Grab-Posisi: An Extensive Real-Life GPS Trajectory Dataset in Southeast Asia[*]

Xiaocheng Huang[#], Yifang Yin[+], Simon Lim[#], Guanfeng Wang[#], Bo Hu[#]
Jagannadan Varadarajan[#], Shaolin Zheng[#], Ajay Bulusu[#], Roger Zimmermann[+]
[#]GrabTaxi Holdings, Singapore
{xiaocheng.huang,simon.lim,guanfeng.wang,bo.hu,jagan.varadarajan,shaolin.zheng,ajay.bulusu}@grabtaxi.com
[+]National University of Singapore
idsyin@nus.edu.sg,rogerz@comp.nus.edu.sg

ABSTRACT

Real-world GPS trajectory datasets are essential for geographical applications such as map inference, map matching, traffic detection, *etc.* Currently only a handful of GPS trajectory datasets are publicly available and the quality of these datasets varies. Most of the existing datasets have limited geographical coverage (a focus on China or the USA), have low sampling rates and less contextual information of the GPS pings. This paper presents *Grab-Posisi*, the first GPS trajectory dataset of Southeast Asia from both developed countries (Singapore) and developing countries (Jakarta, Indonesia). The data were collected very recently in April 2019 with a 1 second sampling rate, which is the highest amongst all the existing open source datasets. It also has richer contextual information, including the accuracy level, bearing, speed and labels trajectories by data acquisition source (Android or iOS phones) and driving mode (Car or Motorcycle). The dataset contains more than 88 million pings and covers more than 1 million kms. Experiments on the dataset demonstrate new challenges for various geographical applications. The dataset is of great value and a significant resource for the community to benchmark and revisit existing algorithms.

CCS CONCEPTS

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

KEYWORDS

Datasets, GPS, Trajectory

1 INTRODUCTION

Real-world GPS trajectory datasets are essential for geographical applications [12] such as map inference, map matching, traffic prediction *etc.* Map inference refers to the task of reconstructing road networks from GPS trajectories [11]. The process of map matching aims to snap GPS trajectories onto a road network [20, 31, 40, 41]. Finally, the goal of traffic prediction is to forecast the traffic situation at a future time [42, 44]. All of these applications depend on the presence of a real-world GPS trajectory dataset.

However, currently only a handful of GPS trajectory datasets are publicly available and the quality of the datasets varies (see Section 2 for details). Most of the existing open source datasets have at least one of the following drawbacks:

- they are outdated;
- they are geographically limited (coming from China and USA, they focus only on structured scenarios); none of the top 10 cities with the worst traffic in the TomTom traffic index [5] is included;
- they cover only a small area of a city;
- they have a low sampling rate (seconds to minutes);
- they contain less contextual information (*e.g.*, no accuracy level, bearing, speed); or
- they are small scale.

This paper addresses all the above issues and presents *Grab-Posisi*[1], the first GPS trajectory dataset of Southeast Asia from both developed countries (Singapore) and developing countries (Jakarta, Indonesia), where Jakarta is ranked at the 7th place in TomTom's traffic index [5]. The data were collected from Grab drivers' phones while in transit. Grab is Southeast Asia's leading ride-sharing company [3]. The term *Posisi* refers to *position* in Bahasa, hence we name the dataset **Grab-Posisi** capturing its essential Southeast Asian nature.

The whole dataset contains in total 84K trajectories that consist of 88,847,080 GPS pings, which cover 1,003,510 km over a duration of 30,104 hours. The data were collected very recently in April 2019 with a 1 second sampling rate, which is the highest amongst all the publicly available datasets. It also has richer contextual information, including the *accuracy level*, *bearing* and *speed*. The accuracy level is important because GPS measurements are noisy and the true location can be anywhere inside a circle centered at the reported location with a radius equal to the accuracy level. The bearing is the horizontal direction of travel, measured in degrees relative to true north. Finally, the speed is reported in meters/second over ground. We will elaborate in Section 4 how such contextual information can be leveraged to improve existing solutions of various geographical applications.

In addition, this dataset was collected from smartphones and labeled by data acquisition source. Compared with existing datasets, where data were collected mostly from taxis' dedicated devices, GPS

[*] The dataset must not be assumed to indicate any of Grab's business interest

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

PredictGIS'19, November 5, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6964-0/19/11…$15.00
https://doi.org/10.1145/3356995.3364536

trajectories collected from smartphones are less stable and more challenging to handle [29]. Since all the trajectories were collected from Grab drivers' phones while in transit, we label each trajectory by phone device type being either Android or iOS. To the best of our knowledge, this is the first dataset which differentiates such device information. Experiments in Sections 3 and 4 demonstrate very different data characteristics between Android and iOS.

Furthermore, we label the trajectories by driving mode (Car or Motorcycle). While cars are common in developed countries such as the USA, motorcycles on two-wheels are one of the main modes of transportation in Southeast Asia. Ride-sharing and delivery industries (*e.g.,* food delivery and logistics) in Southeast Asia mostly use motorcycles [21, 23]. Our dataset provides an opportunity for the research community to make a significant impact by performing deeper investigations into the motorcycle ecosystem.

The remaining parts of this paper are organised as follows. Section 2 presents related work that compares the pros and cons between the existing datasets and *Grab-Posisi*. Section 3 describes the methodology, volume, data format and statistics of the dataset. Section 4 elaborates on how the dataset can be used in various applications and also exemplifies new challenges in real-world applications. Section 5 concludes the paper.

2 RELATED WORK

Only a handful of GPS trajectory datasets are publicly available, but the quality of the datasets varies. Besides latitude, longitude and timestamps that all the datasets include, Table 1 compares the datasets' geography, data source, production year, sampling rate and the contextual information contained (accuracy level, bearing, speed). The data acquisition sources are in general classified into two major categories: smartphones and dedicated GPS devices. As the authors of [29] pointed out, compared with dedicated GPS devices, the data quality of GPS trajectories generated from smartphones is less stable, which poses new challenges in geographical applications. Previously most of the GPS trajectory datasets were collected from cars [11, 14, 22, 25, 38, 45–47], a method which was not common in developing countries. As a trend, with the wide adoption of smartphones, more and more GPS data have become available [8].

Microsoft [43, 45–47] published a GPS trajectory dataset with transportation mode labels (driving, taking a bus, riding a bike and walking) collected in 2008. It covers 28 big cities in China and some cities in the USA, South Korea, and Japan. The GPS devices are composed of stand-alone GPS receivers (Magellan Explorist 210/300, G-Rays 2 and QSTARZ BTQ-1000P) and GPS-equipped phones. The sampling rate is between 2–5 seconds. Each GPS ping contains the information of latitude, longitude, altitude, speed and bearing. However, at current time, this dataset is rather outdated.

