Road Freight Transport Travel Time Prediction

Ksenia Sigakova
This thesis is submitted to the School of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Computer Science. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:
Author:
Ksenia Sigakova
E-mail: ksenia1986@inbox.ru

University advisors:
Johan Holmgren, Ph.D.
School of Computing

Gideon Mbiydzenyuy
School of Computing

School of Computing
Blekinge Institute of Technology
SE – 371 79 Karlskrona
Sweden

Internet : www.bth.se/com
Phone   : +46 455 38 50 00
Fax      : +46 455 38 50 57
ABSTRACT

Context. Road freight transport travel time estimation is an important task in fleet management and traffic planning. Goods often must be delivered in a predefined time window and any deviation may lead to serious consequences. It is possible to improve travel time estimation by considering more factors that may affect it.

Objectives. In this thesis work we identify factors that may affect travel time, find possible sources of information about them, propose a model for estimating travel time of heavy goods vehicles, and verify this model on real data.

Methods. Two main methods, used in this study, are the literature review and the experiment. The literature review resulted in a number of relevant articles found in scientific databases and few research reports published by the leading organizations in the area of Intelligent Transport Systems.

Results. The factors that may have influence on travel time and the possible sources of information about them were identified. The model for estimating travel time of heavy vehicles was proposed and verified on real data.

Conclusions. Experiments showed that considering time related and weather related factors, it is possible to improve accuracy of travel time estimation. Also, it was shown that the influence of a particular factor on travel time depended on the considered road segment. Furthermore, it was shown that different data mining algorithms should be applied for different road segments in order to get the best estimation.

Keywords: travel time estimation, Intelligent Transport Systems, data mining
ACKNOWLEDGMENTS

First, I would like to thank my supervisors, Johan and Gideon, for their support, advices and valuable feedback. Secondly, I would like to express my gratitude to the Swedish Transport Administration for kindly provided information about the road network and historical weather conditions. I am also grateful to the trucking company, as well as the fleet management company that provided historical truck traveling data. Thirdly, I would like to thank my friends for their help and understanding during all this time when I was working on this thesis project and for their support during the presentation. Finally, I would like to express my gratitude to my parents and my dear Tima, Pushinka, Murka and Vasya for their trust and love.
CONTENTS

1 INTRODUCTION ................................................................................................................................. 1
1.1 RELATED WORK.................................................................................................................................. 1
  1.1.1 Site-based measurements ............................................................................................................. 2
  1.1.2 Vehicle-based measurements ................................................................................................... 3
  1.1.3 Comparison ................................................................................................................................. 3
1.2 PROBLEM STATEMENT ...................................................................................................................... 4
1.3 AIMS AND OBJECTIVES .................................................................................................................. 4
1.4 RESEARCH QUESTIONS ..................................................................................................................... 5
1.5 RESEARCH METHODOLOGY ............................................................................................................. 5

2 FACTORS AND SOURCES OF INFORMATION ........................................................................... 7
2.1 COUNTRY AND REGION .................................................................................................................... 7
2.2 TYPE OF THE AREA (URBAN VS. RURAL) ................................................................................... 7
2.3 VEHICLE CHARACTERISTICS ...................................................................................................... 8
2.4 ROAD CHARACTERISTICS ............................................................................................................. 8
2.5 WEATHER CONDITIONS .................................................................................................................. 8
2.6 TIME RELATED FACTORS ............................................................................................................. 10
2.7 TRAFFIC CONGESTION .................................................................................................................. 10
2.8 ROAD ACCIDENTS .......................................................................................................................... 12
2.9 SCHEDULED ROAD CLOSURE ...................................................................................................... 12
2.10 UNSCHEDULED ROAD CLOSURE (OTHER THAN ACCIDENTS AND ROADWORK)... 12
2.11 TEMPORARY DIVERSIONS AND ROADWORKS ........................................................................ 13
2.12 DELAY ON CUSTOMS / BORDER CONTROL ............................................................................. 13
2.13 OTHERS ......................................................................................................................................... 13
2.14 CATEGORIZATION ............................................................................................................................ 13

3 TRAVEL TIME ESTIMATION MODEL .......................................................................................... 15

4 DATA FORMATS AND DATA MINING TOOLS USED IN EXPERIMENTS .............................. 17
4.1 FORMATS AND DATA PROCESSING TOOLS ........................................................................... 17
4.2 DATA MINING TOOLS .................................................................................................................... 18

5 IMPLEMENTATION .......................................................................................................................... 19
5.1 INPUT DATA FORMAT ...................................................................................................................... 19
  5.1.1 GPS data ..................................................................................................................................... 19
  5.1.2 Road network information ........................................................................................................ 20
1 INTRODUCTION

Intelligent Transport Systems (ITS) is a rapidly developing domain intended to improve transport situation in the world by applying new technologies from different research areas such as computer science, wireless communication, logistics and others. The main goals are to increase efficiency of traffic management and transport infrastructure utilization, make transport usage more convenient and secure, and at the same time alleviate negative impact on environment [7]. One branch of this domain is dedicated to the development of the area of road freight transportation. Here a lot of improvements could be done in fleet management and traffic planning. For example, it might be possible to improve accuracy of travel time estimation, the process aimed to predict how much time it is required for a vehicle to travel from a given point of departure to a destination. This is an important task as time becomes a crucial factor. In many cases, especially in food industry, goods must be delivered in a predefined time window [7]. Any deviation may cause serious problems for all sides. Moreover, increasing level of intermodal transportation requires even more accurate arrival time prediction, especially if some kind of scheduled transport such as ferry, train or airplane, is used.

One way to predict travel time with higher accuracy is to take into consideration more factors that may influence it. To be able to do this, it is necessary to combine data from different sources such as GPS measurements from probe vehicles, weather data, real-time information about congestion and accidents, forecasts from advanced weather warning systems and icy road warning systems, and others [7, 24]. Then, historical data can be used to analyse the influence of defined factors on travel time and to find possible patterns, while real-time data and forecasts will help to estimate and, if necessary, re-estimate travel time. The last can be useful as dynamic re-estimation of travel time that take into account all available information makes it possible to choose the most optimal route, while avoiding accidents and congestions.

1.1 Related work

There have been many research projects conducted in the area of travel time estimation. One review of travel-time prediction methods along with explanation of basic concepts and terms related to this area can be found in [6]. This article also names and describes two main categories of travel time measurement techniques, namely, site-based and vehicle-based measurements, data from which can be further used by other vehicles or dedicate systems to make predictions. A site-based measurement system is located in one specific place near a road and always provides observations of the same road segment, while a vehicle-based system is situated inside a travelling vehicle, therefore, it provides data from different road segments but only about this particular vehicle [6, 22]. One example of site-based system is a loop detector system that consists of two detectors, which are located at some distance from each other and register all vehicles passing by. By combining data from both detectors, it is possible to get information about each vehicle such as its length and speed, as well as general information about the road segment between detectors such as its occupancy, flow speed, and so on [21, 22].
These two categories of travel time measurement systems differ a lot in abilities, limitations, format and amount of gathered data. Thus, methods that use data collection techniques from one category are often not suitable for another. That is why this section has two subsections, first of which presents research projects that use site-based measurements, and the second - vehicle-based measurements.

1.1.1 Site-based measurements

Research works described in this subsection use data from site-based devices such as loop detectors, Automatic Vehicle Identification (AVI) reader stations, license plates recognition stations, and so on. One of the most popular application areas here is short-term prediction presented in [2, 3, 4, 5, 10, 26]. It assumes that a vehicle is travelling along a considered road segment departing from the start point at a known time $T_1$. The whole road segment is divided into a sequence of links by several measurement points where the vehicle’s speed can be calculated by using, for instance, data from loop detectors, automatic vehicle identification and so on. The main purpose of short-term prediction is to predict time required to traverse one or more subsequent links based on time observations from the preceding measurement points. Many different methods and techniques developed for such prediction are presented in scientific literature.

For example, in [3], the authors present counter propagation neural network (CPN) that uses travel time on $N$ (up to 10) previous links to calculate forecast for $N$ future links. The performance of developed CPN is compared with back propagation (BP) neural network. They also investigate the relation between the length of the forecasting time step (up to 30 min) and the average forecasting error. A similar approach can be found in [4], but here a spectral basis artificial neural network (SNN) is applied and compared with other methods such as historical average, a real-time profile, an exponential smoothing model and Kalman filtering model.

Another technique is described in [5], where the authors consider state-space neural network (SSNN) and claim that a traditional SSNN may give bad results as soon as some traffic conditions have changed from those in training set, thus, their research is aimed to investigate if it is possible to train SSNN in online mode, which implies that SSNN will be corrected each time new data instances become available. For doing this, they developed two SSNN models with two different online algorithms, each of which is based on the extended Kalman filter (EKF). Although the results show that performance of offline algorithms is slightly better than of online algorithms, in case when input-output pattern was presented in a training dataset, the authors name and discuss several advantages of their approach.

Although the abovementioned methods usually give quite accurate estimation, they have one common limitation, namely, they can be used only when a considered vehicle is already on road, thus, it is impossible to predict travel time before departure. It does not mean that these methods should not be used, on the contrary, they could provide good tools for analysis of real-time road situation and travel time recalculation, but there is a need for another kind of tools to be able to predict time of travels starting in some future.

One example of such methods is described in [26]. It is a linear regression method aimed to predict travel time in the nearest future that combines a historical mean and
the current status travel time. The last parameter represents time required to traverse a given road segment under current road conditions that can be estimated using data from, for example, loop detectors. The results show that such combination of current and historical traffic information lead to more accurate estimation. The same idea of combining historical and real-time data can be found in [2], but here another type of regression, namely, support vector regression (SVR), is applied in comparison with current travel-time prediction method that computes travel time from the real-time data and historical mean prediction method that computes average travel time of the historical traffic data at the same time of day and day of week.

1.1.2 Vehicle-based measurements

This subsection is dedicated to research works that use data from probe vehicles as input information for travel time prediction.

