Analysis of vehicle route choice during incidents

Joakim Janmyr
Daniel Wadell

2018-06-13
Analysis of vehicle route choice during incidents

Examensarbete utfört i Transportsystem vid Tekniska högskolan vid Linköpings universitet

Joakim Janmyr
Daniel Wadell

Handledare David Gundlegård
Examinator Clas Rydergren

Norrköping 2018-06-13
Upphovsrätt

Detta dokument hålls tillgängligt på Internet – eller dess framtida ersättare – under en längre tid från publiceringsdatum under förutsättning att inga extra-ordinära omständigheter uppstår.

Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns det lösningar av teknisk och administrativ art.

Upphovsmannens ideella rätt innefattar rätt att bli nämnd som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannens litterära eller konstnärliga anseende eller egenart.

För ytterligare information om Linköping University Electronic Press se förlagets hemsida http://www.ep.liu.se/

Copyright

The publishers will keep this document online on the Internet - or its possible replacement - for a considerable time from the date of publication barring exceptional circumstances.

The online availability of the document implies a permanent permission for anyone to read, to download, to print out single copies for your own use and to use it unchanged for any non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional on the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility.

According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement.

For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its WWW home page: http://www.ep.liu.se/

© Joakim Janmyr, Daniel Wadell
Abstract

During incidents in a road network, the travel patterns changes. The users in the network then use alternative routes to avoid the congested areas. To have knowledge about routing behaviors during different circumstances in a network could be helpful when redirecting traffic and optimizing signal schemes in the traffic signals.

The use of Global Positioning System (GPS) observations for investigating routing behaviors in a network can be a good alternative to using more traditional traffic simulation models. Real GPS observations gives a direct description of travel pattern used in the network and can be used for analyzing different routes, as well as a calibration tool for different types of traffic models and algorithms.

In this thesis, a method for defining trips from GPS observations was implemented. The method includes defining trips, map matching and inferring the links used between each GPS observation. A penalty-based route set generation algorithm was also implemented, and the inferred paths from the GPS observations were used as a calibration tool for the route set generation algorithm. The investigated network is part of the Interstate 210 freeway east of Los Angeles, USA, and the input data were collected the first half of 2014. The aim is to gain a better understanding about the routing behavior during different circumstances in the network. The purpose is to be able to make more efficient decisions in the future regarding redirection of traffic, and other actions to reduce congestion during incidents.

The investigated scenarios describe the case of regular days and an incident day. The incident investigated occurred April 24th, 2014 and blocked all eastbound traffic on the Interstate 210 freeway.

The results regarding the travel patterns from the GPS observations shows significant differences in number of eastbound travelers choosing to travel north of, south of, and on the freeway during regular days compared with the incident day. The route travel times are also higher during the incident day. Different travel times as costs on the links have a large impact on the results from the route set generation algorithm.

The conclusion is that the implemented methods can be used to gain a better understanding about routing behavior. However, to use the results for decision making, more input data with better precision should be used. On the other hand, the results from this thesis can be used as support when analyzing results from other methods for analyzing traffic.
Acknowledgements

We are grateful and honored for have been given the opportunity to be a part of the Connected Corridor project at Partners for Advance Transportation Technology at the University of California, Berkeley. We sincerely thank David Gundlegård and Clas Rydergren for introducing this thesis for us, and Alexandre M. Bayen for making it possible. We would also like to express our very great appreciation to our mentor Anthony D. Patire and his colleagues at PATH for their enthusiasm and for providing us with support and valuable knowledge, as well for making our stay in Berkeley one to remember.

Our special thanks are extended to our supervisor David Gundlegård and examiner Clas Rydergren at Linköping University for all the support and inspiring discussions. Your guidance and trust have played an important role in our thesis. We also like to thank Rasmus Ringdahl at Linköping University for helping us developing the map matching algorithm and for help with the database.

Furthermore, we want to thank Carina Jönhill Nord and Erik Fransson at Sweco Society, Norrköping for showing interest in our thesis.

Finally, we would like to thank our family, and everyone involved in our thesis for all the support and motivation. Our visit to Berkeley have been a truly great experience, which has helped us to improve ourselves, both on a personal and professional level.
Contents

List of Figures ............................................................................................................................ v
List of Tables ............................................................................................................................ vii
List of Abbreviations ............................................................................................................... viii
1 Introduction ......................................................................................................................... 1
  1.1 Background .................................................................................................................. 1
  1.2 Aim and Purpose ......................................................................................................... 2
  1.3 Research questions ...................................................................................................... 2
  1.4 Methodology ................................................................................................................ 3
  1.5 Limitations ................................................................................................................... 3
  1.6 Outline ......................................................................................................................... 4
2 Introduction to traffic modelling ......................................................................................... 5
  2.1 Traffic flow models ..................................................................................................... 5
  2.2 Microsimulation and Aimsun ...................................................................................... 6
3 Map matching and positioning ............................................................................................ 8
  3.1 Map matching algorithms ............................................................................................ 9
    3.1.1 Map matching algorithm for single observations ............................................... 11
    3.1.2 Hidden Markov Model by using Viterbi algorithm ........................................... 13
4 Route set generation algorithm ......................................................................................... 14
  4.1 Dijkstra’s algorithm ................................................................................................... 15
  4.2 Overlap between routes ............................................................................................. 16
5 Methods for route set estimation ....................................................................................... 18
  5.1 Map matching algorithm ........................................................................................... 18
  5.2 Inferred paths from GPS observations ....................................................................... 18
  5.3 Route set generation algorithm .................................................................................. 23
6 Experimental setup and tools ............................................................................................ 25
  6.1 Computer tools .......................................................................................................... 25
    6.1.1 QGIS ................................................................................................................... 25
    6.1.2 pgAdmin and PostgreSQL ................................................................................. 25
    6.1.3 Python ................................................................................................................. 26
    6.1.4 MATLAB ........................................................................................................... 26
  6.2 Extracted network from Aimsun ............................................................................... 26
  6.3 GPS observations ....................................................................................................... 29
  6.4 Incident day ............................................................................................................... 31
  6.5 Experiments ............................................................................................................... 32
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>The fundamental diagrams of traffic flow</td>
<td>6</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Example of a map matching process</td>
<td>10</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Flow chart of the map matching algorithm</td>
<td>11</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Example of a simple HMM trellis graph</td>
<td>13</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Flow chart of the method for defining trips from the GPS observations</td>
<td>19</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Flow chart of the link estimation algorithm based on GPS observations</td>
<td>22</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Flow chart of the route set generation algorithm</td>
<td>24</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Extracted network from the microsimulation model in Aimsun</td>
<td>27</td>
</tr>
<tr>
<td>Figure 9</td>
<td>The TAZ in the network with the containing centroids</td>
<td>28</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Historical PeMs data</td>
<td>30</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Historical photo from the day of the major incident, April 24th, 2014, together with its location in the traffic network</td>
<td>31</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Network divided into 5 investigation zones</td>
<td>32</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Simplification of the investigation zones with possible eastbound paths</td>
<td>33</td>
</tr>
<tr>
<td>Figure 14</td>
<td>The investigated cases</td>
<td>33</td>
</tr>
<tr>
<td>Figure 15</td>
<td>The two zone pairs used for calibration of the route set generation algorithm</td>
<td>34</td>
</tr>
<tr>
<td>Figure 16</td>
<td>Distribution of the difference between system time and sample time</td>
<td>35</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Distribution of number of GPS observations, aggregated by hour</td>
<td>37</td>
</tr>
<tr>
<td>Figure 18</td>
<td>Percentage of GPS observations provided by Fleet and Consumer respectively during 23 regular days and the incident day</td>
<td>38</td>
</tr>
<tr>
<td>Figure 19</td>
<td>Heat map of the density of the GPS observations during 23 regular days, for 24 hours</td>
<td>38</td>
</tr>
<tr>
<td>Figure 20</td>
<td>Heat map of the density of the GPS observations during the incident at April 24th, 2014, for 24 hours</td>
<td>39</td>
</tr>
<tr>
<td>Figure 21</td>
<td>Heat map of the density of the GPS observations during 23 regular days, 1 p.m. to 8 p.m.</td>
<td>39</td>
</tr>
<tr>
<td>Figure 22</td>
<td>Heat map of the density of the GPS observations during the incident at April 24th, 2014, 1 p.m. to 8 p.m.</td>
<td>40</td>
</tr>
<tr>
<td>Figure 23</td>
<td>Selection of mapped GPS observations with their corresponding resemblance interval</td>
<td>41</td>
</tr>
<tr>
<td>Figure 24</td>
<td>Total number of identified trips from zone 1 to zone 4</td>
<td>43</td>
</tr>
<tr>
<td>Figure 25</td>
<td>Number of identified trips in case 1, north of the I-210 freeway</td>
<td>43</td>
</tr>
<tr>
<td>Figure 26</td>
<td>Heatmap of observed paths north of the I-210 freeway for regular days</td>
<td>44</td>
</tr>
<tr>
<td>Figure 27</td>
<td>Heatmap of observed paths north of the I-210 freeway for the incident day</td>
<td>45</td>
</tr>
<tr>
<td>Figure 28</td>
<td>Number of identified trips in case 2, on the I-210 freeway</td>
<td>45</td>
</tr>
<tr>
<td>Figure 29</td>
<td>Heatmap of observed paths on the I-210 freeway for regular days</td>
<td>46</td>
</tr>
<tr>
<td>Figure 30</td>
<td>Heatmap of observed paths on the I-210 freeway for the incident day</td>
<td>47</td>
</tr>
<tr>
<td>Figure 31</td>
<td>Cumulative number of trips on the I-210 freeway</td>
<td>47</td>
</tr>
<tr>
<td>Figure 32</td>
<td>Number of identified trips in case 3, south of the I-210 freeway</td>
<td>48</td>
</tr>
<tr>
<td>Figure 33</td>
<td>Heatmap of observed paths south of the I-210 freeway for regular days</td>
<td>49</td>
</tr>
<tr>
<td>Figure 34</td>
<td>Heatmap of observed paths south of the I-210 freeway for the incident day</td>
<td>49</td>
</tr>
<tr>
<td>Figure 35</td>
<td>Travel times on the main arterials</td>
<td>50</td>
</tr>
<tr>
<td>Figure 36</td>
<td>Speed on the main arterials</td>
<td>51</td>
</tr>
<tr>
<td>Figure 37</td>
<td>Coverage of paths from GPS observations when the threshold parameter is fixed and μ is varying for zone pair 1</td>
<td>52</td>
</tr>
</tbody>
</table>
Figure 38 - Coverage of paths from GPS observations when $\mu$ is fixed and the threshold parameter is varying for zone pair 1.................................................................................................................. 53
Figure 39 - Coverage of paths from GPS observations when the threshold parameter is fixed and $\mu$ is varying for zone pair 2.......................................................................................................................... 54
Figure 40 - Coverage of paths from GPS observations when $\mu$ is fixed and the threshold parameter is varying for zone pair 2.................................................................................................................. 55
Figure 41 - Degree of overlap between routes from the route set generation algorithm, when the threshold parameter is fixed and $\mu$ is varying ........................................................................ 56
Figure 42 - Degree of overlap between routes from the route set generation algorithm, when $\mu$ is fixed and the threshold parameter is varying................................................................................ 57
Figure 43 - Coverage of paths from GPS observations for different combinations of the parameters, zone pair 1.................................................................................................................. 59
Figure 44 - Coverage of paths from GPS observations for different combinations of the parameters, zone pair 2.................................................................................................................. 59
Figure 45 – Occurrences of each link in the route set in scenario 1............................................. 61
Figure 46 – TAZ affected by the incident of April 24th, 2014 in the route set in scenario 1........ 62
Figure 47 - Occurrences of each link in the route set in scenario 2............................................. 63
Figure 48 - Occurrences of each link in the route set in scenario 3............................................. 64
List of Tables

Table 1 - Percentage of GPS observations that were filtered out for each day ......................... 36
Table 2 - Results from the map matching algorithm ................................................................. 41
Table 3 – Results from the path inferring process of the GPS observations ......................... 42
Table 4 - Summary of the results from the set of experiments for zone pair 1 ......................... 57
Table 5 - Summary of the results from the set of experiments for zone pair 2 ......................... 58
Table 6 - Set of experiments performed with different combinations of the parameters ....... 58
Table 7 - Summary of the results from the set of experiments with different combinations of parameters for zone pair 1 ........................................................................................................ 60
Table 8 - Summary of the results from the set of experiments with different combinations of parameters for zone pair 2 ........................................................................................................ 60
Table 9 – General results from the route set generation scenarios ................................................. 65
List of Abbreviations

**DR:** Distance Resemblance
**GIS:** Geographical Information System
**GPS:** Global Positioning System
**GUI:** Geographical User Interface
**HMM:** Hidden Markov Model
**HOV:** High-Occupancy Vehicle lane
**HR:** Heading Resemblance
**ITS:** Intelligent Transportation System
**PeMs:** Performance Measurement System
**SSSP:** Single Source Shortest Path
**TAZ:** Traffic Analysis Zones
1 Introduction

Traffic incidents can cause severe problems such as major congestion, total stop, and even secondary incidents. In case of congestions and stops at the road stretch during the incident it is possible that travelers at some location upstream from the incident change their planned route to avoid the congestion. What is not as obvious and easy to understand is which routes that are used instead. If the incident is major and many travelers change their planned route it could lead to increased traffic flow at the main arterials in the area around the incident. As for all road networks, the road type, speed limit, number of lanes, signal schemes at the traffic signals etcetera is designed to be able to handle a predicted traffic demand during normal conditions. When travelers change their route due to incidents it is therefore a risk that the arterials around the incidents gets crowded with possible congestion and low throughput as a result. One important aspect when trying to overcome this problem is to understand which routes that are used in the network during incidents. With information about used routes during incidents it is possible to control the traffic conditions in a more efficient way. This can be done by for example redirecting traffic and change signal schemes in intersections, to avoid congestion and improve the traffic flow at the arterials around the incident.

