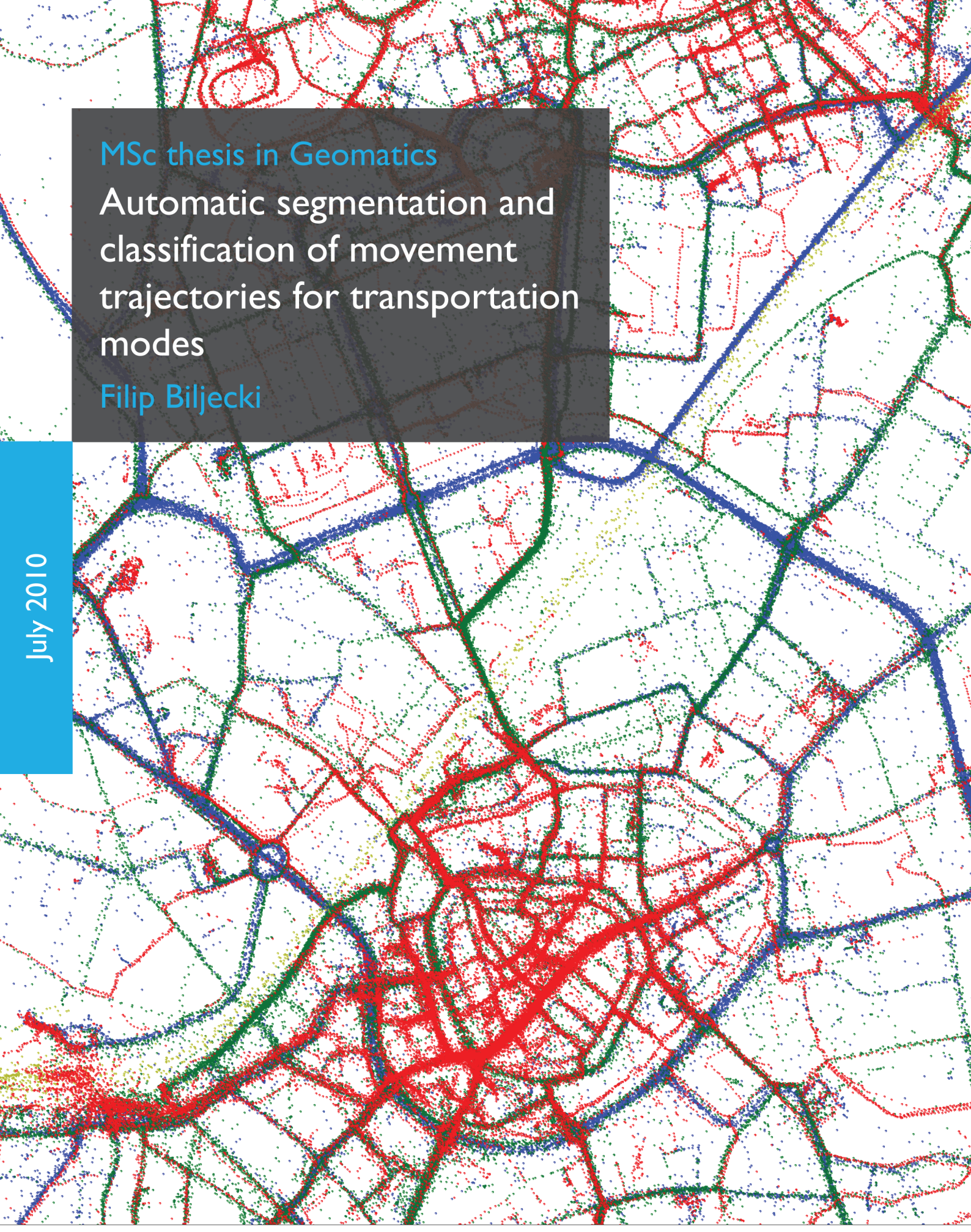


MSc thesis in Geomatics

Automatic segmentation and classification of movement trajectories for transportation modes

Filip Biljecki

July 2010

TU Delft Delft University of Technology

Department of GIS Technology, OTB Research Institute for the Built Environment

AUTOMATIC SEGMENTATION AND CLASSIFICATION OF MOVEMENT TRAJECTORIES FOR TRANSPORTATION MODES

A thesis
submitted to the Delft University of Technology in partial fulfilment of the requirements
for the degree

Master of Science in Geomatics

by

Filip Biljecki

July 2010

Supervisors: Prof. dr. Peter van Oosterom
Dr. Hugo Ledoux
Co-reader: Dr. Bastiaan van Loenen

Biljecki, Filip: *Automatic segmentation and classification of movement trajectories for transportation modes*, MSc thesis, Delft University of Technology, © July 2010

LOCATION:

Department of GIS Technology
OTB Research Institute for the Built Environment
Delft University of Technology
The Netherlands

TIMEFRAME:

26 October 2009 – 25 June 2010

FINAL PRESENTATION:

2 July 2010 at 12:30

EXAMINATION COMMITTEE:

Prof. dr. Dick Simons
Prof. dr. Peter van Oosterom
Dr. Hugo Ledoux
Dr. Bastiaan van Loenen
Dr. Kees Maat

COVER ILLUSTRATION:

The cover shows a classified subset of the dataset used in this thesis, covering the city of Amersfoort, the Netherlands. The image contains 1.1 million points sampled with GPS. Each point was classified for the transportation mode—the blue colour represents cars, while green, red and yellow are for bicycles, walking, and trains, respectively.

The data was acquired in 2007 for a travel behaviour study by the Department of Urban and Regional Development, OTB Research Institute for the Built Environment, Delft University of Technology.

ABSTRACT

The knowledge of the transportation mode used in a movement trajectory (derived in form of timestamped positions) is critical for applications such as travel behaviour studies. This thesis presents a method for segmenting movement data into single-mode segments and their classification with respect to the used transportation mode.

The method relies on concepts found in expert systems, most notably membership functions, fuzzy logic, and certainty factors. A prototype, which may serve as a framework for managing travel behaviour surveys has been built in order to validate the presented theories and to classify the available test dataset. The transportation modes that this system classifies are walking, bicycle, tram, car, bus, train, underground, sailing boat, ferry, and aircraft.

This research also investigates the performance of OpenStreetMap data in solving this problem. This free source of geodata proved to be crucial for the classification, where the ten transportation modes are discerned with various indicators mostly derived from the geodata, for instance, the proximity of the trajectory to the tram network and the information whether the movement has been made on a water surface or not. The classification relies on eliminating unlikely transportation modes by values set with a number of empirically derived fuzzy membership functions, and by using the selected combination of indicators it is possible to distinguish in between transportation modes which exhibit a similar behaviour (e. g. a car and bus in urban areas). Finally, the classification results are attached with a certainty value. The results are supplemented with additional mode-related information, e. g. the name of the departure train station.

The segmentation has been done by detecting potential transition points between two transportation modes as brief stops. After each segment between consecutive potential transition points is classified, adjacent segments with the same classification outcome are merged (and removing the transition point in between), and keeping only the actual transition points where the transportation modes had been changed.

The method solves the problem with noisy data, and traffic congestions which bias the indicators by using additional statistic values. The classification of gaps in the data (e. g. caused by a signal shortage during the logging of a trajectory) derived satisfying results, and segments with only their starting and ending point have been successfully classified. Moreover, thanks to the availability of the OpenStreetMap data, the prototype is not restricted to trajectories acquired in the Netherlands, but it is also able to segment and classify trajectories acquired abroad.

The accuracy of the classification with the developed prototype, determined with the comparison of the classified results with the reference data derived from manual classification, is 91.6 percent.

SAMENVATTING

Het bepalen van de vervoersmodaliteit, die gebruikt wordt tijdens een verplaatsing (afgeleid van posities met tijdstempels) is kritisch voor toepassingen, zoals reisgedragstudies. Deze scriptie presenteert een methode voor het segmenteren van bewegingsdata in enkelvoudige modaliteitssegmenten en hun classificatie ten aanzien van de gebruikte vervoersmodaliteit.

De methode steunt op in expertsystemen gevonden concepten, vooral lidmaatschapsfuncties, fuzzy logic en zekerheidsindicatoren. Er is een prototype, dat dienst kan doen als kader voor het beheer van data t.b.v. reisgedragstudies, gebouwd om de gepresenteerde theorieën te valideren en de beschikbare testdatasets te classificeren. De verschillende vervoersmodaliteiten die dit systeem onderkent zijn: lopen, fietsen, tram, auto, bus, trein, metro, zeilboot, veerboot en vliegtuig.

Verder zijn de mogelijkheden om dit probleem via OpenStreetMap gegevens op te lossen onderzocht. Gebleken is dat deze vrij toegankelijke bron van geodata cruciaal is voor de classificatie, waarbij de tien vervoersmodaliteiten worden onderscheiden met behulp van verschillende uit geodata afgeleide indicatoren, bijv. de nabijheid van het traject ten opzichte van het tramnetwerk of de informatie of de beweging heeft plaatsgevonden over het water of niet. De classificatie is gebaseerd op het elimineren van onwaarschijnlijke vervoersmodaliteiten door middel van waardesets met een aantal empirisch bepaalde 'fuzzy' lidmaatschapsfuncties. Door gebruik te maken van de geselecteerde combinatie van indicatoren is het mogelijk onderscheid te maken tussen vervoersmodaliteiten die een gelijksoortig gedrag vertonen (bijv. auto en bus in stedelijke gebieden). Tenslotte worden de classificatieresultaten voorzien van een zekerheidswaarde. Aan de resultaten worden aanvullende modaliteitsgerealteerde informatie, bijv. de naam van het treinstation van vertrek, toegevoegd.

De segmentatie is gedaan door potentiële overstappunten tussen twee vervoersmodaliteiten als korte stopplaatsen op te sporen. Nadat elk segment tussen opeenvolgende transitiepunten is geclassificeerd worden naburige segmenten met dezelfde classificatieuitkomstensamengevoegd (en de transitiepunten verwijderd), zodat alleen de werkelijke transitiepunten waar de vervoersmoduliteiten zijn gewijzigd overblijven.

Via deze methode worden ruis in de data en verstoringen door files, die de indicatoren beïnvloeden, opgelost door gebruik te maken van aanvullende statistische waarden. De classificatie van gaten in de gegevens (bijv. veroorzaakt door een signaaltekort tijdens het registreren van het traject) heeft tot bevredigende resultaten geleid. Segmenten met alleen hun begin- en eindpunt zijn succesvol geclassificeerd. Bovendien beperkt het prototype zich—dankzij de beschikbaarheid van de OpenStreetMap data – niet tot in Nederland beschikbare trajecten, maar is het ook in staat om beschikbare buitenlandse trajecten te segmenteren en te classificeren.

De nauwkeurigheid van de classificatie van het ontwikkelde prototype op basis van het vergelijk van de geclassificeerde resulaten met de referentiedata, die afgeleid zijn van de handmatige classificatie, bedraagt 91.6 procent.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to the people who directly or indirectly contributed to this thesis. First and foremost, I thank my supervisors Peter van Oosterom and Hugo Ledoux for their guidance, inspiring thoughts, and valuable suggestions. The thanks are extended to Wilko Quak who had greatly helped me in the starting phase of the project by giving advices and directions that saved me a lot of time.

Kees Maat and Wendy Bohte from the Department of Urban and Regional Development contributed to this project by sharing the indispensable data and with constructive remarks. Stefan van der Spek from the Faculty of Architecture shared the data from his project as well.

Although not actively involved in my thesis, other people from the Department of GIS Technology have their share of help as well: Elfriede Fendel, Tjeu Lemmens, Martijn Meijers, Theo Tijssen, Edward Verbree, and Sisi Zlatanova. The friendly but professional environment in the department was very motivating and enjoyable to work in.

Ben Gorte gave interesting suggestions for the development of the solution, while Bastiaan van Loenen helped with assessing the privacy concerns described in the final chapter.

The Mobility, Data Mining and Privacy (MODAP) project, Yücel Saygin, and Monica Wachowicz supported my trip to the 13th AGILE International Conference on Geographic Information Science (Guimarães, Portugal—May 2010) to hold the presentation about the preliminary results of this project. Attendants of the presentation had given valuable suggestions.

I wish to thank also the participants from the Workshop on modelling moving objects held at the University of Amsterdam in April 2010 for their constructive comments about my progress.

I thank my office mate Ken Arroyo Ohori, colleagues from Geomatics, and friends who gave many helpful suggestions on the work, but also made my stay in Delft pleasant and fun.

Finally, I am grateful to my parents Jasminka and Zvonko, the rest of my family, and girlfriend Marija for their continuous support during these two years.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

CF	Certainty Factor
CLIPS	C Language Integrated Production System
DBMS	Database Management System
DOP	Dilution of Precision
ESA	European Space Agency
EU	European Union
FES	Fuzzy Expert System
HDOP	Horizontal Dilution of Precision
IDW	Inverse Distance Weighting
GIS	Geo-information Systems
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GPX	GPS Exchange Format
ISO	International Organisation for Standardisation
KML	Keyhole Markup Language
MF	Membership Function
OGC	Open Geospatial Consortium
OSM	OpenStreetMap
PCA	Principal Components Analysis
PDOP	Position Dilution of Precision
SHA-1	Secure Hash Algorithm 1

SQL	Structured Query Language
SVM	Support Vector Machine
TTFF	Time To First Fix
UML	Unified Modelling Language
VDOP	Vertical Dilution of Precision
WGS84	World Geodetic System 1984
XML	Extensible Markup Language

1 INTRODUCTION

1.1 MOTIVATION

Humans travel by using transportation modes, for example: walking, bicycle, car, and train (Rodrigue et al., 2009).

The knowledge of the used transportation mode by humans is critical for applications such as travel behaviour research (Bohte & Maat, 2009), where researchers aim at understanding human travel behaviour in order to predict travel patterns, and evaluate transport-related measures and policies. Travel behaviour is the way how people travel in space, where they go, how often, which transportation mode do they use, whether they chain trips, which route they choose, and so on. Researchers try to understand the impact of the built environment, the quality of public transport and the cost of various transportation modes. Typical research questions are:

- Do the types of trip origins and destinations differ between short and long trips or trips made by different modes of transportation? (Lester et al., 2008)
- Do people choose the location of their residence based on the preference of their transportation mode? For instance, do people live near a railway station because they prefer traveling by train, or do car-based households prefer to live in the suburbs? (Cao et al., 2009; Maat, 2009)

Another important knowledge in travel behaviour research is the one of the travel purpose (e. g. commuting, shopping, and leisure), which is often analysed along with the knowledge of the used transportation mode.

This knowledge can also be used for transport planning and traffic management (Asakura et al., 2000), for instance, the measurement of traffic flows (Ranjitkar et al., 2002), where population travel patterns are modelled and analysed. The information about the used transportation mode may have other applications, for instance, urban planning, location-based services, and navigational applications where the user interface of a navigational device and the route suggestion can be changed depending on the transportation mode currently used, and the the device might show points of interests which are only accessible by a specific transportation mode.

The data required by travel behaviour researchers is acquired in travel surveys, involving randomly sampled individuals. In the past, researchers collected the information of the used transportation mode and trip purpose through paper diaries filled by participants or telephone surveys, which often resulted in underreporting of short trips, and inaccurate and incomplete data (McGowen & McNally, 2007).

Recent advancements in positioning technologies has enabled inexpensive and straightforward acquisition of movement data with handheld positioning devices in a different form: timestamped positions, i. e.

$$(x_1, y_1, z_1, t_1)$$

$$(x_2, y_2, z_2, t_2)$$

$$\vdots$$

$$(x_n, y_n, z_n, t_n)$$

Several authors praise the advantage of these techniques over travel diaries since underreporting of trips is less likely, the data are immediately available in the digital form, and in general more data are available, for instance, travelled paths which can be integrated in a Geo-information Systems (GIS) environment for additional analyses, precise time data, trip distances, and trip duration (Bricka & Bhat, 2006; Draijer et al., 2000; Wolf, 2000). Further, most researchers conclude that these receivers completely replace, rather than supplement, traditional travel diaries (Wolf et al., 2001). Several travel surveys with positioning loggers have already been done worldwide (Axhausen et al., 2004; Bohte & Maat, 2008; Draijer et al., 2000), almost exclusively in Europe and North America.

However, the aforementioned data such as transportation mode and trip purpose can never be acquired directly with a positioning receiver, unlike with travel diaries or telephone surveys, since only timestamped positions of movement are available. Combining the two methods would be a high burden for participants of these surveys (Wolf et al., 2001), and since the datasets are usually vast, manual classification may not be possible. Hence, a method for the automated detection of transportation modes and trip purposes from the movement data has to be developed.

1.2 OBJECTIVES

This thesis concentrates on the determination of the transportation mode from a set of timestamped positions:

$$\begin{gathered}(x_1, y_1, z_1, t_1)\\(x_2, y_2, z_2, t_2)\\\vdots\\(x_n, y_n, z_n, t_n)\\\Downarrow\\\text{Transportation mode}\end{gathered}$$

where a set of important and most frequently used transportation modes is determined. Since a trajectory may contain multiple transportation modes, the problem is extended to the segmentation of the movement data:

$$\left.\begin{gathered}(x_1, y_1, z_1, t_1)\\\vdots\\(x_i, y_i, z_i, t_i)\end{gathered}\right\}\ \text{1st transportation mode}$$

$$\left.\begin{gathered}(x_{i+1}, y_{i+1}, z_{i+1}, t_{i+1})\\\vdots\\(x_j, y_j, z_j, t_j)\end{gathered}\right\}\ \text{2nd transportation mode}$$

$$\vdots$$

$$\left.\begin{gathered}(x_k, y_k, z_k, t_k)\\\vdots\\(x_u, y_u, z_u, t_u)\end{gathered}\right\}\ \text{n-th transportation mode}$$

Since the classification of the trip purpose is a problem of itself, it is not a part of this thesis. However, it is discussed further in the conclusion as future work (section 7.3.14 on page 96). Hence, the objective of this thesis is to develop a method that automatically segments and classifies movement data with respect to the used transportation mode. The objective and main research question that this thesis attempts to answer is:

How to segment and classify a movement trajectory for the used transportation mode(s)?

In order to answer the research question a prototype has been implemented. The developed prototype also serves as a framework for managing large-scale travel behaviour surveys—it facilitates storing, retrieving, analysing, classifying, and visualising the collected and classified movement data.

Obstacles and difficulties arisen in the research, and the nature of the approach require investigating several aspects with additional research subquestions:

- How to assign a certainty value to each classification result, and give multiple classification outcomes sorted and filtered by confidence?
- If a transportation mode was changed during logging (e. g. from a bicycle to a train), how to correctly detect the transition, segment the log and correctly classify both transportation modes?
- Can geo-information, such as roads and railway tracks, be supplemented to the movement data to considerably improve the classification? What is the most appropriate source of geodata?
- How to overcome the shortcomings of current technologies, for instance the inaccuracy of positioning loggers, noisy measurements, and their signal shortages, for deriving a reliable classification result?
- How to solve ambiguity in the classification process and discern similar transportation modes?

1.3 TERMINOLOGY

In this section, the definition of the important terms used in this thesis, with which not all readers may be acquainted, is presented. The definition of a *transportation mode* is given in the beginning of this chapter. The selection and list of transportation modes considered in this thesis is given in the section 3.1 on page 23.

Moving objects are all objects that may change their position through time, e. g. people. In this case their position can be often represented with a point, without losing valuable information. This thesis concentrates on the movement of people, which are often called *users*, or *respondents* in the context of travel surveys.

During their existence, moving objects experience *journeys*, each one occupying a time interval in the object's lifespan and taking the object between two relevant locations—bird migration, satellite orbit, daily commuting, and mail service. They can be perceived as countable traveling units, i. e. *trajectories*—"a record of the evolution of the position of an object that is moving in space during a given time interval in order to achieve a given goal." (Spaccapietra et al., 2008):

$$\text{trajectory} : [t_{\text{begin}}, t_{\text{end}}] \rightarrow \text{space}.$$

It is important to note that the time space function that describes the object's position is defined over the whole lifespan of the object, but a trajectory is given

by restricting the function to a specific time interval $[t_{\text{begin}}, t_{\text{end}}]$, included in the lifespan of the object. The movement of an object may be *segmented* into trajectories between two relevant locations depending on the application. For example, the movement of a truck of a delivery company can be segmented into daily movements, but also into movements between customers.

In this thesis, two varieties of segmentation are considered. First, the segmentation into separate *trips* or *journeys*—connections between two relevant locations related to an individual or household, e. g. movement from home to work and from work to shopping (Maat & Timmermans, 2006), and since trajectories can be completed with the use of different transportation modes, a second segmentation is established for obtaining single-mode trajectories, simply called *segments*, but sometimes also referred to as stages. In the segmentation of the trajectories for different transportation modes, the points where the segmentation occurs are defined as *transition points*.

A representation of a trajectory in 2D is given in Figure 1.

Figure 1: An example of movement in 2D.

The record of movement is synonymous with *track* and more applicable in the context of current acquisition technologies. The recording is nowadays generally done with sampling (observing) their position in certain intervals of time or distance, deriving *sampled points*—sequences of positions and timestamps (i. e. position in spacetime) in a specific time interval $[t_{\text{begin}}, t_{\text{end}}]$. An example of a trajectory with the emphasised sampled points, recorded from the previously shown trajectory is shown in Fig. 2.

Figure 2: Sampled points from the trajectory shown in Fig. 1

The predefined interval for sampling positions is the *sampling period*. Due to technological limitations, the sampling period cannot be infinitesimal, hence the trajectory cannot be completely acquired, and it is represented with linear interpolation between samples (Figure 3).

Figure 3: An example of the recorded trajectory in 2D. The trajectory is formed by forming straight lines between sampled points (linear interpolation), which are de-emphasised here.

It can be noted by comparing Figures 1 and 3 that joining the points results in the minimal path traveled between sampled points, not the actual, which may differ. Notice that the distance between points (Figure 2) is different although the sampling period is constant in time due to different velocities in the trajectory.

In order to formalise the presented concepts and related terms with their relations, a Unified Modelling Language (UML) class diagram, inspired by the work of Verbree et al. (2005), is given in Figure 4 on the facing page.

A sampled point is part of a single-mode segment. Each point has the basic data (timestamped positions), but with additional data it is possible to derive additional information, for instance, it is possible to calculate its speed from the distance and time difference to the subsequent point. Since a segment is a collection of such points in sequence made with the same transportation

mode, additional information may be computed, such as distance of the segment, duration, number of points, minimum speed, and so on. The first and last point of a segment are transition points, which separate the segment from adjacent segments completed with other transportation modes. The segment is a part of a journey, another collection of points, but related to a purpose of movement (between two relevant locations).

A movement archive contains all journeys of an individual in a recorded timeframe.

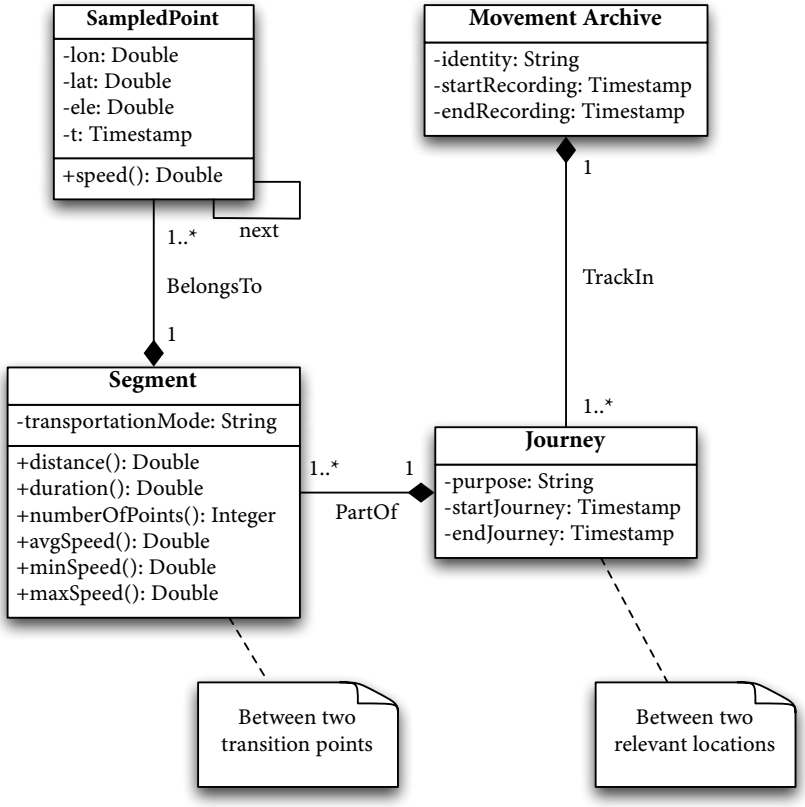


Figure 4: UML class diagram formalising the presented concepts relevant to segmentation and classification of movement trajectories. Adapted from (Verbree et al., 2005).

1.4 ACQUISITION OF MOVEMENT TRAJECTORIES

1.4.1 *Global Positioning System*

Positioning data can be acquired outdoors with several technologies, of which the most commonly used is the Global Positioning System (GPS). Other notable and near-future acquisition methods are trilateration of mobile devices, accelerometer traces, Galileo from the European Union (EU) and European Space Agency (ESA), and the Russian GLONASS. The trajectory classification method presented in this thesis should be operational for any relatively accurate method in future, however, in this thesis the used data are acquired with GPS, and GPS is emphasised.

GPS is a United States Global Navigation Satellite System (GNSS), comprising 24 to 32 satellites orbiting at approximately 20 200 km. Its three main products are:

- Position, in World Geodetic System 1984 (WGS84). Its accuracy considerably depends on several factors, especially the type of the receiver. Today's budget GPS receivers, which are used in travel surveys, are able to derive the position in the accuracy of less than 10 m in 95 % (2σ) of the measurements.

  It is important to mention that less sensitive receivers may derive noisy positions. In order to give an example of inaccurate and noisy movement data for the early and easier understanding of the problem, Figure 5 gives an obvious example of such data.

Figure 5: An example of noisy and inaccurate movement data layered over a satellite image for better presentation. It is obvious that the person could not have travelled in the way the recorded trajectory shows. The imagery is copyrighted by Aerodata International Surveys and Google (2010).

- Time. GPS has become the world's principal supplier of accurate time.
- Velocity, with accuracy of 0.1 m/s. Receivers calculate the velocity by measuring the frequency shift (Doppler effect) of the GPS signals (Misra & Enge, 2006).

1.4.2 *Sampling period*

Most of the receivers compute their position and velocity every one second, however not every of these observations is logged. The sampling period τ is the programmed time between two observed and logged positions ($\tau = t_i - t_{i-1}$).

However, if $t_i - t_{i-1} > \tau$, the sampling was interrupted, most likely by signal shortage, causing missing time intervals and disruption in the data.

It is important to have as frequent as possible sampling period, while logging a trajectory (Byon et al., 2007). The major difficulty of longer sampling periods is that the actual traveled path between the two sampled points may not be a straight line. This problem is visible when analysing trajectories in roundabouts, interchanges and longer turns. Another problem is the variability of the behaviour between two points, e. g. the speed.

Figure 6 shows two segments derived from the same movement, but with different sampling periods. The first segment, depicted in black, had a sampling

period of τ, while the second segment shown in red is sampled with the period 5τ. The differences between the representation of the two segments is obvious, and arising problems are evident, for instance, longer sampling periods will always result in shorter segments.

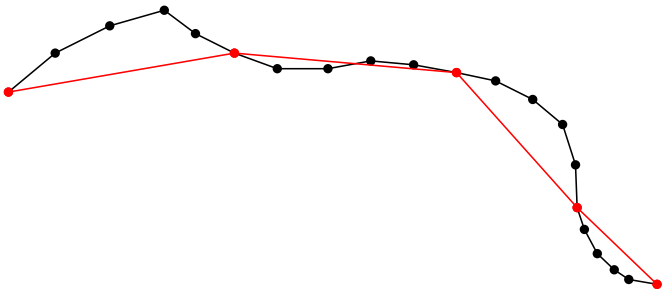


Figure 6: Two segments derived with different sampling periods (τ and 5τ).

1.4.3 *Exchanging the GPS data*

The GPS data are usually delivered in the GPS Exchange Format (GPX) format, a Extensible Markup Language (XML) schema and the *de facto* standard for lightweight interchange of GPS data. It consists of a collection of coordinate tuples with corresponding timestamps (observations), and various additional data for each, e. g. speed and number of satellites. The horizontal position is expressed with latitude and longitude in WGS84, and the elevation in metres in the same system. Time is formatted according to the International Organisation for Standardisation (ISO) 8601 standard, and it is usually rounded up to the nearest second.

As an example, a fragment of a GPX file representing one sampled point is presented below

```
<trkpt lat="52.196537" lon="5.413356">
  <ele>51.475254</ele>
  <time>2007-03-11T12:50:47Z</time>
  <course>220.490177</course>
  <speed>10.932674</speed>
  <fix>3d</fix>
  <sat>4</sat>
  <hdop>22</hdop>
  <vdop>20</vdop>
  <pdop>29</pdop>
  <quality>1</quality>
</trkpt>
```

It is visible that the GPS receiver in question is able to store additional data such as the speed, heading, Dilution of Precision (DOP) values, and number of satellites, which can be used for assessing the quality of the data. Although various receivers are able to derive these additional information in their observations, for saving memory they are not storing them in the output.

It should be noted that the number of significant digits of each measurement is not standardised and does not depend on the precision of the measurements. The following excerpt shows the GPX output for a point of a Garmin GPSMAP© 60C receiver, which is not significantly more accurate than the device used to generate the previous example:

```
<trkpt lat="51.996836336329579" lon="4.355507791042328">
   <ele>33.3028564453125</ele>
   <time>2009-12-10T18:10:49Z</time>
</trkpt>
```

In this project the developed prototype supports all GPX outputs, as long as they provide the position with the corresponding timestamps, which is the minimal requirement. Therefore, the prototype should be able to segment and classify movement data with no additional information other than timestamped positions in form of (x, y, z, t), i. e. the structure of latter GPX example.

As an example, the Figure 7 shows a trajectory layered over a satellite image, with the information for one sampled point ($\tau = 2\,\mathrm{s}$).

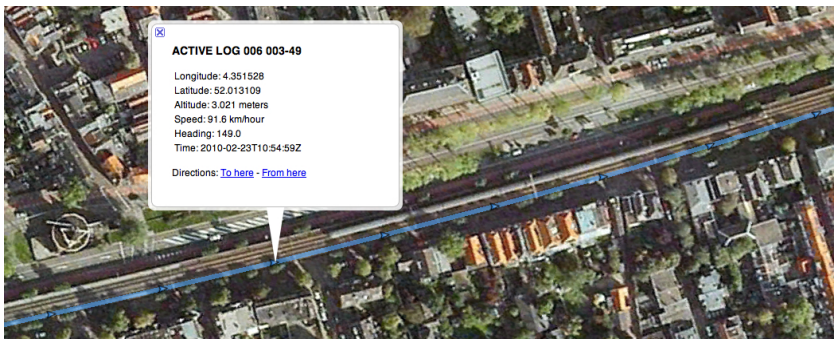


Figure 7: An example of a GPS trajectory with the information for one sampled point, layered over a satellite image. The imagery is copyrighted by Aerodata International Surveys and Google (2010).

1.4.4 *Test data*

In this thesis, in order to test the developed method and prototype, various movement datasets were used. Apart from the possibility of testing the algorithm, there are two other benefits from the data: validation of the final results of the algorithm since reference data of the used transportation mode was made available, and derivation of various travel characteristics needed in order to investigate the relation to a particular transportation mode, for instance, the speed behaviour for a particular transportation mode.

There are four available main sources of data:

- The data from the survey conducted in the Netherlands by Bohte & Maat (2009) (Department of Urban and Regional Development, OTB Research Institute for the Built Environment, Delft University of Technology) as part of a travel behaviour study focused on residential choice was be used. It had been acquired over 15 weeks in 2007 with an average sampling period of 6.5 s, and the raw data (a collection of GPX files of four gigabytes) is available for the purpose of this project. A GPS logger from the Dutch company Amaryllo with a SiRFstar II chip was used. This chip is prone to noise and signal shortages, but was chosen because of low power consumption.

  The data has been classified in an interpretation-validation process in which the system first made a preliminary segmentation and classification (described in § 2.1.4 on page 14 in more details), after which the data have been validated by the respondent in a web-based questionnaire (Figure 8 on the facing page). The researchers are interested in removing or shorten the validation process by improving the segmentation and classification process, which is one of the motives for the initiation of this project. The data validated by the respondents may be used for experiments and validation of the method developed in this thesis. However, the classified

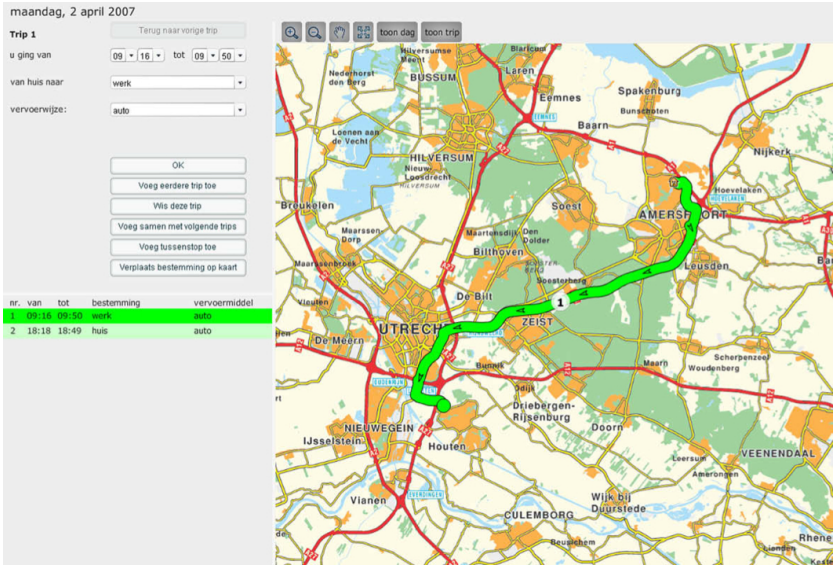


Figure 8: The web-interface for validating the classification results. Courtesy of Bohte & Maat (2009).

data has a few problems, most notably—it is not segmented accurately (the times of the transitions are not accurate), and several short segments are left-out. Hence, manual classification of such data as a supplement had to be done for validation purposes.

In addition to the mentioned benefits by using these data, the algorithm can be tested for robustness in handling vast datasets produced in travel behaviour surveys.

These data accounts for more than 90% of the test data in this project.

- Own data acquired during this research, with a Garmin GPSMAP© 60C receiver, with variable sampling periods, in the Netherlands and abroad. This was done in order to simulate several specific situations that may arise and cause difficulties in the segmentation and classification process.

- Raw files from the OpenStreetMap (OSM) project, with variable sampling periods. Their benefit is that they include more trajectories from abroad, and transportation modes which may be seldom included in the first two datasets. The files are shared under the Creative Commons Attribution-Share Alike licence, hence these data can be freely used in this project. Each uploaded file in OSM is tagged with keywords, hence, it is straightforward to search for trajectories made by specific transportation modes.

- Data from the project of Van der Spek et al. (2009) from the Department of Urbanism, Faculty of Architecture, Delft University of Technology which concentrates on collecting data on pedestrian movement in city centres. The project addresses the topic of improving city centres for pedestrians, especially for shoppers and tourists (Van der Spek, 2010).

  The sampling period of the trajectories is 5 s. The movement data mostly covers the area of the city of Delft, and its benefit to this project is the large quantity of walking segments used for investigating behaviour of walking. The dataset also contain other transportation modes, mostly bicycles and

cars, and it is delivered segmented and classified (from a manual method supported by questionnaires of respondents).

To quantify the available test data, in Table 1 are presented various statistics of all the datasets together for an overview.

Table 1: Statistics of the raw GPS data available for this project.

Points	17 624 899
Files	9585
Individuals	1369
Distance	539 587 km
Duration	1253 days

As an impression of the dataset and its size, Figure 9 shows the city of Amersfoort "mapped" from the available movement data. Each point of the dataset in the presented spatial extent was plotted. Frequently used paths (e. g. highways) can be observed by the aggregation of multiple points (i. e. thicker lines).

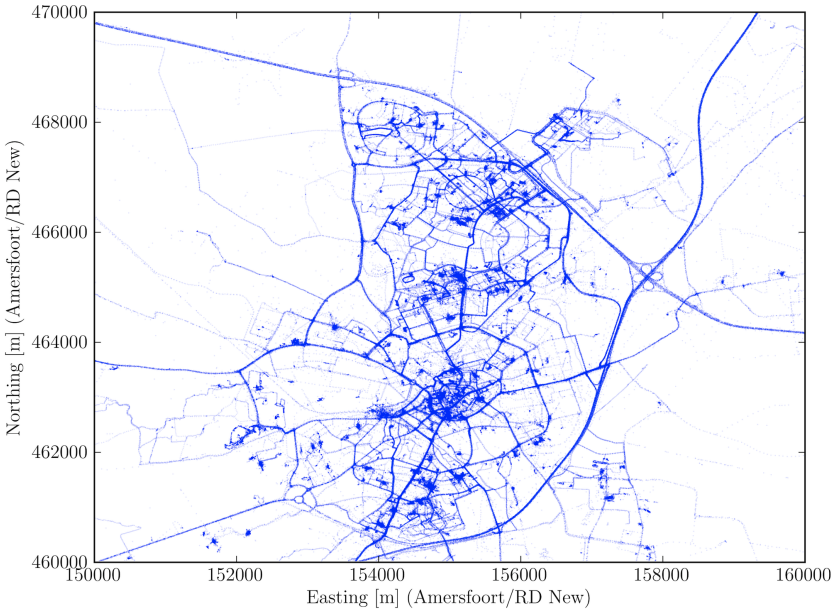


Figure 9: Visualisation of GPS data in Amersfoort. This spatial extent contains 1.2 million points.

