GLOBAL ROAD NETWORK DATASET FUSION BASED ON TRACEABILITY MECHANISM

Chenchen Wu [1], Hongwei Zhang [1], Xiao Du [1], Zhichao Li [2], Shangwei Lin [1], Jianwei Liu [1] *

[1] National Geomatics Center of China, Beijing, China- (wucc, hwzhang, duxiao, linshangwei, liujw)@ngcc.cn
[2] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University, Wuhan, China - 2019206190034@whu.edu.cn

Commission IV, ICWG IV/III

KEY WORDS: Data fusion, Road network dataset, Traceability mechanism, Confidence.

ABSTRACT:

Accurate road network information is the foundation of urban construction, traffic planning and emergency response, and also provides necessary assistance for public travel. How to provide global road network data products with complete elements, rich attributes, and high precision is a problem that must be considered in the reserve of basic geographic information resources.This paper studies the geometric information fusion of multi-source data based on OSM road network data, mainly carries out relevant research work on the identification and matching of homonymous entities, and proposes a method based on traceability mechanism to solve the problem of inconsistent geographical location in different data sets after topology preprocessing of homonymous features, so as to improve the integrity and accuracy of road network data after fusion.

This paper proposes a method based on traceability mechanism, and explores a technical processing flow of global road network data fusion. The test process greatly improves the matching effect of entities with the same name in different data sets, greatly ensures the original topology relationship, and improves the integrity and accuracy of the fused road network data. The data test result shows that this method has certain application value and can basically meet the actual needs of multi-source road network data fusion from a global perspective.

1. INTRODUCTION

As an important part of spatial data framework, road data has the characteristics of rapid development and rich attributes. Accurate road network information is the foundation of urban construction, traffic planning and emergency response, and also provides necessary assistance for public travel. How to provide global road network data products with complete elements, rich attributes, and high precision is a problem that must be considered in the reserve of basic geographic information resources. At present, the production mode of road network data mainly includes field survey based on mobile devices like GPS and feature extraction based on remote sensing images. However, these two modes are not applicable to all regions of the world, especially in production of large-scale road network data: 1) it is impossible to carry out large-scale field measurement in overseas areas; 2) Affected by various factors such as technology and equipment, the integrity of automatically extracted road network data is often not high due to the interference of buildings, overpasses, trees and other factors, especially in urban areas.

In recent years, with the rapid development of Internet technology, network information has shown explosive growth. Open source data provide rich geographic information, such as Open Street Map(OSM), Bing map and Google map, which provide vector maps containing features of boundary, transportation, waterway, buildings and POI, etc. National real estate information websites contain house POI information. Longitude and latitude information and detailed address information can be parsed from the web page source code. GeoNames integrates free data from various sources and store it into a database.

Among them, OSM was paid high attention in voluntary geographic information (VGI) dataset, due to its advantages of wide coverage, substantial content, strong current situation and complete disclosure. However, its road network data also has some disadvantages from a global perspective, such as uneven geometric accuracy, uneven data distribution and inconsistent topological relationship. If other road network dataset can be fused into OSM and verified each other, it will become an idea of rapid production for global road network data.

Therefore, this paper studies the geometric information fusion of multi-source data based on OSM road network data, mainly carries out relevant research work on the identification and matching of homonymous entities, and proposes a method based on traceability mechanism to solve the problem of inconsistent geographical location in different data sets after topology preprocessing of homonymous features, so as to improve the integrity and accuracy of road network data after fusion.

