On processing GPS tracking data of spatiotemporal car movements:

a case study

Author 1: Xiaoyun Zhao
Editor: Hasan Fleyeh
Nr: 2014:11
On processing GPS tracking data of spatiotemporal car movements: a case study

Xiaoyun Zhao*

School of Technology and Business Studies, Dalarna University, Sweden

This version: 2015-07-13

The advancement of GPS technology has made it possible to use GPS devices as orientation and navigation tools, but also as tools to track spatiotemporal information. GPS tracking data can be broadly applied in location-based services, such as spatial distribution of the economy, transportation routing and planning, traffic management and environmental control. Therefore, knowledge of how to process the data from a standard GPS device is crucial for further use. Previous studies have considered various issues of the data processing at the time. This paper, however, aims to outline a general procedure for processing GPS tracking data. The procedure is illustrated step-by-step by the processing of real-world GPS data of car movements in Borlänge in the centre of Sweden.

Keywords: GPS tracking data, visualized map, road network, map-matching

1. Introduction

Global Positioning System (GPS) technology has developed enormously in the last few decades and it continues to improve. The use of a portable device, such as a smartphone or other communication devices with built-in GPS for navigation and orientation is nowadays common.

GPS devices can also track mobile objects with regards to longitude, latitude, time, speed and altitude at regular time intervals. This spatial and temporal information can be used for investigating the activities of people and their induced effects. It is possible to use the unprocessed GPS tracking data; however, its usage is limited to rather few aspects namely those only requiring recorded speed,

* Xiaoyun Zhao is a PhD-student in Micro-data analysis and corresponding author: xzh@du.se, phone: +46 23-778509
coordinates and time. Several studies have addressed certain issues that arise in processing GPS tracking data. For instance, Kharrat et al. (2008) proposed an algorithm (NETSCAN) for mobile object clustering and applied it in an environment constrained by a network. Giannotti et al. (2011) presented a query and data mining system named M-Atlas, but noted that it is difficult to transform GPS tracking data into mobility knowledge. Etienne et al. (2012) provided a method for detecting outliers of spatiotemporal trajectories with primary applicability for travel behaviour analysis.

However, no study has attempted to discuss all issues related to processing GPS tracking data simultaneously, let alone provided a procedure or open source code with detailed information for doing so. The aim of this paper is to address several of the issues arising in processing GPS tracking data and thereby outline a general procedure for the data processing. Moreover, the program code in the process is public as a reference for use of interest\(^1\). The study is carried out by using real-world GPS tracking data of some 300 cars that were originally collected for the purpose of studying CO2-emissions induced by retailing.

The processing of the GPS tracking data requires a clear definition of movement, a detailed understanding of the capability of the GPS device and the output of the GPS logger, access to digital data of the road network as well as methods for matching the GPS data and the network. All this is discussed in the paper. Descriptive statistics and visualized maps are used to summarize and illustrate the mobility patterns. The technical documentation of the data processing in this paper is detailed in the interest of readily being replicable on the same or similar type of data. The data in this paper are freely available upon requesting it from the author.

\(^1\) Requested by email to xzh@du.se or from URL https://github.com/Stefangemfind/MapMatching
Section 2 of this paper gives an overview of the related literature. Section 3 provides details of the data collection. In section 4 definition of movement is given and the processing of the data in the plane is described. In section 5 the movements are further processed to obey the restrictions imposed by the network and the mobility pattern is visualized by maps. Section 6 concludes the paper.

2. Literature review

The application of GPS has increased in location based services and intelligent transportation system as a consequence of the popularity of portable, low-cost GPS devices. There is a large body of studies that have integrated GPS in the areas of ecology(Steiner et al., 2000; Tuner et al., 2000; Cagnacci et al., 2000), agriculture(Stafford, 2000; Auernhammer, 2001; Zhang et al., 2002) and sports (Coutts and Duffield, 2010; Aughey, 2011). The main common feature of these research areas is that there is no underlying road network that confines the mobile objects. However, the road network is a confinement in many mobility studies relying on GPS tracking data (Van Schaick 2010). This paper focuses on GPS tracking data restricted to a road network; there are three broad aspects that have been of concern in this area of research.

