Road Network and GPS Tracking with Data Processing and Quality Assessment

Xiaoyun Zhao
Road Network and GPS Tracking with Data Processing and Quality Assessment

Xiaoyun Zhao

Microdata Analysis
School of Technology and Business Studies
Dalarna University
Sweden
2015

ISBN: 978-91-89020-92-4
Abstract

GPS technology has been embedded into portable, low-cost electronic devices nowadays to track the movements of mobile objects. This implication has greatly impacted the transportation field by creating a novel and rich source of traffic data on the road network. Although the promise offered by GPS devices to overcome problems like underreporting, respondent fatigue, inaccuracies and other human errors in data collection is significant; the technology is still relatively new that it raises many issues for potential users. These issues tend to revolve around the following areas: data reliability, data processing and the related application.

This thesis aims to study the GPS tracking from the methodological, technical and practical aspects. It first evaluates the reliability of GPS-based traffic data based on data from an experiment containing three different traffic modes (car, bike and bus) traveling along the road network. It then outline the general procedure for processing GPS tracking data and discuss related issues that are uncovered by using real-world GPS tracking data of 316 cars. Thirdly, it investigates the influence of road network density in finding optimal location for enhancing travel efficiency and decreasing travel cost.

The results show that the geographical positioning is reliable. Velocity is slightly underestimated, whereas altitude measurements are unreliable. Post-processing techniques with auxiliary information is found necessary and important when solving the inaccuracy of GPS data. The densities of the road network influence the finding of optimal locations. The influence will stabilize at a certain level and do not deteriorate when the node density is higher.

**Key words:** GPS tracking, Reliability, Road network, Visualized map, Map-matching, P-median Model, Network density
Contents

List of papers ....................................................................................................................... III

I. Introduction .......................................................................................................................... 1

II. Reliability evaluation (Paper I) .......................................................................................... 4

III. Data processing (Paper II) ................................................................................................ 6

IV. Optimal Location (Paper III) .......................................................................................... 7

V. Conclusion and Future Research ..................................................................................... 8

PAPER I .................................................................................................................................... 12

PAPER II ............................................................................................................................. 42

PAPER III ............................................................................................................................. 69
List of papers

This thesis is based on the work contained in the following papers, referring to by Roman numerals in the text:


Paper II- Zhao, X. (2014). On processing GPS tracking data of spatiotemporal car-movements: a case study. *Journal of Location Based Services* (Submitted for publication)


My contributions to the listed papers were as follows:

Paper I- Experiment design, data collection, data process and analysis, manuscript writing and revising.

Paper II- Data process and analysis, procedure development, manuscript writing and revising.

Paper III- Data process and analysis, manuscript writing and revising.
I. Introduction

Global Positioning System (GPS) is a fast-growing, technologically sophisticated field combined with a satellite navigation system that broadcasts location information (latitude and longitude, speed, heading, altitude, etc.) across the planet. GPS was originally designed for military use; the technology was declassified and released to the public in the year 2000. Much like personal computers, the technology quickly became faster, smaller and cheaper. In less than a decade, GPS technology has spread like wildfire and is used in a wide array of applications. The most common applications have been land, air and marine navigation, and surveying. It has become an integral part of daily life for many individuals and geographic information systems, as well as businesses, construction, resource, environment and agriculture.

GPS technology can be embedded into many portable, low-cost electronic devices nowadays to track the movements of mobile objects. This implication has greatly impacted the transportation field by creating a novel and rich source of traffic data. Wolf (2000) concluded that GPS devices could be used to substitute, rather than supplement, the traditional travel diary. GPS devices have since then become an essential contributor to location-based services and intelligent transportation systems for traffic management and control, transportation routing and planning, as well as transportation policy and travel behavior analysis.

Although the promise offered by GPS devices to overcome problems like underreporting, time inaccuracies, respondent fatigue, and other human errors in data collection is significant, the fact that the technology is relatively new raises many issues for potential users as well. These issues tend to revolve around the following areas: reliability, data processing and the related application of the results.
GPS hardware is evolving rapidly with smaller size, higher compact units and lighter weight to improve the accuracy of data. A key issue in the accuracy of GPS devices is the number of available satellites. Research to-date suggests that, for travel mobility analysis, a GPS device should be capable of simultaneously tracking four or more satellites in order to maintain an acceptable accuracy. As GPS devices become more accurate, efficient, and cost-effective, can it be entirely reliable in real applications? There are shortcomings found in the GPS data, for instance:

- Inaccuracy: Most modern low-cost GPS receivers have a stated accuracy of 5 meters in geographical positioning. This implies a precision in instantaneous speeds calculated from this data to be $\pm 18$km/h, if a 1 second sampling interval is used.
- Complexity: The inaccuracies outlined above mean that for any real useful purpose, complex rules must be imposed when analyzing the data in order to try to reflect the individual’s mobility. Furthermore, the reliability evaluation is more crucial in transportation applications due to the inherent restriction from the road network.

While the reliability of GPS traffic data is influential for its applications in intelligent transportation systems, there is also considerable effort and expense involved in processing the data with detailed information. Specifically, the data processing is required to:

1) format and store raw data tracked by the GPS device;
2) process the data and generate user output, or reformat the raw data for input into other analysis software;
3) provide visualization of the data or link the data to a geographic information system (GIS);
4) map-match the data to a digital road network for correction and analysis;
5) compress the data for storage and retrieval.
This process could be cumbersome and time-consuming, even negating many of the potential benefits offered by GPS. This is particularly important since GPS devices have the potential to generate a significant amount of data with relatively little effort. This potential will be of little use if post-processing becomes too burdensome. Unfortunately, no standard software packages or procedures are available that support all processing of GPS data for transportation studies. What is needed is to outline a general step-by-step process for processing GPS tracking data to visualize the data and for further use. This especially involves filtering outlying positions and matching positions to the road network.

Along with the reliability evaluation and data processing, applying processed and reliable GPS data for mobility analysis suggests that individuals have strong preferences for optimal travel routes along the road network. The location of a travel destination is one crucial factor in determining people’s travel behavior and mobility pattern. The induced effect, such as pollutant emission, traffic congestion and construction change can vary enormously due to the different choices of facility locations, especially in a complex road network. However, it could be troublesome to efficiently find the optimal location of facilities using a specific method (for example \( p \)-median model) for geographically distributed demands in a dense road network. This prompts us to consider the influence of different densities of the road network in choosing the optimal location of a facility.

The main goals of this thesis are therefore the following:

- **Methodological goal:** evaluate the reliability of GPS-based traffic data. This evaluation has been conducted based on data from an experiment containing three different traffic modes (car, bike and bus) traveling along the road network.
- **Technical goal**: outline the general procedure for processing GPS tracking data and discuss related issues that are uncovered. This procedure is carried out by using real-world GPS tracking data of 316 cars.

- **Practical goal**: investigate what the influence of road network density is when finding optimal location. In particular, how does the method $p$-median model perform in a complex and dense road network?

In order to reach these goals, three studies have been done respectively in this thesis. Section 2 summarizes Paper I. Section 3 and Section 4 present Paper II and Paper III. Section 5 summarizes conclusions based on these three studies and proposes possible studies for future research.

### II. Reliability evaluation (Paper I)

The availability and usability of GPS devices in geo-positioning and tracking mobile objects has grown enormously in the past decades and is still increasing. Theiss et al. (2005) identified a wide range of applications of GPS tracking data including timing, logistics, traffic management, and weather forecasting and concluded that it will change the way companies and organizations run their business.

GPS tracking technologies have been applied extensively in transportation studies, in particular, for studying the routes of motorized vehicles (Zito et al., 1995; Quiroga and Bullock, 1998; Murakami and Wagner, 1999). GPS provide detail information to study the travel pattern and prediction of human mobility (Ashbrook et al. 2002, 2003). GPS data can also be applied in environment control. For instance, Jia et al. (2013) studied the induced pollutant emissions of CO2 from car movements by using a GPS tracking data of car movements.
Gathering information of spatial-temporal mobility by GPS device is still subject to critical reflections, even though GPS tracking data opens up for various applications. Leduc (2008) examined recent developments in road traffic data collection and discussed the potentials and bottlenecks related to new GPS technologies. Moreover, Van der Spek et al. (2009) concluded that GPS offers a widely useable instrument to collect invaluable spatial-temporal data on different scales and in different settings, adding new layers of knowledge to urban studies. However, the use of GPS technology and deployment of GPS devices still offers significant challenges for future research. Besides, the enormous use of GPS tracking technologies hinges critically on the functioning of the device.

