

AntMapper-based workflow for the crowdsourced Mapillary data preprocessing

Meihui Wang*, James Haworth[†]

Department of Civil, Environmental Geomatic Engineering, University College London

January 17, 2022

Summary

Mapillary imagery is a novel crowdsourced data offering street-level imagery and GPS trajectory data simultaneously. Image inconsistency and measurement error of GPS trajectory data are two main issues obstacle Mapillary data applications in urban studies. In this paper, proposed workflow of crowdsourced Mapillary data cleaned random images and matched the GPS trajectory to correct road segment in Inner London by using AntMapper algorithm. This preprocessing work ensures the data quality and consistency of the crowdsourced Mapillary dataset for future spatio-temporal urban analytics in various scenarios.

KEYWORDS: Crowdsourced Data, GPS Trajectory, Map Matching, Mapillary

1. Introduction

Mapillary, one of the worldwide crowdsourced street-level imagery platforms, does not only provide street-level images depicting built environment elements, but also capture the GPS positions of people's movement simultaneously (Ding et al., 2021). The high spatial coverage and fine temporal granularity made Mapillary data becoming a critical data source for addressing challenging and dynamic research issues in urban studies.

However, few studies have utilized the Mapillary dataset in urban applications to date. Different from Google Street View imagery, Mapillary data is contributed by worldwide individuals. They are playing both roles as data user and producers (Mahabir et al., 2020). Dirty and noise data acquired by inexperienced users in the crowdsourced dataset is common and heterogeneity among images is inevitable (Biljecki and Ito, 2021; Ridzuan and Zainon, 2019). Moreover, since the data was generated by GPS-enabled mobile devices, the measurement error cannot be ignored due to the urban canyon effect and other signal interferences (Gong et al., 2018).

Map matching is the process to match a sequence of raw GPS trajectory data onto the correct road network on a digital map (Lou et al., 2009; Gong et al., 2018). It is a fundamental step for GPS trajectory data preprocessing. Existing methods were separated into two major groups, the local method and the global method. Local methods normally consider the position of each single GPS point at a time. They have faster-running speed, but these algorithms overlooked the spatial and temporal relationships between two points in the same trajectory. Therefore, the performance may be insufficient for road networks with high

* meihui.wang.20@ucl.ac.uk

[†] j.haworth@ucl.ac.uk

complexity and the intersections with complicated connectivity. The global approaches process the entire sequence data from a global perspective based on the distance to the road segment and similarity towards the road network. The global methods are less sensitive to the sampling rate and can get better results in complex scenarios (Gong et al., 2018; Peng et al., 2019). However, it is time-consuming for a massive amount of high-sampling frequency Mapillary trajectory data.

To overcome above challenges, we developed a scientific workflow including data preparation work and an Ant Mapper map matching algorithm. The map matching method considers both local geometric and topological attributes and global similarity between entire trajectory and road network. In this paper, we use Inner London road dataset to showcase the proposed workflow successfully and effectively cleaned the crowdsourced Mapillary data and matched the sequences to the corresponding road segments. This preprocessing work improve the data consistency and ensure the data quality of crowdsourced Mapillary dataset for the potential urban scenarios.

2. Data preparation

2.1 Road network

The road network is acquired from Ordnance Survey, including 100,892 road segments in Inner London. The attribute table was revised by considering the vertical connectivity. Consequently, a road network MultiDiGraph was constructed based on following attribute table (see Table 1).

Table 1 Attributes of Road network

Name	Datatype	Description
edge_id	String	The unique id of the road segment
from	String	Corrected start node id considering the vertical connection Format: start node + start grade
to	String	Corrected end node id considering the vertical connection Format: end node + end grade
direction	Integer	0: single direction 1: both directions
bearing	Double	The bearing value of road segment
length	Double	The length of road segment (unit: Meter)
bbox	Geographic coordinates	The minimum bounding box of the road segment
geometry	Geographic coordinates	The line string of each road segment

2.2 Mapillary GPS trajectory dataset

Mapillary trajectory data was collected from September, 2009 to September, 2021, containing 3,182 sequences with 400,355 points in Inner London. Data was collected through the mini bounding box to in JSON or GeoJSON format by using the JSON API. The attributes of each Mapillary GPS point are shown in Table 2.