Tsinghua University [25] has provided a taxi trajectory dataset of Beijing collected in May 2009 with 129 million GPS pings. 75.36% of the pings are at the highest sampling rate of 1 minute. Bearing and speed are provided. Additional information such as if a taxi is vacant or not is also provided. However, the sampling rate is too low for many applications.

ETH Zurich [32] referred to a GPS trajectory dataset collected by a private sector company in 2009 from 4,882 participants using an on-person GPS logger for 6.65 days on average. Zhejiang University [19] used a dataset collected by Hangzhou City Traffic Bureau, which were generated by GPS devices on 7,475 taxis from April 2009 to April 2010. The dataset has about 3 billion records and is sampled at a frequency of about 1 minute. Unfortunately we are not able to locate either of the two datasets on the Internet.

The University of Illinois at Chicago [14] published a GPS trajectory dataset collected by its University of Illinois at Chicago shuttle buses in 2012. The sampling rate is by far the highest at 3 seconds, covering 2,869 km. Nonetheless, this dataset only covers a small area of the city.

In [38], Beijing Jiaotong University mentions a GPS trajectory dataset from Xi'An, China. However, the dataset does not seem to be publicly available on the Internet. Similarly, in [22], the Tsinghua-Berkeley Shenzhen Institute mentions a GPS trajectory dataset from Beijing, China. However, the dataset also does not seem to be open source.

Another publicly available GPS trajectory dataset was collected by the authors of [11] in 2015. It covers two major cities, namely Athens and Berlin. The Athens dataset was obtained from a school bus, with a sampling rate of 20 seconds to 30 seconds, covering 7,224 km. The Berlin dataset was obtained from a taxi fleet, with a sampling rate of 40 seconds, covering 41,116 km. Nonetheless, this dataset only covers a small area of each city.

The authors of [27] published a GPS trajectory dataset generated by a single user jogging in Joensuu between 2014-11-16 and 2015-04-25. This dataset is rather small with 43,891 GPS pings. None of the values for bearing, accuracy level or speed are provided.

Didi Chuxing, as one of the biggest ride-sharing companies in China [8], launched the Didi Chuxing GAIA Initiative to share their drivers' GPS trajectories. Currently it covers two cities in China, Xi'an city and Chengdu city. The GPS trajectories shared are from 2016 with a sampling rate of around 2 to 4 seconds. Latitude and longitude of GPS pings are provided while bearing, accuracy and speed are not.

One study by Bolbol *et al.* [15] mentions a few small-scale GPS trajectory datasets. However, due to their small size they are less useful for experiments. There is also a Beijing Taxi trip dataset available in the IEEE DataPort [4]. However, the IEEE DataPort is a subscription service which many researchers may not be able to access.

It is worth noting that OpenStreetMap [2] maintains a crowd-sourced GPS trajectory repository. Users all over the world are free to upload their GPS tracks. Due to its crowd-sourcing nature, the GPS trajectories collected have a broad spectrum of characteristics and thus in order to utilize them a significant amount of effort would be required for data cleaning and data preprocessing.

This paper presents the first GPS trajectory dataset of Southeast Asia for both developed countries (Singapore) and developing countries (Jakarta, Indonesia). The data were collected recently during April 2019 with 1 second sampling rate, which is the highest amongst all the existing open source datasets. Furthermore, it contains rich contextual information such as bearing, accuracy level and speed.

Table 1: Available Datasets

Dataset	Geography	Data Source	Year	Sampling Rate	Bearing	Accuracy	Speed
Microsoft (GeoLife) [45–47]	China,USA, South Korea,Japan	GPS receivers and phones	2008	2-5 seconds	Y	N	Y
Tsinghua University [25]	China (Beijing)	Taxi	2009	1 minute	Y	N	Y
ETH Zurich [32]	Switerland (Zurich, Winterthur, Geneva)	Phone	2009	-	-	-	-
Zhejiang University[19]	China (Hangzhou)	Taxi	2010	1 minute	-	-	Y
University of Illinois at Chicago[14]	Chicago	School bus	2012	3 seconds	N	N	N
Beijing Jiaotong University [38]	China (Xi'An, Chengdu)	Taxi	2014	30 seconds	Y	N	Y
Tsinghua-Berkeley Shenzhen Institute [22]	China (Beijing)	Taxi	2012,2014,2015	1 minute	Y	N	Y
http://www.mapconstruction.org [11]	Athens, Berlin	School bus, Taxi	2015	30-40 seconds	N	N	N
University of Eastern Finland [27]	Finland (Joensuu)	Phone	2015	30 minutes	N	N	N
Didi Chuxing GAIA Initiative[8]	China (Xi'An)	Phone	2016	2-4 seconds	N	N	N
IEEE DataPort [4]	China (Beijing)	Taxi	2018	-	-	-	-
OpenStreetMap[2]	World-wise	-	-	-	-	-	-
This paper	Southeast Asia (Singapore/Jakarta)	Phone (Android and iOS)	2019	1 second	Y	Y	Y

3 DATASET

3.1 Methodology, Volume and Data Format

The dataset is sampled from Grab drivers' trajectories with the drivers' personal information encrypted and the real start/end locations removed. The collection dates range from 2019-04-08 to 2019-04-21 (inclusive, UTC) with 6,000 trajectories gathered per day. The trajectories were collected from drivers' phones during driving. The trajectories must not be assumed to indicate any of Grab's business interests.

Table 2: Trajectory Category

City	Mode	Device	Total Trajectories
Singapore (SIN)	Car	iOS	14K
Singapore (SIN)	Car	Android	14K
Jakarta (JKT)	Car	iOS	14K
Jakarta (JKT)	Car	Android	14K
Jakarta (JKT)	Motorcycle	iOS	14K
Jakarta (JKT)	Motorcycle	Android	14K

Table 2 shows the trajectory categories, which show the geographical coverage, the driving-mode and the device variation. We cover two cities in Southeast Asia, Singapore and Jakarta, representing a developed and a developing country, respectively. We also for the first time label trajectories by their driving mode being either Car or Motorcycle. We further categorise if the trajectories are collected from Android or iOS devices and will demonstrate in a later section that GPS quality varies by reporting devices. For each category we have collected 1,000 trajectories per day for 2 weeks from 2019-04-08 to 2019-04-21, and therefore each category includes 14,000 trajectories in total.