Nowadays, the internal system of a portable, inexpensive GPS tracking device is designed in a complex way due to the desire for precision and accuracy. Configuration of a GPS device when conducting field tracking is becoming more complicated. How well do the concurrent GPS devices perform in tracking vehicle mobility? To what extent can the accuracy information provided by the manufactures be trusted?

As argued by Shoval (2008), the device can function as an effective and reliable tool for data collection only if it does not affect the nature, quality or authenticity of the data collected. 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, as 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).
Following this, the assessment of the reliability of GPS tracking needs to be scrutinized. This paper examines how well GPS tracking data matches the travelled route for a bike, a car, and a bus, for which the route, the speed, and the altitude are pre-set within the experiment. In the experiment, we vary the type of vehicle, speed, altitude, sampling frequency, and filtering level.

The geographical positioning is found reliable, but with an error greater than postulated by the manufacturer and a non-negligible risk for aberrant positioning. Velocity is slightly underestimated, whereas altitude measurements are unreliable.

### III. Data processing (Paper II)

With the knowledge of how reliable the GPS-based traffic data are, the recorded positions and instantaneous velocities from a portable, low-cost GPS device can be applied with fairly good reliability. However, direct use is limited with the risk of negating valuable information and introducing error. The vulnerability of GPS data needs to be supported by additional information to obtain the desired accuracy, integrity and availability for applications.

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 behavior analysis.

No study has attempted to discuss all issues related to processing GPS tracking data simultaneously, let alone provided a procedure for doing so. This paper aims to address several of the issues arising in processing GPS tracking data and thereby outline a general procedure for the data processing. The study is carried out by using real-world GPS tracking data of 316 cars.
that were originally collected for the purpose of studying CO2-emissions induced by retailing. Descriptive statistics and visualized maps are used to summarize and illustrate the mobility patterns.

This paper confirms that a general procedure in GPS data processing is necessary to have a detailed understanding of the capability of the GPS device and the output of the GPS logger, to generate a clear definition of movement, to visualize the data pattern as well as match the GPS data on the digital network.

IV. Optimal Location (Paper III)

Road network exhibits its key function in the previous two studies. This informs us that in transportation analysis, optimal travel routes are influenced by roads but are determined by destinations. Whether the headed facility is optimally located or not is crucial in route optimization. However, when the road network becomes more complex, finding the optimal location could be troublesome. This paper aims to investigate the density of the road network in influencing the performance of p-median model in finding optimal location of facilities.

The p-median model is a corner-stone in location science. Hakimi (1964) outlined the p-median model in the network space and showed that the optimal solution is found at the nodes of the network (Hakimi, 1965). The objective function is Σ_{qεN} w_q min_{pεP} \{d_{qp} \}, where N is the number of nodes, q and p indexes the demand and the facility nodes respectively, w_q is the demand at node q, and d_{qp} is the shortest network distance between the nodes q and p. Since the p-median problem has been proven NP-hard by Kariv and Hakimi (1969), solutions are generally found by use of some of the many heuristic algorithms proposed in the literature.

Algorithms, spatial aggregation of demand points, and choice of distance measure have been studied extensively. However, few studies have scrutinized the density of the road network with
the \( p \)-median solution. In particular, Han et al., (2013) studied the \( p \)-median solutions when the density of a road network was varied from 500 to 70,000 nodes. For a density beyond some 10,000 nodes, they found a gradual worsening in solutions. This study checks their finding by using a competing heuristic (vertex substitution) and replicating their study. We reject their finding. The solutions stabilize at about 10,000 nodes; they do not deteriorate in higher node density.

This study complements the research of Han et al. (2013) by replicating their study and including an alternative heuristic algorithm to check their surprising finding of poor solutions for very dense networks. This provides a better understanding in optimally locating facilities on the road network where the complexity is continuously increasing nowadays.

**V. Conclusion and Future Research**

This thesis summarizes the assessment of GPS-based traffic data and its related use for human mobility on the road network. The main goals of this thesis are first to evaluate the reliability of the GPS-based traffic data, and then to outline a general procedure for processing this type of data. Based on these works, the third aim of the thesis is to assess the density of a complex road network in influencing the performance of \( p \)-median model on finding optimal locations.

The specific contributions are driven by the goals above, and they include:

Firstly, a well-designed field experiment is conducted to assess the reliability of traffic data based on GPS devices as traffic sensors. No evaluation has been done on traffic modes of car, bus and bike simultaneously. The results show that the geographical positioning is reliable, but it has an error greater than postulated by the manufacturer and a non-negligible risk for aberrant positioning. Velocity is slightly underestimated, whereas altitude measurements are unreliable. This evaluation method can be applied to assess other types of GPS-based traffic sensors as well.
Secondly, the analysis and documentation of general procedures is presented for processing GPS-based traffic data. No study has attempted to discuss all issues related to processing GPS tracking data simultaneously, let alone outline a procedure for doing so. Based on the understanding of the reliability of GPS data, a processing procedure is provided by using real-world GPS tracking data of 316 cars. In particular, post-processing techniques with auxiliary information is found necessary and important when solving the inaccuracy of GPS data.

Thirdly, a connected investigation of optimal locations is studied based on the understanding of the importance of a road network. The solutions of the p-median model of finding optimal locations will stabilize at about 10,000 road nodes; they do not deteriorate when the node density is higher. This could aid in optimizing travel routes and minimizing travel cost as optimal locations of a travel destination is crucial to influence travel behavior.

In future research, further analyses would be required in examining different types of GPS devices in tracking different traffic modes on the road network. A promising research question would be to examine the influence of residential relocation in spatial urban planning based on GPS data. Travel behavior and its induced effects could be one of the focuses. The use of GPS in the sports field to examine the speed and route choices would be an interesting direction, to check the relationship between competition strategy and outcome in the absence of road network restriction.

References


A field experimental evaluation on the reliability of GPS based traffic data

Xiaoyun Zhao
School of Technology and Business Studies at Dalarna University, Sweden

Kenneth Carling
School of Technology and Business Studies at Dalarna University, Sweden

Johan Håkansson
School of Technology and Business Studies at Dalarna University, Sweden

GPS tracking of mobile objects provides spatial and temporal data for a broad range of applications including traffic management and control, transportation routing and planning as well as transportation policy and travel behaviour analysis. Previous transportation research has focused on GPS tracking data as an appealing alternative to travel diaries. Yet, the GPS based data are gradually becoming a cornerstone for real-time traffic management. Tracking data of vehicles from GPS devices are however susceptible to measurement errors – a neglected issue in transportation research. By conducting a randomized experiment, we assess the reliability of GPS based traffic data on geographical position, velocity, and altitude for three types of vehicles: bike, car, and bus. We find the geographical positioning reliable, but with an error greater than postulated by the manufacturer and a non-negligible risk for aberrant positioning. Velocity is slightly underestimated, whereas altitude measurements are unreliable.

Key words: Transportation, GPS tracking device, Reliability, Road network

1 A: Dalarna University, 791 88 Falun, Sweden T: +46 23 77 8509 E: xzh@du.se
2 A: Dalarna University, 791 88 Falun, Sweden T: +46 23 77 8967 E: kca@du.se
3 A: Dalarna University, 791 88 Falun, Sweden T: +46 23 77 8573 E: jhk@du.se
1. Introduction

Global Positioning System (GPS) is a Global Navigation Satellite System (GNSS) for geopositioning. The availability and usability of GPS devices in geo-positioning and tracking mobile objects has grown enormously in the past decades and is still increasing. The GPS has emerged for civilian use in the 1990s as the space geodetic technique being accurate and affordable (Zumberge et al., 1995). Theiss et al. (2005) identified a wide range of applications of GPS tracking data including timing, logistics, traffic management, and weather forecasting and concluded that it will change the way companies and organizations run their business.