One example could be found in [8], where the authors present a dynamic travel time prediction model that is based on Kalman filtering and which takes historical and real-time data obtained from probe vehicles as input. The developed model is further used to compare link-based and path-based approaches. In the first approach travel time is calculated by summing up respective times on all links belonging to a considered road segment, while in the second it is calculated directly for the whole segment. The authors show that path-based approaches perform better than link-based approaches and discuss possible reasons behind this result.

In [17] the authors compare two models for travel time calculation that are based on GPS data: the speed model and the proportional model. In the speed model all instantaneous speed values from GPS data reported by a probe vehicle while it is inside a given road segment are averaged and travel time is calculated as a segment length divided by obtained average speed. At the same time the proportional model does not consider speed from GPS data at all. Travel time is calculated based on GPS timestamps and locations. The results show that the proportional model performs much better than the speed model. One of the found reasons for that was that in the speed model even a short stop or slow down may lead to significant reduction of calculated average speed if a vehicle reports GPS data at this moment.

Another interesting example is a recent research work presented in [9]. Here a route is considered as a set of stops and moves, which, in turn, are divided into segments specified by changing in a vehicle’s speed. Segment parameters are used to build a model for travel time prediction based on Support Vector Regression. As input for the model the authors use time, along with derived values for the hour of the day, the day of the month and the day of the week, and coordinates from GPS data, as well as a set of predefined parameters, namely, vehicle and driver identifiers and a region. As a result the model produces travel time estimation for a predefined vehicle driven by a given driver in a given region. In conclusion it is said that the model may give even more accurate estimation if more factors are considered.

1.1.3 Comparison

A good comparison of applicability of site-based and vehicle-based measurements in case of freight transport is given in [18]. The authors compare GPS derived
information with Freeway Surveillance (loop detectors) data. They argue that, as Freeway Surveillance systems report only average performance of all kinds of transport, it is a bad source of information concerning freight transport. One reason for this is that trucks, in most cases, move slower than usual cars, especially, if they are fully loaded. Another is that Freeway Surveillance system presents information about traffic situation for a whole road, without considering that conditions on different lanes can be totally different. This also could be a source of inaccuracy, as trucks usually travel on the right, “slow”, lane. Furthermore, trucks have different characteristics than passenger cars. They need more time to slow down and more time to speed up, they cannot easily change a lane and so on. Thus, mixing them with other cars leads to bad accuracy in estimation.

The report further states that although GPS data is more real, it shows individual truck performance that may also differ from general traffic conditions. To deal with this problem, it is necessary to have GPS data from several trucks operating on a considered segment. In conclusion, the authors discuss problems and possibilities of integration of data from different sources such as GPS information, loop detectors, and so on, as well as from different authorities and organizations. They show the need for such integration to obtain more accurate and reliable results.

1.2 Problem statement

Although there were many research project dedicated to travel time estimation, most of them consider only one source of data, for example GPS data or data from loop detectors. It was expected as this area is quite new and many possible sources of information appeared very recently. Nevertheless, authors of several scientific articles argue that combining data from different sources of information may improve accuracy of travel time estimation as more factors could be taken into account [9, 18]. Therefore, there is a need to develop a model for travel time estimation that can combine data from several relevant information sources. This thesis work is intended to address this problem.

1.3 Aims and objectives

The main aim of this project is to develop a model for estimating travel time of a heavy vehicle for a given road segment in a transport network.

This aim will be achieved by fulfilling the following objectives:

- identify factors that may affect the travel time of a heavy freight vehicle
- determine the available sources of input data about these factors
- develop a model for travel time estimation
- verify the proposed model on real data and analyse results
1.4 Research questions

In order to achieve the main aim described above, it is necessary to answer the following research questions:

**RQ1.** What are the main factors that may affect the travel time of a heavy vehicle?

**RQ2.** What types of input data sources could provide information about defined factors?

**RQ3.** How can data from defined sources be combined in order to build an appropriate model for estimating travel time of a heavy goods vehicle?

1.5 Research methodology

In order to answer the formulated research questions, it was decided to divide this thesis project into two parts. In the first part, RQ1 and RQ2 were answered by the literature review. In the second part, the experimental approach was used to answer RQ3.

As the area of Intelligent Transport Systems is not a pure computer science area, but includes some parts of transport studies, logistics and others, the first step of this research work was to obtain a good understanding of all aspects of this area in general, and, at the same time, to gather information about different kinds of factors affecting travel time and possible sources of information about them. In order to do this, a literature review was conducted. First of all, the search was performed among scientific literature from the database such IEEE, ACM and so on, but it appeared that there was not enough general information in scientific articles. However, the found articles provided references to relevant organizations and projects from the area of Intelligent Transport Systems. These references were used to find more specific publications and research reports. Furthermore, they helped to find several projects dedicated to implementation of systems that could provide additional input data for travel time estimation. As the amount of gathered information to that moment was enough to answer the first and the second research questions and other sources of information were considered as less reliable, it was decided to finish this cycle of the literature review.

The purpose of the next step was to find previous research work dedicated to travel time estimation. It would help to understand what was already done in this area, what kinds of input data, tools and techniques were used, and what was recommended for future research. This purpose was achieved by conducting one more literature review but using other key words. This time the amount of information found in scientific databases was sufficient.

Taking into consideration all gathered information, a travel time estimation model was proposed and the next step was to verify it on real data. During the second literature review it was found that many research projects dedicated to travel time estimation used the data mining approach as it provided many ways to analyse, process and explore big amounts of data, and, also, to build patterns, to find relationships between different concepts and to present results in appropriate forms.
Because of this, it was decided to use visualisation tools and data mining techniques in the current study as well, and the search for description of such tools and techniques was performed. Based on results of this search and available data, the scheme for estimation process was developed and two experiments were conducted according to this scheme. Since there were several appropriate data mining algorithms, it was decided to use all of them and then compare results. This also made it possible to understand if the same algorithm could be used for all road segments or it would be better to automatically choose the best algorithm for each particular segment.
2 FACTORS AND SOURCES OF INFORMATION

The first step in building a model for travel time estimation is to determine what factors may affect travel time. The list of groups of such factors, prepared based on the results of the literature review [6, 9, 10, 15, 16, 24], is presented below. Each group combines factors that have similar properties such as sources of information that can be used to get data about them. Also, each group is presented along with a short description, explanation of its importance for travel time estimation, examples of factors that it may include, and possible sources of information about these factors.

It should be noticed that all gathered factors most probably have some influence on travel time but this is not known for sure. It could be that some of them have effect only in certain combinations with other factors or even not have any effect at all. However, to understand that there is a need for further investigation, thus, they should be considered.

2.1 Country and region

Effect of the same set of factors may vary significantly depending on the country or a particular area. For example, heavy snowfall may not cause long delays in areas where it is a usually occurring phenomenon as there are enough snow removal machines and most drivers are used to this type of weather conditions. On the other hand, even light snowfall may stop all traffic in regions that usually do not have snow at all. Also, the time of reaction of emergency services in case of some problems on roads may differ depending on the country.

Possible sources of information: this factor is constant for each particular route and known in the planning phase.

2.2 Type of the area (urban vs. rural)

The next predefined factor that may affect travel time is the type of the area where a considered road segment is located. An urban area could be characterized by lower speed limits, a big number of traffic lights and pedestrian crossings, a higher level of road occupancy and traffic density. These are reasons for more frequent vehicle speed-down and stops, longer traffic queues, a higher number of road incidents and congestions. All mentioned above characteristics make it more difficult to predict the road situation in middle-size and big cities beforehand. Fortunately, most freight transport routes go around cities or on the outer parts of them, thus, the effect of such kinds of problems are usually alleviated for heavy goods transport.

Possible sources of information: this factor is constant for each particular route and known in the planning phase.
2.3 Vehicle characteristics

Vehicle characteristics may affect travel time in several ways. First of all, they determine a possible set of routes that can be taken between a point of departure and a respective destination. Sometimes the shortest route is not suitable for a particular vehicle because its characteristics do not meet the specific weight or height restrictions of some road segments along the route. This is especially a case in presence of tunnels or bridges. Furthermore, the same vehicle can either satisfy or not satisfy a weight restriction depending on its load.

Secondly, load, as well as a type of a shipment, directly affect speed. It is more likely that the driver of a truck with hazardous or fragile freight will drive more carefully and slower than if it would carry any other kinds of goods.

Examples: vehicle type (single trailer, double trailer), vehicle height and weight, load (full load / half load / empty), type of shipment (hazardous or fragile freight).

Possible sources of information: this factor is constant for each particular vehicle and known in the planning phase. Information about different kinds of restrictions imposed on vehicles could be found in official publications from The Road Administration. For example, weight restrictions applicable to the Swedish road network are described in the brochure “Legal loading” published by The Swedish Transport Agency [27].

2.4 Road characteristics

Characteristics of a considered road segment should also be taken into consideration. Besides obvious parameters, such as traffic amount and density, there are several other attributes that may influence travel time. One of them is a road category that defines if a road is a European highway, a national road, a primary county road or a secondary county road. This is an important parameter as most other road characteristics depend on it. Another attribute is the allowed speed limit that puts the upper bound on a vehicle’s speed.

Also, vehicle may spend time waiting on a railroad crossing, thus, in the case if there is one along the considered segment, knowing a schedule of trains crossing it could result in more accurate travel time estimation.

Examples: road category(European highway, national road, primary or secondary county road), bearing capacity class, speed limits, usual amount of traffic, frequency of accidents and congestions, presence of railroad crossings, tunnels, bridges or mountain passes.

Possible sources of information: this factor is constant for each particular route and known on the planning phase.

2.5 Weather conditions

Probably, the most obvious but at the same time critical factors leading to deviation in travel time are related to road weather conditions [6, 7, 10, 15, 24]. During normal conditions travel time is mostly defined by allowed speed limits, vehicle
characteristics and occupancy of a road segment, while bad weather conditions may result in a wide range of consequences from small reduction in speed to the full block of a road. Moreover, they may cause other problems such as accidents and congestions that, in turn, lead to further speed reduction and worsening of a road situation. Thus, weather related factors are very important for travel time estimation. Of course, it is unrealistic to consider all this great number of specific, sometimes highly correlated and difficult to be measured factors, especially taking into account that many of them are applicable only in some local areas and under particular circumstances, but even information about a limited set of the most relevant and easy-to-get factors could be very helpful in traffic planning and management as shown in [15].