1.5 SCOPE OF THIS THESIS

The main objective of this thesis is the segmentation and classification of movement trajectories. A prototype is built in order to test and validate the developed ideas.

This thesis does not deal with data acquisition and the theory of travel behaviour research, although a few trajectories have been acquired for tests, and basic information on the acquisition methods is given in this chapter. The classification of trajectories for trip purpose is a topic that can be aligned to this

project, and although it is a separate problem which would require additional research, the preprocessed and segmented data in the scope of this project could be used for such project as well. The applications of classified results are not part of this thesis, although they are listed in the beginning of this chapter, and a few recommendations for other possible applications are given in the final chapter (e. g. § 7.3.13 on page 95).

The simplified scope of this thesis is given in Figure 10.

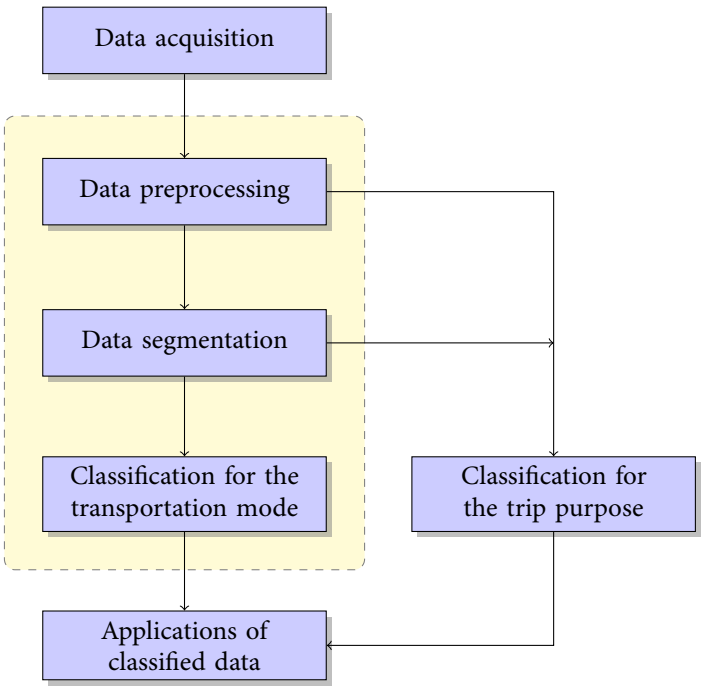


Figure 10: A diagram showing the scope of the thesis (in yellow) with the closely related topics.

1.6 OUTLINE OF THIS THESIS

Chapter 2 (on pages 13–20) presents the related work in the frame of segmentation and classification of movement trajectories for transportation modes. The available methods are summarised and their shortcomings are listed as a guideline for building a method which is designed for solving the difficulties encountered in segmentation and classification for transportation modes.

Chapter 3 (pp. 21–48) in general describes all aspects of the methodology, segmentation-classification solution and the used tools. The chapter lists and justifies the considered transportation modes. The selection of the source of the required geodata is given in this chapter as well. The chapter also contains practical information needed in order to support the presented theory.

The 4th chapter (pp. 49–56) describes the implementation of the prototype in the process of importing and preprocessing the data required for the segmentation and classification process. The import of the geodata (from the OpenStreetMap project) is described as well.

Chapter 5 (on pages 57–78) continues with the description of the prototype in the frame of segmentation and classification, and gives more technical details of the solution. A few aspects, such as resolving gaps in the data due to practical details is given in this chapter rather than in Chapter 3 about the meth-

odology. Additionally, the chapter shows that it is possible to derive additional transportation mode-oriented information.

The validation of the results and the performance of the classification system (experiments) are given in the 6th chapter (pp. 79–83).

The conclusion with recommendations for future work is stated in Chapter 7 (on pages 85–96). This chapter also discusses the possible contributions of this thesis.

Appendix A (on pages 97–107) discusses the possibilities of removing noise from movement data and interpolation of missing time intervals. This two approaches are not actively used in this thesis due to the reasons discussed in the corresponding sections. However, they are given in the end of the thesis as a guideline for future work.

2 RELATED WORK

There are previous attempts to solve the key problems of this thesis. The summary of each solution is given in the section 2.1. For an overview they are presented in the Table 2 on page 18. The conclusions of this overview are given in the section 2.2, along with the identification of the problems that have to be solved in this thesis.

2.1 SUMMARY OF CURRENT SOLUTIONS

In this section, eleven existing methods for segmentation and classification of movement trajectories for transportation modes are presented. The names of the methods are derived from their most important characteristics.

2.1.1 *Using neural networks*

The approach of Byon et al. (2009) uses neural networks-based and fuzzy logic-based methods to determine the transportation mode, and it is built upon previously published methods (Byon et al., 2006; Chung & Shalaby, 2005; Shalaby et al., 2006). As the input data the speed, acceleration and the number of satellites in view during the sampling are used. The authors claim that the available number of satellites helps to identify the transportation mode since different modes have different sizes of ceilings and windows, thus differently obscuring the view to the clear sky. For example, a GPS receiver placed in a car should have better reception through its large windshield, than a receiver in a tram.

In the process of supervised learning, the classifier of the neural network is trained with the validated data. This approach distinguishes between four transportation modes: car, bus, tram, and walking. Although the authors list the accuracies between 60 and 97 percents depending on the identified transportation mode, a single accuracy figure is not reported. The average of all reported figures is approximately 82 %. The accuracy is defined as the ratio of correctly classified classes and total number of trials.

I do not second the idea of using the visible number of satellites as an input parameter since the receiver in a car will certainly have a different average number of satellites in view in urban areas with high buildings comparing to a countryside. Moreover, a receiver positioned near a window in a tram might have a better reception than a receiver in a car in a urban area, hence I do not see much benefit using this parameter, in addition to the fact that not all devices log the number of visible satellites, and the fact that this approach restricts the method to the GPS.

2.1.2 *Fuzzy-logic and raw data*

Schüssler & Axhausen (2009) have a relatively minimalistic approach, using only positions and corresponding timestamps from the collected data and rejecting any other possibly available data from the logger.

For transportation mode detection, a fuzzy-logic approach is used in detecting five modes: walk, cycle, car, train, and urban public transport, that is, bus and tram in a joint category. The membership functions deriving three descriptive

values—low, medium and high are established from the input variables: the median of the speed, the maximum speed and acceleration distributions.

After the membership functions assign the values for each variable, the fuzzy system with 16 various rules is applied for mode detection. One example of a rule is: if the median speed is low, *and* if the maximum acceleration is low, *and* if the maximum speed is medium then the urban public transport had been used.

The authors do not report the accuracy of their method.

2.1.3 *Decision-tree based inference model*

The approach described in (Zheng et al., 2010, 2008a,b) detects four transportation modes: car, walking, bus and cycling. It consists of change point-based segmentation method for partitioning the data into single-mode segments and a decision-tree-based inference model, with which an accuracy of 75 % is achieved.

In the segmentation, the authors stress that a person usually walks while changing a transportation mode, hence the detection of walking segments leads to the detection of points where the transportation mode was changed.

In the inference model, five variables are used—speed, acceleration, heading change rate, stop rate, and velocity change rate.

The method takes advantage of the fact that people driving a car or taking a bus are not flexible with changing the direction as users which are walking or cycling, hence they calculate the heading change rate for each point. Another indicator of a transportation mode is that people taking a bus or walking, are likely to stop more times than while driving, hence a stop rate, that is, stop frequency within a unit distance is calculated and used in the transportation mode identification. The last hint is the ratio of the speed difference between a point and its adjacent point and the speed at that point (velocity change rate), a dimensionless value conceptually similar to the acceleration:

$$\frac{v_{i+1} - v_i}{v_i} \tag{2.1}$$

The authors state that the last two clearly capture the difference among various transportation modes. In the end, the segments are classified and the results are sorted by probability.

The method is patented at the United States Patent and Trademark Office (Zheng et al., 2009).

2.1.4 *Deterministic approach*

The deterministic method of the previous attempt for classification of the test dataset available for this thesis is described in (Bohte & Maat, 2008, 2009; Bohte et al., 2008). The researchers use a decision tree considering the mean speed, and the maximum speed. For instance, if the average speed is between 10 and 25 km/h and the maximum speed is between 14 and 45 km/h then it is estimated that the person cycled. Since train and car have similar speed behaviour, GIS network data for the railways was used in addition—if at least one third of the points lie in the 50 m buffer of the railways, then the trajectory is estimated to be made with a train. The trajectories have been segmented by analysing the distance to stations.

The method uses the location of several points of interests to determine the trip purpose, e. g. shopping or work. The accuracy of their classification system is 70 %.

2.1.5 *Deterministic approach 2*

In an internal report, De Boer (2008) describes his work, which is similar to the method described in §2.1.4, however it uses additional variables: acceleration, journey distance, duration, and knowledge of transportation mode availability for each user. For example, if the user does not possess a car (or does not have access to it), then the car is removed from the system as a possibility. Although the work is not finished, it provides empirical information about average speeds and journey distances that can be expected for each considered transportation mode.

2.1.6 *Using Support Vector Machines*

In a more abstract approach, Dodge et al. (2009) are aiming on the general classification of movement trajectories based on a variety of different characteristics (e. g. simulation and extraction of movement patterns), loosely concentrating on the transportation mode detection, using it only as an example of their general theory. The paper also emphasises eye-movement data and biology and ecology as application domains, not only GPS data described in other methods so far.

They start with preprocessing the data, most notably resampling the data to a regular time interval using linear interpolation of fixed time intervals. This implies that the authors concentrate on tracks, rather than sampled points, and the method is limited to a fixed sampling period.

The principle of the method is to compute the different behaviour of statistical descriptors, in our case acceleration and speed, over time, in contrast to the other methods where absolute descriptors are computed. First, the global descriptive statistical parameters (e. g. mean acceleration over the whole data) are computed, and afterwards profile decomposition into segments with similar movement character is done. This is realised with the deviation from the median line of different movement parameters over time, and sinuosity, in order to extract local information at finer resolutions.

As the method concentrates on the behaviour of statistical descriptors (variability), relative (normed) values of the descriptors in the range [0,1] are computed rather than in units such as km/h. For instance, a speed of 34 km/h might have the same value as 60 km/h in a different dataset.

The classification is done using Support Vector Machine (SVM), a set of related supervised learning methods (Hsu et al., 2003), while Principal Components Analysis (PCA) is used for reducing the number of original features (Pearson, 1901). However, no detailed information are given about this process.

One notable descriptor, not used by other methods is the sinuosity of the trajectory which authors claim exhibits an association with a certain transportation mode. Authors do not seem to use any GIS network data.

In the end, the Spearman rank correlation coefficients for different descriptors are computed, such as the correlation between speed and acceleration, or acceleration and distances between sampled points. This theory seems not applicable in data with variable sampling periods, unless it is resampled with linear interpolation, which may not be favourable in all cases.

In an example, four transportation modes (motorcycle, car, bicycle, and pedestrian) are used. The authors claim that the accuracy of their method for classification of movement trajectories is 82 %.

2.1.7 *Hidden Markov model*

Reddy et al. (2010) built a prototype classification system consisting of a decision tree followed by a first-order hidden Markov Model, which achieves accuracy level of 93.6 %. A hidden Markov Model is the simplest dynamic Bayesian network, more described in (Rabiner, 1989).

The system uses the acceleration variance, speed, and the likelihood of transitions between particular transportation modes. However, their method distinguishes four transportation modes—walking, cycling, running and motorised transport, of which the latter accounts for all motor vehicles, i. e. motorcycle, car, and bus (Reddy et al., 2008). This method seems to neglect the problematic part or distinguishing between a car, bus and train, although it is interesting that it is able to distinguish between cycling and running which have similar speed distribution, thanks to the acceleration variability between the two modes. Their approach concentrates on mobile phones, which nowadays in addition to the GPS receiver often include an accelerometer, that is significantly more accurate and sensitive in deriving accelerations than built-in GPS receivers, which calculate the acceleration from the speeds of two points in sequence. If only a GPS receiver is used, and the accelerations are derived from it, the accuracy of this method drops to 74.4 %. This paper implies that using the accelerations from GPS receivers may not be valuable, which is also a finding of this research (see § 3.5.2 on page 45).

2.1.8 *Hierarchical Markov model*

Liao et al. (2007) introduce at hierarchical Markov model that can learn and infer a user's transportation routines from GPS data for building personal maps. Beside the transportation mode, the project has a wide scope of derived information from a journey (e. g. location of transportation destinations), prediction of future movements, and detecting user's errors, such as boarding a wrong bus. The process of inferring the transportation mode is not detailed to make any further conclusion, however, they provide us with an useful advice for segmenting multimodal trajectories—by analysing the proximity to potential-transition locations such as bus stops and parking lots (Liao et al., 2006). Three modes are classified (bus, walking, and car), and in addition to the locations of potential transitions, also the speed is used for classification.

2.1.9 *Neural networks 2*

Gonzalez et al. (2008) investigated the usage of neural networks for identification of the transportation mode only from critical points, i. e. a minimum set of GPS fixes required to accurately reconstruct the user's path. The critical points are derived from stops, sudden turns and changes in acceleration. The following input variables are used: the average and maximum acceleration, the average and maximum speed, the ratios of the number of critical points over the total distance and the total time of the movement, the total distance of the movement and the average distance between critical points. The authors additionally advise using the standard deviations of distances between stops locations, and the dwell times when classifying using all recorded points, not just critical points (Gonzalez et al., 2010).

The method distinguishes between car, walking and bus, with the reported accuracy of 91 %.

2.1.10 *Concentrating on accelerations*

The method from Lester et al. (2008) classifies trajectories into four transportation modes: bus, car, walking, and cycling, using speeds, accelerations from an accelerometer device (in addition to GPS readings), and the location of bus stops, in order to differentiate between a bus and a car. However, since transportation mode determination is not the focus of the paper, no additional information are available.

2.1.11 *User's knowledge*

The method of Stopher et al. (2007) focuses on seven modes—walking, rail, ferry, underground, cycling, bus and car, which is initialised by analysing supplementary information such as mode ownership and driving licence possession, and thus removing non available modes from possible solutions. Subsequently, the following input is used for assigning weights for each probable mode: maximum accelerations and speeds, GIS network data for detecting public transportation and distinguishing between a car and bus. Underground rail trajectories are identified separately, in the repair process, when gaps are found in the record that correspond to a rail trip with no GPS data available during the rail segment.

Their method is successful in 95 % of the tested cases (Stopher et al., 2008). The accuracy of the method when not using supplementary information is not reported.

2.1.12 *Overview of the present solutions*

The tabular overview of the presented solutions is shown in the Table 2 on the next page. The methods are referenced with the number of the section they are described in.

The majority of the methods classify between three or four transportation modes on average, with approximately four criteria. A third of the authors do not report the accuracy of their solutions, while the average accuracy of the reported results is higher than 80 percents. The majority of the methods do not use GIS data in the classification. There is no apparent relation between the number of transportation modes or criteria with the reported accuracy.

2.2 OBSERVATIONS

2.2.1 *Fundamentals*

This literature review helped to gain the essentials for developing a new method. The following valuable knowledge is used from the reviewed methods:

- All methods use the speed as the primary variable for mode detection, implying that the speed gives the highest indication of a transportation mode. Because of comparable speeds, additional knowledge is essential in order to distinguish a car, a bus, and a train, and other modes with similar distribution of speeds.
- Nearly maximum values should be used rather than maximum values of speeds and acceleration in order to make the method robust for noisy measurements (Schüssler & Axhausen, 2009; Stopher et al., 2007). Nearly maximum values are usually calculated with 95th percentiles.

Table 2: Comparison of the reviewed methods for transportation mode identification. The dash represents unknown information.

Method	Modes	Criteria	GIS data usage	Accuracy (%)
§2.1.1	4	3	✗	82
§2.1.2	5	3	✗	—
§2.1.3	4	5	✗	75
§2.1.4	4	2	✓	70
§2.1.5	7	6	✓	—
§2.1.6	4	3	✗	82
§2.1.7	4	3	✗	74
§2.1.8	3	2	✓	—
§2.1.9	3	8	✗	91
§2.1.10	4	3	✓	—
§2.1.11	7	4	✓	95
Average	4.5	3.8	5 of 11	81.3

- GIS data may be used not only for detecting line infrastructure features (e. g. roads and railways), but also for determining potential mode transition points such as railway stations (Liao et al., 2006).
- Underground modes can be detected by finding signal shortages with last known points around the stations (Shalaby et al., 2006; Stopher et al., 2008).
- A decision tree based method is the most straight-forward method for solving this problem, but it delivers single results without a value of certainty, and does not deal with ambiguity when two modes have similar behaviour.

2.2.2 *Disadvantages, missing information and problem identification*

However, I have observed several shortcomings and drawbacks in the presented methods. Some publications have missing key information. For instance, a few methods use approaches such as SVM, PCA, Neural networks and Hidden Markov models, which are not elaborated further, hence no conclusion can be derived, and the transfer of the information from the authors is limited. The reviewed approaches seem to be more theoretical, without a stated implementation since the used tools are seldom listed and there are not many references to the developed experimental software, and the application domain of the results. Addressing the problem in a more technical way is usually missed. For instance, how does the reasoning using acceleration perform in trajectories with poor sampling periods?

The majority of the methods do not segment a trajectory into single-mode segments, which may result in a wrong classification. The rest have scarce explanation of the segmentation approach, most useful that a person usually walks

while changing a transportation mode, and in order to detect a transition, first it is required to lookup for walking segments.

Unfortunately, there is no standardised data or mechanism for testing the claimed accuracy of each method. The authors often do not describe the quality, size and structure of their data. At this point, there is a concern that most of the authors use only flawless data with a frequent sampling period which contain only a single transportation mode, and without outliers and data gaps. Further, there is not a standardised classification and absolute number of modes to be distinguished. Some authors use a single class for all motorised modes, while some use a single class for public transportation, making further comparison between the developed methods difficult.

Papers with the detection of the mode as a secondary objective seem to heavily underestimate the problem and exaggerate about their non-implemented conceptual achievements.

The following problems are inferred from the cited papers, but also from own analysis of the available test data:

1. All authors admit the difficulties with traffic jams, congestions and heavy traffic which lower some of the statistical values used in the criteria (e. g. mean speed), resulting in inaccurate classification.

2. Problems with distinguishing modes with comparable statistical descriptors and behaviour such as speed and acceleration.

3. Most methods consider a limited number of modes, which may be trivially distinguishable in most circumstances due to their very different behaviour in movement. Methods which incorporate more modes usually do not report high accuracy.

4. Only a limited number of criteria is used in each method, although a few authors give recommendations for further rules in future work, these criteria are already used in some other works. A method having implemented all the presented criteria does not exist yet for unknown reasons.

5. Although several concepts are presented, and a few implementations, still there is not a complete system constructed for serving large-scale projects (e. g. travel behaviour studies) that result in huge datasets.

6. Although most of the methods use a fuzzy approach, the modes are in the end determined deterministically without a value of certainty attached to the classification outcome, giving no alternative to other possibly used modes.

7. It is uncertain if the presented methods cope with data gaps caused by signal shortages since there are no records about it. The change of mode during longer data gaps may impose problems with the detection of the transportation mode. Moreover, in case the mode was changed during the shortage it may be very difficult to detect the transition and extrapolate where and when it occurred.

   It is only known that Shalaby et al. (2006) take advantage of signal shortages by detecting underground modes, which is investigated as well, but also for other modes which may involve signal shortage during the journey.

8. Geodata seems to be seldom used in this field. When used, it is not used to full potential, rather for a narrow application such as detecting nearby parking lots, and no further details are given (e. g. source of the data and used operations).

9. The detection of short trajectories is not addressed, where a limited amount of points is available.

10. The methods do not appear to be robust, taking into account several constraints and special cases, and their performance is not reported. Usually, there is no information about the developed prototypes which could verify the presented concepts.

11. GPS inaccuracy and bad signal may give noisy data, and reparation (or rejection) of such data is not frequently discussed.

12. Sea and air transportation modes are generally left out from the classified modes. A complete classification solution should include these modes as well.

13. The presented methods have small application domains, they are usually concentrated on small, local, areas such as a single city or province, which imply local cases such as the number of bus stops and local GIS data, but also cultural and other difference, which tend to be different between countries. In the end they are tested on small datasets, and their worldwide applicability is doubtful.

My aim in this thesis was to build a method which is robust to most, if not all, of these problems. Each of the problems is addressed in the following chapters with the suggested solution. The complete list of all answers to these problems is given in the conclusion of this thesis (§ 7.1 on page 85).

3 METHODOLOGY

In this chapter, the methodology for solving the objectives and answering the research questions of this thesis is described. Based on it, in order to test the theory but also to classify the available test data, a prototype was implemented, described in the next two chapters in more details with technical information (Chapter 4 on page 49, and Chapter 5 on page 57).

The list of considered transportation modes and the reasoning between the selection is given in section 3.1 on page 23, while the list of the software used for the implementation of the prototype is given in section 3.4 on page 35.

I set up a few requirements and guidelines to be followed for designing the method for solving the key problem in this thesis, but also for building a functional prototype for facilitating movement behaviour surveys:

- The developed method should be *robust*, where many different situations that may arise (e. g. seldom transitions between particular modes) should be anticipated and taken into account.
- Problems such as missing time intervals, noisy measurements, variable sampling periods, broadly different travel behaviour for each mode, and segmentation of multi-journey and multimodal trajectories require the method to be *adaptive* and *"smart"*. Lower sampling periods should be especially handled since they may cause inacurate classification.
- It should be *universal* where any trajectory can be read, from any outdoor logging device in any location worldwide, but it also should be designed that may be used outside of the scope of travel behaviour research, i. e. for segmenting and classifying the trajectories for virtually any application that requires the knowledge of the used transportation mode.
- The method should be designed to be expanded if needed, i. e. as a stack of virtually independent operations, that can be removed for computational and other reasons, but also be added in future work. For instance, if the user has a specific preference in smoothing the trajectories, it should be possible to activate (or deactivate) the component in question. Moreover, it should be possible to add additional transportation modes in future, and modify or upgrade the classification system. Therefore, the method should be *extendable*.
- Travel surveys involve very large datasets. The method should be *efficient*, be able to handle the datasets, and anticipate various difficulties that may arise in these cases.

The workflow is divided into four parts: importing the data, preprocessing, segmenting, and classifying it for transportation modes.

The central part of the system is a database where the data are stored after each operation. Data may be stored in different ways, and here a Database Management System (DBMS) is used for:

- Better organisation of the data.
- Best performance with big datasets.

- Better integration in a GIS environment, with various spatial operations and spatial indexing methods.
- For later easier querying, and updating of the trajectories.

In the workflow, first raw GPS data (in GPX format) are imported in the database, after which are preprocessed for calculating the *indicators*—various statistical values (e. g. speed) that may indicate a particular transportation mode. The indicators are the input of the classification system for deriving the transportation mode, and are calculated from the (x, y, z, t) series of timestamped positions and geodata. They are listed in § 3.5 on page 36 with the reasoning behind their selection.

Since geodata is used in the preprocessing part, for instance, in operations such as calculating the proximity to the nearest railway and nearest bus stop, corresponding geodata of the covered area has to be imported. The import of the GIS data has to be executed only once, before processing any GPS file. The selection of the GIS data source is given in 3.4.2 on page 35.

Section § 3.2 on page 24 explains the methodology for the segmentation of the trajectories, and gives more practical details in order to facilitate the explanation. After the segmentation of trajectories is completed, the segmented data in form of single-mode trajectories in single-journeys are passed to the classification system (§ 3.3 on page 29). The classification is empirically manually trained with training data (supervised learning), in which relationships are found from the calculated indicators to a particular transportation mode. The estimation and definition of empirical data, which contains mathematical functions and several constants, is critical for the accuracy of the classification system.

For a conceptual view of the methodology, the reader is referred to Figure 11.

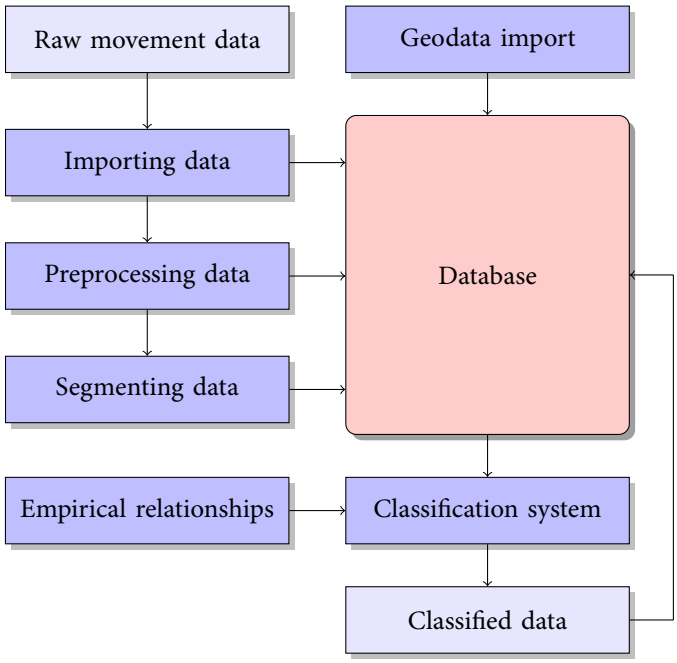


Figure 11: Overview of the methodology of the project.

3.1 SELECTION OF CONSIDERED TRANSPORTATION MODES

In literature, there is no generally accepted and complete list of transportation modes. Extensive publications often include seldom-used modes in the Netherlands such as camel, submarine, carriage, ski, balloon and sledge. Since there is not much benefit in considering these modes in this thesis, and since their inclusion could unfavourably affect the complexity and final results of the developed algorithm, in this thesis only the most frequently used transportation modes in the Netherlands are considered. Their list is composed from the last Dutch National Travel Survey by the *Centraal Bureau voor de Statistiek* (Ministerie van Verkeer en Waterstaat, 2009a), with the addition of sea and air transportation modes and underground: walk, bicycle, scooter, car, tram, bus, train, underground, sailing boat, ferry, and aircraft.

With the referred modes, I covered the vast majority of the used modes in the Netherlands, but also in many other countries, facilitating the segmentation and classification of trajectories acquired abroad. The entire list of all transportation modes is significantly longer, and can be found in various publications and on the web.

However, my experiments have shown that it is not possible to distinguish between a bicycle and scooter to a satisfying certainty, since there are too many overlapping and common characteristics of behaviour for the two (e. g. speed and used infrastructure). Therefore, scooter is left out of the classification system as a separate mode, and is merged with the bicycle. Discerning between scooter and bicycle is difficult for a human classification, let alone for a machine automated system.

The introduction of sea and air modes is novel in comparison to the related work. Although these modes are not in the focus of movement research, the dataset contains journeys made with the mentioned modes, and their classification might be useful for other applications.

For technical reasons, I introduce the *mode* standing for describing segments which are logged on approximately the same location. Motionless segments usually have the purpose of waiting and may give a good indication of the transition of a transportation mode. Examples for this are standing at a station for waiting a bus or train. Although standing may seem very straightforward to detect, occasional walking may occur, and/or noisy measurements may cause points not to have the same position. However, standing is not emphasised and it is not listed as an additional "transportation mode".

As it is shown in the rest of the thesis, in a few cases discerning between a certain subset of the listed modes is not possible with a high certainty. Therefore, a hierarchy of transportation modes is introduced in order to give an accurate result for a group of similar modes, which is more acceptable than returning inaccurate or uncertain results. Three layers of transportation modes are generated, and the classification system returns classification results for each layer. The third layer is the most detailed layer which contains the above listed transportation modes, except scooter.

The first basic layer contains the most general grouping of the mentioned modes into land, sea, and air. The second layer requires more explanation. Distinguishing the following group of modes:

- Bus, tram, and car.
- Sailing boat and ferry.

in some cases may be very complex. To overcome the possible errors of the classification system, the second layer contains groups of these *complicated* modes. Hence, the second layer comprise seven transportation modes: walk, bicycle,

car/tram/bus (a single *mode*), train, underground, boat (comprises sailing boat and ferry), and aircraft.

Although the underground mode could be merged with train in the second layer of the transportation mode hierarchy due to similar characteristics, the achieved results are confident enough to completely separate these two modes already in the second layer.

The list of classified transportation modes if presented in the Table 3, with their hierarchy.

Table 3: List of considered transportation modes by layers.

1	Land	Sea	Air
2	Walk	Boat	Aircraft
	Bicycle		
	Car/tram/bus		
	Train		
	Underground		
3	Walk	Sailing boat	Aircraft
	Bicycle	Ferry	
	Car		
	Tram		
	Bus		
	Train		
	Underground		

3.2 SEGMENTATION

Since trajectories may contain multiple journeys and may have been completed with multiple modes, the trajectories have to be segmented for each, therefore there are two consecutive types of segmentation:

1. first the segmentation of trajectories to single-journeys segments (between two meaningful locations), and
2. segmentation of journeys into single-mode segments (also known as stages).

Although both segmentations technically derive segments, the segments in the first segmentation are referred to as *journeys*, and the latter simply as *segments*. Once a trajectory is segmented, it is ready for classification.

This section explains the concept of the segmentation into journeys (§3.2.1), and discusses the possibilities of different methods of detection of transition points and segmentation into single-mode trajectories (§3.2.2). Section 3.2.3 on page 29 discusses the possibility of modelling probabilities of transitions between particular transportation modes.

3.2.1 *Segmentation into journeys*

Different journeys are often separated by an interruption in logging the data, caused by either a signal shortage (individual in a building) or a turned off device. Moreover, the departing point of the next journey is usually close to the arrival point of the previous journey. By examining the datasets I concluded that most of the journeys are mutually separated by longer period such as a working shift (8-9 hours) or a night, hence they are easy to detect. However, the real challenge is to separate journeys which are very close in time, and discern gaps in the data which are caused by "regular" signal shortage during movement, or by shortages caused by turning off the device or entering a building. The last mentioned problem is kindled by the fact that GPS receiver do not get an immediate fix (positioning solution) when turned on, rather after a period defined as Time To First Fix (TTFF), hence the first point after the signal shortage might be shifted in space in relation to the last known point before the data gap. This situation causes ambiguity—has the data interruption been caused by different journeys, or by a usual signal shortage during movement such as entering a tunnel or train.

TTFF is strongly correlated with time after the last fix, hence in case of different journeys the spatial shift is usually directly correlated with the time difference of the adjacent points in the data interruption. Figure 12 shows a case of data interruption. The last known point before the shortage is recorded at t_i, and the first known point afterwards at t_{i+1}.

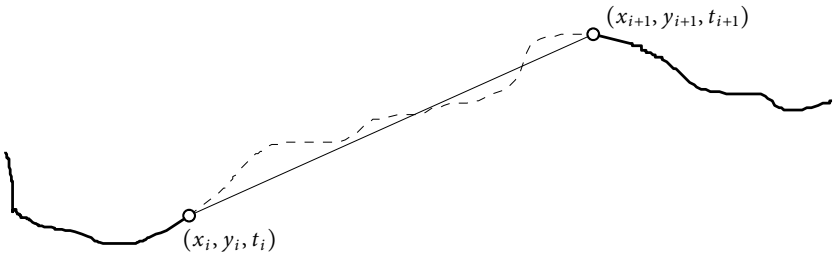


Figure 12: Data interruption may be caused by both signal shortage during travelling or in between two journeys (stay in a building).

Consider that the user ended the first journey at t_i, and started another afterwards at t_{i+1} - TTFF. The spatial shift $d = \| \mathbf{x_{i+1}} - \mathbf{x_i} \|$ might imply that the user continued movement which was not recorded. Although the exact trajectory may not be known, we can calculate the mean speed of the distance *as the crow flies*. It is to expect that due to long standing at the same point between two journeys, the speed is very low in case of two journeys. The threshold was put to 2 km/h, which is lower than the usual minimum speed of the slowest transportation mode in the system—walking. If the value exceeds the threshold, the trajectory is not segmented for multiple journeys, i. e. there is an indication that the data was interrupted during movement.

The system considers two thresholds for segmenting a trajectory into multiple journeys: a short break (10 min) and a long break (8 h). These thresholds (lower bounds) can be easily altered in the system if a user wants to have a more sensitive segmentation between two journeys. I have made tests that have shown that these values are sensitive enough to cover virtually all cases. In the short break, when the disruption in the data exceeds the mentioned threshold, the gap in the data is marked as a potential separation between the two journeys. As explained in the methodology, in that case the average speed is calculated between the boundary points of the two adjacent segments (before and after the shortage), and if it is below 2 km/h, then there is a strong indication that a new journey

was commenced. If the disruption passes the threshold of the long break, the trajectory is in any case split into two journeys since it is not nearly possible to reconstruct the movement and activities in more than 8 h of not recording. This threshold may be raised in cases of very long train or aircraft stages that result in signal shortage, but no such case was found in the test dataset, hence the presented thresholds are optimal, and are empirically derived after extensive examination of the test dataset and different situations.

3.2.2 *Segmentation into single-mode segments*

The second segmentation is more challenging—the transitions between modes occur much faster than transitions between journeys and in most of the cases are not *crisp*, and require a different approach. There might be three approaches in the segmentation into single-mode segments, described in the following sections.

3.2.2.1 *Segmentation with fixed-rate partitions*

In this approach, a trajectory may be first partitioned into segments of a constant size (either by time or distance). For instance, every one minute or 100 m (Figure 13).

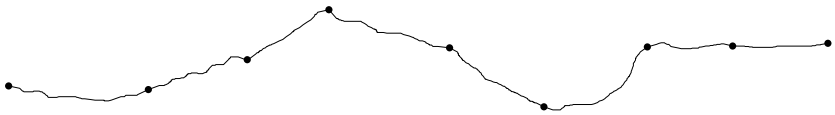


Figure 13: Partitioning of a trajectory into fixed-rate segments.

The classification system would classify each segment for the used transportation mode (Figure 14).

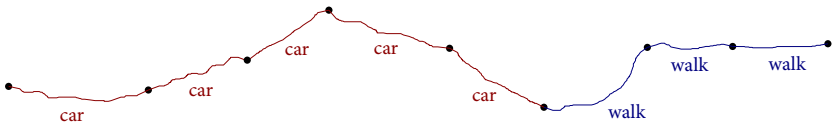


Figure 14: Classification of fixed-rate segments.

Segments with the same classification outcome would be merged into one segment (Figure 15).

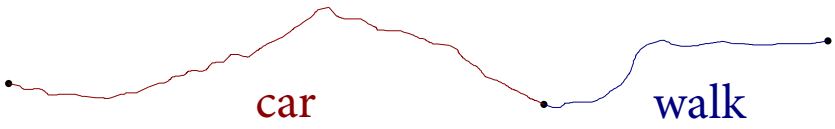


Figure 15: Merging adjacent fixed-rate segments with the same transportation mode.

This method is simple to implement, however, it has drawbacks which are not acceptable in this project:

- It is slow and not efficient.
- If the transition occurred in the between the partition points (most of the cases), the classification of that segment might be inaccurate, and the position and time of the transition not be accurately determined.

- If the segment where a transition occurred was marked as a transition segment, the further refinements of that segment would be inefficient and the classification, due to the low number of sampled points, difficult.

Although the method could be built in the way that starts by classifying large segments, and refining them due to uncertain results (that may be an indication of multiple transportation modes), it may not be possible to detect the combination of more modes in a segment, and the classification may even attribute another transportation mode.

3.2.2.2 *Segmentation by sensing changes in behaviour*

Mountain & Raper (2001) report that the indication that the user may have changed the transportation mode is a rapid and sustained change in direction or speed. Therefore, a segmentation algorithm would require detecting sudden changes in movement behaviour and detecting such points as transition points.

As an example, Figure 16 presents the speed behaviour of two transportation modes (train and car) with corresponding statistical descriptors for the whole journey (for both modes together) and the marked transition.

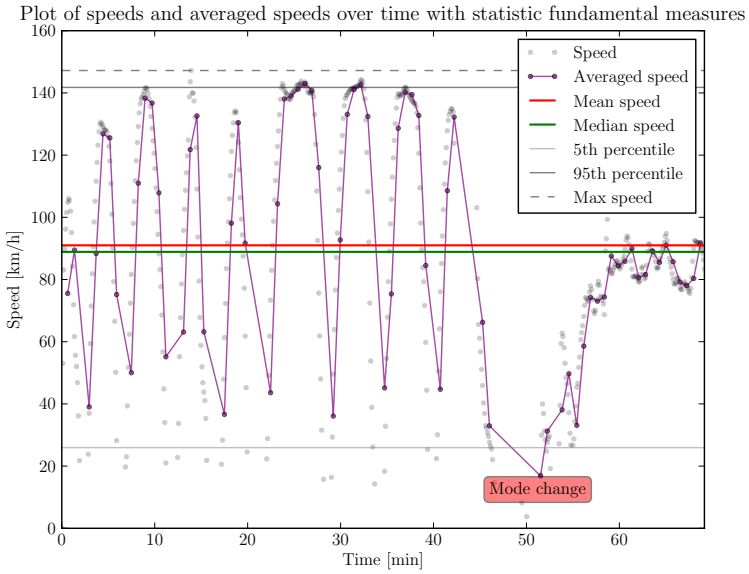


Figure 16: Differences in speed and speed behaviour for two different transportation modes, with statistical descriptors. For a better overview, the speed is averaged every 10 points and plotted in purple. The transition between the two modes is marked at the 50th minute.

Although for humans it is quite easy to segment the presented movement, the development of a segmentation algorithm sensing movement changes may be complex.