Firstly, GPS tracking has been conducted for the purpose of improving the quality and the quantity of travel data. For instance, Wagner (1997), Casas and Arce (1999), Draijer et al.(2000), Doherty et al. (2001) respectively have conducted comprehensive data collection with GPS in Lexington, Austin, Quebec City and the Netherlands to test this method versus ordinary travel diaries. They found that sufficient and valuable travel information could be obtained.

Wolf (2000) checked if GPS data could substitute, rather than supplement, the traditional travel diary. In a later study, Wolf et al. (2001) used GPS data to collect travel data in personal
vehicles and demonstrated that it is possible to derive trip purpose from the data. Gruteser and Grunwald (2003) studied whether it is technically feasible to reduce the privacy risk in location identification. Leduc (2008) conducted a snapshot of the development of traffic data collection methods and discussed the potentials and challenges related to emerging technologies.

Secondly, using GPS data over a certain period of time to analyse human mobility and travel behaviour on a road network by is important. The prime advantage of using GPS is that it provides real-time spatial and temporal information of the entire trip (Grengs et al., 2008), up on which it is possible to identify travel time and distance, origin and destination as well as stops. Patterson et al. (2003) applied GPS tracking to classify a user’s transportation mode of car, bus or walk as well as to predict the individual’s most probable route. Askbrook and Starner (2003), Krumm and Horvitz (2006) and Liao et al. (2007) aimed to understand individuals’ outdoor movements by using GPS data and to extract individuals’ significant places and predict their movements.

Li et al. (2004) inspected the travel time variation in commuting trips, the route choice and the effects on departure time based on GPS data. Zheng et al. (2009, 2010) provided approaches to identify culturally important locations, travel sequences and to differentiate between walking, driving, taking a bus and riding a bike. Huang and Levinson (2012) analysed the influence of movement on a road network and clustered their destinations based on GPS data in the Twin Cities; they found that higher accessibility and diversity of retail services around the destination are more attractive. Schönfelder et al. (2006) concluded that the use of GPS data for travel behaviour analysis could provide unique insight into the structure, size, and stability of human activity spaces.
Thirdly, evaluation of GPS data performance is necessary. Positioning technologies based on stand-alone GPS receivers are vulnerable and have to be supported by additional information to obtain the desired accuracy, integrity and availability (Skog and Handel, 2009).

It is difficult to obtain accurate GPS data since its performance depends not only on the features of the sensor, the GPS receiver and the vehicle model but also on the trajectory dynamics and environments. It is even more challenging in urban environments, buildings may block satellite signals, forcing the GPS receiver to work with a poor geometric constellation of satellites, thereby reducing the accuracy of the data (Huang and Tan, 2006; Modsching et al., 2006; Godha and Cannon, 2007). Marias et al. (2005) found that multipath propagation of the radio signal due to reflection in surrounding objects could lead to decreased position accuracy of the GPS receiver. Schlingelhof et al. (2008) confirmed that development of intelligent transport system applications and location based services require not only higher accuracy GPS but also better reliability and integrity with auxiliary information.

Map-matching is a commonly used solution to improve the accuracy of GPS data by matching positions and trajectories to a road using a digital map of a road network. Greenfeld, (2002), Bruntrup et al. (2005) and Wenk et al. (2006) applied an incremental algorithm for matching GPS positions to their most probable locations on a road network. Brakatsoulas et al. (2005) proposed three map-matching algorithms where the trajectory nature of the data was used to improve accuracy. Mustière and Devogele (2008) provided an approach for matching networks with different levels of detail to determine one-to-many links between networks. Most map-matching studies assumed that the digital map is of high accuracy; however there are many situations in which this is unlikely to be the case. For instance, White et al. (2000) and Ochieng et al. (2009) studied map-matching algorithms
to reconcile inaccurate data with a poor digital road network. Quddus et al. (2007) conducted a thorough survey of the existing map-matching algorithms and found that enhancement is needed to improve the performance of map-matching in dense urban areas with complex road networks.

