Challenges and Opportunities in Exploiting Large-Scale GPS Probe Data

Yin Wang, Yanmin Zhu, Zhaocheng He, Yang Yue, Qingquan Li

HP Laboratories
HPL-2011-109

Keyword(s):

Transportation; GPS; probe; data analysis; map matching

Abstract:

Global positioning system (GPS) receivers are widely deployed in navigation and tracking devices installed in taxis, buses, utility vehicles, and smart phones. These GPS receivers can provide probe data that illuminates automotive traffic conditions for applications including traffic analysis, travel time estimation, map building, and congestion and accident detection. The characteristics of real GPS probe data are not well understood, however, and these characteristics may pose difficulties for existing algorithms. For example, map matching is often a prerequisite for GPS trace-related applications. Existing solutions often rely on assumptions, e.g., Gaussian GPS noise, that are not necessarily true of data collected on metropolitan streets. In this paper, we study probe data from tens of thousands of taxis in the three largest cities in China. We comprehensively characterize the data, quantify noise present in it, gauge the applicability of existing map matching algorithms to the data, identify challenges in processing large-scale GPS data, and suggest research opportunities.

External Posting Date: July 21, 2011 [Fulltext] Approved for External Publication
Internal Posting Date: July 21, 2011 [Fulltext]

Challenges and Opportunities in Exploiting Large-Scale GPS Probe Data

Yin Wang
Hewlett-Packard Labs
Palo Alto, CA
yin.wang@hp.com

Yanmin Zhu
Shanghai Jiao Tong University
Shanghai, China
yzhu@cs.sjtu.edu.cn

Zhaocheng He
Sun Yat-Sen University
Guangzhou, China
hezhch@mail.sysu.edu

Yang Yue
Wuhan University
Wuhan, China
yueyang@whu.edu.cn

Qingquan Li
Wuhan University
Wuhan, China
qqli@whu.edu.cn

ABSTRACT

Global positioning system (GPS) receivers are widely deployed in navigation and tracking devices installed in taxis, buses, utility vehicles, and smart phones. These GPS receivers can provide probe data that illuminates automotive traffic conditions for applications including traffic analysis, travel time estimation, map building, and congestion and accident detection. The characteristics of real GPS probe data are not well understood, however, and these characteristics may pose difficulties for existing algorithms. For example, map matching is often a prerequisite for GPS trace-related applications. Existing solutions often rely on assumptions, e.g., Gaussian GPS noise, that are not necessarily true of data collected on metropolitan streets. In this paper, we study probe data from tens of thousands of taxis in the three largest cities in China. We comprehensively characterize the data, quantify noise present in it, gauge the applicability of existing map matching algorithms to the data, identify challenges in processing large-scale GPS data, and suggest research opportunities.

1. INTRODUCTION

GPS probes are ubiquitous. For example, smart phones integrate GPS navigation systems and can transmit location data through applications [1, 10]. Fleet management systems using GPS tracking devices are widely deployed in taxis, buses, utility and commercial vehicles [4, 13]. As such, there is considerable interest in the application of the GPS data, e.g., traffic analysis [4, 20], traffic forecasting [11], hotspot analysis [14], real-time trip planning [12], travel time estimation [5], and map building [7, 8].

However, the characteristics of large-scale GPS data are not well understood, which may limit the applicability of existing solutions if the latter rely on assumptions that do not hold in practice. For example, map matching GPS coordinates to a passable route is often the first step in traffic-related applications. Most existing algorithms are designed for GPS navigation devices with relatively frequent sampling rates and high precision [18]. GPS probes, on the other hand, may sample infrequently (e.g., once per minute) to reduce communication costs [15]. Similarly, noise (position error/uncertainty) varies with the device. Recently, Hidden Markov Model (HMM) approaches to map matching noisy, low-sample-rate GPS data have proven effective [16, 17]. HMM approaches require knowledge of the probability distribution of GPS measurement errors, which are typically assumed to follow Gaussian distributions. But the real noise statistics in large-scale probe data remain unknown. Even if the installed GPS tracking unit has been properly studied, with thousands of them deployed, inconsistent behaviors are likely to occur. These unknown factors could significantly affect the accuracy of map matching algorithms, which in turn affect other GPS trace-based applications.

Existing empirical studies of GPS probes often focus on feasibility for traffic sensing and do not thoroughly characterize the probe data employed. Furthermore they are often limited in scale. For example, GPS-equipped mobile phones have been proven effective for traffic sensing, based on 100 devices from a 10-mile stretch of a freeway for eight hours [10]. A similar study examined bus GPS data from a 2.5-mile corridor for two days [3]. A study on taxi GPS data emphasized communication reliability [13].

In this paper, we analyze taxi GPS data from the three largest cities in China: Beijing, Shanghai, and Guangzhou. We collected the data from all 28 thousand taxis in the Beijing database for one week in 2008, and sampled nearly 5 thousand taxis each from the Shanghai and Guangzhou databases, for one week in 2006 and 2011, respectively. The taxi fleets in these cities are some of the largest in the world, and these data sets likely cover most common issues with taxi GPS probes.

In China, taxi companies install GPS devices initially for the purpose of phone call dispatching and theft protection. These devices communicate the location, speed, direction, occupancy, and other status information to the central server through cellular phone networks. The data is typically stored in relational databases. Lately, local governments collect and aggregate the data from these companies for vehicle and traffic regulation, e.g., speed violation detection. While there is considerable interest in using the data for applications such as traffic estimation and travel time prediction, we are not aware of any mature system that produces reliable and credible results in the metropolitan scale.

Although large in scale, these taxi GPS devices often provide low quality data compared to probes used in research experiments. Taxi dispatching does not require high precision, and cost efficiency is often the primary concern. For example, the GPS tracking unit is sometimes integrated into LED advertisement boards at no additional cost. The qual-

City	ID(key)	taxi ID	GPS time	svr time	latitude	longitude	speed	direction	status	effective
Beijing		cell phone	Unix time		integer	integer	integer	integer	bytes	binary
Shanghai	integer	integer	text		double	double	byte	byte	binary	
Guangzhou	integer	license plate	text	text	xx.xxxxx	xxx.xxxxx	xxx	xxx	bytes	binary

Table 1: GPS record table fields from different cities

ity of both the device and the installation can affect the quality of measured position data. Heterogeneous devices used by different companies or installed at different times pose another challenge. Our analysis of the data reveals issues that are usually ignored by existing solutions.

To the best of our knowledge, we are the first to systematically characterize large-scale GPS probe data. The fleet size, geographical range, and time span are unprecedented. We thoroughly study measurements available in our three data sets and analyze location noise in detail. We describe limitations in existing algorithms for map matching and other analyses, and suggest promising directions for the development of scalable GPS data processing systems and practical algorithms for valuable applications.

The rest of this paper is organized as follows. Section 2 presents the statistics and features of various measurements in our GPS records. Section 3 analyzes GPS noise distribution and the root cause of various outliers. Section 4 discusses the challenges of map matching using real GPS probe data and proposes research directions. Section 5 concludes.

2. DATA CHARACTERISTICS

The type of data reported by the GPS probes is largely consistent, even though there are different data formats and sampling characteristics. This section introduces the formats and studies the statistics of all measurements. We abbreviate the three city names as BJ, SH, and GZ from now on for the sake of succinct presentation.