2. RELATED RESEARCH

On the matching of geometric features between road network data, relevant scholars at home and abroad have carried out relatively more research and achieved some research results. Saalfeld proposed using distance to describe the overall similarity of line elements (Saalfeld, 1999); Zhang Qiao proposed the

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B4-2022-633-2022 | © Author(s) 2022. CC BY 4.0 License.

similarity calculation method of line entity based on the middle area method and the similarity calculation method based on the azimuth of broken line segment; Sui Haigang discussed the matching problem of homonymous line entities in the research of feature-based automatic change detection method of road network, and mainly adopted the matching technology of homonymous entities based on buffer growth search; Xiong proposed a semi-automatic data matching method for network data integration; According to the matching characteristics of broken lines in road network, Chen Yumin and others proposed a broken line node distance matching algorithm based on grid index, which transformed the calculation of geometric similarity between complex broken lines into a matching method to calculate the distance from node to broken line, which reduced the computational complexity and improved the computational efficiency (Chen et al., 2007); Tang Luliang uses European distance to measure the similarity of two line elements; Based on the probability based matching algorithm, Tong Xiaohua and others proposed a multi index weighted fusion based on probability theory matching algorithm; Deng min extended the Hausdorff distance and proposed a spatial object matching method based on the extended Hausdorff distance; Li Qingquan proposed a hierarchical path planning algorithm suitable for lbs; Fu Zhongliang proposed a spatial matching method combining coarse matching and fine matching in the study of multi-scale spatial database updating (Fu et al., 2007); Wu Jianhua proposed a weight based similarity calculation model of spatial elements, and matched the elements under complex spatial relationships based on this model; By introducing the environmental similarity of point and line entities, Wu Jianhua proposed a multi feature combination matching method considering the environmental similarity, which effectively improved the accuracy of entity matching; Ying Shen proposed a top-down matching method; Hu Yungang proposed a matching method based on analytic hierarchy process, which expresses the matching relationship between road targets with different scales and tenses through three matching levels: decomposition, basic and abstract of spatial entities; Safra Proposed a method to find homonymous points from two data sets based on location algorithm; Zhao Binbin uses the method of intersection of minimum circumscribed rectangle and target elements to find candidate matching sets; Guo Li uses the multi-step buffer method to match the entities with the same name; Huang Wei and others proposed three matching modes for multi-scale vector simple geometric entity data to effectively realize cross-scale vector spatial data matching method; And building materials put forward an adjustment and merging algorithm based on topological relationship to obtain the optimal spatial location of unmatched entities; An Xiaoya used a shape multilevel chord length function to describe the shape of line elements and surface elements respectively. By combining with the center distance function, they constructed the geometric similarity model of vector elements, and realized the shape matching of line elements and surface elements respectively; An Xiaoya proposed a mesh feature matching algorithm for map data with different scales based on similarity measurement. In the study of similarity evaluation of linear elements, Liu Pengcheng derived the similarity evaluation model of linear elements from the shape description model of planar elements; Ma Huang group put forward a hierarchical division method of road network based on hierarchical random graph; Zhang Yunfei, Yang song and Zhao Dongbao used the probability relaxation method to match linear elements on the basis of determining the candidate matching set; Luan Xuechen proposed a road network node matching method based on structural pattern; Gong Xianyong and others used the colony advantage of ant colony algorithm to find the globally optimal matching scheme of homonymous entities in road network. Tong extended on the basis of Li proposed an improved road entity matching method of optimization and iterative logistic regression model matching algorithm (Tong et al., 2007); Safra proposed a line matching method to match the endpoint rather than the whole matching line, so as to improve the matching efficiency; Luo Guowei and others proposed an optimal combination matching method based on grid division, which selects the best matching object from the candidate elements by comprehensively comparing the spatial and semantic features of the combination object. Fu Zhongliang and others put forward the algorithm of multiple logistic regression to realize road network matching on the basis of Tong and others. Guo Qingsheng proposed a stroke partial matching algorithm which can change and update road data at different scales (Guo et al., 2017). Liu Chuang proposed a road network linkage matching method considering the similarity of superior and subordinate spatial relations. Zhang Jianchen and others proposed a road network matching algorithm combining global and local optimization to improve the M: n matching pattern from a global point of view. Zhu Di and others focused on the low-frequency floating vehicle trajectory data and abstracted the map matching problem. Guo Ningning and others selected the similarity of the four spatial features of the distance, direction, shape and length of the road section as the index to measure whether the road section is matched, and selected the radial basis function neural network to realize the automatic matching of the road network. Wang Zhiguo and others combined the matching and fusion theory of vector elements to design a matching and fusion algorithm suitable for VGI data of road network.

