Grab-Posisi-L: A Labelled GPS Trajectory Dataset for Map Matching in Southeast Asia

Zhengmin Xu
GrabTaxi Holdings
zhengmin.xu@grab.com

Yifang Yin
NUS, Singapore
idsyin@nus.edu.sg

Chengcheng Dai
GrabTaxi Holdings
chengcheng.dai@grab.com

Xiaocheng Huang
GrabTaxi Holdings
xiaocheng.huang@grab.com

Robinson Kudali
GrabTaxi Holdings
robinson.kudali@grab.com

Jinal Foflia
GrabTaxi Holdings
jinal.foflia@grab.com

Guanfeng Wang
GrabTaxi Holdings
guanfeng.wang@grab.com

Roger Zimmermann
NUS, Singapore
rogerz@comp.nus.edu.sg

ABSTRACT

Map matching has long been a fundamental yet challenging problem. However, there are currently only a few public small-scale map matching benchmark datasets. Both the GPS trajectories and the road network in the existing map matching datasets are represented by location only, which cannot support the development of data-driven and semantic-enriched map matching algorithms that have increasingly emerged in recent years. To bridge the gap, we present the first large-scale attribute-rich map matching benchmark dataset covering two cities in Southeast Asia (*i.e.*, Singapore and Jakarta). Our GPS trajectories contain rich contextual information including the accuracy level, bearing, speed, and transport mode in addition to the latitude and longitude geo-coordinates. The underlying road network is a snapshot of the OpenStreetMap where roads are associated with rich attributes such as road type, speed limit, *etc.* To ensure the quality of our dataset, the annotation of the map-matched routes has been conducted by a team of professional map operators. Analysis on our dataset provides new insights into the challenges and opportunities in map matching algorithms.

CCS CONCEPTS

• **Information systems** → *Geographic information systems.*

KEYWORDS

GIS, datasets, map matching, GPS trajectories, digital maps

ACM Reference Format:
Zhengmin Xu, Yifang Yin, Chengcheng Dai, Xiaocheng Huang, Robinson Kudali, Jinal Foflia, Guanfeng Wang, and Roger Zimmermann. 2020. Grab-Posisi-L: A Labelled GPS Trajectory Dataset for Map Matching in Southeast Asia. In *28th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '20), November 3–6, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397536.3422218

1 INTRODUCTION

Map matching algorithms aim to determine the correct route where a vehicle has traveled on given the noisy GPS traces. Though the correction of the raw positioning data has been important for a variety of downstream applications, there is currently no public large-scale benchmark dataset for map-matching evaluation [9]. The few existing map matching datasets [7, 8] have limitations such as 1) the number of GPS trajectories is small (*i.e.*, less than 100), and 2) both the GPS traces and the map data are represented by location only. These drawbacks make the existing map matching datasets infeasible for the evaluation of recently proposed data-driven or semantic-enhanced map matching algorithms where large-scale historical GPS trajectories with rich attributes are required [4, 12]. Therefore, researchers mostly have to evaluate their contributions based on private, simulated, or augmented datasets, which makes the comparison between different methods highly difficult.

To solve the above issues, we present *Grab-Posisi-L*[1], the first large-scale real-world GPS trajectory dataset with digital map and manually labelled map-matched routes in two cities (*i.e.*, Singapore and Jakarta) in Southeast Asia. The GPS dataset contains both motorcar and motorcycle drivers' trajectories with rich sensor data including latitude, longitude, accuracy level, bearing, speed, and timestamp. All GPS trajectories are sampled from in-transit drivers of Grab, which is the leading ride-sharing company in Southeast Asia [6]. The digital map we provide is a snapshot of OpenStreetMap (OSM) [5] in April 2020. In addition to the road network topology, OSM roads are also associated with semantic attributes such as road type, speed limit, number of lanes, *etc.*, which provide contextual information that can be used to enhance the map matching results.