GPS tracking technologies have extensively been applied in transportation studies, in particular for studying the routes of motorized vehicles (Zito et al., 1995; Quiroga and Bullock, 1998; Murakami and Wagner, 1999). For instance, Schönfelder (2002) presented an approach to collect GPS longitudinal travel behaviour data on humans and described the complexity of their daily life with the interaction between periodicity and variability. Stopher et al. (2007) demonstrated that GPS can be used successfully to supplement travel diary surveys.

GPS is also applied to study the travel pattern and prediction of human mobility (Ashbrook et al. 2002, 2003). For instance, Jia et al. (2012) confirmed the scaling property and identified the Levy flight characteristic of human mobility by using the GPS tracking data of car movements. GPS data is also applied in environment control. For instance, Jia et al. (2013) studied the induced pollutant emissions of CO2 from car movements by using a GPS tracking data of car movements.

Even though GPS tracking data opens up for interesting applications, gathering information of spatial-temporal mobility by GPS is subject to critical reflections. Leduc (2008) examined recent
developments in road traffic data collection and discussed the potentials and bottlenecks related to new GPS technologies. Moreover, Van der Spek et al. (2009) concluded that GPS offers a widely useable instrument to collect invaluable spatial-temporal data on different scales and in different settings adding new layers of knowledge to urban studies, but the use of GPS-technology and deployment of GPS-devices still offers significant challenges for future research. Besides, the enormous use of GPS tracking technologies hinges critically on the functioning of the device.

Nowadays, the internal system of a portable, inexpensive GPS tracking device is designed in a complex way due to the desire for precision and accuracy. Configuration of a GPS device when conducting field tracking is becoming more complicated. How well do the concurrent GPS devices perform in tracking vehicle mobility? To what extent can the accuracy information provided by the manufactures be trusted? As argued by Shoval (2008), the device can function as an effective and reliable tool for data collection only if it does not affect the nature, quality or authenticity of the data collected.

Following this, the assessment of the reliability of GPS tracking needs to be scrutinized. In this paper, we examine how well GPS tracking data matches the travelled route for a bike, a car, and a bus for which the route, the speed, and the altitude are pre-set within the experiment. In the experiment, we vary the type of vehicle, speed, altitude, sampling frequency, and filtering level.

Section 2 provides a review of research using GPS tracking data with a focus on studies in which the reliability of such data is examined. Section 3 presents the experimental design and the data collection process. Section 4 gives the experimental results. Section 5 ends the paper with a concluding discussion of the findings.
2. Literature review

We have conducted a thorough search for literature relevant to the use of GPS based traffic data. It goes without saying that the use of GPS has penetrated into various transportation applications, such as mobility pattern recognition, vehicle navigation, fleet management, route tracking, and schedule information systems. As a consequence, there is a vast body of studies reporting on applications of GPS with a brief discussion about the reliability of the data. In Table 1 we list most such studies under *Applications*. The discussions contained in these studies do not add any new knowledge to the data reliability and we therefore turn to studies with reliability as the primary concern.

Table 1: Relevant empirical studies on the reliability of GPS based data for transportation

<table>
<thead>
<tr>
<th>Area of research</th>
<th>Research topics</th>
<th>Typical method</th>
<th>Main findings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hardware</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[9], [17], [22], [24], [29], [34], [41], [44-45]</td>
<td>How does the configuration of the hardware affect the precision?</td>
<td>Deductive reasoning</td>
<td>GPS device with advanced technology settings and ideal conditions (open area, many available satellites, and augmentation systems) generates higher accuracy.</td>
</tr>
<tr>
<td></td>
<td>What is the effect of the surroundings?</td>
<td>Laboratory studies</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Observational data for a single device</td>
<td></td>
</tr>
<tr>
<td><strong>Correction methods</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[6-7], [14], [28], [40], [42-43], [48-49], [53-54], [59-60], [65-67]</td>
<td>Map-matching</td>
<td>Theoretical calculation</td>
<td>The inaccurate information acquired from GPS devices can be rectified.</td>
</tr>
<tr>
<td></td>
<td>Differential GPS</td>
<td>Simulation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dead reckoning</td>
<td>Observational data tests</td>
<td></td>
</tr>
<tr>
<td><strong>Empirical assessments</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[4-5], [15], [19], [27], [29], [55], [58], [63], [70], [74], [80-82], [87]</td>
<td>How well are the objects positioned?</td>
<td>Deductive reasoning</td>
<td>The positional accuracy varies from a few centimetres to hundred meters. The error in velocity is 1% or much more.</td>
</tr>
<tr>
<td></td>
<td>Do the recorded velocities coincide with the speedometer?</td>
<td>Laboratory studies</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Observational data for a single device</td>
<td></td>
</tr>
<tr>
<td><strong>Applications</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[2], [8], [10-13], [16], [20-21], [23], [26], [29], [32-33], [35-39], [48-50], [56], [62], [66], [71-72], [74], [76-79], [84-86]</td>
<td>Travel data collection</td>
<td>Field test</td>
<td>Stand-alone GPS devices are helpful and useful in transportation analysis, but vulnerable and need to be combined with auxiliary information for accuracy and integrity</td>
</tr>
<tr>
<td></td>
<td>Vehicle navigation</td>
<td>Post-processing analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fleet management</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Route tracking</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mobility pattern recognition</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: The number refers to the reference in the reference list in the end of the paper.*
Obviously the quality of the hardware and the surroundings where the GPS is being used may affect the reliability of the device. There are some studies that have looked into these aspects (Hardware in Table 1). The starting point is typically that a GPS device requires a clear sight with at least four satellites to determine spatial positions. The accuracy may be enhanced by advanced hardware chipsets, dual-frequency receivers, and carrier-phase measurements supported by augmentation systems (e.g. SBASs, WAAS, EGNOS and MSAS). Under optimal conditions it is possible to have a real-time positional accuracy within decimetres. The required receivers are however too expensive for the use in, for instance, commercial in-car navigation systems. Moreover, optimal conditions do not only call for sophisticated GPS devices, sensors, vehicles, and map information, but also puts requirements on trajectory dynamics and surrounding environment (Skog and Handel, 2009). In urban environments, buildings may partly block satellite signals, forcing the GPS device to work with a poor geometric constellation of satellites, thereby reducing the accuracy of the positional estimates. Multipath propagation of the radio signal due to reflection in surrounding objects may further lead to decreased positional accuracy without notification by the GPS device, thereby reducing the integrity of the navigation solution.

Another strand of the literature presumes erroneous recordings of the GPS device and focuses on methods for correcting the error (Corrections methods in Table 1). The Dead Reckoning (DR) system and map matching algorithms integrated with differential GPS (DGPS) are examples of commonly used hybrid systems for enhancing the positioning of vehicles on land. The DR system can smooth the error of the GPS and provide continuous positioning even in times when the GPS is unavailable (Meng, et al., 2004). The DR produces however an accumulating drift in
the error, but this can be corrected by the DGPS. In situations where an underlying network is
available, map matching has become a popular solution to remedy the inherent error of the GPS.
In essence, map matching is to use a digital map of the road network to impose constraints on the
GPS navigation and tracking recordings (Skog and Handel, 2009). Quddus et al. (2007) reviewed
the currently existing map-matching algorithms and their limitations. Map-matching has been
predominantly applied in post-processing GPS data (e.g., Marchal et al., 2005; Schüssler and
Axhausen, 2009a, 2009b).

The literature most relevant to our study is the studies attempting to assess the reliability of
GPS data by comparing them to known conditions (Empirical assessments in Table 1). These
studies aims at evaluating the reliability of GPS device, but are typically not conducted as
experiments. They examine one transportation mode, one environment, one aspect of tracked
information, and one configuration of the device. The studies are also examining the static
accuracy using small samples without controlling for external condition. A notable exception is
the recent work of Schipperijn et al. (2014). They tested the dynamic accuracy of a GPS device
(Qstarz Q1000XT portable GPS receiver) for the use in public health applications under varying
real-world environmental conditions, for four modes of transportation, and at three levels of
sampling frequency. They found that not even a half of the positional recordings were within 2.5
meters of the actual position with the proportions varying by travel mode and area.