From the above analysis, the data fusion between OSM road network and other road networks has achieved some research results at home and abroad. But there are still deficiencies: most of the data used in the existing fusion methods are professional spatial data, which has unified production standards, high consistency of topological relations. However, OSM data mainly marked by volunteers, it has the characteristics of inconsistent topological relationship, uneven deformation and density distribution. There are some difficulties in the matching of homonymous entities.

3. METHODOLOGY

3.1 Problems to be solved

The global road network data fusion method is mainly based on the road network matching algorithm. It judge whether a line elements from different data sources logically correspond to the same physical road. Then non-homonymous entities can selected as supplement to the original road network data source, homonymous entities can be selected to verified original road network data source. In this way, the richness and availability of road network data can be improved.

Due to different data collection means, business orientation and emphasis on content in production of road network, it is necessary to preprocess the original road data to reduce the impact of quality on data fusion as much as possible.

In the global road network data fusion process, the preprocessing method of breaking at major turning points, relevant intersections, road intersections, river boundaries and administration boundaries is used. A complete road in the original road network is interrupted into different segments, which improves the accuracy of road network fusion, but also introduces the following problems:

(1) What should be added is not added: Only a part of the whole road is supplemented, and the whole road is not completely supplemented. As a result, there are intermittent discontinuous line segments in the map, as shown in the fig.1. In figure.1-a, the black line segment represents the original data, the green line segment represents the data to be fused, and the fusion result is shown in figure.1-b. The red part is incremental data, and some part of the whole original data is missed.

(2) What should not be added is added instead: According to the theoretical method, in the case of height matching with the original road network, some new road data are still selected as fusion elements. Figure.1-c shows the black original data and the green data to be fused. As shown in figure.1-d, after fusion processing, the red part is incremental data, and the highly overlapping data is wrongly supplemented.

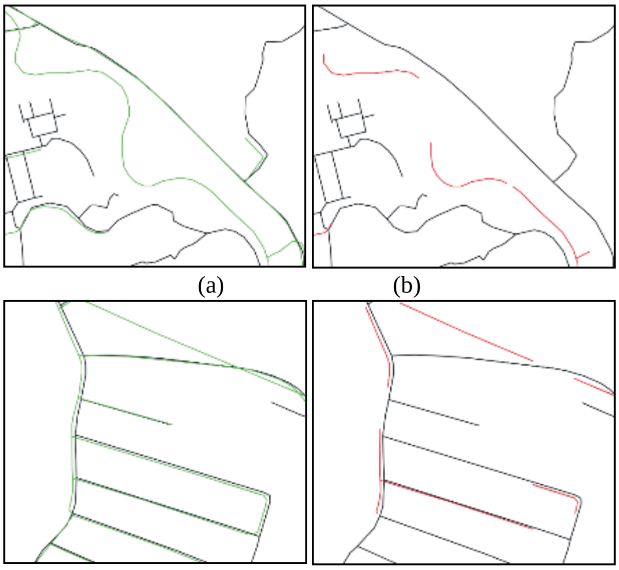

(a) (b)
(c) (d)

Figure 1. Errors in fusion process

This paper proposed a global road network data fusion method based on traceability mechanism, effectively solves the above two problems, and achieves good results.