Many cities have a simulation model of the traffic system. These models are important in the process of trying to estimate future capacity problems, which is a threat since the population and amount of road users are increasing. The models are important for analyzing effects of large modifications of the infrastructure, like rearrange intersection types or building new roads. However, there are other possible methods than simulation models for analyzing aspects as route diversion and route sets. The use of observed traffic measurements in the network during incidents can improve the estimates of the likely route sets, while also being a validation tool for the simulation models.

1.1 Background

This thesis is part of the Connected corridors project in California, USA. The work has been performed in the research group Partners for Advanced Transportation Technology at the University of California, Berkeley. One main area of work in the research group is to investigate travel patterns and routing behavior in a corridor near the Interstate 210 freeway, east of Los Angeles. That is, to investigate which routes the travelers use in the traffic network during different circumstances. One method for this is to use a simulation model of the traffic system. Today, the research group has developed a detailed simulation model of the corridor in a
software called Aimsun, which is used for simulation of the traffic during different circumstances. Using a simulation model is only one of the methods for analyzing traffic. Another possible method for analyzing the traffic is to use real traffic observations. The research group has for several years been provided with GPS observations from a company called HERE technologies. Each GPS observation gives information about for example, position and speed. Until today, the GPS observations have not been used for analyzing travel patterns and routing behavior due to lack of time and higher priorities in the research group. This thesis focus on developing a method for inferring paths from these GPS observations, to be able to analyze travel patterns and routing behavior in the corridor, during both regular days and during specific days with known incidents. A route set generation algorithm will also be implemented. A route set is a set of possible routes between two points in the network. The routes identified from the GPS observations will be used for calibration of the route set generation algorithm. The network that will be used in this thesis will be extracted from the simulation model in Aimsun.

1.2 Aim and Purpose

The aim of this master thesis is to gain a better understanding about the routing behavior in the network during regular days and during incidents. To achieve this a method for generating trips from GPS observations will be developed. The routes identified from the GPS observations will be used as a calibration tool for a more general route set generation algorithm that will be tested during different circumstances in the network to gain a better understanding about which routes that have been used.

This knowledge can lead to the possibility to make more efficient decisions in the future regarding redirection of traffic and changes in signal schemes of the traffic signals in the areas around the incident, to reduce congestion during incidents.

1.3 Research questions

- Is it possible to use observed GPS observations for investigating route sets in a network during regular days and days with known incidents? And in that case, how does those results differ from the results from a calibrated route set generation algorithm?
- How important is the accuracy of the GPS observations when mapping them to the network? And how can incorrectly mapped GPS observations be handled in the trip generation method?
• How does the route travel times based on observed GPS observations in the network differ between regular days and an incident day?
• How does the route sets from the route set generation algorithm change when using travel times from the GPS observations as costs on the links?

1.4 Methodology
To analyze the route sets in the network the proposed method is to extract the network from the Aimsun model into an open source Geographic Information System (GIS) software. With this network including nodes, links, and Traffic Analysis Zones (TAZ), a route set generation algorithm will be implemented in Python to generate a route set for each TAZ-pair in the network. This route set generation algorithm will be either link elimination based, or link penalty based. This means that the overlap of the different routes in the set will be calibrated to a reasonable value. The generated routes will then be compared with the paths identified from the GPS observations. To identify paths from the GPS observations, a method for generating trips and inferring links between each GPS observation will be implemented in Python. The comparison between the route set generation algorithm and the routes from the GPS observations will be done both for regular days without incidents, and for one day with a known incident at a specific location. The results will be evaluated based on percentage of matched links between the routes, number of generated routes, and on the overlap in the route set from the route set generation algorithm. The method for calculating percentage of matched links and overlap will be implemented in MATLAB. In the route set, those pair of TAZ that are affected by the incident area will be used to generate different route sets during other circumstances.

1.5 Limitations
Regarding the travel patterns from the GPS observations only a set of eastbound travel patterns will be investigated. The input data will only be collected from the first half of 2014, due to that the investigated incident in this thesis occurred during that period. The route sets generated with the route set generation algorithm will be independent between different TAZ-pairs. This means that when a route set is generated for a certain TAZ-pair, the cost on the links will be reset before generating the route set for the next TAZ-pair. Only the travel time during different circumstances will be used as generalized cost on the links in the route set generation algorithm. When extracting the network all transit lines will be removed and the type of vehicle that the GPS observations come from will not be investigated.
1.6 Outline

The report begins with an introduction to traffic modelling in Chapter 2, where general information about traffic flow models is presented. This is followed by a theoretical background about map matching algorithms and route set generations algorithms, in Chapter 3 and 4. A description of the method for the map matching algorithm, together with the methods for inferring paths from GPS observations and the route set generation algorithm are presented in Chapter 5. Chapter 6 presents the experimental setup, computer tools used in the thesis and the input data. The results and analysis are presented in Chapter 7, followed by discussion, future work, and conclusions in Chapter 8 and 9.
2 Introduction to traffic modelling

Models describing the traffic conditions in a specific area can be used when describing the reality in a simplified way. The results from a traffic model can be used for analyzing and visualizing the outcome of a project or a future investment in the road network or infrastructure in general. To obtain models that represent reality in a good and relevant way it is important to have good input data to the model. This input data is often real traffic observations such as flow count data, GPS observations and/or information about the investigated area. Information about the area could be for example population, work places and residential areas. However, the model needs to be calibrated over time to represent the changes of the real traffic situation. This means that the input data for the model must be updated to obtain results from the model that represent the time-period that is desired to be investigated. There are different types of traffic models that can be used depending on what kind of information that is desirable to obtain, and the size of the investigated area and network. Two common types of models are macroscopic models and microscopic models. The models on macroscopic level is often used for larger areas such as cities or large parts of cities with aggregated averages values of flow, speed, and density as output. Microscopic models can be used for small intersections or road sections and gives information about the dynamics of the individual vehicle such as queue length and waiting time (Barcelo 2010).

2.1 Traffic flow models

Traffic flow models can be described as fluid. Both traffic and fluid can be described with the fundamental equation, which describes the relationship between flow, density, and speed. See Equation (1) (Holmberg & Hydén 1996).

\[
q = k \cdot v
\]  

(1)

Where \( q \) is the flow, \( k \) the density, and \( v \) the speed.

Figure 1 shows the fundamental diagrams based on Equation (1).
The left graph in Figure 1 shows the relationship between the speed \( v \), and the vehicle density \( k \). \( V_f \) is in this case the free flow speed when the network is empty, that is, when the density is zero. When the density increases, the speed decreases until the density reach the jam density, \( k_j \). The notations \( v_m \) and \( k_m \) are the maximum speed and density before the flow starts to decrease instead of increase. The middle graph in Figure 1 shows the relationship between the flow \( q \), and density \( k \). \( q_m \) is the maximum flow and \( v_m \) indicates the optimum speed before the flow starts to decrease. The right graph in Figure 1 shows the relationship between the speed and flow. All notations in these three graphs relate to each other between the graphs (Milojevic et al. 2015).

Most traffic models are built with the fundamental equation in mind together with a simplified network over the area that is investigated. A network is usually represented by nodes, such as intersections and roundabouts; links, such as roads and sidewalks; and centroids, which are the zones in the network where a trip is generated and/or attracted. Each represented link in the network holds individual properties, such as direction, speed limit, and capacity, just like in the real network (Barcelo 2010).

### 2.2 Microsimulation and Aimsun

A general microscopic simulation model is often used to simulate the traffic conditions in detail in smaller networks where the demand is defined as input flows, turning proportions, and exit sections. The traffic flow in a microscopic model is described by each individual vehicle in the network. To achieve this each possible action of those vehicle must be modelled. Such actions are acceleration, deceleration, lane changing and car following behaviors. Each of these actions is modelled mathematically and calibrated to obtain real life behavior of each individual vehicle (Barcelo 2010). There are several different software for simulating traffic.
Aimsun is a traffic modelling software and was originally created during a long-term research program for microscopic traffic simulations (Barcelo 2010). The focus when developing the software was to be able to analyze microscopic models, but today Aimsun can offer mesoscopic and macroscopic models as well as hybrids of these models. Aimsun comes in two different versions, which are Aimsun Live and Aimsun Next. Aimsun Live is a decision support system for real-time traffic management and allows the traffic operators to make real-time decisions regarding the current and upcoming traffic condition (Aimsun 2018a). Aimsun Next is a well-developed traffic management tool and the latest version offers a four-step transportation planning process (Aimsun 2018b). In this thesis, the used network is modelled in Aimsun Next.
3 Map matching and positioning

Real traffic observations in form of vehicle positions, or the number of vehicles passing some certain points in the traffic network can be useful when developing methods for traffic analysis. The data are in many cases complex and it is not obvious how to process and analyze the data for a specific project. Accurate vehicle localization is important for many Intelligent Transportation System (ITS) applications to get reliable and correct analyzes and results. According to White, Bernstein & Kornhauser (2000) there are mainly three different ways to determine a vehicle’s location.

The first is to use some form of dead reckoning system. This method uses the vehicle speed and direction of movement in order to continuously update the location of the vehicles. Many dead reckoning system use a compass and/or a gyro system to enhance the determination of the absolute heading. A big problem with dead reckoning is that it will produce uncertain positioning when the trips becomes long, due to that the absolute positioning error will grow proportionally with the distance traveled (Greenfeld 2002).

The second way is to use some form of terrestrial beacon that broadcast radio waves to locate the vehicle’s position (Iwaki, Kakihara & Sasaki 1989). When the vehicles reception to the beacon becomes weak, the positioning will get low accuracy. According to Iwaki et al. (1989) there is a possibility that the real vehicle position is far away from the estimated position even though it is a detailed network and the estimated position of the vehicle is located on the road.

The third and probably most used method to determine a vehicle’s location is to use some form of radio/satellite positioning system, such as GPS. GPS devices are not so expensive and can provide in addition to the position, the speed and heading (Greenfeld 2002). The position of a GPS observation is determined with a latitude and longitude value, which can be converted into an X and Y value with respect to a two-dimensional surface. According to Jagadeesh, Srikandan & Zhang (2004) there is a strong need for exploring a solution for vehicle location that relies on a GPS receiver which only purpose is to position the observation without complex computations. The accuracy of a GPS position may vary, and it depends on several factors, such as satellite geometry, radio interference, signal reflections and blockage due to obstacles for example. The US Government (2017) states that GPS-enabled smartphones are typically accurate within a 4.9 m radius under open sky, but the accuracy can worsen near obstacles.
3.1 Map matching algorithms

To infer travel trajectories from the GPS observations, a map matching algorithm could be of good use. The main goal with a map matching algorithm is to map the observed vehicle position to a link in a traffic network. There are many ways to design a map matching algorithm in which each has advantages and disadvantages. The complexity level of the map matching algorithm depends a lot on the purpose of the application and the quality and availability of the data. Greenfeld (2002) means that there are three complexity levels that a map matching algorithm has to resolve. The lowest level of complexity is needed when the vehicles travels on fixed routes in the traffic network. An example of this could be buses or other public transport vehicles, which always travelling on the same links between each station. The only thing the algorithm needs to do is to locate the buses to one of the links that makes up the bus route.

The second level of map matching algorithm is to map GPS observations to links when the origin and destination of the route is known in advance. The algorithm will assume that the vehicles follows some of the suggested routes and will then map the observation to one of these routes. If the deviation between the observation and the suggested route is large, the idea is to reconstruct a new suggested route from new the position to the destination (Greenfeld 2002). The disadvantages with these kinds of algorithms is that it can result in incorrect matches, if the deviation between the observations and the known route is small. For example, if a vehicle is travelling on a parallel link with a link on the known route, the algorithm will most likely locate the vehicle to the link on the known route.

The third and most complex level is when the map matching algorithm does not have any knowledge or information about the system except the networks structure in form of links and the positions of the observations. However, the goal for all of these three levels of map matching is to map the GPS observation to the most likely link in the network.

Some map matching algorithms are designed to map the observations one by one. These are relatively fast and easy to implement and can give reasonable results depending on the purpose of the application and the quality of the data. Jagadessh et al. (2004) propose a map matching algorithm of this kind. It is explained more in detail in section 3.1.1. There are also more advanced map matching algorithms, which maps a sequence of observations. This means that the algorithm will locate an observation based on the previous and the following observations to map the observation more accurate to the most likely link. A well-known method for mapping
a sequence of observation is the Hidden Markov Model (HMM) based on the Viterbi Algorithm. This method is presented in section 3.1.2.