3.2.2.3 *Segmentation with calculating the proximity to the nearest potential transition points*

Liao et al. (2006) segment multi-modal trajectories by analysing the proximity to potential-transition locations such as bus stops. This method presents another interesting use of GIS data. However, their approach may have difficulties in areas with dense traffic features (especially in the Netherlands), where the distance

between potential transition points for various modes may be in the range of GPS errors, hence this method is used only partially in order to discern between cars, busses and trams (it is discussed in more details in § 5.6.2 on page 67).

3.2.2.4 *Segmentation by detecting stops and signal shortages*

A few researchers (e. g. Zheng et al., 2010) indicate that a person usually walks or stops during the transition. By examining the test dataset and observed the same behaviour, I choose to second their conclusions, and to follow this logic. However, by examining the available data I noticed that the transitions often cause data interruption (signal shortage under the roof in a train station, or entering a bus), hence I have to add signal shortages to the list. They are used as an additional indication for a potential transition.

All stops which are longer than a specific threshold, and last points before a signal shortage, are considered as *potential transition points*. These events indicate that the transportation mode *might* have changed, and the segment should be terminated and classified.

In forming the presented threshold for *stopping* a segment, it should be taken into account that oversegmentation of the trajectories is better than undersegmentation:

- In oversegmentation very short segments are possible and the segmentation algorithm detects more stops, but errors are more likely to happen, while in
- undersegmentation, although the resulting number of segments is lower and computationally more acceptable, it is eventual that fast transitions may pass undetected—e. g. exiting a tram/bus, and immediate departure of some modes,

hence a sensitive threshold should be established. For instance, a case when a person is running to a tram which immediately departs should be registered as a stop as well. This kind of fast transitions should be detected with a low threshold. By setting a low threshold for short stops, it is possible to ignore walking segments as an indication of the transition between two modes. Hence this segmentation method uses stops and standing segments as an indicator for the transition between two transportation modes.

Since many single-mode trajectories contain stops, e. g. cars stopping for traffic lights, initially the trajectory may be segmented in a high number of segments and points of potential transition. However, after each segment is classified, adjacent segments with the same mode are merged (Figure 17).

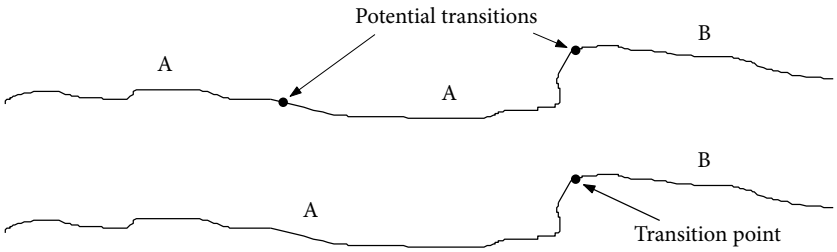


Figure 17: All stops and signal shortages in a trajectory are first marked as points of potential transition, and the segments are classified separately. If two adjacent segments have the same classification outcome, they are merged into one segment.

Each segment is terminated after a stop or signal shortage is encountered. The threshold for a data disruption is set to 30 s, while a stop is considered when for

more than 12 s there is no movement. Since in a stop the position might not be recorded at exactly the same position and there might be slight movement, to be more precise, a stop is marked when consecutive points in 12 s do not have a speed higher than 2 km/h. Since the majority of journeys start with standing to half of a minute, in the beginning of each trajectory the segmentation algorithm assures that the first five points of a trajectory cannot be segmented.

3.2.3 *Probability of transition between two modes*

There is a possibility of modelling the transitions between two modes in terms of probability. For instance, if it is known that most people after cycling make a transition to walking, a transition matrix describing the probability of transition between two modes may be built, and used to predict the next transportation mode. This part was investigated, but it was not used since I concluded that it may degrade the results. In the test dataset, there are several "exceptions" which would not be correctly classified in this case, e. g. a case of a transition from a train to a car is rare, but it appears in the dataset. Moreover, modelling such transitions is outside the scope of this thesis since it requires considerable research. Therefore this system should not greatly rely on such assumptions before a segment is considered.

3.3 CONCEPT OF THE CLASSIFICATION

This section describes the concept of the classification of single-mode segments for transportation modes. The solution is built upon the theory of expert systems, but since it does not include complex reasoning as expert systems do, it may not be considered as a "typical" expert system. However, since many concepts are used from the expert systems, this concept may be considered as rather an expert system alike solution.

3.3.1 *Basics of expert systems*

There are various general approaches for building a classification system, as it can be concluded from different existing approaches described in the preceding chapter. In this thesis, an expert system approach is chosen because of its maturity, but also to check the performance of expert systems in this problem, since there are no records of them being used. Although expert systems did not witness much development in the last decade, they are still current, mature enough and suitable for solving this problem—they are well known for their suitability to all kind of classification problems (Holzmann et al., 1999; Rearden et al., 2007; Wentz et al., 2008).

An expert system is a software package that can reason through complex situations. It comprises the knowledge of an expert in a certain field to provide answers to problems (Buchanan & Duda, 1982). They are applicable to specific problems, and developed in the frame of artificial intelligence in order to substitute experts, usually in a very narrow field. The typical usage of expert systems is in medicine (Grazia, 2006). Expert system are used in GIS, for instance in the Cartographic Expert System (Alkemade, 2000; Kotte, 2002; Van Oosterom et al., 2001), and an expert system for polygon classification of topologically structured topographic data converted from spaghetti data (Van Oosterom, 1999).

Fundamentally, expert systems consist of a knowledge base (facts), and an inference procedure (rules), which derive conclusions. Rules are usually of a basic construction (input-output mapping relationship):

IF *e* is observed THEN *h* is true

where *e* is a condition or evidence, and *h* is a hypothesis. For example: IF (the streets are wet) THEN (it rained).

A rule-based system requires some kind of program to manipulate the rules. The procedure that does this is called an inference engine, because in many rule-based systems, the task of a system is to infer something, e. g. a classification, from the data using the rules (Wilson, 2003).

An expert system can be realised in any programming language. However, there are available tools, for instance the C Language Integrated Production System (CLIPS), the most widely used expert system tool (Giarratano & Riley, 1998), and its Java descendant Jess (Friedman-Hill, 2003), which facilitate the implementation of expert systems.

Another important concept in expert systems is (un)certainty, which occurs when one is not absolutely certain about a piece of information (Nickles & Sottara, 2009). The degree of certainty, introduced by Shortliffe et al. (1975) is represented by a numerical value—Certainty Factor (CF), a quantification of the confidence that an expert might have in a conclusion or hypothesis *h* that s/he has arrived at from an evidence *e*—CF(*h*, *e*). In some expert systems, the range of CF is usually from 0 to 1, or from -1 to 1, where negatives would indicate disbelief, evidence that contradicts a hypothesis. It is important to note that this term should not be confused with probability, since certainty factors are relative and approximate measures determined by an expert, and the whole concept is not strictly defined. For instance, there is no rule that multiple CFs should sum up to one, in contrast to the probability theory.

In case of multiple CF from a set of conclusions (hypotheses), the obtained CFs should be propagated through a reasoning chain, i. e. combined, where several inference methods had been established. For instance, in MYCIN, an early expert system developed in the early 1970s at Stanford University, when two CF are ANDed (conjuctive reasoning), the joint CF is the minimum value of the two (Shortliffe & Buchanan, 1975):

$$\mathrm{CF}[A \cap B] = \min(\mathrm{CF}[\mathrm{A}], \mathrm{CF}[\mathrm{B}])$$

in contrast to the situation when they are ORed (disjunction), when the maximum value is taken:

$$\mathrm{CF}[A \cup B] = \max(\mathrm{CF}[\mathrm{A}], \mathrm{CF}[\mathrm{B}])$$

These two approaches may be also seen as a pessimistic and an optimistic assessment, respectively.

When CF occur in a sequence, the result CF is found by multiplying the CF in the chain (Grimshaw, 2001; Lucas, 2001). Certainty factors are a practical and straightforward alternative to Bayesian reasoning, however, it is not mathematically pure and lacks mathematical correctness of probability theory (Negnevitsky, 2005). Nevertheless, this approach is suitable for being used in this thesis.

In relation to this project, an example of a fact is the mean speed of a trajectory—30 km/h. By considering only the mean speed of a trajectory, there is suggestive evidence that the value of the speed *probably* represents a car:

IF (mean speed is 30 km/h)
THEN (mode = car)
WITH CF = 1.0

In this system the facts are realised with indicators and are described in § 3.5 on page 36 in more details with practical information.

However, one of the limitations of classic expert systems is that they are based on crisp logic and do not provide the expression of imperfection, i. e. the absence of fuzziness (Ghorbel et al., 2009), hence Fuzzy Expert System (FES), which uses fuzzy logic rather than Boolean logic should be introduced (Garibaldi, 2005). Fuzziness occurs when the boundary of a piece of information is not clear-cut (*crisp*), and when there is no single quantitative value to classify an information. For instance, the concepts young, tall and good are fuzzy and depend on the context (Orchard, 1998), hence, respective membership functions, assigning partial membership to a class, should be defined.

A fuzzy set A in a universe of discourse U is characterised by a membership function

$$\mu_A : U \rightarrow [0,1] \tag{3.1}$$

which associates with each element x of U a number $\mu_A(x)$ in the interval [0,1] which represents the grade of membership of x in the fuzzy set A (Zadeh, 1975).

3.3.2 *Assigning certainties in the classification*

The fuzzy expert system developed for this thesis uses fuzzy logic to derive certainty factors, i. e. *fuzzy variables are used to assign certainties to each derived hypothesis*. Consider the following case of a rule as an explanation of the concept. If the maximum speed in a segment is 118 km/h, from common sense we could build a rule that with a high certainty may conclude that the used mode is a car:

IF (max. speed is 118 km/h)
THEN (mode = car)
WITH CF = 1.0

Higher speeds for cars are seldom, however, they should not be discarded as possibilities, since there still *might* exist a possibility that the segment was completed with a car. In order to retain the reasoning, but give it less weight, this is done by assigning a lower CF:

IF (max. speed is 138 km/h)
THEN (mode = car)
WITH CF = 0.6

Therefore, the certainty factors in this fuzzy expert system are a function of available evidence: CF = $f(e)$, i. e. membership functions which are empirically defined by investigating travel behaviour for each transportation mode in the training data, a subset of the test data used for that purpose.

Each available fact should be used for each considered transportation mode (class) in the system, for instance, extending the use of the information of the maximum speed for trains:

IF (max. speed is 118 km/h)
THEN (mode = train)
WITH CF = 0.4

Therefore, each rule in the system determines an array of certainty factors, one for each mode considered:

IF (max. speed is 118 km/h)
THEN (mode = {car, train, …})
WITH CF = {1.0, 0.4, …}

In case of multiple facts, the final CF is determined as a conjunctive CF since the rules are not fired in a particular sequence:

IF (max. speed is 55 km/h)
THEN (mode = tram)
WITH CF = 0.85

IF (proximity to tram network is 4933 m)
THEN (mode = tram)
WITH CF = 0

→ CF(tram) = min(0.85, 0) = 0

This is done for each mode respectively. It is visible that one rule in my FES could completely eliminate the possibility of a transportation mode based on one fact only, which is a positive "byproduct" for this thesis. Therefore, the presented expert system works on the *elimination of unlikely modes* by assigning them CFs of zero for each evidence that is against a hypothesis.

In order to formalise the presented concepts an overview is given. For each transportation mode m (e. g. train) of the n considered modes, the classification system contains k membership functions f_m^i, where k is total number of indicators (facts) used as the input of the classification and i marks the designation of the indicator, e. g. f_2^3 or $f_{\text{train}}^{\text{max.speed}}$. For each segment, k indicators $i_{1\ldots k}$ are calculated (e. g. i_3 or $i_{\text{avg.speed}}$) and passed to the respective membership functions for each transportation mode (e. g. $f_{\text{train}}^3(i_3)$, $f_{\text{car}}^3(i_3)$, $f_{\text{bicycle}}^3(i_3), \ldots$) from which certainty factors $\text{CF}_m^i = f_m^i(i)$ are calculated. The total number of the membership functions and corresponding certainty factors is the product of the number of indicators k with the number of the considered transportation modes n.

After computing the k certainty factors for each transportation mode, the system determines the minimum value for each and considers it as a the "final" CF, i. e. confidence that the mode in question was used to complete the classified segment:

$$
\begin{array}{lllll}
\text{CF}_1^1 = f_1^1(i_1) & \text{CF}_1^2 = f_1^2(i_2) & \ldots & \text{CF}_1^k = f_1^k(i_k) & \Rightarrow \text{CF}_1 = \min(\text{CF}_1^1, \ldots \text{CF}_1^k) \\
\text{CF}_2^1 = f_2^1(i_1) & \text{CF}_2^2 = f_2^2(i_2) & \ldots & \text{CF}_2^k = f_2^k(i_k) & \Rightarrow \text{CF}_2 = \min(\text{CF}_2^1, \ldots \text{CF}_2^k) \\
\vdots & \vdots & \vdots & \vdots & \\
\text{CF}_n^1 = f_n^1(i_1) & \text{CF}_n^2 = f_n^2(i_2) & \ldots & \text{CF}_n^k = f_n^k(i_k) & \Rightarrow \text{CF}_n = \min(\text{CF}_n^1, \ldots \text{CF}_n^k)
\end{array}
$$

For more insight in the theory of expert systems, the reader is referred to the publications cited in this section. Technical details of the implementation of the system and definition of membership functions are given in Chapter 5 on page 57. The membership functions are conceptually discussed further in the next section.

3.3.3 *Types of trapezoidal membership functions*

A common construction of a Membership Function (MF) is trapezoidal, as it is shown in Figure 18.

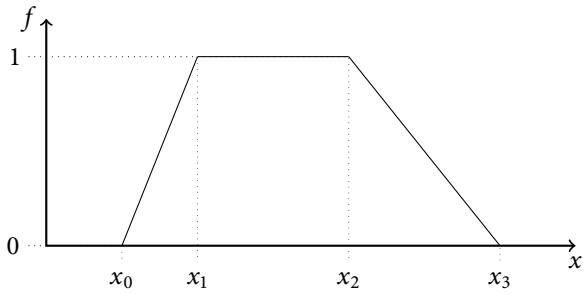


Figure 18: A trapezoidal membership function.

This is also one of the simplest constructions, and it is suitable for this FES approach. It requires the definition of four points, where x_0 and x_3 correspond to a certainty of zero, while x_1 and x_2 to one. Every value in between is considered as fuzzy. It is important to note that in this concept the range of the derived values by the MF is [0,1], and the input of the indicators is always non-negative:

$$f(x) = \begin{cases} 0, & \text{if } x \leq x_0 \\ \frac{x-x_0}{x_1-x_0}, & \text{if } x_0 < x < x_1 \\ 1, & \text{if } x_1 \leq x \leq x_2 \\ \frac{x_3-x}{x_3-x_2}, & \text{if } x_2 < x < x_3 \\ 0, & \text{if } x \geq x_3 \end{cases}$$

There are a few types of MFs used in this thesis:

1. The standard trapezoidal MF described above (Fig. 18). It may start at zero ($x_0 = 0$), and it is the most frequent type of MF used in the definitions.

2. The MF which is used in occasions where all cases are possible (Fig. 19), i. e.

$$f(x) = 1, \forall x.$$

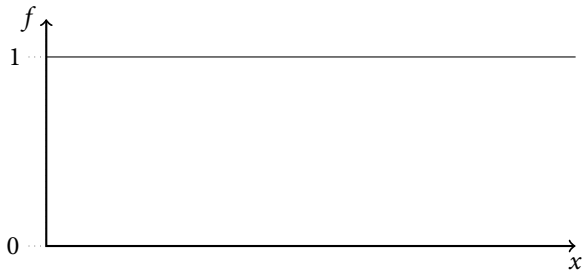


Figure 19: A trapezoidal membership function resulting in a CF of one for all the input values.

   For instance, in the indicator of the proximity to the nearest bus line, in the MF for cars any input value returns a CF of one, since cars are not related to a bus line, and in any case the certainty for a car should not be affected.

3. The MFs used for the indicators of the proximity for a certain mode usually start at zero, but it is not limited at a certain value as buffers are (Fig. 20).

Since there is not a crisp boundary for possible GPS errors, it smoothly drops after a value, depending on the infrastructure:

$$f(x) = \begin{cases} 1, & \text{if } x_{0,1} \leq x \leq x_2 \\ \frac{x_3 - x}{x_3 - x_2}, & \text{if } x_2 < x < x_3 \\ 0, & \text{if } x \geq x_3 \end{cases}$$

For instance, the four values for car in the indicator of road proximity are 0, 0, 15, and 40 (metres).

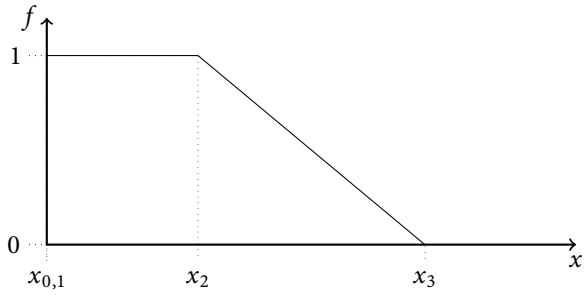


Figure 20: A trapezoidal membership function which starts with a CF of 1 and decreases its value until x_3.

4. In some cases, a MF is not bounded on the right side ($x_{2,3} = \infty$):

$$f(x) = \begin{cases} 0, & \text{if } x \leq x_0 \\ \frac{x - x_0}{x_1 - x_0}, & \text{if } x_0 < x < x_1 \\ 1, & \text{if } x \geq x_1 \end{cases}$$

This is useful in order to reject some modes for trajectories which are very close to a certain proximity, for instance, a trajectory very close a railway may not be made by a car or bus. See Figure 21 for an example.

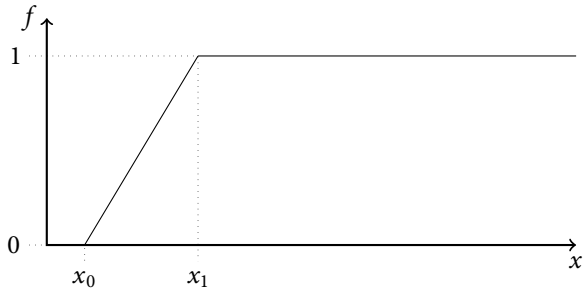


Figure 21: A trapezoidal membership function which is not bounded on the right side.

5. A type of MF which inputs a boolean value and returns a boolean value, i. e. a CF of either zero or one. This is used in the indicator of the knowledge of the position of a trajectory over water surfaces. For instance, in the MF for land modes, the value `True` returns the CF of `0.0`.

3.4 USED TOOLS

3.4.1 *Software*

The prototype was implemented in Python, which was selected because of its share in the M. Sc. Geomatics curriculum, and availability of modules for GIS operations and statistics (Numpy and Scipy). Although Fuzzy Expert System solutions are available, I decided to build my own FES in Python completely from scratch, not using present solutions in order to tailor the system for this specific problem. Additional reasons for this approach are:

- To have complete control and overview over the processes during the reasoning in the segmentation and classification.
- The implementation of my approach, with fuzzy certainty factors and in an form of an array, and reliable elimination of unlikely modes.
- To gain more knowledge in the field of expert systems, fuzzy logic, machine learning, and artificial intelligence.
- Usage of an available FES tool requires exporting the processed data from Python and then reimporting the results. In my approach, the whole process from importing the data to delivering the segmentation and classification results is done in Python.

The data (both raw and preprocessed) are stored in PostgreSQL, as the used geodata. PostgreSQL is an object-relational DBMS, with a notable share in Geoinformation Systems thanks to its extension PostGIS which adds support for geodata types.

It is interesting to mention that these two solutions are open-source. Hence, this system was completely built on free software, which is also available for all three mainstream operating systems nowadays (Windows, Mac OS X, and Linux), making possible the usage of the developed prototype on virtually all machines nowadays.

For this project, a dedicated server for hosting the developed system and storing the data was available: with an Intel Core Duo processor, running Ubuntu Linux 9.10. Additionally, for reviewing the GPX files on imagery and better presentation, Google Earth (former EarthViewer 3D by Keyhole, Inc.) was used, from which a few screenshots are included in this thesis.

3.4.2 *GIS data*

Geographical data used for this project should fulfil the following requirements:

- They must be available for free usage for this project.
- They should contain all needed features (e. g. bus infrastructure).
- The data should have worldwide coverage so trajectories from abroad could be segmented and classified as well.

One of the few sources that is suitable for usage in this project is OpenStreetMap (OSM), a collaborative project to create a free editable map of the world. OSM licensing permits the free usage of the data for this project, it is widespread worldwide, and it contains all the geographic features required by this project. The Figure 22 on the next page shows an example map from the OSM project.

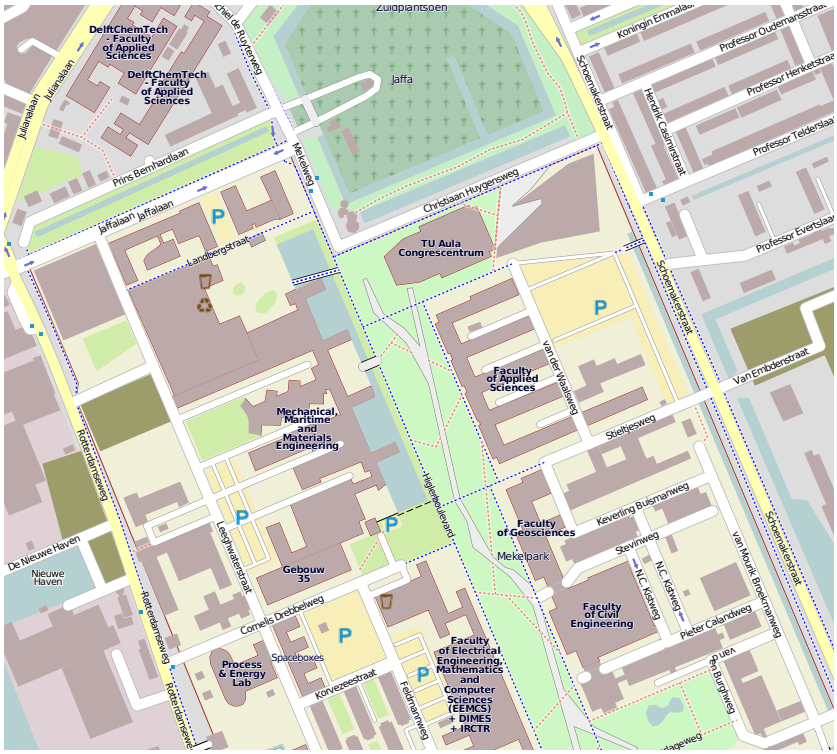


Figure 22: Rendered OpenStreetMap data, showing the northern part of the TU Delft campus. Situation as of 27 April 2010.

The data in the project is up to date and accurate enough, especially in the Netherlands since in 2007 the Dutch mapping company Automotive Navigation Data donated the entire street map of the Netherlands to OSM.

More technical details of OSM is given in the next chapter, along with the description of the import process.

3.5 SELECTION OF THE INDICATORS

Indicators are statistical descriptors derived from the data, from a set of points, that might be informative about the used transportation mode in a segment. They are represented as either a single float or boolean value i, from which a certainty factor for a particular transportation mode m is derived through a corresponding membership function:

$$\mathrm{CF}_m^i = f_m^i(i)$$

For instance, an example of an indicator is the mean speed $\overline{v}$ of a segment:

$$\left.\begin{array}{c}(x_1, y_1, z_1, t_1, v_1)\\(x_2, y_2, z_2, t_2, v_2)\\\vdots\\(x_n, y_n, z_n, t_n, v_n)\end{array}\right\} \quad \frac{1}{n}\sum_{i=1}^{n} v_i = \overline{v}$$

$$\overline{v} \implies \begin{aligned} \mathrm{CF}^{\overline{v}}_{\mathrm{train}} &= f_{\mathrm{train}}(\overline{v}) \\ \mathrm{CF}^{\overline{v}}_{\mathrm{car}} &= f_{\mathrm{car}}(\overline{v}) \\ \mathrm{CF}^{\overline{v}}_{\mathrm{ferry}} &= f_{\mathrm{ferry}}(\overline{v}) \\ &\vdots \end{aligned}$$

However, as it is shown throughout this section, many transportation modes have overlapping values for each indicator, e. g. a car and train may have the same speed, hence the indicators are used primarily for rejecting unlikely transportation modes. For instance, if a trajectory is over a water surface, then all the land transportation modes such as walk, car and train may be ignored as a possibility. Another example is to remove the possibility of cars in areas where there are no streets that may be used by cars.

In order to comply with the efficiency requirement, there is a need to find the minimum possible relevant and reliable indicators that would lead to an accurate classification. Additional indicators may serve as a check, however, their redundancy might not be necessary and lead to higher computational complexity. Moreover, the chosen indicators proved to be reliable in the classification and elimination of unlikely modes.

The indicators are in general divided into two categories based on data dependency: values that can be resolved solely from the trajectories (from the timestamped positions), and values that require supplementary geodata, and both are presented in the continuation of this section. The behaviour of several indicators is investigated, but not all presented indicators are actively used in this thesis due to the reasons discussed in the following sections. Hence, the overview of the indicators is separated into two sections: indicators that are used in the prototype for classification of the trajectories (§3.5.1), and indicators whose behaviour is investigated and discussed, but due to several shortcomings not used (§ 3.5.2 on page 45). All such indicators have been calculated for each processed trajectory for research purposes, but later ignored due to the reasons discussed in respective sections. They are left in the prototype and the populated database for possible future work. A brief conclusion is given in § 3.5.3 on page 48.

3.5.1 *Indicators that are used for the classification*

SPEED One of the most important and obvious indications of a transportation mode is its speed (Bohte et al., 2008). With various statistical descriptors derived from the speeds, such as the mean speed and the maximum speed of a segment, it is straightforward to distinguish between modes such as walk, bicycle and car which have the significantly dissimilar speed characteristics and behaviour. However, various modes have overlapping speed characteristics, hence speed shall not serve as the sole indicator.

Speeds can be calculated either directly from the GPS receiver (from the Doppler shift), or from the difference between two subsequent coordinates with corresponding timestamps:

$$v_i = \frac{\| \mathbf{x_i} - \mathbf{x_{i-1}} \|}{t_i - t_{i-1}} \tag{3.2}$$

Both calculations are subject to noise, however, speeds derived directly from the coordinates are less accurate for reasons stated in §1.4.2 (p. 6), and in addition:

- The acquired positions always contain errors, hence the distance between two points (the numerator in Eq. 3.2) may not be always accurate to a satisfying level.

- Both timestamps are usually rounded to the nearest seconds, thus the time difference between two points is usually not correct (it is an integer). An example is given below. On the left two timestamps are expressed with more significant digits, i. e. milliseconds, and on the right side the rounded timestamps are presented. The difference between time differences of timestamps is notable.

```
2007-01-23T10:55:30.634Z      2007-01-23T10:55:31Z
2007-01-23T10:55:37.349Z      2007-01-23T10:55:37Z
------------------------      -------------------
                  6.715s                        6s
```

- The speed between two sampled points may be variable, rather than constant.

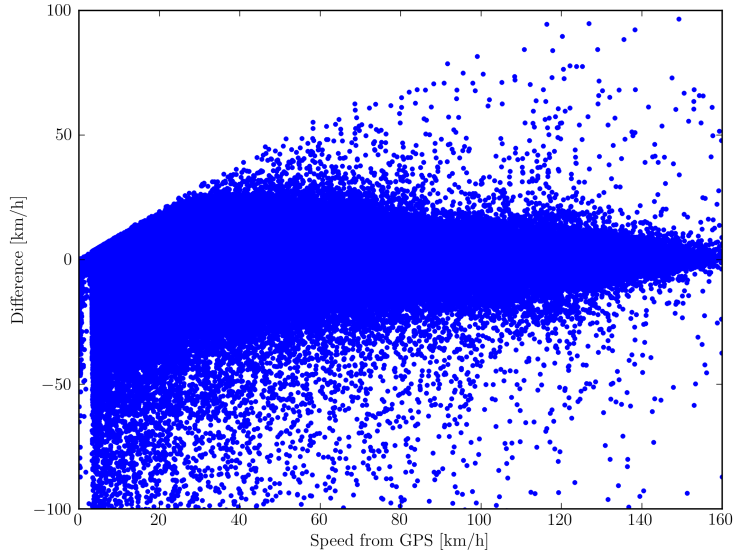
This leads to a conclusion that contrary to the speeds calculated from the GPS receiver, the speeds calculated from two subsequent coordinates is rather an approximation of the average speed of the shortest path between two points. This is further elaborated in the §A.2.2 on p. 103 and Fig. 58 (p. 105).

As stated earlier, although virtually all GPS receivers available on the market immediately derive the speed from the GPS measurements, some of them do not store it in the GPX output, hence the speeds in these cases should always be calculated from the coordinates.

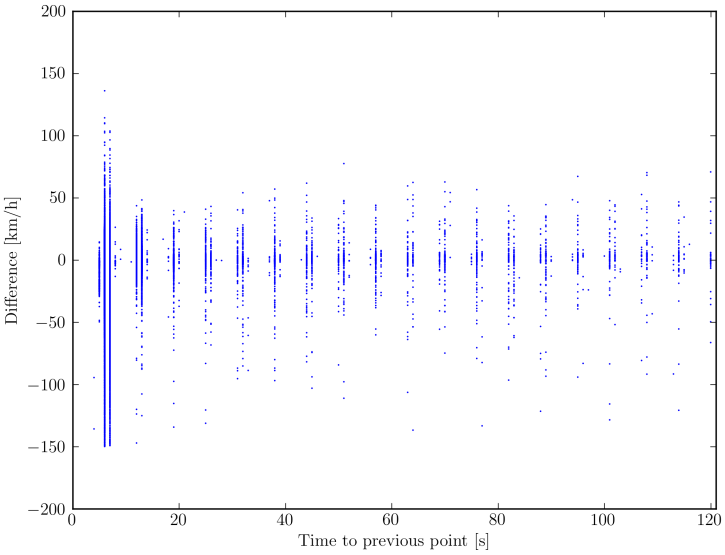
In order to investigate this approach and compare the two speeds derived from different sources, the speeds have been calculated from the coordinates for all points in the test dataset where the speed from the GPS receiver was available. The reader might be interested in the two aspects: are the differences between the speeds higher for movements with higher speeds, and are the differences positively correlated with the length of sampling periods (do longer sampling periods yield a higher difference since the presented variations may occur more frequently over longer periods). The differences are computed only for receivers in motion, with the intention of eliminating the potential bias caused by motionless points, where the difference is zero anyway.

Calculations in the prototype and Figure 23 on the next page show that there is no correlation between the difference of the two calculated speeds, and the sampling period or speeds. The discontinuities visible in Fig. 23b are due to the integer values of the sampling periods. The distribution of differences follows the Laplace distribution (see Fig. 24 on page 40). Although the three presented plots show that the differences may be as high as above 20 km/h, which can be considered as a high value by all standards, the frequency of these occurrences is below 1 %, which is clearly visible in the referred histogram. In this thesis three indicators derived from the speeds are used—the maximum speed in a segment, the mean moving speed, and the mean speed. Each is described in more details below.

The *maximum speed* in a segment is an important factor for the elimination of modes, e. g. if the maximum speed of a segment is 80 km/h, walk and bicycle may be safely ignored under all circumstances. However, the maximum speed should not be directly considered as it is subject to noise, hence the nearly-maximum speed, realised with 95th percentiles, should be used (Schüssler & Axhausen, 2009), i. e. the value of a variable below which a 95 % of the observations fall. Figure 25 on page 40 shows the problem with noise and benefit of this approach. The segment contains a noisy point with the value of the speed above 160 km/h which is unrealistic with respect to the other sampled points in the segment. By taking into account the maximum value in the subset, the indicator would be biased because of this single, but protruding value. Using the 95th percentile in



(a) Differences in relation to GPS speeds.



(b) Differences in relation to the sampling period.

Figure 23: The differences between calculated and GPS speeds have no correlation with neither the speed or sampling period.

this case, the nearly-maximum speed gives a safer approximation of the highest value in the segment.

Since I strived to also classify short segments where just one noisy point may compromise the calculations, 95th percentiles in that cases are prone to noise. Therefore, 70th percentiles are used. The threshold for shorter segments is put to 15 points. For segments that contain more than 15 points, 95th percentiles are used.

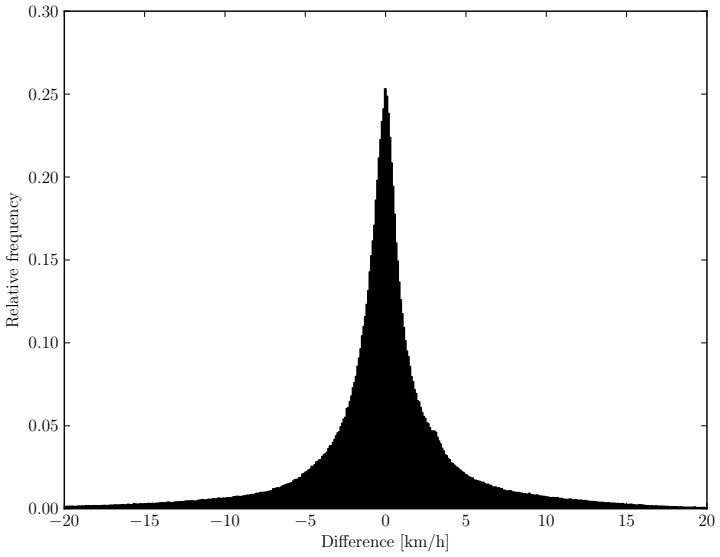


Figure 24: The distribution of differences between calculated and GPS speeds follows the Laplace distribution.

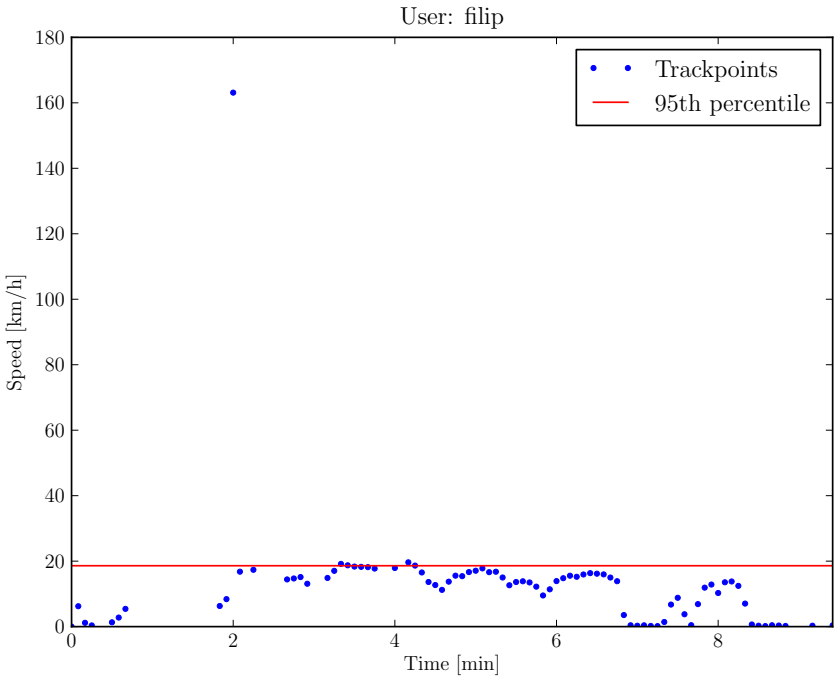


Figure 25: Usage of 95th percentile rather than maximum values is a straightforward solution for reducing noisy observations. This segment, which contains one noisy point, had been completed with a bicycle.

The mean speed is an useful indicator for getting a general idea of a transportation mode in a segment. For instance, although trains and cars may have the same maximum speed in a trajectory, trains, in general, tend to be faster than cars, hence in such cases a higher CF is put for a specific transportation mode.

However, most of the presented transportation modes involve stopping which bias the mean speed. Although the calculated mean speeds including stops is technically a mean speed, it is not acceptable in this consideration. For instance, long stops of trains at large railway stations, in contrast to highways without stops, may result in a significantly lower train's mean speed, contrary to the previous statement. Figure 26 shows the quantity of zero-values in a trajectory made by car.

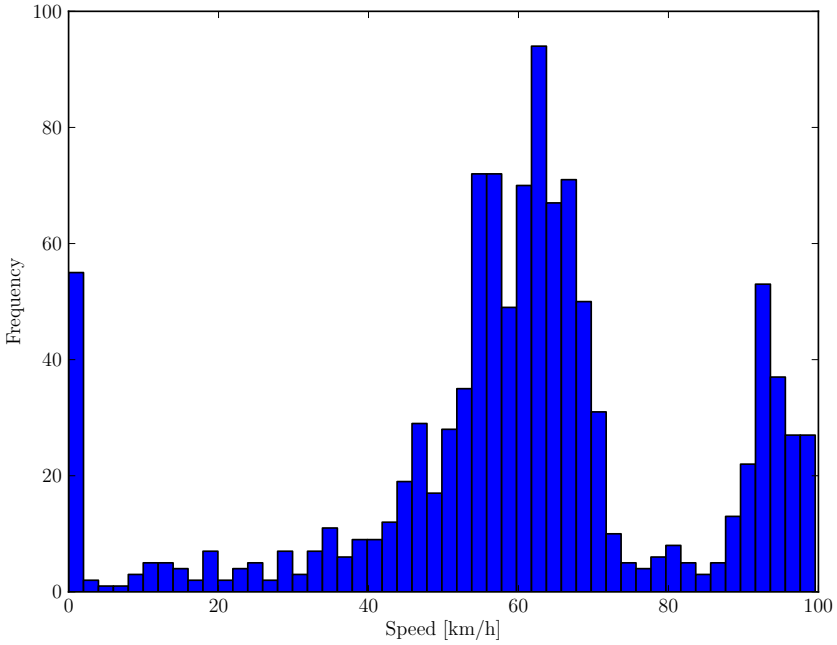


Figure 26: Histogram of speeds during a trajectory made with a car. The stops are visible at zero.

