A Macroscopic Comparison Method for Road Network Datasets Using Geographical Traffic Distribution

Hengyi Zhong, Toru Seo
Department of Civil and Environmental Engineering, Institute of Science Tokyo

In transportation network analysis, various types of road network data (e.g., OpenStreetMap (OSM), simplified networks, authoritative data) can be used even when focusing on the same region. Since different road network datasets can make different performance in analyses, it is necessary to compare them and make appropriate selections in a qualitative manner. However, many of the existing methods for comparing road network datasets are limited to specific topological evaluations and do not consider transportation. One challenge for such comparison is that it is hard to match links and nodes among datasets, which makes difficulties on comparing traffic status locally and globally in some road networks. This study proposes a method for quantitative comparison of different road network datasets with explicit consideration for traffic flows on them. The method first conducts a static traffic assignment with hypothetical demand for each dataset, and then compare the results using Wasserstein distance on two dimensional plane. Case study on OSM and its simplifications suggests the potential use of the proposed method in evaluating and selecting road network datasets.

Keywords: transportation network analysis, Wasserstein distance, road network simplification

1. Introduction

In transportation network analysis, road network data plays an important role as the foundation where all travelers move and traffic flows. Road network data is a type of geographical information system (GIS) data which organizes real-world roads into a network structure with their connections and shapes. Nowadays, an increased accessibility to various types and formats of road network data, such as OpenStreetMap (OSM) and Transportation Networks (Transportation Networks for Research Core Team, accessed 2025), has dramatically improved the efficiency of research on road networks, including transportation network analysis.

The comparison of different road network datasets is essential for understanding the quality of data and evaluating their accuracy and usefulness. In transportation analysis, features for comparison can be varied. For example, simplified road networks may not accurate, but tends to computationally efficient and are expected to keep the original traffic status.

Previous studies mainly focused on the characteristics of network, such as positional accuracy, completeness and topological correctness (Boeing 2024, Hakley 2010, Hashemi 2017). However, these methods only focus on a specific feature of network and may not be appropriate to directly apply to transportation network analysis, in which various local and global features are essential. For the above-mentioned example, existing methods could only evaluate the accuracy but not traffic and computational issues, thus the usefulness of data could not be evaluated properly.

In this paper, we developed a comprehensive method of comparing road network datasets that can capture the effects of local and global features for transportation analysis. Traffic distribution on a road network is considered as a geographical distribution weighted by traffic parameters. Then differences between two road networks is measured by a traffic-weighted distance between two corresponding distributions. Also, a possible application on evaluating different road network data in network simplification by proposed method was discussed.

2. Methodology

To evaluate road network data on a macroscopic scale, we consider the entire network represented by the data as one single distribution. In this study, we represent road network data as geographic distribution with traffic. Traffic-weighted distance between two road network datasets is defined as the Wasserstein distance between two corresponding distributions.

2.1 Traffic status

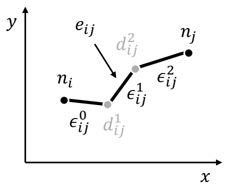
To obtain possible traffic status on road network data, static traffic assignments with user equilibrium were utilized for compared networks respectively. Necessary traffic parameters (e.g. capacity and max speed) and OD demands can be hypothetically assumed from road network data.

Traffic parameters are set by the road type for each link of the network. Links of higher-class road type such as trunk and motorway are set with large capacity and high max speed, and vice versa.

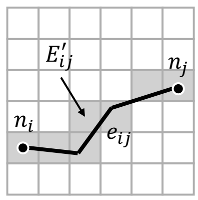
OD demands can be calculated from trip generation and attraction volumes by the gravity model. Divide the area of road network data into zones with appropriate sizes and set centroid nodes for each zone. Travel cost is the shortest travel time between two centroids. To simulate a high number of external trips by passing expressways, set high trip generation and attraction volumes for zones with major roads connected to the outside of the area, and low volumes for the other zones.

2.2 Geographic distribution of road network data

Let road network data $\mathcal{D}$ as a form of network consists of a set of nodes $\mathcal{N} = \{n_i\}$ and a set of links $\mathcal{E} = \{e_{ij}\}$. n_i represents an intersection in road network and e_{ij} represents a road section between n_i and n_j. Dummy nodes d_{ij}^k are used to illustrate the shape of e_{ij} by dividing it into small line segments ϵ_{ij}^a, as it shown in **Figure 1**(a). Each node and dummy node have its coordinates. Each link has its traffic characteristics such as link length, capacity and max speed.