Figure 2 illustrates a small example of the map matching process and its possible disadvantages.

![Figure 2 - Example of a map matching process](image)

The left part in Figure 2 illustrates a small traffic network with observations from one vehicle. The blue dots represents the observed positions, while the stars represent the true positions. The right part of Figure 2 shows the same network but also includes estimated locations from two different methods of map matching. The green dots represents estimated observations from an algorithm that only consider one observation at a time, while the red dots represents estimated observations with an algorithm that consider a sequence of observations. It is not obvious if the vehicle has used link BE or CF, since measurement 4 and 5 are closer to CF than BE. Measurement 6 indicates that link BE is the most probably used. In these situations it is hard to decide which link and route the vehicle actual used. Poor accuracy of the GPS observations and low number of measurements, together with a network with many links close to each other is difficult to handle. The map matching process can be improved by using the observation attributes, like speed and heading. The heading of the observations can be compared with the direction of the links, which implies that many links in the opposite directions can be rejected. Shen, Tang & Zhang (2015) means that the point speed of the observations together with the timestamps can be useful to estimate a reasonable distance between each observation. Observations close to challenging areas, like intersections and parallel links can prevent the map matching process to estimate the position correct if the heading and position of the observations are deceptive (Hashemi & Karimi 2016). With this in mind, a sequential map matching algorithm could be useful.
3.1.1 Map matching algorithm for single observations

Jagadeesh et al. (2004) propose an efficient map matching algorithm with fast computations. The algorithm is relatively easy to implement and can conceptually be divided into two parts. The goal is to correctly identify the road section and determine the exact location of the vehicle on the identified road section (Jagadeesh et al. 2004). A brief overview of the map matching algorithm is presented in the flow chart in Figure 3.

![Figure 3 - Flow chart of the map matching algorithm](image)

The algorithm will go through each observation, one by one. It consists mainly of 6 steps, which are described below.

**Initialization**
Before the algorithm can be used, all observations to be mapped must be set to *not parsed*.

**Step 1: Find the nearest links**
The algorithm will find the $x$ nearest links to the observation within a radius of $y$ meter, with respect to the observation’s position in latitude and longitude.

**Step 2: Distance resemblance**
For each of the $x$ links, a matching score based on the difference in distance between the observation to the links will be calculated in meters. The matching score goes from 0 to 1 where
1 is considered to be exactly on the link. Equation (2) presents how the matching score for the distance resemblance ($DR$) is calculated.

$$DR = 1 - \left( \frac{\text{distance between link and observation}}{y} \right)$$

(2)

**Step 3: Heading resemblance**

For each of the $x$ links, a matching score based on the difference in degrees between the heading of the observation and the links will be calculated. A matching score of 1 indicates the exact same heading of the observation and the link. Equation (3) and (4) presents how the matching score for the heading resemblance ($HR$) is calculated.

$$\text{heading difference} = \min \left\{ \left( \frac{\text{heading}_{\text{link}} - \text{heading}_{\text{observation}} + 360}{360} \right) \mod 360, \left( \frac{\text{heading}_{\text{observation}} - \text{heading}_{\text{link}} + 360}{360} \right) \mod 360 \right\}$$

(3)

$$HR = 1 - \left( \frac{\text{heading difference}}{90} \right)$$

(4)

The $\%$ is the notation for modulo operations and is in this case used to obtain differences in degrees lower or equal to 360 degrees.

**Step 4: Resemblance**

For each of the $x$ links, a final common matching score will be calculated by taking the lowest matching score of the distance resemblance and the heading resemblance. This is presented in Equation (5)

$$\text{Resemblance} = \min \{DR, HR\}$$

(5)

**Step 5: Filtering process**

If the final matching score is less than a threshold value, then the observation is too uncertain and will not be mapped to any link.

**Step 6: Assign the observation to a link**

The link with the highest matching score is considered to be the most likely link, which induce the observation to be mapped to that link.
3.1.2 Hidden Markov Model by using Viterbi algorithm

A well-known technique of map matching that is based on a probabilistic sequence model is HMM. In the HMM based approach, a number of candidate links are represented as states (nodes) in an HMM trellis graph for each observation (Jagadeesh & Srikanthan 2016). An example of a simple HMM trellis graph is presented in Figure 4.

![Figure 4 - Example of a simple HMM trellis graph](image)

The edges in Figure 4 represents the transitions between states. The most likely sequence of states can be discovered by applying a dynamic programming algorithm called Viterbi algorithm. According to Forney (1973), the Viterbi algorithm is a recursive optimal solution to the problem of estimating the state sequence of a discrete time finite state. In other words, the underlying process of the Viterbi algorithm is to find the so called “Viterbi path” between each state. This is done by calculating the probabilities of all the transitions, which are calculated by using the shortest path. According to Wei, Wang, Forman, Zhu, Guan (2012), the shortest path calculations are the most computational expensive part of the map matching process. With this in mind, the HMM by using Viterbi algorithm can be very cost expensive in form of computation time compared with other simpler map matching algorithms. This problem can be improved by limit the number of states for each observation.
4 Route set generation algorithm

When analyzing the route diversion and route flows in a network, the first step is to investigate which routes that are most likely to be used by the travelers between each OD-pair. In a large urban network there are many different routes that can be used, but most of them may be unreasonable in practice and very circuitous (Bekhor, Ben-Akiva & Ramming 2006). One method for generating routes is the \( k \) shortest path algorithm which creates the \( k \) shortest routes between an OD-pair in the network (Bekhor et al. 2006). However, when applying this kind of method for generating the route set it is not regulated how much the different routes will differ from each other. This means that two different routes in the set can be identical except for a few links, which can be rather far away from how the routes looks in practice. It is on the other hand possible to generate a set with completely unique routes by applying a shortest path algorithm such as Dijkstra’s algorithm, and remove the links used in the first route from the network and apply the algorithm again, to obtain the second shortest route and so on. A scenario with completely unique routes can also be rather far away from a reasonable route set in practice. Most likely, when travelling from point A to point B in a network on different routes, some of the links included in the routes will be used in more than just one route. When designing route set generation algorithms, a problem is to achieve a reasonable balance with overlapping links in the route set.

One approach for achieving reasonable overlapping between routes is proposed by Nassir, Sall, Ziebarth & Zorn (2014) and can be categorized as an iterative penalty-based \( k \) shortest path algorithm. This algorithm adds costs on the links in a generated route according to a penalty term proposed by Bovy, Bekhor & Prato (2008). The penalty term can be calculated by Equation (6).

\[
p(a) = \frac{l_a}{\mu} \ln \sum_{j \in C_t} \delta_{aj}
\]

where:

- \( a \): Current link
- \( l_a \): Generalized cost of using link \( a \)
- \( \mu \): Parameter
- \( C_t \): Set of generated routes
- \( \delta_{aj} \): Equals 1 if link \( a \) is in route \( j \), 0 otherwise
By using this penalty term in the route set generation algorithm, the penalty on each link gets higher for each route that the link is already used in. The pseudo code for generating $T$ routes is shown below (Nassir et al. 2014).

**Initialization step:** $t := 1; C_t := \emptyset$;
Set the link cost to the generalized cost. The penalty term $p(a)$ is set to 0.

**Step 1:** Find the route with the lowest cost for the OD-pair and assign it to path $i$. Path $i$ is replaced with the latest shortest path each iteration.

**Step 2:** $C_t := C_t \cup \{i\}$

**Step 3:** Update the penalty term $p(a)$ for each link $a$ based on Equation (6)

**Step 4:** $t := t + 1$; if $t < T$ go to step 1, otherwise stop

The generalized cost for each route can be based on several different factors and combinations of these. This could be for example number of lanes, free flow speed, types of intersections, type of road, and other measurements of attractiveness.

When the route set for each OD-pair is generated, the demand is assigned to the routes between the OD-pairs. The assignment, as mentioned in section 2.2.4 can be done by using Wardrop’s principle of user equilibrium. The demand can also be assigned by using some kind of route choice logit model.

### 4.1 Dijkstra’s algorithm

The Dijkstra’s algorithm is a shortest path algorithm categorized as a Single Source Shortest Path (SSSP) algorithm which means that it is designed to calculate the shortest path from one single node to any other node in the network (Thorup 1999). Dijkstra’s algorithm does not handle negative costs on the links in the network. For handling negative costs there are other shortest path algorithms such as the Bellman-Ford algorithm (Lundgren, Rönqvist & Värbrand 2010). Dijkstra’s algorithm can obtain the shortest path from start node $n_s$ to end node $n_t$ in a network with a set of nodes $N$ and a set of links $B$. The algorithm can be explained in the following 4 steps.

**Step 0:** Divide the set of nodes into two subsets $A$ and $D$. Subset $A$ is the set of searched nodes and $D$ is the set of unsearched nodes. Set: $A = \emptyset$ and $D = N$. Mark start node $n_s$ with
\((p_s, y_s) = (\text{previous node, node price}) = (\mathbf{-}, 0)\). That is, the node does not have any previous node and that the node price is \(y_s = 0\). All other nodes obtains an initial node price of \(y_j = \infty\).

**Step 1:** Identify node \(i \in D\), which has the lowest node price: \(y_i = \min_{k \in D} y_k\).

**Step 2:** Search node \(n_i\), that is to investigate all outgoing links \((i, j) \in B\) from node \(n_i\). If \((y_i + c_{ij}) < y_j\) is true, then a cheaper way from \(n_s\) to \(n_j\) through \(n_i\) has been found. \(c_{ij}\) is the cost of the arc going from \(i\) to \(j\). Then mark node \(n_j\) with \((p_y, y_j) = (i, y_i + c_{ij})\).

**Step 3:** Move node \(n_i\) from subset \(D\) to subset \(N\). This means that \(n_i\) has been search and the algorithm will not consider that node anymore.

**Step 4:** If every node in \(D\) have been searched, which is if \(A = N\), then stop. Otherwise, go to step 1.

When the algorithm has searched through all the nodes, the shortest path can be obtained by using the labels of the nodes. Since Dijkstra’s algorithm search through every node in the network, the algorithm will also provide the shortest path from the start node \(n_s\) to all other nodes in network (Lundgren et al. 2010). If the network is large, it requires high computational cost when using this algorithm.

**4.2 Overlap between routes**

When a route set is generated, it is possible to investigate the similarity between the routes in the route set. That is, to investigate the degree of overlap. Depending on application, different degree of overlap can be desirable. It may not desirable to have completely unique routes in the set, while also not desirable to have almost identical routes in the set. One approach for calculate the degree of overlap between routes is proposed by Hu & Chiu (2015).

In this definition of the degree of overlap, the total degree of similarity \(r^i\), for route \(p_i\) is calculated by Equation (7).

\[
r^i = \sum_{j=1}^{n_i} r^i_j / k - 1
\]

The degree of similarity \(r\), that is, the overlap of those paths can be calculated by Equation (8).

\[
r = \frac{\sum_{i=1}^{k} r^i}{\sum_{i=1}^{k} n_i}
\]
The notations in Equation (7) and Equation (8) are:

\( r_f^j: \) The number of times link \( j \) in path \( i \) show up in all other \( k - 1 \) paths

\( n_i: \) The total number of links in path \( i \)

\( r^i: \) Degree of similarity of path \( i \)

\( r: \) Degree of link similarity of the totality of \( k \) paths

The value of \( r \) will be in the interval \([0,1]\) where 0 means that the routes in the route set are completely unique, and 1 means that the routes are identical (Hu & Chiu 2015).
5 Methods for route set estimation

In this thesis two different approaches for estimating likely route sets have been implemented and evaluated. The first method is the implementation of a route set generation algorithm based on Dijkstra’s algorithm, where the cost on the links is updated between each generated route, to obtain several reasonable routes for each OD-pair. This method is less data driven and has only travel times on the links as input. The second method is more data driven where the routes are generated by using the GPS observations from HERE Technologies. The GPS observations in this method consists of GPS measurements throughout the network giving information about position, time, heading and speed etcetera. By defining which of these measurements that belongs to the same trip it is possible to estimate different routes used in the network. The calibration of the route set generation algorithm is done by comparing these routes with the routes obtain from the GPS observations. To handle the large amount of data in this thesis, a PostgreSQL server has been used. The algorithms were implemented in Python where SQL-queries were used to obtain and write data from and to the database.

5.1 Map matching algorithm

In this thesis a simple point by point map matching algorithm was used to map the GPS observations to sections in the network. The algorithm used is described in section 3.1.1. The parameters in the algorithm are set to find the 5 nearest links to the observation within a radius of 20 meter, with respect to the observation’s position in latitude and longitude. This means that a distance of 20 meter will generate a matching score of the distance resemblance of zero. Regarding the matching score of the heading resemblance, a difference equal to 90 degrees will generate a matching score of the heading resemblance of zero. For each of the 5 links, a final common matching score will be calculated by taking the lowest matching score of the distance resemblance and the heading resemblance. A filtering process was also used. If the final resemblance is equal to 0, then the observation is considered too uncertain and will not be mapped to any link, and if the heading resemblance is less than 0.5, that is, a difference in angle more than 45 degrees, then the observation is also considered too uncertain and will not be mapped to any link.