2.1 Data Sources and Formats

We obtained our taxi GPS data from local transportation bureaus. The data from BJ is a complete database dump for the time period from midnight 12/1/2008 to midnight 12/8/2008, which contains 80,991,481 GPS records from 28,002 taxis. Previous work using BJ data includes [19] (a different time period). In SH, our data set is a partial dump from the database, including 4,403 taxis during the period of 10/6 – 10/13/2006, a total of 40,360,033 records. The total number of taxis in the database is unknown. Previous work using the data includes [14, 20]. The database in GZ contains 17,226 taxis as of 04/01/2011. We picked the first 5k taxis based on the alphabetic order of licence plates, and retrieved samples for the time period of 5/16 – 5/23/2011. Only 4,941 taxis transmitted data during the period, with a total of 55,578,763 records. In the raw text format, the file sizes of the data sets are roughly 9.0GB, 2.5GB, and 5.5GB, respectively.

Currently there is no nationwide industrial standard in China for the GPS record format. The province of Guangdong has established and enforced a provincial standard in early 2009, which results in much better data quality as our analysis proves. However, as the data sets from BJ and SH are obtained in 2008 and 2006, respectively, the current data formats and quality can be different. All cities show clear signs of heterogeneous devices and configurations, as a result of data aggregation from different taxi companies.

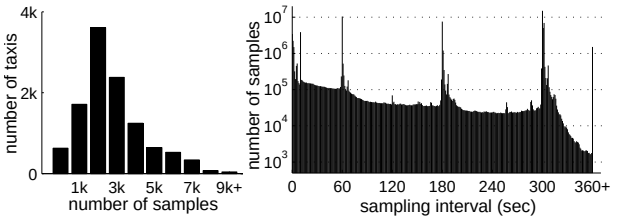
taxi group	count	sampling interval vacant (sec)	sampling interval occupied (sec)	granularity lat/lon (degree)	granularity speed (km/h)	granularity direction (degree)
BJ-A	10597	300-301	300-301	1/100,000	2	3
BJ-B	7,885	301-302	301-302	1/100,000	1	10
BJ-C	3,569	180-181	180-181	1/100,000	2	3
BJ-D	2,670	60-61	60-61	1/100,000	2	3
SH-A	1,838	63-64	63-64	1/60,000	1	2
SH-B	1,773	16	61	1/1,0000	2	45
GZ-A	3,507	20	60	1/100,000	1-2	1
GZ-B	1,244	30-31	30-31	1/100,000	3-4	10

Table 2: Summary of groups with more than 1,000 taxis

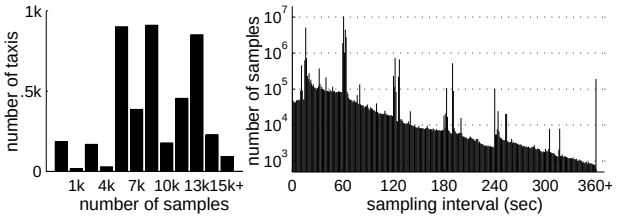
Table 1 shows all fields of GPS records. Overall, BJ mostly uses integer numbers, SH uses various primitive data types, and GZ favors text strings. The record id is the incremental key of the table. The record table in BJ does not use a key. Different databases use different taxi IDs. BJ uses cell phone numbers, indicating the communication through cellular phone networks. SH uses an integer id. These integers are probably assigned to each taxi company as taxis with adjacent ids exhibit similar data characteristics. GZ uses licence plates directly. All databases contain a timestamp for each record. BJ uses an integer that represents Unix time in seconds, and the other two cities use text strings. GZ includes the server timestamp upon message receival, which helps us analyze transmission issues. Latitude and longitude are stored as integers in BJ database, where the last five digits actually represent a fraction of a degree. SH uses doubles to represent decimal degrees for latitude and longitude. GZ stores decimal degrees as text strings. The speed data is also different. BJ uses an integer that represents centimeter per second. SH stores the km/h speed as an unsigned byte, i.e., 255 km/h maximum. GZ uses a string with three characters, i.e., 999 km/h maximum. For the vehicle heading direction, BJ again stores an integer ranging from 0 to 360 degrees. SH uses an unsigned byte ranging from 0 to 180, and the degree is the value multiplied by two. GZ uses a string of length three for the degree value. Status field includes various flags from the vehicle. The most relevant one is the occupancy of the taxi. SH database contains only this flag. The effectiveness field reflects the belief of the GPS device whether the location is accurate, typically judged by the number and strength of satellite signals.

2.2 Sampling Interval Analysis

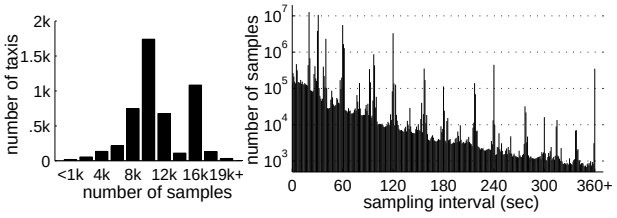
GPS devices in taxis take samples by time (temporally) or distance (spatially) traveled, and the sampling rate can be different depending on if the status is vacant or occupied [13]. Even with spatial sampling, there is often a timeout which causes an update in case of slow movement or waiting. In addition, taxis often take samples during status changes, e.g., occupancy change, and location reestablishment after a signal loss. Nevertheless, we found the sampling methods and sampling rates easy to identify. Temporal sampling usually exhibits one or two intervals with at least an order of magnitude more samples than the rest.



(a) Beijing



(b) Shanghai



(c) Guangzhou

Figure 1: Histogram of sample counts and sampling intervals

Spatial sampling has intervals relatively evenly distributed, with slightly more samples at the timeout interval. Most taxis in our data sets use temporal sampling.

The left column of Fig. 1 shows the linear-scale taxi histograms by total sample counts, and the right column shows log-scale sample histograms by sampling intervals. In BJ, most taxis have only 1-4k samples, which contribute to the spike at 5 min sampling interval. There are also many samples at the intervals of one and three minutes. The corresponding taxis are significantly less in number as lower sampling intervals produce more samples. For example, 70 taxis alone contribute to the spike of 3.75 million samples at the 10 sec sampling interval. In SH, the dominating sampling intervals are 16 sec and 61-64 sec. Manual inspection revealed two groups of taxis, one samples at 63-64 sec interval with 6-8k samples, the other samples at 16 sec when vacant and 61 sec when occupied, with 11-14k samples. There are also two dominating groups in GZ, one samples at 20 sec when vacant and 60 sec when occupied, and the other samples at 30-31 sec. In the sample histograms of both SH and GZ, there are spikes at multiples of these main sampling intervals. Missing samples likely resulted in these spikes, analyzed in the next subsection.

We use a 10% threshold to discover sampling intervals for each taxi, i.e., any interval with more than 10% of the total samples. As a result, each taxi almost always returns no more than two intervals and the total sample percentage of these significant intervals is more than 50% in most cases, indicating good accuracy. Table 2 summaries taxi groups we discovered using this threshold. The different measurement granularity of different groups also confirms our classification. The latitudes and longitudes in BJ and GZ include five digits for the fractional part. While SH-A includes six digits, the last two digits have only six values, representing multiples of 1/60,000 (it is not converted from minute/second representation as numbers differ by exactly 1/60,000 degree exist from one taxi). The granularity of speed and direction is discussed in Sections 2.4 and 2.5.

