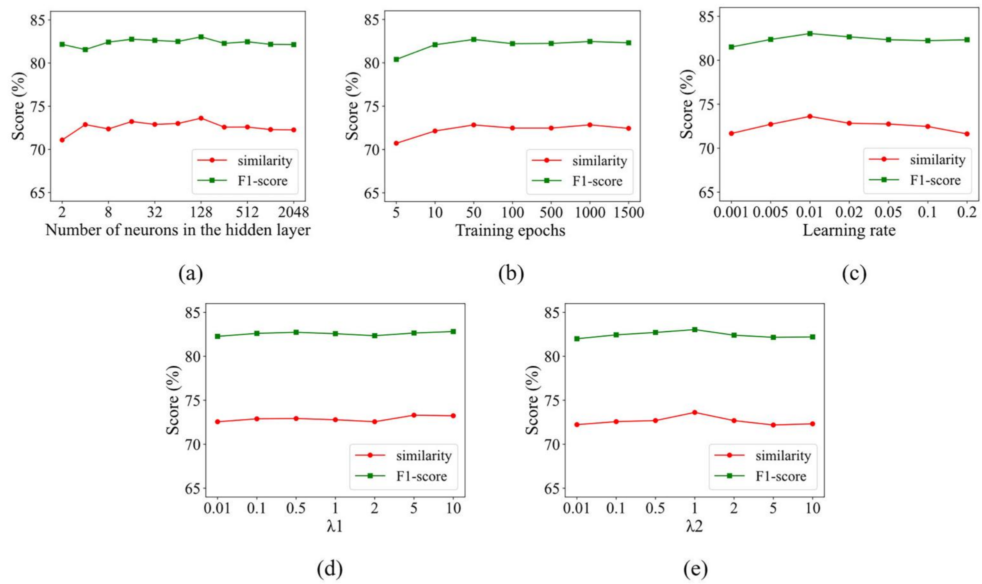


Figure 15. The performance of the proposed road network selection model under different hyper-parameter settings. (a) Impact of different numbers of neurons in the hidden layer, (b) Impact of different training epochs, (c) Impact of different learning rates, (d) Impact of different λ_1, and (e) Impact of different λ_2.

As shown in Figure 15, the F1-score and similarity indicators reached maximum scores at 128 neurons in the hidden layer. They reached relatively stable levels when the training epoch was above 50. The optimal learning rate in the road network selection at 1:200,000 was 0.01. Regarding the hyperparameters of λ_1 and λ_2 used to adjust the connectivity loss and regularization degree, the proper values were 5 and 10 for λ_1 and 1 for λ_2. Overall, F1-score and similarity are not so sensitive to the values of λ_1 and λ_2, likely because they are primarily used to maintain the connectivity of the road selection results.

5. Discussions

This study proposes an automatic road network selection method based on GCNs that considers both graphical structure features and social functional features of roads. In this section, we first analyze the significance of these two types of road features across different scales. We then explore and discuss alternative or complementary datasets that can be used to evaluate the social functional importance of roads. Lastly, we explore the transferability and generalization of the proposed road network selection model through additional experiments.

5.1. Significance of road features across different scales

This paper investigates the impact of human activities on road network selection from the perspective of human travel by leveraging trajectories and POIs. The results illustrate the effectiveness of the proposed road features (i.e. travel path selection probability and attractiveness) at a large or medium scale. However, the effectiveness of the two road features, especially regional attractiveness may decrease at a small scale.

In existing road network selection methods, the graphical structure features (i.e. road length, road level, road density, degree, closeness centrality, and betweenness centrality) are commonly used. These features are conducive to preserving the overall spatial structure of the road network, but they often overlook detailed roads with lower road levels or shorter lengths. As a result, some roads that are crucial for human travel but are shorter or of lower levels may be omitted in the road network selection process. However, by integrating the graphical structure features with functional semantic features, the selected road network not only preserves well topological structure but also includes roads with less apparent graphical characteristics yet vital for human travel. This is also consistent with the requirement for larger-scale maps to retain more detailed information.