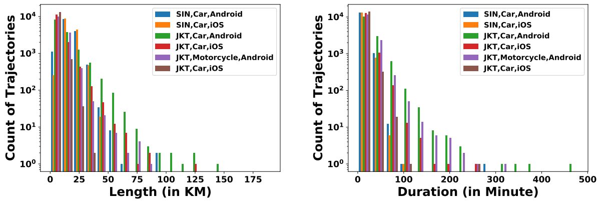
The whole dataset contains in total 84K trajectories that consist of 88,847,080 GPS pings, covering 1,003,510 km over a total span of 30,104 hours. The average trajectory length is 11.94 km and the average duration per trip is 21.50 minutes.

Each trajectory is serialised in a file in Apache Parquet format. The whole dataset size is around 2 GB. Each GPS ping is associated with values for a trajectory_ID, latitude, longitude, timestamp (UTC), accuracy level, bearing and speed (Table 3). The GPS sampling rate is 1 second, which is the highest among all the existing open source datasets.

Table 3: 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

Besides latitude and longitude of a GPS ping, the accuracy level, bearing, and speed provide the context when a GPS point is collected. When smartphones collect GPS pings, the operating system applies fuzzy logic to generate the context information from multiple location providers [7, 9]. The accuracy level indicates the accuracy in the horizontal plane. With Android devices [7], the accuracy level refers to the radius within which the location confidence is 68%. In other words, given a circle centered at the reported latitude and longitude, and with a radius equal to the accuracy level, then there is a 68% probability that the true location is inside the circle. A value of 0.0 indicates an unavailable accuracy. In iOS devices [9], the reported latitude and longitude identify the center of a circle with a radius of the reported accuracy level. The true location is assumed to be randomly distributed inside the circle. A negative accuracy level indicates that the latitude and longitude are invalid. Bearing is the horizontal direction of travel of this device, and is not related to the device's orientation. Bearing is measured in degrees relative to true north. In Android devices [7], the bearing ranges within (0.0, 360.0] degrees, where 0.0 indicates an invalid bearing. In iOS devices [9], if two consecutive GPS pings are at the same location, the bearing is 180. Speed is measured in meters/second over ground. In Android devices [7], 0.0 represents an invalid speed.



(a) Length **(b) Duration**

Figure 1: Trajectory Statistics

3.2 Trajectory Statistics

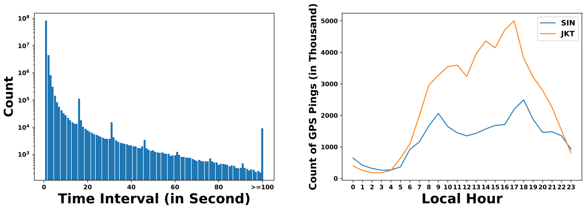
Figure 1 shows the distribution of the trajectory lengths and durations. The length of a trajectory is defined as the sum of the Haversine distance between two consecutive GPS pings that occur within 10 seconds and their Haversine distance is within 2 km. Figure 1a shows that most of the trajectories are within 50 km while the extreme trajectories go as long as 200 km. The average trajectory length is 11.90 km.

Figure 1b plots the distribution of the trajectory durations. Most of the trajectories have a time span within 1 hour. The average trajectory duration is 21.50 minutes.

3.3 Sampling Rate and Temporal Statistics

The GPS sampling rate is 1 second. Though we observe trajectory breakage in some cases due to several reasons, *e.g.*, a vehicle passing through a tunnel. Figure 2a shows the time interval between two consecutive pings. Most (> 90%) of the time intervals are 1 second.

Figure 2b shows the ping counts by local hour-of-the-day. It is clear that Singapore has two peaks, one around 9 am and the other around 6 pm. In contrast, Jakarta's traffic situation is more complex in that the evening peak hours are more evident than the morning peak hours. This observation sheds light on the hyper-active local characteristics of Southeast Asia.



(a) Sampling Rate **(b) Ping Count by Hour**

Figure 2: Sampling Rate and Temporal Statistics

3.4 Accuracy Level Distribution

One of the major contributions of this paper is its inclusion of the accuracy level in the dataset. In this section, we illustrate the different accuracy behaviors between Android and iOS devices. Figure 3 compares the accuracy level reported by Android and iOS devices from Singapore and Jakarta. The x-axis represents the accuracy level in meters. Recall that the accuracy level roughly indicates the radius of the circle within which the true location lies with a certain probability. The lower the accuracy level, the more precise the reported GPS ping is. The y-axis shows the normalised count of the GPS pings. The normalisation is such that the normalised ping count from different categories are comparable. Formally,

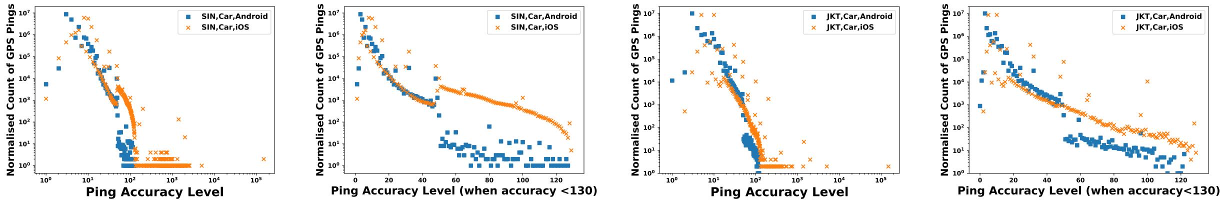
$$normalised\ ping\ count\ of\ accuracy\ level\ x\ for\ category\ cat$$
$$= \frac{ping\ count\ of\ accuracy\ level\ x\ of\ category\ cat}{ping\ count\ of category\ cat} \times C$$

where $C = 20,000,000$.

Figure 3a reports the normalised ping count distribution for categories ⟨*SIN, Car, Android*⟩ and ⟨*SIN, Car, iOS*⟩. It is clear that the accuracy levels reported by Android and iOS devices differ significantly. Firstly, the maximum accuracy level from Android devices is 127 meters while the iOS reported accuracy can be as large as 149 km. Secondly, the accuracy level reported by iOS devices has a long tail, meaning that the reported locations are less trustworthy. Figure 3c reports the normalised ping count distribution for Jakarta, and we observe a similar long-tail pattern.