The report [15] presents the results of a survey aimed to estimate relevance and quality of weather information resources used for decision-making by U.S. Department of Transportation personnel. In the survey the authors use concepts of Products and Product Components. One Product Component represents a particular weather factor such as current air temperature or historical dew point, while Products correspond to sets of weather parameters grouped by a source, time frame and locality, for example, Environmental Sensor Station (ESS) historical information or Road Condition Report. This division makes it possible to evaluate specific factors and at the same time give more general estimation such as how important forecast information is. Based on the results of a survey with 37 participants from 28 U.S. states, the authors conclude that the Pavement Weather Forecast in general has the highest level of importance. Among particular factors, the most relevant ones are precipitation type, precipitation start and end time, snow rate and snow amount. Although these results cannot be generalized to all countries, it is a good evidence of significance of weather information in traffic studies.

One more interesting idea that can be derived based on [15] is that to make more accurate travel time prediction, information about historical, current and expectable future weather conditions should be combined. The historical data could be used for statistical analysis of possible reasons for deviation in travel time with purpose to find out what weather factors most probably affect road situation in general and travel time in particular. After determining the set of relevant factors, their historical values could be used as an input to some estimation algorithm aimed to discover a relationship between weather and travel time. The obtained pattern will help to predict how long it will take to traverse a given road segment in case of forecasted weather conditions.

Also, historical weather forecasts could be compared with respective actual weather data to determine a level of possible deviation and its effect on accuracy of prediction.

Examples: type and amount of precipitation, wind speed, ground surface icing, pavement and air temperature, relative humidity, weather warnings, etc.

Possible sources of information: data from road weather stations for historical analysis, weather forecast from state or private meteorological agencies and information from weather warning system [7] for future prediction.
2.6 Time related factors

It seems that time related factors are the only ones that were considered in many research studies, for example in [9] and [10], and for which it is already proved that they undoubtedly effect travel time. It is known for everybody that, in most cases, there is more traffic during weekdays than on weekends, or that roads out of cities are busy on Friday evening during summer, especially in case of good weather. This knowledge could be summarized in weekly and daily traffic patterns that represent dependences between speed or travel time and the day of the week (for weekly pattern) or the hour of the day (for daily pattern).

It is also known that traffic conditions during holidays are quite different from normal. Nobody will expect much traffic even on main roads in big cities in the morning 1st January (of course, it is applicable only for countries that celebrate New Year at this time). Nevertheless, it is more difficult to understand how to deal with holidays, especially as some of them are celebrated each year on different dates. Furthermore, unusual traffic conditions could be observed in a day before and after holiday, which should be taken into consideration while estimating travel time. One way to do it is to use several Boolean parameters such as “a holiday”, “a day before a holiday”, and “a day after a holiday”.

Examples: the part of the day, the day of the week, season, holidays.

Possible sources of information: the part of the day, the day of the week and a season could be derived from timestamps in GPS data; information about official holidays in a particular country could be found in respective government regulations.

2.7 Traffic congestion

Although everybody seems to understand what traffic congestion is, it is not so obvious to give an explicit definition of it. Some people consider a queue of several cars stopping at traffic light as a congestion, while for others even long queues on all road lanes could be a good road situation just because they move a little bit faster than usually. So, definition of congestion could vary depending on a country, a type of the area (urban versus rural), characteristics of a particular road and so on. Some further discussion on this topic could be found in [16].

There could be many definitions of traffic congestion, but almost each of them follows one of three main approaches. The first approach describes congestion as a road situation when speed of vehicles is lower than the free flow speed that is the maximum allowed speed on a considered part of a road. In the second approach congestion implies a deviation from normal road conditions for the worse. The last approach introduces a new concept, a level of congestion, that indicates to what degree a traffic situation became worse in terms of car queues and speed reduction. This approach already has several implementations that are successfully used by drivers. One of them is a software system Yandex.Traffic developed by Yandex [19]. It works as follows. Drivers download the free mobile application Yandex.Maps that provides maps for navigation along with a lot of other useful information. This application has an option “Send traffic information”. If it is activated, then when drivers use Yandex.Maps, their GPS location, heading and speed are periodically sent to Yandex. Data from all drivers are
summarized and analyzed with purpose to determine levels of congestion on different roads. The results of this process is used to update information about congestions on Yandex.Maps in mobile devices, so, it becomes available to drivers almost in real-time. Figure 2.1 presents an example of how such a map with congestion information looks like. For more detailed information one should refer to the official description of this technology in [19].

Systems such as Yandex could be used as sources of real-time data about congestion. Furthermore, Yandex uses all collected traffic data to build a weekly congestion pattern and provides an additional service that allows to see usual congestion situations at any time of a chosen day of the week. This pattern could be very useful in travel time estimation.

Other companies, such as Google, Navitel, TomTom, Garmin and others, also have developed similar systems with congestion detection.

Possible sources of information: Queue warning systems [24], real-time data from other drivers, systems like Yandex.Traffic [19]. For recurrent congestions historical traffic data could be used to find a pattern. Also, results of traffic congestion studies in a considered area could be a valuable source of information.
2.8 Road accidents

This is a crucial factor in travel time estimation as consequences of road accidents may vary from a small congestion to a temporary road closure. At the same time, it is very difficult to predict accidents and their effect on the road situation and travel time. Nevertheless, it is possible to get information about road accidents in real time. This could help to avoid a place of accident if another route is available. Also, this information could be used to make re-estimation of travel time.

Possible sources of information: police reports, real-time data from other drivers.

2.9 Scheduled road closure

The road can be closed because of roadworks, a festival, a marathon or another special event. The main characteristic of scheduled road closures is that they are supposed to be planned, which means that information about them should be spread out beforehand and to all interested parties, including private drivers, transport companies, public transport users, and so on. The best way to do it is to use as many media as possible. This information could be published on a dedicated web page on the Internet, broadcasted on radio or television, announced by public transport drivers, and so forth. Also, it is necessary to plan alternative routes for all kinds of vehicles that might be affected.

Examples: roadwork, a festival, a sport event such as a marathon, a parade, a demonstration or another special occasion.

Possible sources of information: there is a need for a well-known and publicly available service that provides information about closed roads along with possible alternative routes for different kinds of vehicles in good time. As it is expected that all planned road closures are controlled by the road administration, it should be the primary source of such information.

2.10 Unscheduled road closure (other than accidents and roadwork)

The road can also be closed or blocked because of some sudden event such as a severe accident, pavement subsidence, the fall of a bridge, an avalanche or other natural disasters. In this case the most important thing that should be done first is to warn other drivers as such situations might be dangerous and the best advice for other vehicles should be to turn back and not approach the place of incident.

Examples: a severe accident, pavement subsidence, the fall of a bridge, an avalanche or other natural disasters, a blockage by a group of people due to a strike or a protest.

Possible sources of information: police reports, real-time data from other drivers, weather warning system [7] (if a road is closed because of natural disasters). In Sweden the Swedish Transport Administration provides this kind of information but not always in time.
2.11 Temporary diversions and roadworks

Maintaining a road in a good condition requires periodically roadwork. Main roads with high traffic amount and density require repairing almost all the time. As in most cases roadworks imply speed reductions, it is necessary to consider this in travel time estimation. A road may also be closed because of roadworks, but this case was already described above.

Possible sources of information: the road administration.

2.12 Delay on customs / border control

Although this factors could be included in the group “Road characteristics”, it was placed into a separate group because of special regulation applicable to vehicles passing by. This is especially relevant regarding the border between EU and non-EU countries or between two non-EU countries.

Possible sources of information: historical data, real-time information from other drivers.

2.13 Others

There exist many others factors that were not included in this list for lack of reliable information about them in literature or because no existing ways to get data about them or predict them were found. Examples of such factors include problems with a vehicle, driver behaviour, possible stops, waiting time on railroad crossing, and so on.

2.14 Categorization

All groups of found factors were categorized into three classes: predefined factors that are known on the planning phase, as well as predictable and unpredictable factors (see Table 2.1). As one can see, it is possible that some factors belong to several categories at the same time. For example, recurrent traffic congestion could be predictable, but it is sometimes impossible to predict congestion caused by an accident, roadwork, et cetera. Also, most roadworks are scheduled, thus, they should be known beforehand. Nevertheless, it could happen that a part of a road requires urgent repairing due to, for instance, pavement subsidence, that results in unscheduled roadwork.

The next section presents the proposed model for travel time estimation that takes into consideration the identified factors.
Table 2.1: Categorization of factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>Predefined</th>
<th>Predictable</th>
<th>Unpredictable / difficult to predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country and region</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of the area (urban vs. rural)</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Type and other characteristics of a vehicle</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road characteristics</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weather conditions</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Time of travelling (season, the part of the day, etc.)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic congestion</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Road accidents</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Scheduled road closure</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unscheduled road closure</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Temporary diversions and roadworks</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Delay on customs/border control</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
3 TRAVEL TIME ESTIMATION MODEL

Based on the detected groups of factors that may effect travel time, a model for travel time estimation was developed. Figure 3.1 shows its structure. Here the white boxes represent input data for the system and the gray boxes represent processes inside the system. “Predefined factors” represent all factors from the respective category according to the categorization from the previous subsection. The dotted line from the real-time information indicates that if this information is available, it is used as the additional input.

The model is supposed to be used in the planning stage to get estimation of travel time before departure. Furthermore, the same model can also be used for re-estimation, for example, if a weather forecast changes or if some additional information about roadwork, congestions, accidents, and so on, becomes available. As the real-time information is usually unknown until the last hours before departure, it is considered as the optional part.

The proposed model consists of two main steps. In the first step GPS data from previous trips are filtered by the values of the predefined factors. Then they are combined with respective weather information and other possible historical data to build a predictor with the help of data mining techniques. In the second step, the built predictor takes as input a weather forecast and the values of the predefined factors and gives an estimation of required travel time.