In contrast, smaller-scale maps focus on more macroscopic characteristics and the overall structure of the road network. Therefore, it is crucial to consider the characteristics of road networks at different scales and select appropriate features or assign different weights to the features for road network selection. For example, the weights of road features in different areas (e.g. urban and rural areas, dense and sparse areas) or at different scales (e.g. 1:50,000 and 1:200,000) should be different. The proposed method can effectively estimate the feature weights for roads in different types of areas during the training process, as long as the training dataset covers such roads.

5.2. Exploring alternatives or complements for road functional importance evaluation

The evaluation of road importance is a crucial aspect of road network selection. In the era of ubiquitous social sensing and mobile positioning, the availability of various social sensing data provides opportunities to assess the social functional importance of roads from the perspective of human perception. In this study, we use POIs and trajectory data to evaluate the social functional importance of roads, where POIs serve as potential attractions for human activities carried out on roads, and trajectories (i.e. traffic flow) are the direct reflections of actual road usage. However, it is worth mentioning that trajectory data and POIs are only two types of commonly used data that we can use to measure the social value or functional importance of roads. Other datasets, such as land use data, mobile phone usage data, and social opinion data, which can reflect the social value of roads, can also be employed as alternatives or complements to POIs or trajectory data in our proposed method. This does not affect the overall framework presented in this paper. For example, the integration of land use data, which reveals the distribution of residential, commercial, industrial, and other types of functional areas, can provide information about the types of activities and establishments along specific roads. Comments about POIs from online platforms (e.g. https://www.dianping.com) can be used to adjust the weight of POIs, influencing the importance of roads. Mobile phone usage data provide a more comprehensive perception of road resource usage (Yan *et al*. 2022) and can be used to evaluate road importance for road network selection when such data are available. Additionally, social opinion data can capture public sentiment and perceptions about the importance of certain roads, which can be used to ament road importance assessment.

In summary, datasets reflecting the social functional values of roads can be adopted as alternatives or complements to POIs and trajectories used in this paper. It should also be emphasized that graphical structural features are the dominant features in ensuring the integrity and topological structure of the selected road network, whereas the social functional value of roads serves more as a correction and supplement to the road importance evaluation. In applications, the choice of road features for measuring road importance should be carefully considered according to specific demands.

5.3. Generalization capability and transferability of the proposed method

This section investigates the transferability and the effectiveness of the proposed selection model for road networks in other cities. The model was trained using the road network data in Beijing, and we tested its performance in the selection of road networks in other cities, such as Wuhan, in China. Experiments were conducted within the confines of the Third Ring Road area of Wuhan at two corresponding target scales, with experimental settings identical to those used in Beijing. Specifically, the experiment conducted at the 1:200,000 scale utilized all road data within the Third Ring Road area of Wuhan, while the experiment at the 1:50, 000 scale was confined to road data within the central district of Hankou as an example.

Quantitative results of the road network selection in Wuhan are presented in Table 6. It indicates that the model trained in Beijing also exhibited superior performance in Wuhan, with certain metric values surpassing those observed in Beijing. Figure 16 illustrates the road network selection results at different target scales, which show that the overall structure and shape of the selected road networks were well maintained at both target scales. As shown in Figure 16(a), the selected roads within the red-circled block are more dense than other regions. This observation can be attributed to the area's status as the most prosperous region, characterized by heavy traffic flows and plenty of POIs. The roads within this area are more complex compared to other areas, which is why this area was chosen for the road network selection at the target scale of 1:50,000 in addition to guaranteeing the visual effect. The road network selection results in Wuhan demonstrate the transferability, effectiveness, and great potential of the proposed method. It provides a beneficial and meaningful attempt for Geographical Artificial Intelligence models empowering cartography and demonstrates contributions to intelligent and automatic map generalization.

Table 6. The quantitative evaluation results in Wuhan using the trained model from Beijing.

Scale	Precision	Recall	F1-score	Similarity	Commission error	Connectivity
1:200, 000	0.8644	0.8960	0.8799	0.8153	0.1210	1.0000
1:50, 000	0.8766	0.9422	0.9082	0.8333	0.1287	1.0000

Figure 16. Road network selection results in Wuhan using the models trained in Beijing of (a) at the target scale of 1:200,000, and (b) at the target scale of 1:50,000.