The map-matched routes are manually labelled by a team of professional map operators, who have been trained to label road segments specifically for map-matching purpose. To facilitate the annotation, we generate initial results based on the HMM map matching algorithm. We visualize and present the initial results to the annotators on a map interface in JOSM [2]. The annotators can then check whether the the trajectory perfectly follows the routes or not, and make necessary corrections if there are any incorrect match or missing road segments. Since we have not recorded the actual routes the drivers travelled on, we cannot guarantee absolute correctness of the manual annotations. To ensure the quality of our dataset, we randomly select 10% of the manual annotations to go through a second-round human validation. The results show that around 97% of the sample data has a difference of less than five segments from the original annotation, where the fraction

SIGSPATIAL '20, November 3–6, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8019-5/20/11.
https://doi.org/10.1145/3397536.3422218

[1]The dataset is available upon request sent to grab-posisi.geo@grab.com

of disagreement is quite small as the average segment count of a trajectory is around 281 in our dataset.

2 DATASET COLLECTION

2.1 GPS Trajectories

In this dataset, we collected labelled data for motorcar and motorcycle drivers' trajectories in Singapore (SIN) and Jakarta (JKT). The GPS trajectories of SIN motorcar, JKT motorcar and JKT motorcycle are sampled from Grab-Posisi dataset [6], maintaining the equivalent spatial coverage as the original dataset. To bring more data diversity, we collected Grab motorcycle drivers' trajectories in Singapore as well. The new traces were collected in April, 2020 from drivers' devices. Same as Grab-Posisi dataset, the GPS sampling rate of these trajectories is as high as 1 second except some GPS breakage cases. The rich attributes of the GPS points, *i.e.*, speed, bearing and accuracy, are also provided in the same format as in Grab-Posisi dataset (Table 1).

Table 1: Attributes of GPS Pings

Attribute	Data Type	Remark/Format
Trajectory ID	string	identifier for the trajectory
Latitude	float	WGS84
Longitude	float	WGS84
Timestamp	bigint	UTC
Accuracy Level	float	circle radius, in meter
Bearing	float	degrees relative to true north
Speed	float	in meters/second

2.2 OSM Map Representation

OpenStreetMap (OSM) [3] represents map topology with nodes and ways. Each node, uniquely identified by a node ID, is associated with its latitude and longitude. For instance, node 1790582708[2] refers to a point on Central Expressway (CTE) in Singapore. A way consists of a sequence of nodes depicting the curvature of it, thus it is not necessarily an edge between two intersection nodes. A way is also enriched with various attributes, such as road class (highway or residential), lanes, speed limit, *etc.*

Conventionally, we refer *segments* [3] to be the connection between two consecutive nodes in a way. Segments inherit the underlying way attributes. Each segment can be identified by an OSM node pair, *e.g.*, (1790582708, 5982519924), the order of which indicates the direction of the road it belongs to. We say 1790582708 is the *start node* and 5982519924 is the *end node* of the segment, respectively. We say a segment S_i is *directly connected* to a segment S_j if the shortest path from S_i to S_j consists of only S_i, S_j themselves. We also say there is a *gap* between S_i and S_j if S_i is not directly connected to S_j. The notion of gap is to cover the corner case where S_i's end node is S_j's start node but a turn restriction prohibits S_i to be connected to S_j directly. Consequently, either there is no path between S_i and S_j because of the turn restriction, or a path exists between S_i and S_j by making a u-turn at the end node of S_i and reaching S_j through other segments. In either case, S_i is not directly connected to S_j. The notion of direct connection and gap will be used later in Section 2.3.3 for data post-processing.

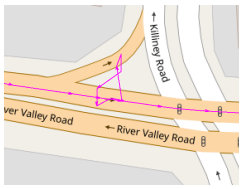
Segments serve as the basic units in our annotation. The reasons why we chose segments over ways are threefold. Firstly, OSM segments are universal. Any map matching algorithm developed based on OSM data can utilise our dataset for performance evaluation. Secondly, with OSM segments, the underlying way and way attributes can be retrieved from OSM for better development of attribute-enriched map matching algorithms. Last but not least, OSM segments can more easily adapt to map version upgrade because OSM is collaboratively edited daily.

To facilitate the usage of this dataset, we also provide the map data which was used for our annotation. The map data we use is a snapshot of the OSM maps of Singapore and Jakarta in April 2020 in standard PBF format, where the nodes and ways are encoded into protocol buffers.