Table 2 Attributes of Mapillary GPS trajectory data

Name	Datatype	Description
caputre_at	Integer	The timestamp of the image collection (unit: millisecond)

image_id	String	The unique id of the image
sequence_id	String	The unique id of each sequence
is_pano	Boolean	Whether the image is panoramic or not
organization_id (optional)	String	The unique id of data producers' organization
geometry	Geographic coordinates	The position of each image when recording

To clean up the raw GPS trajectory data, three rules were set up to clean the dirty and noise data: (1) images number less than ten; (2) average time interval large than 2 min; (3) average distance large than 300 or equal to zero meters.

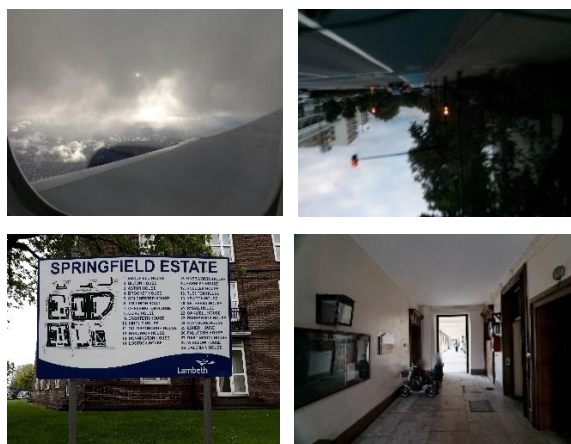


Figure1 Examples of cleaned Mapillary images

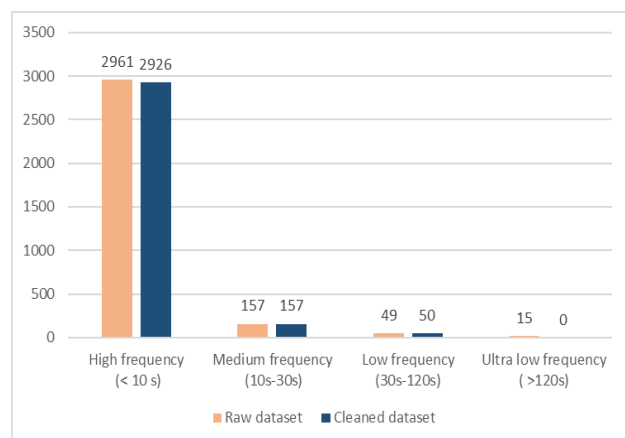


Figure 2 Number of images based on different sampling rate in raw dataset and cleaned dataset

After preprocess, 49 sequences were removed from raw dataset. 398, 884 points in 3,133 sequences are left and the average number of images in each sequence is 125.8. Figure 2 shows the sequence with different sampling rate in raw data and cleaned data. Among cleaned data, 93.39% of sequences are collected at high frequency less than 10 seconds.

3. Map matching

3.1 Problem definition

Each GPS trajectory T is a sequence of spatial points $p_1 \rightarrow p_2 \rightarrow \dots \rightarrow p_n$ in time order. The objective of map matching is to determine the best path $P = e_1 \rightarrow e_2 \rightarrow \dots \rightarrow e_n$ on the road network from the candidates set C_i .

3.2 Ant Mapper map matching method/Model

Ant Mapper is the state-of-the-art map matching algorithm originally proposed by Gong et al., (2018). It introduces a three-step workflow including the candidate calculation, local attributes and global similarity evaluation and path optimization (See Figure 3).