3.2 Main principle

Assuming the road network dataset S as original data, road network dataset T as the data to be fused. Road network dataset A is obtained as incremental data after fusion processing. The specific traceability process is as follows:

3.2.1 In pre-processing before fusion, the line feature after the interruption saves the original line feature ID. In road network dataset T, each geographic line feature is interrupted. The corresponding relationship between the data CID introduced after the interrupt and original line feature ID is retained. As shown in fig.3, the original road feature which ID is 1 is divided into four sub roads, corresponding relationship indicates as FID=1,CID=0 、 FID=1,CID=1 、 FID=1,CID=2 、 FID=1,CID=3. FID is the original road ID before processing, the network data after interruption is expressed by the formula1, formula2.

$$\boldsymbol{S} = [s_0 \quad \cdots \quad s_i \quad \cdots \quad s_n]^T, n = LEN(\boldsymbol{S}) \qquad (1)$$

$$\boldsymbol{T} = [t_0 \quad \cdots \quad t_i \quad \cdots \quad t_m]^T, m = LEN(\boldsymbol{T}),$$
$$t_i = [b_0 \quad \cdots \quad b_j \quad \cdots \quad b_{BLEN(t_i)}] \qquad (2)$$

where s_i = line feature i in road network dataset S
LEN(S) = number of elements in road dataset S
t_i = line feature i in road network dataset T
b_j = feature j after t_i is interrupted
LEN(T) = number of elements in dataset T to be fused
$BLEN(t_i)$ = number of elements of line *i* in dataset T after line feature interruption processing

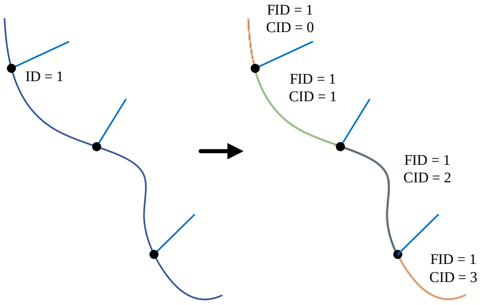


Figure 2. Road feature divided into four sub roads

3.2.2 In the post-processing after fusion, source line is traced according to the original ID carried by the incremental data. The details are as follows:

(1) Traverse the incremental data and save the corresponding relationship between the original ID and the supplementary data set.

$$\boldsymbol{A} = \{a_0, a_1, \cdots, a_{n-1}, a_n\}, n = LEN(\boldsymbol{A}) \qquad (3)$$

$$a_i = \|id_i \leftrightarrow \{b_0, b_1, \cdots, b_{m-1}, b_m\}\|, m = MLEN(t_i) \qquad (4)$$

where a_i = the set of line features which take $\mathrm{id_i}$ as original ID in incremental road network dataset A
$\{b_0, b_1, \cdots, b_{m-1}, b_m\}$ = data be fused into the original dataset.
$MLEN(t_i)$ = number of elements of sub road lines be fussed and which takes $\mathrm{id_i}$ in dataset T

(2) Calculate ratio of the length of lines take id_i as original ID in incremental dataset A, to the length of lines take id_i as original ID in dataset T. Then, length ratio set W is obtained.

$$\boldsymbol{W} = [w_0 \quad \cdots \quad w_i \quad \cdots \quad w_n]^T, n = LEN(\boldsymbol{W}),$$
$$w_i = \left(\sum_{k=0}^{m=MLEN(t_i)} DIS(b_k)\right) / DIS(t_i) \qquad (5)$$

DIS is a function that obtains the geographic length of line features according to the projected coordinate system. When w_i greater than the pre-set threshold *offset* (10m is used in this paper), it indicates that the subline feature set can be supplemented as a whole line feature traceable to the original.

(3) Set the confidence of the entire incremental line feature. The line feature t with successful traceability will establish confidence according to the set of supplementary sub feature set $\{b_0, b_1, \cdots, b_{m-1}, b_m\}$. The ratio of the length of each broken line feature b_i to that of the original line feature t_i, is regarded as the weight value. The confidence of the whole original road is set according to the formula as follows.

$$\boldsymbol{C} = [c_0 \quad \cdots \quad c_i \quad \cdots \quad c_n]^T, n = LEN(\boldsymbol{C}),$$
$$c_i = \sum_{k=0}^{m=MLEN(s_i)} (DIS(b_k)/DIS(s_i) \times COIN(b_k)) \tag{6}$$

where c_i = confidence of the whole integrated line i
$COIN$ = function to obtain the confidence of an incremental line feature

3.3 Workflow

The backtracking process flow chart is shown in Figure.3.