To conclude, GPS tracking data has become a reliable source to continuously provide travel data over a certain period. Although high data quality cannot be guaranteed, approaches such as map-matching have been widely used in the correction of data inaccuracy. The GPS tracking data have been broadly applied for analysis of travel behaviour and mobility prediction by processing the data; however, studies that have attempted to outline a specific procedure for the data processing and address the related issues are deficient. Therefore, efficiently processing the GPS travel data is crucial to make good use of the current data quantity and quality. Consistently adjusting and modifying the data is necessary to ensure the accuracy and validity for the further data analysis. Precisely illustrating the GPS travel data is helpful to analyse human mobility and predict travel behaviour on a road network.

3. Data Collection

The data collection was conducted by using a type of standard Bluetooth GPS data logger named BT-338X. Although using GPS devices to replace traditional travel diaries can reduce the collection burden and improve the data quality, there will still be substantial non-response by randomly selecting a sample of the population because it requires consent of the individual to carry the GPS device. We instead successfully negotiated an agreement with four large sports associations (Domnarvets GOIF, Kvarnsveden Hockey, Stora Tuna IK and Torsångs IP) to recruit car-owning volunteers in conducting the data collection. Each association provided approximately 75 anonymous volunteers with their home addresses. A unique ID made up of the association name and a number
was assigned to each volunteer.

In total 89 devices were shared among these volunteers according to a protocol. The device combined a GPS receiver and a data logger with a Bluetooth interface to record their car movements. Each volunteer's car equipped one device for one or two weeks. The device was always equipped to the same car for the duration of the tracking period. There was no guarantee that the car with the device would only be driven by the registered volunteer because this car could be shared by all the members in the household. This is however not a concern since the car movements were the tracking target.

The volunteers were aware of the atypical situations such as, failed to charge or carry the device, device malfunction or car issues. The data collection was undertaken from March 29 to May 15 in 2011 and the successful compliance attained to be 95%. The device activated tracking every 5 or 30 seconds. The recorded information included date, time, longitude, latitude and speed. There were 309,263 valid positional recordings after removing 5,402 invalid ones due to signal loss. The data were stored in 316 log files, one for each volunteer.

Figure 1 illustrates the residential distribution of the volunteers and all the residents in Borlänge. The volunteers are spread out in Borlänge in a pattern similar to all the residents. Due to the requirement that every volunteer must possess a car, the volunteers will appear less concentrated in the centremost area compared to all other residents in general. The four sport associations shown by the red triangles are dispersedly located in the city. Most of the volunteers reside in Borlänge; however, the spatial extension of their movements covered more than half of the entire territory of Sweden (Jia et al., 2012). The focus of this paper is the processing of the predominant movements in Borlänge city.
4. Processing GPS data on the plane

4.1 Data from the GPS logger file

The original GPS tracking data from volunteers were recorded into DataLogger files. Each DataLogger file consists of three main variables, Date, TP and positional recording. The variable Date notes the latest date and time when the file was loaded from the device to the computer by using the software GlobalSat Data Logger PC Utility. It is in the format of yyyy-mm-dd tt:mm:ss. The variable TP represents the tracks, in which a track is defined as the sequentially linked line based on a number of positional recordings in a specific time period. Each positional recording contains the information in the sequence of latitude, longitude, time, date, speed and altitude. The longitude and latitude are referenced by the World Geodetic System 84 (WGS 84) in the degrees decimal minutes format and are measured with a precision of 5 meters. The time is in the format of tt:mm:ss. The date is in the format of ddmmyy. The speed was measured in the unit of km/h. The altitude was not recorded.
and was assigned value -1.

Figure 2 shows an example of a DataLogger file from volunteer Domnarvet11. The Date shows that the file was loaded at 2011-04-29-13:15:56. The TP 1= 001, 2011-04-05:20:20:27 signifies that the first track was assigned to 001 and it started at date 2011-04-05 and time 20:20:27. The volunteer Domnarvet11 made 17 tracks in total.