As claimed by Schipperijn et al. (2014), mobile objects in free-living studies are likely to
move dynamically. It is therefore vital to know the dynamic accuracy for various travel modes in
changing surroundings. However, Schipperijn et al. (2014) only studied the influence on
positional accuracy by changing the sampling frequency of the GPS device neglecting other
factors possibly affecting the accuracy. To conclude, the number of studies that have evaluated the reliability of standard GPS devices employing different configurations for tracking various types of vehicles on real road networks is limited.

3. Experimental design and data collection

We want to examine how well GPS tracking data matches an actual route travelled. Vehicles are in focus for this study and we therefore assume them being restricted by an underlying road network. We consider the vehicles bike, car, and bus being the dominating means of private transportations. In the experiment, the vehicles travel on pre-set routes of known geographical position and altitude with speeds decided in advance. While they are travelling their mobility is being tracked by a GPS device.

For the experiments, a standard and integrated GPS device that could be broadly used in different vehicles under various circumstances is preferable. Smart phone with GPS application restricted to cellular network or wireless network is therefore not considered. Other important features in selecting the device are that the device is user friendly, easy to operate and has a durable battery. BT-338 (X) was finally chosen after a survey in the product market, this device is a combination of a GPS receiver and a data logger\(^4\). According to the manufacturer, the device should provide a geographical positioning within an error of 5 meters and a measurement error of velocity less than 0.4 km/h. The manufacturer makes no claims about the precision in the measurement of altitude.

Figure 1 illustrates the interface in configuring the device with regard to some of the factors in the experiment. We consider intensive sampling by the device with measurements every one and five seconds as well as sampling every 30 seconds. Note that the latter implies that some of the vehicles will easily travel more than 500 meters between recordings. Such setting implies a coarse assessment of the vehicle’s mobility pattern. Hence, the levels of sampling frequency represent both dense and sparse data. We set the data logging format to track position, time, date, speed, and altitude. The WAAS/EGNOS/MSAS feature is enabled to acquire more precise position as suggested by the manufacturer. We consider both enable and disable data logging when distance is less than the selected radius 20 meters.

Table 2 illustrates the factors and corresponding levels in the experimental design. We are in possession of 15 identical GPS devices with a unique identifying number. They are randomly assigned to one of three groups of equal size for which the sampling interval is set to 1, 5, and 30
seconds respectively. In each group two randomly selected devices have the data logging disabled if distance is less than the radius of 20 meters.

Table 2: Experimental design of collecting GPS tracking data

<table>
<thead>
<tr>
<th>Sampling Interval</th>
<th>1s</th>
<th>5s</th>
<th>30s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device No.</td>
<td>3 29</td>
<td>37</td>
<td>36</td>
</tr>
<tr>
<td>Distance Restriction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bicycle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70km/h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bus</td>
<td>80-100km/h</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

On the bike, all the 15 devices are carried by the rider in a backpack. Moreover, the devices are in the backpack in the back seat of the car while the backpack is kept in the front seat of the bus. The data collection of the bike and the car is undertaken in Borlänge in Sweden. The data collection of the bus is undertaken along the bus line 151 between Borlänge and its neighbouring city Falun.

It was difficult to fix the velocity of the bus in advance as would be preferable. The velocity varied along the scheduled route due to the traffic and the behaviour of the drivers. For this reason, only a segment of the route, where the velocity varied smoothly between 80 km/h and 100 km/h, was used for GPS tracking. Meanwhile the bus trip was filmed. The bike followed a strict setting of velocities ranging from 15-50 km/h in six levels. For the car, velocities of 15-70
km/h were considered. Travel diaries were used to note unexpected changes in route, velocity, and emergent situation. The bike was ridden by the same rider and the driver of the car was the same throughout the experiment.

Data for the bike was collected at noon in order to reduce the risk of deviation from the protocol caused by other people on the route. Likewise, data collection for the car was undertaken between 3 and 4 in the afternoon to avoid peaks in the traffic. The data collection for the bus was conducted after 6 in the afternoon thereby minimizing the variation in velocity due to people waiting at bus stop. The data collection took part on a cloudy summer day with an air temperature of about 22 degrees and almost no wind.

An accurate speedometer of the vehicles is essential for the experiment. To ensure this we first considered the speedometer of the bike. The speedometer works by counting the wheel revolutions per time unit adjusted by the circumference of the tire. Crucial for the accuracy is the measurement of the circumference. The tires were inflated immediately prior to the experiment and the circumference was measured by two different tape measurers. Thereafter we calibrated the car speedometer by riding the bike and driving the car side by side and recording the speeds simultaneously. We checked the relationship between the recordings from the bike speedometer and the car speedometer by means of linear regression: $Car = 1.0385 \times Bike$. The relationship is strong with a correlation of 0.998. The speedometer of the car was adjusted accordingly in the experiment.

The routes for the experiment were chosen having the need for maintaining a constant velocity in mind. In the choice of routes, we tried to avoid places where the GPS signal was likely to be disturbed. This means that the routes do not pass high buildings, strong magnetic fields or are in
valleys. As for the car, we also needed to consider the speed limits of the roads while a bike may be ridden at any speed on a bike path.

Figure 2(a) depicts the route for the bike with arrows indicating the riding direction. The route is about 2 kilometres and it is a paved bike path. The route was used consecutively for each velocity at a time. For instance, at the velocity of 20 km/h the route took 6 minutes meaning that there could be 360, 72, and 12 recordings per GPS device for the three levels of sampling frequency. The variation in altitude of the route is only a few meters.

Figure 2: (a) The bike route; (b) The car route; (c) The bus route
Figure 2(b) depicts the route for the car with arrows showing the directions. The route is segmented by colour representing the attained velocity. The route was travelled several times to ensure sufficiently many recordings per cell in the experimental design. The range in altitude is 40 meters. Maintaining a constant velocity with a car in an ordinary traffic situation is of course difficult. The circles in figure 2(b) represent segments identified in advance as impossible to maintain the speed due to intersections and speed bumps. Afterwards the experimental recordings, pertaining to segments where the intended velocity was not met according to the travel diary, were removed. Figure 2(c) depicts the bus route. This route has a variation in altitude with a range of 37 meters.

All the GPS devices were turned on before initiating the data collection. The reason was that there is acquisition time for the device to start recording. The original GPS tracking data were kept into DataLogger files. The files may be loaded from the device to a computer by using the software Global Sat Data Logger PC Utility. We retrieved the data directly after the experiment was completed. The device number 4 was malfunctioning and did not record any data. The other 14 devices worked well and we obtained in total 25,901 recordings of the car, 9,224 recordings of the bike, and 8,688 recordings of the bus.

As a final remark we note that there is a trade-off between sampling interval and battery lifespan (Ryan et al., 2004). We checked whether the duration of the battery of the device differed for various settings of the sampling interval. The check was conducted by randomly selecting 6 of the GPS devices and letting 3 of them with intervals 1, 5, and 30 seconds and letting the other 3 of them with intervals 1, 5, and 30 seconds and data tracking within 20 meters.
distance radius disabled. It turned out that the duration of the battery was unrelated to these two factors.

4. Experimental results

We begin by examining the positional reliability, followed by examining the reliability of velocity and end with a check on the measurement of altitude obtained from the GPS device.

4.1 Geographical positioning

The geographical positions of the mobile object are necessary to identify the objects trajectory. In the experiment the trajectory of the vehicles is known by the road network and its digital representation. The location and the trajectory of a car are restricted by the road network (Skog and Handel, 2009). As a statistics to assess the reliability of the geographical positioning obtained from the GPS device we measure the concordance of the recordings and the road network. Ideally the positional recordings should be on the underlying road network.

Figure 3 shows by an example some of the positional recordings on the road network. The green circles indicate the recordings that match the road network. The yellow circles indicate recordings on the edge of the road network, by us regarded as matching the road network well enough. The red squares indicate inaccurate recordings off the road network. In this example, 8 of the 42 recordings failed in giving an accurate position of the car. The width of the road is 14-

---


6 The road network is provided by the National Road Data Base (NVDB) and is operated by the Swedish Transport Agency. NVDB classifies the road network into national roads, local roads and streets. The national roads are owned by the national public authorities. The local roads and streets are built and owned by municipalities or companies or private persons. The positional accuracy of the road segments used in this study is within 0.2 meter.
20 meters meaning that an error of 5 meters is tolerated even if one considers that the car was not driven in the middle of the road.