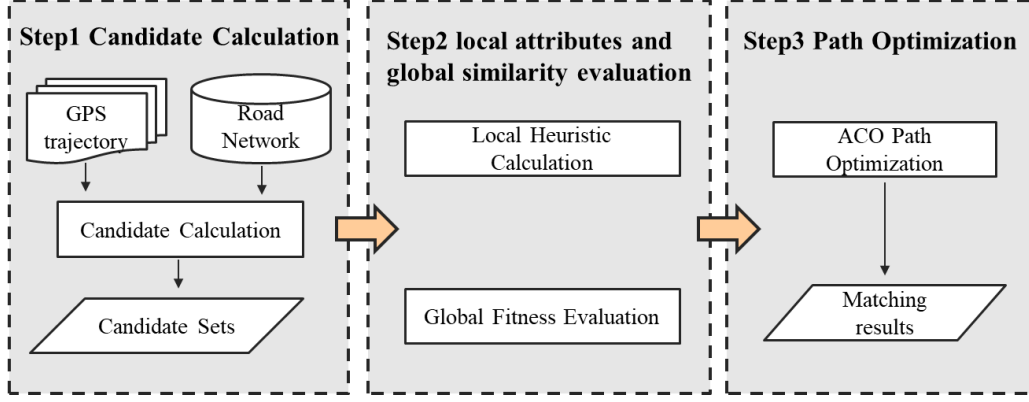


Figure 3 Workflow of Map matching

3.2.1 Candidate calculation

Candidate calculation is to calculate the projected point on the potential road segment within certain distance for the GPS point. As shown in Figure 4, within the searching distance, three candidate points of p_i are c_i^1 , c_i^2 , c_i^3 on e_1 , e_2 , e_i respectively.

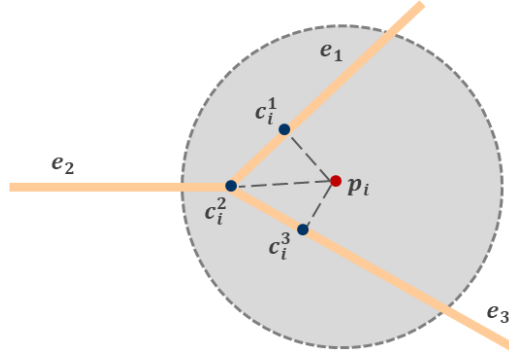


Figure 4 Candidate points for the GPS point p_i

3.2.2 Local attributes and global similarity evaluation

Local attributes and global similarity evaluation are significant processes in map matching. Local attributes consist of geometric attributes and topological information. The heuristic value of local attributes for each candidate link is defined as:

$$H_{(c_i^m \rightarrow c_{i+1}^n)} = h_{1(c_i^m \rightarrow c_{i+1}^n)} + h_{2(c_i^m \rightarrow c_{i+1}^n)} + h_{3(c_i^m \rightarrow c_{i+1}^n)} \quad (1)$$

Where $h_{1(c_i^m \rightarrow c_{i+1}^n)}$ represents the geometric error, which reflects geometric properties; $h_{2(c_i^m \rightarrow c_{i+1}^n)}$ s heading error, which is related to the road network direction; $h_{3(c_i^m \rightarrow c_{i+1}^n)}$ is routing error, which represents road connectivity.

$$H_{norm} = \frac{2 \times \arctan(H_{(c_i^m \rightarrow c_{i+1}^n)})}{\pi} \quad (2)$$

After normalization, heuristic value H_{norm} ranges from 0 to 1. The higher heuristic value, indicates the better map matching result from local perspective.

Global similarity is the similarity between GPS trajectory and road segment. It is measured by fitness value to reflect the quality of global map matching of entire GPS trajectory. It is defined as:

$$F_{norm}(T, P) = \frac{1}{k} \sum_{i=1}^k \min \left\{ \frac{l_i}{e_i}, \frac{e_i}{l_i} \right\} \quad (3)$$

where k is the number of road segments in the path, l_i is the edge between two adjacent points from the trajectory. Fitness value F_{norm} ranges from 0 to 1 and the lower heuristic value indicates the worse map matching result from the global similarity view.

3.3.3 Ant colony-based path optimization

Ant colony (ACO) is a graph-based optimal pathfinding algorithm, that simulates ant behavior in finding the shortest way from their nest to a food source. In this study, the optimal path will be determined considering the local heuristic value and global fitness value. The ACO approach outperforms other path search algorithms in terms of computing efficiency and robustness in a variety of scenarios.

4. Discussion and results

4.1 Results of Map matching

The map matching process was performed in parallel utilizing 20 computation cores on UCL Myriad HPC. The average running time was 11.2 seconds for each sequence. As shown in Figure 5, 3,133 sequences were matched to 29,963 road segments in Inner London.