	Beijing	Shanghai	Guangzhou
duplicate samples	4.00%(24,385)	2.50%(4,228)	0.19%(4,905)
missing samples	0.60%	7.36%	11.7%
reverse order	–	0.30%(4,184)	0.82%(4,888)
GPS time error	–	–	0.028%(1,341)

Table 3: Transmission error (percentage and taxis affected)

```
record id   server time      GPS time         longitude latitude
31728500429,5/20/11 21:26:32,5/20/11 21:18:08,113.23598,23.14391
31728512858,5/20/11 21:27:04,5/20/11 21:18:39,113.23618,23.14401
31709065161,5/20/11 5:27:22 ,5/20/11 21:19:00,113.26282,23.15134
31728524473,5/20/11 21:27:32,5/20/11 21:19:10,113.23941,23.14638
31728548684,5/20/11 21:28:35,5/20/11 21:20:11,113.24556,23.14984
```

Figure 2: GPS timestamp error?

2.3 Transmission Time and Error

Table 3 summaries our analysis of transmission issues. First we discover a significant percentage of duplicate records in the table, with BJ being the worst and GZ the cleanest. As these duplicates cover almost all taxis, network retransmission is likely the root cause.

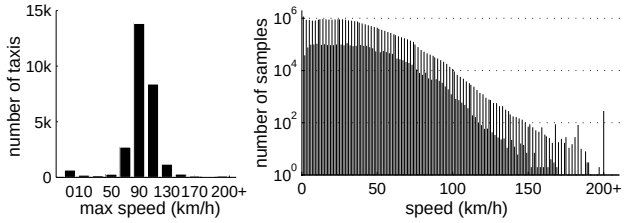
Our calculation of missing samples is approximate as there is no device or network log. We use only taxis in Table 2 for the calculation, as their sampling intervals are more prominent than the rest of taxis. Let n_i be the number of samples at interval i. Assuming t is the sampling interval, the missing rate is defined as

$$n_{2t}/(n_t + 2 \times n_{2t}) \quad (1)$$

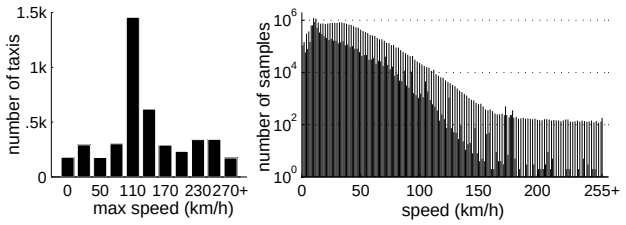
Samples at higher multiples of t are not considered as the count is typically an order of magnitude less. For taxi groups with two intervals for vacant and occupied status, only the smaller one is counted. As the missing rates of different taxi groups are close within the same city, we calculate the average rate in Table 3. The result reverses the rank of duplicate samples. We suspect that the network in BJ retransmits more aggressively while the network in GZ drops more packages.

Reverse order in the third row means that the package arrival order is different from the GPS timestamp order. Only GZ database includes the server time for this calculation. However, as the record id key monotonically increases over time, we use it to determine reverse-ordered samples in SH. Only records from the same taxi are compared and counted. Reverse-ordered samples also affect almost all taxis in both cities, indicating network transmission delays.

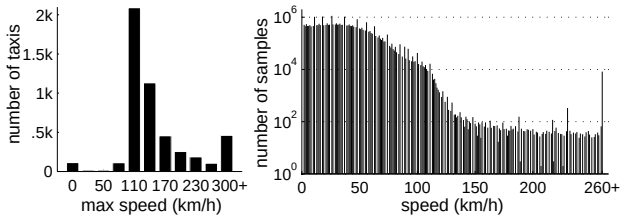
Finally, we analyze the exact transmission delay using the server timestamp available in GZ database. The time difference between the server and the GPS timestamps clearly follows heavy-tail distribution, and is very consistent among all taxis. There are 4,120 taxis that start the distribution at 493 seconds, and another 596 taxis start between 483-503 seconds. This suggests that these GPS devices synchronize their time by satellites, and the server time is 493 seconds faster. The median time of delivery is around 8 seconds, and the 90 percentile is at 15 seconds. The upper bound of the delivery time is around 500 seconds. However, there are



(a) Beijing

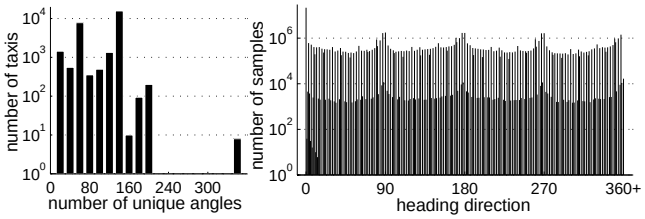


(b) Shanghai

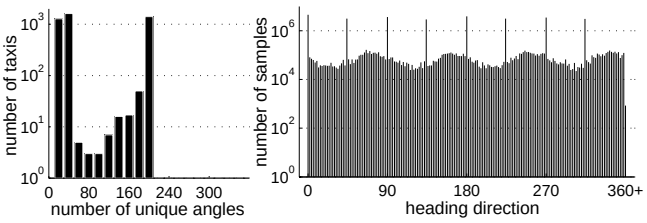


(c) Guangzhou

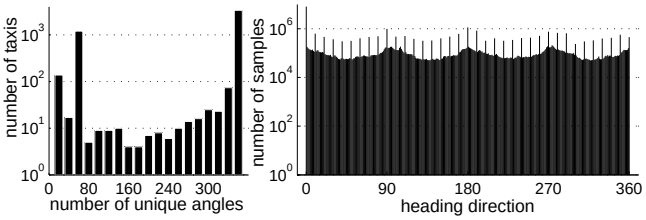
Figure 3: Vehicle measured speed



(a) Beijing



(b) Shanghai



(c) Guangzhou

Figure 4: Vehicle measured direction

very strange outliers that are orders of magnitude distant from the center of distribution. Fig. 2 shows one example. The large time difference in the third record is due to device malfunctioning rather than the server error, because both the sampling interval and coordinates are inconsistent with adjacent samples. Based on the features of the data, we apply an outlier detection algorithm that finds gaps at both ends of the center distribution region. The gap must be at least one order of magnitude larger than the length of the center region, and samples beyond these gaps are outliers. Table 3 shows that 1,341 taxis have outliers that account for 0.027% of all samples from all taxis.

2.4 Measured Speed

Speed information can be obtained from either the odometer or the GPS calculation. We suspect that most taxis use GPS calculation for the convenience of installation. Our analysis supports this observation as excessive speed numbers exist, and in many cases accompanied by abnormal yet consistent GPS coordinates.