5.2 Inferred paths from GPS observations

To obtain actual trips from the GPS observations, they must be filtered and sorted. Figure 5 shows a flowchart of the method for defining trips of the raw GPS observations.
Below is an explanation of the flow chart in Figure 5.

**Load table with raw GPS observations:**
The first step is to load the table containing the raw GPS observations. Each row in this table is one GPS observation. This table can consist of GPS observations from one or several days.

**Filter data:**
The filtering process of the raw data is presented in section 6.3

**Find unique devices:**
All the unique device id’s are filtered out.
Filter devices

All devices with less than 20 GPS observations are removed, due to it is not desirable to have a low number of GPS observations in each trip.

The lower part of the flow chart is then done for each unique device id, for each GPS observation.

Compare time between two GPS observations:

In this step, the difference in sample time between two consecutive GPS observations is calculated. If the time difference between two consecutive GPS observations is less or equal to 2 minutes they are included in the same trip. If the difference is more, they are separated into two different trips. The 2-minute threshold is used to only obtain trips with rather close consecutive GPS observations in time. Otherwise it is hard to say what happened between the GPS observations (Wang, Gao & Juan 2017).

Filter trips:

Here all trips containing less than 10 GPS observations are removed. Without this filter there would be many small trips saved which contain a small number of GPS observations and it is then hard to analyze and evaluate those trips with the travel patterns investigated in this thesis. The travel patterns investigated are explained in section 6.5. The decision to require at least 10 GPS observations was decided empirically.

When the trips have been generated according to the steps above each GPS observation is mapped to the network using the algorithm explained in section 5.1. The next step is to infer the links in the network, used in each trip, between each GPS observation. This can be considered as a sort of estimation of the trips obtained by the GPS observations, since there is not always a mapped GPS observation on each link in the trip. One alternative for estimating these links is to run a shortest path algorithm, Dijkstra’s algorithm for example, between each mapped GPS observation in each trip. The output of this is a sequence of links used between each mapped GPS observation, assuming the traveler choose the shortest path between each GPS observation. However, since some of the GPS observations will be mapped to the incorrect link in the network, a more advanced approach of inferring links was implemented. The main idea is to calculate the shortest path between two consecutive, \( j \) and \( j + 1 \), mapped GPS observations using Dijkstra’s algorithm and save the length and mean free flow speed of the links included in that sub path. This distance and mean speed is then compared with the actual speed between the two consecutive GPS observations. The actual speed is obtained by taking
the time difference between the GPS observations and the total length in the path. If the actual speed is much higher than the mean free flow speed calculated from the links in the path, the sub path is considered unreasonable. An example of this is if the sub path between two consecutive GPS observations is 3 miles, and the mean free flow speed is 45 mph, the time between those two GPS observations should be around: \[ \text{time} = \frac{\text{distance}}{\text{speed}} = \frac{3}{45} = 4 \text{ minutes}. \]

If the time between these GPS observations was 1 minute then the speed must have been around 180 mph, which is unreasonable. The end node (GPS observation \( j + 1 \)) of that sub path is then considered to be mapped on the incorrect link and the sub path is then calculated between GPS observation \( j \) and \( j + 2 \) instead. The algorithm developed will allow an actual speed up to 120% of the mean free flow speed of the links in the sub path. Figure 6 shows the flow chart of the link estimation algorithm implemented in this thesis.
The combined output of the methods for defining trips, mapping GPS observations to the network, and inferring links used between the GPS observations is several paths based on GPS observations with inferred links for each path, together with IDs and geometries for the links. This makes it possible to compare the paths from the GPS observations with the paths from the route set generation algorithm, as well as visualize the paths in QGIS.
5.3 Route set generation algorithm

The method for generating the route set for each OD-pair is based on an iterative process of finding the shortest path using Dijkstra’s algorithm. Between each iteration the cost on each link in the latest created route is increased with a penalty term based on the one explained in section 4. Equation (13) shows the definition of the penalty term $p(a)$ for each link used in this thesis.

$$p(a) = \frac{t_a}{\mu t} \ln(1 + \sum_{j \in c} \delta_{a_j})$$

The difference from this penalty term compared with the one suggested by Bovy et al. (2008) described in Chapter 4, is that in the logarithm, a value of 1 is added. This is to obtain an increased cost on the links starting from the first route created, to avoid identical routes being generated in the beginning of the iteration process.

When the links used in the latest route are updated, the next route will be different due to the increased cost on links in the latest route. The initial generalized cost on the links was set to the free flow travel time based on the length and speed limit. The idea is to get reasonable spread on the different routes between the OD-pairs. However, it may not be desirable to generate the same number of routes for all the OD-pairs. For the longer OD-pairs it may be more reasonable to have more routes, than for the short OD-pairs. To handle this situation the number of routes generated for each OD-pair depends on the cost of the shortest path for the current OD-pair.

The algorithm will generate routes until the cost of the route exceeds the cost of the shortest route multiplied with a defined threshold percentage value. Since all data and information was stored in a PostgreSQL database the Dijkstra’s Algorithm was implemented by using the `pgr_dijkstra` function. This function is included in the PostgreSQL extension `pgRouting` for PostgreSQL. The function’s input is:

- sections in the network with corresponding id, origin node, and destination node
- the column defining the cost on each section
- the column defining the reversed cost (for avoiding routes going the wrong direction)
- OD-pair, origin node and destination node.

The output is a sequence of sections used in the created route. Figure 7 shows the flow chart of the route set generation algorithm.
In this thesis the route set generation algorithm shown in Figure 7 was executed between all origin and destination centroids in each TAZ-pairs in the network, rather than between all origin and destination centroids separately. All possible combinations of the origin centroids and destination centroids in each TAZ was found. If the cost of a generated route exceeds the cost of the shortest path multiplied with the threshold value, the algorithm tries to generate a route between the next possible combination of origin and destination centroid. When all combinations exceed the cost of the shortest path multiplied with the threshold value, the route set is completely generated between that TAZ-pair and the costs are reset before generating the route set for the next TAZ-pair.
6 Experimental setup and tools
This chapter of the report presents the experimental setup of the methods and tools that have been used in this thesis.

6.1 Computer tools
In this thesis a variation of computer software have been used, which will be discussed in the following sections.

6.1.1 QGIS
When it comes to analyze, visualize and editing geospatial data like links and nodes in a traffic network, QGIS is a useful tool. QGIS is an open-source cross-platform desktop software. The software works as a geographic information system, which provides the user with comprehensive possibilities of importing and exporting layers and editing and analyzing different geographical information, like GPS observations. The software is based on the use of different layers. QGIS offers several layer types, but the most common is vector and raster layers. Vector layers consists of three different geometry types, which are points, lines, and polygons (Gisgruppen 2014). These geometry types could either be created in the graphical user interface (GUI) or by typing directly in the python console. Another import feature is the possibility to import shape files into QGIS. A shapefile contains geospatial vector data for GIS software. In a traffic network model, intersections are represented by points, links are presented by lines and zones are represented by polygons. The raster layer consists of a grid or a raster of attribute values. These attribute values could be for example different colors representing the density of GPS observations, shown in a heatmap.

One important advantage in QGIS is the possibility of using plugins and databases. One can easily manage to import or export information to and from a database. QGIS has played a central role in this thesis regarding the visualization of the GPS observations and the network. The network has been created in the Aimsun software and then been imported as a shapefile into QGIS. The Python console, which is a plugin to QGIS has been useful for creating scripts.

6.1.2 pgAdmin and PostgreSQL
pgAdmin is a software for developing and administrate PostgreSQL databases. PostgreSQL is operated as many other types of databases where the information is stored in tables and SQL queries are used to obtain and write data from and to the database. PostgreSQL comes with
many different extensions. Extensions used in this thesis are PostGIS to handle geometry and pgRouting for shortest path calculations (pgRouting Contributors 2017; PostGIS 2018).

6.1.3 Python

Python is an interpreted, object-oriented, high-level programming language designed to be used for general purpose programming (Python Software Foundation 2018). This means that the software can be used in a wide variety of applications. The programming language is an open source software, which makes it freely to use. Python has an efficient and readable syntax, it allows the users to express the code with fewer lines, compared with other programming languages like C++ and Java for example. This is because Python does not use curly brackets or end statements to delimit different code blocks, it delimits the code using whitespace and tabs.

Python is a scripting language. A script is a small program or code block which can be developed to execute some certain tasks or to interact with another computer software. A python script is compiled automatically by being interpreted from the first line to the last line of the code block, without being compiled separately (Sanner 1999). This feature of using python scripts can be very useful for reviewing, running, and testing smaller parts of the program´s functionality.

Python has been used in this thesis to write scripts that interact with QGIS and PostgreSQL database. The scripts make it possible to import and export information to the database. This could be for example to update the cost of the links in the link penalty-based route set generation algorithm and then store the new created paths into a table in the database on a certain format.

6.1.4 MATLAB

MATLAB is a matrix-based high-performance programming language for technical computations. One thing that differ MATLAB from other programming languages is that all variables are structured as arrays or matrices. The language allows matrix operations, visualization of data and implementations of functions and algorithms. In this thesis, MATLAB has been used for the computations regarding the calibration of the route set generation algorithm (MATLAB 2018).

6.2 Extracted network from Aimsun

The network is a road stretch of the Interstate 210 freeway east of Los Angeles with approximately 2.2 miles arterials both south and north of the freeway. The road stretch is
approximately 17 miles long and consists of over 1000 lane-miles of roadway. The network of the model consists of:

- 4064 sections (including the transit line *Golden Line*)
- 1592 centroid connectors
- 1620 nodes
- 389 centroids (with centroid connections)
- 86 Traffic Analysis Zones (TAZ).

The sections, nodes, and centroids are extracted as shapefiles and loaded into QGIS and a PostgreSQL database, both for visualization and for script development in the Python console in QGIS. Figure 8 shows the sections as black line segments, nodes and centroids as green dots of the model imported as shapefiles into QGIS.

![Figure 8 - Extracted network from the microsimulation model in Aimsun](image)

All imported layers (sections, nodes, centroids, TAZ, centroid connectors) have different attributes that both describes the characteristics as well how they relate to each other. The following attributes in the section layer have been used in this thesis:

- **ID**: Unique ID for each section in the network
- **Speed**: Speed limit on the section
- **From node**: Defines which node in the node layer the section starts from
- **To node**: Defines which node in the node layer in the section ends at

The centroids attributes of interest are:

- **ID**: Unique ID for each centroid in the network.
- **Type**: Defines if traffic starts, ends or both starts and ends at the centroid.
In the node layer, the attribute ID has been used to keep track of the unique ID of each node. These IDs matches the ones in the attributes From node and To node in the section layer. Each object also has a geometry attribute defining where they are in relationship to each other. This geometry can be represented by both latitude and longitude as well as more general X and Y coordinates in QGIS. As mentioned, the routes generated in this network have been defined between TAZ rather than between centroids. Each TAZ contains several centroids where trips can start and end. Figure 9 shows the TAZ containing the centroids.

![Figure 9 - The TAZ in the network with the containing centroids](image)

To perform shortest path calculations in this network using PostgreSQL and pgRouting the following modifications were done:

- Removal of the transit line in the network.
- Merge the centroid layer and the node layer. To be able to start and end a trip at a centroid the centroids and nodes where merged into the same layer.
- Connect the centroids with the section layer. This was done by adding nodes between the section layer and the centroid connector layer. This makes it possible for trips to start at a centroid, move out on the centroid connector, and then enter the section layer.
- Add and modify attributes in the section layer. An attribute describing the reversed cost was added in order to handle one-way sections in the network. The cost attribute on the centroid connections were set to a high value to ensure that no routes are using the centroid connections as ordinary sections. Attributes for the travel time based on length and speed were also added on each section.
- Add attribute in the node layer. An attribute describing which TAZ each node belongs to were added in the node layer.
With these modifications, it is possible to perform shortest path calculations in the network using pgRouting, both by using specific origin and destination nodes, and also TAZ to TAZ calculations using the TAZ attribute in the node layer.

6.3 GPS observations

Due to regulations and integrity of the consumers, the company HERE Technologies cannot share specific information about the provider of the data, for example of which company that provides the data or which kind of device that have been used. However, they can provide information if the GPS observations are of the type “fleet” or “consumer”. It could be the case that the “fleet” GPS observations are provided by vehicles of more commercial use. That could be for example, taxi companies, emergency vehicles or carrier companies. The “consumer” GPS observations are most likely private users, which uses navigation applications on their devices. Different “consumer” GPS observations could be provided by different applications.

Each measurement of the GPS observations gives the following information:

- **Latitude and Longitude**: Position of the GPS observations
- **Device ID**: The identification number of the device sending information
- **Sample date**: The date and time when the measurement was taken
- **Heading**: The heading of the trajectory from 0 to 360 degrees
- **Speed**: Point speed in km/h
- **Data provider**: Fleet or Consumer
- **System date**: The date and time when the measurement was read into the system

The device ID of the GPS observations changes with an unknown frequency of time, which means that one vehicle can form a travel pattern with several different device IDs.