A bike-path in NVDB is represented by a line, not a polygon, although its width is 3.5 meters according to the department of motor vehicles in Sweden. In assessing the positional recordings of the bike to the underlying road network we allowed for a tolerance distance of 5 meters.

Table 3 gives the proportion of positional recordings that match the road network. Considering that the manufacturer of the GPS device claims that the error in positioning is at the most 5 meters, it is to be expected that almost all recordings should match the road network. This is generally not the case. 75% to 90% of positional recordings for the bus are accurate. The positioning of the car was more reliable with about 90% of the recordings being accurate. As for the bike, the recordings frequently fail to identify its travel on the network.
Table 3: Proportion of positional recordings matching the road network

<table>
<thead>
<tr>
<th>Factors</th>
<th>Vehicle</th>
<th>Bike</th>
<th>Car</th>
<th>Bus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1s</td>
<td>Distance radius 0m</td>
<td>60.06%</td>
<td>94.97%</td>
<td>75.75%</td>
</tr>
<tr>
<td></td>
<td>Distance radius 20m</td>
<td>68.24%</td>
<td>91.15%</td>
<td>77.21%</td>
</tr>
<tr>
<td>5s</td>
<td>Distance radius 0m</td>
<td>54.90%</td>
<td>87.33%</td>
<td>75.29%</td>
</tr>
<tr>
<td></td>
<td>Distance radius 20m</td>
<td>26.69%</td>
<td>93.27%</td>
<td>74.42%</td>
</tr>
<tr>
<td>30s</td>
<td>Distance radius 0m</td>
<td>73.00%</td>
<td>92.15%</td>
<td>80.95%</td>
</tr>
<tr>
<td></td>
<td>Distance radius 20m</td>
<td>91.18%</td>
<td>92.86%</td>
<td>90.00%</td>
</tr>
</tbody>
</table>

As an overall finding drawing on Table 3, there is no clear pattern emerging from the factors considered in the experiment. Possibly the longest sampling interval tends to lead to better positioning, the device generally gives higher accuracy in positioning for the car but tends to have large variation on bike. However, we have noted a serial correlation of the recordings implying that an inaccurate recording is likely to be followed by another if the time interval is short.

Figure 4: Bike routes in the secondary experiment
The surprising results for the bike prompted us to run a secondary experiment. We speculated that the positional recordings of the bike were interfered by the surrounding environment. Figure 4 depicts the two routes travelled by the bike at a second occasion. One route coincides with the route used in the original experiment while the second route is a part of the car’s route.

In the first experiment, we had numerous inaccurate recordings in the three areas depicted in Figure 4 by a white circle and two triangles. The circled area is nearby power lines to the north. The areas indicated by triangles have trees with a height of 8-10 meters. In the secondary experiments all settings of the GPS devices were kept as in the first experiment, but the bike travelled both routes at a speed of 20 km/h.

Table 4 gives the proportion of accurate recordings on the two routes. Although the proportion of accurate recordings on the original bike route is higher in the second experiment, it is still rather low. Again most inaccurate recordings happened at the three areas previously identified as problematic. The positional recordings on the car’s route were substantially better. This exercise illustrates that the GPS device may generate (infrequent) errors due to the interferences with the surroundings such as trees and built-ups in a non-obvious way (Modsching et al., 2006).

<table>
<thead>
<tr>
<th>Factors</th>
<th>Original bike’s route</th>
<th>On the car’s route</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Distance radius 0m</td>
<td>73.83%</td>
</tr>
<tr>
<td></td>
<td>Distance radius 20m</td>
<td>58.79%</td>
</tr>
<tr>
<td>1s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5s</td>
<td>Distance radius 0m</td>
<td>50.38%</td>
</tr>
<tr>
<td></td>
<td>Distance radius 20m</td>
<td>69.29%</td>
</tr>
<tr>
<td>30s</td>
<td>Distance radius 0m</td>
<td>71.49%</td>
</tr>
<tr>
<td></td>
<td>Distance radius 20m</td>
<td>80.13%</td>
</tr>
</tbody>
</table>
4.2 Estimating the velocity

It goes without saying that it is more difficult to estimate a changing velocity than a constant velocity. Drivers (and riders) need to adjust their speed in line with the traffic but also at intersections, roundabouts, tortuous locations, and traffic lights (Jia et al., 2012). This is also true in conducting an experiment of this kind. We used the travel diary of the car and the bike to delete recordings where the intended constant velocity was not possible to maintain. As for the bus, the films were used for deleting recordings where the velocity was not constant. Figure 5 illustrates how the recorded velocity varies around the pre-set constant velocity.

![Figure 5: Recorded velocity versus actual velocity as measured by one GPS device for the car](image)

The figure shows the recordings from one device in the car where the device was set to record the velocity in intervals of 30 seconds. There is a tendency that the recorded velocity is generally lower than the actual velocity. Recall that the manufacturer claimed that the error in velocity
should be within 0.4 km/h. Table 5 further shows the statistics for the recorded velocity as the average, the standard deviation, and the root mean square error (MSE). The velocity is underestimated by about 5% and the standard deviation exceeds by far 0.4 km/h. The relative error in the recorded velocity seems not however to be related to the setting of the GPS device.

Table 5: Statistics of recorded velocity for bike, car and bus

<table>
<thead>
<tr>
<th>Signal Interval</th>
<th>1s</th>
<th>5s</th>
<th>30s</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distance Restriction</strong></td>
<td><strong>Mean</strong></td>
<td><strong>SD</strong></td>
<td><strong>v(mse)</strong></td>
</tr>
<tr>
<td><strong>Bicycle</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15km/h</td>
<td>14.13</td>
<td>0.90</td>
<td>1.25</td>
</tr>
<tr>
<td>20km/h</td>
<td>19.73</td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td>30km/h</td>
<td>29.01</td>
<td>0.91</td>
<td>1.34</td>
</tr>
<tr>
<td>40km/h</td>
<td>38.54</td>
<td>1.37</td>
<td>2.00</td>
</tr>
<tr>
<td>50km/h</td>
<td>43.82</td>
<td>0.52</td>
<td>1.29</td>
</tr>
<tr>
<td><strong>Car</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15km/h</td>
<td>15.34</td>
<td>0.69</td>
<td>0.77</td>
</tr>
<tr>
<td>20km/h</td>
<td>19.45</td>
<td>0.94</td>
<td>1.09</td>
</tr>
<tr>
<td>30km/h</td>
<td>29.33</td>
<td>1.37</td>
<td>1.52</td>
</tr>
<tr>
<td>40km/h</td>
<td>30.60</td>
<td>1.91</td>
<td>1.67</td>
</tr>
<tr>
<td>50km/h</td>
<td>43.36</td>
<td>0.93</td>
<td>1.89</td>
</tr>
<tr>
<td>60km/h</td>
<td>48.21</td>
<td>0.95</td>
<td>2.03</td>
</tr>
<tr>
<td>70km/h</td>
<td>57.40</td>
<td>0.91</td>
<td>2.76</td>
</tr>
<tr>
<td><strong>Bus</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80km/h</td>
<td>79.18</td>
<td>2.13</td>
<td>2.27</td>
</tr>
<tr>
<td>90km/h</td>
<td>84.80</td>
<td>1.40</td>
<td>1.41</td>
</tr>
<tr>
<td>100km/h</td>
<td>88.62</td>
<td>1.47</td>
<td>2.02</td>
</tr>
<tr>
<td><strong>15km/h</strong></td>
<td>94.95</td>
<td>1.28</td>
<td>1.27</td>
</tr>
</tbody>
</table>

We have conducted analysis of variance (ANOVA) to formally test for the factors. The error between the recorded and actual velocity was the response variable. The error increased with the velocity. There was no significant difference for whether the distance restriction was on or off. The sampling interval was unrelated to the error, except for the recordings of the bike. In this case the longer sampling interval was associated with a (marginal) increase in the error.

We also checked for a relationship between the error in velocity and the geographical error as discussed in section 4.1. We did so by labelling all positional recordings on the road network as
accurate and all those off the road network as inaccurate. Thereafter we repeated the ANOVA including the factor Accurate in the model. It was strongly significant suggesting a greater underestimation of the velocity if the positional recording was inaccurate.