The first track contains 16 positional recordings with numerators from 1 to 16. Specifically, 1=60298968, 15282927, 182027, 50411, 6240, -1 indicate that the latitude is 6029.8968, longitude is 1528.2927, the time is 182027 (which is 18:20:27), the date is 50411 (which is 05-04-2011), the speed is 62.40 km/h and the altitude is filled as -1. The listed time is 2 hours earlier than the actual local time due to the change of the summer time; therefore, the listed time plus two hours is the actual local time in recording the positions.

Table 1 shows the number of valid GPS DataLogger files from the volunteers. There are 48 from Domnarvet GOIF, 59 from Kvarnsveden Hockey, 58 from Torsång IP and 71 from StoraTuna.
IKA. Additional 80 volunteers from StoraTuna were recruited during the data collection and were assigned as the group of StoraTuna IK B.

Table 1. Number of valid GPS DataLogger files from each sport association

<table>
<thead>
<tr>
<th>Valid GPS Logger files</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Domnarvet GOIF</td>
<td>48</td>
</tr>
<tr>
<td>Kvarnsveden Hockey</td>
<td>59</td>
</tr>
<tr>
<td>Stora Tuna IK A</td>
<td>71</td>
</tr>
<tr>
<td>Stora Tuna IK B</td>
<td>80</td>
</tr>
<tr>
<td>Torsång IP</td>
<td>58</td>
</tr>
<tr>
<td>Total</td>
<td>316</td>
</tr>
</tbody>
</table>

Further, we parse the original data into a matrix with eight variables. The Date variable is excluded because it does not provide any information regarding to the car movements. In this matrix, the variable TP is named as TRACK_ID and the variable positional recording is represented by six variables named as PR_ID, LATITUDE, LONGITUDE, TIME, DATE and VELOCITY. The abbreviation of PR_ID means the positional recording ID. The identification for a volunteer is displayed as USER_ID. Figure 3 shows this structure and all the variables.

Figure 3. Example of the matrix structure from volunteer Domnarvet11
4.2 Descriptive statistics of the processed GPS data

There were 316 volunteers who made 5,180 tracks with 309,263 positional recordings according to the reorganized data. Table 2 exhibits that the volunteers made at least 1 and at most 66,531 positional recordings during the tracking period. In total 73 single positional recordings that cannot compose a track are deleted. The median number of positional recordings in each track is 79; while the minimum is 2 and the maximum is 95. The number of tracks varies from 1 to 734 and 75% of the volunteers have made less than 17 tracks.

Table 2. Descriptive statistics of positional recordings for tracks and volunteers

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP number per Volunteer</td>
<td>1</td>
<td>7</td>
<td>11</td>
<td>17</td>
<td>734</td>
</tr>
<tr>
<td>Positional Recording Number</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>per Track</td>
<td>2</td>
<td>19</td>
<td>79</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>per Volunteer</td>
<td>2</td>
<td>278</td>
<td>517</td>
<td>809</td>
<td>66531</td>
</tr>
<tr>
<td>Distance (m)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>per Track</td>
<td>2</td>
<td>1481</td>
<td>7837</td>
<td>26921</td>
<td>117722</td>
</tr>
<tr>
<td>per Volunteer</td>
<td>3767</td>
<td>101712</td>
<td>186840</td>
<td>349276</td>
<td>2471518</td>
</tr>
<tr>
<td>Speed (m/s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>per Track</td>
<td>0.01</td>
<td>5.8</td>
<td>11.4</td>
<td>15.3</td>
<td>41.6</td>
</tr>
<tr>
<td>per Volunteer</td>
<td>0.03</td>
<td>11.6</td>
<td>13.3</td>
<td>15.9</td>
<td>25.9</td>
</tr>
<tr>
<td>Time Span (s)</td>
<td>5</td>
<td>5</td>
<td>30</td>
<td>30</td>
<td>342775</td>
</tr>
</tbody>
</table>

The raw time and date were recorded separately in the GPS log file and cannot be used for calculations such as the time span between certain positional recordings or the time differences among tracks. Therefore, the Unix Time Stamp is used to convert the recorded date and time into the number of seconds that have elapsed since 00:00:00 Coordinated Universal Time (UTC), Thursday, 1 January 1970, not counting leap seconds.

