Evaluating the accuracy of GPS-based taxi trajectory records

Zhong Zheng, Soora Rasouli, Harry Timmermans*

Eindhoven University of Technology, PO Box 513, Eindhoven 5600MB, Netherlands

Abstract

Taxi data are an underused source of travel information. A handful of research has been concerned with the processing of raw taxi GPS data to minimize random error. The study of methods that systematically detect erroneous data has, however, received less attention. Generally, an origin and a destination are identified when the taxi occupancy status (occupied/vacant) shifts. Information would be wrongly recorded if taxi drivers incorrectly operate their device or missed signals when the status of their taxi changes. It leads to extremely short trips or long trips. This study proposes a set of criteria to evaluate the accuracy of trips, imputed from taxi GPS data. In particular, attributes such as inaccurate signal, mismatch of movement and speed, abnormal average speed, and mismatch of trips length measured on maps and calculated from records are suggested. Taxi data should pass these tests if trips have been identified accurately. Using these suggested criteria, the accuracy of 150 million GPS records, collected in Guangzhou, China is evaluated.

Keywords: Taxi data, GPS data, OD trips, data processing

1. Introduction

Travel demand models play an important role in urban and regional planning. They serve to predict future demand for infrastructure and facilities, or are used to estimate changes in transport and activities over time. Most transportation-related problems, including traffic congestion, crash frequency, energy consumption, and vehicle
emissions, are directly related to vehicle usage rates and driver behaviour. It articulates the relevance of activity-based models, which have been formulated as an alternative to four-step models for a variety of reasons. Ignoring a detailed discussion of the specific reasons, it did result in an approach that is much more detailed in time and space, compared to the traditional traffic zones. Giving the resulting need for accurate and detailed trip trajectories, GPS technology has gained increasing attention in transport research. It has the advantage of collecting more accurate and massive data on locations, time, routes and other driving related information compared to traditional methods.

There are two modes of GPS data collection: vehicle-based and person-based. For vehicle-based data collection, vehicles are equipped with GPS collectors, reporting location information with a certain time interval. In a personal GPS data collection survey, individuals are required to take GPS loggers or GPS-enabled smart phones with them or use an app in their smart phone. Taxi GPS data are an example of the first kind of data collection and may be a reliable source for researching vehicle behaviour. The GPS records track taxi information rather than personal information, and hence this data collection approach does not involve any privacy issues. The GPS loggers are always turned on when taxi drivers are working. Furthermore, it is convenient to record the start and end point of a trip as loggers are related to taxi meters. However, all other transportation modes are omitted when only collecting taxi GPS data. Moreover, if the purpose is to collect daily diaries, taxi data tend to generate incomplete diaries.

Thus, taxi GPS data may be useful in travel behaviour research. Compared to the use of GPS loggers and smart phones, relatively little is, however, known about the quality of taxi GPS data. While an abundant amount of research has been conducted on the accuracy of GPS data, embedded in smart phones or as stand alone devices and their use in the collection of activity-travel data, only a few studies have examined the accuracy of taxi GPS data. Little is known about the quality of taxi GPS data, which will contain device, information and system errors. Accuracy evaluation and subsequent filtering of the data is a necessary step for any analysis. It should be realized in this context that the accuracy of person-based imputed activity-travel diaries is often achieved by administering a prompted recall instrument asking respondents to check and if needed rectify the imputed data. Taxi data prohibit such use of prompted recall instruments, implying a higher importance of error detection, filtering and possibly correction algorithms.

Existing research has proposed outlier detection methods but a deep understanding of the distribution and root causes of data error is still limited. In contributing to the assessment of the usefulness and limitations of taxi GPS data, we propose a trip-based evaluation method, which uses a set of criteria to examine the accuracy of trip information, derived from raw taxi GPS data. The criteria relate to different attributes of taxi trips. Data should pass these criteria to be considered accurate data. In addition, causes for erroneous data are identified.