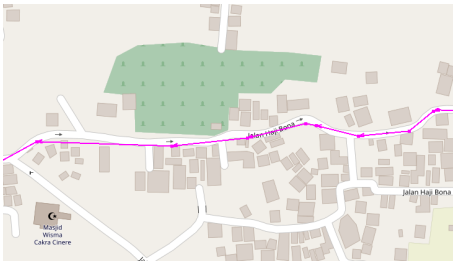
2.3 Route Annotation Design

2.3.1 Annotation Challenges. Although this task appears to be simple and well-defined, there are some subtleties within. The first problem is whether we annotate each GPS point with a segment or annotate a trajectory with a sequence of segments. Ideally, each GPS point in a trajectory corresponds to exactly one segment. When doing the annotation practically, however, the segment which a point belongs to is not always obvious. For instance, there may exist multiple reasonable segments matching a GPS point falling in an intersection (Figure 1a). Moreover, the right label of a jumping-back GPS point (Figure 1b) is also arguable. We could match it to the segment matching its previous point, or mark it as a match failure. Due to these ambiguities, we decided to annotate the whole trajectory, without explicit point-to-segment correspondence.

(a) Two points falling in an intersection where it is both fair to label them to the pink road and the yellow road.



(b) Jumping-back points near a Y-split in a trajectory moving from left to right.



(c) Trajectory moving from right to left, conflicting with the one-way road it seems to be on.

Figure 1: Annotation challenges

[2]https://www.openstreetmap.org/node/1790582708

The second challenge is maintaining the map data consistency throughout the annotation process (3 months in our case). As OSM is collaboratively updated daily, it can be anticipated that some nodes become invalid (*e.g.*, because of deletion) and some new nodes are added in the midst of our annotation. Due to the limitation of the visualisation tool we use (JOSM [2]), it is infeasible to load and visualise the entire map when the map size is too large. Therefore, we break it down into two stages, annotating on OSM real-time maps and post-processing to keep all the segments consistent with one single map version. We explain the procedure of post-processing in Section 2.3.3.

Another practical issue is that some trajectories conflict with the road restrictions or rules defined by the map. Typical cases include driver moving in the opposite direction to a road which is tagged one-way (Figure 1c), driver making a u-turn at an intersection which doesn't allow that, *etc.* It is beyond this paper's scope to fix driver behaviors or map imperfections. For those cases, our annotaters filter out trajectories that have a large portion of points having those issues.

2.3.2 Annotation Workflow. It is tedious to annotate from scratch. So we first map match all the GPS trajectories using the HMM algorithm [8], de-duplicate segments in each trajectory, and save segment node IDs of each trajectory in a CSV file for annotators to make fixes. Annotators would then visualise the trajectory and segments in JOSM[2] and check whether the segments perfectly follow the trajectory. Figure 2 shows a visualisation example. They can refer to the latest map downloaded from OSM and edit the segment node IDs in the CSV file if any segment is incorrect or missing.

The annotation was undertaken by a team of professional map operators, who have worked on OSM for years and are trained to label segments for map-matching's purpose.

2.3.3 Data Cleaning and Post-processing. For data cleaning, we select a map version and remove all the segments that are invalid in this map. Recall that segments might become invalid due to OSM deletion or the inaccessibility as indicated by their tags in OSM.

It is worth noting that the annotation workflow might result in gaps in a segment sequence where two consecutive segments are not directly connected. We post-process the sequence to fill the gaps with *reasonable* assumptions. The process we develop is based on heuristics from our empirical experience, as follows:

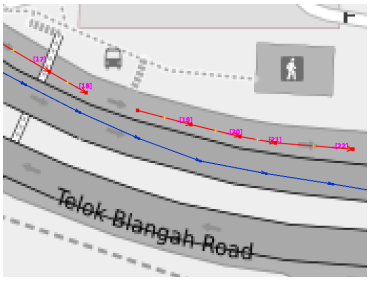


Figure 2: Visualization of a GPS trajectory (blue) and the segments (red) generated by a map-matching algorithm in JOSM. Numbers indicate the indices of the segments.