Therefore, only speeds from moving points in a segment are considered in the calculation of the mean speed of a segment, resulting in the *mean moving speed*. The threshold is set to 2 km/h, rather than zero, to account for GPS noise which result in small but notable non-zero speeds.

The second benefit of this approach is to solve the mentioned problems with traffic congestions. Since stops are therefore not counted, the mean speed is not significantly lowered in a traffic jam, and very slow movement is not registered.

Although the mean speed of a segment is not used in general due to above discussed reasons, it is used in order to distinguish standing and walking. As mentioned earlier, standing may involve occasional walking which results in moving speeds, but it should still be considered as standing. Calculating the mean moving speed does not contribute the classification since it is equal for both modes. However, standing, due to a higher number of zero speeds, implies a significantly lower mean speed.

With this approach, also occasional GPS jumps due to noise, frequent for standing-related locations such as roofed railway stations are accounted.

Although most of the researchers agree that the speed is the most useful indicator (e. g. Byon et al., 2006), I have to add that it is hard to model travel behaviour only in speeds since all modes have a long spans of possible and usual speed, and many cases of infrequent travel behaviour should be taken into account for best classification results, e. g. car segments of a speed significantly higher than local speed limits (in some cases above 170 km/h).

PROXIMITY TO A NETWORK Several transportation modes are constrained to a network infrastructure, e. g. trains travel on railways, hence geodata such as railway infrastructure and highways can help automatically detect corresponding modes. This obvious approach is used in a number of methods (De Boer, 2008; Stopher et al., 2008), for instance where a car and bus are distinguished by the locations of the bus stops (Chung & Shalaby, 2005). On the other hand, by analysing the proximity of the trajectory to a network, these indicators could be used to rather reject particular modes, such as trains in railway *null* zones, i. e. outside a railway buffer. In this thesis, the following networks have been taken into account:

- Railway (all types).
- Tram lines. In OSM they are usually stored in conjunction with railway, hence they had to be separated.
- Roads (all types).
- Bus network (superimposed attribute on the road infrastructure). It is used in order to rather reject possibilities for buses on roads not included in a bus network.
- Underground lines (with segments that are not underground where might be GPS reception).

In order to determine the thresholds on how to limit the proximity to a network in order to consider particular modes (build a buffer), but also how to model the respective membership functions in the fuzzy domain, two foreseen deviations should be taken into account:

- Unavoidable GPS errors cause deviations from the infrastructure. It should be taken into account that the distribution of GPS errors follows a Chi-square distribution (Teunissen et al., 2008), hence membership functions should be modelled with respect to the distribution. The magnitude of errors is variable between different types of receivers and situations.

  Figure 27 shows two segments made with a train on the same route, on two different days. One segment lies in the bounds of the railway area, however,

Figure 27: Two segments made on the same railway in different timeframes. Their difference is approx. 80 metres. (Imagery © Aerodata International Surveys and Google (2010).)

  another is approx. 80 m away from the first segment, even positioned on a road, which may result in an incorrect attribution of the used mode to a car or bus. This case indicates that GPS errors should be seriously accounted since the accuracy may be inferior to the manufacturer's claimed accuracy and cause significant errors in the classification.

- The network infrastructure in OSM is modelled with lines. Hence, without the presence of GPS errors, a trajectory on a road always contains a deviation from the corresponding infrastructure data, which is especially notable in roads with more than one lane. The distribution of these differences depends on the width of the infrastructure, and how the infrastructure was modelled. In OSM usually no data are available on this.

Additionally, OSM data contains various errors from modelling. It is evident that membership functions should be extensive enough to compensate the sum of these errors, and take into account the worst-case scenario. As an example, Figure 28 shows that the distribution of distances from a road, in a trajectory made by a car.

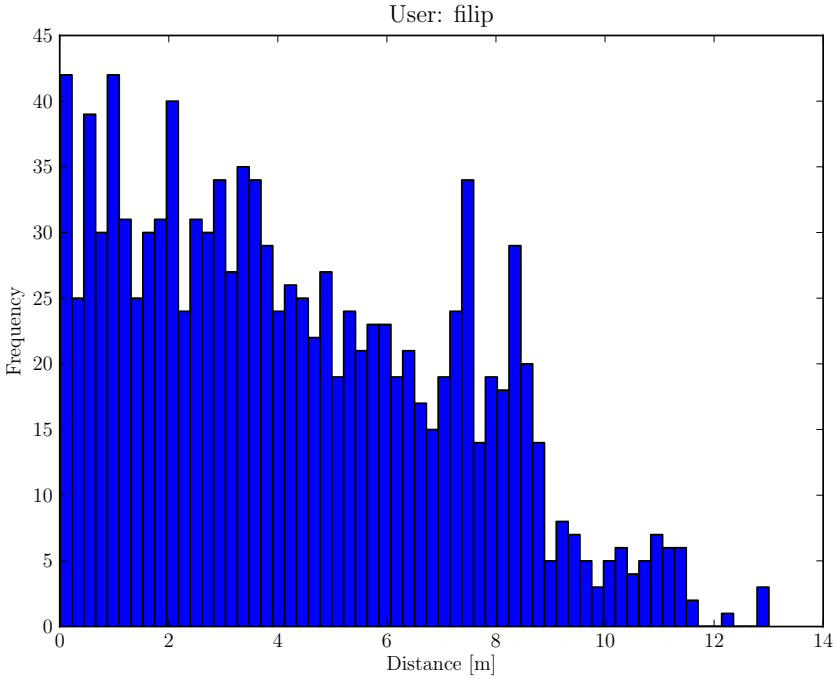
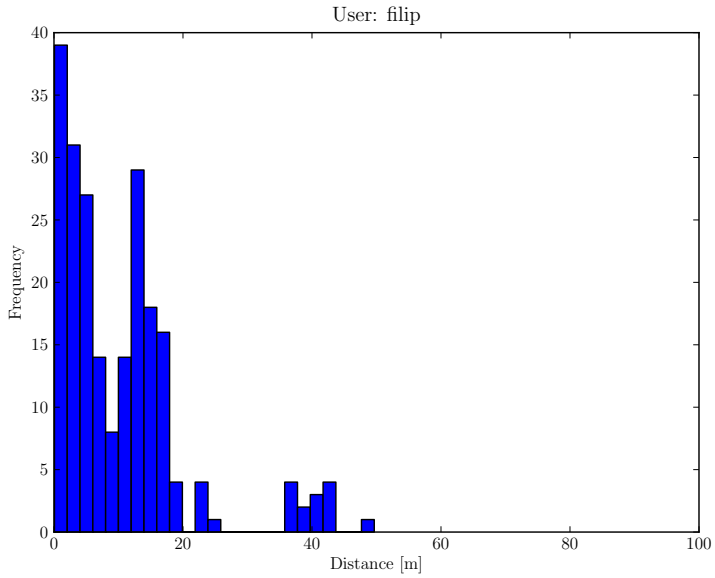


Figure 28: Distribution of distances from a road during a 20 minute car trajectory.

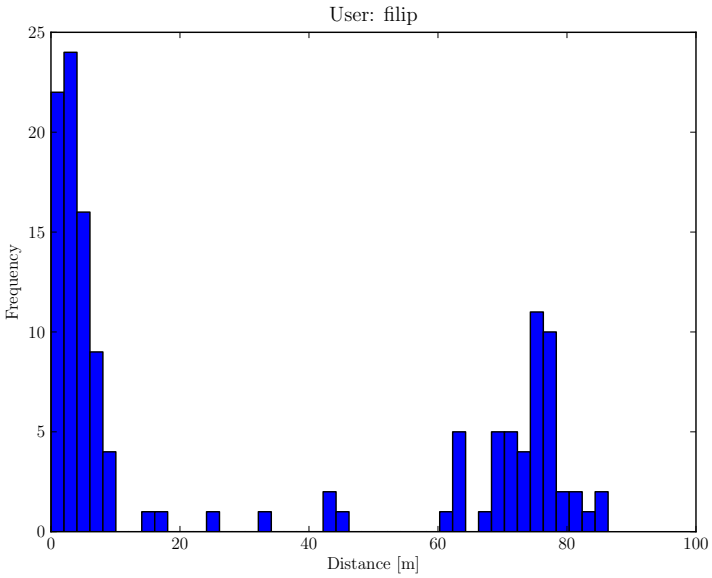
However, since it is not possible to separate the listed deviations from a measurement, they should be consolidated in one model of a membership function. After conducting experiments, I concluded that due to many possible cases an unique model cannot be found even between same infrastructures for the same mode. An example is shown in Figure 29 on the next page, where two histograms show the proximities of trajectories made with the same mode (train) on the same route to the infrastructure (railway). Even here with two very similar, if not equal, cases there are significant differences.

The quantified indicator is the mean distance of the segment to the nearest network feature of each considered network, i. e. six mean values are calculated for the following modes: train, tram, car, bus, underground, and aircraft.

OSM data also contains features as footways, pedestrian zones and cycleways, i. e. infrastructure for modes walking and cycling, however after implementing them I concluded that they should not be taken into account for two reasons: the density of these features is too high, causing the corresponding membership functions to cover almost the whole area anyway, and since one can walk and cycle virtually anywhere, even outside the infrastructure covered by OSM, these two modes should not depend on an infrastructure, except for rejecting them if they are over water surfaces, which leads the reader to the next indicator.



(a) First segment.



(b) Second segment.

Figure 29: Differences between two histograms of proximities from the nearest railway, for two trajectories made by a train on the same route but on different days.

WATER SURFACES The location of water surfaces, i. e. the information is the trajectory on water, can help to directly detect water transportation modes, but also discard all land modes. For each point it can be computed is it on water or not, as shown on Figure 30.

This indicator is a bit different from the indicators presented so far since it involves boolean values, however, its return value is the ratio of the number of points marked on water over the number of all points in the segment, e. g. 96 %.

However, it should be noted that there are several cases to be handled, for instance, a trajectory crossing a bridge is marked as on water due to the structure of OSM data (again see Figure 30), and due to GPS errors, travelling near a

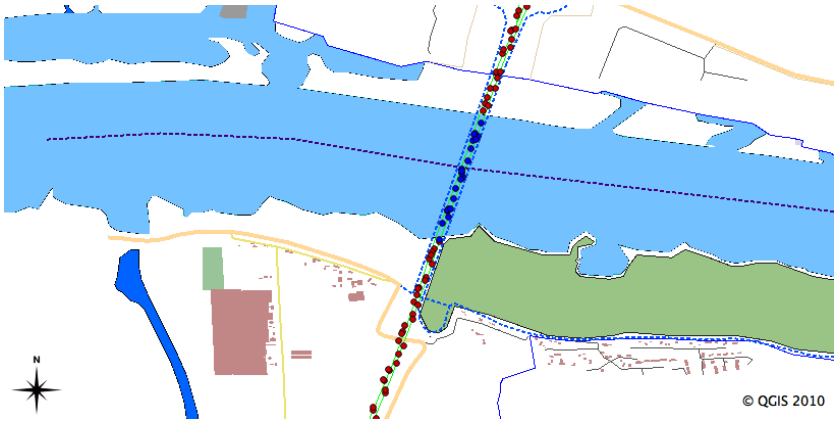


Figure 30: Points detected on water in contrast to points detected on land.

water surface may cause the trajectory to be marked over water. This problem is addressed in the implementation of the classification system (Chapter 5).

This indicator is not dependent on the sampling period.

POTENTIAL TRANSITION POINTS The locations of parking lots, stations and airports may be used for detecting the location where a transportation mode was changed. Hence, in addition to the computed distance for each point to the nearest network infrastructure, the proximity to the nearest stations for each mentioned network (e. g. bus) was calculated as well. Parking lots, as a few researchers suggest (e. g. Liao et al., 2006) are not considered, since the number of these features is very high, and not all of them are included in OSM, for instance, private garages.

However, as the presented segmentation method for detecting brief stops shown to be a reliable and a sufficient method for marking potential transition points, this indicator will be used only partially for distinguishing between car, tram and bus as it is further elaborated in section 5.6.2 on page 67.

This indicator is not dependent on the sampling period as long as at least one relevant point is recorded.

3.5.2 *Indicators that are investigated but are not used for the classification*

ACCELERATION The acceleration might contribute in distinguishing between modes with similar speeds, such as train and car. For example, a train has a lower acceleration and deceleration when departing or approaching stops (i. e. stations), than a car (i. e. traffic lights or parking lots) (Byon et al., 2007; Gonzalez et al., 2008; Stopher et al., 2008). The acceleration is determined from the speeds at two adjacent points:

$$a_i = \frac{v_{i+1} - v_i}{t_{i+1} - t_i} \tag{3.3}$$

Figure 31 on the next page gives the example of acceleration behaviour of the same trajectory presented earlier in plot showing speed behaviour (Fig. 16 on page 27).

The acceleration and deceleration should be viewed as two separate indicators since the absolute value of deceleration is usually not equal to the acceleration. The following values have been considered in the prototype and computed for each segment: the mean acceleration, mean deceleration, nearly maximum acceleration, and nearly maximum deceleration. After investigating the behaviour

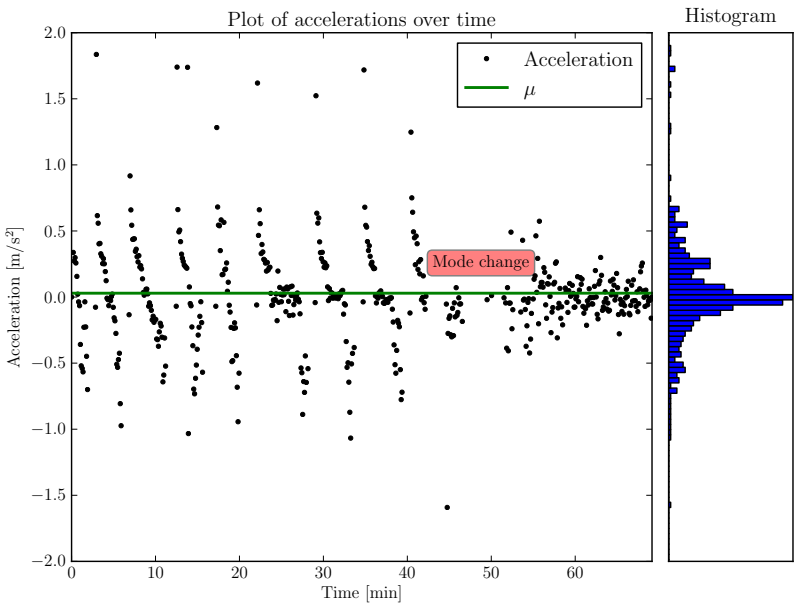


Figure 31: Differences in acceleration and acceleration behaviour for two different transportation modes.

of these values for different transportation modes, I can conclude that there is a noticeable, but slight relation to each particular mode in typical behaviour. For instance, the average nearly maximum value of the acceleration for a few trajectories made by bicycle was 0.7 m/s^2, while for car the value was 0.6 m/s^2. The common case for pairs of modes where the difference was higher is that other indicators already sufficiently contribute to their distinguishing (e. g. speed for train and walk, or intersection with a water polygon for ferry and bicycle).

However, beside the usually small difference in acceleration and deceleration between modes, there are two major difficulties for which neither the acceleration and deceleration are used as an indicator in this thesis:

1. Many segments contain specific situations such as sudden stops to avoid accidents and unexpected pedestrians, or to stop at traffic lights, hence the maximum deceleration values are comparable in the majority of transportation modes. In the test datasets, there are cases where two segments on the same road made by the same person with the same mode, but in a different time frame, proved to have significantly different acceleration and deceleration values.

2. Due to the variability of the acceleration, acceleration cannot be accurately estimated from segments recorded with long sampling periods τ. In the prototype, the acceleration was not calculated for points separated by more than 15 s, which is a threshold that can be changed in the developed prototype. From my examination of the data, I can conclude that the acceleration should be used only in segments with very frequent sampling periods (below two seconds), which are rare—only 0.25 % of the test dataset has a sampling period lower than or equal to 2 s. The presence of GPS errors further degrade the calculation of the acceleration.

STOP RATE Gonzalez et al. (2010) conclude that the number of stops in a fixed distance may be different for certain transportation modes, e. g. a car which shortly stops on traffic lights and a train which stops on train station

less frequently. While it is not possible to implement this approach due to my segmentation technique which automatically segments the trajectories between stops and classifies each segment separately, it is not useful too, especially in the Netherlands. Although busses in general indeed tend to stop more often than cars, since in addition to traffic lights have to stop at stations, at many places, especially in sparsely populated areas (e. g. Flevoland) the stops are not frequent, and there is not much difference between certain modes.

Moreover, this indicator was implemented in the prototype for research purposes. After the segmentation was disabled for this purpose, the prototype was fed with single-mode trajectories and stop rates for a variety of trajectories and transportation modes was calculated. However, not much benefit from using the stop rate was found due to non-existing links to a particular transportation mode.

HEADING CHANGE RATE Another possibility for an indicator is to calculate the distribution and rate of change of headings for each segment (Axhausen et al., 2004). This would help to distinguish a train from a car since a train never makes sudden turns such as car, but also the total rate of heading changes is different for most of transportation modes. This indicator is conceptually without doubt useful, however, in sampling periods higher than a few seconds the results from the prototype suggested it was not precise enough to be used.

VELOCITY CHANGE RATE Similarly to the previous heading change rate, also the velocity change rate could be computed (Zheng et al., 2010). This is different from the acceleration, since first involves calculating the following value for each point:

$$v_\Delta = \frac{v_{i+1} - v_i}{v_i} \tag{3.4}$$

Then the number of sampled points whose v_Δ is greater than a certain threshold v_r per unit distance.

After the implementation, I can conclude that as with acceleration, this approach is not useful in segments which are not sampled very frequently (1-2 s).

ELEVATION The elevation of a point is available directly from the raw data. It may contribute in the elimination of certain modes, e. g. sea modes that always have a very low elevation, but it is redundant since more reliable indicators are used for this purposes, explained in the following sections. Moreover, the available elevation value is usually of poor quality.

Although the elevation was not actively used in the classification due to these reasons which were well-known before, the elevation from each observation is stored in the database for future work or other purposes.

JOURNEY DISTANCE AND JOURNEY DURATION The travelled distance and duration of a journey may give a good hint on the transportation mode. For instance, if a journey covers 20 km it is more probable that a car was used rather than a bicycle.

In general these two indicators cannot be used in the full potential for the following reasons:

- They cannot be precisely linked to a transportation mode, they rather depend on a person's behaviour—for shorter journeys all modes can be used, but also that applies for longer trips.

- Recent studies shown that commuting times, and therefore journey distances, greatly vary between different countries (The Economist, 2009)

However, because of the used segmentation, the segments are rather short and these indicators are not applicable. Nevertheless, these values are saved for each segment and are used for informative purposes.

3.5.3 *Conclusion*

Several potential indicators have been evaluated in this section. Most of them have been discarded because of the low performance with sampling periods longer than a few seconds which is frequent in the test dataset. However, all of the indicators have been implemented in the prototype and are left for future work when data with more frequent sampling periods may be available.

The following indicators are used as an input to the classification: mean speed, mean moving speed, nearly-maximum speed, proximity to the nearest networks (bus, tram, train, underground, and streets that are accessible to cars), and the location of a segment with respect to a water surface. The proximities to the nearest stations as potential transition points are not used as indicators. They are used rather as facilitators for distinguishing in between a few modes only in uncertain situations.

4 IMPORTING AND PREPROCESSING THE DATA

This chapter concentrates on the technical aspects of importing and preprocessing the data, conceptually described in the previous chapter. The import and processing of the GIS data is described in §4.1.

The trajectories are first imported in the database (described in § 4.2), and then preprocessed for calculating the input values of the segmentation and classification system—the indicators. As described earlier, the preprocessing part (§4.3) can be split into two parts—calculating the indicators solely from the sampled points (§ 4.3.1 on page 52), and calculating the indicators that require supplementary GIS data (§ 4.3.2 on page 53). Although this is a prototype, the algorithms are written to make the method robust for various problems that may arise, and tests involving large datasets shown that the prototype successfully handled all the arose problems. The computational aspects of the developed import-preprocessing procedure in the prototype is given in section 4.4 on page 53.

4.1 IMPORTING THE GIS DATA

In order to do the computations involving geodata, it has to be imported first. As discussed in the preceding chapter, the data from the OpenStreetMap (OSM) project was used.

Daily snapshots of OSM data are freely available on the web, and in this thesis the snapshot of 17 March 2010 was used. Any newer version of the data can be updated at any time.

Basically, the data are organised separately into geometry (in the WGS84 coordinate reference system) and corresponding attributes (tags), in a generic form `key=value`, for instance in case of the building of the OTB Research Institute:

```
amenity=university
name=OTB Research Institute
addr:street=Jaffalaan
addr:number=9
building:use=office
```

An object may contain as much attributes as needed, or it may be even left represented without any tag. As geometry, the tags may be edited by anyone which is the prominent characteristic of the OSM project.

The data was imported in PostGIS with a Python script, however, it was organised in separate tables for better organisation. For instance, there are separate tables for railways, roads, and water surfaces.

4.2 IMPORTING THE TRAJECTORIES

The travel survey conducted by Bohte & Maat (2008) has the GPX files organised in directories for each respondent, where the name of a directory is his/her id (for privacy but also simplicity, names are withheld). Furthermore, the files are organised in subdirectories for each week of survey and contain various files with other information, not relevant to this project (e. g. work address of a respondent). Therefore the algorithm for importing the data scans all subdirectories in the directory where the data are stored for files with the `.gpx` and `.GPX` extensions and starts importing them taking into account the name of the current directory

for determining the user's id. Both extension are taken into account since some GPS loggers produce the data in the capitalised extension. An extract of the structure of the available dataset is presented in Fig. 32. All additional datasets are organised in this fashion.

```
|-- week1
|   '-- logs
|       |-- an1
|       |   |-- 05JA1258.gpx
|       |   |-- 11JA0926.gpx
|       |   |-- 12JA0950.gpx
|       |   |-- 13JA1056.gpx
|       |   |-- 14JA0836.gpx
|       |   |-- 15JA0818.gpx
|       |   |-- 16JA0815.gpx
|       |   |-- 17JA0743.gpx
|       |   |-- 18JA1200.gpx
|       |   '-- 19JA0922.gpx
|       |-- an10
|       |   |-- 05JA1247.gpx
|       |   |-- 11JA0649.gpx
|       |   |-- 12JA0652.gpx
|       |   |-- 13JA1114.gpx
|       |   |-- 14JA0841.gpx
|       |   |-- 15JA0652.gpx
|       |   |-- 16JA0626.gpx
|       |   '-- 17JA0639.gpx
```

Figure 32: An excerpt of the dataset, showing the organisation of the GPX.

I observed that for some reasona a few trajectories were duplicated in different directories (probably due to a simple human error while organising the data), hence before importing the file the algorithm checks if the file was already imported in the database in order to ignore it. This is important to avoid double import and preprocessing, and later possible conflicts of duplicate data. The check is done producing the digest of the content of a file by Secure Hash Algorithm 1 (SHA-1), a 160-bit cryptographic hash function designed by the National Security Agency (NSA):

```
$ sha1sum 16JA0658.gpx
1b2a14db9cc1d0533cf79fee1c55945dd85044bc  16JA0658.gpx
```

Even if the filename was changed for some reason, the duplicate file was successfully detected and marked as a duplicate. The likeliness that two different files are marked as duplicates is called hash collision, and in case of SHA-1 is 2^{63}, which means that for one million files in the database the chance that two have the same digest is 1.08×10^{-13}, hence in this case we can be quite confident with the assumption that each file has its unique fingerprint. The duplicate files are reported in the log of the prototype:

```
   Importing file: data/week1/logs/az19/14JA1014.gpx for user az19
---WARNING: The trip was already imported in the database
(fileid: 12636, user: az41. Imported on 2010-04-08 11:02:25.442277+02).
Skipping import...
```

Apart from duplicates, there is a number of invalid files in the dataset. Some files are empty, while many files have their schema broken (usually due to the device shut down after power cut off). The empty files are easily detected by

calculating the size of the file, and the broken XML schema are detected by detecting the errors in parsing. These files are ignored, and marked as invalid for the user:

```
Entering directory: /data/week14/z1416
   Importing file: /data/week14/z1416/01JL1302.gpx for user z1416
---WARNING: Invalid file (broken XML schema). Skipping...
```

After these checks, if the file was determined as eligible for import, it is parsed, with the `xml.dom.minidom` library. For each point the longitude, latitude, elevation and timestamp are stored. The algorithm also searches for `speed`, `course`, `fix`, `sat`, `hdop`, `vdop`, `pdop` tags in the `gpx` file and stores them in the database. Additionally, the points located in the Netherlands are converted to the Amersfoort/RD New coordinate system for later easier visualisation and possible future applications.

The geometry of the sampled point is inserted as 2D in PostGIS, while the elevation is stored as an attribute, since it is not used as much as the horizontal geometry.

All the points in a file are inserted in the database in one transaction, hence the importing of the files can be stopped at any time, and continued later. By running again the script, the inserted files are not compromised, since the algorithm checks for already imported files with the hash algorithm. After importing, the file is registered as imported by inserting the hash of that file in the table of GPX files. Therefore the importing algorithm can be considered as robust as it handles invalid data, it can be stopped at any time without creating conflicts in the data, and it can be run in multiple occasions without generating duplicate or missing points.

4.3 PREPROCESSING THE POINTS

The indicators, the umbrella term of all statistical descriptors that might contribute in the classification of the transportation mode, are derived in the preprocessing part, after the points are imported in the database. The preprocessing is done for each user separately, in order to analyse all available trajectories. Due to the fact that additional trajectories may be imported in multiple steps, compromising some of the already calculated indicators, the previously preprocessed data are deleted from the database, and the whole trajectory is preprocessed again. This makes the method more robust, and enables the user to append new trajectories in the database, even the ones in the middle of already imported trajectories. Some indicators are not recalculated and are associated with the newly imported data.

During preprocessing, additional checks are made. The two most frequent errors were:

- Duplicate points, and duplicate timestamps. Bugs in the loggers caused identical points to be logged more than once. Such points have the same timestamps which are easily detectable. However, different points may have same timestamps, because of the rounding to the nearest second which was especially frequent in the very short sampling periods. These points are detected as well, and only the first point was preserved:

  ```
  WARNING: The point 24803232 has the same timestamp as
  the point 24803210 --> 2007-05-31 15:48:45
  Check the corresponding GPX file. Skipping point 24803232 ...
  ```

- Some GPX files do not have the points in sequence, hence the algorithm presumes that adjacent points in the log were not adjacent in reality. This

is solved by sorting the points by time in the database prior to any preprocessing.

The following sections describe the calculation of the indicators for each point—indicators that can be derived solely from the movement data, and indicators that use GIS data.

4.3.1 *Indicators derived from the trajectories*

The indicators are calculated in three passes through the points, due to unavailability of data, e. g. accelerations cannot be calculated without the knowledge of the speeds. In the first pass the following is computed for each point:

1. The id of the previous point.
2. Time difference between sorted adjacent points.
3. Distance between the current and previous point.
4. The speed, from the two above values.
5. The order of the point in the file and user's trajectory.

After more data are available, the second pass computes:

1. The "smoothed" speed (used in outlier detection, described in appendix A.1 on page 97), defined as $\overline{v_i} = \frac{1}{3}(v_{i-1} + v_i + v_{i+1})$.
2. Distance to the next point.
3. Time to the next point.
4. The acceleration.
5. The heading to the next point.
6. The heading change rate.

The third pass of preprocessing a point is designated only for calculating the smoothed acceleration (similarly to the smoothed speeds). Each pass involves outlier detection which was not actively used in the prototype and it is described in more details in section A.1 on page 97. Values required for outlier detection are calculated for investigating the outlier removal, but are not removed from the prototype since smoothing the speeds and accelerations does not significantly affect the performance of the process and in future these data may be useful. The same applies to other indicators (e. g. heading change rate) which are calculated for research purposes and are not used due to the reasons stated in § 3.5.2 on page 45. If a user of the prototype prefers not to compute certain indicators, the prototype leaves the option to remove certain calculations for better performance. Since the aim of this research was to investigate all available indicators, including averaging some of them, they have been calculated for all the test data and are left in the database.

4.3.2 *Indicators that require GIS data*

The developed algorithms for matching the points to the infrastructure are not applicable to this problem (Greenfeld, 2002; Taghipour et al., 2008; White et al., 2009), since they are focused on single modes, hence for each point the distance to the nearest infrastructure of a specific transportation mode is calculated. Due to complexity this part is implemented in a separate script, but may be run together with the previously described procedures (both possibilities are implemented).

Basically, each distance was computed with PostGIS, and stored in the database. For instance, the distance to the nearest bus stop for the point with the id 2841292 was calculated with the following Structured Query Language (SQL) query:

```
SELECT min(distance_sphere(
openstreetmap_roadelements.geometry,
trackpoints.geom))
FROM trackpoints, openstreetmap_roadelements
WHERE openstreetmap_roadelements.elementtype = 'bus_stop'
AND trackpoints.pid = 2841292
AND openstreetmap_roadelements.geometry &&
    Expand(GeomFromText('POINT(5.533354 52.027023)', 4326), 0.002)
GROUP BY openstreetmap_roadelements.name;
```

Although the coordinates have been transformed to the local (RD New) system, as it can be seen in the above SQL statement, the distances are computed in WGS84 since several trajectories in the dataset have been acquired abroad.

In order to build a small bounding box around the point for faster searching of matching spatial features in the database, the `Expand` function was used with a threshold of 0.002 degrees, and the point's coordinates have been retrieved from the database before the above referenced query and inserted as an argument in the function. This completes the whole process tremendously faster since the database contains a great amount of matching features.

For each of the point, 13 similar queries are executed. For computing if the point on water, a query checked whether the point intersects with one of polygons representing any water surface where traffic is possible (e. g. ocean, sea, and waterways). In addition, the distance to the nearest polygon was calculated, in order to compensate for GPS errors, i. e. if a point lies below than 10 m from the nearest *water polygon*, then it is assumed on water.

As an example, Figure 33 on the next page shows a scatter plot of all points in a spatial extent with their proximity to the nearest railway. Many points are left blank because of `NULL` values because of using bounding boxes with the `Expand` function, which assigns no values for features outside the bounds of the bounding box.

4.4 OPTIMISATION AND COMPUTATIONAL ASPECTS

Extensive work has been done to improve the efficiency of the presented implementation, to speed up the whole import-preprocessing procedure, since the amount of data and related computations is vast. For instance, for each point 25 SQL queries are made in total, which means that for the whole test dataset almost half of billion of queries should be executed. Each step was benchmarked in order to analyse bottlenecks, and after the optimisations the average import time for a single point is 4.2 ms, thus the whole training dataset was imported in the database in 21 hours.

Preprocessing the data for calculating the non-geographic indicators took on average 60 ms per point, hence the complete dataset was preprocessed for these indicators in 12 days. Due to extensive PostGIS computations, the average time

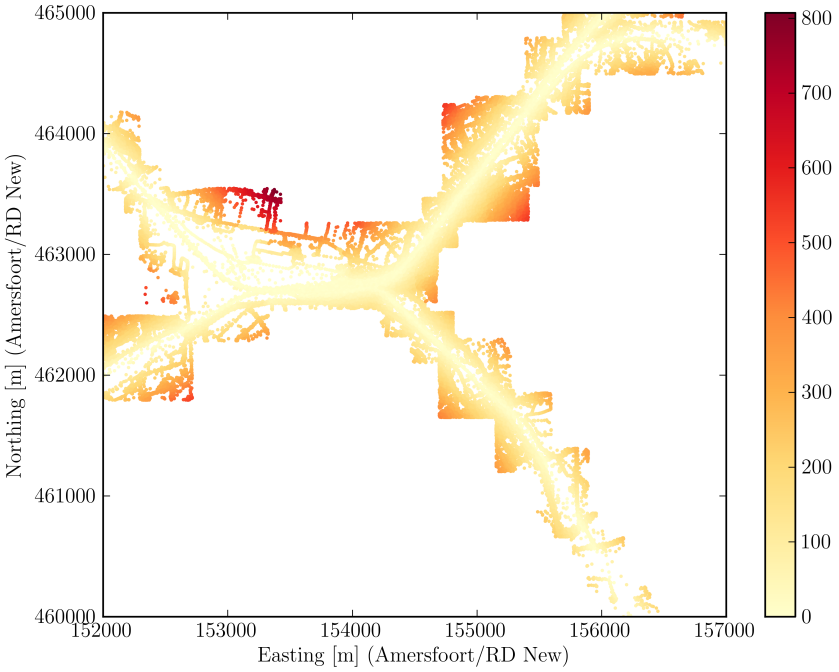


Figure 33: Scatter plot of values of the proximity to the nearest railway (in metres) for all points in the extent (475 495 points found). The location of the railway is visible from low values. The crispness of edges is due to the usage of bounding boxes while searching for the nearest railway feature for each point.

for calculating the proximity for one point to all available network infrastructure features is 0.13 s, i. e. the proximity was calculated for the whole dataset in 27 days of full-time processing. These values may seem high to the reader, however, considering that large-scale travel surveys span over a few months, and the continuously acquired movement data can be preprocessed during the survey, this processing time should not cause waste of time.

By timing the main components of the script, it was interesting to analyse where most of the computer time was spent. Transformation to the local RD/New system was executed in C with PROJ.4, the cartographic projections library, and takes 0.003 ms on average, hence on one million points only 3 seconds are used for executing the conversion of the coordinates from WGS84 to RD New, needed for the secondary options of the prototype, for instance, visualisation of plots for this thesis. Computing, checking and inserting the hash for a single file (track) took 4.5 ms, hence on the whole dataset, 40 seconds (0.0005 %), which is a very reasonable tradeoff to achieve the robustness of the algorithms.

The preprocessing time of 60 ms per point can be broken into the passes. The first pass on average takes 19 ms per point, while the second and third pass take 18 ms and 10 ms, respectively. The remaining 13 ms are spent on computing metadata of each file (e. g. start and end time, distance travelled), which was inserted in the database as well.

The merging of the SQL statements to insert all the points in one transaction is made by merging strings with the data. I have done tests with merging strings with different methods, and the default method in Python (`str + str`) shown the best performance.

XML parsing of the files took most of the importing time. Although there are several Python libraries for processing XML data, usage of different libraries did not show any significant improvement. The code was later converted to C with Cython, but no speed improvement was noticed either. Also, the Psyco module

for optimising the code was loaded, and no significant speed boost was visible, hence the code cannot be significantly improved further, and since the resources are taken mostly by PostgreSQL (querying, inserting, updating) the optimisation of the database was further observed.

Although there is not much that can be done by optimising `SELECT`, `INSERT` and `UPDATE` statements, a few things were done:

- Grouping as much data as possible to insert it in one transaction. For instance, all points in the import process are inserted at once per each file, as already noted.
- PostGIS distances were calculated on a sphere, not spheroid as a geodetic line. Although the spheroid yields more accurate distances there is a processing speed tradeoff (it is roughly 50 % slower). Since the distances are on average just 0.5 % more accurate, there is not much benefit investing more processing time. The Fig. 34 shows the differences in distances between the two commands over distance.

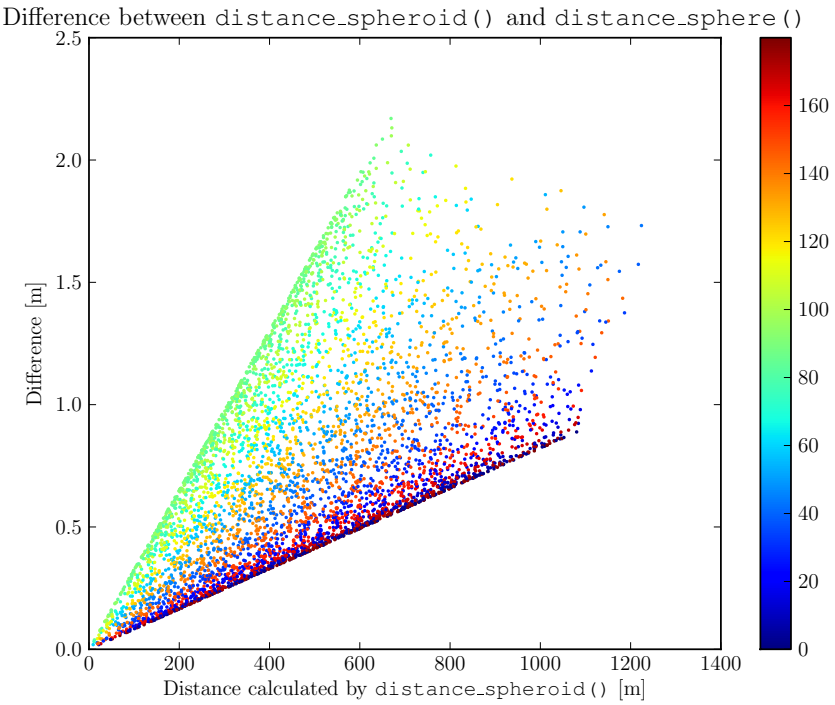


Figure 34: The differences between distances calculated with `distance_sphere()` and `distance_spheroid()` do not exceed 0.5 %, which is acceptable in this project. The colours represent the absolute azimuth between two points (in range from 0° to 180°).

- Spatial indexing and small bounding boxes are used for each query involving spatial data.
- Indices (b-tree) are put also on all attributes which are frequently sorted, beside on primary keys, for example on timestamps when sorting the points by time for the sequences (as it is used in preprocessing for the indicators that are computed from the sampled points).

