MASTER THESIS

INCORPORATING DRIVER PREFERENCE IN ROUTING

A REAL LIFE IMPLEMENTATION AND EVALUATION OF THE 'PERSONALIZED ADAPTIVE ROUTING ALGORITHM'.

STATUS: FINAL REPORT

Author: Ernst Jan van Ark
Student number: S1022970

July 10, 2013

University of Twente
Faculty of Engineering Technology
Master Traffic Engineering & Management
Centre for Transport Studies

Graduation committee

Dr. M.H. Martens
Faculty of Engineering Technology,
Centre for Transport Studies.
Dr. Ir. L.J.J. Wismans
Faculty of Engineering Technology,
Centre for Transport Studies.
Ir. W.P. van den Haak
TNO, research group Smart Mobility

UNIVERSITEIT TWENTE.

TNO innovation for life

PREFACE

This thesis represents a culmination of work and learning that has taken place over a period of over one year (May 2012 until July 2013). It marks the conclusion of my study Civil Engineering and Management at the University of Twente.

During the master Traffic Engineering & Management, I received the opportunity to participate in a variety of courses in which several state of the art facets of traffic and transport were introduced and elaborated. The topics ranged from the policy perspective, the development of traffic models towards the application of intelligent transportation systems. During one of the final courses I was challenged in a project in which an indicator framework was to be developed that could be used to observe, assess and reward the traffic safety performance within the driving behavior of users by means of a flexible insurance premium. During the course of the project it became apparent that user acceptance and the attitude from the driver towards a measure is a pivotal factor determining the success and failure of a proposed measure. Especially when developing and implementing a system in which the insurance premium is based on the safety performance, that seeks the boundaries in terms of individual privacy, the discussion concerning user acceptance becomes even more delicate and complex.

In my opinion the conjunction between the user and the traffic system is intriguing, in a world in which most traffic engineers tend to think in technical-physical solutions it is worthwhile to utilize a different viewpoint in which the user receives priority.

During one of the last theoretical courses of the curriculum I came in contact with the Smart Mobility department of the Netherlands Organization for Applied Scientific Research (TNO). During the initial conversations it became apparent that TNO is developing a smartphone application which aims too “relieve” the driver by minimizing the discomfort and surprises while travelling. Within the near future the application aims to supply tailor-made travel information which is relevant for the individual user and his current trip. Currently TNO is investigating the behavioral implications, the user acceptance and the personal attitudes towards the application and invited me, as part of my final thesis, to participate in this interesting and challenging field of research.

I would like to take this opportunity to thank all those who have contributed to this thesis. First and foremost I would like to thank my supervisors from the University of Twente: Dr. Marieke Martens, Dr. Ir. Jing Bie and Dr. Ir. Luc Wismans. Jing Bie has not been able to accompany me to the end of my thesis, but his support helped me to develop and crystallize my research. Luc has taken over the supervision of Jing, and despite the fact that Luc joined halfway through my research his comments and ideas have been of great importance. I also would like to thank my supervisor at TNO; Ir. Paul van den Haak. His support, enthusiasm and expertise offered additional insights, directions, tools and the positive energy to carry on.

Last but not least I would like to thank my colleagues, friends, family and everyone else who was willing to think along, to read documents and help me in any other possible way!

Every end is a new beginning. The conclusion of my study marks the first step towards a new stage of life with new opportunities and challenges, both professionally but much more important privately, to which I am really looking forward.

Ernst Jan van Ark, July 2013

MANAGEMENT SUMMARY

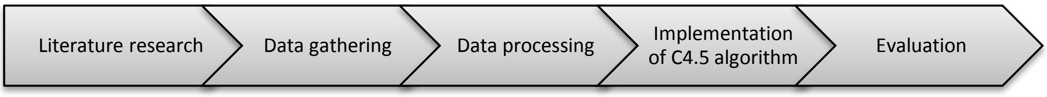
Transportation is a vital part of our economy; when designed, realized and utilized efficiently transportation systems provide economic and social opportunities. Based on this statement the 'management of the traffic operation' became a permanent discipline within the traffic and transport policies. The effectiveness of measures aimed at mobility management depends on the extent to which end-users are willing and able to assess and change their behavior. From a theoretical perspective the decision making process is often defined as a systematic assessment of the merits and value of objects. This assumption is often translated in random utility models which are based on the principles of micro-economics. However, in reality, humans often evaluate objects and decisions in an unsystematic manner that does not reflect the merits and value of the objects in an objective truthful and maximum utility manner. Choices will mainly be made based on ease, motive, comfort and cost but especially on emotion.