To know if a day can be considered regular, historical Performance Measurement System (PeMS) data have been used. This information and categorization have been obtained from previous research at the University of California, Berkeley (UC Berkeley | ITS PATH 2018). An example of the historical PeMs data is shown in Figure 10.
To categorize a day as regular, the a.m. and p.m. measurements must be of the type regular, for both westbound and eastbound traffic. Figure 10 shows that January 15\textsuperscript{th}, 2014 can be considered a regular day.

The GPS observations represent regular days from 6 Tuesdays, 11 Wednesdays, and 6 Thursdays. This sums up to 23 regular days in total. These days were the only days without any reported incidents during the first half of 2014. Mondays, Fridays, and weekends are not included in the data set, due to the possibility of differences in traffic flow and also since one major incident that will be investigated occurred on a Thursdays. In order to get more reliable results from the raw data, it needs to be filtered. If the difference between the sample time and system time is too large it could indicate that the data is unreliable (University of California, Berkeley 2018). To find an acceptable threshold value of the difference, an investigation of the distribution of the difference will be done. Before this, all GPS observations with system time earlier than the sample time will be removed, since this indicates that something went wrong when logging the data to the system.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|c|}
\hline
& WB & & EB & \\
\hline
måns & 01/13/2014 & incident & regular & regular & regular \\
\hline
tis & 01/14/2014 & regular & regular & regular & incident \\
\hline
ons & 01/15/2014 & regular & regular & regular & regular \\
\hline
tor & 01/16/2014 & regular & regular & incident & incident \\
\hline
fre & 01/17/2014 & regular & regular & regular & regular \\
\hline
\end{tabular}
\caption{Historical PeMs data}
\end{table}
6.4 Incident day

The major incident that will be investigated took place April 24\textsuperscript{th}, 2014. The incident involved two trucks, where one of the trucks ended up on the electrified Metro Gold Line, and the other one landed on its side on the freeway. All eastbound lanes were blocked for approximately 3 hours, from 1 p.m. to 4 p.m. After 3 hours the right three lanes were opened. The last lanes were opened late in the evening (Day 2014). Figure 11 shows a historical photo from the day of the incident and the location in traffic network (Google Earth Pro 2018).

![Figure 11 – Historical photo from the day of the major incident, April 24\textsuperscript{th}, 2014, together with its location in the traffic network](image)

As shown in Figure 11, the incident occurred on the eastbound lanes of the I-210 freeway at Allen Avenue. The red dot at the bottom of Figure 11 represents the location of the incident in the traffic network.
6.5 Experiments

To investigate the structure of the GPS observations, which will be a calibration tool for the route set generation algorithm, a set of experiments will be performed. All experiments will be performed for regular days and the incident day separately. Initially, experiments regarding the GPS observations will be done to investigate the quantity of data that is available aggregated by hour, as well as the ratio between fleet and consumer provider. This can answer if there are any significant differences between regular days and the incident day regarding number of GPS observations obtained. A geographical investigation of the GPS observations throughout the network will be done to know where the density of the GPS observations is in the network. Regarding the map matching algorithm, all GPS observations will be mapped if they satisfy the criteria in the algorithm. Therefore, statistics from the map matching algorithm can be obtained, as well as general information about average heading and distance resemblance. The method for inferring paths from GPS observations will be used for all trips identified from the GPS observations to answer general questions about the trip generation, such as average length and duration of the trips, and number of trips generated etcetera. To investigate which main arterials and proportion of these that are used in the network, the network will be divided into investigation zones, see Figure 12.

![Network divided into 5 investigation zones](image)

A simplification of the zones in Figure 12 with possible eastbound paths is shown in Figure 13.
Different cases describing in which order the travelers passes these zones in the network will be investigated for the GPS observations. This will give information about proportions of the travelers using different travel patterns in the network, during different circumstances. The investigated cases with description are shown in Figure 14.

Note that the cases in Figure 14 are only eastbound. The choice of only investigating eastbound travel patterns is due to most of the traffic going eastbound during the investigated hours of the incident, which are 1 p.m. to 8 p.m. The incident also blocked the eastbound lanes.

To calibrate the route set generation algorithm, the paths from the GPS observations and the route set generation algorithm will be compared. Different parameter setups in the route set generation algorithm will be tested to investigate which setup that gives the best result regarding percentage of matched links in the different routes, number of generated routes, and the degree of overlap in the generated route sets. The method used for calculating the degree of overlap is described in section 4.2. To get more significance in the comparisons, two different zone pairs will be used. One short zone pair where a rather low number of paths were identified from the
GPS observations, and one longer zone pair where a larger number of paths were identified. Figure 15 shows the two zone pairs that were used for calibration.

![Figure 15 - The two zone pairs used for calibration of the route set generation algorithm](image)

The final parameters will be decided by looking on how many routes that are generated, percentage of matched links, and the degree of overlap in the route sets. According to Hu & Chiu (2015) a reasonable degree of overlap in a route set should be around 25-35%. To understand how the parameters in the route set generation algorithm affect the results, experiments will be performed with one of the parameters fixed, while varying the other parameter. Depending on the application of the route set, different goals and performance metric could be suitable. In this thesis it is desirable to generate as few routes as possible, while still having a high percentage of matched links, and also an overlap between 25-35%.

When the route set is generated the zones affected by the incident area will be filtered out. Between these affected zones, new route sets will be generated with both free flow travel times and with travel times based on GPS observations during the incident day. The requirement for forming an average speed on a link in the network will be set to at least 5 GPS observations mapped to that section. The links that does not meet that requirement will be assigned a percentage of the free flow speed. The percentage will be based on the results from the paths identified from the GPS observations. For the route sets representing the incident day, no routes will be allowed on the incident links.
7 Results & Analysis

In this chapter, the results and analyses will be discussed. First, some results regarding the GPS observations in general will be presented. The results from the map matching algorithm and route set estimation using GPS observations will handle both regular days and the incident day that took place April 24th, 2014. In section 7.4, the results from the route set generation algorithm will be presented, with comparisons of the estimated routes from the GPS observations. The route sets analyzed from the route set generation algorithm will represent both regular days and the incident day.

7.1 GPS observations

When presenting the results regarding the GPS observations from HERE technologies, no numbers showing the number of GPS observations, sample rate, or penetration rate will be presented. This is due to the terms of use of the GPS observations from HERE technologies. In some cases, graphs will be presented, but with the axis showing any of these measurements blank, just to present the general pattern of the data.

As mentioned, 23 regular days and one incident day have been used as input data. To get more reliable data, it was filtered before used in the methods. As mentioned in section 6.3, all GPS observations with system time earlier than the sample time was removed. Figure 16 shows a histogram of the distribution of the difference between the system time and sample time (positive difference).

![Figure 16 - Distribution of the difference between system time and sample time](image)

---

35
Figure 16 shows the frequency of the GPS observations on the y-axis with the difference in 10 seconds interval on the x-axis. The blue line in the histogram shows the break point in the data set at approximately 360 seconds. Therefore, all GPS observations with a difference larger than 6 minutes was removed from the data. This means that approximately 16.9 % of the GPS observations were removed with this requirement. The break point was defined visually where the graph tends to converge.

Table 1 shows the percentage of GPS observations that were filtered out for each day.

<table>
<thead>
<tr>
<th>Day</th>
<th>Percentage filtered out</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-01-15</td>
<td>29.8 %</td>
</tr>
<tr>
<td>2014-01-29</td>
<td>50.9 %</td>
</tr>
<tr>
<td>2014-02-06</td>
<td>45.8 %</td>
</tr>
<tr>
<td>2014-02-11</td>
<td>9.5 %</td>
</tr>
<tr>
<td>2014-02-25</td>
<td>8.6 %</td>
</tr>
<tr>
<td>2014-02-26</td>
<td>8.6 %</td>
</tr>
<tr>
<td>2014-02-27</td>
<td>11.6 %</td>
</tr>
<tr>
<td>2014-03-05</td>
<td>7.9 %</td>
</tr>
<tr>
<td>2014-03-06</td>
<td>9.3 %</td>
</tr>
<tr>
<td>2014-03-11</td>
<td>3.5 %</td>
</tr>
<tr>
<td>2014-03-12</td>
<td>5.6 %</td>
</tr>
<tr>
<td>2014-03-13</td>
<td>3.7 %</td>
</tr>
<tr>
<td>2014-03-18</td>
<td>4.4 %</td>
</tr>
<tr>
<td>2014-03-19</td>
<td>4.3 %</td>
</tr>
<tr>
<td>2014-03-26</td>
<td>19.5 %</td>
</tr>
<tr>
<td>2014-03-27</td>
<td>24.7 %</td>
</tr>
<tr>
<td>2014-04-08</td>
<td>9.4 %</td>
</tr>
<tr>
<td>2014-04-09</td>
<td>3.7 %</td>
</tr>
<tr>
<td>2014-04-16</td>
<td>3.0 %</td>
</tr>
<tr>
<td>2014-04-24 (Incident)</td>
<td>18.0 %</td>
</tr>
<tr>
<td>2014-04-30</td>
<td>7.9 %</td>
</tr>
<tr>
<td>2014-05-15</td>
<td>2.5 %</td>
</tr>
<tr>
<td>2014-05-20</td>
<td>1.8 %</td>
</tr>
<tr>
<td>2014-05-28</td>
<td>57.1 %</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>16.9 %</strong></td>
</tr>
</tbody>
</table>
Table 1 shows that the percentage of data being filtered out for each day is in the interval 1.8% to 57.1%. However, there are three rather obvious outliers in the set of days in the input data with a percentage from 45.8% to 57.1%. This of course increase the average percentage. Without these outliers the average percentage of filtered data would decrease to 9.4%. As mentioned, the number of GPS observations obtained for each day cannot be presented due to the terms of use. However, the number of GPS observations obtained during the incident day does not deviate from the number of GPS observations during the regular days.

Figure 17 shows the distribution of number of GPS observations, aggregated by hour, for the 23 regular days and for the incident day at April 24th, 2014. The y-axis showing the number of observations is blank due to the terms of use for the GPS observations.

![Distribution of the GPS observations, aggregated by hour](image)

As expected, Figure 17 shows that the number of GPS observations obtained is highest during daytime with small peaks around the morning and afternoon peak hours. However, the result also shows that during some of the incident hours from 2 p.m. to around 6 p.m., the number of GPS observations obtained is 2-3 times more during the incident day, compared with a regular day. This result is most significant for the Consumer provider. One possible explanation for this could be that during the incident the network is more congested with higher density of vehicles on the road sections. This could lead to more users in the network that can send GPS data to the system. Another explanation could be that users tends to use their GPS devices more when there is an incident in the network, due to the increased desire for suggested reroutes.

Figure 18 shows the percentage of Fleet and Consumer GPS observations during 23 regular days and during the incident day at April 24th, 2014, aggregated by hour.
As shown in Figure 18, the majority of the GPS observations are obtained from the Fleet provider.

Figure 19 shows a heat map of the density of the GPS observations throughout the network during 23 regular days, and all 24 hours.

Figure 19 shows that the density is highest on the I-210 freeway. This result is expected because the corridor is frequently used by users commuting to and from Los Angeles.

Figure 20 shows a heat map of the density of the GPS observations throughout the network at the incident day April 24th, 2014, for all 24 hours.
Figure 20 shows small difference in the heat map during the incident day, compared with the regular days. The density of GPS observations seems to be a bit higher west of the incident area, and a bit lower on the east side of the incident area. However, this comparison could give a larger difference when only investigating the actual time of the incident, which was roughly about 1 p.m. to 8 p.m.

One thing to mention regarding the heatmap comparisons between regular days and the incident day in this study is that the colors in the heatmaps does not correspond to the same density, when comparing regular days with the incident day. This is rather to give an indication of where the dense areas in the network are during different circumstances.

Figure 21 shows a heat map of the density of the GPS observations throughout the network for 23 regular days, during the time of the incident, 1 p.m. to 8 p.m.

As Figure 21 indicates, no large differences in the heat map can be seen for the regular days during 1 p.m. to 8 p.m., compared with the heatmap in Figure 19, during all 24 hours.

Figure 22 shows a heat map of the density of the GPS observations throughout the network at the incident day April 24th, 2014, 1 p.m. to 8 p.m.
In Figure 22, a significant difference in the density of the GPS observations can be seen compared with the regular days. During the incident 1 p.m. to 8 p.m. a very low density of GPS observations can be observed east of the incident area, and a very high density west of, and around the incident area. A possible reason for the low density east of the incident area could be that the incident blocked all eastbound traffic. Since the incident took place on the afternoon, many travelers going home from work in Los Angeles was held up in the congestions west of the incident. The number of travelers going westbound to Los Angeles during the afternoon is much lower than the number of travelers going eastbound on the afternoon. This is because most of the travelers that are commuting, lives in the outer areas of Los Angeles, with their workplaces downtown Los Angeles, and not the other way around.

7.2 Map matching algorithm

Figure 23 shows a selection of 300 mapped GPS observations in the network. The legend describes the intervals of the resemblance.
As shown in Figure 23 the red GPS observations have the lowest resemblance but are still a sufficiently good match to be mapped to the network. Approximately one third of this selection of GPS observations were mapped with a high resemblance of 0.8207 – 0.9918.