![Figure 3.1 Travel time estimation model](image-url)
In order to verify the proposed model, it was decided to implement it based on GPS data from probe vehicles combined with information from the Swedish road database NVDB, and data from the road weather stations (VViS). This implementation would also help to investigate different types of weather and time related factors. Furthermore, it may reveal hidden problems and opportunities that could be important for further research. The following sections describe this experiment part in detail.
4 DATA FORMATS AND DATA MINING TOOLS USED IN EXPERIMENTS

This section gives a short description of tools and algorithms used in the experiment part of this thesis project. According to the model presented in Figure 3.1, before getting estimation of time required to traverse a particular road segment, it is necessary to build a predictor, which will be specific for the selected road segment. This process could be divided into two main parts. The first part is dedicated to gathering required data from all available sources, filtering them according to given parameters and recording results in a special format. In the second part data mining tools and algorithms are applied to this resulting set of data in order to build a predictor. Throughout the rest of this section, the tools used in each of these two parts are introduced.

4.1 Formats and data processing tools

As it was mentioned above, in the first part all necessary information should be collected from different sources. To do this, it is necessary to understand file and data formats used for data transfer in each particular source of information. In this thesis project the following file formats were used:

- Tab Separated Value format. In this format each row in the file is corresponding to one instance of data or one record from a database. Values of parameters inside a row are separated by a tab spacing. The example of data in this format is given in Figure 4.1. This is a part of the file with GPS measurements used as input data in the experiments presented in Section 5.

- Keyhole Markup Language (KML) format. This is an XML based format used to describe geographic data for visualization in maps or Earth browsers such as Google Earth. It was adopted by the Open Geospatial Consortium (OGC) as an international standard for geographic visualization that could be used in all Earth browser implementations. The complete specification of KML format along with examples of its usage in Google Earth could be found in the official documentation, available at the Google Developers web portal (https://developers.google.com/kml/documentation) and the Open Geospatial Consortium web portal (http://www.opengeospatial.org/standards/kml).

The detailed description of data formats and sources of input data used in the experimental part of this thesis is given in Subsection 5.1.

Figure 4.1 Example of a file in Tab Separated Value format.
It could be seen in Figure 3.1 that one of the main sources of historical information for the proposed model is GPS data collected during previous trips. According to results of the literature review, processing and filtering of this kind of information could be done much easier with the help of some Geographic Information System (GIS). In [16, 18, 23, 25] one can find examples of usage of geographic information systems for traffic data interpretation, as well as discussions about advantages of this approach. Although there are some problems concerning processing GPS data in GIS such as map matching, inaccuracy in GPS data and maps used in GIS, different coordinate systems and so on, in [23] one can find successful solutions for these problems that are currently used in practice. Based on these results, it was decided to use Google Earth as an additional tool for data visualization.

4.2 Data mining tools

In the second part it was decided to use Weka, a software suit containing a collection of data mining algorithms and tools [20]. As the result should be a numeric value, it is necessary to use algorithms for numeric prediction. The following subset of such algorithms presented in Weka was chosen to be used in this thesis project (their complete description could be found in [20]).

- DecisionStump - a tree with the only node indicating the most significant factor
- M5 pruned model tree
- M5 pruned regression tree
- RepTree – a regression tree built using variance reduction and reduced-error pruning
- M5Rules – rules obtained from a model tree
- Linear Regression

This choice was made because it was important to understand how the considered factors affect travel time, thus, the obtained predictor should be easy to analyze. In future research other algorithms, such as neural networks, could be used as well.

After building predictors with the help of above listed algorithms, it is necessary to evaluate them in order to choose the best one that can be further used for prediction. Weka provides several ways to do it. First of all, input data could be automatically split into training set and test set according to a proportion defined by a user. It is also possible to manually select data for evaluation, put them into a separate file and create a “supplied test set” from this file. These methods are simple and easy to understand, but they will work well only if there is enough data available as both training and test sets should be representative in order to build a good predictor and give an accurate evaluation. Unfortunately, it is not always feasible to collect large amounts of data. In this case, other ways should be used. One of such way, provided by Weka, is cross-validation. In k-fold cross-validation input data is divided into k parts. Then a predictor is built and evaluated k times. Each time one of these k parts are used for evaluation and the rest of the data is used as a training set. In the end, results of k evaluations are summarized. Probably, the most common variant of cross-validation, namely 10-fold cross-validation, was used in this thesis.
5 IMPLEMENTATION

5.1 Input data format

The purposes of the experiment part of our project was to verify the proposed travel time estimation model, as well as to investigate what kinds of factors may affect travel time and how information about them can be used to improve travel time estimation. As it was unrealistic to consider all types of factors mentioned in Section 2 because of many reasons, such as the absence of sources of information about many factors, time limitation, and others, it was decided to restrict the investigation to time-related and weather-related factors, as they seem to be the most relevant ones according to the results of the literature review. Nevertheless, most methods and techniques developed in this project will work with other kinds of factors as well.

Considering the abovementioned restrictions the main types of input data required for this thesis work are as follows:

- GPS data collected by heavy goods vehicles
- Road network map
- Weather data

The subsections below will describe each of these types of data along with their sources, used in this project, in more detail.

5.1.1 GPS data

The first type of input data was GPS data collected by two trucks during a period of time from 26/01/2011 to 2/03/2012. The format of the data is as follows.

- Vehicle – id of a truck
- Javatimestamp– current time in an SQL timestamp format, i.e. in milliseconds since January 1, 1970, 00:00:00 GMT. For example, 1296003255 represents Wed, 26 Jan 2011 00:54:15 UTC.
- Longitude – longitude of a reference point
- Latitude – latitude of a reference point
- Altitude – altitude of a reference point
- Speed – instantaneous speed of a truck at the time when the GPS point data is recorded
- Heading – the current direction of a truck at the time when the GPS point data is recorded
The data is stored in tab separated value format in a text file. Each row in the file represents one GPS measurement. All measurements for each truck are ordered by the time when they were recorded.

5.1.2 Road network information

To be able to match a particular GPS position to a respective road (so called map matching), it was needed to find a source of road network information. As GPS data described in the previous subchapter were gathered by trucks operated in Sweden, it was necessary to find relevant information about Swedish roads. The source of this information should be accessible, updatable and reliable. The chosen National Road Database, NVDB (Swedish, Nationell VägDataBas), meets all these requirements.

NVDB is a database containing basic information about roads in Sweden, such as their location, names, speed limits, length and so on. The Swedish Transport Administration (Trafikverket) is responsible for maintenance and timely updates of the information. It is required that all changes in the real road network must be reflected in the NVDB within a predefined time interval. This gives the NVDB a great advantage over other road databases. Other advantages and possible usage of NVDB could be found in documents of the “NVDB Contents” series (available at the NVDB web-portal, http://www.nvdb.se).

Each road in NVDB is presented as a, so called, reference line that consists of a sequence of straight lines in three-dimensional space. Roads interconnection in NVDB is described in terms of nodes and reference links, where a node represents a real intersection, end of a road or another relevant point on a road and a reference link is defined as a road segment between predetermined start and end nodes (but it may contain other nodes as well). Each road segment in NVDB is stored along with its basic parameters (the format is presented in Appendix A.1).

![NVDB link 20450 consisting of 9 intervals as it is viewed in Google Earth](image)

![NVDB link 20450 consisting of 9 intervals as described by its coordinates in the NVDB KML file](image)
The geometry of a reference link is described as a curve consisting of a sequence of intervals presented by coordinates of their start and end points. An example of such description is shown in Figure 5.1. It should be noted that road parameters, including segment length, are stored only for a whole link, not for a particular interval.

5.1.3 Weather data

Weather data for our project was obtained from Swedish Road Weather Information Systems (Väg Väder informations Systemet, VViS). The data is collected from a set of weather stations adjacent to Swedish roads and stored in a tab separated value file. The format of data is presented in Appendix A.2.

For using this information as a source of historical weather data in our travel time estimation method, we need to be able to find a row in this file corresponding to weather conditions for a given road segment in a given time. As it could be seen, there is no coordinates or other information about location of a particular weather station except its name that usually indicates the nearest settlement. Thus, there is no easy way to automatically find a weather station that is situated close to a considered road segment. Although it could be possible to develop an algorithm solving this problem by geocoding, but it is not a trivial task and it is beyond the scope of this thesis work. So, it was decided for this project to find a respective weather station manually. This could be done by using the current version of the map with weather stations provided by the Swedish Transport Administration that is available on http://gis.vv.se/iov.

5.2 Input data processing

It should be noted that all information below is given in regards to the experiment part of the current project, i.e. we consider the Swedish road network as it is presented in NVDB and data from other sources described in the previous subsection.

Figure 5.2 Example of possible error caused by inaccuracy in GPS receiver
The next remark is that we consider all coordinates in two-dimensional space, i.e. we ignore altitude even if it is presented in data. The main reason for that is that altitude is not reliable, according to the information provided by the Swedish transport administration.

When we get new GPS data, they must be processed before being used for building a predictor. It is assumed that the route corresponding to a given GPS data is known, thus, it will be possible to filter road links from NVDB, to which considered GPS positions could belong. Here the filtering is a necessary process that enables to avoid many errors such as cases when a particular GPS position lies closer to the link representing the lane with opposite direction as it is shown in Figure 5.2. Here the red lines represent NVDB links and the yellow point represents a GPS position that was taken by a truck travelling on the right side of the road. As it could be seen this GPS position seems to be close to the left side of the road, thus, without filtering this will lead to an error.

After preparing the set of relevant links, it is necessary to process each GPS position and to find a respective NVDB road link, to which this GPS position belongs (map matching problem). This is done by using a geometric approach. Assume that a GPS position is a point and a road link is an interval in two-dimensional space. Then, the distance from a point to an interval can be found by the following algorithm.

In Figure 5.3a $P$ is a point and $AB$ is an interval under consideration. The distance $d$ from $P$ to $AB$ is equal to $d_0$, a length of an perpendicular $PH$, if $H$ lies inside the interval $AB$ (Figure 5.3 b). Otherwise, $d$ is equal to the minimum of $d_1$ and $d_2$, i.e., the lengths of the intervals $PA$ and $PB$ respectively (Figure 5.3 c).