The time span between two neighbouring positional recordings was mostly 5 or 30 seconds if the car did not go to a tortuous location (Jia et al., 2012) or stayed at the same location for a long time.
37.7% of the recordings have a time span of 5 seconds and 54.3% have a time span of 30 seconds. The maximum time span was 342,775 seconds. The reason for the very large time span was that if the car has stopped moving but the device was kept on, the tracking would pause. If the number of previous recordings in that track was less than 95, the next positional recording would be added when the car continued to move and tracking started again.

The Euclidean distance between two neighbouring positional recordings in one track is calculated and added together. The sum is the distance of this track in the plane. This measurement of the distance underestimates the real network distance that the car has travelled because it is impossible to always directly drive in straight lines from one location to another when travel on the real road network. The underestimation error could become smaller when the positional recordings are more intensive. However, it is difficult to constantly acquire all positions that the car has covered. The distance measure can be varied in this step and should be chosen carefully in this step considering the application focus and the trade-off between the frequency of the positional recordings and the desired accuracy.

In this specific case, the underestimation could be regarded as acceptable considering that 92% of the positional recordings are tracked with a fairly high frequency of 5 or 30 seconds. There are huge variations in travel distance as is shown in Table 2. The minimum distance for one track was 2 meters while the maximum was 117,722 meters. The total distance that the volunteers had travelled varied from 3,767 meters to 2,471,518 meters.

The recorded instantaneous speed is the speed that the car has at the moment of recording. The average speed of the car on a track segment is calculated by using the distance and the time length between two neighbouring positional recordings. The average speed of a volunteer can be derived in
the same way. A conversion from km/h to m/s is done in order to be consistent with the measurement of distance (m) and time (s). The median of the average speed for all tracks was 11.4 m/s while for all volunteers it was 13.3 m/s.

We randomly select 10 tracks from those 5,180 tracks, and then generate the scatter plot with the linear regression line between the instantaneous speed and the average speed as shown in Figure 4. Most of the points line up in the fairly straight red line, the slope approximately equals to 1 compared to the straight green line. The scatter plot indicates that there is a strong positive linear association between the instantaneous speed and the average speed although the relation is weaker in the low velocities than in the higher ones.

![Figure 4. Differences between the instantaneous speed and the average speed of a track](image)

Purposive locations are positions with drastic changes in time, distance or angle along the movement trajectories of the individual volunteers (Jia et al., 2012). It is understandable that a track consists of purposive locations and this leads to the ambiguous issues in defining tracks. It is reasonable to assume a volunteer to stop and finish a certain activity in about or more than 10 minutes.
Therefore, locations where the time interval exceeds a threshold of 550 seconds are identified and the tracks are thereafter redefined.

If there is no time span over 550 seconds between two neighbouring positional recordings through the whole track, then keep the information of the start and end points, then assign a `TRACK_ID` to this track. If at least one time span over 550 seconds is identified and in addition, the distance between the neighbouring positional recordings is less than 2 km, the old track will then be redefined. As is shown in Figure 5, the time span between positions A and B is larger than 550 seconds; A will be regarded as the end point for the first track while B which happens straight after A will be regarded as the start point for the second track. This original track will then be segmented into two tracks and each track will be assigned a unique `TRACK_ID`. In total 6,534 time spans are identified and there are 8,736 tracks after the redefinition.

![Figure 5. Illustration of redefining a track based on time span between neighbouring positions](image)

5. Processing GPS data on the road network

5.1 Linking positional recordings to tracks

The longitude and latitude of GPS data are referenced by the world geographic coordinate system WGS 84 in the format of degrees decimal minutes. We first convert the WGS 84 degrees decimal minutes into the WGS 84 decimal degree. Considering that the Swedish projected coordinate system SWEREF99_TM is used in the digital map of Dalarna road network from the National Road Database (NVDB) in Sweden, the transformation from the WGS 84 decimal degree to the SWEREF99_TM is then conducted. Figure 6 (a) illustrates the distribution of 309,190 positional recordings from the
volunteers; they are intensive and highly overlapped in the centre area. The small enlarged map in
Figure 6 (a) illustrates how the positional recordings are arranged. Figure 6 (b) illustrates the tracks by
linking the positional recordings sequentially based on the time of occurrence.