4.3 Altitudes

The GPS device is presumably able to record the altitude of the vehicle as it travels. However, the manufacturer is not specific about the precision in the recorded altitude. We expect the precision of altitude to be poorer that the geographical position considering for instance the requirement for connection to additional satellites for estimating altitude.

In order to check the precision in the recorded altitude, we first acquired the geo-information of altitude in Borlänge from the national altitude database (NNH)\(^7\). We applied spatial join in Arc GIS 10.1 to join the attribute table of the actual altitude layer to the attribute table of the recorded altitude layer. Each position of the vehicle where a recorded altitude occurred is related to the nearest point in the actual altitude layer. The maximum distance between the position of the recording and the actual altitude layer is 21 meters. This is an inconsequential approximation as the road network covered in the experiment does not contain any steep up- and down-hills. Another (trivial) approximation is the fact that the devices were carried by the rider in a backpack, in the back seat of the car, and in the front seat of the bus. Hence, the altitude of the devices was 1-2 meters above the level of the road network.

The error in recorded altitude with respect to the actual altitude is large. Most of the time the error was within the range of -50 meters and 50 meters, but frequently the error exceeded 100

---

\(^7\) The altitudes data is provided by Sweden’s Mapping, Cadastral and Land Registration Authority (www.lantmateriet.se). The altitude model is made by laser scanning and has an average altitude error of 0.1 meter and 0.4 meter in the plane.
meters. Considering for instance that the bike path travelled in the experiment was essentially flat such a magnitude in error is enormous.

Moen et al. (1996) discussed the concepts of 2-D and 3-D fix and argued that a 3-D fix should offer a greater precision in estimating the altitude. The GPS device used in the experiment generates a 3-D fix. All the same, the results are not impressive.

5. Concluding Discussion

This paper focuses on a method for evaluating the reliability of portable, standard GPS devices in tracking vehicles. The experiment was conducted by equipping a GPS tracking device BT-338(X) on vehicles being car, bike, and bus and then track the geographical position, velocity, and altitude of the vehicles in the road network. Pre-processing and cleaning of the data was necessary and auxiliary information needed.

The GPS tracking data identified the actual positions of the vehicles fairly successfully. The surroundings of the experiment had no obviously interfering attributes like high built-ups, forests, magnetic fields, and so on. The partially poor identification of the bike’s positions by trees and in the vicinity of magnetic fields shows however that the positional error of the GPS is highly vulnerable to the surroundings (see also Modsching et al., 2006). Fortunately, this problem can be rectified by using map-matching algorithms as proposed by Brakatsoulas et al. (2005), Taylor et al. (2006), and Quddus et al. (2007). We believe that the GPS tracking data on position may be useful for routing, mobility patter recognition, and general navigation of travellers. However, the accuracy is insufficient in cases requiring high geographical precision such as parking, emergency rescue, and the like.
The tracked instantaneous velocities are quite accurate with a tendency of underestimation. The error between recorded velocity and actual velocity is monotonically increasing with the speed. It should however be noted that we did not study the accuracy regarding acceleration and deceleration which are common phenomena in ordinary traffic.

Concerning the recorded altitudes in the tracking data, we found it to be highly inaccurate and we suggested disregarding this parameter in practical use until further investigations.

The reliability seems to be unrelated to the sampling frequency. Of course, intensive positional recordings provide more details regarding the mobility pattern. However, it comes at the expense of more aggressive data rendering communication, storage, data processing, data mining, and data analysis more costly. Balancing between these aspects is necessarily specific to the domain of application.

There is drawback of GPS devices due to a short effective lifespan (Ryan et al., 2004). The data collection part of the experiment in this paper lasted at the most for two hours; the duration of the device was not a concern here as the operational time for the device is about 11 hours after being fully charged and in continuous mode. However, the lifespan may be a costly drawback in full-scale applications.

Finally, this study examined one specific standard GPS device. It would be interesting in the future to conduct further analyses including other types of GPS device by using the experimental method outlined in this study.
References

A field experimental evaluation on the reliability of GPS based traffic data


A field experimental evaluation on the reliability of GPS based traffic data


A field experimental evaluation on the reliability of GPS based traffic data


Zhao, Carling and Håkansson
A field experimental evaluation on the reliability of GPS based traffic data


A field experimental evaluation on the reliability of GPS based traffic data


A field experimental evaluation on the reliability of GPS based traffic data


On processing GPS tracking data of spatiotemporal car movements: a case study

<table>
<thead>
<tr>
<th>Journal</th>
<th>Journal of Location Based Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>TLBS-2014-0050</td>
</tr>
<tr>
<td>Manuscript Type</td>
<td>Research Paper (Under Review)</td>
</tr>
<tr>
<td>Keywords</td>
<td>GPS tracking data, visualized map, road network, map-matching</td>
</tr>
</tbody>
</table>

URL: http://mc.manuscriptcentral.com/tlbs E-mail: jlbs@tendf.co.uk
On processing GPS tracking data of spatiotemporal car-movements: a case study

Xiaoyun Zhao

Micro-data analysis, School of Technology and Business Studies, Dalarna University, Sweden

Abstract

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, velocity 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 velocity, 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 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. 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.

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, agriculture and sports (Steiner et al., 2000; Tuner et al., 2000; Cagnacci et al., 2000; Stafford, 2000; Auernhammer, 2001; Zhang et al., 2002; Coutts and Duffield, 2010; Aughey, 2011). Common to these research areas is that there is no underlying 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). In this paper, the focus is limited to GPS tracking data on a road network. In this area of research, there are three broad aspects that have been of concern.

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, the analysis of human mobility and travel behaviour GPS data over a certain period of time 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 in car, bus or foot 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 predicting 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.

3. Data Collection

The data collection was conducted by using a type of standard Blue-tooth 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 Blue-tooth 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 velocity. 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.

Figure 1. Spatial distribution of the volunteers as well as all the residents in Borlänge

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, velocity 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 ttmmss. The date is in the format of DDMMYY. The velocity 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.

![Figure 2. Example of GPS data from volunteer Domnarvet11](image-url)
The first track contains 16 positional recordings with numerators from 1 to 16. Specifically, 1=6029.8968, 1528.2927, 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 velocity 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>
<thead>
<tr>
<th>Sport Association</th>
<th>Valid GPS Logger files</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.
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.

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

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRACK_ID</td>
<td>PR_ID</td>
<td>USER_ID</td>
<td>LATITUDE</td>
<td>LONGITUDE</td>
<td>TIME</td>
<td>DATE</td>
<td>VELOCITY</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Domnarvet11</td>
<td>60298906</td>
<td>15232275</td>
<td>182027</td>
<td>549111</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>Domnarvet11</td>
<td>60299726</td>
<td>15238763</td>
<td>182057</td>
<td>549111</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>Domnarvet11</td>
<td>60301304</td>
<td>15239746</td>
<td>182127</td>
<td>549111</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>Domnarvet11</td>
<td>60302047</td>
<td>15301304</td>
<td>182157</td>
<td>549111</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>Domnarvet11</td>
<td>60302738</td>
<td>15308870</td>
<td>182227</td>
<td>549111</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>6</td>
<td>Domnarvet11</td>
<td>60302546</td>
<td>15314649</td>
<td>182257</td>
<td>549111</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>7</td>
<td>Domnarvet11</td>
<td>60304964</td>
<td>15317971</td>
<td>182327</td>
<td>549111</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>8</td>
<td>Domnarvet11</td>
<td>60307263</td>
<td>15325359</td>
<td>182357</td>
<td>549111</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>9</td>
<td>Domnarvet11</td>
<td>60306866</td>
<td>15324840</td>
<td>182427</td>
<td>549111</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>10</td>
<td>Domnarvet11</td>
<td>60305765</td>
<td>15324637</td>
<td>182457</td>
<td>549111</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>11</td>
<td>Domnarvet11</td>
<td>60306312</td>
<td>15328032</td>
<td>182527</td>
<td>549111</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>12</td>
<td>Domnarvet11</td>
<td>60307001</td>
<td>15335369</td>
<td>182557</td>
<td>549111</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>13</td>
<td>Domnarvet11</td>
<td>60308224</td>
<td>15338033</td>
<td>182627</td>
<td>549111</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>14</td>
<td>Domnarvet11</td>
<td>60307548</td>
<td>15341480</td>
<td>182657</td>
<td>549111</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>15</td>
<td>Domnarvet11</td>
<td>60306057</td>
<td>15342688</td>
<td>182727</td>
<td>549111</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>16</td>
<td>Domnarvet11</td>
<td>60304570</td>
<td>15345359</td>
<td>182757</td>
<td>549111</td>
</tr>
</tbody>
</table>

Figure 3. Example of the matrix structure from volunteer Domnarvet11
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.