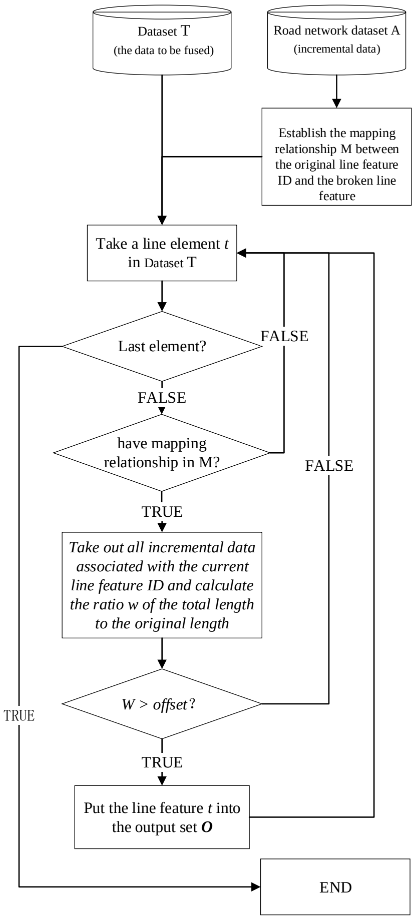
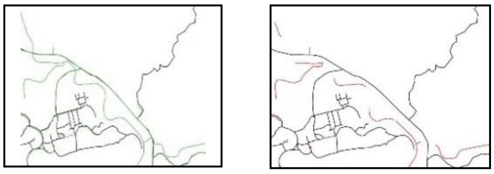

Dataset T
(the data to be fused)
Road network dataset A
(incremental data)
Establish the mapping relationship M between the original line feature ID and the broken line feature
Take a line element *t* in Dataset T
Last element?
FALSE
FALSE
have mapping relationship in M?
FALSE
TRUE
Take out all incremental data associated with the current line feature ID and calculate the ratio w of the total length to the original length
W > offset?
TRUE
TRUE
Put the line feature *t* into the output set ***O***
END

Figure 3. Work flow of the dataset fusion

3.4 Test results

As shown in the Figure 4, fusion result with traceability mechanism shows that the connectivity and rationality of incremental data is improved, and the data errors caused by fusion are reduced.



Original data

Fusion without traceability mechanism

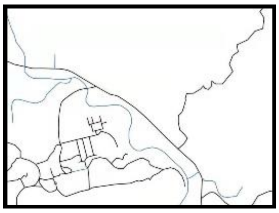


Figure 4. Comparison diagram of test results

In order to verify the efficiency of the methodologies proposed in this paper, South America was taken as the test area. It took one week to complete the integration of the whole region. As shown in Figure 5, the black line segment is the original road network data, and the red line segment is the incremental road segment supplemented. After fusion, the road data density is increased by about 8%.