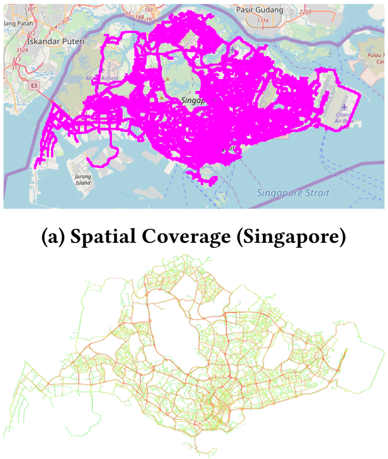
Figure 3b zooms into the accuracy level of less than 130 meters for ⟨*SIN, Car, Android*⟩ and ⟨*SIN, Car, iOS*⟩. We observe for both Android and iOS devices, as the accuracy level increases, the ping count first increases and then decreases around the accuracy level of 10 meters. Interestingly, at an accuracy level of 50 meters, the trend changes for both but in the opposite direction. The same pattern is also observed in Figure 3d for Jakarta, except that the trend for iOS is rather smooth.

The observations from the dataset about the accuracy levels should caution the usage of the accuracy level in applications.



(a) iOS vs Android in Singapore **(b) Singapore (Zoomed)** **(c) iOS vs Android in Jakarta** **(d) Jakarta (Zoomed)**

Figure 3: Accuracy Differs by Device



(a) Spatial Coverage (Singapore)

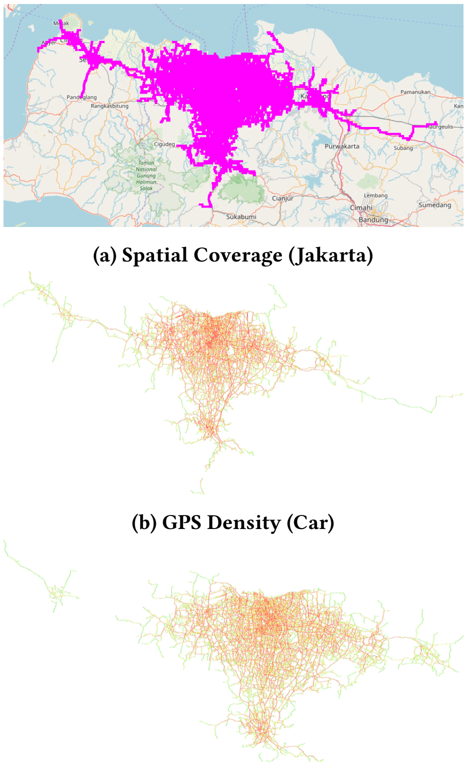
(b) GPS Density. Highways have more GPS.

Figure 4: Spatial Coverage (Singapore)

3.5 Spatial Coverage

Figures 4 and 5 show the spatial coverage of the dataset. Compared with existing datasets, which only cover a specific area of a city, the *Grab-Posisi* dataset encompasses almost the whole island of Singapore (Figure 4a). On the right (east) of Figure 4a, we can see that even the roads surrounding Changi Airport, which are more relevant for commercial traffic, are covered. Figure 4b depicts the GPS density in Singapore. Red color represents high density while green represents low density. Expressways in Singapore are clearly visible because of their dense GPS pings.

Figure 5a illustrates that the *Grab-Posisi* dataset encloses not only central Jakarta but also extends to external highways. Figure 5b depicts the GPS density of cars in Jakarta. Compared with Singapore, trips in Jakarta are spread out in all different areas, not only concentrated on highways. A similar pattern can be observed for motorcycles (Figure 5c). When comparing cars (Figure 5b) and motorcycles (Figure 5c) in Jakarta, it is worth noting that motorcycles tend to avoid highways. This hyper-local characteristic can be explored for better route planning by delivery industries in Indonesia.



(a) Spatial Coverage (Jakarta)

(b) GPS Density (Car)

(c) GPS Density (Motorcycle)

Figure 5: Spatial Coverage (Jakarta)

4 APPLICATIONS

4.1 On Map Inference

The quality of map data has a significant impact on the effectiveness of geo-applications including Geographic Information Systems (GIS), Intelligent Transportation Systems (ITS), Location-based Services (LBS), *etc.* However, the frequent, dynamic update of road networks in the traditional, manual way can be very time-consuming and labour-intensive, resulting in issues such as important roads being missed or real-time traffic conditions being unavailable. Recently, a significant number of research efforts have concentrated on reconstructing road networks from GPS trajectories automatically [10, 16, 33].

A high quality real-world dataset is essential for evaluating and comparing proposed map inference approaches [11]. As mentioned in Section 2, current public GPS datasets mostly have at least one of the following shortcomings: 1) small scale of the dataset, 2) sparse coverage, 3) low sampling rate, or 4) a single attribute where only GPS coordinates are available. The *Grab-Posisi* dataset addresses all these issues and provides the community with a large-scale, attribute-rich GPS trajectory dataset for the development of new map inference algorithms. Next, we showcase the basic use of the dataset in map topology and geometry inference, and in road attribute mining.

4.1.1 Map Topology and Geometry. A road network is generally denoted as a directed graph, where the edges represent road segments, and the vertices represent the starting and ending points of the road segments. Map inference refers to the problem of transforming a set of noisy GPS trajectories to the geographical graph structure to represent the road network. In this section, we evaluate the map inference algorithm proposed by Davies *et al.* [18] for an area in Jakarta (the region identified by geohash location value [1]=qqggg).

The map inference algorithm [18] works as follows. It first splits the 2D space in the horizontal plane into square cells units and generates a 2D histogram indicating the number of GPS pings found in each cell. The intuition is that the frequency value attached to each cell represents the confidence that the cell is part of a road. Note that it is likely that a cell with frequency value of 0 is surrounded by cells with a high(er) frequency. A blur filter is then used to fill small gaps by averaging cell values with neighboring cells. Next the algorithm binarizes the frequency values to be either 0 or 1, indicating whether there is a road or not, by thresholding. In our experiment we use Gaussian adaptive thresholding [6]. We generate a grey-scale image such that a pixel value equals 255 (*i.e.*, white) if the frequency value is 1. Figure 6 shows the generated grey-scale image for Jakarta in the region of geohash. As illustrated, it identifies most of the roads in the area.