6. Conclusions

In this paper, we introduced two new road features of travel path selection probability and regional attractiveness to measure human cognition of road functional importance and use these features for road network selection. We established a GCN-based model to learn how many roads should be selected and which roads to select at the target scale, thereby reducing the subjectivity of parameter setting and improving the intelligence and accuracy of road network selection. By considering the social functional values of roads, the road network selection results are more consistent with human cognition, which can better meet practical application requirements. We compared the performance of our method with four representative methods and investigated the effects of the proposed two functional features on road network selection. The experimental results show the effectiveness and superiority of the proposed method. We found that the travel path selection probability and regional attractiveness are more conducive to road network selection at a large scale than at a small scale. We also explored the transferability and generalization ability of the proposed method for road network data in different cities. The results show that when the road network selection model trained on the road network data of Beijing is applied to one new city, i.e. Wuhan, the model also performed well under different scales, and has great potential for automatic selection of road networks.

Although the proposed method exhibits some superiority over the available methods, some limitations still exist. First, the map-matching pre-processing operation for computing the travel path selection probability is time-consuming. Parallel computing and high-performance computing can be used to improve computing efficiency in the future. Second, the construction of road network datasets for training the proposed method requires labeling part of the road network, which is slightly more laborious. In future work, we will modify our proposed method to further improve the automation of road network selection by leveraging advanced semi-unsupervised techniques.

Moreover, experiments on the road network selection in more different cities and even countries still need to be conducted to further explore the applicability and generalization of our proposed method on more different and complex road network structures and more target scales.

Acknowledgments

The authors gratefully acknowledge the comments from the editor and the reviewers.

Disclosure statement

The authors reported no potential conflict of interest.

Funding

This study was supported by the Funds of the National Key Research and Development Program of China (No. 2022YFB3904203), the National Natural Science Foundation of China (No. 42271462, 42171441, and 42271485), the Science and Technology Innovation Program of Hunan Province (No. 2024AQ2026) and the Hunan Provincial Natural Science Foundation of China (No. 2024JJ8343, 2021JJ40727).

Notes on contributors

Jianbo Tang is an associate professor at Central South University. His research interests include trajectory data analysis, geographical scene modeling, and spatiotemporal data mining. He contributed to the conceptualization, methodology, writing (review & editing), and funding acquisition.

Min Deng is a professor at Central South University. His current research interests include spatiotemporal data mining and map generalization. He contributed to the conceptualization, methodology, resources, and funding acquisition.

Ju Peng is a Ph.D. candidate at Central South University. Her research interests include map generalization, trajectory data mining, and spatiotemporal data mining. She contributed to the conceptualization, data processing, experiment analysis, visualization, and writing (original draft, review & editing).

Huimin Liu is an associate professor at Central South University. Her research interests include map generalization and spatiotemporal data mining. She contributed to data processing, data collection, visualization, and funding acquisition.

Xuexi Yang is an Associate Professor at Central South University. His research interests include knowledge graph, and spatiotemporal data mining. He contributed to data collection, visualization, and funding acquisition.

Xueying Chen is a postgraduate student at Central South University, Changsha. Her research interest includes map generalization and trajectory data analysis. She contributed to data collection, experiment analysis, and visualization.

ORCID

Jianbo Tang http://orcid.org/0000-0002-9780-0439
Ju Peng http://orcid.org/0000-0001-7520-1534

Huimin Liu http://orcid.org/0000-0002-7419-5838
Xuexi Yang http://orcid.org/0000-0001-5531-5841

Data and codes availability statement

The data and codes that support the findings of this study are available on 'figshare.com', with the identifier at the public link: https://doi.org/10.6084/m9.figshare.23654001.

References

Ai, T., 2022. Some thoughts on deep learning empowering cartography. *Journal of Geography and Cartography*, 5 (2), 25–40.

Benz, S.A., and Weibel, R., 2014. Road network selection for medium scales using an extended stroke-mesh combination algorithm. *Cartography and Geographic Information Science*, 41 (4), 323–339.

Brandes, U., 2008. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30 (2), 136–145.

Brintrup, A., Ledwoch, A., and Barros, J., 2016. Topological robustness of the global automotive industry. *Logistics Research*, 9 (1), 1–17.

Chen, J., *et al.*, 2009. Selective omission of road features based on mesh density for automatic map generalization. *International Journal of Geographical Information Science*, 23 (8), 1013–1032.

Çolak, S., Lima, A., and González, M.C., 2016. Understanding congested travel in urban areas. *Nature Communications*, 7 (1), 10793.