(1) Leave gaps larger than 1KM (haversine distance) untouched because they are more likely to be caused by missing GPS signal.
(2) Fill in a gap within 1KM between two consecutive segments. Given a segment sequence from annotations, we scan it in sequential order to identify a gap between consecutive segments S_i and S_{i+1}.
   - We find the shortest route between S_i and S_{i+1} using Dijkstra's algorithm. If the routing distance is less than 2.5 times of the haversine distance, we insert the shortest route in between.
   - Otherwise, we expand the route search to be between the segment before S_i *i.e.*, S_{i-1} and the segment after S_{i+1} *i.e.*, S_{i+2}. The route search terminates after maximum 4 expansions.
   - Note that while segment S_{i-1} is directly connected to S_i, there is no guarantee that S_{i+1} is directly connected to S_{i+2}. We terminate early when a gap between S_{i+1} and S_{i+2} is found.
(3) Repeat the process for the resulting segment sequence until no more gaps can be filled automatically.
(4) Annotators will have a second review on the remaining gaps. We observe that most of such cases are caused by map issues, such as missing roads, the inaccessibility of segments in between, or road restrictions.

Note that this procedure can be easily extended to any future OSM map version to keep the dataset up-to-date.

2.3.4 Quality Control. We control the annotation quality by checking randomly sampled 10% of the annotations. 97% of the sample data has a difference of less than 5 segments from the original annotation.

2.3.5 Annotation Format. The segments we provide are OSM node ID pairs in CSV format. Each trajectory corresponds to one CSV file named as the trajectory ID containing a sequence of segments (Figure 3). Segment ID indicates the index of each segment. Start node ID and end node ID indicate the OSM node IDs of each segment.

```
Segment ID,Start Node ID,End Node ID
0,4693679859,4693679841
1,4693679841,4693679842
2,4693679842,4693679846
...
```

Figure 3: Sample Annotations

3 DATASET ANALYSIS

We provide map-matched route annotation for 2,937 GPS trajectories covering 2 cities in Southeast Asia (Singapore and Jakarta) in 2 transport modes, *i.e.*, motorcar and motorcycle. Table 2 shows the summary statistics of the dataset. There are a total number of 2,055,252 GPS points in all trajectories and 827,241 OSM segment labels. Note that for JKT trajectories, total segments are more than total pings. This is because the road curvature in Jakarta, especially

Table 2: Statistics Summary

Category	Trajectories	Total Points	Total Segments
SIN Motorcar	1,774	1,921,127	583,130
SIN Motorcycle	300	59,233	46,530
JKT Motorcar	446	40,273	99,720
JKT Motorcycle	417	34,619	97,861
Overall	2,937	2,055,252	827,241

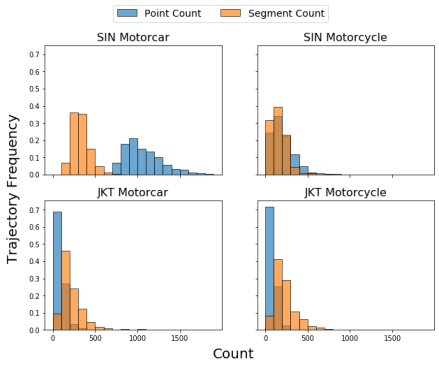


Figure 4: The histogram of trajectory length in terms of GPS point count and segment count.

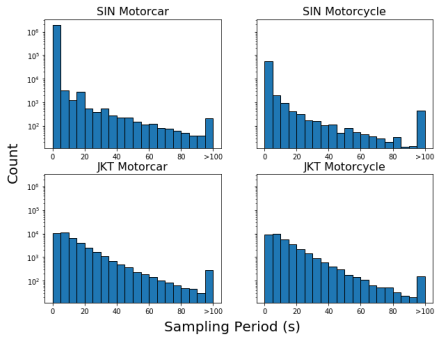


Figure 5: The distribution of sampling period in each category (log scale).

suburban area, is high and hence more segments are needed to capture the shape of the roads. Figure 4 further shows the distribution of trajectory length in each category in terms of GPS point count and segment count. Overall, the average segment count of a trajectory is around 281. Due to the annotation challenges mentioned before, there are a small number of gaps in the route annotation. In all four categories, there are below 0.2% of the gaps.