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><strong>Volunteer TP number</strong></td>
<td>1</td>
<td>7</td>
<td>11</td>
<td>17</td>
<td>734</td>
</tr>
<tr>
<td><strong>Positional Recording Number</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Track</td>
<td>2</td>
<td>19</td>
<td>79</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>Volunteer</td>
<td>2</td>
<td>278</td>
<td>517</td>
<td>809</td>
<td>66531</td>
</tr>
<tr>
<td><strong>Distance (m)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Track</td>
<td>2</td>
<td>1481</td>
<td>7837</td>
<td>26921</td>
<td>117722</td>
</tr>
<tr>
<td>Volunteer</td>
<td>3767</td>
<td>101712</td>
<td>186840</td>
<td>349276</td>
<td>2471518</td>
</tr>
<tr>
<td><strong>Velocity (m/s)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>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>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><strong>Time Span (s)</strong></td>
<td>5</td>
<td>5</td>
<td>30</td>
<td>30</td>
<td>342775</td>
</tr>
</tbody>
</table>

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. As is illustrated in Figure 4, this measurement of the distance underestimates the real distance that the car has travelled on the road network. The underestimation error could become smaller when the positional recordings are more intensive. It is easy to measure the Euclidean distance; however, it is difficult to constantly acquire all positions that the car has covered considering the trade-off between the frequency of the positional recordings and the accuracy of the distance measurements.

Moreover, 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.

![Figure 4. Differences between the Euclidean distance and the Network distance](image)

The recorded instantaneous velocity is the velocity that the car has at the moment of
recording. The average velocity of the car on a track segment is calculated by using the
distance and the time length between two neighbouring positional recordings. The average
velocity 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 velocity 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 velocity and the average velocity as
shown in Figure 5. 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 velocity and the
average velocity although the relation is weaker in the low velocities than in the higher ones.
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. 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 6, 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 6. Illustration of redefining a track based on time span between neighbouring positions

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 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. The transformation from the WGS 84 decimal degree to the projected coordinate system SWEREF99_TM is then conducted; because the SWEREF99_TM is used in the digital map of Dalarna road network from the National Road Database (NVDB) in Sweden.

Figure 7 (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 7 (a) illustrates how the positional recordings are arranged. Figure 7 (b) illustrates the tracks by linking the positional recordings sequentially based on the time of occurrence.
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 7 (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 8, 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 8. Positional recordings on the road network](image)

Figure 9 illustrates the situation after zooming in on the area that has the highest density of the tracks. It is difficult to see any potential relationship between the data and the road network due to the messy visualization.
Figure 9. The density of tracks with the underlying road network before removing all single tracks

One cause for the messy visualization in Figure 9 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 overlapped. 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.

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. The result is illustrated in Figure 10.

![Map Matching](image)

**Figure 10.** The highest density of tracks with the underlying road network after removing single tracks

### 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 sampling frequency of 5 or 30 seconds, considering the time complexity, we applied a global map-matching algorithm with a subset of 285 GPS positions and a road network section of 1458 vertices and 677 road segments. Figure 11 illustrates an example of matching an off-road track to the road network, in which Figure 11 a) shows that the off-road track is matched to the road and the correction is shown in Figure 11 b). Figure 11 c) shows the trip after the match.

Figure 11. 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
5.2.2 Spatial Join

Another crucial part is to show the spatiotemporal constraint of the tracks. A Spatial Join tool such as in Arc GIS is one of the geo processing tools that are recommended for showing the features of movements if the datasets are large or complex, or both. In this procedure, 5,071 volunteers’ tracks from 306,664 positional recordings are matched with 3,521 road segments.

Figure 12 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 12. Frequency of positional recordings on the road network](image_url)

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 13 illustrates the variation of velocity 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.

Figure 13. Variation of velocity 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. 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 a reference 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


A note on network density and \( p \)-median solutions

Xiaoyun Zhao\(^*\), Kenneth Carling, Johan Håkansson

Abstract

The \( p \)-median model is commonly used to find optimal location of facilities for geographically distributed demands. So far, only a few studies have considered the importance of the road network for the model. In particular, Han, Håkansson, and Rebreyend (2013) studied the \( p \)-median solutions when the density of a road network was varied from 500 to 70,000 nodes. For a density beyond some 10,000 nodes, they found a gradual worsening in solutions. This study checks their finding by using a competing heuristic (vertex substitution) and replicating their study. We reject their finding. The solutions stabilize at about 10,000 nodes; they do not deteriorate in higher node density.

Keywords: \( P \)-median Model, Vertex Substitution, Simulated Annealing, Dense Network

1. Research question

The \( p \)-median model is a corner-stone in location science. Hakimi (1964) outlined the \( p \)-median model in the network space and showed that the optimal solution is found at the nodes of the network (Hakimi, 1965). The objective function is \( \Sigma_{q \in N} w_q \min_{p \in P} \{d_{qp}\} \), where \( N \) is the number of nodes, \( q \) and \( p \) indexes the demand and the facility nodes respectively, \( w_q \) is the demand at node \( q \), and \( d_{qp} \) is the shortest network distance between the nodes \( q \) and \( p \). Since the \( p \)-median problem has been proven NP-hard by Kariv and Hakimi (1969), solutions are generally found by use of some of the many heuristic algorithms proposed in the literature.

\(^*\) Xiaoyun Zhao is a PhD-student in Micro-data analysis and corresponding author: xzh@du.se, phone: +46 23-778509. Kenneth Carling is a professor in Statistics and Johan Håkansson is a professor in Geography. All are at the School of Technology and Business Studies at Dalarna University, Sweden
Algorithms, spatial aggregation of demand points and choice of distance measure have been studied extensively. However, few studies have scrutinized the density of the road network with the $p$-median solution.

Consider Figure 1 as an illustration of the issue. Figure 1a shows the national road network in the Swedish region Dalarna. It is a sparse network of 5,437 kilometres with 1,548 nodes. Figure 1b, on the other hand, also imposes the local streets and subsidized private roads that open to private and commercial transportation use. This represents a dense road network with an extension of 20,240 km and 67,020 nodes. In a practical location problem, is the dense network always warranted for a better $p$-median solution?

Figure 1. Map of Dalarna region showing (a) national road system, and (b) national road system with local streets and subsidized private roads. Source: Carling, et al. (2012b).

Han et al. (2013) set out to answer this question by analysing how the solutions to the $p$-median problem changed when the density of the network was varied. In Figure 2 the solution at a given density (i.e. number of candidate nodes) is compared to the best solution ever found in the network for a location problem of 5, 10, 20 and 40 facilities. Surprisingly, the solutions are not monotonically improving with a denser network. In fact, Figure 2
suggests solutions to be poor in a very dense network. Han et al. (2013) solely used simulated annealing in their experiments. Consequently, the non-monotonic function depicted in Figure 2 may not only be due to the density of the road network, but also the performances of the algorithm.

This study complements the research of Han et al. (2013) by replicating their study and including an alternative heuristic algorithm to check their surprising finding of poor solutions for very dense networks.

Figure 2. Excess in distances (per cent) compared to the best solution in the network. Number of nodes (x-axis) and the relative difference between solution for a given number of nodes and the best solution in the network as \( \left( \frac{|\text{current solution} - \text{best solution}|}{\text{current solution}} \right) \times 100\% \) (y-axis). Source: Han, M. et al. (2013).

2. Vertex Substitution (T&B)

Optimal solutions to large combinatorial problems such as the \( p \)-median problem are difficult to obtain (Al-khedhairi, 2008). In a pre-work, we investigated the performances of four algorithms; Greedy Search, Vertex Substitution, Lagrangian Relaxation and Simulated Annealing. All of them solved the \( p \)-median problem for the data of this study in the cases of 7 and 11 facilities. However in the following, we use Vertex Substitution as it consistently outperformed the other three competitors. To ensure that we replicate the study of Han et al.
(2013), we also employed Simulated Annealing (SA) with the same values of the parameters as them.