Courtial, A., Touya, G., and Zhang, X., 2023. Deriving map images of generalised mountain roads with generative adversarial networks. *International Journal of Geographical Information Science*, 37 (3), 499–528.

Foerster, T., Stoter, J., and Kraak, M.J., 2010. Challenges for automated generalisation at European mapping agencies: a qualitative and quantitative analysis. *The Cartographic Journal*, 47 (1), 41–54.

Gülgen, F., and Gökgöz, T., 2011. A block-based selection method for road network generalization. *International Journal of Digital Earth*, 4 (2), 133–153.

Guo, X., *et al.*, 2023. A method for intelligent road network selection based on graph neural network. *ISPRS International Journal of Geo-Information*, 12 (8), 336.

Han, Y., *et al.*, 2020. Application of AHP to road selection. *ISPRS International Journal of Geo-Information*, 9 (2), 86.

Hu, Y.G., *et al.*, 2007. Selective omission of road features based on mesh density for digital map generalization. *Acta Geodaetica et Cartographica Sinica*, 36 (3), 351–357.

Jeatrakul, P., and Wong, K.W., 2009. Comparing the performance of different neural networks for binary classification problems. In: *2009 Eighth International Symposium on Natural Language Processing*, 111–115.

Jiang, B., 2015. The fractal nature of maps and mapping. *International Journal of Geographical Information Science*, 29 (1), 159–174.

Jiang, B., and Claramunt, C., 2004. A structural approach to the model generalization of an urban street network. *GeoInformatica*, 8 (2), 157–171.

Jiang, B., and Harrie, L., 2004. Selection of streets from a network using self-organizing maps. *Transactions in GIS*, 8 (3), 335–350.

Jiang, B., Zhao, S., and Yin, J., 2008. Self-organized natural roads for predicting traffic flow: a sensitivity study. *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (07), P07008.

Karsznia, I., Wereszczyńska, K., and Weibel, R., 2022. Make it simple: effective road selection for small-scale map design using decision-tree-based models. *ISPRS International Journal of Geo-Information*, 11 (8), 457.

Kingma, D.P., and Ba, J., 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv: 1412.6980*.

Kipf, T.N., and Welling, M., 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, Z., and Openshaw, S., 1992. Algorithms for automated line generalization1 based on a natural principle of objective generalization. *International Journal of Geographical Information Systems*, 6 (5), 373–389.

Li, Z., and Zhou, Q., 2012. Integration of linear and areal hierarchies for continuous multi-scale representation of road networks. *International Journal of Geographical Information Science*, 26 (5), 855–880.

Liu, X., Ai, T., and Liu, Y., 2009. Road density analysis based on skeleton partitioning for road generalization. *Geo-Spatial Information Science*, 12 (2), 110–116.

Liu, X., Zhan, F.B., and Ai, T., 2010. Road selection based on Voronoi diagrams and "strokes" in map generalization. *International Journal of Applied Earth Observation and Geoinformation*, 12, (2010), S194–S202.

Liu, Y., and Li, W., 2019. A new algorithms of stroke generation considering geometric and structural properties of road network. *ISPRS International Journal of Geo-Information*, 8 (7), 304.

Lou, Y., *et al*., 2009. Map-matching for low-sampling-rate GPS trajectories. In: Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, 352–361.

Mackaness, W.A., and Beard, K.M., 1993. Use of graph theory to support map generalization. *Cartography and Geographic Information Systems*, 20 (4), 210–221.

Mathew, B.S., and Isaac, K.P., 2014. Optimisation of maintenance strategy for rural road network using genetic algorithm. *International Journal of Pavement Engineering*, 15 (4), 352–360.

Peng, J., *et al*., 2023. Exploring crowd travel demands based on the characteristics of spatiotemporal interaction between urban functional zones. *ISPRS International Journal of Geo-Information*, 12 (6), 225.

Peng, J., *et al*., 2024. A movement-aware measure for trajectory similarity and its application for ride-sharing path extraction in a road network. *International Journal of Geographical Information Science*, 1–25.

Porta, S., Crucitti, P., and Latora, V., 2006. The network analysis of urban streets: a dual approach. *Physica A: Statistical Mechanics and Its Applications*, 369 (2), 853–866.