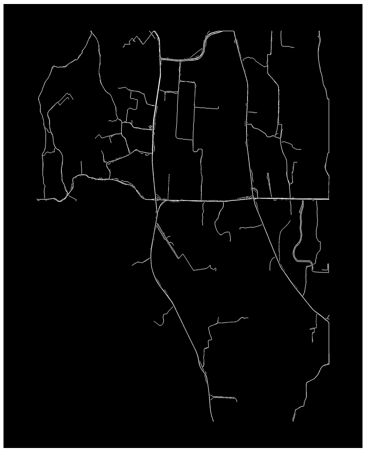


Figure 6: Grey-scale Image after Adaptive Thresholding (geohash=qqggg)

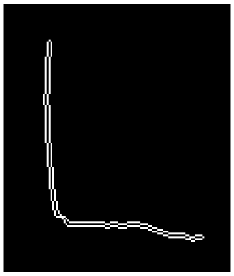
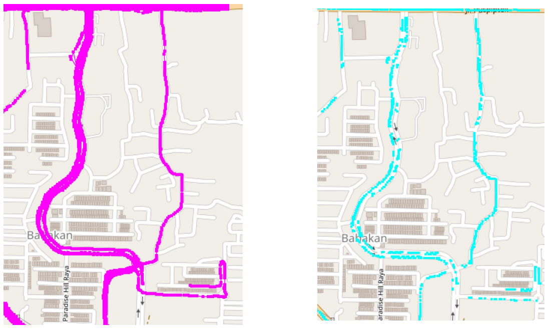


Figure 7: An example of a contour (corresponding to part of the top right corner of Figure 6)

From the grey-scale image we can now infer the road geometry. The shape of the roads is identified by applying a classic contour detection method, where a contour is a curve joining all the continuous points (along the boundary). Figure 7 shows one of the extracted contours. We can see that contours do not necessarily consist of straight line segments and can be of any shape. The center-line of a contour is deemed to be the road geometry. Davies *et al.* [18] use a Voronoi graph for center-line detection.

Figure 8 shows a snippet of the inferred map (Figure 8b) from our GPS trajectories (Figure 8a). We can see that the algorithm of [18] roughly infers the skeleton of the underlining map. However, the inferred curved road is broken into dis-joined pieces. Further, it miss-interprets the shape of the roundabout in the bottom right corner.

It is worth emphasizing that in countries like Indonesia, the road network can be very intricate and complex. The *Grab-Posisi* dataset can be of great value for benchmarking and improving existing map inference algorithms in such challenging real-world situations.

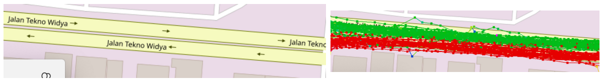


(a) Raw GPS Trajectories **(b) Inferred Map**

Figure 8: Map Inference Method using Kernel Density Estimation (KDE) [18]

4.1.2 Road Attributes. With the availability of dense historical GPS trajectories, it is also possible to automatically infer road semantics (*e.g.*, the directionality of roads such as one-way, the type of roads such as bridges, speed bumps, *etc.*) based on data-driven approaches. First, map matching algorithms need to be applied to infer the most likely road segment onto which each GPS point belongs. Next, the road attributes can be inferred based on all the GPS trajectories that traversed it.

Figure 9 exemplifies a road attribute inference, namely a *one-way directionality detection* of individual roads. Figure 9a depicts two one-way roads in Jakarta. Figure 9b overlays the *Grab-Posisi* GPS trajectory dataset on the two one-way roads, where red color indicates west direction of GPS trajectories and green color indicates east direction. It is quite evident that the *Grab-Posisi* GPS dataset can be used for one-way road detection.



(a) One-way roads

(b) GPS Trajectories (Red=West, Green=East)

Figure 9: Road Attribute Inference

4.2 On Map Matching

The problem of map matching refers to the task of automatically determining the correct route where the driver has traveled on a digital map, given a sequence of raw and noisy GPS points. The correction of the raw GPS data has been important for many location-based applications such as navigation, tracking, and road attribute detection as aforementioned [20, 31, 40, 41]. In this section, we demonstrate how rich attributes can be used in map matching algorithms.

Among the advanced statistics-based map matching algorithms, the Hidden Markov Model (HMM) is one of the most widely used methods, which models the road emission and transition probabilities based on the measurement noise level and the road network layout [30, 34]. It has been shown that HMM-based map matching obtains very good matching accuracy when dealing with high-sampling-rate GPS trajectories (*e.g.*, sampling intervals less than 30 seconds). More specifically, let $L = \{l_1, l_2, \ldots, l_n\}$ denote a sequence of raw GPS locations, and $E = \{e_1, e_2, \ldots, e_m\}$ denote a set of road segments of the road network. The emission probability is modeled as

$$P(l_t | p_t = e_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{dist(e_i, l_t)^2}{2\sigma^2}} \tag{1}$$

where σ is the standard deviation of GPS measurements, $dist(e_i, l_t)$ represents the minimum great circle distance between road segment e_i and GPS measurement l_t. $P(l_t | p_t = e_i)$ is the emission probability for road segment e_i, which can be interpreted as the likelihood that the GPS measurement l_t would be observed if the vehicle were actually on road segment e_i.

In practice, it is not feasible to calculate $P(l_t | p_t = e_i)$ for all $e_i \in E$. A common approach is to select only road segments within a distance radius $d_{threshold}$ of the GPS location l_t. It is debatable if $d_{threshold}$ is a constant. If $d_{threshold}$ is set to be small, then we may miss the real segment that the GPS location is on. This happens especially for noisy GPS locations that occur far away from their true locations. One the other hand, if $d_{threshold}$ is large, then we need to process a large number of possible segments in central city areas, where road networks are dense. This results in unnecessary computations and consequently higher running time.

Table 4: Snippet of Noisy Raw GPS Pings

Latitude and Longitude	Accuracy Level
1.2851133,103.8443881	45
1.2852771,103.8444593	42
1.2853581,103.8443996	45
1.2854508,103.8443068	44
1.2855187,103.8441327	43
1.2856159,103.8439763	46
1.2854326,103.8439028	28
1.2853808,103.8437961	22
1.2853013,103.8437471	17
1.2853081,103.8436901	16
1.2853379,103.8436345	15

The accuracy levels provided in the *Grab-Posisi* dataset can be of great use to address this issue. In this section, we demonstrate a simple usage of accuracy levels by setting $d_{threshold} = accuracy\ level \times 2$, such that different GPS pings utilize different distance thresholds. In addition, we set $\sigma = d_{threshold}$.