The Vertex Substitution was first discussed as a local search heuristic by Teitz and Bart (1968) and it is also known as T&B. This classical interchange heuristic begins with randomly selecting an initial configuration. That configuration will be replaced by a better solution found from its 1-neighborhood. The process iterates until the present configuration cannot be improved in its 1-neighborhood. Hence, the algorithm always terminates at an optimum, possibly a local one. The implementation is summarized by the following steps:

1. Randomly select $p$ nodes from the candidate nodes as the initial configuration $S$;
2. For solution $S$, calculate the objective function value abbreviated as $OFV_S$;
3. Construct a set $C$ of all candidate nodes not in $S$;
4. Construct the 1-neighborhood configuration of $S$ (for each vertex $s_i$ in $S$ substitute $s_i$ with every point $c_i$ in $C$) and select a new configuration $S_{\text{new}}$ from the 1-neighborhood so that $OFV_{\text{new}}$ has the smallest value of all the 1-neighborhood configurations;
5. If $OFV_{\text{new}} < OFV_S$, substitute $S$ with $S_{\text{new}}$ and go to step (3); otherwise stop the search.

T&B (as well as SA) starts at a random configuration, thereby inducing variation in the solution. To reduce the risk of a solution merely being a local optimum, we start with 4 random configurations in each experiment and select the solution with the smallest objective function value among these four solutions. In the implementation, the program was coded in C and compiled using GCC on a Linux (Ubuntu) system. The computer had a memory of 7.9 G and a CPU of Intel Core i5 3.3 GHz.
3. Data processing and results

Carling, Han and Håkansson (2012a) examined the effects of distance measures in the region under study having asymmetric distributions of road network and population. We briefly discuss the data here and refer to their work for more details. The data is the complete digitalized representation of the real world road network and geo-coding of the population of Dalarna in Sweden. The population data is from Statistics Sweden as of 2002. The residents are geo-coded with a precision of 175 meters and amounts to some 275,000. Figure 3 depicts the spatial distribution of the population in the region.

![Map of the Dalarna region showing one-by-one kilometre cells where the population exceeds 5 inhabitants.](image)

The road network of Dalarna (see Figure 1) is stored in two shape files, one of them includes all the information of the speeds and the directions; the other contains the road classes. We use the c-shape file library to process the road network file. There are 1,797,939 nodes and 1,964,801 road segments. This is the road network used for travelling between the residence and the nearest facility.
The Dijkstra algorithm (Dijkstra, 1959) was used to calculate the shortest distance between each potential location node to all the nodes of the population. The algorithm starts from a node and then calculates the distance between it and all the other nodes. After the computation, we found 9,020 nodes not connected with the main part of the network. These nodes and corresponding road segments were deleted. The matching between the residents’ locations and the network is based on the network after deletion.

Carling et al. (2012a) concluded that travel time and network distance both give the similar configurations for the optimal location of multiple facilities. The only difference between this study and Han et al (2013) is that travel time as the distance measure is employed rather than network distance.

The coordinates of the residents do not perfectly coincide with the nodes of the road network. We approximate the distance between the resident and the facility by using the resident’s nearest node in the network. This approximation potentially introduces an error in the computation. However, the average distance between the residence and the nearest network node is only 62 meters which is substantially less than the geo-coding error. To compute the travel time, the speed is needed. We find that 84 % of the roads have a speed limit of 70 km/h. There were 168 road segments with speed missing for which we imputed a speed limit of 70 km/h.

For the location problem, Han et al. (2013) did not consider all the 1,797,939 nodes as candidates for locating a facility. They considered candidate nodes ranging from some 70,000 down to as few as 500. As a side-remark, the network with 70,000 nodes is the densest one we have encountered in the literature. They reduced the number of nodes to 67,020 by grid aggregation, i.e. two or more nodes close to each other within the same grid were collapsed into one node being the centre of the grid. Thereafter, the nodes were further reduced by
imposing restriction on permissible road classes. The road classification is hierarchical in that the lowest class consists of road of highest quality: the sparsest road network that is classified as road class 0 only includes the European highways, whereas road class 9 consists of narrow dirt-roads. Table 1 summarizes the number of nodes upon varying the restriction on road classes. As a consequence, the number of candidate nodes ranges from 1548 to 67020.

**Table 1.** The average travel time (in seconds) from the demand points to their nearest facility. The number of facilities ($p$) and candidate nodes for locating facilities are varied.

<table>
<thead>
<tr>
<th>Road classes</th>
<th>Nodes</th>
<th>$p=5$</th>
<th>$p=10$</th>
<th>$p=20$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SA T&amp;B</td>
<td>SA T&amp;B</td>
<td>SA T&amp;B</td>
</tr>
<tr>
<td>0-1</td>
<td>1548</td>
<td>964.84 964.84</td>
<td>586.62 586.62</td>
<td>419.07 418.56</td>
</tr>
<tr>
<td>0-2</td>
<td>2237</td>
<td>957.09 961.88</td>
<td>582.55 581.66</td>
<td>387.42 384.45</td>
</tr>
<tr>
<td>0-3</td>
<td>3135</td>
<td>957.09 961.77</td>
<td>582.52 581.66</td>
<td>386.81 382.38</td>
</tr>
<tr>
<td>0-4</td>
<td>5673</td>
<td>957.10 957.09</td>
<td>582.79 581.66</td>
<td>382.00 376.29</td>
</tr>
<tr>
<td>0-5</td>
<td>11112</td>
<td>957.64 956.87</td>
<td>582.32 580.41</td>
<td>383.55 371.63</td>
</tr>
<tr>
<td>0-6</td>
<td>11259</td>
<td>956.87 956.87</td>
<td>582.97 580.41</td>
<td>385.66 371.61</td>
</tr>
<tr>
<td>0-7</td>
<td>19556</td>
<td>959.46 956.87</td>
<td>585.30 580.41</td>
<td>383.27 371.54</td>
</tr>
<tr>
<td>0-8</td>
<td>44296</td>
<td>960.58 956.87</td>
<td>591.40 580.70</td>
<td>394.03 371.54</td>
</tr>
<tr>
<td>0-9</td>
<td>67020</td>
<td>961.57 956.87</td>
<td>593.61 580.70</td>
<td>400.38 371.54</td>
</tr>
</tbody>
</table>

Table 1 also gives out the results from the computational experiments of T&B and SA when the number of facilities ($p=5, 10, 20$) and the density of road network are both varied.

We readily replicated the results of Han et al. (2013) as the outcome from SA is almost identical to their results. Specific to our question of this paper, T&B provide solutions similarly to SA up to the density of road class 5, i.e. about 10,000 nodes. Thereafter upon increasing the number of nodes, solutions of T&B are stable whereas the solutions of SA deteriorate. The reason for the poor performance of the SA algorithm in these complex problems might be either the fact that the maximum number of iterations was fixed at 20,000 (in accordance with Han et al, 2013) or the re-heating scheme of the algorithm or both.
Anyway, we may conclude that the best solution to the $p$-median problem will not always be found in the most detailed network. Yet given an efficient algorithm and unlimited computing time, the solution to $p$-median model improves monotonically with the density of the network.

4. Conclusion

The $p$-median model is commonly used to find optimal location of facilities for geographically distributed demands. Han et al (2013) studied the $p$-median solutions when the density of a road network was varied from 500 to 70,000 nodes. For a density beyond some 10,000 nodes, they found a gradual worsening in solutions. In this study we rejected their finding: the solutions stabilize at about 10,000 nodes, they do not deteriorate when the node density is higher.

As a secondary finding, we note that the SA and T&B algorithms perform differently. While they produced identical solutions in a sparse network, upon solving a $p$-median problem with more than 10,000 candidate nodes the SA performed poorly whereas the solutions of T&B were stable. It should be noted however that the vertex substation required substantial computing time to get a solution, e.g. in the case of $p=20$ and the most dense network T&B required more than 40,000 s whereas SA only cost about 60 s.

Acknowledgement

The authors would sincerely thank Zhiguang Dan in helping with the writing, testing and tuning the Simulated Annealing and Vertex Substitution algorithms.
References