As an alternative, various methods in the field of knowledge discovery and data mining have been developed. These methods result in more flexible frameworks that are able to represent the effects of an attribute in comparison with random utility models. In the past thirty to forty years methods concerning Decision Tree Learning (DTL) algorithms have been developed. DTL algorithms are decision support tools that uses a tree-like graph to model decisions and their possible outcomes.

Especially in route choice algorithms it has been difficult to include the personal attributes within the traditional modeling approaches. In 2007 Park, Bell, Kaparias and Bogenberger stated that, due to more widespread usage of personal navigation devices, incorporating user preference within the route guidance is one of the most desired features to improve the user satisfaction towards navigation systems. By suggesting and simulating a learning model that employs the C4.5 DTL learning algorithm, the driver route choice behavior with respect to driver preferences is represented. Currently no real-life implementation of the decision tree learning algorithm has been implemented and evaluated to model the route choice in the field of traffic and transport. This master thesis aims to go beyond the scope and results from prior work by including non-simulated and real life user data to examine the regularities in user preferences within routing behavior.

The objective of this study is to contribute to the understanding of route choice modeling by implementing and evaluating a real-life framework, inspired by the research of Park et al. (2007), that includes the C4.5 decision tree learning methodology to integrate user preference in a so called 'Personalized Adaptive Routing algorithm'. Main aim of the adaptive routing algorithm itself is to aggregate criteria values to determine the relative weight of route attributes such as directness, familiarity, travel time, travel time reliability, aversion, complexity and travel distance. The algorithm will deduct these attributes from the historically observed route choice behavior and will apply the relations to improve the output of future route requests.

The objective above is translated towards the following main research question: **How can the C4.5 decision tree learning algorithm be applied and utilized to identify and integrate individual preference mechanisms within route choice algorithms and how does this algorithm perform in relation to traditional routing algorithms?**


Literature research
Data gathering
Data processing
Implementation of C4.5 algorithm
Evaluation

The research strategy comprised of five key stages, within the first phase a literature review has been conducted to gain insight in the theoretical background of traveler information systems, travel behavior and especially route choice. The second stage aimed to generate a large scale database which described the revealed travel behavior with respect to the trips that a user made. This database is generated based on the results of a field operational test in which a 'global positioning system' based data acquisition platform is employed to gather disaggregated travel data. In the third stage the data gathered from the second stage was processed, main aim was to process the raw data from the data gathering phase towards a structure that supports the implementation of the personalized adaptive routing algorithm. The aim of the fourth stage was to actually implement the decision tree algorithm based on the data structure that is derived from the third phase. This stage particularly focused on the comparison of the performance of the DTL based 'Personalized Adaptive Routing Algorithm' with the traditional shortest path and multi-attribute routing algorithm. The fifth and last phase focusses on the results in a broader perspective to make an inventory of the added value of data mining algorithms in the field of traffic and transport.

The data to facilitate the implementation of 'Personalized Adaptive Routing Algorithm' is supplied by the KATE mobile data acquisition platform, which is developed by TNO. The main element for this thesis is the location tracing algorithm which automatically records the time stamped local coordinates of the device. This algorithm dynamically utilizes various sensors (GPS, WiFi or network triangulation) and update intervals to preserve battery power while stationary and to improve the data quality while travelling. Between the 25th of September 2012 and the 31st of March 2013 95 users were equipped with a smartphone application that is based on the KATE platform.

Based on the locational traces, the timestamps and the subsequent distance travelled between these traces the individual trips were deducted from the data. The collected locational traces were snapped to the infrastructural network by means of a MapMatching algorithm which linked the locational traces with the infrastructural network. The total number of trips that was detected was 11.490 of which 1172 trips were made by train. The total distance travelled was 277.021 kilometers. The data is obtained without direct user input, so the trip data for each users represents his or her naturalistic travel behavior. Especially this last characteristic distinguishes the data from the KATE platform from the traditional methodologies based on travel diaries.

Within this thesis it was found that, based on the current technologies, we are able to cost-effectively gain insight in the travel behavior of participants. Although the group in this thesis was not representative for a large population, the results demonstrate the potential when the technology is scaled up and supplied to a broader user group.

The main input for the 'Personalized Adaptive Routing Algorithm' consisted of a set of maximally disjoints possible paths that represented the origin and destination of each trip that was detected. All paths have been evaluated in terms of travel time, travel distance, directness, complexity, travel time reliability, familiarity and aversion. These attributes are the main input for the learning process within the 'Adaptive Personalized Routing Algorithm'.

It is not possible to use the specifically observed absolute values for each attribute within the learning model. In the learning model therefore relative values of the attributes over the shortest route (in time) of each origin destination pair are used.