**Step 1:** Determination if $H$ lies inside or outside the interval $AB$.

A linear equation of the line $AB$ that is given by two points $A$ and $B$ can be defined as

$$\frac{x - x_2}{y - y_2} = \frac{x_1 - x_2}{y_1 - y_2} \tag{1}$$

that in the slope-intercept form is

$$y = \frac{y_1 - y_2}{x_1 - x_2} \cdot x + y_2 - \frac{x_2(y_1 - y_2)}{x_1 - x_2} \tag{2}$$
or \( y = k \cdot x + b \) with \( k = \frac{y_1 - y_2}{x_1 - x_2} \) and \( b = y_2 - \frac{x_2(y_1 - y_2)}{x_1 - x_2} \).

Then, the equation of a perpendicular to this line through the point \( P \) is
\[
y = -\frac{1}{k} \cdot (x - x_P) + y_P
\]

(3)

As \( H \) is the intersection of \( AB \) and \( PH \), its coordinates can be found by solving the system of equations (2) and (3). The result is
\[
x_H = (y_p + \frac{x_p}{k} - b) \cdot \frac{k}{1 + k^2}
\]
\[
y_H = k \cdot x_H + b
\]

At the same time, the line \( AB \) can be presented in the parametric form as
\[
\begin{align*}
  x(t) &= x_1 + (x_2 - x_1) \cdot t \\
  y(t) &= y_1 + (y_2 - y_1) \cdot t
\end{align*}
\]

Then, putting coordinates of the point \( H \) in one of this equations one can find parameter \( t \):
\[
t = \frac{x_H - x_1}{x_2 - x_1}
\]

If \( H \) is inside the interval \( AB \), then \( 0 \leq t \leq 1 \).

Step 2: Distance calculation.

If \( H \) is inside the interval \( AB \), then the distance from \( P \) to \( AB \) is \( d_0 = d(P, H) \). Otherwise, it is the minimum of \( d_1 = d(P, A) \) and \( d_2 = d(P, B) \), where the function \( d(P1, P2) \), where \( P1(x_1, y_1) \) and \( P2(x_2, y_2) \) represent any two points, can be calculated as follows:
\[
d(P1, P2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}
\]

Step 3: Finding the nearest link

To find a link that is closest to a given GPS point, it is necessary to calculate distances to all relevant NVDB links, as it was described in steps 1 and 2, and choose the minimum among them. If this minimum is less than some predefined value, the respective link is announced as the nearest, otherwise, it is assumed that the given GPS position does not belong to a considered part of road network. The last situation may occur, for example, in presence of errors in GPS.

In conclusion, it should be pointed out that in this algorithm we use Euclidean geometry instead of spherical geometry, which is considered to be an acceptable replacement in situations when the considered area is small in comparison with the squared radius of a sphere that is the Earth in our case. This is the known fact about
spherical geometry (for complete explanation of this fact one should refer to the respective sources). To prove that it is applicable in our case, we conducted several experiments with hundreds of GPS positions and in all these cases the proposed algorithm found right nearest links. Furthermore, as a result of this phase, we create an additional KML file that displays GPS positions along with the found nearest links, that can be used to visually check correctness of results.

5.3 Experiment description

Before describing the proposed estimation process in detail, we would like to briefly outline its general scheme.

Having a considered road segment and a proposed time of travelling, the first step is to find relevant historical data, corresponding to previous travels along the considered road segment, in available GPS data. Further in this thesis such travels will be called “historical trips”. After that, to each trip we will add weather data corresponding to the date of travel. Values for some additional time related parameters will be derived from GPS timestamps. The resulting set of trips along with weather and time related parameters will be used as an input for the data mining part, in which several predictors are built and the best one is chosen. This will be the last step in the experiment part of this thesis, as it enables to verify the proposed model. When using the predictor for estimation of future travel time, the final step should be to put all known parameters about the considered future travel into the chosen predictor and get respective estimation.

Although the main aim is to estimate travel time $T_s$, the designed model will actually predict an expected speed $V$ on a considered road segment. After that the respective time could be easily calculated by the following formula:

$$T_s = D_s / V,$$

where $D_s$ is the segment length, which is equal to the sum of the lengths of all belonging to the segment links.

The reason for predicting speed lies in the characteristics of GPS measurements. As they are taken periodically, then, depending on this period, it may happen that lengths of found historical trips as well as their start and end points will vary significantly, thus, building a model based on travel time may lead to poor accuracy of estimation. Furthermore, speed is more universal concept and makes it possible to compare traffic characteristics on different road segments.

The following subsection provides full description of each step of the estimation process used in the experiments.

5.4 Estimation scheme

The process of travel time estimation is divided into several steps presented below along with expected outcomes of each step:

1. Choice of a road segment
In this step it is necessary to choose a segment of the road network, on which estimation will be made. At one time only one direction should be considered as the opposite side may have different characteristics such as speed limits, number of lanes and so on. If estimation is intended to be done as a part of another activity, for instance, a search of an optimal route, this step could be executed by a dedicated software program.

**Output:** coordinates of start and end point of a chosen segment, and a set of main numbers of all roads included into the segment.

2. **NVDB file filtering**

Here a respective NVDB file with road network information is processed to filter roads with given main numbers. As in our case NVDB information for each county is given in a separate file, it probably will be necessary to combine information from several files if a considered road segment lies in more than one county.

**Output:** filtered NVDB file in KML format that is used to visually check correctness of the filtering and a plain text file with information about chosen roads that is used as an input in further procedures.

3. **Finding a sequence of road links**

This task arises from the way how road links are presented in NVDB. Although each link has a unique identifier (the field ObjectID in an NVDB file), it does not imply that adjacent links are numbered in consecutive order, as shown in Figure 5.4. This makes it impossible to decide whether a particular link belongs to a considered road segment or not based on its identifier. At the same time it is necessary to get a sequence of road links representing the road segment in NVDB to be able to calculate a distance between GPS points, which is needed to evaluate the vehicle’s speed. The only found way to do it based on available information is to work with link coordinates.

As it was described in section 5.1.2, each NVDB link is presented by at least two points defined by their coordinates. Furthermore, if two links are adjacent, then the coordinates of the end point of one link must be the same as the coordinates of the start point of the other link. This fact serves as the basis for the proposed algorithm.

Assume that \( S_i \) and \( E_i \) are coordinates of the start and the end points of link \( L_i \). If \( L_i \) and \( L_j \) are adjacent, then \( S_j = E_i \). In the input file each row presents a link in the format “\( S_i \ E_i \ L_i \ C_i \)”, where \( C_i \) denotes some additional characteristics of the link. The task is, knowing \( S_1 \) and \( E_N \), to find a sequence \( L_1, L_2, \ldots, L_N \), where \( S_i = E_{i-1} \) for \( i \in [2, N] \) and \( N \) is the number of links in the resulting sequence.

With a close look, one can easily see that for this step the road network could be considered as a graph presented by a list of edges, where \( S_i \) and \( E_i \) are nodes (some of them coincide) and links are edges. In this case, the aim is to find a path between two vertices in a given graph. This is a well-known task and in this thesis it is solved by a depth-first search.
Figure 5.4. A part of E22 near Norjeboke as an example of inconsecutive numeration of links. NVDB links are coloured in blue, green and red. The numbers near links represent their identifiers.

Figure 5.5. The output file with the sequence of links for the example from Figure 5.4. The first column represents link identifiers and the second column represents lengths of links.
As a result of this step, we get a file with a required sequence of road links between two given points. The example of such a file for the case shown in Figure 5.4 is presented in Figure 5.5.

The only problem with the proposed algorithm appears in the presence of several routes between start and end points. To deal with this, we create an additional file in KML format that displays a found sequence of links. This enables to check if this sequence matches the considered segment or not. The last means that another route is found. In this case, one possible way to get an expected sequence is to delete any link that belongs to the found route but not to the segment and repeat the search.

**Output:** a text file with a sequence of NVDB links representing a considered road segment and a file in KML format that is used to display the found sequence to check its correctness.

4. *Finding relevant historical GPS data*

The purpose of this step is to find relevant historical trips. Then, each trip could be described by a set of GPS measurements taken from inside the segment. As the segment could be presented as a sequence of NVDB links, it is needed to find all sets of consecutive GPS measurements that were taken near some of these links. This is not a difficult task, as according to the input data processing algorithm described in Section 5.2, nearest links for all historical GPS positions should be found before this step.

5. *Speed calculation*

Now it is time to calculate average speed for each trip found in the previous step. This can be done by applying the following formula:

\[ V_{av} = \frac{D}{T}, \]

where \( D \) is the length of the part of a road between the first and the last GPS points belonging to a considered trip and \( T \) is the difference between timestamps of these GPS points. Further, the process will be explained with the help of the illustration in Figure 5.6. Here \( P_1 \) is the first GPS point of a considered trip and \( P_2 \) is the last point. \( L_0 \) and \( L_N \) are the nearest NVDB links to \( P_1 \) and \( P_2 \) respectively. \( E_0 \) is the end point of the link \( L_0 \) and \( S_N \) is the start point of the link \( L_N \). The dashed line represents a sequence of links between \( L_1 \) and \( L_{N-1} \). In this case, the length of the part of the road between \( P_1 \) and \( P_2 \) can be approximated to
\[ D' = d(P_1, E_0) + d(S_N, P_2) + \sum \text{length}(L_i) \mid i = 1, N-1, \]

where \( d \) is the distance in a spherical coordinate system and \( \text{length}(L_i) \) denotes the length of link \( L_i \) as given in NVDB.

Using the file with a sequence of road links representing the road segment in NVDB obtained in the previous step, one can calculate the last summand. For this it is necessary to sum up all lengths (values from the second column) starting from the row with \( L_1 \) and finishing on the row with \( L_2 \). For example, to calculate the length between the start of link 20567 and the end of link 316634 from the example illustrated in Figure 5.5, it is needed to sum up all values from the second column from 2nd to 6th rows of the file shown in Figure 5.6. The resulting length will be 314.478 metres.