The transition probability of the HMM model is defined as

$$P(p_t = e_j | p_{t-1} = e_i) = \frac{1}{\beta} e^{-\frac{d_t}{\beta}} \tag{2}$$

where $d_t = |dist(l_t, l_{t+1}) - dist\_route(l_t, l_{t+1})|$. Functions $dist(l_t, l_{t+1})$ and $dist\_route(l_t, l_{t+1})$ calculate the Haversine distance and routing distance in meters between the two GPS measurements, respectively. The transition probability expresses the probability of a vehicle moving from road segment e_i to road segment e_j. In our experiment, we empirically set $\beta = 10$.

Finally, the Viterbi algorithm is utilized to compute the optimal path, by using dynamic programming to find the path that maximizes the product of the emission and transition probabilities. The recovered path provides a candidate road segment for each location measurement in L.

Figure 10 shows the map matching results (in fuchsia) from both a noisy GPS track (Figure 10a) and a high quality GPS track (Figure 10b) extracted from the *Grab-Posisi* dataset. Figure 10a clearly shows that the raw GPS pings are far away from their true locations (in light blue). Table 4 lists the accuracy levels of the GPS pings. We can see that this kind of GPS noise is captured by the accuracy levels. The first few GPS pings have much higher accuracy levels than the last few GPS pings. By taking into account of accuracy levels, we successfully match the GPS pings onto the road network (in fuchsia).

Figure 10b shows a high quality GPS track where all the GPS pings have an accuracy level equal to 5 meters. There are a few roads near the GPS pings, but with an accuracy level as small as 5 meters, the map matching algorithm confidently eliminates segments that are slightly further, *e.g.*, the Jalan Jati Waringin road and the Jalan Tol Bekasi-Cawang-Kampung Melayu road. As a positive benefit, in this case the running time is at least halved compared to a noisy track.

This experiment demonstrates the benefits of having and utilizing the accuracy levels in map matching. We encourage the research

Upper Cross Street
Eu Tong Sen Street
New Bridge Road
Chinatown
Yue Hwa
Upper Cros
Jalan Jati Waringin
Kali Malang
Jalan Tol Bekasi-Cawang-Kampung Melayu

(a) Noisy GPS Track. (Table 4)
High accuracy levels help enlarge search radius to match successfully.

(b) Good GPS Track.
Low accuracy levels help reduce search radius to reduce running time.

Figure 10: Accuracy Levels Help Improve Map Matching
Light blue: raw GPS track (Direction: from East to West)
Fuchsia: map matched result

community to further explore the rich attributes provided in this dataset to improve map matching algorithms.

4.3 On Traffic Detection and Forecast

In addition to the inference of a static digital map, the *Grab-Posisi* GPS dataset can also be used to perform real-time traffic forecasting, which is very important for congestion detection, flow control, route planning, and navigation [42, 44]. Some examples of the fundamental indicators that are mostly used to monitor the current status of traffic conditions include the average speed, volume, and density on each road segment. These variables can be computed based on drivers' GPS trajectories and used to predict the future traffic conditions. Accurate traffic forecasting has received increasing research attention in recent years, as it is a challenging and crucial component in the development of intelligent traffic systems.

Recently, Yu *et al.* proposed a novel Spatio-Temporal Graph Convolutional Network (STGCN) to predict traffic from GPS trajectories [42]. Meanwhile, a similar framework was presented by Zhao *et al.*, namely a temporal graph convolutional network (T-GCN) that combines a graph convolutional network with a gated recurrent unit to simultaneously learn the complex topological structures for spatial dependence and the dynamic traffic changes for temporal dependence. While traditionally, statistical models such as the autoregressive integrated moving average have been usually used for time-series analysis [35], such methods have been recently challenged by machine learning based techniques, where better prediction accuracy can be obtained. The development of such data-driven machine learning, especially deep learning techniques, rely heavily on the availability and quality of traffic datasets. In fact, traffic in Jakarta is notoriously bad and the city is ranked at the 7th place in TomTom's 2018 traffic index [5]. We believe that the release of the *Grab-Posisi* large-scale GPS trajectory dataset will be beneficial for the community, as it will provide a foundation for benchmarks for real-time traffic prediction and its open source nature will facilitate the comparison among different traffic forecasting techniques.

In addition, the dataset provides an approximate ground-truth of the free-flow speed of expressways in Singapore and Jakarta. Free-flow speeds are widely used as a road quality measure, and also as a default routing/traffic value when real-time data are too sparse or not available.

4.4 On Trajectory Completion and Next Location Prediction

The performance of the aforementioned applications usually relies on having dense trajectories. For example, the map matching accuracy of HMM-based methods will decrease significantly when being applied to sparse trajectories. However, due to the power restrictions and bandwidth limitations on mobile devices, it is likely that only sparse trajectories are collected to reduce cost. To solve this issue, a few efforts have been made to investigate trajectory completion, *i.e.*, given low-sampling-rate GPS trajectories, we would like to predict the points in-between and recover high-sampling-rate routes [24]. Li *et al.* presented a knowledge-based trajectory completion method [24], which completes the geometry of trajectories' historical data without knowing the information of the underlying road network. As the *Grab-Posisi* dataset has a dense coverage in the geospatial domain, it provides large historical trajectories for the estimation of traffic flows. And therefore, it can be used as a good evaluation dataset for trajectory completion problems.

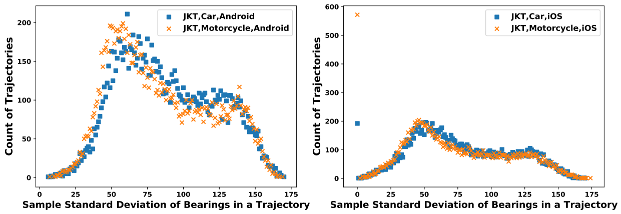
Related to trajectory completion, another location prediction problem is defined as, given a sequence of GPS locations, we would like to predict the location where the driver will go next. Existing work on next location prediction mostly focuses on check-in data from social networks [26, 28, 39]. Liu *et al.* proposed a novel spatio-temporal recurrent neural network to model the local temporal and spatial contexts to improve the accuracy of the next location prediction [26]. Yao *et al.* further presented a semantics-enriched recurrent model that jointly learns the embeddings of multiple factors including user, location, time, and keywords [39]. The *Grab-Posisi* dataset provides large-scale real-world GPS trajectories of people for fine-grained next location prediction. It would be interesting to evaluate the existing methods, and develop new algorithms based on GPS trajectories in addition to people's check-in records at social networks.