To generate the initial model the first set of 10 routes were selected as input to build an initial decision tree. Subsequently the remaining trips was used to test and update the model. By classifying each of the possible paths for each trip the model deducts the routes that resemble the personal attributes of each specific user. If the predicted route corresponds with the route that is actually taken by the user the model is accepted, if the predicted route does not correspond with the revealed route the model is updated. This process continues until all the route choice data for a specific user has been used. The process of applying individual trips can be regarded as a time series data of a driver that travelled, subsequently the model performance (the percentage of predictions that corresponds with the revealed route) represents the predictive accuracy of the 'Personalized Adaptive Routing Algorithm'.

The results of the 'Personalized Adaptive Routing Algorithm' in terms of predictive performance were compared with two traditional routing algorithms of which one is based on a single attribute shortest path (travel time based) assessment criteria and the other is based on a multi-attribute utility function.

The results of this study point out that, in its current implementation, the 'Personalized Adaptive Routing algorithm' is not able to achieve a higher predictive performance than the traditional shortest path algorithms. Based on the 3407 test trips the traditional shortest path algorithm achieves a 19% predictive performance while the 'Personalized Adaptive Routing Algorithm' achieves a predictive performance of 4%. The multi-attribute utility function scored a predictive performance of 0%.

One of the main factors that impeded the results of this study was the coherence between the revealed route and the set of possible paths. In almost 60% of the tests the routing algorithm failed because none of the proposed routes was similar to the revealed route. Moreover it was found that the route scores of each possible path varied only very slightly, the differences in the most important sections of the route on the underlying road network were averaged out as noise due to the disproportional distance on the high level network.

Subsequently a number of analyses have been applied to gain further insight in the learning behavior of the C4.5 adaptive routing algorithm. An apparent contradiction was identified; on the one hand it seems that the chosen methodology and architecture effectively describes the past (historic) behavior of the users but on the other hand the model fails to predict the future behavior of the users. One of the main factors that impeded the results was the imbalance in classes. While the algorithm had to classify 15 alternative routes the final test result of the implementation relied on the correct classification of one of these routes as the 'predicted' route (true positive) while the results of the routes that were correctly classified as 'un-attractive' were disregarded (true negatives). This assumption however represents reality in which the final user is only interested in the correctness of the predicted route which directly influences the satisfaction with the proposed system. The correct prediction of route that will not be chosen are not of any interest for the user.

In addition two different experiments were carried out to test the influence of individual attributes within the algorithm and moreover various pruning thresholds have been applied. During these tests no significant improvements in terms of predictive performance have been achieved. Based on these results we can conclude that further researches in alternative methodologies are perhaps more successful than optimization efforts of the current methodology.

In conclusion we can state that, although the results of this study did not substantiate a clear added value of the ' Personalized Adaptive Routing Algorithm', there is a clear ground for further research. Especially because the performance of all the models that were applied were limited we can conclude that the factors behind human decision making are clearly complex and are insufficiently integrated in the current routing algorithms.

The primary recommendation to further research the implementation of a DTL based routing algorithm is to split up the route prediction process in separate sections. By independently modeling the sections from the origin towards the motorway, the motorway itself and the section from the motorway towards the final destination the route is divided into pieces that are more similar in terms of road types and travel distances. Secondly it advised to improve the route generation algorithm; a link between this algorithm and the personal factors derived from the DTL algorithm, for example by forwarding locations that a user often passes, can improve the coherence between the revealed and predicted routes. Especially in combination with the segmented route prediction algorithm, the two proposed recommendations can reinforce each other to significantly improve the performance of the 'Personalized Adaptive Routing Algorithm'.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of abbreviations

AHP	– Analytical hierarchical process
ANP	– Analytical network process
ATIS	– Advanced traveler information systems
AVE	– Aversion
COMP	– Complexity
DIR	– Directness
DTL	– Decision tree learning
ECU	– Electronic control unit
ETP	– Enabling Technology Program
FAM	– Familiarity
FM	– Frequency modulation
GIS	– Geographic information system
GPS	– Global positioning system
ITS	– Intelligent transport systems
ITIS	– Intelligent traffic information systems
IMEI	– International mobile equipment identity
KATE	– Keen Android Travel Extension
KM	– Kilometer
MIN	– Minutes
MNL	– Mixed multinomial logit
OD	– Origin-destination
PND	– Personal navigation device
REL	– Reliability
SCM	– Sensor City Mobility
SP	– Shortest path
TD	– Travel distance
TIS	– Traveler information systems
TMC	– Traffic message channel
TNO	– Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek (Netherlands Organisation for Applied Scientific Research)
TT	– Travel time
UM	– Utility maximization

1. INTRODUCTION

This chapter introduces the subject of this thesis project, starting with a prologue in section 1.1. The problem definition is provided in section 1.2 which leads to the elaboration of the research relevance in section 1.3. In section 1.4 the project is defined and delineated by means of the research objective and the corresponding research questions. Lastly section 1.5 will provide an outline of this document.