Each SQL statement was analysed with the `EXPLAIN ANALYSE` prefix command, however, I conclude that not much can be done further and the import

and preprocessing time is reasonable. The PostgreSQL (and PostGIS) queries are reasonably fast, their performance is as expected, and their number cannot be reduced. However, another possibility to further speed up the import-preprocessing process is to reduce the points for which the indicators are calculated, i. e. calculate the statistical indicators for instance every 5th point, rather than for every point. Statistical indicators in frequent sampling periods should not significantly vary between adjacent points, hence for accurate results it not necessary to compute the indicators for each point, especially in trajectories with a higher number of points. The developed prototype leaves this option open for additional optimisation and efficiency. However, the reduced number of indicators will compromise the segmentation algorithm in detecting the stops.

The current total import and preprocessing capacity of the dedicated server is on average half of million points per day, which corresponds to a 37 day trajectory acquired with a sampling period of 6.5 s, or 900 one-hour trajectories at the same sampling period. The computational complexity of the import-preprocessing algorithm for a trajectory is O(n), where n is the number of points in the trajectory.

The size of the database, with the imported and preprocessed trajectories and with the imported OSM is 29 gigabytes. However, almost half of the size is taken by the indices, especially in tables with spatial data. Putting the indices on various columns, such as timestamps, improves the performance in the segmentation and classification algorithm where the points fetched from the database have to be ordered by time.

5 THE SEGMENTATION AND CLASSIFICATION SYSTEM

The concept of the segmentation and classification is given in Chapter 3. This chapter deals with its implementation, practical information and examples, i. e. details concerning the developed experimental software. The classification of gaps in the data is explained in this chapter (§ 5.5 on page 63) since it requires more practical information. The same applies to distinguishing in between transportation modes with similar behaviour of the indicators (§ 5.6 on page 65).

The implementation of the segmentation (conceptually described in § 3.2 on page 24) appeared to be trivial, hence it is not discussed in more details, except for merging adjacent segments with the same classification outcome (section 5.1.1). Practical details about deriving the indicators and the certainty factors are given in §5.2. The supervised learning process is described in section 5.4 on page 60. For that purpose an application for inspecting the behaviour of the indicators for each considered transportation mode is developed (§ 5.3 on page 59).

The computational performance is discussed in § 5.7 on page 69, similarly as in the previous chapter for the import-preprocessing part of the prototype. The user interface of the prototype is presented in section 5.8 on page 70, while the types of outputs of the classification system are described in section 5.9 on page 71.

Moreover, this chapter presents an experimental feature of deriving additional mode-related information (§ 5.10 on page 75) and explains the process of adding new transportation modes to the system (§ 5.11 on page 76) to show its extendability.

5.1 SEGMENTATION OF THE TRAJECTORIES

The segmentations into journeys and single-mode trajectories are already well explained in § 3.2.1 on page 25 and § 3.2.2.4 on page 28, respectively. After the system segments a trajectory into single-journey segments, the segmentation for single-mode segments can fire. Each segment as a list of points is then passed to the classification system for the classification. The following section briefly discusses practical details about merging adjacent segments with the same class.

5.1.1 *Merging adjacent segments with the same classification outcome*

Since trajectories are first segmented for potential single-mode stages, adjacent segments which are classified for the same mode have to be merged in one. This happens often since sensitive thresholds have been put, and basically every stop terminates a segment, for instance, movement in an urban area due to many stops (e. g. traffic lights) may result in many segments.

After merging the two segments with the same classification outcome, the new parent segment is classified, i. e. no joint result from the underlying subsegments is used. This is useful in refining and improving the classification result. Consider an example from the test dataset—a 6 km movement made with car is split into two segments due to a stop in the beginning of the journey. The first segment (0.5 km) is classified for a car with a CF of 0.4, and for a bicycle with a small but still non-zero CF of 0.1. This less certain, but still correct classification is due to the very low speed of a car in the residential zone. The second longer segment (5.5 km) was more representative for a car with its nearly maximum

speed of 63 km/h, but due to the overlap of the used road with a bus line, both modes were returned in the classification result. A CF of 1.0 was assigned to a car, and 0.3 for a bus. The classification of both segments derived equal (and correct) classes, hence the segments are processed for merging and classification of the resulting segment. However, an interesting result is obtained with the refinement and reclassification of the parent segment. This time for the whole trajectory the car scored a CF of 1.0, while both bicycle and bus were eliminated from the classification. This is due to the nearly maximum speed of the whole segment—63 km/h (the parent segment inherited the nearly maximum speed of the second underlying segment) which eliminated the possibility for a bicycle, and since in the first underlying segment there is no bus network, the bus was removed as a possibility, too. Therefore, merging and reclassifying segments improves the results and confidence of the classification.

This behaviour was encountered many times in the classification of the test dataset. Segments which were classified with a lower, but still highest CF have been greatly improved by merging with adjacent segment of the same class and (re)classification. Therefore, the small difference between CFs of multiple results does not represent a problem as long as the class with the highest CF is the correct class.

5.2 DETAILS ABOUT THE INDICATORS AND CERTAINTIES

In the system, there is a total number of 9 indicators—mean speed, mean moving speeds, nearly-maximum speed, five proximity values to the nearest networks (railway, tram lines, roads accessible to cars, bus lines, and underground lines), and the boolean value of the overlap of the trajectory with a water polygon (see § 3.5.1 on page 37). Since 3+8+11 transportation modes are considered (including standing; see Table 3 on page 24 for the rest of the list), there are 198 resulting membership functions (22 for each indicator). Since there is duality in the indicator of the presence of the segment over water (for aircrafts both `True` and `False` possibilities result in the CF of one), a new supplementary indicator `not_on_water` is introduced with corresponding new 22 MFs, increasing the total to 10 indicators and 220 membership function.

This indicator has another aspect that should be taken into account. If a person walks near a water surface (coast or bank of a canal) or crosses a bridge with any land mode, due to GPS errors and structure of OSM data, the segment may be marked as over water (`True`); see Figure 30 on page 45 for an example of crossing a bridge. This problem is solved by calculating the ratio of points which are detected over water with the total number of points in the segment. The threshold was set to 90 % which proved to work well, since segments made with land modes in some cases had most of the points marked as over water (up to 70-80 %) due to their proximity to the nearest water surface. This problem is almost exclusively encountered in the Netherlands. Therefore, the boolean indicator of crossing of the trajectory with a water area is derived from an intermediary float indicator r, the share of points in the segment that are on water:

$$i_{\text{water}}(r) = \begin{cases} \texttt{True}, & \text{if } r \geq 0.9 \\ \texttt{False}, & \text{if } r < 0.9 \end{cases}$$

While segmenting a trajectory, the data about the indicators is collected for the current segment, since it is already used in the segmentation process (e. g. speeds for the detection of stops), and it is saved for the classification process.

As mentioned earlier in § 3.5.1 on page 37 in case of speeds there are two data available: speeds from the GPS receiver, and speeds calculated from the

adjacent timestamped coordinates. If both speeds are available in the database, the system takes the first mentioned with the standard `COALESCE` SQL function which returns the first non-`NULL` value from the list, i. e. if the first field is empty the second is returned.

Before passing the values of the indicators to the classification system, they are sanitised for `NULL` values and other discontinuities, since the fields of the values of the proximities to the nearest infrastructures contain `NULL` due to the small bounding boxes that are used in preprocessing.

Continuing the example with the indicators derived from the speeds, after speeds of all points in the segment are retrieved, they are stored in a one-dimensional array which is passed to the classification (list of points, similar to the concept presented in the beginning of § 3.5 on page 36). As the system operates by taking single float (or boolean) values $i_{1\ldots k}$ and returning single float or boolean values $\mathrm{CF}_m^i = f_m^i(i)$, in case of speeds the two relevant indicators (mean moving speed, and nearly maximum speed) are extracted by filtering the array for zero values in order to remove standing points, after which the mean and 95th percentiles are computed. In case of segments shorter than 15 points, 70th percentile is used since it is less vulnerable to noisy observations. The third indicator related to the speed—mean speed is used for distinguishing between walking and standing, as already described. However, this indicator has MFs which return certainty factors of one for all modes except standing and walking for all input values (see second example in § 3.3.3 on page 32 and Figure 19 on page 33).

A single float or boolean value is determined for each indicator and it is now input in all corresponding membership functions. For instance, the value of the proximity to the nearest tram line is used by 22 different trapezoidal membership functions in order to derive the certainty factor for each mode in each layer for the corresponding indicator. Each layer in the hierarchy of transportation modes contains a separate one-dimensional array with the derived certainty factors from each indicator.

After functions derived a certainty factor for each mode per indicator, corresponding CFs by mode are concatenated in an array, i. e. a list of 10 certainty factors (10 indicators) for each class. For each class, the minimum value is reported (conjunctive reasoning) as its final certainty factor. Modes may be rejected by assigning a CF of zero by at least one indicator, which shown to be very convenient for the performance of the classification.

All classes with the CF of zero are disregarded from the system, and only classes with non-zero CFs are shown. In case all modes in a layer are eliminated, the system reports that the mode is unknown and the classification is not possible. This is rather an infrequent case.

5.3 DEVELOPING THE TRAJECTORY INDICATORS APPLICATION

In the framework of this thesis, an utility application was developed to facilitate the overview and inspection of the statistical indicators for each trajectory and transportation mode in order to establish empirical relationships to a particular transportation mode.

The input of the application are the identification of a user and $[t_{\mathrm{begin}}, t_{\mathrm{end}}]$, i. e. the time interval of the user's (respondent's) trajectory that needs investigation of the behaviour:

```
Please enter the user's id: v17
The range of the track of the user v17 is from
2007-01-05 12:39:05 to 2007-01-18 14:24:26
Specify the range of the track to be plotted. For the
start or end bounds leave empty the corresponding fields.
```

```
Start time (format: YYYY-MM-DD HH:MM:SS): 2007-01-12 23:00:00
End time (format: YYYY-MM-DD HH:MM:SS): 2007-01-12 23:25:00
There are 226 points in the specified range.
```

After searching for all points which are in the interval, the application returns statistical information and plots for each indicator. An example is shown in Figure 35 where the proximity to the road infrastructure for all the points in the specified time interval is given.

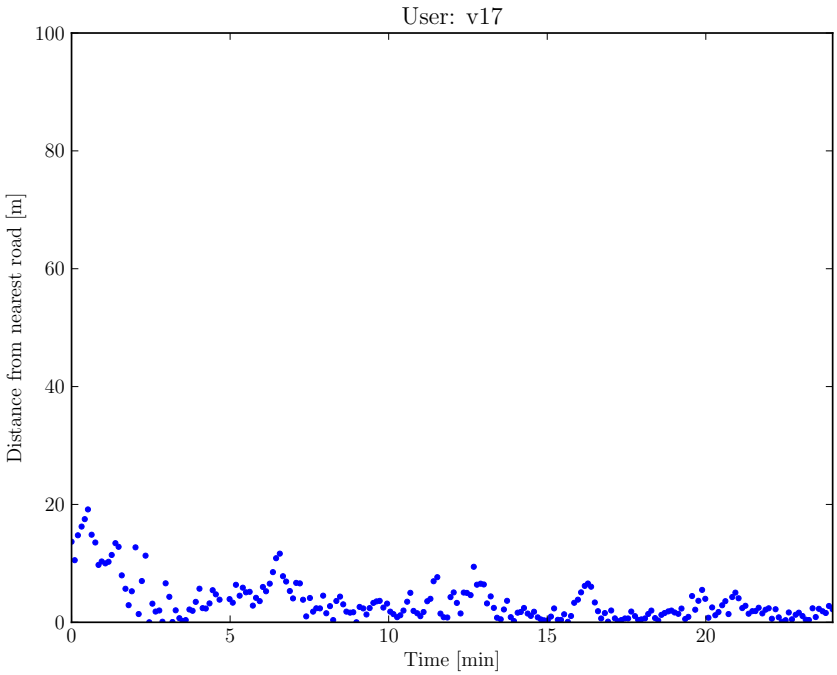


Figure 35: A generated plot from the developed application. Proximity of the trajectory to the road infrastructure and its stability in the given interval serve as a good indicator for cars.

From the plot, it is possible to deduce that the trajectory is adjacent to the road network, which is a strong indicator for a car. However, as stated before, in virtually all cases additional indicators are needed, and specifically in this example still a possibility of a bus, tram and bicycle remains open.

The statistics for the road proximity (values in metres) for this trajectory are given below:

```
-Road prox.-

Mean:       3.56
Median:     2.37
5th perc.:  0.34
95th perc.: 11.39
```

As noted, in the classification system, the mean distance from an infrastructure element is used.

This application, included in the prototype, was indispensable in investigating the behaviour of the indicators for each considered transportation mode.

5.4 TRAINING THE SYSTEM

A subset of the available data was selected as training data in order to investigate empirical relationships and train the classification system. The training

trajectories are diverse, from various people, representing all the considered transportation modes in various situations. The "iterative" process of defining the MFs and various constants, i. e. a manual process of analysing errors in trials of the classification was completed when a satisfying classification result was obtained from the training data.

Since human behaviour while travelling is diverse, and many different situations arise, it was difficult to find an optimal and unique model that would work for all situations and individuals. A homogenous model that would work in all cases is not possible, but finding a model that works best for the training dataset and leaves untypical situations in a minority is possible. For instance, in the MF of the nearly maximum speed for cars, the fourth point where the value (CF) approaches zero (see x_3 in Fig. 18 on page 33) could be safely set to 180 km/h in the vast majority of the cases to correctly reject the possibility of a car in segments with speeds above that threshold. As there might be a trajectory in which the 95th percentile of the speeds is 190 km/h, the classification would not be correct. Therefore, it is not easy to compact all the possible movement behaviour in one model, especially when building a model of a classification system that requires worldwide operation, as I strived for. My test have shown that trajectories made by the same person with the same mode on the same route on different days may have significant discrepancy in the indicators. While modelling the functions, general and frequent behaviour was taken into account, hence such differences may cause inaccuracies in the classification results.

The empirical definition took into account all cases in the training dataset, however, since new situations may emerge in new datasets it is by no means definitive, and it may immediately be concluded that a definitive model may not be even possible. A definitive model would require extensive investigation of travel behaviour for each considered transportation mode in each region covering the trajectories and type of infrastructure (e. g. behaviour of cars in residential zones or highways), and so on, which is outside the scope of this thesis. Nevertheless, the classification proved to work well with a MF of a simple trapezoidal construction, defined by four points (see § 3.3.3 on page 32 for the list of membership functions).

5.4.1 *Storing the domain knowledge*

Each MF, i. e. the definition of four points of the trapezoid is stored in an XML file. The benefit of such approach is that the definition could be easily edited at any time, and new indicators or new transportation modes could be added in the system.

The XML is divided into indicators, whose children nodes are transportation modes divided into layers of the described hierarchy. Its XML schema is given below

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified">
  <xs:element name="definition">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" ref="indicator"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="indicator">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" ref="mode"/>
      </xs:sequence>
```

```
      <xs:attribute name="name" use="required" type="xs:NCName"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="mode">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="values"/>
      </xs:sequence>
      <xs:attribute name="layer" use="required" type="xs:integer"/>
      <xs:attribute name="name" use="required" type="xs:NCName"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="values" type="xs:string"/>
</xs:schema>
```

The file was successfully validated to meet the related standards. An excerpt of the XML file with MFs for the indicator of the maximum speed is given below as an example:

```
<indicator name="maximum_speed">
  <mode layer="1" name="land">
    <values>0,0,150,250</values>
  </mode>
  <mode layer="1" name="sea">
    <values>0,0,30,50</values>
  </mode>
  <mode layer="1" name="air">
    <values>200,400,900,1200</values>
  </mode>
  <mode layer="2" name="stand">
    <values>0,0,4,7</values>
  </mode>
  <mode layer="2" name="walk">
    <values>3,4,12,16</values>
  </mode>
  <mode layer="2" name="bicycle">
    <values>7,10,25,33</values>
  </mode>
  <mode layer="2" name="car">
    <values>25,30,125,160</values>
  </mode>
  ...
```

In order to facilitate the definition of MFs, a few aliases are established, e. g. `all`, indicating that the MF returns the value of one for all the values (see Fig. 19 on page 33). For instance, a car does not use a bus network infrastructure and in any case of the proximity of the segment to the bus line, the CF for a car in that rule is one:

```
  <mode layer="2" name="car">
    <values>all</values>
  </mode>
```

A few membership functions use boolean values, rather than float values, and are defined accordingly instead of four float values as in the trapezoidal MF. For instance, the indicator if the segment is on water:

```
<indicator name="on_water">
  <mode layer="1" name="land">
    <values>false</values>
  </mode>
  <mode layer="1" name="sea">
    <values>true</values>
  </mode>
  <mode layer="1" name="air">
    <values>true</values>
  </mode>
  ...
```

In order to show the standard overlapping of the membership functions in a single indicator, Figure 36 depicts the overlap of several functions for the indicator of the nearly-maximum speed.

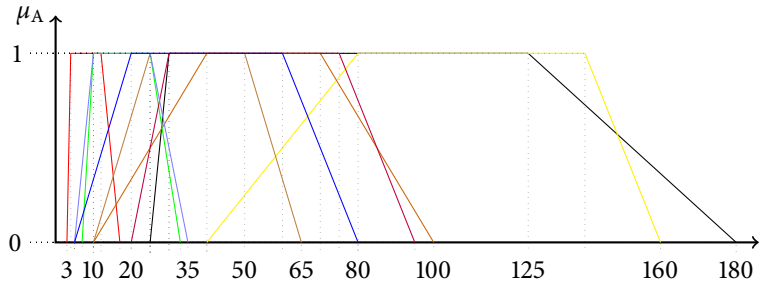


Figure 36: The membership functions usually overlap. This is an example for the membership functions for nine modes used in the indicator of the nearly maximum speed (in km/h). The following modes are plotted: car (black), train (yellow), walk (red), bicycle (green), tram (brown), bus (purple), sailing (light blue), ferry (blue), and underground (dark orange). The classes standing and aircraft are left out for aesthetic reasons.

5.5 DEALING WITH DISRUPTIONS IN THE DATA

Signal shortages which cause disruption in the acquisition of data (i. e. gaps) are frequent, and hard to handle, since these data does not exist, therefore we are dealing with the *classification of non-existing data*. These gaps may not be important to classify, especially if the transportation mode was not changed in between.

As noted, data is marked missing when no samples are recorded in more than 30 s. Moreover, since the signal is not regained instantly, due to TTFF and location in the new transportation mode (e. g. receivers have no or bad reception in aisle seats of trains) the transitions sometimes cause longer spatial shifts and longer gaps in the dataset. The problem is complex since there are numerous different cases. The classification attempts of these cases are explained in this separate section.

Resolving the gaps requires investigating many possible cases that occur in practice, and that is the reason why this part was not discussed in the chapter about the methodology. In addition to these problems, this method takes advantage of gaps, since the underground mode does not have any reception, and it is detected by the disruption of signal in between entrances to the two underground stations, similar to the methods of Stopher et al. (2008) and Shalaby et al. (2006).

The following distinct cases account for most, if not all occurrences of gaps, and their reconstruction was conceptually developed and implemented in the prototype. All cases are depicted in Figure 37 on the following page.

(a) "Regular" gaps, during which the transportation mode was not changed. They are usually short, and caused by entering in tunnels or a random temporary loss of signal (Figure 37a).

(b) The signal was lost during the transition between two modes, and the whole segment in between is not recorded (Figure 37b). This case occurs frequently, especially in trains. For instance, a person cycles to a train station, where the signal is lost, and it is not regained later in the train. The logging resumes after a person makes a transition to a third mode,

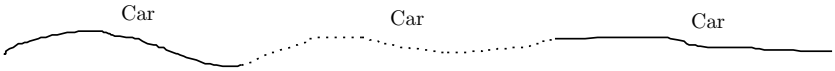
usually walking or again bicycle. Even if the two boundary modes are the same, another mode in between might be detected with the presented technique.

(c) The sampling is interrupted during a segment, and it is resumed at a transition to another mode (Figure 37c). This case is similar to the inverse case where the sampling is interrupted at a transition from one mode to another and resumed afterwards in the segment before another transition occurs.

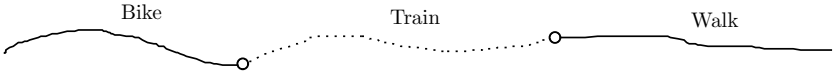
(d) Trajectories with fragments of data, without recorded transitions are frequent in some train stages (Figure 37d).

(e) Two transitions which are not recorded add more challenge to this problem (Figure 37e).

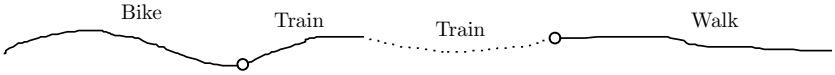
(f) Two transitions in a disruption double the problem (Figure 37f).



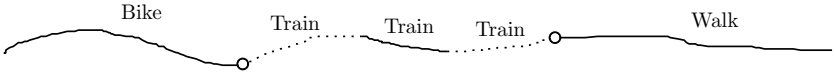
(a) "Regular" gaps where the mode was not changed.



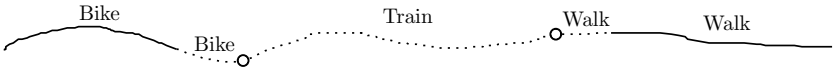
(b) The whole segment done with another mode is not recorded.



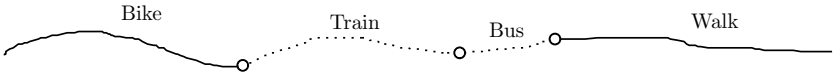
(c) The first transition is recorded, but not the rest of the segment and the second transition.



(d) Neither of the transitions is detected, but a small fragment of the segment is recorded in between.



(e) Signal shortage occurs during both transitions, the mode is not recorded, and the disruptions exist in the earlier and next segments.



(f) Two segments are not recorded, including their transitions.

Figure 37: General cases of data interruption.

Since the points are not sampled, the indicators in these cases are limited, the system is left to *guessing* the situation in the gap and used transportation mode(s). The distance between the two adjacent recorded segments is known, along with the time difference. From these, the average distance may be computed, although this is rather a rough approximation due to the potential sinuosity of the travelled path. As one might suggest, proximity to the stations for certain modes are available for the points on the edge of the gap. Although it is possible to take into account the proximity to the stations, these cases have something more in common—they occur on the network of each corresponding mode.

Hence the location of networks instead of stops are used, which are already available from the preprocessed procedure.

Before each disruption in the data, the system stores the classification result of the preceding segment, and the distance from the last known point to all considered infrastructures. This is also done for the first point after the gap.

The reasoning system starts analysing matching infrastructures from the buffers of the boundary points of the segments (points adjacent to the gap—last recorded point in the previous segment and first in the subsequent segment). These values have been computed in the preprocessing part. If two infrastructures match (e. g. if both points fall in the buffer currently set to 30 m), then the corresponding mode is assigned. This is especially useful for underground modes since it is the only way to classify them. Some underground modes may have short travels on the surface, but always involve data missing time intervals.

In case of the match of multiple infrastructures, the average speed of the gap and the knowledge of the previous transportation mode prevails. If neither condition is met, the system analyses the average speed of the gap, and the average speed of the first 20 points of the next segment. If either speed is higher than 300 km/h, the gap is marked as it is done by air. The value of 20 points was taken in order to preserve the travel behaviour closer to the gap, since the subsequent segment in some cases could be long and exhibit different behaviour (e. g. higher speed) than in its part closer to the previous segment. This value can be easily changed in the prototype.

This solves the presented cases (a), (b), (c), (d) by a unified approach. The method is also useful in segments where due to bad signal reception only fragments of data are available.

Case (e) is resolved only in instances where the time difference and distance to the occurred transition is small. In other cases the segment is marked as unknown as it involves too much ambiguity. The same applies for (f) which is a case that is hard to solve even with human intervention.

Another specific case could not be solved is that if a person lost GPS signal while boarding a ferry, and reappeared after the segment was finished, the in between sea mode could not be resolved since both boundary points fall onto land. Reasoning that the person crossed a water polygon in between requires complex GIS operations. This problem is doubled by the possibility that the person could have crossed a bridge between two land polygons.

Although machine reasoning in signal shortages is complex, it may return satisfying results. Since these data do not exist, it is hard to consider its classification as reliable. However, the results from the prototype are promising, especially in train segments where most of the signal shortages occur.

5.6 DISTINGUISHING MODES WITH SIMILAR BEHAVIOUR

The biggest challenge in the classification of certain modes is to distinguish in between the following sets of modes:

- Standing and walking
- Car, tram, and bus

Each method is described in the following two sections.

5.6.1 *Standing and walking*

To a reader, standing and walking may seem easy to discern. However, standing is rarely concentrated on one position, due to GPS noise and occasional walking

not longer than a few seconds, e. g. moving for a few meters while waiting for a train, which is still considered as standing.

Figure 38 helps in the clarification of this problem. Although the depicted points are all from standing at a station, they are displaced mostly due to noise. Observe the thicker edges of a few points—they mark a stack of multiple points at a same position (two of them are shown with red arrows), hence the displaced points are in the minority. The points in the shape of an arrow are moving points and their directions.

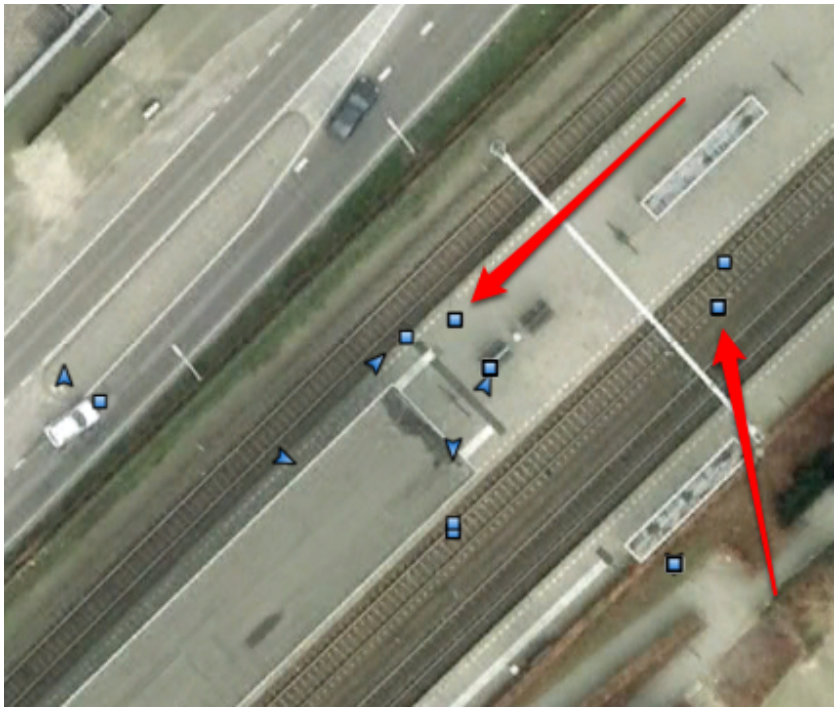


Figure 38: Standing at a station may involve a few moving points due to occasional walking and GPS noise. The two red arrows show examples of the aggregation of multiple points at the same position, while the points with arrows represent moving points and their direction. Imagery is copyrighted by Aerodata International Surveys and Google (2010).

In this case, a special rule is built, using a mean speed, rather than the mean moving speed. In all modes, except these two, the MF value for all the values (mean speeds) is 1. Hence, this individual MF is used only to discern between standing and walking, and the computed CF affects only these two modes.

For instance, if we define the two MF for the indicator of the mean speed with the following values

```
<mode layer="3" name="stand">
  <values>0,0,1,2</values>
</mode>

<mode layer="3" name="walk">
  <values>0,1,8,10</values>
</mode>
```

the classification system is successful in distinguishing the two modes.

In order to have other modes not influenced by this special indicator introduced just for discerning standing and walking, all other MFs for the remaining modes should be defined too, to return CFs of 1 in all the cases, for instance, for bicycle:

```
<mode layer="3" name="bicycle">
  <values>all</values>
</mode>
```

in contrast to the indicator of mean moving speed, in which the mean moving speed for every mode is considered:

```
<mode layer="3" name="bicycle">
  <values>7,10,20,25</values>
</mode>
```

5.6.2 *Car, tram, and bus*

Cars and buses use the same infrastructure—roads, and their speed in urban areas, where most of the bus traffic occurs, is comparable and not notably different. Cars are easily classified on roads without a bus line, i. e. the bus is eliminated as a possibility where a CF of zero is assigned by a MF considering the proximity to the bus network, as it is shown in the following example in Fig. 39. In a segment,

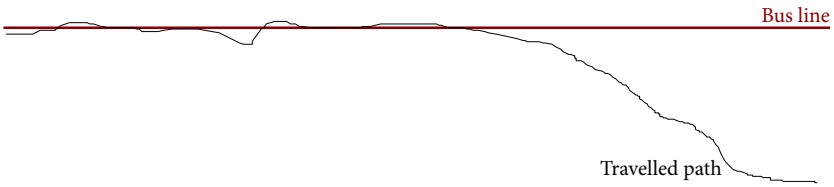


Figure 39: Cars are easily classified in case of the high distance from the nearest bus line.

a car moves along a bus line, however, at one point turns on another road which is not frequented by buses. Although a part of the segment was made on a road included in the bus network, the mean distance to the bus line in the whole segment is high, and a CF of zero is added by a corresponding MF.

Trams have a similar speed behaviour, but use a different infrastructure. However, tram segments are often adjacent to roads, in the range of the standard accuracy of GPS errors, and are sometimes used by bus lines, too. Therefore in segments where bus and/or tram lines exist and completely overlap with the travelled segment, there is ambiguity between the two or three modes, and the derived CFs are comparable.

Here the knowledge of the location of bus and tram stations is used. Basically, if the starting point of a segment falls into the buffer of a bus or tram station, then the possibility of the corresponding transportation mode is augmented in the classification system by *injecting* additional CF values to each corresponding mode. Figure 40 on the next page clarifies this theory. When a person commences a new segment, the prototype calculates the distance to the nearest bus and tram station. If the new segment is started in a buffer of a station, then the corresponding mode gets a CF *boost* in the subsequent segment. The value is currently put to 0.2 since I noticed that virtually all discrepancies between these three modes in the classification are less than 0.2. The size of the buffer was currently set to 20 m which compensates the size of the station and GPS noise.

However, there are various situations that may result in undesirable results. These situations are taken into account. In general, injection of supplementary CFs for one of these two transportation modes could happen in three cases:

- A person walks to a station, and starts a bus or tram trip, i. e. a new segment.
- The bus segment starts after the person stands for some time at a station.

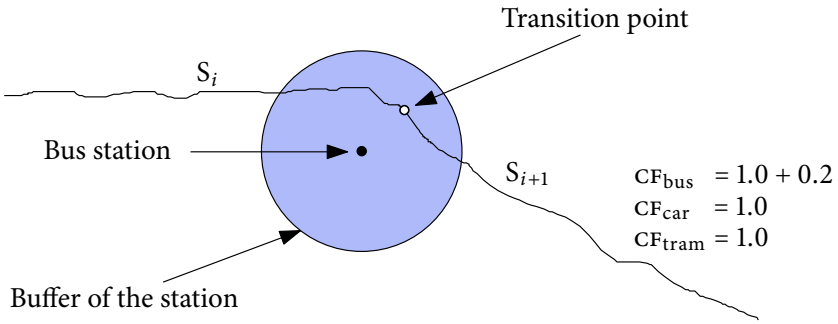


Figure 40: Injecting certainty factors supplement for segments which commence at a station for bus or tram contribute to the distinction of the modes car, bus, and tram.

- A person is already in a bus or tram which stops a station, and a new segment is automatically created because of the stop.

Hence, the knowledge of the transportation mode in the previous segment is retained. The CFs of 0.2 are added only if the previous segment was classified as standing, walking, or—bus or tram, respectively. For instance, if the previous segment was made by a tram, the CF of 0.2 is added for tram only to the derived CF in the upcoming segment commenced near a tram station.

This is also very useful in cases where a person stops with a car for a traffic light which is very close to a bus/tram station. Since the previous segment was made by a car, this time an additional CF for a car is injected to the classification system (Figure 41).

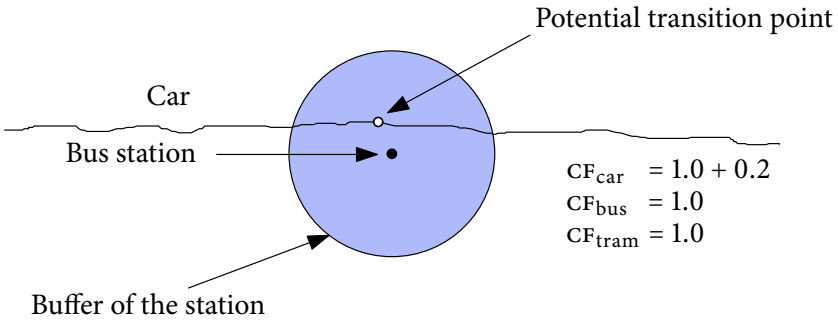


Figure 41: The classification of a segment is influenced by the result of the classification of the previous segment.

This is an important aspect, since in the Netherlands due to the density of the traffic features, stops may be ambiguous, since some of them, e. g. traffic lights, are used by all three modes.

Another important case which is taken into account is a transition from walking to a bicycle which occurs in the buffer of a tram or bus station. Since the previous segment was made by walking, the next segment qualifies for additional CF for either bus or tram. However, the CF of 0.2 is added only when in the next segment an uncertainty between car, tram or bus exist, hence the next segment is correctly classified as made by a bicycle. In case of a bus trajectory which stops for a traffic light, despite the new segment starts outside the buffer of a bus station, again the additional CF value is added to the new segment, making the classification more correct.

The cases considering signal shortage are taken into account as well. In some occurrences, a GPS receiver loses the signal while entering a bus or tram. The

knowledge of the position of the last known point and uncertainty between car, tram and bus of the next segment, automatically assigns a corresponding transportation mode. The same applies if a signal shortage appears in the middle of a bus/tram ride—the knowledge of the transportation mode used before the signal shortage has a great benefit.

Despite these efforts, there are additional, but seldom cases which may degrade the performance of the classification. For instance, if a transition from standing to a tram was made in the buffer of both a tram and bus station, then both modes get the augmented CFs, and the classification relies on the usual indicators (proximity to the network), which still in most cases manage to make a distinction between the two modes, however, in cases where the two infrastructures overlap, classification may be ambiguous and not possible. Another complicated instance is a segment made with a car which stops at a bus station (e. g. a person may be left at a bus station), and if a new segment starts immediately with a bus, but without a standing segment, the new segment may be incorrectly classified as made by a car. The same difficulty is found in a reverse case where a person is picked up with a car at a bus station after a segment made by a bus.

In this section, I have shown that many cases are taken into account, and the distinction between car, bus and tram is successful to a certain level, especially in cases of longer segments where the chance of overlapping of multiple infrastructures is low. However, the last examples show that not all cases may be solved, and may impair the classification result.

5.7 COMPUTATIONAL PERFORMANCE AND OPTIMISATION

The segmentation and classification of the trajectories is much faster in comparison to the import and preprocessing processes. Once the data are stored, here the advantages of using a DBMS as a principal storage method are fully evident. The indicators for each segment, sorted by time, are returned very quickly, due to using a b-tree index on each relevant column, including timestamps.

The computational complexity of the classification of a segment is O(n). Since merging and reclassifying the trajectories is unforeseen and depends on the case, here a unified expression cannot be presented.

The whole segmentation and classification process (with reading and storing the results back in the database) varies for the number of points to be classified. On average, a file with 1800 points (current average per user in the test data—3.3 hours of movement) was segmented and classified in 15 s. Hence, the complete trajectory of a user (on average 7 files—one per day) is usually classified in 1.7 min. The whole test dataset was segmented and classified in 41 hours.

It may be useful to present the results of the segmentation. For the whole test dataset, 245 218 segments between stops were generated. The threshold for segmentation was put to 12 s, as mentioned earlier. The average distance of a segment is 2.2 km, while the duration is 7.4 min. The shortages, which were not included in these calculations, account for additional 35 602 segments. Each track has 26 segments on average, which are later merged in a considerably smaller number of segments.

Further optimisation in terms of speed of the segmentation and classification algorithm would not yield significantly better results, due to the amount of operations and the reasoning chain. The biggest leap would be achieved by running the algorithm directly in the database, or a migration to a faster compiled programming language such as C++.

5.8 USER INTERACTION WITH THE SEGMENTATION AND CLASSIFICATION SYSTEM

Although the process of segmentation and classification is fully automated, there are options for the interaction with the user, for instance:

- The `-a` flag segments and classifies all the unclassified trajectories in the database.
- With the `-u` flag, the system classifies only trajectories of a specified user.
- In conjunction with the user flag, the flags `-s` and `-e` define the timeframes of the trajectories to be segmented and classified.
- The flag `-r` reports the following information: results from both segmentations (timeframes for journeys and for single-mode segments), the values of all calculated indicators for each segment, and the classification outcome for each segment. If multiple segments have to be merged and reclassified this is reported as well.
- If the user of the prototype changes a value in the definition of the membership functions or thresholds, the prototype may be forced (`-f` flag) to rerun the classification of an already classified trajectory.

An example is shown below, but since the whole reporting is too long to insert it in the example, only the report of the segmentation for the journeys and the classification outcome of an arbitrary segment are shown. For each classified complete user's trajectories, no matter of the status of the `-r` flag, the program prints the progress, in this case only for a single user.

```
$ python classification.py -u an25 -s '2007-01-30 06:00:00'
-e '2007-01-30 20:00:00' -r

1  Segmentation and classification for user an25

Number of trips in the track: 2
1 2007-01-30 06:50:09 - 2007-01-30 07:52:25
2 2007-01-30 16:29:41 - 2007-01-30 16:36:33
...
```

The classification outcome for one segment in the trajectory is shown below. The transportation modes with CF of zero are not shown for aesthetic reasons. In each layer of the hierarchy, the CFs are normalised for easier analysis of the relation between certainties of the derived classes.