The remainder of this paper is arranged as follows. Section 2 briefly introduces related work on GPS error detection in general and of taxi trip records in particular. Methods to improve data quality and understanding of errors are discussed in this section as well. Section 3 presents the proposed method, which includes four criteria to evaluate data accuracy. Section 4 outlines the data collection process and describes the data structure. Section 5 provides the findings of an application of the suggested approach. Error distributions, effectiveness of filtering and causes of error are also presented. Finally, Section 6 draws conclusions and sketches avenues of future work.

2. Related work

A handful of studies explored missing data, uncertainties and errors in taxi GPS records. Incorrect records may be caused by mismatching GPS data to map coordinates, low accuracy of GPS navigation devices, or low sampling frequency. To minimize random error, spatial smoothing methods have been developed. Smoothing techniques such as Gaussian kernel filters may be employed for removing systematic and random errors. However, problems of wrong data are not addressed and thus still exist. Because of signal sheltering and unstable GPS collectors, some records may have problems, such as missing data, invalid locations, or wrong time. Operating errors of drivers may also have resulted in erroneous records. Spatial smoothing methods do not help much in evaluating and excluding wrongly recorded information.

A substantial amount of research has concentrated on noisy records (outliers/uncertainty/error information) detection and filtering. The most common detection concerns outliers. Taxi trips may involve long distances; cases where taxis travel beyond provinces or even countries may happen. Veloso et al. discard travel distances larger than 30 km. Very short trips may also be problematic because passengers are more likely to take transport modes other than taxis for such distances. Therefore, scholars have discarded trips of less than 200 meters or 500 meters.
Measuring attributes of travel records (location, distance, speed, duration, direction, etc.) is another effective approach to identify possible errors. These attributes should fall within a reasonable range and be self-matched for correct records. Straight line speed, abnormal coordinates, map accuracy and Gaussian noise measurement have been used to examine data accuracy in a large-scale GPS probe data study\(^3\). A survey conducted in New York\(^4\) used origins and destinations of trips to identify outliers. If a taxi trip has either a pickup location or drop-off location that cannot be snapped to a street segment within a reasonable distance threshold, it is considered as type I outlier. If the ratio of computed shortest path and recorded distances is greater than a threshold, the records are marked as type II outliers. Fortunately, we have trajectory data, which are more detailed than OD data. Therefore, it is convenient and proper to calculate the measured lengths and compare these with recorded lengths. In addition, we will offer more criteria to filter out outliers.

Besides the self-matching of measurements, measurements should be consistent with external constraints. For example, Bertini and Tantiyanugulchai \(^12\) compared bus travel data extracted from the dispatch system and data collected by GPS sensors on a 2.5 miles corridor. A study used coordinate errors and communication errors to reveal the feasibility of two sets of taxi GPS data\(^13\). Different from self-mismatching, domain-oriented approaches consider locations outside certain administrative regions or within regions of certain land use types (e.g. river/lake) as outliers through geospatial analysis such as point-in-polygon testing or nearest neighbour computing\(^2\). However, this approach requires accurate map matching. Spatial points located in impossible terrain, such as a river or sea, suggest erroneous data.

It is also important to know the statistics and distribution of error. Hidden Markov Model (HMM) approaches on mapping error GPS data assume that the errors follow a Gaussian distribution\(^14, 15\). However, the true distribution of errors remains unknown.

3. Method

We define the proposed method as a GPS outlier detection process (Fig. 1). As illustrated in table 1, a taxi GPS record has the following attributes: location, time, speed, direction, and occupancy. Taxi trips can be identified from a change in occupancy status. The occupancy (STAT field) shifting from vacant to occupied defines the origin of a trip, while a shift from occupied to vacant marks the destination of a trip. By processing and aggregating data to trips, additional information, such as average trip speed and trip distance, can be derived. Our method for the evaluation of data accuracy is based on four criteria and the idea that correctly recorded information should be pass all these criteria.

The first criterion is to detect low accurate signals. Information about signals accuracy is given in the raw data records (EFF field), where the value ($V_{eff}$) is 1 for high accuracy and 0 for low accuracy. For a global positioning system, at least three satellites should be accurate, while less than three satellites may cause error and low accuracy.

The second criterion is a mismatch of movement and speed. Sometimes a taxi’s movement status may be contradictory according to different fields in the GPS records. The GPS record may indicate zero speed, although the location is changing. The GPS record may also be larger than zero speed, while the location remains the same. Both mismatches point are erroneous data, and such data should be filtered out.