4.5 On Mode Detection

Transportation mode detection refers to the task of identifying the travel mode of a user (some examples of transportation mode include walk, bike, car, bus, *etc.*), which is a fundamental yet challenging GPS trajectory classification problem [37]. In recent decades, research on transportation mode detection has been widely discussed, with a considerable number of approaches having been proposed. For example, Xiao *et al.* introduced an approach based on ensemble learning to detect the transportation mode using GPS data only [36]. Hand-crafted features such as speed, acceleration, and turn angle were estimated based on GPS trajectories, which were used to perform the mode classification. Dabiri and Heaslip presented a deep learning based approach, which utilizes convolutional neural networks to automatically extract high-level features from the raw GPS inputs to improve the precision accuracy [17].

However, most of the existing work has developed their algorithms based on only one or two types of sensor data, where GPS and accelerometer values are the top two widely used data types. Walk, car, and bus are the three transportation modes that have been studied the most in previous work, while only a few efforts have been made on the investigation of less popular modes such as motorcycle and subway. The *Grab-Posisi* dataset will be beneficial for the research community in the following two aspects. First, the GPS trajectories are associated with rich attributes including GPS accuracy, bearing, and speed in addition to the latitude and longitude of geo-coordinates. Second, our dataset contains trajectories of both car and motorcycle. We also provide the device type (iOS or Android) that is used for data collection. It would be interesting to study the impact of the different sensor and device types on the performance of transportation detection.

As an example, we illustrate different bearing patterns in Figure 11. Figure 11a compares the distribution of sample standard deviation of bearings in trajectories between Car and Motorcycle from Android devices in Jakarta. It is clear that bearing could be a good indicator to classify driving mode. Figure 11b observes similar effects in iOS devices. Interestingly, the iOS category sees a significant number of trajectories with a small (< 5) sample standard deviation of the bearings. This finding should caution the usage of bearings if the device information is not provided. Bearings could be explored further for better mode classification.



(a) Android **(b) iOS**

Figure 11: Bearing Pattern Differs by Driving Mode

4.6 Economics Perspective

The real-world GPS trajectories of people reveal realistic travel patterns and demands, which can be of great help for city planning. Recently, Bao *et al.* proposed a data-driven approach for bike lane planning based on large-scale real-world bike trajectories [13]. As there are some realistic constraints faced by governments such as budget limitations and construction inconvenience, it is important to incorporate both the planning authorities' requirements and the realistic travel demands mined from trajectories for intelligent city planning. For example, the trajectories of cars can provide suggestions how to schedule highway constructions. The trajectories of motorcycles can help the government to choose the optimal locations to construct motorcycle lanes for safety concerns.

5 CONCLUSIONS AND FUTURE WORK

This paper presents the first GPS trajectory dataset of Southeast Asia from both developed countries (Singapore) and developing countries (Jakarta, Indonesia), covering more than 1 million km. The data have been collected recently with a sampling rate as high as 1 second. It also has richer contextual information, including the accuracy level, bearing and speed that can be leveraged to improve existing technologies. The trajectories are also labeled by data acquisition source (Android or iOS phones) and driving mode (Car or Motorcycle), which makes a valuable dataset for supervised mode classification. The dataset is of great value and a significant resource for the community for benchmarking and revisiting existing technologies.

In future, we plan to add more labels to the dataset. For instance, label the ground truth of map matching results.

REFERENCES

[1] 2008. Geohash. http://en.wikipedia.org/wiki/Geohash

[2] 2010. OpenStreetMap Public GPS traces. https://www.openstreetmap.org/traces

[3] 2012. GrabTaxi Holdings. https://assets.grab.com/wp-content/uploads/media/Grab-Corporate-Profile.pdf

[4] 2018. BEIJING TAXI TRIP DATA (IEEE DataPort). https://ieee-dataport.org/documents/beijing-taxi-trip-datasample

[5] 2018. TomTom Traffic Index. https://www.tomtom.com/en_gb/traffic-index/ranking

[6] 2019. Adaptive Thresholding. https://docs.opencv.org/3.4.0/d7/d4d/tutorial_py_thresholding.html

[7] 2019. Android Developer Document. https://developer.android.com/reference/android/location/Location.html

[8] 2019. Data source: DiDi Chuxing GAIA Open Dataset Initiative. https://outreach.didichuxing.com/research/opendata/en/

[9] 2019. iOS Developer Document. https://developer.apple.com/documentation/corelocation/cllocation

[10] Mahmuda Ahmed, Sophia Karagiorgou, Dieter Pfoser, and Carola Wenk. 2015. A Comparison and Evaluation of Map Construction Algorithms using Vehicle Tracking Data. *Geoinformatica* (2015), 601âĂŞ632.

[11] Mahmuda Ahmed, Sophia Karagiorgou, Dieter Pfoser, and Carola Wenk. 2015. A comparison and evaluation of map construction algorithms using vehicle tracking data. *GeoInformatica* 19, 3 (2015), 601–632.

[12] Rajesh Krishna Balan, Khoa Xuan Nguyen, and Lingxiao Jiang. 2011. Real-time trip information service for a large taxi fleet. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*. ACM, 99–112.

[13] Jie Bao, Tianfu He, Sijie Ruan, Yanhua Li, and Yu Zheng. 2017. Planning Bike Lanes Based on Sharing-Bikes' Trajectories. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1377–1386.

[14] James Biagioni and Jakob Eriksson. 2012. Map Inference in the Face of Noise and Disparity. In *International Conference on Advances in Geographic Information Systems*. 79–88.

[15] Adel Bolbol, Tao Cheng, Ioannis Tsapakis, and James Haworth. 2012. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems* 36, 6 (2012), 526 – 537.

https://doi.org/10.1016/j.compenvurbsys.2012.06.001 Special Issue: Advances in Geocomputation.

[16] Chen Chen, Cewu Lu, Qixing Huang, Qiang Yang, Dimitrios Gunopulos, and Leonidas Guibas. 2016. City-Scale Map Creation and Updating Using GPS Collections. In *International Conference on Knowledge Discovery and Data Mining*. 1465–1474.

[17] Sina Dabiri and Kevin Heaslip. 2018. Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transportation research part C: emerging technologies* 86 (2018), 360–371.

[18] J. J. Davies, A. R. Beresford, and A. Hopper. 2006. Scalable, Distributed, Real-Time Map Generation. *IEEE Pervasive Computing* 5, 4 (Oct 2006), 47–54. https://doi.org/10.1109/MPRV.2006.83

[19] Guande Qi, Xiaolong Li, Shijian Li, Gang Pan, Zonghui Wang, and Daqing Zhang. 2011. Measuring social functions of city regions from large-scale taxi behaviors. In *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*. 384–388. https://doi.org/10.1109/PERCOMW.2011.5766912

[20] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P. J. Nordlund. 2002. Particle Filters for Positioning, Navigation, and Tracking. *IEEE Transactions on Signal Processing* (2002), 425–437.

