

Tang J, Deng M, Peng J, et al. Automatic road network selection method considering functional semantic features of roads with graph convolutional networks[J]. International Journal of Geographical Information Science, 2024, 38(11): 2403-2432.

Paper implementation details document

Due to data privacy and national policy issues, the multi-scale road network data and taxi trajectory data used in the paper cannot be publicly shared. To facilitate the reproduction and application of the method proposed in this study, the implementation details of are provided below.

(1) Study Area 1: Within the Fifth Ring Road of Beijing, China. The data processing for Study Area 2 (Wuhan) follows the same workflow. And the model trained in Beijing was directly applied when selecting roads in Wuhan to verify the generalization ability of the proposed method.

(2) Experimental Data: Road network data at three scales: initial scale 1:10,000, target scale 1 1:5,000, target scale 2: 1:20,000. The initial scale road network data comes from OSM, and the target scale road network data comes from Tiandi Map (the initial OSM data can be processed based on the online map)

(3) Data Preprocessing:

a) Single line road generation (road centerline extraction): Since the downloaded OSM data is at the road level and contains road information such as different directions, primary roads and auxiliary roads, and existing road selection research is mainly based on single-line roads, the complex OSM multi-line roads are first processed into single-line roads.

b) Road topology check and construction: Check and reconstruct the topology of the generated roads, including deleting isolated roads, processing road pseudo nodes, etc.

c) Target scale road network data processing: The corresponding target scale road network data can be obtained from Tianditu Map (if any), and then align the target scale roads with the original scale roads to ensure that the target scale data within the original scale data. If the multi-scale road network data from Tianditu Map or other national office geographical information agency are unavailable, manually process the road network with reference to the corresponding scale from online Tianditu Map (time-consuming but necessary and useful). Alternatively, you can also use multi-scale road network data in other research areas as long as you have.

d) Repeat steps a)-(c) on the road network data at target scale 1. This aims to achieve road selection from target scale 1 to target scale 2.

(4) Road dual graph construction:

a) The road dual graph organizes the road network data in a graph structure, where road segments are expressed as nodes on the graph and the connection relationships between roads are expressed as edges on the graph.

b) To construct the road dual graph, you first need to obtain the road segments and their road connection nodes, and assign them unique ID numbers, and then construct the graph file data with the ‘.gpickle’ suffix using the provided code or your own implementation.

c) Specifically, the road is abstractly expressed as a dual graph ($G=(V, E)$), where V (graph node) represents the road segment and E (graph edge) represents the connection relationship between road segments. In this study, the Data module provided by torch_geometric is used to organize and express the data, where x represents the road segment

and its characteristics, and edge_index stores the connection relationship between road segments.

(5) Road Network Feature Calculation

- a) Road length: Calculated based on the road geometry.
- b) Road grade/level: OSM data includes road grade information, but the classification needs to be converted to numerical grades. For roads with missing road grade information, the road grade information is supplemented based on surrounding roads, road names or online maps.
- c) Road density: Calculated based on the road density calculation paper referenced in our article. Other road density calculation methods can also be used.
- d) Road topology information: Calculated based on the constructed road dual graph using the python networkx package, including:
 - i. Centrality
 - ii. Betweenness centrality
 - iii. Proximity centrality:
- e) Regional attractiveness: Extract POIs within a certain buffer zone of the road and calculate the regional attractiveness value of each road according to the formula provided in the article.
- f) Travel path selection probability (at least one week of vehicle trajectory data is required): Match the vehicle trajectory data to the road using map matching algorithm, then count the traffic flow on the road segment, and calculate the selection probability of each road according to the formula provided in the article.

(6) Model construction and training:

A simple two-layer graph convolutional neural network is used as the road network selection model. We adopt the cross entropy loss function (which measures the effect of road selection results in terms of accuracy), the connectivity loss function proposed in this study (which measures the effect of road selection results in terms of the connectivity of the road segment) and a regularization term. In each epoch of training, the connectivity loss function ($L_{connectivity}$) is calculated as follows:

- a) Given the selected road graph (corresponding to used_G in the code) and the road selection beachmark (corresponding to G in the code)
- b) For each node s (road) in used_G, if there are roads connecting to both ends of road s, $I(s)=0$ (corresponding to code lines 117-118), if there is only one end of road s connected to it (hanging road), then $I(s)=1$; if there is no road connecting to both ends of road s (isolated road), then $I(s)=2$.
- c) After evaluating all nodes in used_G, construct the connectivity loss of the road selection result (corresponding to code line 129).
- d) In essence, the cross entropy loss function determines whether the road itself is selected and multiplies it with the true road label to calculate the entropy, while the connectivity loss function in this study determines whether the first-order neighbors of the selected roads are chosen, and calculates the connectivity loss function according to (b).
- e) In the specific implementation process, to enable the connectivity loss function to participate in back propagation, the value of node s (0, 1, 2) calculated according to step (b) needs to be converted into a differentiable continuous function, which is similar to the cross entropy loss function of binary classification (the classification result is 0 or 1). The Sigmoid function is often used to map the output of the model to the range of (0, 1). Therefore, in this study, we use Sigmoid function to map the output of the model to the range of (0, 1) according to the state (0 or 1) of the adjacent nodes (first-order neighbors) of road s. And then the connectivity scores of all selected road nodes are superimposed

as the connectivity loss term of the entire graph.

(7) Topological refinement of road selection results

Even with the above method, some roads in the road selection results may still be isolated or suspended. Therefore, this study proposes to further optimize the topology of the road selection results based on the construction principle of stroke, that is, directly delete short, isolated roads and very short, suspended roads, fill the gaps between adjacent roads, etc. This procedure can greatly ensure the connectivity of the road network selection results, and we have provided an automated processing method. It can also be processed manually in ArcMap or QGIS following this refinement rule.

In summary, the above two strategies can well ensure the topological structure and overall connectivity of the road selection results while considering the road selection characteristics. Additionally, the topological features of roads used in existing studies, such as centrality, betweenness centrality and adjacency centrality, also consider the influence of the first-order neighboring road nodes on the road nodes to a certain extent.

Note: The connectivity loss of the roads in this study cannot be used alone, as it only calculates the connectivity of the selected roads. To ensure the selected roads have good connectivity, it needs to be used in combination with other loss functions that measure the accuracy of road selection (such as the binary cross entropy loss function used in this study).

Written by Ju Peng Email: daisy_pj@csu.edu.cn

论文实现细节说明文档

由于数据隐私和国家政策等问题，论文中使用的多比例尺路网数据和出租车轨迹数据等无法共享。为了便于论文复现和对论文中方法的使用，现将论文的实现细节进行介绍。

1. 中文版

- (1) 研究区域 1: 北京五环内（研究区域 2 武汉数据的处理方式相同，选取时直接将北京训练好的模型应用于该区域）
- (2) 实验数据：三种比例尺的路网数据。初始比例尺 1:10,000，目标比例尺 1 1:5,000，目标比例尺 2: 1:20,000。初始比例尺数据来源于 OSM，目标比例尺数据来源于天地图（可以根据在线地图对初始 OSM 数据进行处理得到）
- (3) 数据预处理：
 - a) 单线道路生成（道路中心线提取）：由于下载的 OSM 数据为道路级别的，包含了不同方向、主干道和辅路等道路信息，而现有的道路选取主要是基于单线道路，因此首先将 OSM 复杂多线道路处理成单线道路。
 - b) 道路拓扑检查与构建：将生成的道路进行拓扑检查与构建，删除孤立道路、处理道路伪节点等。
 - c) 目标比例尺数据处理：可从天地图获取对应目标比例尺的数据（如果有），然后将目标比例尺与原始比例尺数据进行对齐，确保目标比例尺数据存在于原始比例尺数据中。如果无法获取到天地图的数据，可以参考在线天地图对应比例尺下的路网，手动处理获取目标比例尺（比较费时费力）。另外，如果拥有其他研究区域的多比例尺路网数据，也可以使用。
 - d) 对目标比例尺数据 1 进行拓扑检查与重建：主要是为了实现从目标比例尺 1 到目标比例尺 2 的道路选取。
- (4) 道路对偶图构建：
 - a) 道路对偶图是指将用图的方式组织道路数据，其中道路段表达为图上的节点，道路之间的连接关系表达为图上的边。
 - b) 为了构建道路对偶图，首先需要获取道路段及其道路连接节点，并对其进行唯一 ID 编号，然后根据提供的代码进行构建，可以得到后缀名为 `.gpickle` 的图文件数据。
 - c) 具体来说，将道路抽象表达为对偶图 $G=(V,E)$ ，其中 V （图节点）表示路段， E （图边）表示路段之间的连接关系。在本研究中，使用 `torch_geometric` 提供的 `Data` 模块对该数据进行组织表达，其中 `x` 表示道路段及其特征，`edge_index` 存储了路段之间的连接关系。
- (5) 路网特征计算
 - a) 道路长度：根据道路的 `geometry` 计算。
 - b) 道路等级：OSM 自带，但是需要将 OSM 道路分类等级转化为对应的数字等级。对于缺失道路等级信息的道路，根据周围道路、道路名称或者在线地图等对补全道路等级信息。
 - c) 道路密度：根据论文中参考的道路密度计算论文进行计算。也可以使用其他的道路密度计算方法。
 - d) 道路拓扑信息：根据构建的道路对偶图，采用 python 的 `networkx` 进行包计算。