The third criterion is abnormal driving speed. An operational definition of abnormal driving speed is a threshold $\gamma$ speed. In China, the maximum speed limit is 120 km/h on highways and 80 km/h in the city. If a trip’s average speed is higher than say 120 km/h, it may be signal error in de data. Negative speeds may also occur. These trips with an abnormal average speed should be removed from the data set before further analysis.

The fourth criterion is a mismatch between the distance measured from a map ($D_{map}$) and distance recorded from GPS data. The latter distance is calculated by multiplying speed and time $S_{avg} \times T_{avg}$. The ratio of these two distances is not necessarily equal to one because speed information can be wrongly recorded. Thus, trips for which this distance ratio deviates substantially from one should be filtered out. In the present study, we set the tolerance ratio value $\delta$ as 2. Setting the value equal to 2 is arbitrary; it was inspired by a similar study working on OD shortest length\(^4\).

The suggested outlier detection process thus involved first testing the accuracy of the signals, and then processing the traces to calculate trip-level related indicators. Next, outliers are detected by testing each piece of information against a set of three criteria, which basically set thresholds to filter out abnormal records and inconsistent data. Only those trips, which pass the tests on all four criteria, are considered valid; all other data are filtered out and not included in the analysis. Although this method is not perfect in the sense that some erroneous data may slip through, we contend that many errors will be filtered out.
4. Data

The suggested approach is illustrated using taxi GPS data collected in Guangzhou, China (Fig. 1). Guangzhou is the largest city in South China and the capital of Guangdong Province. The city is in the process of rapid urbanization and globalization. It has a population of almost 13 million, an area of 7434 km². The traffic Research Center of Sun Yat-sen University provided 6 days of GPS data. In Guangzhou city, every taxi installs a GPS collector, which sends signals every 20 seconds containing information about location, time, speed, carrying status, etc. The GPS collector is activated when a taxi driver is at work, whether the vehicle is moving or stationary. There are over 100 million GPS records in six days. The current analysis is based on a sample of 1.5 million records. After ordering the data by time and license, we selected the top 1.5 million records from the database pertaining to 11th May, 2009. Although the selected records make up the top 1.5 million records for that day, this data set still represents a random sample of taxi licenses. A field description of the GPS records and an example record are shown in table 1 and table 2 respectively.
Table 1. Field description.

<table>
<thead>
<tr>
<th>Field name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>License</td>
<td>Gives a unique identification for every taxi. For data protection purpose the given example in table 1 gives a fake license number.</td>
</tr>
<tr>
<td>Date</td>
<td>The date of the record.</td>
</tr>
<tr>
<td>Time</td>
<td>The time of the record.</td>
</tr>
<tr>
<td>Longitude</td>
<td>The X coordinates of the location.</td>
</tr>
<tr>
<td>Latitude</td>
<td>The Y coordinates of the location.</td>
</tr>
<tr>
<td>Speed</td>
<td>The instant speed of the record.</td>
</tr>
<tr>
<td>Direction</td>
<td>The direction of the record, value from 0 to 360, beginning from north and increasing clockwise.</td>
</tr>
<tr>
<td>EFF</td>
<td>Gives data accuracy according to satellites, 0 for low accuracy and 1 for high accuracy.</td>
</tr>
<tr>
<td>Stat</td>
<td>Occupancy status of the taxi, 0 for no status, 1 for prevent robbery, 2 for sign-in, 3 for sign-off, 4 for occupied, 5 for vacant, 6 for ignition, 7 for flameout.</td>
</tr>
</tbody>
</table>

Table 2. Taxi GPS record data example.

<table>
<thead>
<tr>
<th>Taxi GPS Record</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Field</strong></td>
</tr>
<tr>
<td>Example</td>
</tr>
</tbody>
</table>

For illustration purposes, we have mapped the distribution of all sample taxi records (Fig. 2a) and those related to the inner city area (Fig. 2b) using open street map. The sample records cover a much larger area than Guangzhou city. Fig. 2b clearly shows that the majority of sample records are in midtown Guangzhou. A few are located in the neighbouring city Foshan, and few in other cities.