```
CLASSIFICATION

--- First layer
Land : 1.0  --> 100.0%

--- Second layer
Car/tram/bus : 1.0  --> 100.0%

--- Third layer
Bus : 0.3  --> 24.85%
Car : 1.0  --> 75.15%
```

When passing the flag `-a`, the program first searches for all unsegmented and unclassified trajectories in the database, and passes them to the segmentation-classification system.

```
$ python classification.py -a
Number of users in the database: 1289

1  Segmentation and classification for user an25
2  Segmentation and classification for user v201
...
```

As with the import-preprocessing application, this process may be stopped at any time and resumed later. Classified segments are not reclassified after restarting the process, unless the `-f` flag is passed, which is convenient if the user prefers to test the classification results for different values in the membership functions or thresholds.

5.9 GENERATING THE OUTPUT OF THE CLASSIFICATION SYSTEM

Apart from the presented `-r` flag, the system has three other forms of output of the classified data:

- Timeframes of a segment $[t_{\text{begin}}, t_{\text{end}}]$ with the classification outcome.
- Transportation mode for each point in the database (useful for visualisation purposes).
- A Keyhole Markup Language (KML) file, suitable for visualisation and integration in a GIS environment.

described in the following sections.

5.9.1 *Descriptive data*

Each classified segment was stored in the database with its corresponding mode, but also basic information (duration of the segment, distance travelled, its first and last points). Shortages are also included in the database with their attempted classification. An example of querying the database for the classified segments in a specific time interval for a specific user is given below.

```
thesis=# SELECT starttime, endtime, mode3, remark FROM segments
WHERE userid = 'an713'
AND starttime >= '2007-04-03 16:10'
AND starttime <= '2007-04-03 16:40'
ORDER BY starttime;

      starttime      |       endtime       |  mode3  |  remark
---------------------+---------------------+---------+----------
 2007-04-03 16:12:21 | 2007-04-03 16:12:59 | Stand   |
 2007-04-03 16:12:59 | 2007-04-03 16:13:31 | Unknown | Shortage
 2007-04-03 16:13:31 | 2007-04-03 16:31:15 | Car     |
 2007-04-03 16:31:22 | 2007-04-03 16:36:51 | Car     |
 2007-04-03 16:36:57 | 2007-04-03 16:39:04 | Car     |
 2007-04-03 16:39:10 | 2007-04-03 16:39:43 | Walk    |
 2007-04-03 16:39:49 | 2007-04-03 16:39:55 | Stand   |
(7 rows)
```

A similar table was created for storing information about journeys, and the data may be listed (a related output is shown on p. 70).

In the table for points, each points was assigned with its classified transportation mode, for instance:

```
thesis=# SELECT time, speed, mode1, mode2, mode3 FROM trackpoints
WHERE userid = 'an713'
AND time >= '2007-04-03 16:32'
```

```
AND time <= '2007-04-03 16:33'
ORDER BY time;

        time         |   speed   | mode1 |    mode2     | mode3
---------------------+-----------+-------+--------------+-------
 2007-04-03 16:32:00 | 37.620197 | Land  | Car/tram/bus | Car
 2007-04-03 16:32:06 |  58.16003 | Land  | Car/tram/bus | Car
 2007-04-03 16:32:13 | 70.443092 | Land  | Car/tram/bus | Car
 2007-04-03 16:32:19 | 75.805954 | Land  | Car/tram/bus | Car
 2007-04-03 16:32:25 | 72.396904 | Land  | Car/tram/bus | Car
 2007-04-03 16:32:32 | 67.707054 | Land  | Car/tram/bus | Car
 2007-04-03 16:32:38 | 74.726105 | Land  | Car/tram/bus | Car
 2007-04-03 16:32:44 | 73.014954 | Land  | Car/tram/bus | Car
 2007-04-03 16:32:51 | 68.518707 | Land  | Car/tram/bus | Car
 2007-04-03 16:32:57 |  69.47834 | Land  | Car/tram/bus | Car
(10 rows)
```

These two presented outputs are beneficial for extending the system for visualisation purposes, better travel diaries with routes, generating maps, and so on. The classified data can be read from the database and used in another system for other products or purposes. An example for this benefit is shown in the next section, where these data are used to generate a KML file.

5.9.2 *Generating a KML file*

An output format which should be more interesting for travel behaviour researchers is KML, which is an XML schema for geodata, primarily used in Google Earth, and an open standard by Open Geospatial Consortium (OGC). KML supports different styles for each feature, which is useful for representing segments made of different transportation modes with different colours. For this purpose, a utility to generate KMLs from a database was programmed. Its usage is straightforward and follows the convention presented in the main program for the segmentation and classification. For instance, the command:

```
$ python kmlbuilder.py -a
```

generates KML files for all the users in the database (one for each), while

```
$ python kmlbuilder.py -u z709 -s '2007-04-04 08:00:00'
-e '2007-04-04 21:00:00'
```

generates a KML file (in this case `classified_track_z709.kml`) for a specific user in a specific time interval. The flag `-a` could be combined with the flags `-s` and `-e` for restricting the movement in a time interval, making it possible to generate the movement of all available respondents in a particular time interval, which might be useful for some applications, e. g. distribution of travel on a public holiday.

The different colours for each transportation mode are generated by the program with the following convention:

```
<Style id="Walk">
  <LineStyle>
    <color>FF0000FF</color>
    <width>5</width>
  </LineStyle>
</Style>
```

Where the identifier of the style for each transportation mode is the name of the mode itself. The colours are defined as `aabbggrr`, where `aa` is alpha (transparency), and `bbggrr` are the colours blue-green-red in hexadecimal form

(in alphabet, rather than the usual red-green-blue). This makes it easy to add additional transportation modes in the future.

The features (linestrings representing each segment) are automatically generated in the following form

```
<Placemark>
  <name>Walk</name>
  <description>
     A classified stage.
     User: az37
     Segment:
     2007-01-23 07:28:57 -
     2007-01-23 07:29:16
     Duration: 0.21 km.
     Duration: 0.3 min.
  </description>
<styleUrl>Walk</styleUrl>
  <LineString>
     <altitudeMode>absolute</altitudeMode>
       <coordinates>
         5.390341,52.153428
         5.391164,52.152935
         5.392414,52.153823
         5.394007,52.156099
       </coordinates>
  </LineString>
</Placemark>
```

Note that the `<styleUrl>` tag refers to the predefined style for each particular transportation mode, passing the information to an application in order to print each transportation mode in different colours for an easy overview. In addition, each segment shows its basic information (user, start and end time, duration and distance). Figure 42 shows an example of a generated KML file for a respondent layered over a satellite image in Google Earth, with the shown information for one segment. The transition between two different modes is clearly visible, as the different style for each transportation mode. Additionally, the example shows that very short segments, such as this one—walking from a building to a bicycle for just 10 m, are correctly segmented and classified.

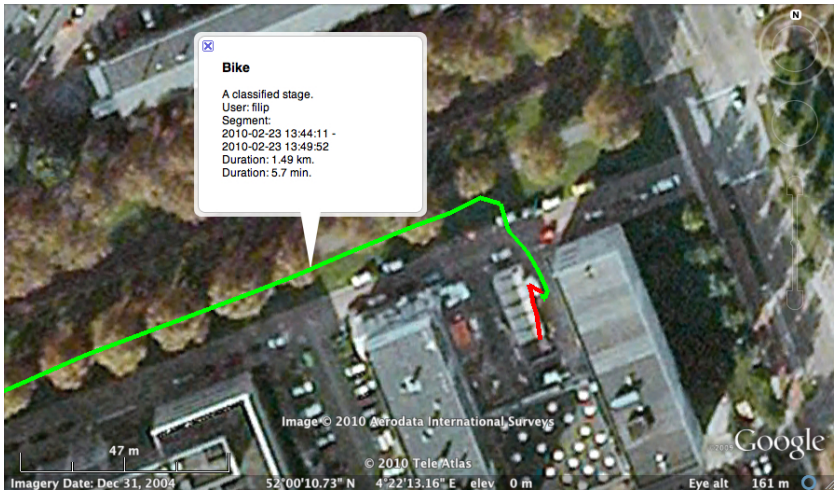


Figure 42: A part of the generated KML visualised in Google Earth over a satellite image. The transition between two modes (walk and bicycle) is visible, as the information for the second segment. This example also shows that the segmentation and classification of very short segments may be successful. Imagery is copyrighted by Aerodata International Surveys, and Google (2010).

In case the transportation mode is unknown, a thin white line is shown, as it is in signal shortages, connecting the two points shared with the previous and the following segments since no points are available. However, the latter has the result of its attempted information in the description, and an example is shown in Figure 43. This journey was done by a train (the classified data is shown in yellow), and contains many disruptions, hence this specific trajectory may be considered as fragments of movement. Nevertheless, the shortages were correctly classified as train on the basis of available data.

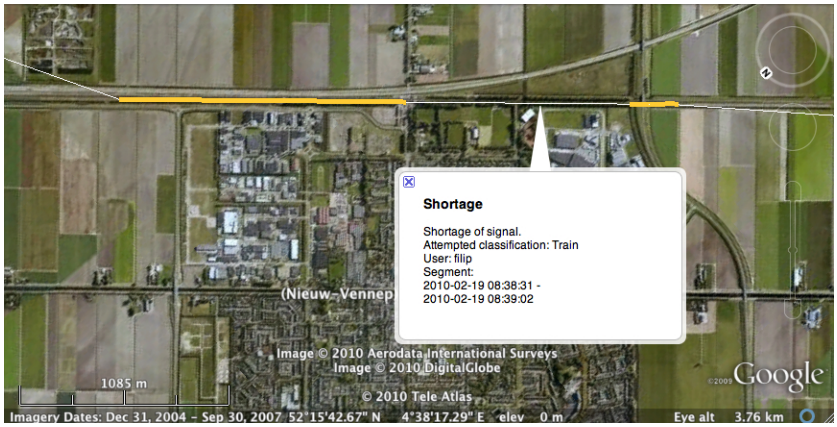


Figure 43: Segments with signal shortages are shown in thin white lines connected with the boundary points of the adjacent segments, with the result of the attempted classification. Imagery is copyrighted by Aerodata International Surveys, DigitalGlobe, and Google (2010).

Since the generated KML file contains a list of segments as distinct features, it is easy to browse through the segments and remove a particular segment if needed. Figure 44 shows the screenshot of the list of classified segments in a KML file in Google Earth. Google Earth, as the *de facto* default application for viewing KML files is the optimal application for using it in conjunction with this output since it contains worldwide satellite and aerial imagery and it is easy to use. However, if needed, KML could be used in other GIS software or converted to other formats since it is widely supported and an OGC standard.

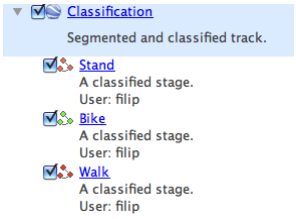


Figure 44: A generated KML file consists of different classified segments.

The generation of the KML is fast (computational complexity is O(n)) and the performance is 5 ms per point—usually a few seconds for a trajectory. Naturally, the size of a generated KML highly depends on the number of contained points, e. g. a trajectory with 1000 points takes less than 30 kilobytes. Thus, the generation of KMLs for all the classified trajectories in the database, considering the taken time and space is not consuming and advisable for better overview of the segmented and classified data.

Figure 45 shows the visualisation of some of the generated KMLs on the area of the Netherlands. The big amount of disruptions in the data due to signal shortages is visible (straight white lines).

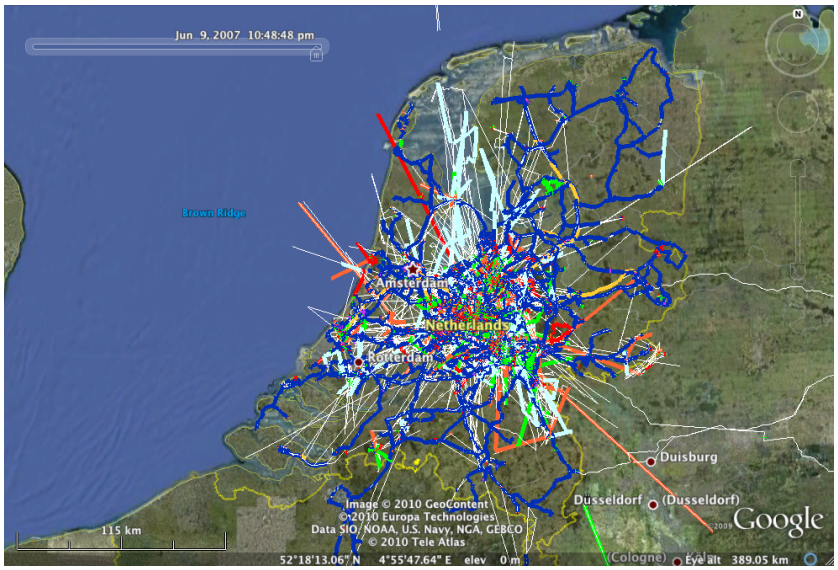


Figure 45: Visualising classified trajectories in the Netherlands with the generated KML files. Imagery © GeoContent and Google (2010).

5.10 DERIVING ADDITIONAL MODE-RELATED INFORMATION

During the course of this thesis, as data and first results became available, I have noticed that various additional information may be appended to the classification result, for instance:

- Names of departure and arrival stations in a journey (train, tram, bus, and underground).
- Line designations (bus/tram line number and name of the transportation company).
- Train type (e. g. the train classification in the Netherlands: *sneltrein, stoptrein, intercity*).
- Airlines, flight designations, and aircraft type.

Additionally, for standing segments, the reason for standing could be deduced from the proximity to certain features found in the OSM data.

Although this part was not an original objective of this thesis, it was conceptually investigated and implemented to an extent with satisfying results. This feature may not be of interest in transport behaviour research, but it is interesting for possible future applications.

In this section, a few examples of the results are given. All the data are derived from the OSM data, except data for airline routes which is obtained from the application Airline Route Mapper. The data are in the public domain, hence it could be used for this purpose.

Segments operated by trains report:

```
Segment done by train.
Arrival station: Den Haag HS
```

In case of an aircraft segment, the system detects the departure and arrival airports, and the airline companies which operate the route:

```
The previous segment was done by an aircraft.
Departure:  Copenhagen, Denmark  (CPH)
Arrival:    Amsterdam, Netherlands  (AMS)
Carrier(s): Scandinavian Airlines System
            KLM Royal Dutch Airlines
```

This even works with signal shortages—e. g. if the whole segment of the aircraft or train is missing (not a single point of the segment is recorded), by analysing available data, the basic information of the complete missing segment is recovered and enriched with additional information:

```
Segment done by train.
Departure station: Nijkerk
Arrival station: Amersfoort
```

The database of flight schedules required to retrieve the flight number is not available due to licensing issues and high price.

The reason for standing is deduced from the proximity of the nearest bus, tram, train, and underground stations or airport, and the knowledge of the next transportation mode:

```
Possible reason for standing:
Train station: Delft
```

In case of long standing in between car segments, the system reports that the cause of the standing is a traffic light or traffic jam.

This feature, although in its experimental stage, demonstrated its performance and potential for future applications. Because of promising results it is left enabled in the prototype. This feature does not significantly affect the performance of the system.

5.11 ADDING NEW TRANSPORTATION MODES AND NEW INDICATORS

In this section it is shown that the developed system is extendable in the way that new transportation modes or indicators could be added. In some applications, new transportation modes may be important and should be included in the system. Analysing the preliminary results, I concluded that virtually all modes that appear in the dataset are already covered in the list of transportation modes. Thus, it was not easy to make a selection of a mode which could be additionally added in the system to show its extendability. In some journeys made by an aircraft I noticed that the bus ride from the aircraft to the airport gate (and vice-versa) is as expected always incorrectly classified. This happens because such *special* bus lines are not included in the OSM data, and these segments are automatically discarded as bus (the CF in the MF for bus proximity was determined as zero) since it is located far away from any bus network in all cases. Although this mode is not frequent in the dataset and may not be relevant to any application, it is the right candidate to insert it in the system as an example. In such segments the classification system usually attributes the mode to a bicycle, since bicycles are not constrained to any infrastructure and their speed behaviour is lower but still comparable to the one of such busses:

```
CLASSIFICATION

--- Third layer
Bicycle : 0.6 --> 100.0%
```

The first step is to select a few segments as training data, in order to investigate and model the behaviour of the mode through various indicators—speed, acceleration, proximity to a certain infrastructure, and think of new indicators that might help in this case:

- To eliminate conflicts with other modes with similar characteristics.
- Contribute the attribution of this mode.
- Elimination of this mode where it does not operate, so it does not affect the classification results of the existing system.

It is important to investigate the overlapping of the new transportation mode with existing modes in the database. For instance, such bus has an overlap with the busses used in public transportation considering the speed range. However, this does not pose problems since the modes are exclusive—the airport bus never uses existing bus lines, while the *normal* bus is constrained to it, hence when considering the proximity to the bus network as an indicator, these modes are exclusive and one of these is always eliminated.

A set of new membership functions for the existing indicators should be added in the system, for instance for the nearly maximum speed

```
<mode layer="3" name="airport_bus">
  <values>5,10,40,50</values>
</mode>
```

and the proximity to the nearest bus line

```
<mode layer="3" name="airport_bus">
  <values>30,50,inf</values>
</mode>
```

which means that if the distance to the nearest bus line is below 30 metres, the mode is rejected. Here the threshold might be put too low since the two infrastructures seldom approach each other that close, but it is left to cover all the possibilities. To make this example simple, the new mode is added only to the third layer in the hierarchy of transportation modes in the system.

After investigating the behaviour of that specific mode, I realised that a new indicator to the system could be added. Airport busses (obviously) operate at airports, which are available in OSM data, therefore the indicator of the proximity to the nearest airport is introduced. Since this indicator is considered in all other modes, new MFs for each existing (and the new) modes should be defined:

```
<indicator name="airport_prox">

  <mode layer="3" name="airport_bus">
    <values>0,0,1200,1600</values>
  </mode>
```

The values are in metres. Since cars do not depend on this infrastructure, the MF returns a CF of one in any case:

```
  <mode layer="3" name="car">
    <values>all</values>
  </mode>
```

This is done for all other existing modes in the system. In the second step, the expert system should be guided to lookup for this new indicator from now on. This is done by adding a new array which contains the values of certainty factors determined by a single indicator for each transportation mode, in this case the indicator `airport_prox`. Running the classification system for the same segment, now we get the correct results:

```
CLASSIFICATION

--- Third layer
Bicycle : 0.6 --> 37.5%
Airport bus : 1.0 --> 62.5%
```

If the user of the system prefers to generate KML files, a new style for the new mode should be created. This is simply done by appending the following XML nodes. The colour and line width can be easily defined to distinguish the new mode from present modes in the system:

```
<Style id="Airport bus">
  <LineStyle>
    <color>FFFFFFCC</color>
    <width>5</width>
  </LineStyle>
</Style>
```

This example presented the extendability of the method, and the straightforward workflow for adding new transportation modes in the system, and new indicators.

Although the prototype is successful in the classification of this newly introduced transportation mode, it is not counted as an additional class in the conclusion and total count of the considered transportation modes since it does not occur frequently.

6 EXPERIMENTS AND VALIDATION

6.1 INTRODUCTION

After the training of the segmentation and classification system was finished, the whole test dataset was segmented and classified. Since the developed method and prototype should be checked for accuracy with a random subset, several experiments were done to validate and assess the described work. The test dataset was already segmented and classified by another technique as well (see § 1.4.4 on page 8), and these data was used in the validation. However, these data cannot be used in full extent to validate the whole test dataset for the following reasons:

- The segmentation had not been done precisely, at the same level of detail as in this work where even short segments are marked and where explicitly points are marked as transition points. The available segmented trajectories of the test data used for validating this method have rather a rough timeframe of the segments, while shorter segments are omitted.
- The hierarchy of modes is different, e. g. tram, underground and bus represent one transportation mode, and does not go in a detailed layer with separate modes.
- Very short segments are not classified. Because of oversegmentation in the presented project, the segmented and classified dataset has many short segments.

Hence, as *ground truth* I have considered the mentioned validated data in the database and checked it further manually. In the validation dataset, some very short segments are missing, hence they are classified manually.

In the experiments, a random subset of 202 segments was chosen for experiments. The segmentation and classification results from the prototype have been manually compared to the validated (reference) dataset.

The calculated classification accuracy depends on the number of samples that are correctly classified, i. e. $\frac{t}{n}$, where t is the number of sample cases correctly classified, and n is the total number of sample cases.

However, rather than just reporting a single accuracy figure, the reader might be interested in the following aspects as well:

- What is the accuracy per each layer in the hierarchy of modes?
- What is the accuracy in case of flawless data, and the accuracy in case of noisy GPS data and very short segments?
- Does the system rather report that the classification is unknown/not possible or assigns a wrong transportation mode?
- What is the performance of the segmentation and do some transitions go undetected?
- Analysis and list of errors.

Although the subset was randomly selected to eliminate bias and have representative trajectories, for better accuracy overview I added a few trajectories with some specific situations which are not frequent in every day life. For instance,

transition from a car to a bus, and then train. In addition, a few datasets from outside the Netherlands (Denmark, Norway, Belgium, Germany, and Croatia) have been added to test the performance of the method for usage with movement data acquired abroad.

6.2 RESULTS OF THE VALIDATION

To separate segments which may have a negative impact on the classification system due to various difficulties such as very few observations and high GPS noise, the subset have been split simply into *good* and *bad* data. As mentioned earlier in the thesis I considered a segment short when the number of points is below 15. Noisy GPS segments are stages where the classification result can be completely influenced by these deviations. Since both cases are problematic, they are considered together in the latter category. On the other hand, *good* data are segments which do not have the two previously mentioned properties, however, they also contains GPS errors and noise, but to a smaller and *normal* extent. Hence, this subset is still not *perfect* as a reader might have the impression, and its classification in these cases may be hard as well.

Table 4 shows the classification accuracy for the mentioned subsets by each layer of transportation modes, and joint figures for the whole subset used in the experiments.

Table 4: Accuracy of the developed classification system (experiments using a random validation subset). The ratio of correct classifications is expressed in percents.

Quality of input data	Layers		
	1	**2**	**3**
Good GPS data	99.1%	94.5%	93.6%
Bad GPS data	99.0%	91.4%	89.2%
Total (all data)	99.0%	93.1%	91.6%

The share of good and bad input data was 54% and 46%, respectively. The high amount of the latter subset is mostly caused by small segments caused by oversegmenting the trajectories (splitting the trajectory at each stop at a very sensitive threshold). However, the accuracy for short and noisy segments is comparable to the accuracy of segments of data higher quality. These results show confidence in oversegmentation of the trajectories. The method for segmentation shown very good results. Oversegmentation of trajectories is evident, and transitions seldom pass undetected (in less than a few percents of the cases). These errors are due to the combination of bad sampling periods and fast transitions. Fortunately, these cases are not frequent.

As expected, the accuracy of the system drops with with respect to the layers, i. e. increased number of classes.

Cars on longer journeys are classified most correctly—at a 100 % rate. This is due to a high number of observations, and a high amount of data enough to reject all other transportation modes. Even in this large validation subset, not a single error was detected in the classification for cars in segments longer than a few kilometres. Since car is the most frequently used mode in the test dataset and in overall according to the Dutch National Travel Survey (Ministerie van Verkeer en Waterstaat, 2009a), this result is assuring. Very short segments, such as walking from a building to a car (usually less than 40 m and 10 points) are

classified correctly in most of the cases, although such segments are irrelevant for travel behaviour research, and may be even discarded, but may serve as a good proof of the precision of the segmentation and classification of very short stages.

An example of the performance for classification of very short segments is shown in Figure 46.

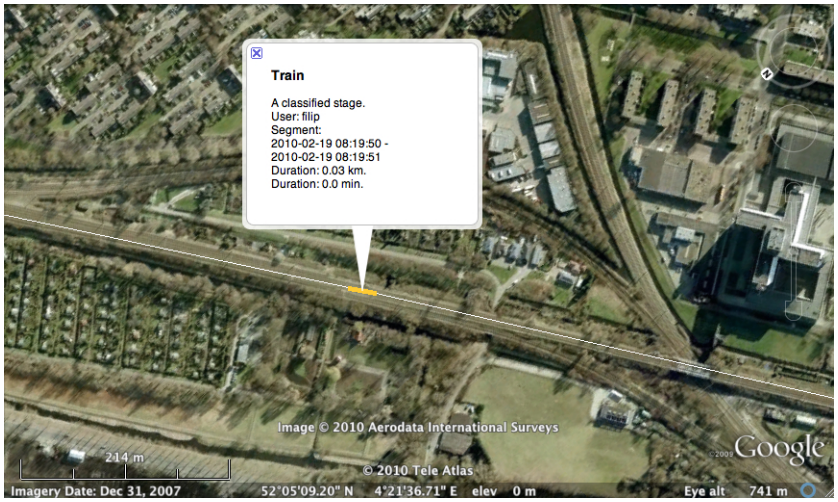


Figure 46: A segment with two points of length of 2 s (thick yellow line), rather a small fragment of movement, is correctly classified, as the adjacent gaps (shown in white thin lines). Imagery © Aerodata International Surveys and Google (2010).

Trajectories acquired outside the Netherlands have been classified at the same accuracy as the rest of the data. No deviation in the accuracy was noticed in regard to such data. In addition, the experimental feature of deriving additional mode-oriented information (e. g. name of departure and arrival train stations) derived mostly correct results as well.

The cause of the 8.4 % of total errors was analysed and it is discussed in the next section in details. Twenty-five percents of this cases were marked as uncertain (unknown class), i. e. a classification is not possible since all modes were rejected (no non-zero CF was present), while the rest was incorrectly attributed as another mode, mostly a mode with similar behaviour and comparable membership functions (e. g. a car instead of a bus).

6.3 ANALYSING THE ERRORS

In this section the causes of all the encountered errors are described. Most of the errors are caused due to noisy data, lack of or bad GIS data and specific situations which could not be easily foreseen, or their modelling would be too complicated and would compromise the existing methodology and results. As noted, upon turning on a GPS receiver usually the first few fixes are very noisy and cause positional jumps in hundreds of meters. In the presence of a large number of fixes, the noisy data are normally compensated with the techniques presented in this thesis, e. g. using the nearly maximum speed with the 95th percentiles, and does not pose problems. However, in very short segments of less than 10 m, where only two or three points are available, the errors in the classification may be notable, and it may not possible to build a reliable method of classification since the data are of very poor quality and even a human intervention is not beneficial. Bus stages which commenced near a bus station, but due to high GPS

errors which caused points to fall outside the buffer of a train station, resulted in wrong classification as a car. Short standing, but with a majority of noisy points resulted in the wrong classification as a bicycle.

As expected, a few errors have been caused by the inability to correctly discern between a car and bus. Here the benefit of using a joint category in the second layer shows its strength, and the two modes have always been correctly classified when combined together. One error was made in a bus segment where the nearly maximum speed was above 90 km/h. Since that is not a typical speed for busses, a possibility for a bus was eliminated and the segment was classified as made with a car. A few very slow (below 20 km/h) car segments in school or residential zones were classified as a bicycle, and running to a bus station to catch a departing bus was classified as cycling. However, the transition in between was successfully determined.

Since the speed is the only indicator to discern between walking and cycling, very slow cycling (less than 7 km/h) in one case was incorrectly classified as walking. In a sailing segment, probably due to the lack of wind, there were a few motionless segments on sea marked as standing. Technically this is correct—the person was standing at the same position for some time, but it is hard to choose whether this case belong to sailing or rather standing on sea. The same happened when a boat was anchored near an island for a few hours. Although these are not errors from the classification system, these cases are mentioned in order to show that there exist fuzzy situations between two classes, and for some situations it is hard to determine to which mode and status belong. The same applies to the interesting observation that when a person is in a car and waits to pick up someone, the system detects that timeframe as standing.

Coastlines in OSM are due to the compression not precisely stored (the features are simplified), hence walking and cycling near sea or ocean may cause wrong attribution to sailing, since the whole trajectory is marked as made on water. One such case was encountered on the shore of one of the West Frisian islands.

Some walking stages were classified as standing. This is an unclear situation due to the fuzziness in between the terms standing and walking, and no clear distinction between the two. For instance, a tourist in a city or a person in a park while walking with a dog may move very slowly or may stop several times, but these stops may still be very short and below the threshold to be considered it as a stop. This is caused by the definition of the *mode* standing, and the user of the system should set his/her acceptable thresholds for distinguishing the two. Walking is one of the modes which are most correctly classified in the database, but untypical behaviour on its lower edge combined with frequent standing complicates the situation. This should not result in significant problems since the adjacent segments are usually correctly classified as walking, and by visualising the segmented and classified KML in Google Earth the result is clearly evident.

It can be argued that these specific situations should not be considered as a flaw of the method and the developed prototype. Another inconvenience is the state of GIS data and changes in the infrastructure over time, such as removal or displacement of tram and bus stations, or abolishment of services. Since most of the test data is from the beginning of 2007, and the GIS data represents the state in the beginning of 2010, one segment made with a tram was incorrectly classified due to that discrepancy. The segment started at a tram station which was removed in the meantime, and it is not present in the current OSM data snapshot. Therefore the injection of a supplementary certainty factor for a tram was missed, and the segment was classified as made with a car. For future work, it is advised storing different versions of GIS data. However, that may be computationally very expensive and demanding due to the frequent update rate of OSM data. This disadvantage can be partially resolved by obtaining the OSM snapshot of a

particular time period, when most of the trajectories that have to be classified have been acquired.

Most of the mentioned situations cannot be easily modelled since that would degrade the present results and involve uncertainty in other already successfully classified stages. For instance, if the threshold for car speed is put lower in order to make a distinction between bicycle and car in residential and school zones, ambiguity between bicycle and car at these speeds would arise in many more other segments. Hence, it can be concluded that by modelling and correctly classifying these specific cases, the accuracy of the system would be impaired by an inaccurate classification in other cases which have a greater share in the dataset.

In trajectories with many segments, in general only a few are incorrectly classified. This should not represent significant problems since a journey, when examined as a whole by travel behaviour researchers, is mostly correctly classified, and the small discrepancies are not important and are easy to ignore. If these shortcomings which are clearly not caused by a flaw in the method or prototype are disregarded, the accuracy of the classification system increases.

As the system in many cases returns multiple results sorted by confidence, if the first result was not correct, in almost 100 % of cases the combination of the first two results was correct. This is one of the notable advantages of having multiple classification results sorted by certainty factors.

7 FUTURE WORK AND CONCLUSIONS

The final chapter of this thesis consists of a conclusion (section 7.1), possible contribution of this research (section 7.2 on page 88), and ideas and recommendations for future work (section 7.3 on page 90).

7.1 CONCLUSION

This thesis has shown how to segment and classify a movement trajectory for the used transportation mode(s). This project has been initiated by the need of a classification solution for the data acquired for the travel behaviour study conducted by the Department of Urban and Regional Development at TU Delft (Bohte & Maat, 2009). Their dataset is now classified with a significantly higher accuracy than the existing method that was used for classification, hence the interpretation-validation procedure described in § 1.4.4 on page 8 and Fig. 8 on page 9 may be eliminated if the presented accuracy of the prototype fulfils the requirements of the study.

Apart from travel behaviour research, this method with the developed prototype can find its application in other disciplines that require the segmentation and classification of movement trajectories for transportation modes.

In the frame of the theory of fuzzy expert systems, it was found that it is possible to assign a certainty value to each classification by defining empirical membership functions, and present multiple results with certainty. Expert systems proved that are still actual and supreme for reasoning in complex situations. Development of these solutions is flexible and can be tailored to a specific problem, as shown by this thesis where a system, inspired by the theory and foundation of fuzzy expert systems, was designed with an own concept and implemented from scratch in Python with the support of a PostgreSQL Database Management System (DBMS).

Although it is very difficult to take into account all possible cases in the real-world, the prototype yielded satisfying results, especially in the segmentation and classification of data with the quality degraded by noisy observations, and/or small number of points. It can be concluded that a model for unifying specific mode behaviour is determined to cover the vast majority of all possible situations. The errors are usually not caused by the imperfection of the system, rather by specific situations whose modelling would be either complicated or would impair the existing classification performance. The development of a system that takes into account virtually all possible situations in movement may not be possible.

Geodata is indispensable for the solution of this problem. In this thesis, the majority of the indicators are formed by analysing the proximity to the nearest infrastructure for all modes which are constrained to such. Additionally, geodata is used to resolve gaps, ambiguity in between car, bus and tram, and to enrich the trajectories with additional information about the transportation modes. Without geodata this would not be possible.

OpenStreetMap proved to be a suitable source of such data for this problem. Its strength are availability (anyone can download and use the data), no costs (it is available for free, even for commercial applications), frequent updates of the data (daily snapshots of the data are available for download), and all the needed features are included.

The classification of data missing time intervals is complex, and a solution using the knowledge of the nearest infrastructure is presented. Since this movement is not registered, it may not be possible to reconstruct it and deduce the critical information needed for the classification, especially in longer disruptions. The developed method showed promising results, but the problem is made more difficult in the combination with the usual difficulties imposed by noisy GPS data. On the other hand, the latter problem is solved as visible from the results of the experiments—bad GPS data degrade the accuracy of the classification by just a few percents. Usage of various techniques presented in this thesis results in a method that is less sensitive for noisy and inaccurate data. This is evident in the segments with a high amount of points. There is a considerable positive correlation with the number of points in a segment and accurate classification. Nevertheless, most of the test data used in this thesis are produced with a GPS receiver from 2005, gradually discontinued. Current and especially coming GPS devices are available with sensitive chips that produce data of higher quality—less noise and less signal shortages. It was observed that test data acquired in 2010 with newer receivers had a significantly better quality than the subset acquired from the receivers produced more than a few years ago. In general, due to the better reception, signal shortages are less frequent and the accuracy of acquisition is higher with less occurrences of noisy observations. Moreover, due to higher memory capacity in new receivers, shorter sampling periods are possible for longer trajectories. Therefore, although this thesis solves the problems caused by data of bad quality, such data may not be frequent in close future.

This method manages to distinguish in between a significant number of transportation modes, higher than any other method so far, to the extent of my knowledge. Although a lot of overlapping characteristics between modes exist, with a careful selection of the indicators and modelling of corresponding membership functions the classification of a large number of modes was made possible. Discerning between car, bus, and tram is done thanks to a developed technique of injecting supplementary confidences based on previous knowledge.

The method is further extendable for new transportation modes and indicators, as made clear in the closing section of Chapter 5. The developed experimental software and the validation procedure shown that the method can be used both in the Netherlands and abroad (at least in Europe), as long as quality OSM data containing the needed features is available.

Segmentation was beside the classification the most important aspect of this thesis. Although this work is not novel in realising that virtually all transitions require stopping, it removed the condition for a walking segment before a transition, and introduced a very precise segmentation method which was kindled in combination with the successful classification of very short segments. Since fast transitions may not be left undetected, the segmentation algorithm was set to a very sensitive threshold which in general oversegmented the trajectories, where some of the segments are shorter than 30 m, i. e. a few points, but correctly classified and merged with adjacent segments in case of equal classification outcomes.

Beside answering the research questions, this thesis aimed at solving the 13 problems which are common for existing methods (listed in § 2.2.2 on page 18). Traffic jams and congestions are solved by ignoring very low speeds and attribute such segments with a standing transportation mode. Using mean moving speeds, the bias in the mean speed caused by standing is eliminated, and the moving segments are classified as car segments (1). By eliminating certain modes based on GIS data indicators, modes with similar behaviour are successfully distinguished. In distinguishing car, bus and tram, the knowledge of the previous mode and proximity to the bus/tram stations is used (2). The prototype successfully classifies 10 modes usually at a high certainty without notable ambiguity (3). In this

research all the criteria used by present work is investigated in order to assess their performance and potential for further refinement (§ 3.5 on page 36). All criteria that shown to be relevant are retained and new criteria are introduced, e. g. overlap of a trajectory with a water surface. Although there is a higher number of indicators in comparison to the related work, there is no redundancy between them (4).

The developed prototype for segmentation and classification of movement trajectories is the main product of this research. It is a functional solution which aims at travel behaviour studies, but may be used for most, if not all other applications requiring the knowledge of the transportation mode in a movement trajectory. The developed prototype may serve as a framework of managing trajectories—their import, preprocessing, storing, visualising, segmentation and classification, and it can be easily extended for new transportation modes, new indicators or even new features. Researchers may find the prototype useful for other applications and various statistics about the data. For instance, see Figure 47 where the daily activity of the respondents was generated from the test data, or Figure 48 on the following page where histograms of all speeds of cars, bicycles and trains in the test dataset were generated. Beside a specific transportation mode, the latter could be generated for a specific respondent, time of day, or a journey. Moreover, all the plots presented in this thesis are a result of the programmed features of the developed prototype and may be used for numerous additional representations of the movement data (5).

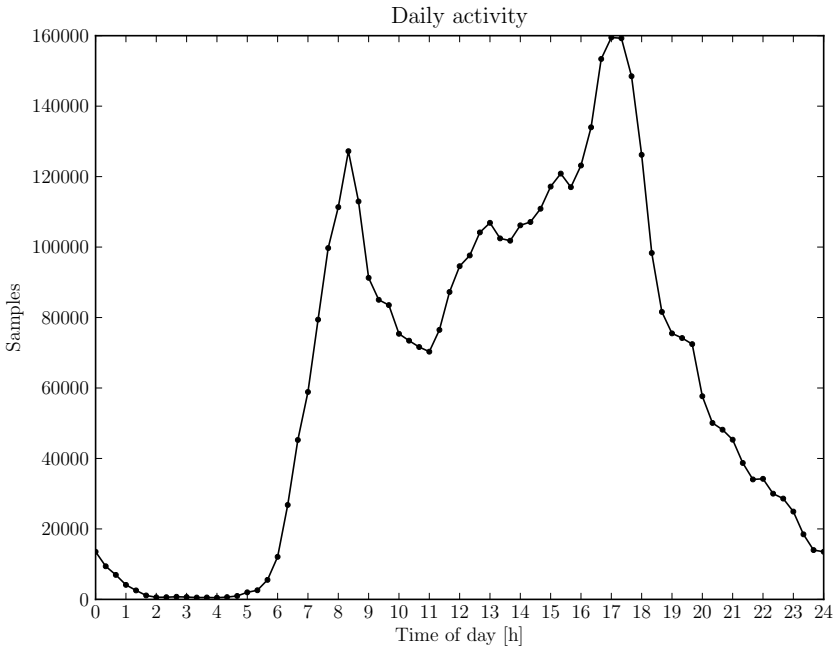


Figure 47: Daily activity of respondents in the test dataset. Adjusted from GPS time to Central European Time and Daylight Saving Time, respectively.

In the classification all results are presented with a certainty measure. This approach has shown that by considering the second result in addition to the main result with the highest certainty, an accuracy very close to 100% can be achieved (6). Data missing time intervals may be very difficult to resolve, even with human intervention. This thesis produced a functional solution, which although does not yield an accuracy comparable to the accuracy of classification of normally recorded trajectories, gives confident results and produces classification results of additional segments which would be ignored in many applications.

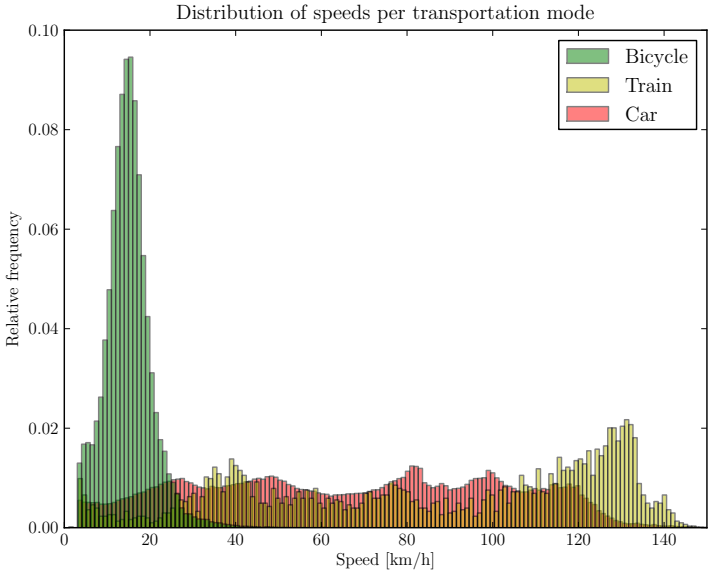


Figure 48: Histogram of speeds of all points in the test dataset classified for bicycle, car, and train.

Moreover, interpolation of such data is investigated (in the appendix—§ A.2 on page 102) (7). Most of the indicators are derived from GIS data, hence it can be concluded that this method relies on values derived from the GIS data as its primary input for the classification. Furthermore, the usage of GIS data is given in details and the usage of OpenStreetMap as an optimal source is motivated (8). As referenced during the course of this report, short segments are easily classified to an accuracy comparable to the one of longer segments (9). The import-preprocess-segmentation-classification process was run for the whole test dataset without any error in the code, and various errors in the data (e. g. duplicate timestamps) are resolved (10). The possibility and implementation for outlier removal is given in the appendix (§ A.1 on page 97) (11). The classification for sea and air modes was introduced in this thesis. Further, the system distinguishes between ferry (incl. motor boats), and sailing boats (12). Experiments involving movement data from different countries (see § 6.2 on page 80) showed that the classification is not degraded by classifying trajectories from abroad thanks to the coverage of OSM data (13).

7.2 SUMMARY AND IMPROVEMENTS WITH RESPECT TO THE EXISTING METHODS

This thesis claims it brings some improvement and contribution to the field of automatic segmentation and classification of movement trajectories for transportation modes. As expected, no improvement could be done regarding basic understandings of the problems, but some solutions have been presented. This section lists the key characteristics of this method and the developed prototype.

1. A higher number of transportation modes in a single system are introduced, i. e. ten, excluding standing and the *special* mode of airport busses introduced at the end of Chapter 5. Further, sea and air modes which have been left out so far are normally included in this system, and successfully

classified. See § 3.1 on page 23 for the list of the considered transportation modes.

2. This method manages to classify basic movement data with only time-stamped positions (x, y, t). Since the elevation is not used as an indicator, it is ignored. This method should work for any present and future acquisition technique (e. g. Galileo) which delivers positions with timestamps in an acceptable accuracy and frequent sampling period. No additional user information are needed (e. g. ownership of a mode), eliminating the costs and process of their acquisition.

3. A functional prototype is developed and presented (see Chapters 4 and 5). The prototype is tailored for both small (e. g. a single trajectory) and vast datasets (e. g. travel research surveys), and it may serve as a framework for managing movement data acquired with travel surveys—importing, pre-processing, storing, visualising, and their segmentation and classification for the used transportation modes.

4. The thesis attempts to solve the problem with gaps in the data, and the implementation delivers usable classification results (see § 5.5 on page 63). The methods of Shalaby et al. (2006) and Stopher et al. (2008) for classifying underground transportation modes are extended to all considered transportation modes in order to attempt the classification of data missing time intervals, usually caused by bad reception of the GPS signal. Moreover, transitions to modes which are not recorded at all are detected.

5. The method and the developed prototype are simple and efficient, and can be extended in future work. Many options are left for investigation in future work, such as reliable outlier removal (§ A.1 on page 97), or acceleration as a future indicator (§ 3.5.2 on page 45). Discussed and calculated, but unused indicators could be used in the future.

   New transportation modes and new indicators may be added by respecting the predefined XML schema, which was shown with the introduction of a special mode (see § 5.11 on page 76).

6. The developed fuzzy expert system is designed and built completely from scratch, not using other tools, for better tailoring for this problem (see § 3.3 on page 29 and Chapter 5 on page 57). The thesis has shown that expert systems and their concepts are very useful for solving various problems in Geomatics.

7. Transportation modes are grouped in similar categories, for giving joint results, rather than inaccurate results in the bottom, more detailed layer. See Table 3 on page 24 for the hierarchy of the considered transportation modes.

8. The presented segmentation technique (§ 3.2 on page 24) is reliable, efficient and "smart", causing very few errors in specific situations. The positions of transition points and transition times are successfully determined.

9. The system is sampling period adapting, and the data are not required to have a fixed sampling period. The presented indicators (see § 3.5.1 on page 37) are adaptive to most of the sampling periods.

10. All the presented technologies (Python and PostgreSQL) and data source (OpenStreetMap) are available for any mainstream operating system and

this solution works on virtually any computer. Moreover, they are available for free to anyone without notable restrictions (see § 3.4 on page 35).

11. The method has applicability in multiple countries. A subset from abroad (a few countries in Europe—see end of §6.2) was segmented and classified at a comparable accuracy as the trajectories from the Netherlands. This is mostly due to the OSM data which is a worldwide project, and its standard of modelling the data that does not require further modifications of the prototype if importing geographic data from other countries.

12. The system successfully deals with uncertainty. The classification process reports certainty factors and multiple results sorted by confidence (see § 3.3.2 on page 31).

13. The method primarily works on the elimination of unlikely modes. This is done through a reliable and efficient rejection of modes through assigning a CF of zero.

14. The use of GIS data and related operations is emphasised. Most of the indicators are derived from GIS data (see § 3.5.1 on page 37). This thesis proved that OSM is the optimal source of geodata for such projects, and that it can contribute with accurate results in this problem. Moreover, the data are updated very often, and it is available to anyone for free.

15. The developed prototype adds additional information about each segment, i. e. name of the arrival bus stop (§ 5.10 on page 75).

16. Classification of short journeys and detection of short segments is done with a success, e. g. even a short walk from a building to a nearby parking lot before the transition to a car (see § 6.2 on page 80).

7.3 FUTURE WORK AND RECOMMENDATIONS

This section gives several recommendations for future work in the field of automatic segmentation and classification of movement trajectories for transportation modes, and ideas for projects that may be started or related to the results of this thesis and the developed prototype.

7.3.1 *Improve the prototype to a complete software*

Considering that although the developed software is in its experimental phase, it may be considered as functional and might already serve as a usable solution for travel behaviour research and other applications not only requiring segmentation and classification but also managing the acquired movement trajectories. However, for future work, the prototype could be improved with a graphic interface and presented to the market. Migration to a faster programming language (e. g. C++) is advised as well.

7.3.2 *Deriving additional transportation mode related information*

Enriching the segments with additional information related to the used transportation mode was not originally a part of this thesis, but it was given conceptually and currently it is in experimental stage in which it delivered satisfying results, presented in § 5.10 on page 75.

Other than the implemented attributes, several other could be implemented, such as designation of the used bus/tram line, flight number (currently not

available due to licensing issues), type of train (train schedules are required), and other reasons for stopping and standing which are included in OSM data—e. g. pedestrian crossings, bridges, and roundabouts.

This has a high potential for some applications. For instance, the additional information derived for public transportation modes may have applications for public transportation companies, or start-up transportation companies exploring new market possibilities. There is a possibility of using this information in the problem of the classification of the trajectories for the trip purpose.

7.3.3 *Removing noisy samples from the trajectories*

I attempted to build a method for detecting noisy observations, however, the obtained results were not satisfying and the approach has not been used in the prototype. A report on this work is included in the appendix (§ A.1 on page 97) and may serve as a guideline for future work. Further investigation on this topic is encouraged since it may directly result in an improved performance of the classification system.

7.3.4 *Modelling of exceptional cases of behaviour without compromising the current results*

A lot of effort was put in modelling specific mode behaviour in a unified model and taking into account several cases that may occur, however, as visible from the presented errors, some cases should be additionally investigated. Since behaviour may vary between regions, for each a special set of membership functions could be set up. However, this requires extensive work in modelling and training the system, and due to the amount and small frequency of such cases it may not even be possible.

7.3.5 *Classification in real-time*

Some applications may require immediate knowledge of the transportation mode, i. e. classification of movement in real-time, hence a method should be developed to determine the transportation mode *on-the-fly*. One of the applications are that the route suggestion on a handheld navigational device might be changed depending on the currently used transportation mode, and several points of interest may be removed from the map, e. g. gas stations in case of walking. With such approach, in future, travel behaviour researchers would be able to acquire the needed information in real-time during a travel survey, saving time and delivering rapid results.

Since real-time classification was not a part of this thesis, its possibility was not further investigated and implemented, however a few guidelines are given below.

Supposedly, such method would require a classification system that would be able classify the currently used transportation mode on the basis of a few most recent points. The presented method should be ready for real-time classification since it is able to classify the trajectory with a limited number of points, i. e. a few points. Another requirement of such system is high performance, since it should be able to calculate several indicators, including many with GIS operations, in a very short amount of time. OSM data are already available for handheld devices making such project possible in the near future.

7.3.6 *Sinuosity and shape of the trajectory as an indicator*

The shape of a trajectory may be an additional indicator, but it is an approach that due to its complexity is outside of the scope of this thesis. For instance, in urban areas walking may have smaller sinuosity than a car due to the traffic regulations. The reader is referred to (Dodge et al., 2009) who further investigate this behaviour.

7.3.7 *Bayesian inference*

As noted in Chapter 3, an expert system returns certainty factors, as a value of confidence for a particular hypothesis. Although the used inference mechanism derived satisfying classification results, it might not be mathematically correct and accurate to deliver probabilities from a simple combination of results from such system, i. e. a probabilistic interpretation of certainty factors is needed. Therefore Bayesian Networks are introduced and in future work their implementation should be considered. Bayesian inference is in usage in GIS, e. g. pattern classification in topographic vector data (Lüscher et al., 2009).

However, the implementation of Bayesian inference was out of scope of this thesis, since most expert system application areas do not have reliable statistical information, and modelling the probabilities is too difficult and time consuming to obtain.

The reader is referred to (Heckerman & Shortliffe, 1992) and (Lucas, 2001) for additional information.

7.3.8 *Classification using Support Vector Machines*

Support Vector Machine (SVM), described in (Chang & Lin, 2001; Hsu et al., 2003), is also a suitable approach for the classification of movement trajectories, as visible in the overview of related work (§ 2.1.6 on page 15). Since the method is different from expert systems and rather a stand-alone method, SVM cannot be used to complement the shortcomings of a expert system classification without significant additional work. However, in future work, it might be valuable to integrate the two methods and activate one when the second does not yield results to a satisfying level of certainty.

7.3.9 *Segmentation by detecting change of "behaviour" in the indicators*

Although the used segmentation technique delivered good results, a segmentation algorithm sensing the changes in behaviour of the indicators (e. g. speed) could be developed in future work as a supplementary method. See the discussion in section 3.2.2.2 on page 27 for more information.

7.3.10 *Extending the benefit of GIS data*

The OSM data used in this project proved to be very useful in the solution of the problem. In general GIS data accounts for most of the indicators and without it the presented solution would not be possible. However, additional work may be done. Although the classification system takes advantage of most of the available geodata, it does not fully exploit the OSM data which in some areas might have useful attributes, such as the type of a road or speed limit. For instance, if it is known that in the Netherlands most of the traffic jams occur on highways, then the classification system may take that fact into account and compensate the speed on such locations. Historical data about traffic jams might be also used,

although, they are not available in OSM. Taking into account the speed limit on a road which is adjacent to a railway would help to alter *on-the-fly* the MF for the maximum allowed speed for cars and correctly classify these cases. The number of lanes of a road, which is conceptually available in OSM but often not acquired, could be used to alter the MF for the proximity to a road (i. e. a wider road would require a wider MF).

7.3.11 *Data mining*

It may be possible to classify the trajectories by using the classification outcome of another similar movement. However, this recommended work involves pattern recognition which is outside of the scope of this thesis.

The proposed approach is tackled in the following two sections, based on the extent of the data—using the trajectories from the current user, or trajectories from all available users. This method may be also used to supplement the classification results from the method described in this thesis.

7.3.11.1 *Involving user's historical data*

In travel behaviour research, a respondent is generally tracked in more than a few days, generating considerable movement data from which his/her travel habits may be deduced. Since a respondent's trajectory ordinarily contains repetitive journeys, not only in space but also in time and usually with the same transportation mode(s) (e. g. every-day commuting), historical user data may be used to additionally catalyse the classification in uncertain situations.

Consider the following case. A trajectory with an individual's travel from home to work may be classified without any flaw and with high certainty. Due to bad quality or incomplete data, another trajectory completed on another day may not be classified with the same certainty. However, if the trajectory was made during a comparable time of day and has a similar pattern in space, the classification system may use a better result from the user's historical data.

Figure 49 on the next page shows a spacetime (2D + time) plot of a respondent's movement in various journeys. It is visible that two journeys have similar spatial patterns. Note that the time in between the journeys is removed in order to avoid blank space.

7.3.11.2 *Involving all users' data*

In movement research surveys, extensive data from numerous respondents is available. By modelling patterns and transportation modes from a group of similar movements, it may be possible to facilitate the classification by searching for a similar trajectory in the database and assign the transportation mode from existing trajectories (classified patterns) in the database. Beside travel surveys, trajectories could be acquired with other projects, e. g. the currently debated kilometre tax project in the Netherlands (Custers & Kuiper, 2009; Ministerie van Verkeer en Waterstaat, 2009b), trajectories from the OSM project, and several leisure-related sources of movement data, i. e. web sites containing paths for runners and cyclists. Further, data from multiple travel surveys could be combined, as it was done in this project.

However, in this approach, a focus should be put on privacy since movement of other people from several sources will be used, and some information may be considered as personal data. Therefore, such trajectories should be anonymised and all information which may indicate to an individual should be removed. Privacy regulations regarding spatial data vary between countries, e. g. in the Netherlands personal data may be generalised to the zip codes level and in

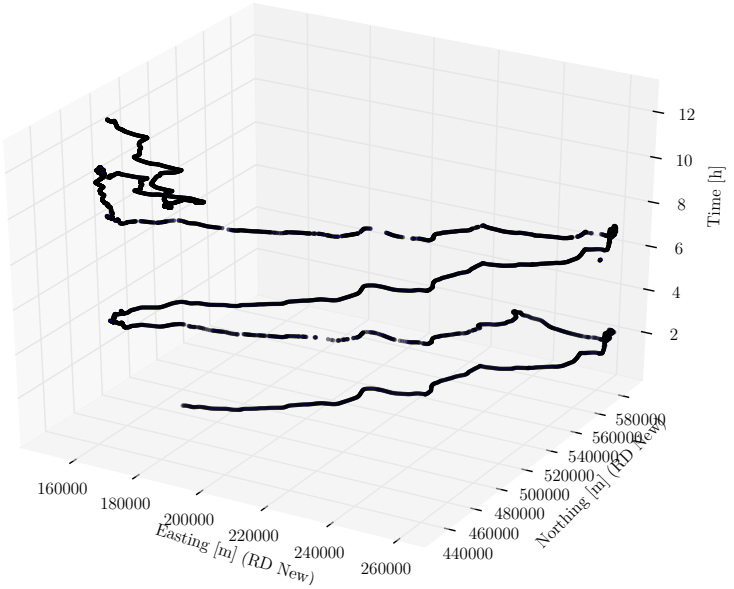


Figure 49: A spacetime plot of a respondent's movement in various journeys. Two of them have similar spatial patterns.

Denmark to a level containing at least 100 addresses, hence a portion of, say, 500 m may be removed from both the beginning and end of each journey in order to remove personal data, but still retain a usable pattern for the purpose of this approach.

Personal data has various interpretations, and although privacy is not a component of this thesis, it may be beneficial to include the definition of personal data not only for the sake of this section, but also for the presented work. In the EU it is defined by the Directive 95/46/EC (of 24 October 1995) on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Article 2a):

> 'Personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.

7.3.12 *Additional data about the user as an indicator*

Although the objective of this thesis was to classify the movement trajectories solely from the timestamped positions (raw GPS data), the classification in travel behaviour studies may be improved with the use of additional data, as Stopher et al. (2007) suggest.

- Preferred transportation mode
- Mode ownership and access to certain modes
- Occupation
- Driving licence possession

The presented data was not available in the survey data used in this thesis.

7.3.13 *Enriching "incomplete" spatial data*

Although the main applications of the segmented and classified data are clearly stated, I propose an additional application from the derived data: enhancing and supplementing existing spatial data. Once the transportation modes and other information of a vast dataset of trajectories are known, it might be possible to use them for repairing or improving OSM data. For instance, roads without the attribute of its type might be classified depending on the recorded average speed of all the points in the vicinity made by a car.

As an example, Figure 50 shows the tessellation of the same spatial extent presented in Figure 9 on page 10. For each square the mean speed of all containing points was calculated and visualised. Just from this derived information, various attributes of features could be enriched (e. g. the road with each intersecting tessellation element may be analysed with the nearby classified data for cars). Related to the speed, the speed limits (both minimum and maximum) could be estimated, and the traffic usage.

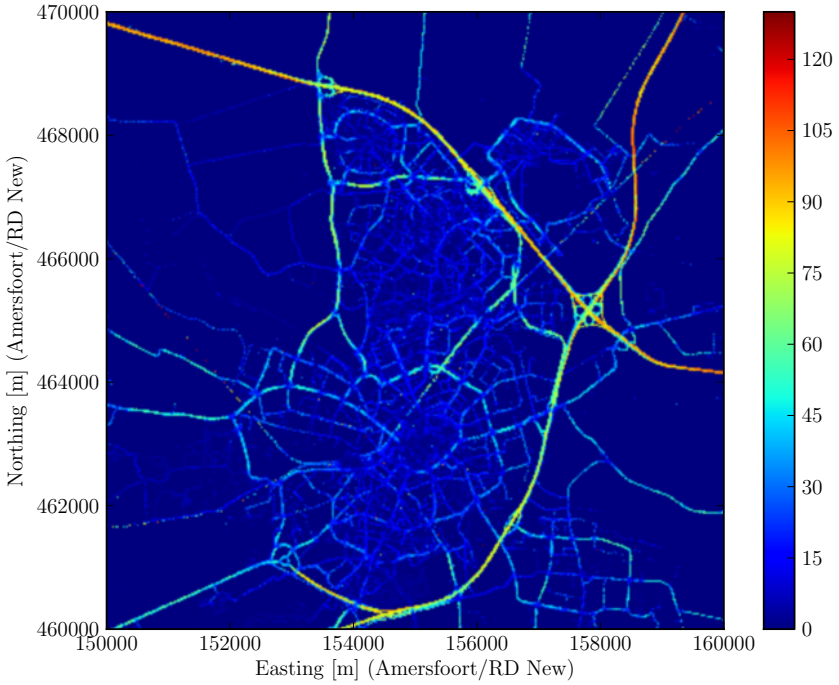


Figure 50: Average speed in the area of Amersfoort from the test dataset. Tessellation with squares of 25 m × 25 m.

Examples of other attributes regarding roads that could be populated: one/two way street (from the direction of travel), number of lanes (from the distribution of the deviations from the infrastructure), traffic lights (congestions of stops), speed bumps (from massive sudden deceleration at one point), tunnels (frequent shortages of signal), and bridges (grouped trajectories over water made by a land mode).

The presented examples are given for roads, but other possibilities for other infrastructure are possible as well, for instance mapping ports, ferry lines, and bicycle paths.

7.3.14 *Classification of the trip purpose*

As stated in the introduction of this thesis, the classification of the movement trajectories for trip purpose is a partially related project that could not be integrated in this MSc thesis' topic. However, work from this thesis may be used for assistance in that problem. For instance, the developed prototype already imports the data and preprocesses it for various indicators which would be required in a project for the determination of trip purpose as well. The result of this project, the knowledge of the transportation mode is already probably a good indicator of the purpose of a journey, as the developed reasoning system for deriving the reason for standing and additional mode-related information about each segment.