Fig. 3 shows the characteristics of vehicle measured speed. The histograms on the left count the number of taxis by their maximally reported speed numbers. The histograms on the right count all samples at each speed value. Taxis in BJ behave relatively similarly. Most taxis report speed within 130 km/h. Although the maximum speed reported is 1,741 km/h, there is a gap at 185 km/h and only 301 samples are beyond it. Among these sample, 181 has the effective flag set, which is slightly above the average effective sample rate. Therefore there is no correlation between GPS location effectiveness and abnormal speed measurement. There are another 49 samples of negative speed not shown on the histogram. The histogram exhibits two curves at the boundary. Taxi groups BJ-A, BJ-C, and BJ-D in Table 2 contribute to the outer curve as their speed measurement is accurate to 2 km/h only, i.e., reporting only even numbers. Group BJ-B contributes to the inner curve with the accuracy at 1 km/h. In SH, the overall sample histogram has a similar shape with BJ's, but there are significantly more samples beyond 150 km/h, which do not fall off as the speed increases. The maximum speed reported is 254 km/h, enforced by the unsigned byte type. However, the frequency at 254 is on par with frequencies between 200-254. The histogram also contains two layers in SH. The out layer is from SH-B that reports only even numbers, and the inner layer is from SH-A that reports all numbers. Interestingly, most samples with excessive speed are from SH-B. In GZ, there is again a significant number of taxis with maximum speed beyond 180 km/h. The maximum number is 999 km/h, as allowed by the data format. The histogram drops two orders of magnitude at 120-130 km/h, then it remains relatively flat. If we consider only effective samples, there is another drop at 260 km/h. Only 98 out of 8,257 samples beyond 260 km/h are effective. Correspondingly, if we drew the taxi histogram on the left with effective samples only, the count at 300+ would reduce from 384 to only 18 taxis. The bars of other bins change very little. Finally, the sample histogram is also bimodal. GZ-A group skips a number after one or two consecutive numbers, and GZ-B group skips two to three numbers for every number that has samples. Together, speed numbers skipped by both have zero sample and the overlapped numbers exhibit spikes in the histogram.

2.5 Measured Heading Direction

The distribution of the direction measurement is also dif-

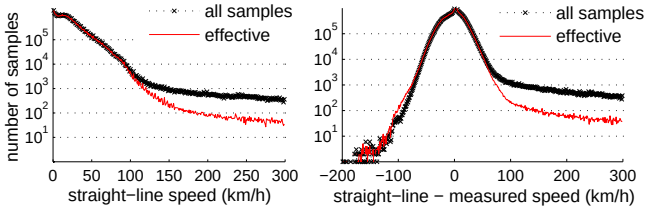
	Beijing	Shanghai	Guangzhou
occupied samples	13.0%	19.2%	28.7%
effective samples	60.4%	–	82.5%
speed 0 samples	65.8%	33.1%	49.3%
speed 0 (effective)	26.7%	–	34.2%
repeated location	45.5%	45.3%	27.8%
fixed location taxi	351	534	107

Table 4: Miscellaneous statistics

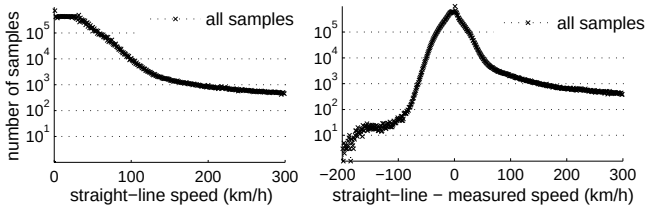
ferent among different taxi groups, displayed in Fig. 4. In this set of figures, the taxi histograms on the left also use log-scale to reveal details of low-count bins. BJ-B group reports only 36 unique degrees, and BJ-A, BJ-C, BJ-D groups have 120 unique values. So the granularity is at 10 and 3 degrees, respectively. These groups contribute to the outer layer in the sample histogram. The inner layer is from a small group of taxis that reports only even numbers, e.g., 180 unique counts. There are 13,905 samples below 0 or above 360. Most of these samples are from a few taxis that use a different scale, 0-36,000. Format standardization could have cleaned the data. SH-A reports only even numbers, as enforced by the data format, and SH-B reports mostly eight directions. There are 850 illegal samples beyond 360 degrees. Samples from GZ contain no illegal direction. GZ-A reports every degree number and GZ-B reports 36 unique degrees. The histograms of different taxi groups in different cities are relatively evenly distributed, with four peaks aligned with the directions of horizontal and vertical roads in each city. More specifically, as roads in BJ are well aligned with longitudinal and latitudinal lines, the peaks are very sharp at the four orientations. In SH, the roads are often aligned with Huangpu River, and therefore the peaks are not located at any major orientation. GZ is somewhat in between.

2.6 Other Statistics

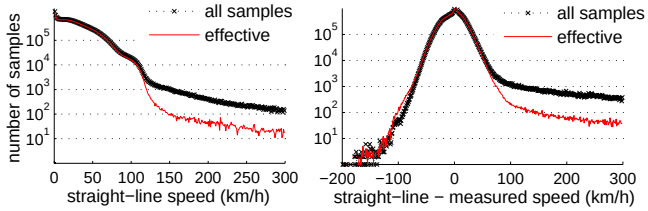
Table 4 summarizes a few more statistics relevant to the data. The occupied sample percentage shows the utilization of taxis. As we count samples rather than time, the actual time percentage can be higher due to lower sampling rate when occupied. Effective samples count those whose effective flags are set to true. Speed zero means samples reporting zero speed. BJ has more samples with speed zero than the other two cities but more than half of them are ineffective. In fact, 65.8-26.7=39.1% of all samples are both ineffective and zero speed, which is almost the same as the total number of ineffective samples, 100-60.4=39.6%. The number is also close in GZ, i.e., 49.3-34.2=15.1% vs. 100-82.5=17.5%. Therefore, taxis almost always report zero speed whenever the location is not confirmed. The measured direction in ineffective samples is mostly zero as well, as there are significantly more samples at zero degree in the histograms of BJ and GZ in Fig. 4. If we count the percentage of zero speed samples among effective samples only, BJ and GZ are close, 26.7/60.4=44.2% and 34.2/82.5=41.5%, respectively, indicating similar average waiting time. On the other hand, the speed and direction measurements are not zeroed out in SH even if the GPS device cannot confirm its location. There are significantly less samples at zero speed and zero degree. Furthermore, considering samples that repeat the previous location, BJ and SH have significantly more samples than GZ. The percentage in SH even exceeds the percentage of zero speed samples. We found that samples repeating the same location can report non-



(a) Beijing



(b) Shanghai



(c) Guangzhou

Figure 5: Straight-line (great-circle distance) speed

zero speed, indicating errors in the measurement. Finally, there are hundreds of taxis in each database that do not update their locations, with the percentage in SH significantly larger than the other two, 534/4403=12.1%.

3. GPS LOCATION NOISE ANALYSIS

Analyzing GPS measurement noise in raw samples without map matched traces is challenging. We develop the result through a series of studies.

3.1 Straight-Line Speed

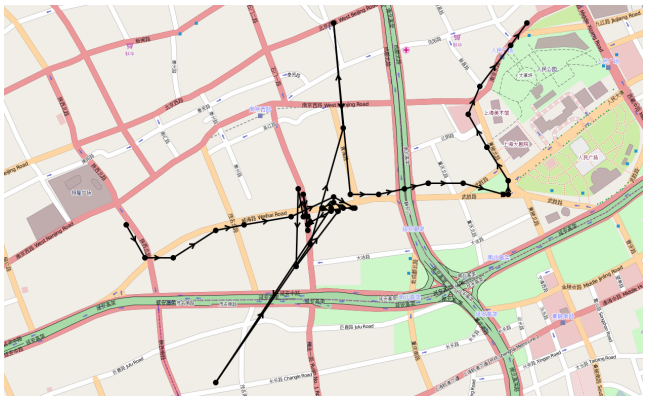
The reported speed mostly measures the instantaneous speed at the sampling time, as zero occurs even when the coordinates have changed significantly. Here we calculate the straight-line (great-circle distance) speed for comparison. Fig. 5 shows the straight-line speed on the left and the difference with the measured speed on the right. Both use log scale. For BJ and GZ where the effective flag is available, we add the curves that count only effective samples.