5. Application

In this section, we map the four types of outliers and compare the data quality before and after the filtering process. The reasons for outliers are also concluded.
5.1. Distribution of invalid records

The statistical pattern of noisy records in large-scale data remains unknown, although some research has assumed the records follow Gaussian distributions\textsuperscript{13,14}. To better understand the distribution of invalid records, it is essential to map the data.

In this section, the purpose of mapping the data is to identify and examine outliers in detail and explore hotspots of noises in the downtown area. Patterns of the four kinds of invalid records are compared.

The spatial scale of the low accurate signals criterion is regional, whereas the scale of the other three criteria is Guangdong Province. For most invalid records located in the downtown, the hotspot map is a good way to present the noise pattern. Kernel density estimation\textsuperscript{16} is a popular method to compose the hotspot map, but taxi data are constrained to road networks, and therefore data near road intersections may generate fake hotspots. Using road density (assigning taxi records to nearest roads) would also be problematic because there are still some records not on roads (parking lot, underground, e.g.). Considering our purpose of finding hotspots to further explore noisy data details, other than getting an accurate result, simply showing points on the map with 98% visual transparency is an option. The transparency is set to 98%, meaning only 2 percentage of colour can be seen for a single point on the map. The heavy colour then indicates large amount of spatial records.

- Distribution of low accuracy outliers

In the raw GPS data, we observe some outliers that are far beyond the destinations where they could be (Fig. 3). As shown in Fig. 4a, these points can be identified by low accurate signals. At the city scale, points off the roads are also found. In these outliers, it is common that a single point shown on the map would be a set of points with the same location. For example, there are 224 records (ID num. 1102006-1102230) indicated by the solid square (trip #1) in Fig. 4a, most of which have 0 instant speed. The other information (time, speed, direction, occupancy) is valid. These errors can be seen as position function failure of the devices considering they are continuous records. Similar extreme cases happen at coordinate (0, 0). We observe 9219 records located at the (0, 0) coordinate, and they are several sets of continuous records.

Records in dotted square (ID num. 644546-644557) have very large instant speed (over 200 km/h) and no occupancy status. In all 267 298 low accurate records, 83.5% have 0 instant speeds. It implies that GPS signals would be more uncertain for stationary taxis than moving ones. But Fig. 5 also illustrates that a too large instant speed is also problematic. In the downtown area, several hotspots are found (Fig. 4b). Hotspot 1 is the location of the main railway station, which has the highest taxi number of taxis in the city. At hotspot 2 and hotspot 3 many taxis have the “flameout” status. The flameout status refers to the run-down of a jet engine caused by the extinction of the flame in the combustion chamber\textsuperscript{9}. Though the reason for the “flameout hotspot” location is unknown, we can conclude that signal flameout identifies erroneous records.
Distribution of mismatched movement and speed outliers

Using the next three criteria, we identified some remote outliers, mostly in the city (see Fig. 6-8). These three criteria are mostly relevant for the city as operate more on driving behaviour than locations. The mismatch of movement status, identified by contradictions in speed and location changes, can detect errors in the GPS ‘speed’ field. A moving taxi with 0 speed must be wrong, while larger than 0 speed for a static taxi is also impossible.

9764 (6.5%) records suggest stationary taxis with speed larger than 0. These mistakes are mainly caused by quick shifts in occupancy status. For example, a taxi’s occupancy changes during the last time interval, and changes again to a different status in the next time interval, leading to a trip containing only one record. The travel length of this kind of single-pointed trip is 0. These data do not represent actual trips. The quick shift in occupancy status can be explained by wrong operations of drivers.

Records with 0 speed but moving locations represent 93.5% of the errors based on this criterion. 82.3% of their travel lengths is less than 1000 meter. These errors may be caused by the devices unable to detect the speed of slow moving taxis. Records in the solid square (Fig. 6a) are from the same taxi (ID num. 485484-458700). According to
their 0 instant speed and flameout status, these records should have the same location. But strangely these points move on the map. Records (ID 1131887-1131900) in dotted squares move slowly and randomly. On the heat map (Fig. 6b) the hotspots are spatially more randomly distributed than strictly constrained to the road network. Such data indicate errors/noise in the GPS traces.