The OSM data, used in this project, contains the data which would be required by the classification for trip purpose (features such as supermarkets and leisure centres), and an algorithm for calculating the proximity to a nearest feature related to a trip purpose, could be easily integrated in the developed prototype. A new project may be built on top of this, since a complete solution for travel behaviour research should classify for both the transportation mode and trip purpose.

A REPARATION OF MOVEMENT TRAJECTORIES

This chapter discusses the possibility of removing the noise from the data, and interpolation of data missing time intervals. Since the presented techniques have not been actively used in the project, they are presented in the appendix. The concept and preliminary results are given, with the reasons of their omission from the main part of the thesis. Both concepts have been implemented in the preprocessing stage of the prototype and may be activated for future work.

A.1 REMOVING NOISE FROM THE MOVEMENT DATA

This section gives an overview of the possibilities for removing noise from the recorded trajectories. Four techniques are implemented in the prototype and the results of experiments are presented (§A.1.2 to A.1.5 on pages 98–101). The conclusion is given in § A.1.6 on page 101.

A.1.1 *Introduction*

GPS measurements are subject to random errors augmented by bad signal reception and bad satellite geometry, hence some trajectories contain noise which is often manifested by sudden positional jumps. The deviations from actual positions may exceed 100 m which may impair the results of the classification method, hence it is expected to remove such outliers from the dataset in the preprocessing stage of the prototype. I consider a point valid if its position is determined under 2σ of the standard GPS errors, that is, the position is recorded with the usual (expected) accuracy determined by the manufacturer.

In detecting the outliers, if available, it is logical to analyse the DOP value, which gives an indication of the precision of the measurements from the geometry of the satellites, and it is often available along with its variants Horizontal Dilution of Precision (HDOP), Position Dilution of Precision (PDOP), and Vertical Dilution of Precision (VDOP). However, by analysing several datasets containing outliers, I have not found any relation between noise and any of these values.

Schüssler & Axhausen (2009) use two approaches for making their method robust for outliers. First, the authors observe that the outliers, beside sudden jumps in 2D position, are also manifested with a sudden jump in altitude, thus they remove all points which have the altitude outside a specific range in the area (for Switzerland they determined the range of [200 m, 4200 m]), but in the conclusion admit that this method is not sufficient, hence the mean and maximum speeds and accelerations are not considered as variables in the fuzzy system to make the method robust for any remaining outliers, instead 95th percentiles are used. In the second approach, they remove all segments which have the speed higher than 180 km/h, similarly to Auld et al. (2008) who suggest to filter all points above the threshold of 160 km/h.

By analysing outliers in my datasets, I have found no correlation between the sudden jump in 2D position and their altitude difference to adjacent valid points. Further, such filtering would discard points from aircrafts which not only have high elevation values, but also highly fluctuating elevations.

Filtering all points with the speed above a certain threshold would also discard points which are from a high-speed train or an aircraft, hence it cannot be used in this project. Setting a transportation mode specific threshold, for instance

40 km/h for a bicycle, is not possible since the transportation mode is not known in the importing and preprocessing stage of the prototype.

In relation to high values of speeds, acceleration can be added into consideration. However, analysing the mentioned datasets which contain noise, I have found no relation between the value of the acceleration and the positional jumps, either. High accelerations are normally found in trajectories with highly variable speeds such as a car in urban area, i. e. valid points.

In a different approach by Taylor et al. (2006), unrelated to transportation mode detection, it is possible to detect noise by matching the points to a GIS network and finding isolated points not matched to the same road. However, this approach has big disadvantages in our case: it is valid only for specific modes where geo-information is available (i. e. cars and trains), and at this point the mode is not known so the selection of the corresponding network is not possible.

There are additional approaches I designed for detecting noisy points as points with:

- sudden jumps in the speed,
- sudden jumps in the acceleration,
- sudden changes in heading, and
- points which are outside an error buffer.

For testing these theories, I have implemented four noise detection algorithms and I have used subsets of points from two segments which contain a few outliers. The first segment A, made with a car, contains 12 points denoted with $A_0 \dots A_{11}$. Points which are considered as outliers are adjacent: A_4, A_5, A_6, A_7. The speeds were determined from the GPS receiver.

The second subset B, which contains 16 points, was made with a bicycle, and the speeds were computed from the positions and timestamps. This is different from the dataset A, in order to test is there a difference with the different methods for calculating the speeds, and therefore accelerations. There are three outliers (B_5, B_6, B_{14}), which are not adjacent, unlike in the previous dataset. The outlier B_{14} has a relatively small distance from the actual value, however, it was included to test the implementation for outliers which have smaller jumps.

By using these datasets, my assumptions were tested for a wider detection of outliers—for different modes, different speeds, and datasets that may have multiple and adjacent outliers. Each theory is described in a separate section in the continuation.

A.1.2 *Detecting variations in speeds*

For the detection of high jumps in speeds, we start by *smoothing* the speed of each point by averaging it with the values from the two adjacent points:

$$\overline{v_i} = \frac{1}{3}(v_{i-1} + v_i + v_{i+1}) \tag{A.1}$$

Comparing the smoothed speed with the reported speed may help detect noise. However, not all points were smoothed. Because speeds can be highly variable in longer periods of time and may still be valid, only points which are sampled in below 10 s are smoothed. Then, by taking the difference between the original and smoothed speeds ($|\overline{v_i} - v_i|$), it should be possible to detect unrealistic values. The threshold was set to 10 km/h. Figure 51 shows an example of the differences

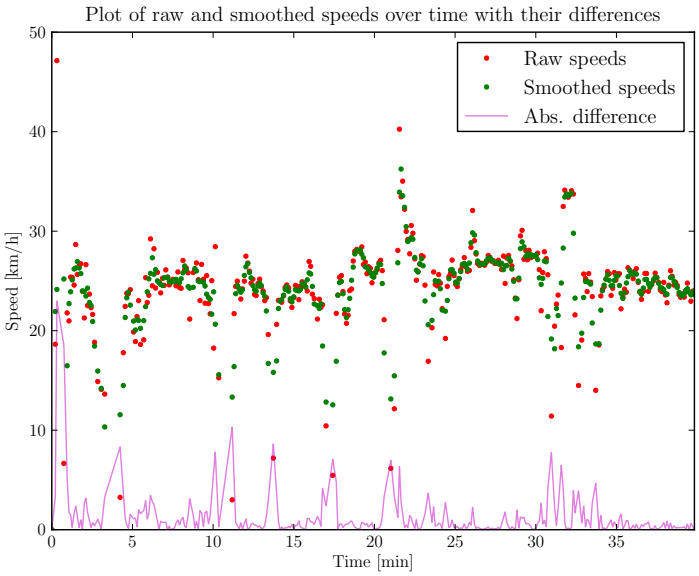
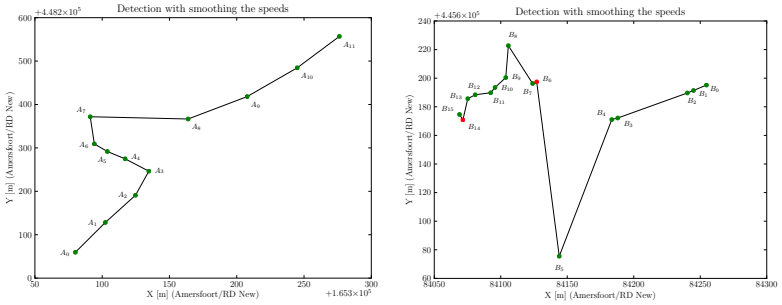


Figure 51: Differences between raw and smoothed speeds in a random segment.

between raw and smoothed speeds. Note the outlier on the beginning of the segment.

However it was observed that a point with a sudden jump in the speed also biases the smoothed speeds of the two adjacent points, hence only the middle point of three consecutive points exceeding the threshold was marked as an outlier. The results are shown in Figure 52. The invalid points are shown in red.



(a) Dataset A. (b) Dataset B.

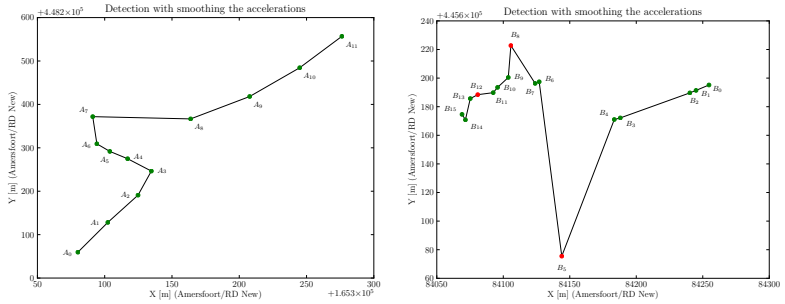
Figure 52: Detecting outliers with smoothing the speeds and calculating the differences.

The algorithm did not detect any outlier in the first example, since outliers did not exhibit high fluctuation of speeds, probably because the speeds were not calculated from the positions, unlike in B. In the second example, the algorithm did detect the outlier B_{14}, however not the two more evident outliers. Also a valid point (B_6) was marked as an outlier. Lowering the threshold does not bring better results since it detects more valid points, especially points that are close to regular variations of the speed, for example, stops at traffic lights.

A.1.3 *Detecting variations in the accelerations*

This part is analogue to the latter method. The accelerations were smoothed in the same fashion as speeds, and the differences were computed. The threshold for detecting outliers in this case was set to $0.6\,\mathrm{m/s^2}$.

From Figure 53 we can see that similarly to the previous method, no outliers were detected in the dataset A, however it was more successful in the second dataset. Outliers B_5 and B_8 were correctly marked, however, a valid point B_{12} was detected as an outlier as well. I observed that many other valid points are detected as outliers in other datasets, hence this method is not reliable as well.

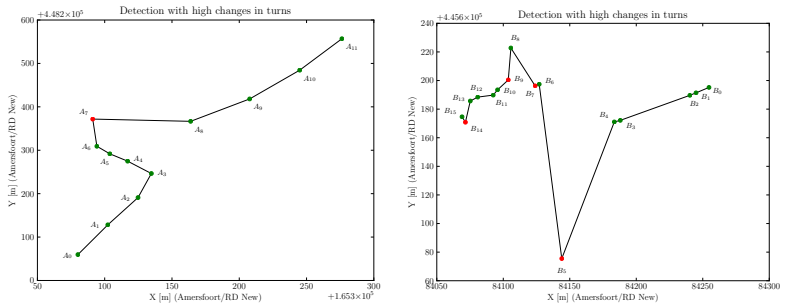


(a) Dataset A. (b) Dataset B.

Figure 53: Detecting outliers with smoothing the accelerations and calculating the differences.

A.1.4 *Detecting variations in heading*

From the presented figures, it can be observed that outliers may be detected with analysing the azimuths of the points with adjacent points. In this case, for each point the azimuth to the following point was computed. Note that a single outlier by deviating from the actual trajectory causes the azimuth at itself and the previous point to deviate by a high value as well, therefore I assumed that if two consecutive points have azimuths differing by more than 30° from the preceding points, than the second point is probably an outlier. The Figure 54 shows the result of this approach for the two test datasets.



(a) Dataset A. (b) Dataset B.

Figure 54: Detecting outliers with detecting sudden turns.

In the dataset A, the farthest point from the actual movement was detected as an outlier, however, its foregoing points which are also outliers were not detected. In B, the algorithm correctly detected B_5 as an outlier, however interestingly it

rather detected adjacent points of B_8 as outliers than the point itself, since of its proximity to the outlier B_5. In the end of the trajectory, B_{14} was correctly marked as an outlier.

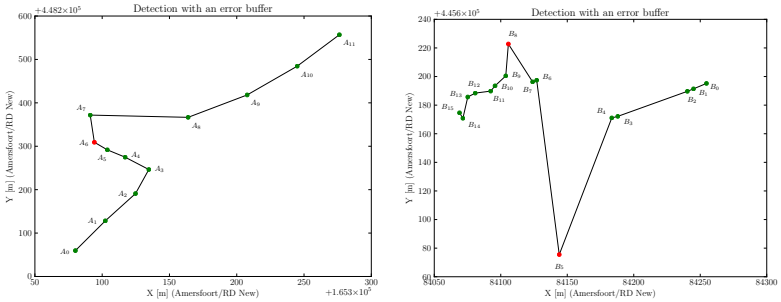
This method brings more confidence in detecting outliers, however, it causes a significant amount of valid points to be marked as outliers as well since sudden changes in headings are possible for some modes and various cases, for instance, usual and longer turns in roundabouts and highway junctions.

A.1.5 *Error buffer method*

Since the transportation mode is not known at this point, it is not possible to set up a constant error buffer where if a point is too far from the adjacent points it is marked as an outlier. Therefore, I created buffers with variable sizes, depending on the speed of the preceding point:

$$r = 2v_{i-1}\Delta t \tag{A.2}$$

That is, if the point is more than twice the predicted distance away from the previous point, it is an outlier. The predicted distance is calculated as the multiplication of the speed at the previous point and the time difference. Even with sudden accelerations and increase in the speed, valid points should still lie in the buffer. An additional constraint was put, that the buffer is calculated only for speeds higher than 10 km/s since lower speeds would results in a small buffer that is vulnerable to GPS errors. The results are shown in Figure 55.



(a) Dataset A. (b) Dataset B.

Figure 55: Detecting outliers with setting speed dependent error buffers.

Although not all of the outliers were detected, valid points were not marked incorrectly as invalid.

A.1.6 *Conclusion*

In this section I have presented and discussed nine options for detecting noise in the data, including four methods which I designed and implemented myself. The tests have shown that none is robust enough to detect the majority of the outliers without marking valid points as outliers too. In some promising methods, the aspect of lowering the thresholds has been investigated as well, however, that causes too many valid points to be filtered out. Methods may be reliable in detecting single outliers, however the presence of fairly close outliers shows the vulnerability of all methods. Where the speeds are detected from the GPS, as in more than 99 % of the test data, outlier detection is difficult since the speeds, and later the accelerations, are not subject to high variations as the positions are.

Furthermore, implementing all of these methods shown that it is computationally very expensive to detect and remove noise. Detecting outliers in the data takes approximately 50 % of the total time of importing the data, and causes repeating the whole import process after their removal since several attributes have to be recalculated for points which were not adjacent in the first import, which is problematic with huge datasets.

Since in the presence of a high number of points, and by incorporating various techniques for calculating the indicators, a small number of outliers did not pose significant problems, in this project outlier filtering was not performed. To make the method robust for outliers and remove their effect from the indicators, 95th percentiles have been used, i. e. *almost maximum* values of statistical descriptors, as Schüssler & Axhausen (2009) and Stopher et al. (2008) suggest. The calculation of mean values in noisy subsets is usually due to the larger number of points not biased by outliers.

The presented methods for removing noise can be activated any time in the developed prototype if the user prefers so, but in this thesis the showed results are obtained without removing outliers.

A.2 INTERPOLATION OF MISSING TIME INTERVALS

Here I discuss the possibility of interpolating missing data in order to improve the classification. Missing data is considered as intervals which are for some reason, usually because of signal shortage, not sampled at the expected rate and interrupted at some point. In addition, data between two normally sampled points in longer sampling periods may be additionally interpolated for more data and possibly better segmentation and classification results.

Since the speeds and the position of a point (in order to calculate its proximity to the nearest network infrastructure) are the two most important information, I investigated their interpolation. One might argue that the speeds can be derived from the (interpolated) positions, however, the interpolation of the speeds is considered separately due to the one-dimensional case which may be easier to handle, but may also give more reliable results.

A.2.1 *Speed interpolation*

Inverse Distance Weighting (IDW) and Kriging appear to be the two suitable approaches for testing the interpolation of the speeds.

INVERSE DISTANCE WEIGHTING IDW is an interpolation method based on assigning the weights to each (nearby) observation with respect to its distance to the unknown point (Shepard, 1968):

$$u(\mathbf{x}) = \sum_{k=0}^{N} \frac{w_k(\mathbf{x})}{\sum_{k=0}^{N} w_k(\mathbf{x})} u_k \tag{A.3}$$

where u is the interpolated value at a location $\mathbf{x}$, and u_k the observations. The weight w_k for the observation k are determined with

$$w_k(\mathbf{x}) = \frac{1}{d(\mathbf{x}, \mathbf{x_k})^p} \tag{A.4}$$

In this one-dimensional case, I consider that the observations of speeds are separated by time, rather than distance, since my side tests involving both time

and distance showed that interpolations taking the distance between two points yield worse results.

For testing purposes, a journey made by a car, with $n = 301$ points is used. The average sampling period is 6.5 s, hence the duration of the recorded trajectory was approximately 32 min. With the jackknifing technique (Quenouille, 1956), each point from the dataset was temporarily removed and was interpolated with the remaining $n - 1$ points from the training dataset, the subset of the initial dataset without that single point ("leave-one-out"). The method was assessed by computing the error between each observed (*real*) and estimated (interpolated) value. In this trajectory, the errors are normally distributed around $\mu = 0.06$ km/h, with the standard deviation $\sigma = 13.27$ km/h, and are shown in the Figure 56a along with the recorded and estimated values. The mean absolute error of the predictions was 10.44 km/h. The standard deviation is quite high considering that the speeds rarely exceed 50 km/h, and that the gaps are usually not larger than 13 s, that is, each interpolated point is in the majority of cases 6 or 7 s away from the two nearest known values. By analysing other trajectories in the same way, it was observed that the error is dependent on the distribution of speeds. In trajectories containing speeds up to 120 km/h, the errors often reach 50 km/h, making this method unsuitable for this project, since it can seriously degrade the results.

In the second test, a longer gap of 5 min was simulated by removing 46 consecutive observations from the original dataset (from the 11.8th to the 16.8th minute of the same journey). The speeds were again interpolated with IDW from the remaining points in the dataset (training data), and the errors were computed from the original dataset (Figure 56b). This time, the errors are not normally distributed, however, their absolute values are comparable to the errors produced by interpolating much smaller gaps (mean abs. error is 14.46 km/h). From the plot, we can see that a high error is present for points without movement (for example see the point at the 14 min). Moreover, the method is highly sensitive to sudden changes of values, i. e. accelerations caused by traffic lights.

KRIGING In the geostatistical approach, two experimental variograms (Eq. A.5) of two different but adjacent 3 km × 3 km spatial extents are computed to show the average dissimilarity of speeds of known observations separated by their distance. The variograms, which are included in Figure 57 on page 105, are significantly different, hence a single model cannot be developed for the interpolation of speed using Kriging techniques.

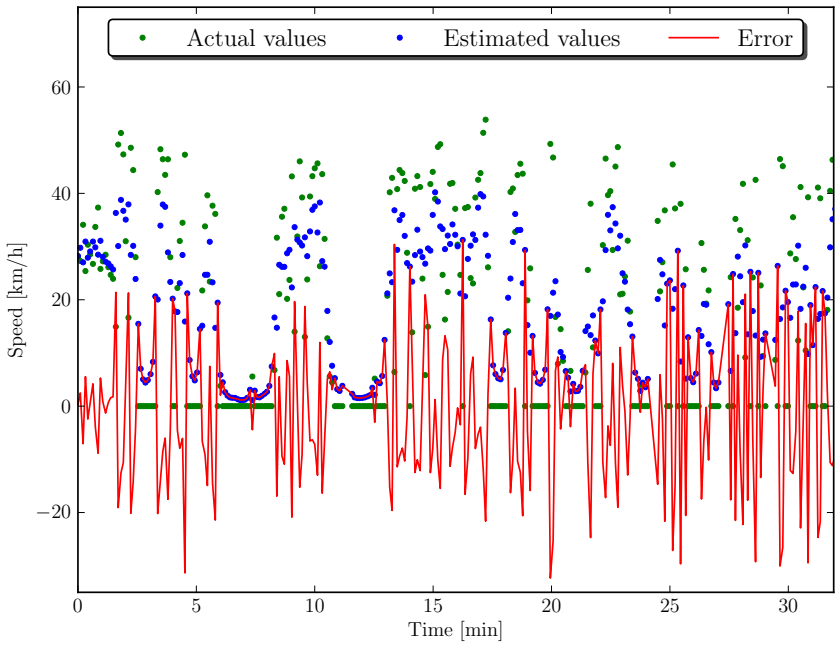
$$2\gamma(x, y) = E\left(|Z(x) - Z(y)|^2\right) \tag{A.5}$$

A geostatistical approach would require a single (or very similar) variogram model for interpolation. Otherwise, each trajectory should be approached individually which is not feasible in large-scale projects.

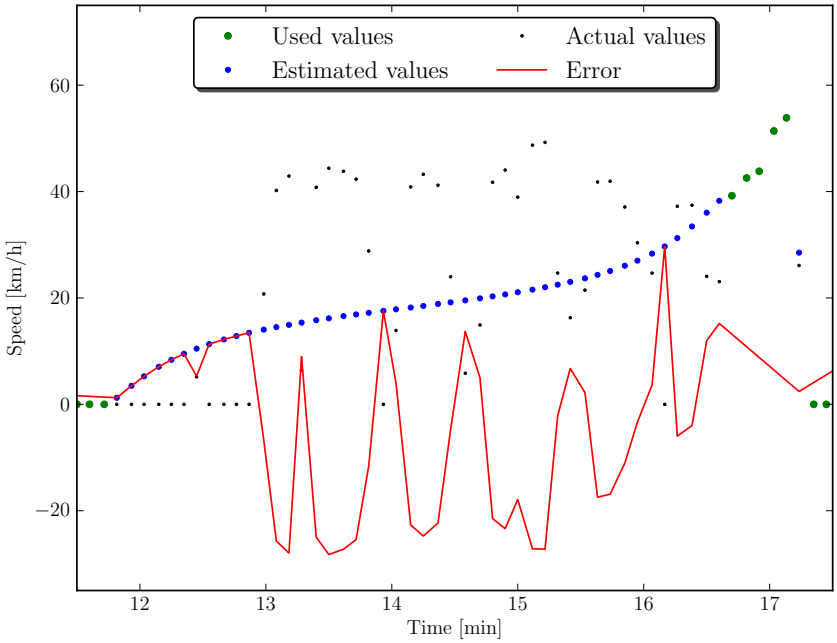
A.2.2 *Location interpolation*

The interpolation of locations of the points missing time intervals is more demanding since the result is multi-dimensional. In the interpolation of locations, the goal is to reconstruct the approximate position of the movement of the user, however, rather than interpolating discrete points, I concentrated on reconstructing the trajectory and area in which the user moved.

One possibility that arises, is to match the sampled points to a GIS network and reconstruct the movement by finding the shortest route between the two points. However, this method cannot be used since at that point the knowledge



(a) Cross-validation of each point.

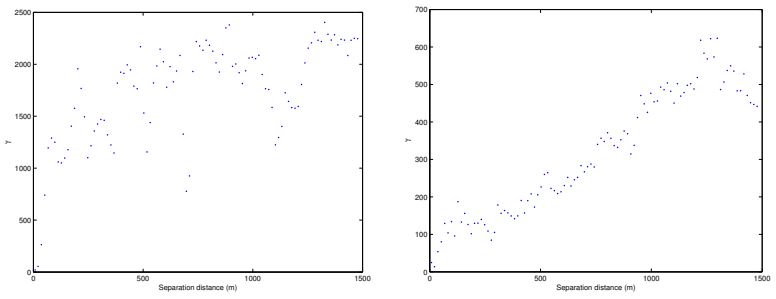


(b) Cross-validation of a longer gap.

Figure 56: Speed interpolation using IDW with quality assessment.

of the used transportation mode is missing, therefore the type of a network that has to be used, i. e. bus network or rail network.

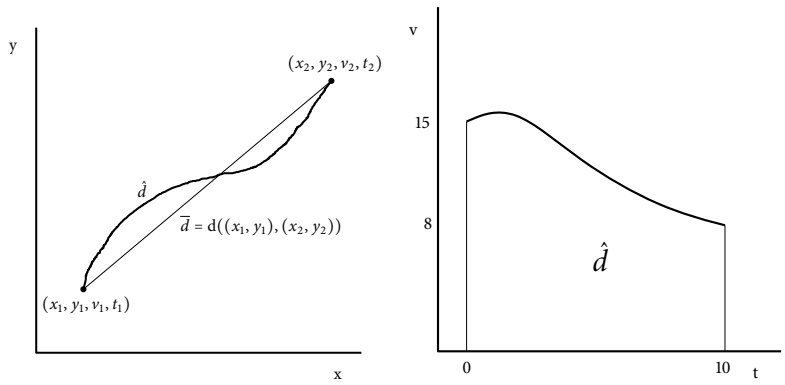
For finding the trajectory, one logical approach is to interpolate a B-spline through the known points. For estimating the area in which the user moved, the *probable* bounds of the movement of the user between two known points could be estimated. It is expected that these bounds form an ellipse connecting the two points.



(a) First variogram. (b) Second variogram.

Figure 57: Variograms of speeds for different spatial extents.

From the location and time of two consecutive points, the distance $\overline{d}$ and time difference Δt can be computed. However, as the Figure 58a shows, the distance $\overline{d}$ can be different from the distance of the travelled path $\hat{d}$, which is especially frequent in longer sampling times. Therefore the computed mean speed $\overline{v}$ is different from the mean speed $\hat{v}$ of the travelled path. Additionally, the speed in the path may be highly variable (Figure 58b). This clarifies the previously raised concerns for calculating the speeds from positions.



(a) Actual travelled distance. (b) Variability of the speed.

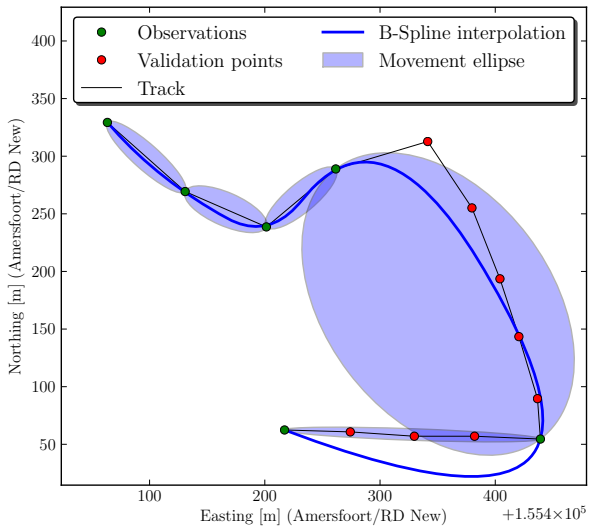
Figure 58: From the sampled points it is not possible to derive the real speed and actual travelled path.

In an example, two points are sampled. Their separation distance is 100 m, while the sampling period was 10 s yielding the mean speed $\overline{v}$ of 10 m/s. If additional knowledge of the speed at the sampled points is present in the GPS data, e. g. $v_1 = 15\,\mathrm{m/s}$ and $v_2 = 8\,\mathrm{m/s}$ the acceleration can be calculated. However, that still does not mean that the user had a constant acceleration during the path, i. e. at one point the user could have reached the speed of 16 m/s. If we assume that the object have undergone the path with constant acceleration, we can calculate the approximation of the distance of the travelled segment:

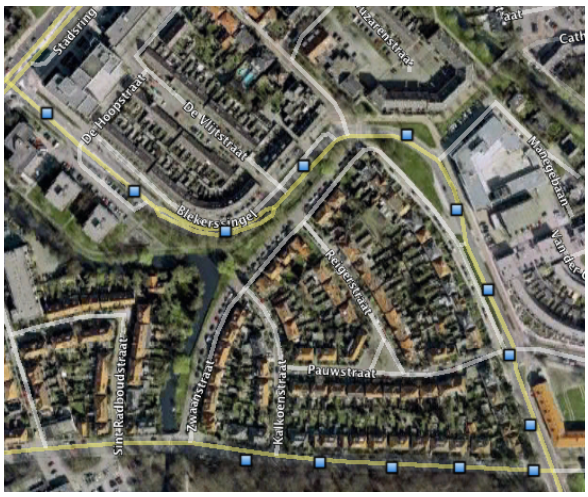
$$\overline{\hat{d}} = \frac{v_1 + v_2}{2} \Delta t \tag{A.6}$$

which in this example equals to 115 m. If we consider this value as the best estimator for the travelled path, then the circumference of the movement ellipse is twice the travelled path, since the user's direction is unknown.

The following example clarifies this theory, but also gives the example of the B-spline interpolation. Two longer gaps are simulated in a short segment made with car with 14 points sampled at the average rate of 6.5 s. The closest two known points from the first gap, from which 5 points are removed are separated by 294 m and 38 seconds, while in the second (3 points removed) the difference is 222 m and 26 s. Taking into account the speeds at the closest known points, the travelled path was estimated to and a movement ellipse was constructed (Figure 59a). The figure also shows the interpolated spline. Additionally, the ellipses were generated for the travelled paths in between the *expectedly* sampled points, i. e. short gaps. For comparison, the sampled points over a satellite image with the road network was added in Figure 59b.



(a) Interpolation of locations.



(b) Data on a satellite image. Imagery © Aerodata International Surveys and Google (2010).

Figure 59: Interpolating locations.

It is visible that this method is sensitive to changes in heading, and the ellipse becomes very large for longer gaps with high estimated travelled distances. Moreover, the area is symmetric since the directions cannot be easily deduced, resulting in high coverage that yields a high error to the result.

By enlarging the ellipse for a higher confidence level, the area becomes larger, additionally having a negative effect on the quality of the prediction, for instance, it may include coverage of too many features non-related to the transportation mode in question, i. e. distant railway which is not connected with the movement in question.

The spline yields similar results—but it is sensitive to changes in heading, and also computationally expensive. The interpolation of the trajectory in the larger gap is promising, however, the interpolation of the shorter gap is quite non-satisfactory due to the sudden change in heading. The interpolation in normally spaced gaps is above initial expectations, however, these gaps often need not to be interpolated for trajectories, even in longer sampling periods, as long the logged points are regularly sampled without data interruption.

A.2.3 *Conclusion*

The investigated interpolation techniques do not give accurate and complete results, hence they have not been used to repair the test data available for this project. The speed interpolation does give high errors even for data with very small gaps which do not need to be interpolated, while the location interpolation with the movement ellipse gives the boundary of the possible movement of the user which is very large to be considered for application. Moreover, the width of such ellipse cannot be calculated precisely because of approximations and complicated calculus. In this section, the circumference is approximated with the Euler's formula which can give errors up to 20 % (Weast et al., 1954). Other, more precise approximations, are computationally expensive and difficult to implement.

The trajectory interpolation is useful in some cases, but in others it may give significant errors which degrade the results. Interpolating both speeds and locations to a higher certainty would require an individual human approach, which is not feasible in large-scale projects, or additional considerable amount of time to develop new reliable interpolation techniques. Further, the fact that the magnitude of GPS errors may result in a high bias even before any interpolation, brings a conclusion that we should concentrate on sampled points, rather then uncertain paths. Since the heading and the travelled path are not much variable during short sampling periods, the deviation is not high, hence it is very important to have a short sampling period in the data. The test data had been sampled every 6.5 s on average, which proved to be satisfying in the scope of this project.

Alkemade, I. (2000). *Beeldschermkartografie ten behoeve van multi-bron internet GIS*. M.Sc. thesis, Delft University of Technology. (Cited on page 29.)

Asakura, Y., Tanabe, J., & Lee, Y. (2000). Characteristics of positioning data for monitoring travel behaviour. In *7th World Congress on Intelligent Transport Systems, Torino*, (p. 8). (Cited on page 1.)

Auld, J., Williams, C., & Mohammadian, K. (2008). Prompted recall travel surveying with GPS. In *2008 Transport Chicago Conference*, (p. 16). (Cited on page 97.)

Axhausen, K., Schönfelder, S., Samaga, U., Wolf, J., & Oliveira, M. (2004). 80 weeks of GPS-traces: Approaches to enriching the trip information. In *Transportation Research Board 83rd meeting*, (p. 28). (Cited on pages 2 and 47.)

Bohte, W., & Maat, K. (2008). Deriving and Validating Trip Destinations and Modes for Multi-day GPS-based Travel Surveys: An Application in the Netherlands. In *Transportation Research Board 87th Annual Meeting*, (p. 17). (Cited on pages 2, 14, and 49.)

Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transport Res C-Emer*, *17*(3), 285–297. (Cited on pages 1, 8, 9, 14, and 85.)

Bohte, W., Maat, K., & Quak, W. (2008). A method for deriving trip destinations and modes for GPS-based travel surveys. In J. Van Schaick, & S. Van der Spek (Eds.) *Urbanism on Track*, chap. 10, (pp. 129–145). IOS Press. (Cited on pages 14 and 37.)

Bricka, S., & Bhat, C. (2006). Comparative Analysis of Global Positioning System-Based and Travel Survey-Based Data. *Transportation Research Record: Journal of the Transportation Research Board*, *1972*, 9–20. (Cited on page 2.)

Buchanan, B. G., & Duda, R. O. (1982). Principles of Rule-Based Expert Systems. In M. Yovitz (Ed.) *Advances in Computers*, vol. 22, (p. 62). Academic Press, New York. (Cited on page 29.)

Byon, Y.-J., Abdulhai, B., & Shalaby, A. (2007). Impact of Sampling Rate of GPS-Enabled Cell Phones on Mode Detection and GIS Map Matching Performance. In *Transportation Research Board 86th meeting*, vol. 07-1795, (p. 21). (Cited on pages 6 and 45.)

Byon, Y.-J., Abdulhai, B., & Shalaby, A. (2009). Real-Time Transportation Mode Detection via Tracking Global Positioning System Mobile Devices. *Journal of Intelligent Transportation Systems*, *13*(4), 161–170. (Cited on page 13.)

Byon, Y.-J., Shalaby, A., & Abdulhai, B. (2006). Travel time collection and traffic monitoring via GPS technologies. In *IEEE Intelligent Transportation Systems Conference, 2006. (ITSC'06)*, (pp. 677–682). (Cited on pages 13 and 41.)

Cao, X., Mokhtarian, P., & Handy, S. (2009). Examining the Impacts of Residential Self-Selection on Travel Behaviour: A Focus on Empirical Findings . *Transport Reviews*, *29*(3), 359–395. (Cited on page 1.)

Chang, C., & Lin, C. (2001). LIBSVM: a library for support vector machines. Tech. rep., Department of Computer Science, National Taiwan University. (Cited on page 92.)

Chung, E.-H., & Shalaby, A. (2005). A Trip Reconstruction Tool for GPS-based Personal Travel Surveys. *Transportation Planning and Technology*, *28*(5), 381–401. (Cited on pages 13 and 42.)

Custers, B., & Kuiper, A. (2009). Data on the Move–Privacy of Road Pricing. *The Journal of Navigation*, *63*(1), 51–59. (Cited on page 93.)

De Boer, A. (2008). Analysis of GPS logs for algorithm design of movement behavior studies. Tech. rep., Delft University of Technology. (Cited on pages 15 and 42.)

Dodge, S., Weibel, R., & Forootan, E. (2009). Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems*, *33*(6), 419–434. (Cited on pages 15 and 92.)

Draijer, G., Kalfs, N., & Perdok, J. (2000). Global Positioning System as data collection method for travel research. *Transportation Research Record: Journal of the Transportation Research Board*, *1719*, 147–153. (Cited on page 2.)

Friedman-Hill, E. (2003). *Jess in Action*. Greenwich: Manning Publications. (Cited on page 30.)

Garibaldi, J. M. (2005). Fuzzy Expert Systems. In B. Gabrys, K. Leiviskä, & J. Strackeljan (Eds.) *Do Smart Adaptive Systems Exist?*, vol. 173 of *Studies in Fuzziness and Soft Computing*, chap. 6, (pp. 105–132). Springer-Verlag. (Cited on page 31.)

Ghorbel, H., Bahri, A., & Bouaziz, R. (2009). Fuzzy Protégé for Fuzzy Ontology Models. In *11th Intl. Protégé Conference, Amsterdam, The Netherlands*, vol. 12, (pp. 18–30). (Cited on page 31.)

Giarratano, J. C., & Riley, G. D. (1998). *Expert Systems: Principles and Programming, Third Edition*. Course Technology. (Cited on page 30.)

Gonzalez, P., Weinstein, J., Barbeau, S., Labrador, M., Winters, P., Georggi, N., & Perez, R. (2008). Automating Mode Detection Using Neural Networks and Assisted GPS Data Collected Using GPS-enabled Mobile Phones. In *15th World Congress on ITS*, (p. 12). (Cited on pages 16 and 45.)

Gonzalez, P., Weinstein, J., Barbeau, S., Labrador, M., Winters, P., Georggi, N., & Perez, R. (2010). Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. In *IET Intelligent Transport Systems*, vol. 4, (pp. 37–49). (Cited on pages 16 and 46.)

Grazia, C. U. (2006). Introduzione ai sistemi esperti. Tech. rep., Sistemi di elaborazione delle informazioni, Dipartimento di Scienze, Università degli Studi "Gabriele D'Annunzio". (Cited on page 29.)

Greenfeld, J. (2002). Matching GPS observations to locations on a digital map. In *Transportation Research Board 81st Annual Meeting*, (p. 13). (Cited on page 53.)

Grimshaw, D. (2001). Certainty Factors. Notes of CPS 820 Knowledge Based Systems, Ryerson University. (Cited on page 30.)

Heckerman, D., & Shortliffe, E. (1992). From certainty factors to belief networks. *Artificial Intelligence in Medicine*, *4*(1), 35–52. (Cited on page 92.)

Holzmann, C. A., Pérez, C. A., Held, C. M., Martín, M. S., Pizarro, F., Pérez, J. P., Garrido, M., & Peirano, P. (1999). Expert-system classification of sleep/waking states in infants. *Medical and Biological Engineering and Computing*, *37*(4), 466–476. (Cited on page 29.)

Hsu, C., Chang, C., & Lin, C. (2003). A practical guide to support vector classification. Tech. rep., Department of Computer Science, National Taiwan University. (Cited on pages 15 and 92.)

Kotte, I. (2002). *Een kartografisch expert systeem ten behoeve van presentatie van gedistribueerde geografische informatie*. M.Sc. thesis, Delft University of Technology. (Cited on page 29.)

Lester, J., Hurvitz, P., Chaudhri, R., Hartung, C., & Borriello, G. (2008). MobileSense - Sensing Modes of Transportation in Studies of the Built Environment. In *International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems - UrbanSense08*, (p. 5). (Cited on pages 1 and 17.)

Liao, L., Patterson, D., Fox, D., & Kautz, H. (2006). Building personal maps from GPS data. In *Annals of the New York Academy of Sciences*, vol. 1093, (pp. 249–265). (Cited on pages 16, 18, 27, and 45.)

Liao, L., Patterson, D., Fox, D., & Kautz, H. (2007). Learning and inferring transportation routines. *Artificial Intelligence*, *171*, 311–331. (Cited on page 16.)

Lucas, P. (2001). Certainty-factor-like structures in Bayesian belief networks. *Knowledge-based systems*, *14*(7), 327–335. (Cited on pages 30 and 92.)

Lüscher, P., Weibel, R., & Burghardt, D. (2009). Integrating ontological modelling and Bayesian inference for pattern classification in topographic vector data. *Computers, Environment and Urban Systems*, *33*(5), 363–374. (Cited on page 92.)

Maat, K. (2009). *Built environment and car travel: Analyses of interdependencies*. Ph.D. thesis, Delft University of Technology. (Cited on page 1.)

Maat, K., & Timmermans, H. (2006). Influence of Land Use on Tour Complexity: A Dutch Case. *Transportation Research Record: Journal of the Transportation Research Board*, *1977*, 234–241. (Cited on page 4.)

McGowen, P., & McNally, M. (2007). Evaluating the Potential To Predict Activity Types from GPS and GIS Data. In *Transportation Research Board 86th meeting*, (p. 21). (Cited on page 1.)

Ministerie van Verkeer en Waterstaat (2009a). Mobiliteitsonderzoek Nederland 2008. Tech. rep. (Cited on pages 23 and 80.)

Ministerie van Verkeer en Waterstaat (2009b). The kilometre price—Different Payment for Mobility. Brochure. (Cited on page 93.)

Misra, P., & Enge, P. (2006). *Global Positioning System - Signal, Measurements, and Performance*. Ganga-Jamuna Press. (Cited on page 6.)

Mountain, D., & Raper, J. (2001). Modelling human spatio-temporal behaviour: a challenge for location-based services. In *GeoComputation - Brisbane*, (p. 9). (Cited on page 27.)

Negnevitsky, M. (2005). *Artificial intelligence: A Guide to Intelligent Systems*. Pearson Education LImited. (Cited on page 30.)

Nickles, M., & Sottara, D. (2009). Approaches to Uncertain or Imprecise Rules - A Survey. In G. Governatori, J. Hall, & A. Paschke (Eds.) *Rule Interchange and Applications*, vol. 5858 of *Lecture Notes in Computer Science*, (pp. 323–336). Springer Berlin / Heidelberg. (Cited on page 30.)

Orchard, R. A. (1998). *FuzzyCLIPS Version 6.04A User's Guide*. Integrated Reasoning, Institute for Information Technology, National Research Council Canada. (Cited on page 31.)

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, *2*(11), 559–572. (Cited on page 15.)

Quenouille, M. (1956). Notes on bias in estimation. *Biometrika*, *43*(3), 353–360. (Cited on page 103.)

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, vol. 77, (pp. 257–286). (Cited on page 16.)

Ranjitkar, P., Nakatsuji, T., Gurusinghe, G., & Azuta, Y. (2002). Car-Following Experiments Using RTK GPS and Stability Characteristics of Followers in Platoon. In *Proceedings of 7th International Conference on Application of Advanced Technologies in Transportation Engineering*, vol. 245, (pp. 608–615). American Society of Civil Engineers, Boston. (Cited on page 1.)

Rearden, P., Harrington, P. B., Karnes, J. J., & Bunker, C. E. (2007). Fuzzy rule-building expert system classification of fuel using solid-phase microextraction two-way gas chromatography differential mobility spectrometric data. *Analytical Chemistry*, *79*(4), 1485–1491. (Cited on page 29.)

Reddy, S., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2008). Determining transportation mode on mobile phones. In *12th IEEE International Symposium on Wearable Computers*, (pp. 25–28). (Cited on page 16.)

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2010). Using Mobile Phones to Determine Transportation Modes. *ACM Transactions on Sensor Networks*, *6*(2), 13–40. (Cited on page 16.)

Rodrigue, J.-P., Comtois, C., & Slack, B. (2009). *The Geography of Transport Systems*. New York: Routledge. (Cited on page 1.)

Schüssler, N., & Axhausen, K. W. (2009). Processing Raw Data from Global Positioning Systems Without Additional Information. *Transportation Research Record: Journal of the Transportation Research Board*, *2105*(4), 28–36. (Cited on pages 13, 17, 38, 97, and 102.)

Shalaby, A., Tsui, A., Chung, E.-H., Byon, Y., & Abdulhai, B. (2006). New Tools for GPS-based Travel Surveys and Traffic Monitoring. In *New Frontiers in Transport Systems*, (p. 33). (Cited on pages 13, 18, 19, 63, and 89.)

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In R. B. Blue, & A. M. Rosenberg (Eds.) *Proceedings of the 1968 23rd ACM national conference*, (pp. 517–524). (Cited on page 102.)

Shortliffe, E., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, *23*(3-4), 351–379. (Cited on page 30.)

Shortliffe, E., Davis, R., Axline, S., Buchanan, B. G., Green, C. C., & Cohen, S. N. (1975). Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Computers and biomedical research*, *8*, 303–320. (Cited on page 30.)

Spaccapietra, S., Parent, C., Damiani, M., de Macedo, J. A., Porto, F., & Vangenot, C. (2008). A conceptual view on trajectories. *Data & Knowledge Engineering*, *65*, 126–146. (Cited on page 3.)

Stopher, P., Clifford, E., Zhang, J., & FitzGerald, C. (2008). Deducing mode and purpose from GPS data. Working paper ITLS-WP-08-06, Institute of transport and logistic studies, The Australian Key Centre in Transport and Logistics Management, The University of Sydney. (Cited on pages 17, 18, 42, 45, 63, 89, and 102.)

Stopher, P., FitzGerald, C., & Zhang, J. (2007). Search for a global positioning system device to measure person travel. *Transport Res C-Emer*, *16*, 350–369. (Cited on pages 17 and 94.)

Taghipour, S., Meybodi, M., & Taghipour, A. (2008). An Algorithm for Map Matching For Car Navigation System. In *22nd International Conference on Advanced Information Networking and Applications*, (pp. 1551–1556). (Cited on page 53.)

Taylor, G., Brunsdon, C., Li, J., Olden, A., Steup, D., & Winter, M. (2006). GPS accuracy estimation using map matching techniques: Applied to vehicle positioning and odometer calibration. *Computers, Environment and Urban Systems*, *30*, 757–772. (Cited on page 98.)

Teunissen, P., Simons, D., & Tiberius, C. (2008). Probability and Observation Theory. Lecture notes AE2-E01, Delft University of Technology. (Cited on page 42.)

The Economist (2009). The daily drudge. Issue Dec 14th 2009. (Cited on page 48.)

Van der Spek, S. (2010). Tracking tourists in historic city centres. In U. Gretzel, R. Law, & M. Fuchs (Eds.) *Information and Communication Technologies in Tourism 2010. Proceedings of the International Conference in Lugano, Switzerland, February 10–12, 2010*, (pp. 185–196). Springer Vienna. (Cited on page 9.)

Van der Spek, S., Van Schaick, J., De Bois, P., & De Haan, A. (2009). Sensing Human Activity: GPS Tracking. *Sensors*, *9*, 3033–3055. (Cited on page 9.)

Van Oosterom, P. (1999). Rule-based Polygon Classification of Topologically structured Topographic Data converted from Spaghetti Data. In *Computational Cartography, Dagstuhl-seminar, 19-24 october 1999*. (Cited on page 29.)

Van Oosterom, P., Tijssen, T., Alkemade, I., & De Vries, M. (2001). Multi-Source Cartography in Internet GIS. In *Proceedings 4th AGILE Conference, Brno*, (pp. 562–573). (Cited on page 29.)

Verbree, E., Maat, K., Bohte, W., Van Nieuwburg, E., Van Oosterom, P., & Quak, W. (2005). GPS-monitored itinerary tracking: Where have you been and how did you get there? *Geowissenschaftliche Mitteilungen*, *74*, 73–80. (Cited on pages 4 and 5.)

Weast, R. C., Selby, S. M., & Hodgman, C. D. (1954). *Mathematical tables from the Handbook of chemistry and physics*. CRC Press. (Cited on page 107.)

Wentz, E. A., Nelson, D., Rahman, A., Stefanov, W. L., & Roy, S. S. (2008). Expert system classification of urban land use/cover for Delhi, India. *International Journal of Remote Sensing*, *29*(15), 4405–4427. (Cited on page 29.)

White, C., Bernstein, D., & Kornhauser, A. (2009). Some map matching algorithms for personal navigation assistants. *Transport Research Part C*, *8*, 91–108. (Cited on page 53.)

Wilson, B. (2003). The AI Dictionary. School of Computer Science & Engineering The University of New South Wales. (Cited on page 30.)

Wolf, J. (2000). *Using GPS data loggers to replace travel diaries in the collection of travel data*. Ph.D. thesis, Georgia Institute of Technology. (Cited on page 2.)

Wolf, J., Guensler, R., & Bachman, W. (2001). Elimination of the travel diary: Experiment to derive trip purpose from Global Positioning System Data. *Transportation Research Record: Journal of the Transportation Research Board*, *1768*, 125–134. (Cited on page 2.)

Zadeh, L. (1975). The concept of a linguistic variable and its application to approximate reasoning–I. *Information sciences*, *8*(3), 199–249. (Cited on page 31.)

Zheng, Y., Chen, Y., Li, Q., Xie, X., & Ma, W.-Y. (2010). Understanding Transportation Modes Based on GPS Data for Web Applications. *ACM Transaction on the Web*, *4*(1), 1–36. (Cited on pages 14, 28, and 47.)

Zheng, Y., Li, Q., Chen, Y., Xie, X., & Ma, W. (2008a). Understanding mobility based on GPS data. In *Proceedings of ACM conference on Ubiquitous Computing (UbiComp 2008), Seoul, Korea*, (pp. 312–321). (Cited on page 14.)

Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008b). Learning transportation mode from raw GPS data for geographic applications on the web. In *International World Wide Web Conference: Proceeding of the 17th international conference on World Wide Web*, (pp. 247–256). (Cited on page 14.)

Zheng, Y., Wang, L., Liu, L., & Xie, X. (2009). Learning transportation modes from raw GPS data. United States Patent Application Publication 20090216704. (Cited on page 14.)

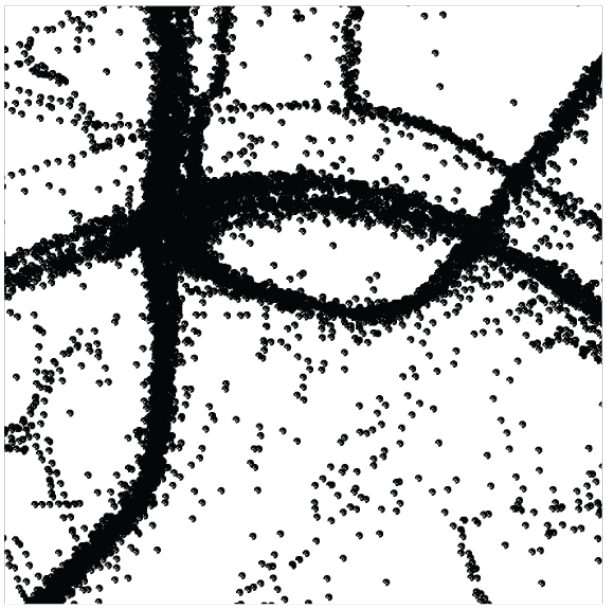
COLOPHON

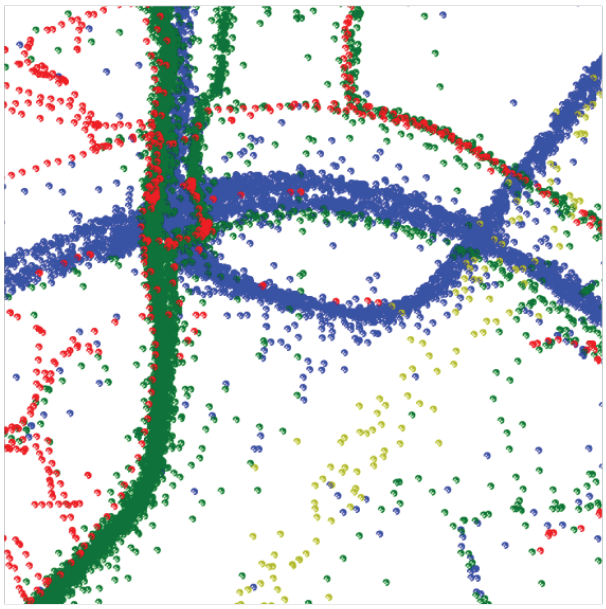
This thesis was typeset with LaTeX 2$_\varepsilon$ on Mac OS X using *Minion Pro* and *Euler* type faces. The code listings are typeset in *Bera Mono*.

The typographic style was inspired by Robert Bringhurst's book *The Elements of Typographic Style* and implemented by André Miede (available via CTAN as `classicthesis`), with some modifications by the author of this thesis.

The illustrations and diagrams were created in PGF/Ti*k*Z and Ipe, while the plots have been generated with Matplotlib, a plotting library for the Python programming language and its NumPy numerical mathematics extension. BibTeX was used to generate the bibliography.

The imagery in the figures is copyrighted by the referred institutions.





Supervised by Prof. dr. Peter van Oosterom
and Dr. Hugo Ledoux.

Department of GIS Technology
OTB Research Institute for the
Built Environment
Delft University of Technology
Jaffalaan 9
2628BX Delft
The Netherlands

Abstract

The knowledge of the transportation mode used in a movement trajectory (derived in form of timestamped positions) is critical for applications such as travel behaviour studies. This thesis presents a method for segmenting movement data into single-mode segments and their classification with respect to the used transportation mode.

The method relies on concepts found in expert systems, most notably membership functions, fuzzy logic, and certainty factors. A prototype, which may serve as a framework for managing travel behaviour surveys has been built in order to validate the presented theories and to classify the available test dataset. The transportation modes that this system classifies are walking, bicycle, tram, car, bus, train, underground, sailing boat, ferry, and aircraft.

This research also investigates the performance of OpenStreetMap data in solving this problem. This free source of geodata proved to be crucial for the classification, where the ten transportation modes are discerned with various indicators mostly derived from the geodata, for instance, the proximity of the trajectory to the tram network and the information whether the movement has been made on a water surface or not. The classification relies on eliminating unlikely transportation modes by values set with a number of empirically derived fuzzy membership functions, and by using the selected combination of indicators it is possible to distinguish in between transportation modes which exhibit a similar behaviour (e.g. a car and bus in urban areas). Finally, the classification results are attached with a certainty value. The results are supplemented with additional mode-related information, e.g. the name of the departure train station.

The segmentation has been done by detecting potential transition points between two transportation modes as brief stops. After each segment between consecutive potential transition points is classified, adjacent segments with the same classification outcome are merged (and removing the transition point in between), and keeping only the actual transition points where the transportation modes had been changed.

The method solves the problem with noisy data, and traffic congestions which bias the indicators by using additional statistic values. The classification of gaps in the data (e.g. caused by a signal shortage during the logging of a trajectory) derived satisfying results, and segments with only their starting and ending point have been successfully classified. Moreover, thanks to the availability of the OpenStreetMap data, the prototype is not restricted to trajectories acquired in the Netherlands, but it is also able to segment and classify trajectories acquired abroad.

The accuracy of the classification with the developed prototype, determined with the comparison of the classified results with the reference data derived from manual classification, is 91.6 percent.