There are more excessive numbers in the calculated speed than the measured one. In BJ, while there are only 281 (167 effective) samples with measured speed above 200 km/h, the count with calculated speed is 504,605 (192,146 effective). In SH, the number is 4,064 vs. 882,836. In GZ, it is 10,003 (1,331) vs. 98,320 (27,001). The effectiveness flag filters out outliers significantly, but not all of them. Comparing with the measured speed, excessive calculated speed affects almost all taxis. Considering effective samples only, the numbers of taxis exceeding 200 km/h in straight-line speed are 14,353, 3,751, and 3,750 in BJ, SH, and GZ, respectively.

3.2 Root Causes for Abnormal Coordinates

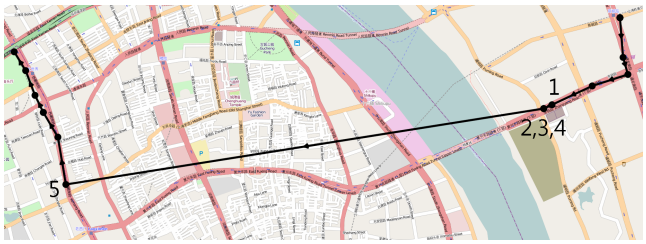
We have shown that both measured and straight-line speed can be excessively large, and the latter affects almost all

(a) A directed series of wrong coordinates while moving



(b) Random wrong coordinates while stopped in city canyon

```
   time     lon      lat     spd dir
1  06:33:08,121.5071,31.2275,66 ,112
2  06:34:10,121.5066,31.2273,0  ,112 // distance=52m
3  06:35:10,121.5066,31.2273,0  ,112
4  06:36:11,121.5066,31.2273,0  ,112
5  06:36:59,121.477 ,31.2233,54 ,157 // distance=2.85km
```



(c) Repeating last-known location in and after river tunnel

Figure 6: Root causes for error coordinates (Shanghai)

taxis. Next we analyze some of these samples to identify the root causes. Fig. 6 shows a few examples from SH. In Fig. 6a, there are 6 samples spanning 7 minutes that lead from the downtown area to the sea and back. The maximum measured speed is 254 km/h, while the straight-line speed exceeds 600 km/h. Based on samples before and after, as well as the speed and distance constraints, there are only a handful of choices for the real drive path. These paths do not have many high-rise buildings nearby, and the taxi was moving at 20km/h on average. It is unclear whether multipath signals resulted in the error. On the other hand, Fig. 6b is most certainly a multipath error inside a city canyon. The taxi was either waiting or slowly moving for 20 minutes in the center area under high-rise buildings, generating a total of 25 samples. These samples are located rather randomly,

(a) Notheast shift of the aerial image in Shanghai

(b) Wrong location of the bridge in Guangzhou

Figure 7: Aerial map misalignment (white dots are samples)

and the measured speed ranges from 0 to 88 km/h (unrealistic in downtown SH). The straight-line speed exceeds 180 km/h in one interval. Fig. 6c shows that tunnels can result in excessive speed too. While inside the river tunnel, the taxi kept reporting its last known location, i.e., the entrance of the tunnel, until it relocated itself after leaving the tunnel. The straight-line speed between samples 4 and 5 is beyond 200 km/h. The sample effectiveness flag, determined by signal strength, can filter out tunnels where there is no satellite signal at all. But multipath errors can pass the check as the signals bounced by obstacles are still received.

Multipath error is difficult to identify and filter out. While excessive measured speed is reported by a small percentage of taxis, as shown in Fig. 3, excessive straight-line speed is common among all taxis. The error is not highly correlated with location either. We divided the map as a grid and checked the percentage of samples with excessive straight-line speed. The result distribution is rather random. Areas of high-rise buildings or tunnels do not exhibit higher error percentage then their adjacent blocks. Finally, straight-line speed alone is not effective in identifying error samples. While extreme cases like those in Fig. 6 can be filtered out by a reasonable threshold, it is unclear how many mild errors exist in the "normal" data. Furthermore, different road types can be close to each other with completely different max speed thresholds. These location errors can confuse map matching algorithms, which in turn affect applications such as traffic estimation and event detection.

3.3 Map Accuracy

We use Open StreetMap (OSM) [2] for our map-related analysis. In China, there are two major data sources for the map. Truck roads are mostly donated by Automotive Navigation Data Inc. Street maps in major cities are often created manually from Bing vertical aerial imagery. In BJ and

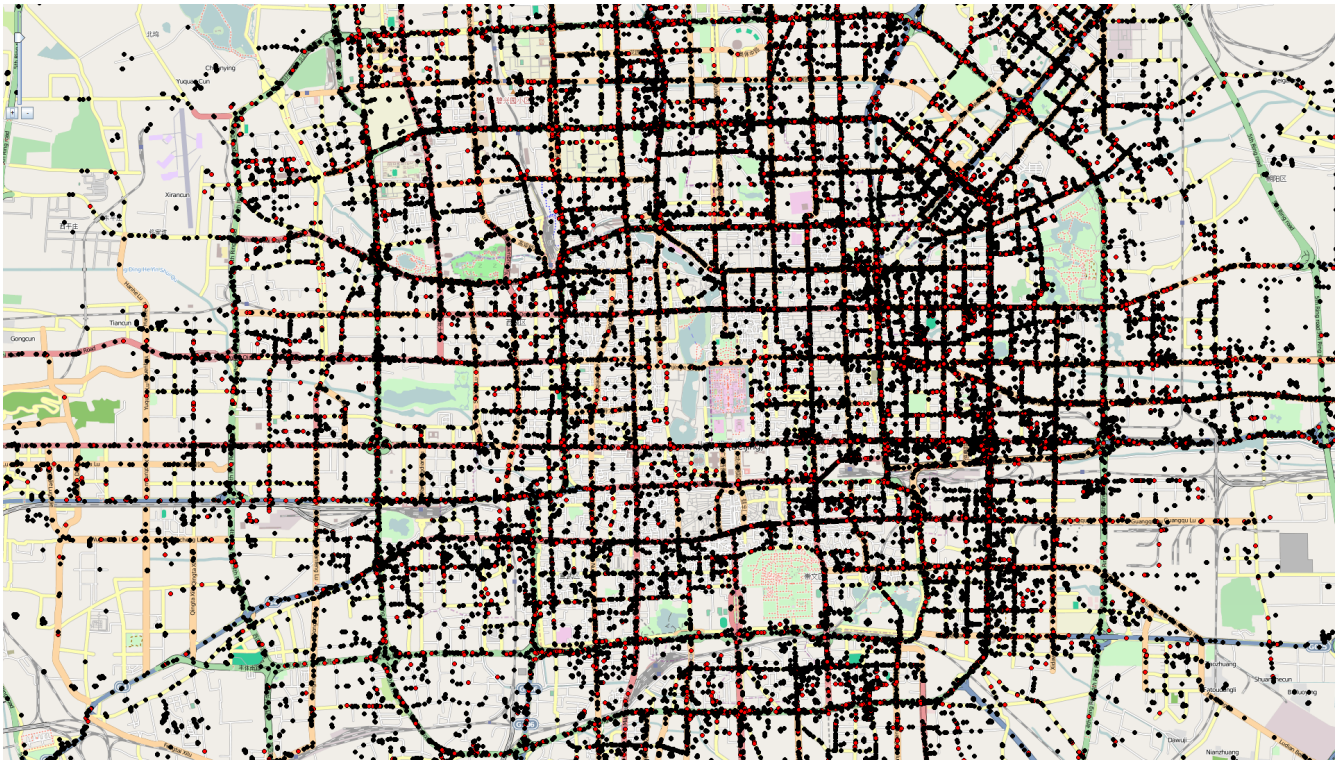


Figure 8: One hour of effective samples in Beijing, red and black dots represent occupied and empty samples, respectively