- Distribution of abnormal speed outliers

The intention of the abnormal speed detection criterion is to remove trips with very large speed. Fig. 7 shows that the number of trips with speed larger than 120 km/h is small (around 200). Other erroneous records have a negative speed. Technically, data were exactly ordered by time, but because of format mistakes in time fields, when exported from the database, the data may have been wrongly read, with the result that some dirty data appear in the series.

These data cannot be trusted as they will cause more mistakes in future calculations. Actually, the minus speed is not a fundamental GPS problem, but we still report it as it does happen.
• Distribution of mismatched distances measured and calculated outliers

The mismatch of trip lengths measured on the map and calculated in terms of multiplication of speed and time also allows the detection of outliers. Points in dotted squares (ID num. 643962-644156) in Fig. 8a are imputed as belonging to the same trip. It is obvious, however, these points do actually not belong to the same trip. It shows the high risk that GPS devices wrongly record taxis with the flameout status. Records in solid squares (Fig. 12a) are also from the same trip. Taxis in the south in solid squares (ID 644331-64439) travel more than 400 km to the north in 40 minutes, which is definitely impossible. The instant speed of these records is very high (over 300 km/h on average) and some are over 900 km/h (ID 644403 and 644404). Application of this criterion can also detect very short trips. Identified from the heat map (Fig. 12b) the travel length of trip ID 12630-12639 is measured as 195 meter and calculated as 55 meter. The taxi was occupied and driving for 3.2 minutes at that time. The calculated distance of very short trips increases the uncertainty of the ratio value, which can be used to exclude very short trips.

Fig. 8. (a) Data filtered by ‘ratio of distances’ (all data); (b) Data filtered by ‘ratio of distances’ (all data).

5.2. Comparison of data before and after filtering

To illustrate the suggested filtering method a comparison of data before and after filtering is necessary. To better understand data quality improvement, statistics (direction, instant speed, average speed, ratio of distance) are compared. Direction and instant speed are based on every single record; average speed and ratio of distance are based on trips.

Heading direction presents a periodic waved pattern, with peak values near 90, 180, 270, and 360 degrees. The explanation is that road networks are square-shaped, and oriented toward the four main directions (north, south, east, west). Noticing that the first bin on the left is heading direction equal to 0, the abnormally large number of 0 degrees is difficult to explain (Fig. 9a). However, after filtering (Fig. 4b) the number of 0 degree directions has decreased to a reasonable level. Records pertaining to the four main directions are distributed uniformly.
Speed information can be obtained from GPS records or geographical measurement. Here we used speed from GPS records, and geographical measurement to compare it with the calculated distance. For instant speed measured by records, the maximum instant speed value has reduced to under 250 km/h after filtering (Fig. 10b), compared to the raw data (Fig. 10a) (noticing the different scale in the vertical bar). The results for average speed measured by trips are improved (Fig. 11a, 11b), suggesting the filtering method improves data stability and certainty.

The ratio of distance measured geographically and calculated is another important indicator of data quality. The two sets of distances are from different sources; data with a ratio of 1 are trustable. Fig. 12a shows that the ratios of some trips are abnormally large (more than 500). After removing trips with a ratio value larger than 2, the results definitely improve. To check how close these values are to 1, we shift the horizontal and vertical bar to show the data after filtering (Fig. 12b). The ratio approximately has a normal distribution around 1, indicating good data quality.
Fig. 12. (a) Ratio of distance measured and calculated (after filtering); (b) Ratio of distance measured and calculated (after filtering).

Table 3. Data accuracy evaluation.

<table>
<thead>
<tr>
<th>Records excluded by</th>
<th>1500000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ineffective signal</td>
<td>267297</td>
</tr>
<tr>
<td>Mismatch of movement and speed</td>
<td>150898</td>
</tr>
<tr>
<td>Abnormal average speed</td>
<td>24132</td>
</tr>
<tr>
<td>Mismatch of distance on map and distance calculated</td>
<td>85531</td>
</tr>
<tr>
<td>Valid records</td>
<td>972142</td>
</tr>
</tbody>
</table>

Fig. 13. Valid GPS data.