[21] Florian Hoppe, Sebastien Lamy, and Alessandro Cannarsi. 2016. Can Southeast Asia live up to its e-commerce potential. *Bain & Company* 16 (2016).

[22] Weiwei Jiang, Jing Lian, Max Shen, and Lin Zhang. 2017. A multi-period analysis of taxi drivers' behaviors based on GPS trajectories. In *20th IEEE International Conference on Intelligent Transportation Systems, ITSC 2017, Yokohama, Japan, October 16-19, 2017*. 1–6. https://doi.org/10.1109/ITSC.2017.8317622

[23] Muhammad Prasetya Kurniawan, Vanee Chonhenchob, Sher Paul Singh, and Sukasem Sittipod. [n. d.]. Measurement and Analysis of Vibration Levels in Two and Three Wheel Delivery Vehicles in Southeast Asia. *Packaging Technology and Science* 28, 9 ([n. d.]), 836–850. https://doi.org/10.1002/pts.2143 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/pts.2143

[24] Yang Li, Yangyan Li, Dimitrios Gunopulos, and Leonidas Guibas. 2016. Knowledge-based Trajectory Completion from Sparse GPS Samples. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPACIAL '16)*. ACM, New York, NY, USA, Article 33, 10 pages. https://doi.org/10.1145/2996913.2996924

[25] Jing Lian and Lin Zhang. 2018. One-month Beijing Taxi GPS Trajectory Dataset with Taxi IDs and Vehicle Status. In *Proceedings of the First Workshop on Data Acquisition To Analysis (DATA '18)*. ACM, New York, NY, USA, 3–4. https://doi.org/10.1145/3277868.3277870

[26] Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2016. Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 194–200.

[27] Radu Mariescu-Istodor and Pasi Fränti. 2018. CellNet: Inferring Road Networks from GPS Trajectories. *ACM Transactions on Spatial Algorithms and Systems* (2018), 8:1–8:22.

[28] Wesley Mathew, Ruben Raposo, and Bruno Martins. 2012. Predicting Future Locations with Hidden Markov Models. In *ACM Conference on Ubiquitous Computing*. 911–918.

[29] Lara Montini, Sebastian Prost, Johann Schrammel, Nadine Rieser-SchÃijssler, and Kay W. Axhausen. 2015. Comparison of Travel Diaries Generated from Smartphone Data and Dedicated GPS Devices. *Transportation Research Procedia* 11 (2015), 227 – 241. https://doi.org/10.1016/j.trpro.2015.12.020 Transport Survey Methods: Embracing Behavioural and Technological Changes Selected contributions from the 10th International Conference on Transport Survey Methods 16-21 November 2014, Leura, Australia.

[30] Paul Newson and John Krumm. 2009. Hidden Markov Map Matching Through Noise and Sparseness. In *ACM SIGSPATIAL*. 336–343.

[31] Mohammed A Quddus, Robert B Noland, and Washington Y Ochieng. 2006. A High Accuracy Fuzzy Logic based Map Matching Algorithm for Road Transport. *Journal of Intelligent Transportation Systems* (2006), 103–115.

[32] Nadine Schuessler and Kay W. Axhausen. 2009. Processing Raw Data from Global Positioning Systems without Additional Information. *Transportation Research Record* 2105, 1 (2009), 28–36. https://doi.org/10.3141/2105-04 arXiv:https://doi.org/10.3141/2105-04

[33] Zhangqing Shan, Hao Wu, Weiwei Sun, and Baihua Zheng. 2015. COBWEB: A Robust Map Update System Using GPS Trajectories. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 927–937.

[34] Guanfeng Wang and Roger Zimmermann. 2014. Eddy: An Error-bounded Delay-bounded Real-time Map Matching Algorithm Using HMM and Online Viterbi Decoder. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 33–42.

[35] Billy M Williams and Lester A Hoel. 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of transportation engineering* 129, 6 (2003), 664–672.

[36] Zhibin Xiao, Yang Wang, Kun Fu, and Fan Wu. 2017. Identifying different transportation modes from trajectory data using tree-based ensemble classifiers. *ISPRS International Journal of Geo-Information* 6, 2 (2017), 57.

[37] Xue Yang, Kathleen Stewart, Luliang Tang, Zhong Xie, and Qingquan Li. 2018. A review of GPS trajectories classification based on transportation mode. *Sensors* 18, 11 (2018), 3741.

[38] Yang Yang, Zhenzhou Yuan, Xin Fu, Yinhai Wang, and Dongye Sun. 2019. Optimization Model of Taxi Fleet Size Based on GPS Tracking Data. *Sustainability* 11, 3 (January 2019), 1–19. https://ideas.repec.org/a/gam/jsusta/v11y2019i3p731-d202077.html

[39] Di Yao, Chao Zhang, Jianhui Huang, and Jingping Bi. 2017. SERM: A Recurrent Model for Next Location Prediction in Semantic Trajectories. In *ACM International Conference on Information and Knowledge Management*. 2411–2414.

[40] Yifang Yin, Beomjoo Seo, and Roger Zimmermann. 2015. Content vs. Context: Visual and Geographic Information Use in Video Landmark Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 11, 3 (2015), 39:1–39:21.

[41] 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*. 7:1–7:10.

[42] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. 3634–3640.

[43] Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang. 2010. T-Drive: Driving Directions Based on Taxi Trajectories. ACM SIGSPATIAL GIS 2010. https://www.microsoft.com/en-us/research/publication/t-drive-driving-directions-based-on-taxi-trajectories/ Best Paper Award.

[44] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. 2018. T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction. *CoRR* (2018). https://arxiv.org/abs/1811.05320

[45] Yu Zheng. 2010. GPS Trajectories with transportation mode labels. https://www.microsoft.com/en-us/research/publication/gps-trajectories-with-transportation-mode-labels/ GeoLife project-Microsoft Research Asia.

[46] Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. 2008. Learning Transportation Mode from Raw Gps Data for Geographic Applications on the Web. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 247–256. https://doi.org/10.1145/1367497.1367532

[47] Yu Zheng, Xing Xie, and Wei-Ying Ma. 2008. Understanding Mobility Based on GPS Data. https://www.microsoft.com/en-us/research/publication/understanding-mobility-based-on-gps-data/