SH, street maps in downtown areas and tourist locations are mostly complete, while suburb areas contain mainly truck roads. The map of GZ lacks detail even in the downtown area. We use a proprietary map in ShapeFile format for the noise measurement in GZ. However, missing roads are still common in all three maps. In addition, as the GPS data from SH is five years old, there are roads in the current map without any sample coverage. In all three maps, a named road is divided into road segments by intersections, and each road segment is represented by a *polyline*. A polyline here is a sequence of vertices connected by straight lines. There are totally 13,466, 14,473, and 15,495 road segments in the maps of BJ, SH, and GZ, respectively. Some of the roads in OSM data have the number of lanes specified. Both directions of major roads are often marked separately as well. In the ShapeFile map for GZ, however, every road segment is marked by exactly one polyline. All these map features affect the accuracy of the noise measurement.

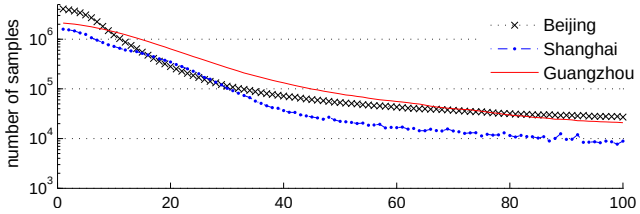
We discover that all three vector maps as well as the Bing aerial image backdrop of OSM do not always agree with GPS samples. Fig. 7 shows two examples from SH and GZ, respectively. White dots represent GPS samples, and there is a line between two consecutive dots if the sampling interval is within 31 seconds and the straight-line speed is less than 50 m/s. Fig. 7a uses three hours of data in SH, and Fig. 7b uses one hour of data from GZ. The Bing aerial image in SH is shifted northeast, while the bridge in GZ is located to the east of the traces. In SH, the OSM vector map is aligned with the aerial image, therefore disagreeing with the GPS traces. In GZ, both the OSM and the proprietary vector maps agree with the GPS traces instead of the aerial image. Furthermore, different locations in both cities show varying degrees of misalignment for both vector maps and areal images. We have not found any map misalignment in BJ. Fig. 8 shows one hour of effective samples.

The above observation suggests that more accurate maps can be derived from GPS traces than aerial imagery. The latter may suffer from alignment errors and distortions due to either camera or map projection. While geological surveys can produce highly accurate maps, it is more costly and less up to date. Finally, we notice that GPS trajectories in GZ are more concentrated than those in SH, as Fig. 7 illustrates (albeit a slightly different scale). This is partly because of the measurement granularity. The trajectories in SH are mostly from group SH-B because of its low sampling interval. Its location accuracy is limited to 1/1,000 of a degree, which is 8.5 meters along the longitude and 11.1 meters along the latitude in SH. Therefore, adjacent trajectories are more distant than those in GZ.

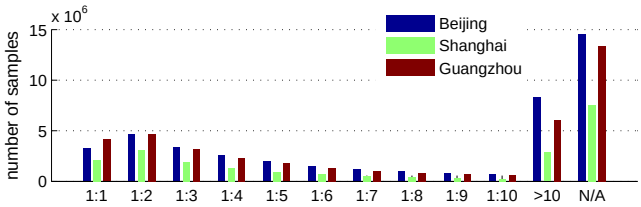
3.4 Gaussian Noise Measure

Statistically speaking, the various outliers in the measurements account for a small fraction of the data. Most samples are located near roads, as can be seen from Fig. 6-8 visually. Assuming Gaussian noise distribution of the GPS coordinates in these "normal" samples, we estimate the distribution parameters in this subsection.

Fig. 9 shows the statistics of the distance between samples and roads. Samples that repeat their previous locations are not counted. Fig. 9a is the log-scale histogram of the distance between each sample and its nearest road. Samples more than 100 meters away from any road account for 8.7%, 5.6%, and 6.8% of total samples in BJ, SH, and GZ, respectively. These isolated samples are mostly due to roads and parking areas not included in the maps. Among all samples within 100 meters to some mapped road, the 95 percentile



(a) great-circle distance to the nearest road (m)



(b) distance ratio between 1st and 2nd matches in 100 m

Figure 9: Distance between samples and their nearest roads

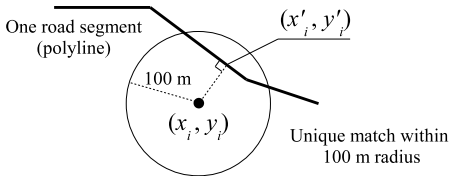


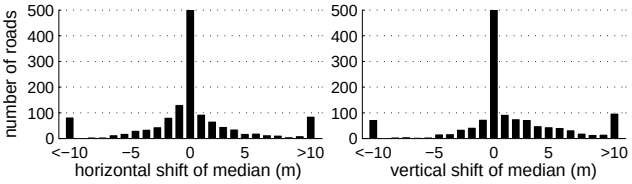
Figure 10: A sample and its nearest point on the road

is reached at 46 m, 38 m, and 52 m, respectively. Fig. 9b shows the ratio between the distance to the nearest and the second nearest roads within 100 meters. The ratio of most samples is more than 1:10 or there is only one road within a 100 meter radius. The latter is displayed as the "N/A" bar in the figure. There are also many samples where the ratio between the two nearest roads is close, likely contributed by taxis waiting for traffic signals near intersections.

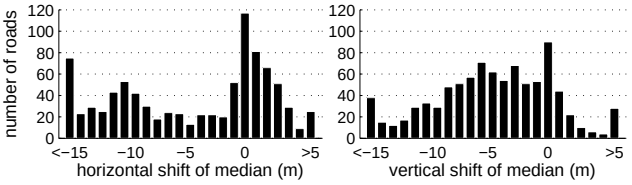
Since a large number of samples reach exactly one road within 100 m radius, assuming the road is the true location of the sample, we use these samples to estimate the noise distribution. Fig. 10 illustrates this strategy. A road segment is a polyilne. If there is exactly one polyline within 100 m radius of a sample, we assign the sample to the road. Considering one day of samples from each city, only those segments with more than 100 samples assigned are included for the analysis. We use this threshold because outliers are quite common in the data. Multipath errors illustrated in Fig. 6 can produce samples at arbitrary locations. As a result, the number of road segments we consider are 1,398 out of 13,466, 892 out of 14,473, and 867 out of 15,495 in BJ, SH, and GZ, respectively. BJ has more road segments of exact match because it has twice more samples than the other two cities. Considering only road segments with more than 200 samples assigned, the number drops down to 981, very close to the other two. For each matched sample (x_i, y_i), we calculate the nearest point on the uniquely matched road segment, (x_i', y_i'). Assuming (x_i', y_i') was the true location of the taxi that generated (x_i, y_i), we analyze the mean and the standard deviation of the noise distribution.