The great circle distance, i.e. the shortest distance between two points over the surface of a sphere, in a spherical coordinate system can be calculated in the following way:

\[ d(P_1, P_2) = r \times \arccos(\sin(y_1) \cdot \sin(y_2) + \cos(y_1) \cdot \cos(y_2) \cdot \cos(x_1 - x_2)), \]

where \( x_1 \) and \( y_1 \) are the geographical latitude and longitude of \( P_1 \) and \( x_2 \) and \( y_2 \) are the latitude and longitude of \( P_2 \), \( r \) is the mean radius of the Earth (we assume that it is 6371 km).

Output: a file in Tab Separated Value format with information about all relevant historical trips including average speed on them.

6. Adding weather information

To add weather information to the found historical travel data, we need to determine the nearest weather station to a considered road segment as described in 5.1.3. Then all weather data obtained from this station should be put into a temporary file. As the size of this file is supposed to be much less than the size of the original file with data from all stations, this will significantly reduce the computation time required for further processing, where we need to match each particular historical trip to respective weather conditions. To do that, it is necessary to find a row of this temporary file corresponding to weather data measured at the moment that is closest to the time of a considered trip. This is not a difficult task as the time of measurement is presented in the weather information. So, we need only to find a row with the smallest difference relative to the timestamp of the first GPS point of the trip. After that, values of weather parameters from this row can be added to trip information.

Output: a file in Tab Separated Value format, each row of which represents a relevant historical trip along with respective values of weather parameters.

7. Data mining part

The main purpose of all previous steps was to prepare all required information about historical trips for the data mining part. In this step this information will be used as an input for some algorithm that will build a predictor for travel time estimation. As it was mentioned in Section 4, this will be done with the help of Weka, but before using it, it is necessary to choose a set of input attributes. This
was done based on results of the literature review and available information. As it was found in [9, 10, 51], time related parameters such as the part of the day and the day of the week often influence travel time. Also, weather factors may affect travel time significantly, as shown in [6, 10, 15], but in this case finding a relationship is a much more complicated task due to a big number and a high correlation between weather parameters. Nevertheless, it was decided to use all parameters presented in data from weather stations except “freezing point”, as there is no data for it. Afterwards, it will be possible to remove some of them if they are not needed.

Table 5.1 presents the preliminarily chosen set of time related attributes used for estimation along with their characteristics. As a description of weather factors can be found in Appendix A2 (Tables A.2 and A.3), here we just list those of them that are used for estimation and indicate their type (see Table 5.2).

Table 5.1: Time related attributes used for travel time estimation

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>season</td>
<td>a year season when the trip occurred</td>
<td>nominal</td>
<td>w – winter</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sp – spring</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sm – summer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>au – autumn</td>
</tr>
<tr>
<td>the day of the week</td>
<td>the day of the week when a vehicle entered the</td>
<td>nominal</td>
<td>0 – Monday</td>
</tr>
<tr>
<td></td>
<td>considered segment</td>
<td></td>
<td>1 – Tuesday</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 – Wednesday</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3 – Thursday</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 – Friday</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5 – Saturday</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6 – Sunday</td>
</tr>
<tr>
<td>the part of the day</td>
<td>the part of the day (in local time) when a</td>
<td>nominal</td>
<td>morning – 7 a.m. to</td>
</tr>
<tr>
<td></td>
<td>vehicle entered the considered segment</td>
<td></td>
<td>11 a.m.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>afternoon – 11 a.m. to</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5 p.m.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>evening – 5 p.m. to</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10 p.m.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>night – 10 p.m. to</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7 a.m.</td>
</tr>
</tbody>
</table>

Table 5.2: Weather related attributes used for travel time estimation

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>surface temperature</td>
<td>numeric</td>
</tr>
<tr>
<td>dew point</td>
<td>numeric</td>
</tr>
<tr>
<td>air temperature</td>
<td>numeric</td>
</tr>
<tr>
<td>code of precipitation</td>
<td>numeric</td>
</tr>
<tr>
<td>type of precipitation</td>
<td>nominal</td>
</tr>
<tr>
<td>amount of precipitation</td>
<td>numeric</td>
</tr>
<tr>
<td>maximum wind speed</td>
<td>numeric</td>
</tr>
<tr>
<td>average wind speed</td>
<td>numeric</td>
</tr>
<tr>
<td>wind direction</td>
<td>nominal</td>
</tr>
<tr>
<td>average wind speed over 30 minutes</td>
<td>numeric</td>
</tr>
<tr>
<td>warning code</td>
<td>nominal</td>
</tr>
<tr>
<td>air humidity</td>
<td>numeric</td>
</tr>
</tbody>
</table>
To be used as input, files with historical information should be processed according to this chosen set of attributes. Furthermore, in order to use Weka in a proper way, all missing or unknown values should be replaced by the symbol ‘?’ as in the weather data each parameter has its own specific notation for representation of unknown values.

After preparing input files, the chosen Weka algorithms, listed in Section 4.2, will be applied to build a set of predictors. As the amount of input information is limited, 10-fold cross validation will be used to evaluate the built predictors.

5.5 Experiments

Based on the developed estimation model, we conducted two experiments on different road segments, namely, the first experiment on the road E22 from Sölvesborg to Karlshamn, and the second one on the road 24 from Melbystrand to Hässleholm. These road segments were chosen based on available data and also because it is important to study different types of road segments in order to investigate what factors mainly influence travel time for heavy goods vehicles. The chosen road segments differ in several parameters. First of all, the road 24 is a national road, while E22 is an European highway (the corresponding parameter in NVDB is “Road category”). Also, functional road class of the road 24 is ‘2’ that indicates that it is less important road than E22, which has functional road class of ‘0’. There are some other differences, not presented in NVDB format, as well. Probably, the main of them is locations of roads. Road 24 is mostly inland road, while E22 is situated along the coast. This is an important distinction as it is expected that weather conditions and their effect on traffic situation are highly dependent on location of a road.

5.5.1 Experiment 1

For the first experiment we have chosen a part of the road E22 from Sölvesborg to Karlshamn (see Figure 5.7).

The NVDB characteristics of the chosen road segment are given below.

- Start link id: 21149
- Start point coordinates: 14.627863547376; 56.0673874332468
- End link id: 377508
- End point coordinates: 14.8210506374241; 56.1871652891096
- Number of links in the segment: 116
- Segment length: 21212.4 metres.

The proposed estimation model was applied to the chosen road segment according to its description with the only difference that in step 7, in the data mining part, we used 10-folds cross validation to evaluate the obtained prediction. During the first three steps of the proposed algorithm, no problems or incorrectnesses occurred.
During forth and five steps, 192 historical trips were found. They were analysed to find outliers and incorrect trips, to understand reasons behind incorrectnesses and to find a way to avoid such situations in the future. As a result, 174 trips were considered as correct trips, while 18 trips were rejected due to the following reasons:

- In 5 trips a vehicle traversed only a part of the considered road segment and then turned to another road. Most such situations could be avoided by using an additional test that checks if distances from the first GPS measurement to the beginning of the segment and from the last GPS measurement to the end of the segment are below some predefined threshold that is defined as the maximum distance that a vehicle can traverse during the period of time between two GPS measurements.

- 5 trips were rejected as it was concluded that a truck had a long stop. The easiest way to detect this is by looking closer on locations of GPS measurements on a map. If two measurements are situated at the same place or the distance between them is so short that even a vehicle moving with very slow speed should traverse greater distance, then it should be concluded that it was a stop. Such situations also could be found automatically by a software. Another question is how to deal with these cases. One solution could be to introduce an additional parameter in the characteristics of historical trips that indicates presence of a stop.

Figure 5.7. The chosen road segment for Experiment 1.
In 3 trips the GPS receiver did not work or was off during a period of time from 15 minutes to two hours. It also could be that the respective measurements were deleted. These situations can be detected and removed from a resulting set of trips automatically. To do this it is necessary to calculate the time difference between each pair of adjacent GPS measurements. If this difference is much greater than the period of GPS measurements, it can be concluded that some measurements are missing.

According to the conducted analysis one more conclusion could be made, namely that obtained average speed in the most incorrect trips was much lower than in correct ones. Thus, low speed could serve as an indicator of probable incorrectness and be used to filter such suspicious trips for further investigation.

The next step of the proposed model implies adding weather information to the found sets of trips. To identify relevant weather stations for this road segment, we used the map with VViS stations from the Swedish Transport Administration (http://gis.vv.se/iov) in combination with a map visualization of a Google Fusion Table that contains data from the weather file described in Section 5.1.3. The map visualization is done by using geocoding to find location of stations by their names. According to these maps, there is only one weather station along the considered road segment, which is located near Mörrum (see Figure 5.8), but this is only a “winter station”, which means that it does not operate during summer, thus, it is necessary to find another station nearby that operates the whole year. One of such stations is located near Kristianstad. The weather data from this station were examined and the results showed that although this station should operate the whole year, in reality it did not work from 19th October 2011 until 31st December 2011. Then it was decided to use weather information from the station in Kristianstad only in case when there is no data available from Mörrum.

Figure 5.8. The chosen weather stations in Mörrum and Kristianstad.
After completing this step, all information about chosen historical trips was used as input for the data mining algorithms described in section 4.2. As it was found during the literature review, most developed algorithms are usually compared with the historical mean prediction method that predicts average travel time of the respective historical traffic data [2, 4, 5, 26]. It was decided to do the same in this thesis. According to the results, all build predictors showed better performance, in terms of mean absolute error and root mean squared error, than ZeroR predicting the historical average (see Table 5.3). According to these results, the best predictor is the regression tree presented in Figure 5.9. Although it is not obvious why the speed is slower if the value of the air temperature is less than 14.05, the deeper analysis reveal some reasons behind it. If the air temperature is high, than it is more likely that it is summer time that means higher amount of traffic with more camping trailers that may result in more congestions, accidents and, consequently, speed reductions. Also, during summer time many drivers of heavy vehicles are on vacation and other drivers replacing them may not have enough experience.