![Positional recordings and tracks](image)

*Figure 6. (a) Positional recordings from all the volunteers (b) Tracks from all the volunteers*

### 5.2 Matching positional recordings to the road network

As the device did not continuously track the position every second but rather with 5 or 30 second
intervals, it is hard to examine how the car has moved during this time span. Moreover, a standard
GPS device is usually sensitive to the surroundings. It cannot continuously provide accurate data but
with an error rate of 5 meters according to the manual. Now we define a trip as the link of all the
positional recordings over which the car has travelled on the road network. The previously defined
tracks on the plane as shown in Figure 6 (b) are therefore not identical to the trips of real car movements on the road network.

It is possible to increase the recording frequency and equip more devices on one car to increase the reliability of data. However, that would be problematic due to the increase of control factors. Additional information and post processing techniques provide the ability to improve the current data performance without inducing any data collection uncertainty. As for the individuals’ travel data, the underlying road network provides reliable auxiliary information to verify the data accuracy and improve the usability. The goal is to match the GPS tracking data of the car movements to the real road network by using a map-matching algorithm and a spatial join tool.

Before the matching, we verify that not all positional recordings are on the road. As is shown in Figure 7, there are positional recordings such as a, b, c and d that off the road with a certain distance. Tracks from linking such positional recordings would then cause a deviation from the real trips.

*Figure 7. Positional recordings on the road network*
Figure 8 illustrates the situation after zooming in on the area that has the highest density of the tracks. In Figure 8 (a), it is difficult to see any potential relationship between the data and the road network due to the messy visualization.

Figure 8. (a) The highest density of tracks with the underlying road network before removing all single tracks  
(b) The highest density of tracks with the underlying road network after removing all single tracks

One cause for the messy visualization is that some movements of the cars are far off the road network due to errors of the positional recordings. The errors vary among different devices. If all the positions that occurred at the same location were recorded correctly and were consistent with the road nodes, the tracks would have been highly overlapping. The distance between each track on the same road would have been less than 14 meters considering the width of the present national two-lane road.
A single track is a track that has only been travelled on once by a single volunteer. Single tracks from a volunteer may occur due to the error in positional recordings. They can also be formed by taking unknown shortcuts or illegal paths since the route choice varies among individuals. Usually, drivers would prefer shorter a distance and an easier path due to fuel consumption, travel time and other costs. A driver may take a shortcut only known to him; therefore, he can avoid taking the detour and the tortuous locations. A driver can also be incorrectly guided if he is not familiar with the roads; he could drive into dead-end roads and then have to turn around. Reasons behind this are complex and difficult to identify. We therefore exclude all single tracks which were only conducted by one volunteer and deviated more than 5 meters from the roads, as is illustrated in Figure 8 (b). There are 39 single tracks identified from the whole 5107 tracks thereafter 5068 tracks remain.

5.2.1 Map-Matching

Map-matching is a commonly used approach for correcting off road positions. Brakatsoulas et al. (2005) concluded that global map-matching algorithms produce better matching results than incremental algorithms. While an incremental method runs fast and performs well when sampling frequency is within 5 seconds (Lou et al., 2009). The running time for incremental and global methods is $O(n)$ and $O(mn \log^2 mn)$, where $n$ is the number of positional recordings in a track and $m$ is the total number of edges and vertices in the road network.