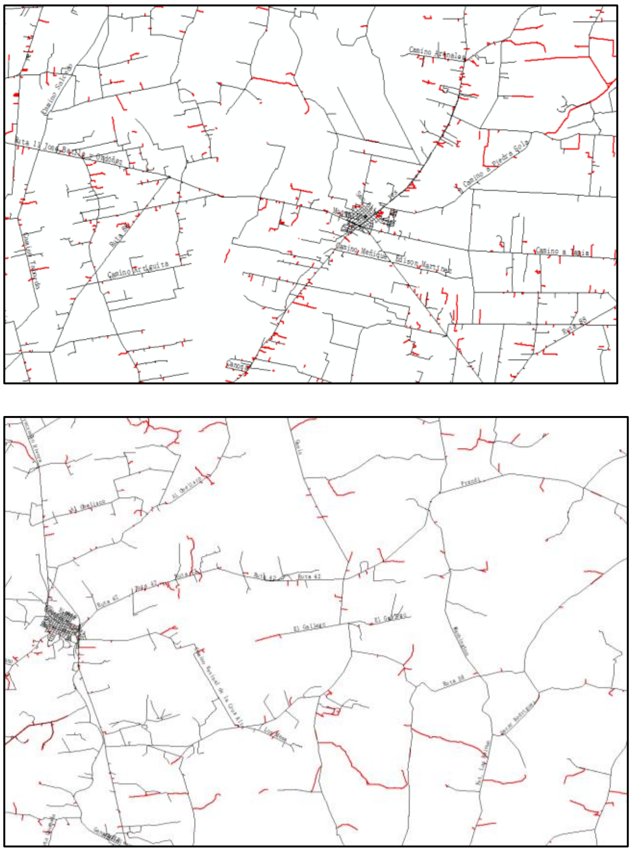


Figure 5. Fusion test results in South American

In addition, the matching relationship generated in the process of fusion can also be used to supplement attributes. For the same road entity after matching, the attribute of the road with name can be assigned to the original road without name. As shown in Figure 6, the green part in the figure is the road section with new attributes according to the matching relationship.

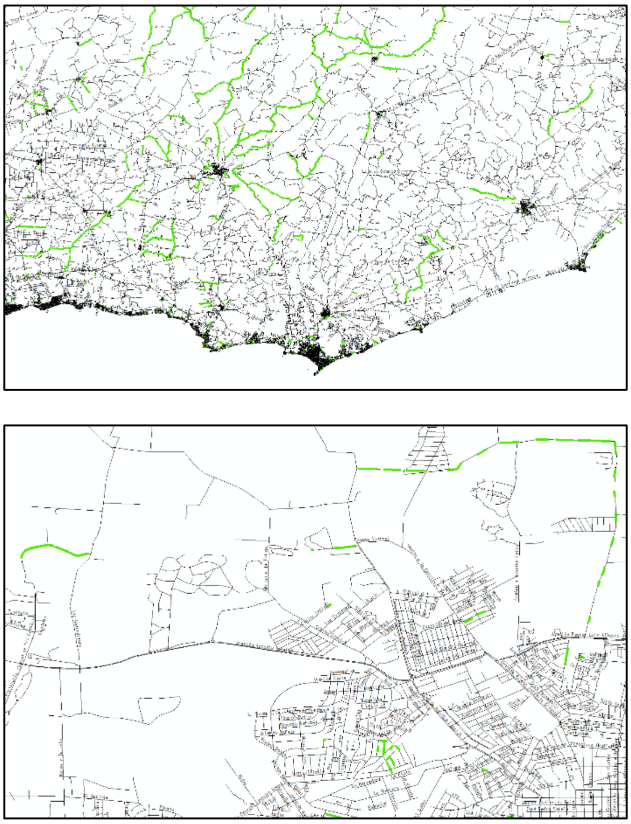


Figure 6. Attribute supplementary test results

The research in this paper has been used in the actual large-scale road network data production, and the produced data results are provided to users through tile map service as shown in Figure 7.

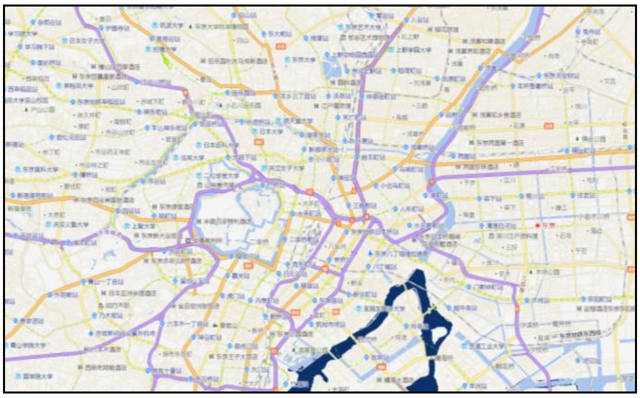


Figure 7. Map service based on fused road network data

4. CONCLUSION

This paper proposes a method based on traceability mechanism, and explores a technical processing flow of global road network data fusion. The test process greatly improves the matching effect of entities with the same name in different data sets, greatly ensures the original topology relationship, and improves the integrity and accuracy of the fused road network data. The data test result shows that this method has certain application value and can basically meet the actual needs of multi-source road network data fusion from a global perspective.