The next best predictor, build by M5 Rules algorithm, is very similar:

\[
\begin{align*}
\text{IF air temperature } &\leq 14.05 \quad \text{THEN} \\
& \quad \text{speed} = +71.7626 \\
\text{ELSE} \\
& \quad \text{speed} = +68.4048
\end{align*}
\]

and the linear regression predictor is

\[
\text{speed} = 64.8441 + \\
2.2057 \times \text{Partday} = \text{night, morning} + \\
6.8582 \times \text{day} = \text{Sunday} + \\
-0.2648 \times \text{surface temperature} + \\
0.2188 \times \text{air temperature} + \\
-6.8601 \times \text{amount of precipitation} + \\
4.8881 \times \text{wind direction} = V, SO, S, NV, SV, NO + \\
3.1784 \times \text{wind direction} = NV, SV, NO
\]
Table 5.3 Performance of the chosen data mining algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZeroR</td>
<td>2.7829</td>
<td>3.5811</td>
</tr>
<tr>
<td>Decision stump</td>
<td>2.6476</td>
<td>3.5077</td>
</tr>
<tr>
<td>M5 pruned model tree</td>
<td>2.6743</td>
<td>3.5587</td>
</tr>
<tr>
<td>M5 pruned regression tree</td>
<td>2.5476</td>
<td>3.3897</td>
</tr>
<tr>
<td>M5 Rules</td>
<td>2.5617</td>
<td>3.3976</td>
</tr>
<tr>
<td>REPTree</td>
<td>2.7134</td>
<td>3.5403</td>
</tr>
<tr>
<td>Linear regression</td>
<td>2.5944</td>
<td>3.4532</td>
</tr>
</tbody>
</table>

According to the built regression tree and rules, the factor that influences travel time on this road segment most is air temperature. The decision stump algorithm also confirms it by building a tree that is very similar to the regression tree from Figure 5.9. At the same time, the linear regression produces a quite different predictor with many factors taken into account. One of the most important factors seems to be wind direction. If the relation between speed and wind direction put on a graph (see Figure 5.10), it is possible to see that, indeed, speed is higher in case of northwest, southwest and northeast wind. Another interesting factor is the day of the week. It follows from the built predictor that on Sunday speed is higher, but this is quite uncertain result as there was only one historical trip performed on Sunday. Although the average speed during this trip was high (see graph in Figure 5.11), it could be just a coincidence.

Figure 5.10. The relation between speed and wind direction in Experiment 1
5.5.2 Experiment 2

In the second experiment we consider a part of road 24 from Melbystrand to Hässleholm, which is presented in Figure 5.12.

The detailed NVDB characteristics of the chosen road segment are given below.

Start link id: 444374
Start point coordinates: 12.9650297359; 56.5040751382
End link id: 15679
End point coordinates: 13.685385534; 56.176748089
Number of links in the segment: 320
Segment length: 66440.43 metres.

Less historical trips were found on this road segment than on the part of E22 from the first experiment and there were more inappropriate trips caused by planned and unplanned stops, breaks in operation of GPS receivers and so on. The detailed description of them is given below.

There were 125 historical trips found. They were analysed with the help of techniques that were suggested based on the results of Experiment 1. For instance, situations when GPS receiver was off were detected automatically this time. Based on analysis results, 87 trips were considered as appropriate ones and 38 were rejected due to reasons listed below.
26 trips were rejected because a truck made at least one stop inside the segment. There were two types of stops. The first type represents planned stops near freight terminals, big warehouses and other places where freight is loaded on and off trucks. These stops are usually long, scheduled and known beforehand. Also, it is easy to see them by examining the GPS data. The other type is unplanned stops that occur because of some problems with a vehicle, when a driver wants to stop to make a call or to buy something in a shop, etc. It also could be a stop at traffic light or at a railway crossing, or in case of an accident. These stops are difficult or sometimes even impossible to predict. At the same time, all long stops could be easily detected by displaying GPS points of a particular trip in a program such as Google Earth and checking if there is any place that contains several points. If such a place is found, by zooming in one can understand what kind of place it is and determine possible reasons behind this stop.

In 13 historical trips some GPS measurements were missing, maybe, because of the GPS receiver did not work or was off during more than 10 minutes. Nevertheless, one such trip was accepted as an appropriate trip because its 15-minutes period without measurements was inside the considered road segment and the average speed was 75,68 km/h that eliminates possibility of any significantly long stop.
In the next step, adding weather information, there were similar problems to those, described in the first experiment. The nearest station was in Hässleholm (see Figure 5.13) but there were no data from 16th May until 1st October 2011. Although there were several more weather stations situated not so far from the road segment, all of them appeared to be “winter stations” and the nearest station operating the whole year was again Kristianstad. Thus, it was decided that the main source of weather information would be data from the station in Hässleholm and when there was no data available from it, weather data from Kristianstad would be used.

Table 5.4. Performance of the chosen data mining algorithms in Experiment 2

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZeroR (historical average)</td>
<td>2.0479</td>
<td>2.8664</td>
</tr>
<tr>
<td>Decision stump</td>
<td>1.7858</td>
<td>2.5113</td>
</tr>
<tr>
<td>M5 pruned model tree</td>
<td>1.8531</td>
<td>2.5939</td>
</tr>
<tr>
<td>M5 pruned regression tree</td>
<td>2.0199</td>
<td>2.8632</td>
</tr>
<tr>
<td>M5 Rules</td>
<td>1.8423</td>
<td>2.5892</td>
</tr>
<tr>
<td>REPTree</td>
<td>1.8009</td>
<td>2.5552</td>
</tr>
<tr>
<td>Linear regression</td>
<td>2.0127</td>
<td>2.8549</td>
</tr>
</tbody>
</table>
After finding all historical trips, information about them was used to build predictors. In this experiment again, all built predictors performed better than ZeroR in terms of mean absolute error and root mean squared error (see Table 5.4). However, two of the best predictors from the first experiment, namely M5 pruned tree and linear regression, in this experiment were the worst ones. Furthermore, the three best predictors look quite different from those in the first experiment.

The trees build by decision stump and REPTree algorithms are presented in Figure 5.14 and Figure 5.15. According to them the main factor for this road segment is the day of the week. From the graphical visualization presented in Figure 5.16, one can easily see the influence of the day of the week on speed. Moreover, the third best predictor, produced by M5 Rules algorithm, confirms this influence as well:

\[
\text{IF day is not Monday or Friday THEN} \\
\text{speed} = 73.7476 \\
+ 0.8278 \times \text{day} = \text{Friday} \\
+ 0.0458 \times \text{air temperature} \\
- 0.0492 \times \text{dew point} \\
- 0.1628 \times \text{maximum wind speed} \\
+ 0.1296 \times \text{average wind speed} \\
\text{ELSE} \\
\text{speed} = 74.1076 \\
+ 3.1965 \times \text{day} = \text{Friday}
\]

![Figure 5.14. The tree built by the decision stump algorithm](image)

![Figure 5.15. The tree built by the REPTree algorithm](image)
Figure 5.16. The relation between speed and the day of the week in the Experiment 2
6 EXPERIMENT RESULTS

This section consists of two parts. The first part presents a discussion about results obtained in the two conducted experiments and the second part raises some validity issues.

6.1 Discussion

Based on two conducted experiments, several interesting conclusions can be made. First of all, it was shown that all built predictors had better performance than ZeroR, representing historical average speed. Furthermore, some of the considered factors, such as air temperature and wind direction from the first experiment and the day of the week from the second, may have quite big influence on vehicle speed.

Another important conclusion is that algorithms that produce the best predictors are different for each road segment. Therefore, it seems to be more beneficial to use several different data mining algorithms and select the best one for every road segment than to use only one algorithm in all cases, especially as such selection could be done automatically by applying tools similar to Weka. Furthermore, some combination of the results of several predictors may provide even better estimations, thus, this can be a topic for future work.

Comparing factors used in resulting predictors in both experiments, it can be seen that in the first experiment the most important factors are related to weather, while in the second experiment they are related to time. It indicates that influence of weather conditions on vehicle speed depends on road location. In case of the road E22, which is situated near the coast this influence is higher than on the inland road 24.

6.2 Validity assessment

Although the conducted experiments showed quite good results, there are several possible sources of errors and uncertainty that should be mentioned. This subsection presents these sources of errors and explains how they were handled in the conducted experiments.

First of all, as the two considered road segments are quite short, 21 and 66 kilometers, it is necessary to mention how the proposed model can be applied if the road segment length is, for example, several hundreds kilometers. In this case it is recommended to divide the whole segment into shorter parts and estimate travel time on each of these parts. This will provide more reliable results as it is expected that more historical trips will be found for a shorter segment. Furthermore, it is a good way to deal with planned stops, just to divide the whole segments into parts between stops.

Next, according to the process for travel time estimation described in subsection 5.4, the first three steps are dedicated to preparing relevant information about the considered road segment. Assuming that information in NVDB is correct, the only error could be if a found link sequence does not correspond to this road segment.
Using visualization tools and GIS technology, such wrong sequences can be easily detected.

Possible problems arising during the fourth and fifth steps are related to preparing information about historical trips. Firstly, the formula for distance calculation in a spherical coordinate system used in step 5 in section 5.4 may give some rounding error. Nevertheless, as the calculated distance is quite small, this error should be small as well. There are several other formulas that can be used instead of this, but a thorough comparison is required to choose the best one, so, it was left for future work.

Another problem is that while detecting incorrect trips, it could happen that not all such trips are found. With more available information, for example, about planned and unplanned stops, accidents, errors in GPS receiver, and so on, results would be more reliable. This information could be provided by drivers or special algorithms could be developed to reveal such situations. Some ideas about these algorithms were already introduced in subsection 5.5. To detect as many incorrect trips as possible, in this thesis all found historical trips were analyzed manually.

One more moment that arouses concern is the choice of weather stations. In both of the conducted experiments one of VViS stations used as a source of weather data is situated at some distance from the considered road segment. Although weather conditions such as air temperature and wind direction are usually similar in a quite wide area, it is recommended for future research to find additional sources of weather information.