Table 2 shows the map matching results from the 23 regular days and the incident day at April 24th, 2014.

<table>
<thead>
<tr>
<th></th>
<th>23 regular days</th>
<th>Incident day at April 24th, 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of mapped GPS observations</td>
<td>70.03 %</td>
<td>73.97 %</td>
</tr>
<tr>
<td>Average distance resemblance</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>Average heading resemblance</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Average deviation in distance (m)</td>
<td>5.46</td>
<td>5.41</td>
</tr>
<tr>
<td>Average deviation in heading (degrees)</td>
<td>4.67</td>
<td>4.94</td>
</tr>
</tbody>
</table>

As shown in table 2, only around 70 % of the GPS observations met the requirements to be mapped to the network. This means that an additionally 30 % of the GPS observations were removed when mapping them to the network. The average deviation in distance between the road section and the observation was 5.46 meters for regular days and 5.41 meters for the incident day. The average deviation in heading was 4.67 degrees for regular days and 4.94 for the incident day.
7.3 Inferred paths from GPS observations

Table 3 shows general results from the trip generation process.

<table>
<thead>
<tr>
<th></th>
<th>23 regular days</th>
<th>Incident day at April 24th, 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of GPS observations used in routes</td>
<td>53.81 %</td>
<td>67.71 %</td>
</tr>
<tr>
<td>Number of routes</td>
<td>48 113</td>
<td>2 776</td>
</tr>
<tr>
<td>Average length of the routes</td>
<td>9.08 mi.</td>
<td>7.36 mi.</td>
</tr>
<tr>
<td>Average true travel time of the routes</td>
<td>15 min, 50 s</td>
<td>17 min, 40 s</td>
</tr>
<tr>
<td>Average free flow travel time of the routes</td>
<td>9 min, 38 s</td>
<td>8 min, 14 s</td>
</tr>
<tr>
<td>Average true speed of the routes</td>
<td>34.44 mph</td>
<td>24.98 mph</td>
</tr>
<tr>
<td>Average free flow speed of the routes</td>
<td>56.63 mph</td>
<td>53.61 mph</td>
</tr>
<tr>
<td>Percentage of free flow speed of the routes</td>
<td>60.82 %</td>
<td>46.60 %</td>
</tr>
</tbody>
</table>

One interesting result from the trip generation from the GPS observations is that even if the average length of the trips during the regular days are longer than for the incident day, the average duration for each trip is shorter for the regular days. During the regular days the average length of a trip is 9.08 miles, with an average duration of 15 minutes and 50 seconds. During the incident day the average length of a trip is 7.36 miles, with an average duration of 17 minutes and 40 seconds. These results can also be seen in the percentage of the free flow speed on the trips. During the regular days, the speed on the trips was around 60.82 % of the free flow speed, while for the incident day, the speed was around 46.60 % of the free flow speed.

When investigating trips going from zone 1 to zone 4 described in Figure 13, section 6.5, not all possible combinations of travel patterns between these zones were investigated in this thesis. However, the total number of identified trips from zone 1 to zone 4 were extracted and are presented in Figure 24.
Figure 24 shows that no significant differences can be seen in total number of trips from zone 1 to zone 4. An explanation for this could be that users tends to stay in queue and wait out the incident, or take reroutes near the I-210 freeway, rather than leaving the corridor completely.

Figure 25 shows the total number of identified trips in case 1, which is trips north of the I-210 freeway.
Even if rather few trips could be identified north of the freeway, a significant difference can be seen when comparing regular days with the incident day. During regular days approximately 1 trip/day could be identified, compared with 5 trips during the incident day.

Figure 26 shows a heatmap of observed paths north of the I-210 freeway for regular days.

![Figure 26 - Heatmap of observed paths north of the I-210 freeway for regular days](image)

Figure 26 shows that the main arterial used on regular days north along the I-210 freeway was Orange Grove Blvd and then down to Rosemead Blvd. The high-density areas indicate queues at intersections with traffic signals. This allows devices to send more data and more vehicle gather upstream of the traffic signal which also increase the number of GPS observations being transmitted from the GPS devices. These “hotspots” in density are often located in connection with intersections and traffic lights and can give valuable information when making response plans and optimizing signal schemes in traffic signals.

Figure 27 shows a heatmap of observed paths north of the I-210 freeway for the incident day.
As shown in Figure 27 the main arterial being used north of the I-210 freeway during the incident day is the same as during the regular days, that is, Orange Grove Blvd down through Rosemead Blvd. On the west part of Orange Grove Blvd there are long queues at the traffic signals in the intersection of Orange Grove Blvd and Lake Avenue. This could give a hint that this intersection should be investigated more during this kind of incidents.

Figure 28 shows the total number of identified trips in case 2, which is trips on the I-210 freeway.
Figure 28 also shows a significant difference in number of trips on the freeway between regular days and the incident day. During the regular days the median number of trips in this case was around 120 trips, while for the incident day 54 trips were identified.

Figure 29 shows a heatmap of observed paths on the i210 freeway for regular days.

![Heatmap of observed paths on the I-210 freeway for regular days](image)

In Figure 29 an even density of GPS observations can be seen along the I-210 freeway during regular days.

Figure 30 shows a heatmap of observed paths on the I-210 freeway for the incident day.
Figure 30 - Heatmap of observed paths on the I-210 freeway for the incident day

Figure 30 shows that during the incident day the density is highest west of the incident. East of the incident shows very low density of GPS observations. This was a rather expected result because the incident blocked all eastbound lanes of the I-210 freeway.

Since the incident occurred on the I-210 freeway, an investigation regarding the cumulative number of trips in case 2 would be interesting. Figure 31 shows the cumulative number of trips on the I-210 freeway.

Figure 31 - Cumulative number of trips on the I-210 freeway
Figure 31 shows that the cumulative graph for number of trips during the incident is rather flat between 1 p.m. – 4 p.m. This is the time interval when all lanes were blocked. After 4 p.m. some of the right lanes were opened for traffic which is indicated by the increased slope of the graph at this time. The reason for why the graph is not completely flat is that the graph shows when the trip starts in zone 1, rather than passing the incident. However, the trips filtered out in this case starts in zone 1, passing through zone 5 (freeway), and then ends in zone 4.

Figure 32 shows the total number of identified trips in case 3, which is trips south of the I-210 freeway.

![Case 3 - Number of identified trips](image)

*Figure 32 - Number of identified trips in case 3, south of the I-210 freeway*

As well as in case 1 and case 2, case 3 also shows a significant difference in number of trips, when comparing regular days with the incident day.

Figure 33 shows a heatmap of observed paths south of the I-210 freeway for regular days.
Figure 33 shows that the main arterial used on regular days south along the I-210 freeway is Colorado Blvd, and then south on Santa Anita Avenue. Some indication of usage can also be seen on California Blvd.

Figure 34 shows a heatmap of observed paths south of the I-210 freeway for the incident day.

Figure 34 shows that during the incident day, a lot more main arterials were used, compared with the regular days. Colorado Blvd is still used but is no longer the main arterial. During the
incident day, Corson St and Del Mar Blvd were used for most of the paths. It can also be seen that Santa Anita Ave is no longer the main arterial for southbound traffic. During the incident Rosemead Blvd and Sunset Blvd are used instead for southbound traffic.

Based on the experiments in case 1, 2, and 3, Figure 35 shows the travel times on the main arterials.

![Travel times on the arterials](image-url)

Figure 35 shows that in general, the travel times during the incident are higher, compared with the regular days. The median of the travel times for the incident day is higher than the upper quantile for the 23 regular days, for all of the arterials investigated. Another interesting result is that the boxes representing the regular days indicates a rather tight distribution of the travel times on the arterials, while some of the distributions for the incident day is very spread. An example of this is the travel times on Orange Grove Blvd. During regular days the travel times on this arterial goes from approximately 2 min/mi. to 3 min/mi. The travel times during the incident day for the same arterials lays in the interval 2.6 min/mi – 5.2 min/mi. approximately. This gives indications that on some of the arterials, the travel times are very similar from day to day during regular days.
Figure 36 shows the speeds on the main arterials.

![Figure 36 - Speed on the main arterials](image)

The speeds shown in Figure 36 is based on the same values as the travel times and are only the inverse of these values.

### 7.4 Route set generation algorithm and parameters

In this section of the report the results from the calibration process of the route set generation algorithm will be presented. Results from the final route sets during different circumstances will also be analyzed and discussed.

#### 7.4.1 Calibration of the parameters

Figure 37 presents the coverage of matched links of routes observed from GPS observations compared with routes from the route set generation algorithm, when the threshold parameter is fixed and \( \mu \) is varying for zone pair 1.
The lines in Figure 37 shows the cumulative percentage of matched links between the 67 routes from the GPS observations and the routes generated from the route set generation algorithm for 8 different experiments. One example is shown in the figure, where it can be seen that 40 of 67 routes from the GPS observations were covered with 60% in the route set, for experiment 1. The straight vertical lines to the left in the figure shows that all the experiments have routes that cover at most 32% of all the links in each of the 67 routes from the GPS observations. The difference in coverage of the paths between the different experiments is rather small when the threshold parameter is fixed and $\mu$ is increasing, while the number of routes increased rapidly from 61 to 358. The percentage of matched links in experiment 1 and 8 respectively, differs with approximately 5 – 12%. With this in mind, the small difference in coverage gives a significant higher number of generated routes.

Figure 38 presents the coverage of matched links from routes observed from GPS observations compared with routes from the route set generation algorithm, when $\mu$ is fixed and the threshold parameter is varying for zone pair 1.
The lines in figure 38 represents the cumulative percentage of matched links between the 67 routes from the GPS observations and the routes generated from the route set generation algorithm for 8 different experiments. The difference of coverage of the paths between the different experiments is rather small when $\mu$ is fixed and the threshold parameter is increasing, except for experiment 1, which stands out from the other experiments. However, the number of generated routes increased rapidly when $\mu$ was fixed and the threshold parameter increased. The explanation for this could be a combination of that the route set generation algorithm does not update the cost of the used links sufficiently and also that the algorithm finds acceptable routes far away from the shortest route. But it can also be seen from the figure that with a low value of the threshold parameter, very few number of routes are generated.

Figure 39 presents the coverage of matched links from routes observed from GPS observations compared with routes from the route set generation algorithm when the threshold parameter is fixed and $\mu$ is varying for the zone pair 2.
14 routes were identified from the GPS observations for zone pair 2. In Figure 39 it can be seen that one of the 14 routes from the GPS observations was not covered by any of the generated routes from the route set generation algorithm for any of the experiments. Experiment 1 deviates from the other experiments in the percentage of matched links. But the difference in coverage between the other experiments is rather small when $\mu$ increasing. Experiment 2 – 8 covers approximately between 82 – 91 % of the links for 11 of the 14 routes from the GPS observations, while experiment 1 only covers around 70 %.

Figure 40 presents the coverage of matched links from routes observed from GPS observations compared with routes from the route set generation algorithm, when $\mu$ is fixed and the threshold parameter is varying for zone pair 2.
From Figure 40, it can be seen that there is a larger spread of generated number of routes between the different experiments when changing the threshold parameter compared with the results when \( \mu \) varied in Figure 39. It can also be seen that the difference in coverage between the experiments is low. Experiment 1 only generated 22 routes but had the same or even better coverage for at least 8 of the 14 routes from the GPS observations compared with the other experiments. For this specific zone pair, a high value of the threshold parameter generates more number of trips while the coverage is almost the same.

Figure 41 presents the degree of overlap between the routes from the route set generation algorithm, when the threshold parameter is fixed and \( \mu \) is varying for both zone pair 1 and 2.
Figure 41 - Degree of overlap between routes from the route set generation algorithm, when the threshold parameter is fixed and $\mu$ is varying

Figure 41 shows that the degree of overlap increase when the value of $\mu$ is increasing. This applies for both of the zone pairs and the reason for this is because of that $\mu$ exist in the denominator of the link penalty definition, which is described in section 5.3. A lower value of $\mu$ than those who have been tested would give a much lower degree of overlap, since the cost term would go to infinity when $\mu \to 0$.

Figure 42 presents the degree of overlap between routes from the route set generation algorithm, when $\mu$ is fixed and the threshold parameter is varying for both zone pair 1 and 2.
Figure 42 - Degree of overlap between routes from the route set generation algorithm, when $\mu$ is fixed and the threshold parameter is varying.

From Figure 42, it can be seen that the degree of overlap decreases when the value of the threshold parameter is increasing. The difference of overlap is bigger for smaller threshold values, which can be seen especially for zone pair 1. The degree of overlap seems to converge against approximately 9% when the threshold value becomes larger and the value of $\mu$ is fixed for these two zone pairs.

Table 4 shows the summary of the results from the set of experiments for zone pair 1. 67 trips were identified from the GPS observations between these zones.