- i. 中心度
 - ii. 中介中心性
 - iii. 邻近中心性:
- e) 区域吸引力: 提取道路一定缓冲区内的 POI, 根据文中提供的公式计算每条道路的区域吸引力值。
- f) 出行路径选择概率 (需要有至少 1 周的数据): 将车辆轨迹数据匹配到道路上, 然后统计道路段上的交通流量, 根据文中提供的公式计算每条道路的选择概率。

(6) 模型构建与训练:

本研究以图卷积神经网络为路网选取的模型, 采用交叉熵损失函数 (从精度方面度量道路选取结果的效果) 以及本研究提出的连通性损失函数 (从路段的连通性方面度量道路选取结果的效果) + 正则化项。在该过程中的每一次 epoch 中, 连通性损失函数 ($L_{connectivity}$) 的计算方式为:

- a) 给定选取的道路图 (对应代码中的 `used_G`) 和道路选取 `beachmark` (对应代码中的 `G`)
- b) 对于 `used_G` 中的每一个节点 s (道路), 如果道路 s 的两端均存在道路与之连接, $I(s)=0$ (对应代码 lines 117-118), 如果仅存在道路 s 的一端有道路与之连接 (悬挂道路), 则 $I(s)=1$; 如果道路 s 的两端不存在任何道路与之连接 (孤立道路), 则 $I(s)=2$ 。
- c) 判断完所有的 `used_G` 中的节点后, 构造道路选取结果的连通性损失 (对应代码 line 129)。
- d) 实质上, 该过程与计算道路选取结果的交叉熵损失函数相比: 交叉熵损失函数判断道路本身是否被选取, 并与道路真实标签相乘计算熵; 本研究中的连通性损失函数针对选取的道路结果, 判断道路的一阶邻居是否被选取, 并根据 (b) 计算连通性损失函数。
- e) 在具体实现过程中, 为了让连通性损失函数可以参与反向传播, 需要将根据 (b) 步骤计算得到的节点 s 的值 (0, 1, 2) 转化为一个可微的连续函数。这与二分类的交叉熵损失函数有异曲同工之处 (分类结果为 0 或 1), 我们通常使用 Sigmoid 函数将模型的输出映射到 (0, 1) 范围内。在本研究中, 可以根据道路 s 的相邻节点 (一阶邻居) 的状态 (0 或 1) 使用 Sigmoid 函数将模型的输出映射到 (0, 1) 范围内, 然后将所有选取的道路节点的连通分数叠加, 作为整个图的连通性损失项。

(7) 道路选取结果拓扑精细化处理

即便通过这样的方式, 道路选取结果的若干道路仍然可能存在孤立或者悬挂的情况, 因此, 本研究进一步基于 `stroke` 的构建原理提出对道路选取结果进行进一步的拓扑优化, 即直接删除短小的孤立道路和悬挂道路, 对相邻道路的 `gaps` 进行填充等, 该过程可以极大保证道路选取结果的连通性。该过程已经提供了自动化的处理方式。也可以在 `arcmap` 或者 `qgis` 中根据此规则手动处理。

综上, 通过上述两种方式可以在顾及道路选取特征的情况下, 很好的保证道路选取结果的拓扑结构和整体连通性。此外, 现有研究中所使用的道路选取特征中的拓扑特征, 如中心度、中介中心性和邻接中心性等一定程度上也顾及了道路的一阶邻近道路节点对道路节点的影响。

***需要注意的是, 本研究中的道路的连通性损失不可以单独使用, 因为其仅仅是对选取的道路的连通性的计算, 是为了保证选取道路拥有较好的局部连通性, 需要结合其他度量道路选取精度的损失函

数使用（如本研究使用的二分类交叉熵损失函数）。

作者：彭举

邮箱：daisy_pj@csu.edu.cn