Finally, probably, the main validity threat was connected with the limited amount of historical GPS data that could lead to poor performance of the built predictor. In order to improve validity, it was decided to use several data mining algorithms in combination with 10-folds cross validation technique for evaluation. If most of the obtained predictors consider the same factor, it could be concluded with higher degree of assurance that this factor really has a big influence. In each of the two conducted experiments the significance of the most important factor, according to the best predictor, was confirmed by others. In the first experiment such factor was air temperature and in the second one it was the day of the week. Furthermore, the influence of the day of the week on speed was also confirmed by the graphical representation that made this result even more reliable.
CONCLUSIONS AND FUTURE WORK

This thesis project was aimed to develop a model for estimating travel time of heavy goods vehicles. In order to identify factors that may have influence on travel time, the comprehensive literature review was conducted. It also enables to determine possible sources of information about the identified factors that can be used as sources of input data. The resulting list of factors was the answer for the RQ1 and the list of sources answered RQ2. Based on the results obtained from the earlier phases of the study, the model for travel time estimation was proposed. In order to verify it, the model was implemented based on GPS data from probe vehicles, information about the Swedish road network, and weather data from the road weather stations. As this implementation combined the data from three different sources and provided the necessary output, it was confirmed that the proposed model for travel time estimation could be a good answer for the RQ3. Furthermore, the results of the experiments showed that, taking into consideration time related and weather related factors, it was possible to improve accuracy of travel time estimation. They also showed that for each road segment the most significant factors, as well as data mining algorithms producing the best predictors, could differ. This conclusion provides one direction for future work. It would be interesting to investigate different road segments, probably not only in Sweden but also in other countries, and for each of them to find factors, which affect travel time the most.

Two more possible topics for future work are connected to the investigation of other types of data mining algorithms, for example, neural networks, and to the selection of an appropriate set of attributes to eliminate correlated attributes such as summer time and air temperature from the first experiment.

Another direction of future work can be to examine the proposed model by adding more factors into consideration. Having an agreement with some fleet management company, it should be possible to get information about other predefined factors, for example, vehicles characteristics. Furthermore, as the area of Intelligent Transport Systems is currently developing, new sources of information about factors can become publicly available. Also, new factors, which were not identified in this thesis, may be revealed. Thus, there is a lot of room for future research.
## APPENDIX

### A.1 Road network format

Table A.1: NVDB format

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>Possible values</th>
</tr>
</thead>
<tbody>
<tr>
<td>ObjectID</td>
<td>an unique identifier</td>
<td></td>
</tr>
<tr>
<td>RouteID</td>
<td>an identifier of a reference link</td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>link length</td>
<td></td>
</tr>
<tr>
<td>From-date</td>
<td>the date when the information about the reference line was created</td>
<td></td>
</tr>
<tr>
<td>To-date</td>
<td>the date from which the information about the reference line became invalid (obsolete)</td>
<td></td>
</tr>
<tr>
<td>Bearing capacity class</td>
<td>permitted bearing capacity class</td>
<td>1 – BK1, 2 – BK2, 3 – BK3</td>
</tr>
<tr>
<td>Functional road class</td>
<td>a classification based on importance of a road</td>
<td>an integer value from 0 (the most important) to 9 (the least important)</td>
</tr>
<tr>
<td>The highest permitted speed</td>
<td>the highest permitted speed according to traffic regulations</td>
<td>5 (walking speed), 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120</td>
</tr>
<tr>
<td>Road width</td>
<td>width of the carriageway</td>
<td>a value in metres with one decimal accuracy, 1.0 – 40.0</td>
</tr>
<tr>
<td>Road manager type</td>
<td>type of organization responsible for road management</td>
<td>1 – national, 2 – municipal, 3 – private</td>
</tr>
<tr>
<td>Road category</td>
<td></td>
<td>1 – European highway; 2 – national road; 3 – primary county road; 4 – secondary county road; 5 – tertiary county road</td>
</tr>
<tr>
<td>Main number of a road</td>
<td>road number according to the county-wise compilation of public roads</td>
<td>1 – 99: European highways and national roads; 100 – 499: primary county roads; 500 – 5000: secondary and tertiary county roads</td>
</tr>
<tr>
<td>Road under number</td>
<td>road under number according to the county-wise compilation of public roads</td>
<td>an integer value, 0 – 99</td>
</tr>
<tr>
<td>County</td>
<td>the county to which the road’s administration belongs</td>
<td>an integer value, 2 – 25</td>
</tr>
<tr>
<td>Road traffic network</td>
<td>type of network: car network or bicycle network</td>
<td>1 – car network</td>
</tr>
<tr>
<td>Road class</td>
<td>type of a road</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 – highway;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 – expressway;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 – expressway, separated</td>
<td></td>
</tr>
<tr>
<td></td>
<td>carriageways;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 – four-lane road;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 – ordinary road;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 – ordinary road, separated</td>
<td></td>
</tr>
<tr>
<td></td>
<td>carriageways;</td>
<td></td>
</tr>
</tbody>
</table>
## A.2 VViS format

### Table A.2: Weather data format

<table>
<thead>
<tr>
<th>№</th>
<th>Name</th>
<th>Description</th>
<th>Value (examples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nr</td>
<td>code of a place from the field “Namn”</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Namn</td>
<td>location of a weather station</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Tidpunkt</td>
<td>time of measurement</td>
<td>2010-01-01 to 2011-12-31</td>
</tr>
<tr>
<td>4</td>
<td>Yttemperatur</td>
<td>surface temperature</td>
<td>[-19,6; 46,2] -99,9 or -99,6: No measurement</td>
</tr>
<tr>
<td>5</td>
<td>Dagpunkt</td>
<td>dew point</td>
<td>[-27,2; 16,9] -99,9: No measurement</td>
</tr>
<tr>
<td>6</td>
<td>Fryspunkt</td>
<td>freezing point</td>
<td>always -99,9: No data</td>
</tr>
<tr>
<td>7</td>
<td>Lufttemperatur</td>
<td>air temperature</td>
<td>[-24,6; 27,8] -99,9 or -99,6: No measurement</td>
</tr>
<tr>
<td>8</td>
<td>Kod</td>
<td>code of precipitation</td>
<td>see Table 5.3</td>
</tr>
<tr>
<td>9</td>
<td>Nederbord</td>
<td>type of precipitation</td>
<td>see Table 5.3</td>
</tr>
<tr>
<td>10</td>
<td>Mangd</td>
<td>amount of precipitation</td>
<td>[0; 44,4] -999: No measurement</td>
</tr>
<tr>
<td>11</td>
<td>Maxvind</td>
<td>maximum wind speed</td>
<td>[0; 1981,5] -99,9: No measurement</td>
</tr>
<tr>
<td>12</td>
<td>Medelvind</td>
<td>average wind speed</td>
<td>[0; 24,5] -99,9: No measurement</td>
</tr>
<tr>
<td>13</td>
<td>Vindriktning</td>
<td>wind direction</td>
<td>N, NO, NV, O, S, SO, SV, V, -9: No measurement</td>
</tr>
<tr>
<td>14</td>
<td>Medelvind_30_min</td>
<td>average wind speed over 30 minutes</td>
<td>[0; 18,8] -99,9: No measurement</td>
</tr>
<tr>
<td>15</td>
<td>Varselkod</td>
<td>warning code</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Luftfuktighet</td>
<td>air humidity</td>
<td>[1,1; 100] -99,9 or -99,6: No measurement</td>
</tr>
<tr>
<td>17</td>
<td>Felkod</td>
<td>error code</td>
<td>0,4,11,800,810,811,812,832,840, 860,870,890</td>
</tr>
</tbody>
</table>

### Table A.3: Types of precipitations

<table>
<thead>
<tr>
<th>Code</th>
<th>Type (original, Swedish)</th>
<th>Type (translation, English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-9</td>
<td>Givare saknas/Fel på givare</td>
<td>No measurement/error</td>
</tr>
<tr>
<td>1</td>
<td>Ingen nederbörd</td>
<td>No precipitation</td>
</tr>
<tr>
<td>2</td>
<td>Regn luft &gt;= -2</td>
<td>Rain, air temperature &gt;=-2</td>
</tr>
<tr>
<td>3</td>
<td>Regn, luft &gt;= -10 (ev underkylt regn)</td>
<td>Freezing rain, air temperature &gt;=-10</td>
</tr>
<tr>
<td>4</td>
<td>Snö</td>
<td>Snow</td>
</tr>
<tr>
<td>6</td>
<td>Snöblandat regn</td>
<td>Sleet</td>
</tr>
<tr>
<td>9</td>
<td>Okänd typ, osäker mgd</td>
<td>Unknown type</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

2.1  Yandex.Maps with information about congestions ........................................ 11
3.1  Travel time estimation model ............................................................................ 15
4.1  Example of a file in Tab Separated Value format ............................................ 17
5.1  An example of NVDB link .................................................................................. 20
5.2  Example of possible error caused by inaccuracy in GPS receiver .................... 21
5.3  Distance calculation ............................................................................................ 22
5.4  An example of inconsecutive numeration of links .............................................. 26
5.5  An example of the output file with the link sequence ......................................... 26
5.6  Representation of a route ...................................................................................... 27
5.7  The chosen road segment for Experiment 1 ....................................................... 31
5.8  The chosen weather stations in Mörrum and Kristianstad ............................... 32
5.9  The regression tree for Experiment 1 ................................................................. 33
5.10 The relation between speed and wind direction in Experiment 1 ................. 34
5.11 The relation between speed and the day of the week in Experiment 1 .......... 35
5.12 The chosen road segment for Experiment 2 ....................................................... 36
5.13 The chosen weather stations in Hässleholm and Kristianstad .................... 37
5.14 The tree built by the decision stump algorithm ................................................. 38
5.15 The tree built by the REPTree algorithm ......................................................... 38
5.16 The relation between speed and the day of the week in the Experiment 2 ...... 39
LIST OF TABLES

2.1 Categorization of factors ................................................................. 14
5.1 Time related attributes used for travel time estimation ......................... 29
5.2 Weather related attributes used for travel time estimation ................... 29
5.3 Performance of the chosen data mining algorithms in Experiment 1 .......... 34
5.4 Performance of the chosen data mining algorithms in Experiment 2 .......... 37
A.1 NVDB format data ........................................................................... 43
A.2 Weather data format data ................................................................. 45
A.3 Types of precipitations ...................................................................... 45
REFERENCES