Figure 5 shows the distribution of sampling period in each category. For SIN Motorcar and SIN Motorcycle, above 90% of the points have a sampling period lower than 5 seconds. For JKT data, 80% of the points have a sampling period lower than 20 seconds.

4 RELATED WORK

There are a number of large-scale GPS trajectory datasets that are publicly available such as the Microsoft GeoLife [11] and the Grab-Posisi dataset [6]. However, without the corresponding map data and the ground-truth of the map matched routes, such GPS trajectory datasets are difficult to be used for the evaluation of map matching algorithms. To the best of our knowledge, there are only two widely used benchmark datasets for map-matching evaluation. Newson and Krumm provided a complete dataset with tracks, map, and map-matched routes of a 3-hour drive in Seattle, WA, USA [8]. Kubička *et al.* selected 100 GPS trajectories all over the world from a publicly available collection called Planet GPX [1] and manually labeled the map-matched routes by human annotators [7]. However, both datasets have drawbacks that cannot support the development of advanced map matching algorithms proposed in recent years. First, the datasets are too small-scale to be used for data-driven map matching approaches [12]. Second, the GPS trajectories and the road network in the benchmark datasets only contain the location information given by latitude and longitude, which hinders the development of semantic-enriched map matching algorithms [4, 10]. Therefore, it is critical to have a large-scale and attribute-rich benchmark dataset to accelerate the evolution of advanced map matching algorithms.

5 CONCLUSIONS

We present a large-scale real-world GPS trajectory dataset with a digital map and manually labelled map-matched routes in support of the development of map matching algorithms. Our dataset consists of 2,937 GPS trajectories, which is the largest among the public map matching benchmark datasets. Moreover, our dataset is attribute-rich where the GPS traces are associated with vehicle's sensor data such as bearing and speed and the road segments are associated with attributes such as road type and speed limit. The dataset is of significant importance for developing and benchmarking advanced map matching algorithms that are based on, *e.g.*, machine learning and semantic analysis.

REFERENCES

[1] 2015. Planet GPX. http://planet.openstreetmap.org/gps/
[2] 2020. JOSM. https://josm.openstreetmap.de/
[3] 2020. Segment entry of OpenStreetMap wiki. https://wiki.openstreetmap.org/wiki/Segment
[4] Heba Aly and Moustafa Youssef. 2015. semMatch: Road Semantics-based Accurate Map Matching for Challenging Positioning Data. In *ACM SIGSPATIAL*. 5:1–5:10.
[5] M. Haklay and P. Weber. 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing* (2008), 12–18.
[6] Xiaocheng Huang, Yifang Yin, Simon Lim, Guanfeng Wang, Bo Hu, Jagannadan Varadarajan, Shaolin Zheng, Ajay Bulusu, and Roger Zimmermann. 2019. Grab-Posisi: An Extensive Real-Life GPS Trajectory Dataset in Southeast Asia. In *ACM SIGSPATIAL International Workshop on Prediction of Human Mobility*. 1–10.
[7] Matěj Kubička, Arben Cela, Philippe Moulin, Hugues Mounier, and Silviu-Iulian Niculescu. 2015. Dataset for Testing and Training of Map-matching Algorithms. In *IEEE Intelligent Vehicles Symposium*. 1088–1093.
[8] Paul Newson and John Krumm. 2009. Hidden Markov Map Matching Through Noise and Sparseness. In *ACM SIGSPATIAL*. 336–343.
[9] Guanfeng Wang and Roger Zimmermann. 2014. Eddy: An Error-bounded Delay-bounded Real-time Map Matching Algorithm Using HMM and Online Viterbi Decoder. In *ACM SIGSPATIAL*. 33–42.
[10] Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. 2016. A General Feature-Based Map Matching Framework with Trajectory Simplification. In *ACM SIGSPATIAL International Workshop on GeoStreaming*.
[11] Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang. 2010. T-Drive: Driving Directions Based on Taxi Trajectories. In *ACM SIGSPATIAL*. 99–108.
[12] Kai Zhao, Jie Feng, Zhao Xu, Tong Xia, Lin Chen, Funing Sun, Diansheng Guo, Depeng Jin, and Yong Li. 2019. DeepMM: Deep Learning Based Map Matching with Data Augmentation. In *ACM SIGSPATIAL*. 452–455.