ACKNOWLEDGEMENTS

We would like to give our acknowledgements to National Geomatics Center for the strong support in this study. This support was valuable.
The authors would like to thank the editors and anonymous reviewers for their valuable comments also, who helped improve this paper.

REFERENCES

Chen, Y. M., Gong, J. Y., Shi, W. Z., 2007. A distance-based matching algorithm for multi-scale road networks. *Acta Geodaetica et Cartographica Sinica*, 36(1), 84-90.

Caset, F., Blainey, S., Derudder, B., Boussauw, K., Witlox, F., 2020. Integrating node-place and trip end models to explore drivers of rail ridership in Flanders, Belgium. *Journal of Transport Geography*, 87, 102796.

Ebisch, K., 2002. A correction to the Douglas-Peucker line generalization algorithm. *Computers & Geosciences*, 28(8), 995-997.

FU, Z., & WU, J., 2007. Update technologies for multi-scale spatial database. *Geomatics and Information Science of Wuhan University*, 32(12), 1115-1118.

Guo, Q. S., Xie, Y. W., Liu, J. P., Wang, L., Zhou, L., 2017. Algorithms for road networks matching considering scale variation and data update. *Acta Geodaetica et Cartographica Sinica*, 46(3), 381.

Jin, F., Wang, C., Li, X., 2008. Discrimination method and its application analysis of regional transport superiority. *Acta Geographica Sinica*, 63(8), 787-798.

Juhász, L., Hochmair, H. H., 2017. How do volunteer mappers use crowdsourced Mapillary street level images to enrich OpenStreetMap. *Proceedings of the 20th AGILE Conference on Geo-Information Science, Wageningen, The Netherlands*, 18-21.

Li, C., Sun, A., Datta, A., 2012. Twevent: segment-based event detection from tweets. *Proceedings of the 21st ACM international conference on Information and knowledge management*, 155-164.

Groenendijk, L., Rezaei, J., Correia, G., 2018. Incorporating the travellers' experience value in assessing the quality of transit nodes: A Rotterdam case study. *Case Studies on Transport Policy*, 6(4), 564-576.

Luo, L., Liu, B., & Liu, X., 2017. Data Quality Assessment and Application Analysis for OpenStreetMap Road Network. J*iangxi Sci,* 1.

McMaster, R. B., 1989. The integration of simplification and smoothing algorithms in line generalization. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 26(1), 101-121.

Neis, P., Zielstra, D., Zipf, A., 2011. The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4(1), 1-21.

Saalfeld, A., 1999. Topologically consistent line simplification with the Douglas-Peucker algorithm. *Cartography and Geographic Information Science*, 26(1), 7-18.

Samsonov, T. E., Yakimova, O. P., 2017. Shape-adaptive geometric simplification of heterogeneous line datasets. *International Journal of Geographical Information Science*, 31(8), 1485-1520.

Shi, W., Cheung, C., 2006. Performance evaluation of line simplification algorithms for vector generalization. *The Cartographic Journal*, 43(1), 27-44.

Tong, X. H., Deng, S. S., Shi, W. Z., 2007. A probabilistic theory-based matching method. *Acta Geodaetica et Cartographica Sinica*, 36(2), 210-217.

Zhang, Y., Marshall, S., Manley, E., 2019. Network criticality and the node-place-design model: Classifying metro station areas in Greater London. *Journal of transport geography*, 79, 102485.