Although map-matching will be time consuming with a large GPS data set in a complex road network, improvement for decreasing time complexity and increasing robustness is possible. This is a recommended procedure for processing GPS data since it improves the data performance with showing the spatial geometric and topological structure of movements along the road network.
In this paper 92% of the data have a high sampling frequency of 5 or 30 seconds, running map-matching on all the positional recordings in the whole road network will have high complexity, we therefore applied a global map-matching algorithm by using a small subset of the whole data as an example for illustration. The subset consists of 285 GPS positions, a road network section of 1458 vertices and 677 road segments. Figure 9 illustrates an example of matching an off-road track to the road network, in which Figure 9 a) shows that the off-road track is matched to the road and the correction is shown in Figure 9 b). Figure 9 c) shows the trip after the match.

![Figure 9](image)

*Figure 9. An example of map-matching an off-road track (a) the off road track (b) the map-matching correction (c) the trip after map-matching*

However, map-matching cannot always provide good results (Quddus et al., 2007). As is shown in the real earth image in Figure 10 (a), the positional recordings from the GPS data show that
the volunteer drove through a school building from the start point S to the end point E, which cannot be true. These positional recordings are then matched to the two nearest roads with correction lines as is illustrated in Figure 10 (b). However, this volunteer can never arrive to E from S if he just travelled on the two peculiar road parts as matched. The error of map-matching in such case could be due to the inaccuracy of the GPS data, the insufficiency of the road network map or the inappropriate use of the GPS device. This problem further exhibits the necessity of data processing to uncover the performance of the data. Studies reviewed in this paper have investigated some possible causes and solutions to improve the data performance of reliability, accuracy, availability and integrity.

Figure 10. (a) A number of positional recordings in the earth image (b) The map-matching gave big error in correction
5.2.2 Spatial Join

Another crucial part is to show the spatiotemporal relationships of the tracks, which can be achieved by spatial join (Orenstein, 1986). Spatial join tool is a technical implementation to compile the functionalities of spatial join. The spatial join tool in Arc GIS is one of the spatial geo processing tools that is recommended for showing the features of movements if the datasets are large or complex, or both. In this procedure, 5,068 volunteers’ tracks from 306,664 positional recordings are matched by 3,521 road segments.

Figure 11 illustrates that most of the roads in the centremost area have less than 100 positional recordings, which happened primarily on the local roads or private streets. Roads that have between 101 and 500 positional recordings are the second most common, which take place mainly on the national roads. This is due to the usage and load capacity of the roads; the maximum number of joined positions to a road is 28,818.

Figure 11. Frequency of positional recordings on the road network
Given a tolerance of 5 meters, 90% of the trips on the plane match the road network. It captures the complexity of the real car movements in urban areas. We can further visualize the variation of average velocities on the road network by connecting velocities onto a map. Figure 12 illustrates the variation of speed when cars drive on the roads of the centremost area given a speed limit of 40 km/h. Most of the cars drive within 40 km/h due to the influence of the surroundings, road conditions, speed limit and other restrictions.

![Velocity variation on road network](image)

*Figure 12. Variation of speed on the road network in the centremost area of Borlänge*

6. Conclusion

This paper aims to outline a general procedure for processing GPS tracking data and to discuss all issues related in the processing. The procedure is illustrated step-by-step by processing the real-world GPS data of 300 car movements that predominantly happened in a centre city of Sweden, Borlänge. The procedure provides a detailed understanding of the capability of GPS devices and the output of the data. In addition, post processing techniques with auxiliary information is found necessary and
important for solving the inaccuracy of GPS data. The procedure applies methods to match GPS data with the road network in order to improve the data performance based on a clear definition of movement.

The processed data and the generated maps from the procedure can be used on a broad range of researches and applications. Processing the same or similar data types can provide valuable information to discriminate mobility patterns, derive accurate inference for environmental control, urban planning, location based services and transportation management. It can also provide useful reference information for adjusting and improving the accuracy of the current GPS tracking devices.

In the future, the time threshold for defining the stops within one track could be changed and differences could be compared. The tolerance used in the reduction of the single tracks may also be altered to minimize the induced bias when precise matching is required. Other sensor information like acceleration rate and dilution of precision (DOP) could be useful in processing GPS data. The performance of the procedure could be evaluated by processing GPS data from other types of GPS devices and transportation modes.

References