(a) continuous distribution



(b) discrete distribution

Figure 1: Distribution of link e_{ij}

$\mathcal{D}$ geographically distributes on a two-dimensional grid space with size $M \times N$, which is defined as the study area $\mathcal{R}: x \in (x_0, x_o + M), y \in (y_0, y_0 + N)$. Geographical distribution E_{ij} of link e_{ij} is as its shape, which consists of all positions (x, y) on all line segments ϵ_{ij}^a.

Assume that all traffic can only distribute on road sections. Traffic distribution with weight w_{ij} on $\mathcal{D}$ only distributes on each E_{ij}. w_{ij} are set from obtained traffic status. To do numerical calculation of Wasserstein distance, we discretize $\mathcal{R}$ into grids so that D can be written as a two-dimensional matrix. As it is shown in **Figure 1**(b), distribution E'_{ij} which reflects the shape of e_{ij} now can be described by all grids (p, q) where corresponding line segments passes through. Therefore, the discrete geographic distribution D' of road network data $\mathcal{D}$ with traffic can be written as follows.

$$D': (p, q) = w_{ij}, \forall (p, q) \in E'_{ij}, \forall e_{ij} \in \mathcal{E} \tag{1}$$

2.3 Calculation of traffic-weighted distance

Distance between two road networks is defined by Wasserstein distance, a distance function evaluates the similarity between two distributions. As it catches the shape well, it is suitable to use Wasserstein distance to measure distance between two geographical distributions (Ambrosio and Gigli 2009).

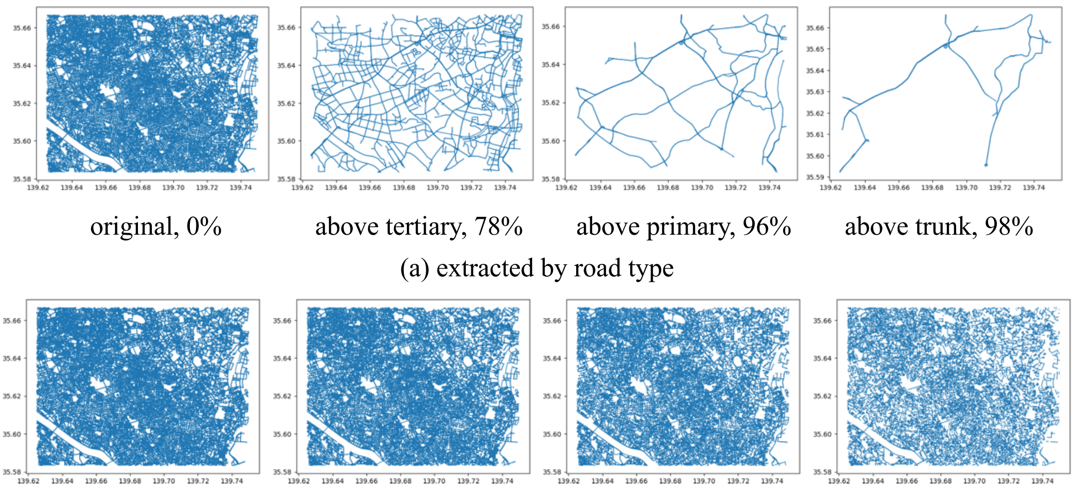
Let the discrete geographic distributions of two road network datasets $\mathcal{D}_A, \mathcal{D}_B$ as $D'_A(\boldsymbol{x}), D'_B(\boldsymbol{y})$ respectively. $\boldsymbol{x} = (p_A, q_A), \boldsymbol{y} = (p_B, q_B)$ denotes positions of grids on D'_A and D'_B respectively. Wasserstein distance $W(D'_A, D'_B)$ is defined as the distance between $\mathcal{D}_A, \mathcal{D}_B$. As geographical distributions are 2-dimentional distribution, this distance can be numerically calculated by the sliced Wasserstein distance $SW(D'_A, D'_B)$ (Bonneel et al. 2015).

3. Case Study

In the case study, we did comparisons of different simplified networks within a same study area. Appropriate network simplification methods, such as extraction, aggregation and abstraction, are widely used in transportation network analysis. In these methods, nodes and links which considered as insignificant are removed to simplify the network but keep the traffic distribution on it.

To confirm whether proposed method could appropriately evaluate the changes of traffic status on different road network data, we calculate $W(D'_0, D'_k)$ by proposed method between original $\mathcal{D}_0$ and simplified network $\mathcal{D}_k$ for each simplification step, using two different simplification methods. One is network extraction method which remains only high-class links. Groups of links were gradually removed from lower classes to the higher classes. The another is randomly removing links and nodes to create randomly edited networks with different reduction rates.