Table 4 - Summary of the results from the set of experiments for zone pair 1

<table>
<thead>
<tr>
<th>Exp.</th>
<th>$\mu$</th>
<th>Threshold</th>
<th>Generated routes</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.50</td>
<td>2.00</td>
<td>61</td>
<td>11.53%</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>2.00</td>
<td>100</td>
<td>12.68%</td>
</tr>
<tr>
<td>3</td>
<td>1.25</td>
<td>2.00</td>
<td>126</td>
<td>12.56%</td>
</tr>
<tr>
<td>4</td>
<td>1.50</td>
<td>2.00</td>
<td>143</td>
<td>12.85%</td>
</tr>
<tr>
<td>5</td>
<td>2.00</td>
<td>2.00</td>
<td>178</td>
<td>13.16%</td>
</tr>
<tr>
<td>6</td>
<td>3.00</td>
<td>2.00</td>
<td>246</td>
<td>13.56%</td>
</tr>
<tr>
<td>7</td>
<td>4.00</td>
<td>2.00</td>
<td>303</td>
<td>13.86%</td>
</tr>
<tr>
<td>8</td>
<td>5.00</td>
<td>2.00</td>
<td>358</td>
<td>14.10%</td>
</tr>
<tr>
<td>9</td>
<td>1.00</td>
<td>1.25</td>
<td>13</td>
<td>50.02%</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>1.50</td>
<td>32</td>
<td>25.24%</td>
</tr>
<tr>
<td>11</td>
<td>1.00</td>
<td>2.00</td>
<td>100</td>
<td>12.68%</td>
</tr>
<tr>
<td>12</td>
<td>1.00</td>
<td>2.50</td>
<td>182</td>
<td>10.64%</td>
</tr>
<tr>
<td>13</td>
<td>1.00</td>
<td>3.00</td>
<td>259</td>
<td>9.82%</td>
</tr>
<tr>
<td>14</td>
<td>1.00</td>
<td>4.00</td>
<td>393</td>
<td>9.22%</td>
</tr>
<tr>
<td>15</td>
<td>1.00</td>
<td>5.00</td>
<td>512</td>
<td>9.05%</td>
</tr>
<tr>
<td>16</td>
<td>1.00</td>
<td>6.00</td>
<td>617</td>
<td>8.95%</td>
</tr>
</tbody>
</table>
Table 5 shows the summary of the results from the set of experiments for zone pair 2. 14 trips were identified from the GPS observations between these zones.

**Table 5 - Summary of the results from the set of experiments for zone pair 2**

<table>
<thead>
<tr>
<th></th>
<th>( \mu )</th>
<th>Threshold</th>
<th>Generated routes</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 1</td>
<td>0.50</td>
<td>2.00</td>
<td>42</td>
<td>9.58 %</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>1.00</td>
<td>2.00</td>
<td>70</td>
<td>10.63 %</td>
</tr>
<tr>
<td>Exp. 3</td>
<td>1.25</td>
<td>2.00</td>
<td>82</td>
<td>10.92 %</td>
</tr>
<tr>
<td>Exp. 4</td>
<td>1.50</td>
<td>2.00</td>
<td>94</td>
<td>11.05 %</td>
</tr>
<tr>
<td>Exp. 5</td>
<td>2.00</td>
<td>2.00</td>
<td>116</td>
<td>11.24 %</td>
</tr>
<tr>
<td>Exp. 6</td>
<td>3.00</td>
<td>2.00</td>
<td>137</td>
<td>11.82 %</td>
</tr>
<tr>
<td>Exp. 7</td>
<td>4.00</td>
<td>2.00</td>
<td>161</td>
<td>12.32 %</td>
</tr>
<tr>
<td>Exp. 8</td>
<td>5.00</td>
<td>2.00</td>
<td>179</td>
<td>12.60 %</td>
</tr>
<tr>
<td>Exp. 9</td>
<td>1.00</td>
<td>1.25</td>
<td>22</td>
<td>16.96 %</td>
</tr>
<tr>
<td>Exp. 10</td>
<td>1.00</td>
<td>1.50</td>
<td>39</td>
<td>12.74 %</td>
</tr>
<tr>
<td>Exp. 11</td>
<td>1.00</td>
<td>2.00</td>
<td>70</td>
<td>10.63 %</td>
</tr>
<tr>
<td>Exp. 12</td>
<td>1.00</td>
<td>2.50</td>
<td>95</td>
<td>10.03 %</td>
</tr>
<tr>
<td>Exp. 13</td>
<td>1.00</td>
<td>3.00</td>
<td>118</td>
<td>9.66 %</td>
</tr>
<tr>
<td>Exp. 14</td>
<td>1.00</td>
<td>4.00</td>
<td>156</td>
<td>9.42  %</td>
</tr>
<tr>
<td>Exp. 15</td>
<td>1.00</td>
<td>5.00</td>
<td>186</td>
<td>9.35  %</td>
</tr>
<tr>
<td>Exp. 16</td>
<td>1.00</td>
<td>6.00</td>
<td>215</td>
<td>9.31  %</td>
</tr>
</tbody>
</table>

The set of experiments shown in table 4 and table 5 were performed to investigate how the parameters affects the results. Based on these results it can be seen in Figure 41, which shows the overlap when varying \( \mu \), that it is hard to achieve an overlap of 25-35 % in the route set with a threshold value of 2. Figure 42 also shows that a threshold value of around 1.5 can be a good starting point when further calibrating the parameters. Regarding the coverage of routes from the GPS observations a highly increased number of generated routes, does not seems to improve the percentage of matched links with the routes from the GPS observations significantly. Based on these results, a set of new combinations of parameters were tested and evaluated. These experiments are shown in Table 6.

**Table 6 - Set of experiments performed with different combinations of the parameters**

<table>
<thead>
<tr>
<th></th>
<th>( \mu )</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 1</td>
<td>1.25</td>
<td>1.50</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>1.25</td>
<td>1.25</td>
</tr>
<tr>
<td>Exp. 3</td>
<td>1</td>
<td>1.25</td>
</tr>
<tr>
<td>Exp. 4</td>
<td>1</td>
<td>1.4</td>
</tr>
</tbody>
</table>

The experiments in Table 6 were performed one at a time to see how the parameters should be calibrated.
Figure 43 presents the coverage of matched links from paths observed from GPS observations, for the experiments in table 6, zone pair 1.

Figure 43 shows that the percentage of matched links for these experiments is similar. However, experiment 1 and 4 seems to perform slightly better regarding the coverage, while experiment 2 and 3 generated fewer routes. A reasonable preference could be experiment 4, which generated rather few routes, but performed rather well regarding the coverage, compared with the other experiments.

Figure 44 presents the coverage of matched links from paths observed from GPS observations, for the experiments in table 6, zone pair 2.
As for zone pair 1, Figure 44 shows that the performance with these parameters for zone pair 2 is also similar. Experiment 1 has the highest percentage of matched links, but that experiment also generated many routes. Experiment 4 is better than experiment 2 and 3 for 9 of the 14 routes from the GPS observations. Also, for zone pair 2, experiment 4 could be a reasonable preference. However, Figure 43 and 44 only present the performance regarding the percentage of matched links and number of routes generated. The overlap is also of interest when deciding the final parameters for the route set generation algorithm. Table 7 shows the summary of the results including the overlap from the set of experiments with different combinations of parameters for zone pair 1.

Table 7 - Summary of the results from the set of experiments with different combinations of parameters for zone pair 1

<table>
<thead>
<tr>
<th></th>
<th>μ</th>
<th>Threshold</th>
<th>Generated routes</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 1</td>
<td>1.25</td>
<td>1.50</td>
<td>38</td>
<td>26.08 %</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>1.25</td>
<td>1.25</td>
<td>16</td>
<td>49.80 %</td>
</tr>
<tr>
<td>Exp. 3</td>
<td>1.00</td>
<td>1.25</td>
<td>13</td>
<td>50.02 %</td>
</tr>
<tr>
<td>Exp. 4</td>
<td>1.00</td>
<td>1.40</td>
<td>23</td>
<td>34.53 %</td>
</tr>
</tbody>
</table>

Table 8 shows the summary of the results from the set of experiments with different combinations of parameters for zone pair 2.

Table 8 - Summary of the results from the set of experiments with different combinations of parameters for zone pair 2

<table>
<thead>
<tr>
<th></th>
<th>μ</th>
<th>Threshold</th>
<th>Generated routes</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 1</td>
<td>1.25</td>
<td>1.50</td>
<td>44</td>
<td>13.01 %</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>1.25</td>
<td>1.25</td>
<td>25</td>
<td>19.09 %</td>
</tr>
<tr>
<td>Exp. 3</td>
<td>1.00</td>
<td>1.25</td>
<td>22</td>
<td>16.96 %</td>
</tr>
<tr>
<td>Exp. 4</td>
<td>1.00</td>
<td>1.40</td>
<td>32</td>
<td>14.43 %</td>
</tr>
</tbody>
</table>

Based on the results with the combinations of parameters, it is shown in table 8 that no settings of the parameters gave an overlap between 25-35 % for zone pair 2. However, for zone pair 1, both experiment 1 and 4 gave an overlap between 25-35 %. Regarding the percentage of matched links for zone pair 1 it can be seen in Figure 43 that the performance for both experiment 1 and 4 is similar, but with a lower number of generated routes in experiment 4. Based on the results in the calibration process of the parameters it was decided to use the parameters $\mu = 1$, and $Threshold = 1.4$.

Even if the overlap for zone pair 2 became under 25 %, zone pair 1 could be considered as more important to calibrate against, due to the higher number of identified trips from the GPS observations. These results also indicate that it is hard to use the same parameters for every zone pair in the network. Due to the variations in length between zones and number of origins
and destinations in each zone it would most probably be more accurate to use different parameters depending of the type of zone pair. However, in this thesis, the same parameters will be used for all zone pairs when generating the final route set for all zone pairs.

7.4.2 Route sets in the corridor

Three route sets were generated to investigate differences during different circumstances and with different travel times as costs on the links. The following route sets were generated:

- Scenario 1: Route set between all zone pairs with the free flow travel times on the links and no incident in the network
- Scenario 2: Route set between zone pairs affected by the incident with the free flow travel time on the links and with the incident links blocked. Zone pairs are considered to be affected by the incident when they have any routes that uses the incident links.
- Scenario 3: Route set between zone pairs affected by the incident with the travel time based on GPS observations on the links and with the incident links blocked

Figure 45 shows a colormap indicating how many times each links in the network have been used in the route set in scenario 1.

Figure 45 – Occurrences of each link in the route set in scenario 1

In Figure 45 the most used links are indicated with darker thicker lines. The figure also shows that the most used links in the route set are on the I-210 freeway. This result was expected due to the high free flow speed on those links, compared with other links in the network. The main arterials along the I-210 freeway such as Orange Grove Blvd, California Blvd, Colorado Blvd, and Del Mar Blvd are also used in many routes. These results were also expected and matches
the results from the investigation of the GPS observations rather well. When this route set was generated all of the zones with routes affected by the incident links were filtered out. A zone is considered affected by the incident if one or several routes from or to the zone includes one of the incident links. These zones are shown in figure 46.

![Figure 46](image)

*Figure 46 – TAZ affected by the incident of April 24th, 2014 in the route set in scenario 1*

The zones affected by the incident summed up to 31 origin zones and 56 destination zones. 3 of the zones did not contain any centroids and are shown in black in Figure 46. The zones affected by the incident were used in scenario 2 which represents a route set with free flow travel time and the incident links blocked. Figure 47 shows a colormap indicating how many times each links in the network occurred in the route set in scenario 2 and 3.
Figure 47 shows that during the incident with free flow speed, the routes in the incident area tend to just take the off-ramp upstream of the incident, then take the parallel arterial (Corson St), and then enter the I-210 freeway right downstream of the incident. This is an effect of only blocking the links where the incident occurred. This was an expected result when using a shortest path algorithm to generate a route set, due to that all surrounding links have the free flow travel time as cost. However, the surrounding links of the incident are also affected in a real scenario. That is, congestions occur upstream and around the incident causing higher travel times on those links, which would make the route set look different. Figure 48 shows a colormap indicating how many times each links in the network occurred in the route set in scenario 3.
Figure 48 shows that when using the speeds from the GPS observations as costs on the links the route set changes significantly. In scenario 3 a larger reroute area around the incident can be seen, compared with scenario 2, when free flow travel time was used as cost. This indicates that inside the area in the zoomed in part of Figure 48 the traffic was slow during the incident day, which leads to low speed and high cost on the links in this area. The arterials around the incident with lower travel times are Orange Grove Blvd north of the I-210 freeway and Cordova/Del Mar Blvd south of the I-210 freeway. Route sets based on speeds from GPS observations can be interpreted as routes the travelers should have used instead, rather than showing the routes actually used. It is also important to mention that in the colormaps in Figure 45, 47, 48 only visualize the number of occurrences for each link in the different route sets and have nothing to do with how many travelers that used each route.
Table 9 shows the general results from the route set generation scenarios.