Richardson, D., and Thomson, R., 1996. Integrating thematic, geometric, and topological information in the generalization of road networks. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 33 (1), 75–83.

Schipperijn, J., *et al*., 2010. Factors influencing the use of green space: Results from a Danish national representative survey. *Landscape and Urban Planning*, 95 (3), 130–137.

Shore, J., and Johnson, R., 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26 (1), 26–37.

Siła-Nowicka, K., *et al*., 2016. Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science*, 30 (5), 881–906.

Thomson, R.C., and Richardson, D.E., 1995. A graph theory approach to road network generalisation. In: Proceeding of the 17th International Cartographic Conference, 1871–1880.

Thomson, R.C., and Richardson, D.E., 1999. The 'good continuation' principle of perceptual organization applied to the generalization of road networks. In: Proceedings of the 19th International Cartographic Conference, 1215–1223.

Tian, J., *et al*., 2016. Road density partition and its application in evaluation of road selection. *Geomatics and Information Science of Wuhan University*, 41 (9), 1225–1231.

Tian, J., *et al*., 2019. A comparative study of two strategies of road network selection. *Geomatics and Information Science of Wuhan University*, 44 (2), 310–316.

Töpfer, F., and Pillewizer, W., 1966. The principles of selection. *The Cartographic Journal*, 3 (1), 10–16.

Touya, G., Zhang, X., and Lokhat, I., 2019. Is deep learning the new agent for map generalization? *International Journal of Cartography*, 5 (2-3), 142–157.

Van Nimwegen, E., Crutchfield, J.P., and Mitchell, M., 1999. Statistical dynamics of the royal road genetic algorithm. *Theoretical Computer Science*, 229 (1-2), 41–102.

Weiss, R., and Weibel, R., 2014. Road network selection for small-scale maps using an improved centrality-based algorithm. *Journal of Spatial Information Science*, 9 (9), 71–99.

Xu, Z., *et al.*, 2012. Road selection based on evaluation of stroke network functionality. *Acta Geodaetica et Cartographica Sinica*, 41(5), 769–776.

Xu, Z., *et al.*, 2018. A method for automatic road selection combined with POI data. *Journal of Geo-Information Science*, 20 (2), 159–166.

Yan, X., *et al.*, 2022. Revealing spatiotemporal matching patterns between traffic flux and road resources using big geodata-a case study of Beijing. *Cities*, 127, 103754.

Yang, B., Zhang, Y., and Lu, F., 2014. Geometric-based approach for integrating VGI POIs and road networks. *International Journal of Geographical Information Science*, 28 (1), 126–147.

Yang, M., Ai, T., and Zhou, Q., 2013. A method of road network generalization considering stroke properties of road object. *Acta Geodaetica et Cartographica Sinica*, 42 (4), 581–587.

Yi, W., Dong, P., and Wang, J., 2018. Node risk propagation capability modeling of supply chain network based on structural attributes. *In: Proceedings of the 2018 9th International Conference on E-business, Management and Economics*, 50–54.

Yu, W., and Chen, Y., 2022. Data-driven polyline simplification using a stacked autoencoder-based deep neural network. *Transactions in GIS*, 26 (5), 2302–2325.

Yu, W., *et al.*, 2020. Road network generalization considering traffic flow patterns. *International Journal of Geographical Information Science*, 34 (1), 119–149.

Zheng, J., *et al.*, 2021. Deep graph convolutional networks for accurate automatic road network selection. *ISPRS International Journal of Geo-Information*, 10 (11), 768.

Zhou, Q., and Li, Z., 2011. Evaluation of properties to determine the importance of individual roads for map generalization. *In*: M.P. Peterson, ed. *Advances in cartography and GIScience*. Heidelberg: Springer, 459–475.

Zhou, Q., and Li, Z., 2012. A comparative study of various strategies to concatenate road segments into strokes for map generalization. *International Journal of Geographical Information Science*, 26 (4), 691–715.

Zhou, Q., and Li, Z., 2014. Use of artificial neural networks for selective omission in updating road networks. *The Cartographic Journal*, 51 (1), 38–51.

Zhou, Q., and Li, Z., 2017. A comparative study of various supervised learning approaches to selective omission in a road network. *The Cartographic Journal*, 54 (3), 254–264.