OpenStreetMap (OSM) road networks of a 10km×10km area in southwest of Tokyo are used in this study. Shapes of example simplified networks and their link reduction rates are shown in **Figure 2**.



original, 0% | above tertiary, 78% | above primary, 96% | above trunk, 98%

(a) extracted by road type

20% reduced | 40% reduced | 60% reduced | 80% reduced

(b) randomly edited

Figure 2: Shapes of different simplifications of OSM networks

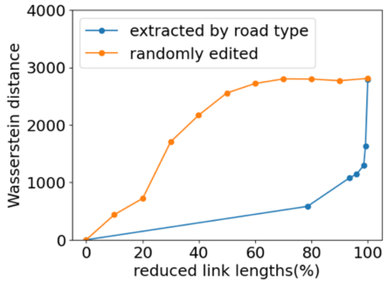


Figure 3: Relationships between link length reduction rates and distances to the original network

Figure 3 shows the relationships between link length reduction rates and distances to the original network of different edited road networks. In appropriately simplified networks, it can be confirmed that those networks have high link length reduce rate but small distance to the original network, refers to small changes on traffic distributions shown in **Figure 4**(a). On the contrast, for randomly edited networks, the distance to the base network rapidly grows since the topology get worse when a part of major roads gets randomly removed, causes large changes on traffic distributions shown in **Figure 4**(b). This result shows the proposed method could catch characteristics that traffic distribution remains stable when simplifying it by appropriate methods.

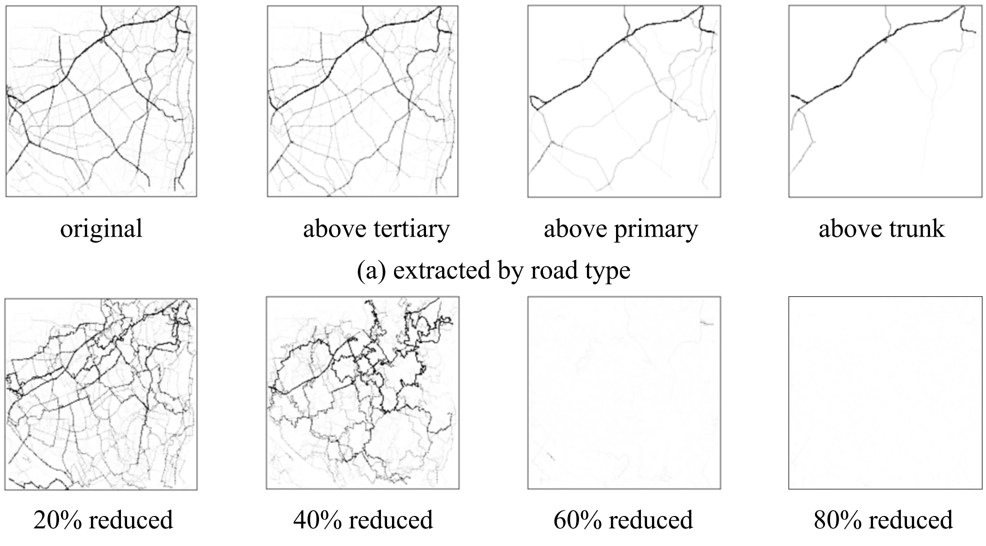


Figure 4: Traffic distributions on different simplifications of OSM networks

4. Conclusions

In this study, we developed a method for comparing road network datasets which can capture the effects of features for transportation analysis in a macroscopic scale. By focusing on the geographical features contained in the road network data and its traffic distribution, road network can be represented as geographic distribution. Based on this representation, difference between road network datasets can be measured as the Wasserstein distance between corresponding geographic distributions. Case study on OSM and its simplifications shows that proposed method could evaluate changes of possible traffic status over different networks, suggests a potential use in evaluating and selecting road network datasets.

References

Ambrosio, L. and Gigli, N.: A user's guide to optimal transport, *CIME summer school*, Italy, 2009.

Boeing, G.: Modeling and Analyzing Urban Networks and Amenities with OSMnx. Working paper. 2024.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H.: Sliced and Radon Wasserstein Barycenters of Measures, *Journal of Mathematical Imaging and Vision*, Vol.51, No.1, pp.22–45, January 2015.

Haklay, M.: How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets, *Environment and Planning B: Planning and Design*, Vol.37, No.4, pp.682–703, 2010.

Hashemi, M.: A testbed for evaluating network construction algorithms from GPS traces, *Computers, Environment and Urban Systems*, Vol.66, pp.96–109, November 2017.

Transportation Networks for Research Core Team. *Transportation Networks for Research*. https://github.com/bstabler/TransportationNetworks. Accessed in January 2025.