Fig. 11 analyzes the misalignment of maps using the median of horizontal and vertical differences, i.e.,

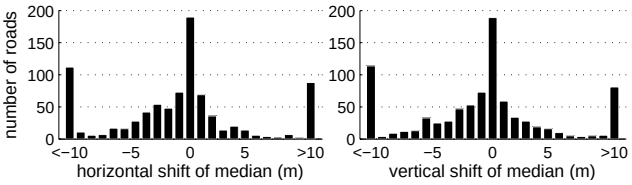
$$\text{median}_i(x_i - x_i')_{great\ circle} \tag{2}$$



(a) Beijing (Open Street Map)



(b) Shanghai (Open Street Map)

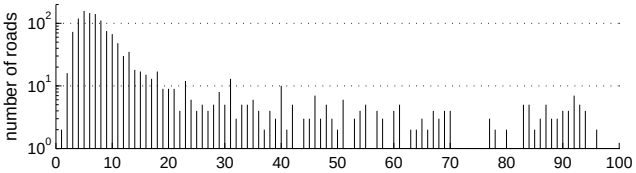


(c) Guangzhou (proprietary map)

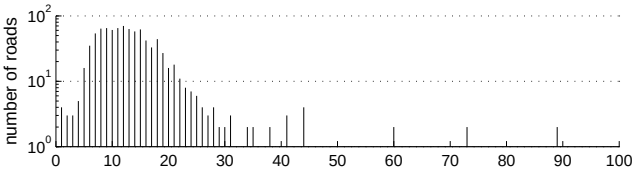
Figure 11: The median of horizontal and vertical map shift

$$\text{median}_i(y_i - y_i')_{great\ circle} \tag{3}$$

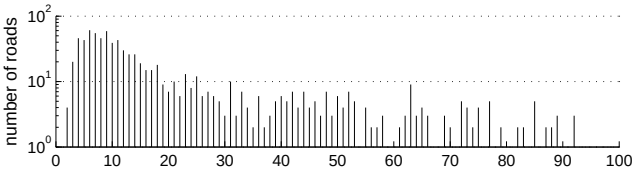
These coordinates are converted using the *azimuthal equidistant projection* centered around each city. Therefore these horizontal and vertical differences closely approximate the great circle distance along latitudinal and longitudinal lines, respectively. The left column in Fig. 11 contains the histograms of (2) for all considered road segments and the right column includes the histograms of (3). The result is consistent with our visual inspection of map accuracy in Section 3.3. In BJ, the map is perfectly aligned with the median of the samples. The outliers where the median shifts more than 10 meters are mostly caused by missing map features, i.e., samples generated from a road or a parking area not included in the map get assigned to the nearest road. Not every missing map feature causes large shifts in both directions. For example, most roads in BJ are well aligned with latitudinal and longitudinal lines. A missing road along the latitudinal line causes zero shift along the longitudinal line. This is another reason of the large concentration of roads with zero median shift. In SH, there are significant numbers of roads with negative medians of samples in both the horizontal and vertical directions, corresponding to a northeast shift of the map. The relatively even distribution of the histogram is likely due to three facts. First, roads in SH are better aligned with Huangpu river than latitudinal and longitudinal lines, also observed in Section 2.5. Therefore a misaligned road more often causes non-zero median in both horizontal and vertical directions. Second, different roads in SH exhibit varying degrees of misalignment with the vector map as well as the aerial image, likely due to image distortion. Last, the measurement is accurate to only 0.0001 degree, or roughly 10 meters. Therefore, in the worst case, the samples are always at least 5 meters away from the road center. The histograms of GZ are somewhat in between. Most roads are well aligned with the samples but there are also significantly more outliers.



(a) Beijing: median of the great-circle distance (m)



(b) Shanghai: median of the great-circle distance (m)



(c) Guangzhou: median of the great-circle distance (m)

Figure 12: Noise estimation using median absolute deviation

Fig. 12 shows the log-scale histogram of the median value of the great-circle distance, i.e.,

$$\text{median}_i(\|(x_i, y_i) - (x_i', y_i')\|_{great\ circle} \quad (4)$$

Assuming Gaussian distribution of noise, this median value can be used for the estimation of standard deviation using median absolute deviation (MAD), i.e.,

$$\sigma = 1.4826\ \text{median}_i(\|(x_i, y_i) - (x_i', y_i')\|_{great\ circle} \quad (5)$$

MAD is a robust estimator for noisy data. It has been used for the estimation of Gaussian GPS noise before [16].

The result exhibits surprising accuracy and consistency in all three cities, despite the heterogeneous GPS devices used across the five-year time span. In BJ and GZ, the median distance for most roads are around 4-10 meters. Beyond 20 meters, the number of roads is an order of magnitude less, likely due to missing map features rather than GPS noise. For example in BJ, 32 out of the 57 roads with the median distance beyond 80 meters have a variance below 100. The combination of large median and small variance indicate that the samples of a road are located in a narrow band far away from it, implying most likely a missing road there. In SH, the center of the histogram is between 6-18 meters. Discounting the map misalignment, the true median distance should be closer to BJ and GZ. The relatively flatter and longer span of the histogram center is consistent with the relatively even distribution of the median horizontal and vertical shift in Fig. 7. There are very few outliers beyond 50 meters, indicating good map accuracy. Overall, we believe the GPS locations are reasonably accurate under normal operating conditions in all cities. Considering varying lanes of different roads as well as inaccurate maps, we take 5 m as the lower bound for the median distance under ideal situation, i.e., a single-lane road mapped accurately. This results in a standard deviation of $5 \times 1.4826 = 7.41$ meters, on par with 4.07 meters reported in [16].

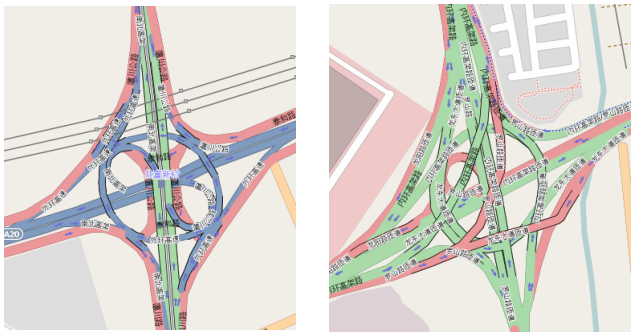


Figure 13: Elevated expressways and their intersections

4. DISCUSSION

Having analyzed the data in great detail, we discuss challenges and opportunities in its relevant applications.

4.1 Map Matching Challenges

Map matching taxi GPS data faces numerous challenges. The sampling rate is low and dynamic. Missing samples and erroneous timestamps are possible. The driving behavior can be different when vacant or occupied. The accuracy of the map may be limited, especially in developing countries. In the metropolitan area, city canyons and tunnels can produce noisy outliers. While these errors are statistically insignificant, their effective identification and removal are very challenging and important for certain applications of the data.