After the data accuracy evaluation process, 972 142 out of 1 500 000 records passed the test and are considered valid records (see table 3 for details). The spatial results are shown in Fig. 3, which clearly indicates that after filtering outliers were successfully removed from the map, though some still remain.

5.3. Causes for abnormal records

An accurate GPS record needs a well-working device, at least 3 satellites in sight and a good surrounding environment. In a tunnel or parking garage, there would be no satellite signal, implying that the last known location,
i.e. tunnel entrance, is reported. High-rising buildings near taxi stops will result in multi-path error, which is difficult to detect and filter out. In our data, we also find evidence of such causes. First, there are cases pointing at failures in positioning. Failure in GPS devices positioning function obviously produces wrong spatial coordinates. As an illustration, case coordinates of 9219 records were (0,0) but other information (time, speed, occupancy, e.g.) was recorded validly. Noticing these error records are continuous, it implies that satellites were receiving wrong signals during a certain period of time.

Second, statistically stationary and flameout taxis have a higher risk of sending wrong signals, as 83.5% of inaccurate signals have 0 instant speed. The situation of stationary taxis is more complicated than that of moving ones. Taxis may stay under high-rise buildings or trees where signals cannot be sheltered or wrongly received. Cold starts of GPS devices will cause signal delay and mistakes. The quality of the first signal of a trip may be not good. Flameout is an abnormal status for a taxi. It leads to either very slow movement (error trip #3 and #4) or very fast movement (error trip #5). Third, wrong operations of drivers create unreal travel. These trips are usually very short, but they represent dirty data for travel analysis.

Table 4. Examples of error trips.

<table>
<thead>
<tr>
<th>Error trip number</th>
<th>Record ID</th>
<th>Filtered by</th>
<th>Cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1102006-1102230</td>
<td>Ineffective signal</td>
<td>Device failure</td>
</tr>
<tr>
<td>#2</td>
<td>644546-644557</td>
<td>Ineffective signal</td>
<td>Device failure</td>
</tr>
<tr>
<td>#3</td>
<td>485484-458700</td>
<td>Mismatch of location and speed</td>
<td>Flameout</td>
</tr>
<tr>
<td>#4</td>
<td>1131887-1131900</td>
<td>Mismatch of location and speed</td>
<td>Flameout</td>
</tr>
<tr>
<td>#5</td>
<td>643962-644156</td>
<td>Ratio of distances</td>
<td>Flameout</td>
</tr>
<tr>
<td>#6</td>
<td>644331-64439, 644397-644504</td>
<td>Ratio of distances</td>
<td>Unknown</td>
</tr>
<tr>
<td>#7</td>
<td>12630-12639</td>
<td>Ratio of distances</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

6. Conclusions

This paper proposed an outlier detection process method to evaluate the accuracy of taxi GPS data. Using four criteria (low accuracy, mismatch of movement and speed, abnormal average speed, mismatch of distance measured on map and distance calculated) 972142 out of 1500000 GPS records in our database appear valid. The filter method is effective as abnormal values in raw data are reduced to a reasonable small number. Using the four filter criteria, different types of erroneous data were identified. Most far away outliers are detected by the low accurate signal criterion. The other three criteria mainly act on erroneous data in the city. Reasons for errors are complicated and difficult to uncover. Some causes are reflected in the statistics and distribution of error data. GPS devices sometimes fail in positioning correct coordinates; stationary and flameout taxis have a higher risk of being wrongly recorded; drivers’ wrong operations will cause very short and unreal travel.

Thus, only 64.8% of the records seems valid, which is a relatively low percentage. This finding shows that taxi GPS data are not without problems. If this percentage is not exceptionally low but representative for taxi data, reported data errors for personal devices and smart phones are much lower. In that sense, taxi data do not offer a good alternative. However, the advantage of taxi data is their sheer volume. Even though the percentage of erroneous trips is relatively high, many data remain. If these remaining data constitute a random sample, they are well suited for trip analysis. The only issue that then remains is to identify the substantial topics for which taxis data are relevant.

References