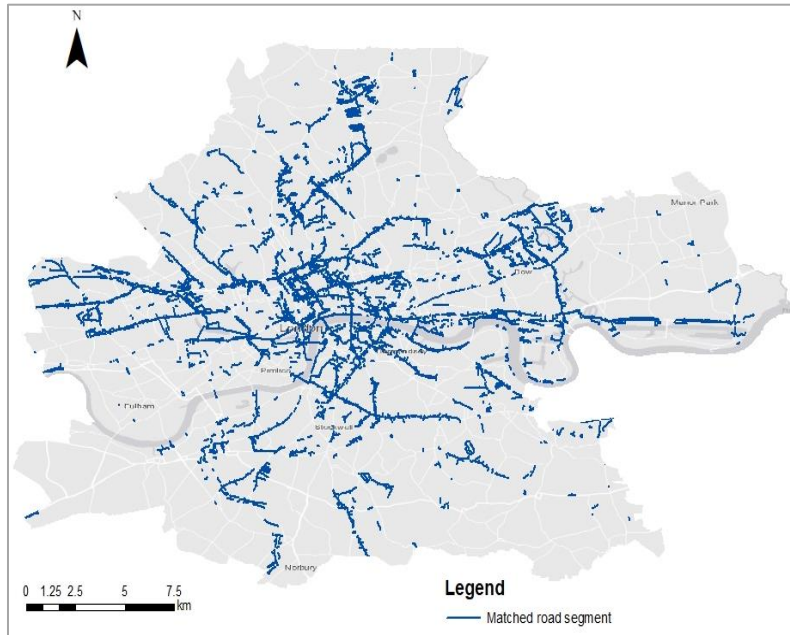


Figure 5 Result of map matching in Inner London

4.2 Discussion and limitations

Despite the good performance of the preliminary result, one significant limitation is that Mapillary image content was not included in the preprocessing work. Mapillary data is street-level imagery data. The heterogeneity of image content is a significant factor in data clean and map matching. Another limitation is

that walkways were not integrated in the road network, despite the fact that a considerable number of sequences were collected when walking. It might result in some mismatched results.

5. Conclusions and future work

This study provides a preprocessing workflow for crowdsourced Mapillary data to clean noise and dirty image data, as well as to match GPS points based on the AntMapper method. This work ensures the data quality and consistency of the crowdsourced Mapillary dataset for future urban analytics in diverse spatiotemporal scenarios. Further research should consider visual content and walkways in the workflow. Furthermore, ground truth data is required to test and validation the performance of AntMapper on Mapillary data.

References

- Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning* 215, 104217. <https://doi.org/10.1016/j.landurbplan.2021.104217>
- Gong, Y.-J., Chen, E., Zhang, X., Ni, L.M., Zhang, J., 2018. AntMapper: An Ant Colony-Based Map Matching Approach for Trajectory-Based Applications. *IEEE Transactions on Intelligent Transportation Systems* 19, 390–401. <https://doi.org/10.1109/TITS.2017.2697439>
- Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., Huang, Y., 2009. Map-matching for low-sampling-rate GPS trajectories, in: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09*. Presented at the the 17th ACM SIGSPATIAL International Conference, ACM Press, Seattle, Washington, p. 352. <https://doi.org/10.1145/1653771.1653820>
- Mahabir, R., Schuchard, R., Crooks, A., Croitoru, A., Stefanidis, A., 2020. Crowdsourcing Street View Imagery: A Comparison of Mapillary and OpenStreetCam. *IJGI* 9, 341. <https://doi.org/10.3390/ijgi9060341>
- Peng, J., Cao, Y., Ding, Z., Yan, J., 2019. A Fast Real-time Map-Matching for Unstable Sampling-rate GPS Trajectories, in: *Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City*. Presented at the ICIT 2019: IoT and Smart City, ACM, Shanghai China, pp. 253–258. <https://doi.org/10.1145/3377170.3377195>
- Ridzuan, F., Zainon, W.M.N.W., 2019. A Review on Data Cleansing Methods for Big Data. *Procedia Computer Science* 161, 731–738. <https://doi.org/10.1016/j.procs.2019.11.177>

Biography

Meihui Wang is a PhD candidate in the Department of Civil, Environmental and Geomatic Engineering at UCL. Her research interest includes street-level imagery, spatio-temporal data mining, deep learning and urban planning.

James Haworth is an Associate Professor, in Spatio-temporal Analytics at UCL, with research interests in spatiotemporal modeling and analytics, in particular applications of machine learning, deep learning and computer vision to the geoinformation sciences.