Existing map matching algorithms can be largely divided into local/incremental methods that match samples one by one [16, 17], and global/geometrical methods that map an entire trajectory at once [6, 15]. Neither solution handles the type of noise in Fig. 6 well. Local methods will map all samples including the outliers. Threshold-based outlier removal is not effective, as discussed in Section 3.2. Global methods have the potential to detect these outliers using geometrical algorithms, but picking the right trajectory to begin with is non-trivial. For example in Fig. 6b, the outlier at the top is located right on the street, much closer than many other "normal" samples. Picking it as a start or end point of a trajectory significantly affects the subsequent map matching procedure. On the other hand, statistical algorithms applied in local methods are more robust against random noise, e.g., HMM [16]. Some combination of both methods may win favor. For example, applying HMM to the cases in Fig. 6 would result in very low or zero probability for the outliers. Then global methods can help confirm the outliers based on normal samples before and after them, therefore restore the correct match.

Finally, complicated road networks pose another challenge. In China, elevated expressways are often laid over existing signal-controlled roads in the metropolitan area. The map becomes quite messy when these double-layer roads intersect; see Fig. 13 for two examples. Without the elevation information of both taxis and roads, GPS measurements alone are insufficient to locate a taxi on either layer. Speed may not help either as a congested expressway can be slower than the local road underneath. Unfortunately as these elevated expressways are arteries of the city, their traffic information is crucial for applications such as travel time estimation.

4.2 Research Opportunities

Section 3 showed that the reported GPS location is often more accurate than map. Therefore, taxi GPS traces can calibrate maps, fill in missing roads, and even build maps from scratch. With a reasonably accurate map matching algorithm, calibrating existing maps is straightforward, as Section 3.4 proves. Map building is more challenging. Previous work in this direction used low-noise GPS devices at the sampling rate of 1 Hz [7]. The result can be extended to handle sampling intervals of 30 seconds or less for straight-line road segments, as Fig. 7 demonstrates. With curved roads or greater sampling intervals, however, the trajectories bear little similarity with the true roads. We plan to address this problem in our future work.

In addition to the accurate location measurements, the heading directions reported are very reliable as well, especially when the taxi is moving. We considered road segments in Section 3.4 where the median of the distance among matched samples is less than 20 meters. The difference between the orientation of the road and the reported direction is often less than 10 degrees when the speed is more than 20 km/h, except for SH-B taxi group that report only 8 directions. Some existing map matching algorithms have considered vehicle orientation information for a better local match of a given sample [9]. We are interested in using the measurement to evaluate map matching accuracy of existing algorithms. Because the ground truth of the actual drive paths is hard to obtain in the large scale, reported direction can be a good alternative. Once the effectiveness of various map matching algorithms is well understood, we can incorporate the direction measurement for even more accurate results. In addition, directions can help with map building, e.g., grouping samples with similar directions and filtering out outliers. Finally, combining map calibration with map matching is interesting. For example, we have shown outliers in Fig. 12 that most likely represent missing map features. Instead of matching the corresponding samples to an existing road, we can fill in the missing features and produce a better match.

Our analysis of GPS data across a five-year time-span shows progress in both data quality and measurement accuracy. Improved devices and standardization certainly helped. Looking into the future, we expect to see more accurate devices at possibly lower cost. National or even global standards are possible. However, various noise and outliers presented in this paper will exist for a long time. This is because of the inherent limitations of GPS measurements, as well as the nature of large-scale systems. Therefore, applications that exploit the data must be resilient to various errors. Statistics can help with reliable results.

5. CONCLUSION

We have shown that large-scale GPS probe data present new challenges and opportunities for the relevant applications. All measurements from the GPS device, including timestamp, can be wrong. The map can be inaccurate too. While these issues are mostly statistically insignificant, inappropriate handling can propagate and magnify the error. On the other hand, the GPS location measurement is accurate under normal conditions and even more reliable than the aerial imagery. With the large quantity of data, the law of large numbers can lead to highly accurate results. Robust statistical algorithms may be the key to success.

6. REFERENCES

[1] Apple, Google collect user data. `http://online.wsj.com/article/SB10001424052748703983704576277101723453610.html`.

[2] Open StreetMap. `http://www.openstreetmap.org`.

[3] R. L. Bertini and S. Tantiyanugulchai. Transit buses as traffic probes: Use of geolocation data for empirical evaluation. *Transportation Research Record*, 1870:35–45, 2004.

[4] A. Biem, E. Bouillet, H. Feng, A. Ranganathan, A. Riabov, O. Verscheure, H. N. Koutsopoulos, and C. Moran. IBM infosphere streams for scalable, real-time, intelligent transportation services. In *ACM SIGMOD*, 2010.

[5] S. Blandin, L. El Ghaoui, and A. Bayen. Kernel regression for travel time estimation via convex optimization. In *IEEE Conference on Decision and Control*, 2009.

[6] S. Brakatsoulas, D. Pfoser, R. Salas, and C. Wenk. On map-matching vehicle tracking data. In *Proceedings of the 31st international conference on Very large data bases*, 2005.

[7] L. Cao and J. Krumm. From gps traces to a routable road map. In *ACM SIGSPATIAL GIS*, 2009.

[8] Y. Chen and J. Krumm. Probabilistic modeling of traffic lanes from GPS traces. In *ACM SIGSPATIAL GIS*, 2010.

[9] J. S. Greenfeld. Matching GPS observations to locations on a digital map. In *Transportation Research Board 81st Annual Meeting*, 2002.

[10] J. C. Herrera, D. B. Work, R. Herring, X. J. Ban, Q. Jacobson, and A. M. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C: Emerging Technologies*, 18(4):568–583, Aug. 2010.

[11] R. Herring, A. Hofleitner, S. Amin, T. Nasr, A. Khalek, P. Abbeel, and A. Bayen. Using mobile phones to forecast arterial traffic through statistical learning. In *Transportation Research Board 89th Annual Meeting*, 2010.

[12] J. Jariyasunant, B. K. D. Work, R. Sengupta, S. Glaser, and A. Bayen. Mobile transit trip planning with realĺctime data. In *Transportation Research Board 89th Annual Meeting*, 2010.

[13] K. Liu, T. Yamamoto, and T. Morikawa. Feasibility of using taxi dispatch system as probes for collecting traffic information. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 13(1):16–27, 2009.

[14] S. Liu, Y. Liu, L. M. Ni, J. F. 0002, and M. Li. Towards mobility-based clustering. In *ACM KDD*, pages 919–928, 2010.

[15] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang. Map-matching for low-sampling-rate GPS trajectories. In *ACM SIGSPATIAL GIS*, 2009.

[16] P. Newson and J. Krumm. Hidden markov map matching through noise and sparseness. In *ACM SIGSPATIAL GIS*, 2009.

[17] O. Pink and B. Hummel. A statistical approach to map matching using road network geometry, topology and vehicular motion constraints. In *IEEE Conference on Intelligent Transportation Systems*, 2008.

[18] M. A. Quddus, W. Y. Ochieng, and R. B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312 – 328, 2007.

[19] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang. T-drive: driving directions based on taxi trajectories. In *ACM SIGSPATIAL GIS*, 2010.

[20] H. Zhu, Y. Zhu, M. Li, and L. M. Ni. Seer: Metropolitan-scale traffic perception based on lossy sensory data. In *IEEE INFOCOM*, 2009.