**Table 9 – General results from the route set generation scenarios**

<table>
<thead>
<tr>
<th></th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of routes</td>
<td>108 411</td>
<td>31 542</td>
<td>44 626</td>
</tr>
<tr>
<td>No. of routes generated between each zone pair in average</td>
<td>15.93</td>
<td>20.21</td>
<td>28.59</td>
</tr>
</tbody>
</table>

Table 9 shows that for the large route set between each of the zone pairs less number of routes in average were generated, compared with scenario 2 and 3. The reason for this is probably that the large route set consists of many short zone pairs in the network, which will decrease the average number of routes. When using average speeds observed from the GPS observations as cost on the links, a significant higher number of routes were generated between the zone pairs affected by the incident, with the incident links blocked.
8 Discussion

The most crucial part in this thesis is the quality of the GPS observations. Almost all the results are based on, or directly affected by the GPS observations. The number of GPS observations provided from HERE technologies varies widely between different days, which could lead to uncertain results when it comes to investigations of single days, like a day with an incident for instance. In order to find the characteristics of regular days and, in addition, also try to reduce the randomness between different days, we used a set of regular days. These days were considered to be regular days according to the historical PeMs data. The set consisted of 23 regular days, where three of them deviated from the others, due to a low number of GPS observations. These three days can also affect the results with some uncertainties regarding the characteristics of a regular day. But since those were considered to be regular days according to PeMs data and they were provided like all other days, we wanted to include them as well. This is something to have in mind when investigating an incident that occurred on a single day. If there is lack of data for that particular day, it could be hard to analyze. But that was not the case for the major incident on April 24th, 2014, that have been investigated in this thesis. That day consisted of more GPS observations than the average of the regular days, which made the day useful and comparable.

The quality of the GPS observations in form of the sample rate and accuracy of the positions can also have a big impact on the results. When the sample rate of the GPS observations is low, it becomes hard to analyze what have happened between them. A fast-moving vehicle which sends information with a low sample rate will probably have traveled a large distance between each GPS observations, which makes it complicated to infer the actual path with high certainty. In the method for defining trips in this thesis, the requirement is that the time difference between two consecutive GPS observations needs to be less or equal to 2 minutes. This requirement can make the method to miss some of the actual trips in the corridor, due to low sample rate, but on the other hand, it will generate more accurate paths of those trips that are identified. Investigations regarding how many trips that the method miss have not been done in this thesis. If the data had higher sample rate in general, more paths would have been observed. The time requirement in the method could differ depending on the purpose of the application. If the aim is for example to only visualize the density of the GPS observations in the network, the requirement of a certain time constraint between each consecutive GPS observation may be unnecessary.
Another factor that affects the filtering process used in this thesis is the difference between the sample time and system time for the GPS observations. That is, the time it takes for the information to reach the system. The desirable maximum difference between these two timestamps could also depend on the purpose of the study. It may be the case that if the purpose is to generate trips or investigate the density of GPS observations throughout the network, this difference is less important than if the purpose is to use the GPS observations for more real time applications. It is hard to know with certainty how a large difference between the sample time and system time affects the results, but it could give a lower credibility in the measurement. It could also be the case that the device time or server time settings are wrong.

The accuracy of the position of the GPS observations is important when it comes to the map matching process. In this thesis, a rather simple map matching algorithm was used. This causes a part of the observations to be mapped to incorrect links, especially if the GPS observations are observed between two parallel links. The network in this study consists of many parallel links. The I-210 freeway has a High-Occupancy Vehicle (HOV) lane reserved for carpooling almost along the whole freeway. The lanes on the freeway are modelled in the network as two separated links, one for the HOV lane, and one for the regular lanes. However, in reality, this is not the case, the HOV lane is an extension to the regular lanes. The HOV lanes are separated from the regular lanes with refuges on some part of the network and with a solid line on other parts. This means that vehicles can on some part of the network change from the HOV lanes to the regular lanes even though it is not legal. Trips like this makes it hard to map the observations due to the structure of the network from Aimsun with separate sections on the freeway. This also causes problem to the method for inferring paths between each GPS observation. If several consecutive GPS observations are mapped to incorrect links, our method will either discards that trip or make unreasonable paths between some of the GPS observations. We have solved this for the majority of the trips by comparing the travel time of the shortest path between each consecutive GPS observations with the actual travel time between the GPS observations. If the travel time is unreasonable, then the method rejects that GPS observation and tries to calculate the shortest path to the next consecutive GPS observation. We investigated this empirically by a visualization of a subset of generated trips and could see that this improvement enhanced the inferred paths significantly. But since there are so many trips from the GPS observations, not all the generated trips were evaluated. It is not known with certainty how accurate this method is for the entire set of trips from the GPS observations. As a result of that, some of the trips from the GPS observations could be partially incorrect, which could affect the general results,
like the number of generated trips, average length on the trips and average speeds etcetera. But a good thing to do would be to implement a method to evaluate all the inferred trips from the GPS observations.

As mentioned, one thing that could solve many of the problems with incorrectly mapped GPS observations is to have a better map matching algorithm. Probably, a map matching algorithm that consider a sequence of GPS observations would give more accurate results. Instead of mapping the GPS observations one at a time, the algorithm should consider a whole sequence of GPS observations from a trip of observations. A good way to do this could be to combine our method of defining and obtaining trips with a more advance map matching algorithm. Then, the input to the map matching algorithm could be a sequence of GPS observations.

The data used in this thesis is from the first half of 2014 and the reason for that is because of the major incident that occurred on April 24th, 2014. However, it is possible that more recent data could be more accurate due to better and more devices in the network sending GPS observations to the system. In the data used in this thesis no information about the accuracy was included with each GPS observation. If information about this were known, GPS observations with low accuracy could have been filtered out, and possibly improve the results.

Generally, when working with large data sets in databases, it is important how the data is structured, to get efficient scripts and algorithms. Correct index on columns in the tables in the database can improve the running time of queries significantly. This have been a challenge during this study. The first idea was to use a server located in Sweden, even though the thesis was carried out in Berkeley, USA. It was early discovered that this was too time consuming and the databases were then stored internally on the computer used in Berkeley. This setup was more efficient, and it is possible to mirror the data base to the server in Sweden when the work is completed. Other possible ways of handling this amount of data would be to store data in Excel, or text files, for handling in for example Matlab. However, when using PostgreSQL, it is advantageously when it comes to visualize and analyze the data in QGIS. As mentioned, the most important aspect when using data bases would probably be how the data is stored, and how the different tables relate to each other.

Regarding the route set generation algorithm developed in this study one important factor is what kind of generalized cost that are used on the links. When calibrating the route set generation algorithm against observed routes from the GPS observations, the results mainly shows that if we generate more routes, more of the observed routes are covered by the route set
generation algorithm. Obtaining a higher percentage of matched links when generating more routes is reasonable, but it may not be preferable to generate many routes depending on application. If this was not a problem, one could simply just generate a route set containing all possible routes for each OD-pair. However, it would be preferable to generate as few routes as possible, but also cover as many of the observed routes as possible. One explanation for why this was hard to achieve in this study could be that only the free flow travel time was used as cost on the links in the network when calibrating the algorithm. This could indicate that there are many other factors affecting which route the traveler will choose. Factors affecting this could be for example number of lanes on the link, number of intersections on the route, number of turns, historical experiences of the individual traveler etcetera. With a more detailed definition of the cost on each link the results may have been different. Another factor that affects the performance of the route set generation algorithm is how many of the routes from the GPS observations that are similar. If most of the identified routes from the GPS observations go on the freeway, and the route set generation algorithm miss routes outside of the freeway, this could affect the results less than if the algorithm missed the routes on the freeway. The way users tend to travel between zones in the networks also affect the results regarding the percentage of matched links. In this thesis trips have been filtered out between two TAZ, without considering that people often have multiple errands between the origin zone and destination zone. This of course, differs a lot from a route set generation algorithm that is based on shortest path calculations. It is also worth mentioning that even if the results regarding the percentage of matched links in the routes from the route set generation algorithm and the GPS observations was rather low, the purpose with that investigation was not necessarily to develop a route set generation algorithm that perfectly matches the routes from the GPS observations. That investigation was performed rather to be able to quantitative set the parameters, and not just pick the parameters randomly.

The desired number of routes generated for each OD-pair also depends on the area of use of the route set. If using a static route set for many different scenarios such as regular days and different type of incidents, it is important that the route set contains a large variation of routes for each OD-pair or TAZ-pair, and if a more specially designed route set is used for a particular type of incident, that route set does not need to cover as many routes. This is important to have in mind when calibrating the route set generation algorithm and evaluating the number of routes for each OD-pair.
It is also important to consider how to interpret the results from the route set generation algorithm. When using travel times from the GPS observations during the incident day as cost on the links the route set indicates which routes the traveler should have used in the network, rather than which routes that were actually used during that day. This is a direct result from using a shortest path algorithm when generating the route sets. That is, the route set with real travel times as costs on the links shows which routes that had the lowest travel time with that input data, even if the areas with low occurrences in the routes set in fact were used a lot during the incident day.
9 Conclusions and future work

9.1 Future work

To further investigate the possible areas of use of the GPS observations, one factor that have large effect on the reliability of the results are the amount and quality of the input data. The methods developed in this thesis could be used for handling a much larger amount of input data. To obtain more routes and more statistical significance in the output one could use more than 23 days representing the set of regular days. It would be possible to for example load data for a whole year. The usage of more recent data would also increase the quality of the GPS observations due to better devices and more users. To obtain more data representing the incident days it would be possible to combine several incident days with the same characteristics and location of incident. If for example two incidents took place on the eastbound I-210 freeway on approximately the same location those two incidents could be combined and give more routes and more reliability in the results.

Due to lack of time in this thesis no comparisons were done with the Aimsun model. The results from a large set of GPS observations could be used as an extra calibration tool for the simulation model. One possible area of future work could be to investigate differences in routing behaviors, travel times, congested areas etcetera between the GPS observations and the Aimsun model. This could be helpful when evaluating the performance and creditability of the simulation model.

Regarding the route set generation algorithm it would be interesting to define more detailed costs on the links. The idea with this would be to reflect the reality in a better way. Factors that could be considered in these costs could be number of lanes, number of intersections, width of the road, historical travel times etcetera. To assign flow to the routes from the route set generation algorithm using for example a logit model and the OD-demands used in the Aimsun model would also be interesting for further research. This extension of the study could make it possible to compare route flows between the Aimsun model and the route set generation algorithm combined with a logit model and historical PeMs data.
9.2 Conclusions

It is possible to use GPS observations for investigating route sets in a network. However, when using this kind of data for making conclusions about the route sets in a network, the amount and quality of data is very important. In this thesis, data from 23 days was used as input when describing a regular day, and more data than this would be preferable to get more significance in the results. This effect becomes even stronger when only evaluating routes during one particular incident day, with data from only that day as input. But even if the amount of data is rather small, significant differences can be seen between regular days and the incident day. Regarding differences between the route sets from GPS observations and the route set generation algorithm, it is clear that the costs used on the links when running Dijkstra’s algorithm dramatically affects the results when investigating an incident day. When using free flow travel times on the links and only blocking the links with the incident, the routes tends to only go around the incident over a very short distance. The routes from the GPS observations shows longer reroutes around the incident area due to congestions and slow traffic upstream and around the incident. Results from the route set generation algorithm more similar to this were achieved when using real travel times from the GPS observations as costs on the links.

The accuracy of the GPS observations is very important when mapping them to a network like the one used in this thesis. If observations are mapped to the incorrect link to a network that in many senses are restricted regarding possible travel patterns, it could lead to that reasonable paths cannot be found between the observations. If the accuracy of the GPS observations are poor, knowledge about the accuracy could be useful. If each GPS observation would have come with information about signal strength or similar, a lot of GPS observations could have been filtered out in advanced. When inferring paths between GPS observations, Dijkstra’s algorithm can be used. To handle some of the incorrectly mapped GPS observations in the paths, comparisons between consecutive observations can be done to investigate if the length of the Dijkstra’s sub path is reasonable with respect to the time difference between the observations. However, not all incorrectly mapped GPS observations can be handled with this method, and due to the high number of routes that cannot be controlled manually. To investigate how accurate the method for inferring paths really are, an automatized method would have to be implemented.

The results from the GPS observations also shows that for regular days, the route travel times were around 60.82 % of the free flow travel times, while for the incident day the route travel times where around 46.60 % of the free flow travel time. The median travel time on the arterials
Orange Grove Blvd, Colorado Blvd, California Blvd, Corson St, and Del Mar Blvd during the incident is higher than the upper quantile for the 23 regular days for all of the arterials.

When it comes to the percentage of matched links between the routes from the route set generation algorithm, and the routes from the GPS observations, the parameter values had some effect, but not significant. However, the parameter values strongly affected how many routes that were generated. The result did not show any significant improvement in percentage of matched links when the number of routes increased with several hundred percent. The $\mu$ value did affect the overlap between the routes and with a lower value of $\mu$, a lower overlap was obtained. The threshold parameters gave a higher overlap when increasing.

The overall conclusion is that the implemented methods in this thesis can be used to gain a better understanding about routing behavior and travel patterns. However, to use the results for decision making, more input data with better precision should be used. On the other hand, the results from this thesis can be used as support when making decisions or when analyzing results from other methods for analyzing traffic.
Bibliography


Google Earth Pro (2018). (April 24, 2014). Pasadena, California, USA. 34°09'08.4"N 118°06'37.9"W, Eye alt 894 m. Google Earth Pro (software). [2018-04-14]