1.1 PROLOGUE

Transportation is a vital part of our economy and impacts the development and welfare of populations. When transport systems are designed, realized and utilized efficiently, they provide economic and social opportunities which results in reinforcing effects such as an improved accessibility to markets, employment and additional investments (Rodrigue, Comtois, & Slack, 2009). Based on the statement above the "management of the traffic operation" became a permanent discipline within the traffic and transport policies which are applied integrally in coherence with other policy measures such as physical interventions.

The effectiveness of measures aimed at mobility management depends on the extent to which the end-users are willing and able to change their behavior. In practice it is difficult to take this behavioral aspect into account. Traditionally challenges and possible measures are approached technologically while assuming 'technology will do the job'. On other occasions policy makers assume that if the right conditions are shaped the expected behavior will manifest itself automatically.

In practice travelers make their own choices; this choice will mainly be based on ease, motive, comfort and cost (rationale considerations) but especially on emotion (irrational considerations). The perceived 'world of mobility' in which travelers make choices does not appear to be coherent and opportunities are not fully utilized. Mobility management strives to organize the decision framework in which users make their choices.

Due to the recent technological developments, more advanced and better tools became available at the disposal of the driver. Navigation systems (also known as personal navigation devices) have become an integral part of the modern equipment of vehicles and aim to assist the driver. Most navigation systems provide a single route based on a single attribute and mainly interpreted the attribute travel time or travel distance to determine the route advice. The question however is, especially in light of the statements above, whether a proposed route, based on travel distance or travel time represents the expectation of a user?

Knowledge related to route choice behavior is limited, it is however save to assume that every individual person is different and that every person approximates his route choice from a different perspective. Often these varieties are related to the attributes he (or she) takes into account. The development of large scale data-sources offers possibilities to better understand route choice behavior, the outcomes can subsequently be used to improve the quality of the route proposals that are generated by route planners.

TNO currently develops the KATE platform that aims to provide a 'research toolbox' to gather knowledge on travel behavior based on applications for mobile devices. Moreover this toolbox provides a communication platform to supply dynamic multi-modal travel information. Within this research the data derived from the KATE platform is utilized to investigate the personal attributes within route choice.

1.2 Problem definition

Traveler information systems are rapidly developing in all modes of transportation and were already recognized in 2002 as the social trend that will have the greatest influence on future transportation systems (Wachs, Social trends and research needs in transport and environmental planning, 2002). By providing pre-trip and en-route travel time information a significant impact on travel behavior could be achieved by enabling drivers to make efficient choices with regard to route choice, transportation mode and departure time. If people are actively triggered to evaluate the available travel options, the utilization of the transportation system can be improved by further incorporating traveler information systems.

Traffic information can be provided by variable message signs, navigation devices, radio broadcasting and dynamic real time travel information. The main disadvantage of these communication instruments is that it is difficult to specify or to segregate specific user groups. Due to the rapid development of (connected) smartphones and the increasing popularity of mobile applications new possibilities arise to more specifically target users and to adjust the information on these users.

A better understanding of the impact of information on travel behavior is a key issue in evaluating traveler information systems. Route choice behavior is a complex decision making process, which incorporates multiple objectives, factors and emotions. From a theoretical perspective the decision making process is often defined as a systematic assessment of the merits and value of object (Scriven, 1991). However this definition says more about what evaluations should be from a theoretical perspective than about how people actually evaluate objects or choices. It might be, that humans often evaluate objects and decisions in an unsystematic manner that does not reflect the merits and value of the objects in an objective truthful and maximum utility based manner.

Random utility models, which are based on the principles of micro-economic theory, have been widely applied in studies that describe travel behavior. The assumption that travelers are utility optimizers, which is adopted in random utility models, is often critically appraised by various behavioral scientists (Garling, Kwan, & Golledge, 1994). The research of Yamamoto, Kitamura and Fujii (2001) for example states that linear-in-parameters utility functions assume that the effect of attributes of an alternative are compensatory. This implies that an increase of one attribute can be compensated by a proportional decrease or increase of another attribute to yield the same utility.

As an alternative, various methods in the field of knowledge discovery and data mining have been developed. These methods have resulted in more flexible frameworks that are able to represent the effects of an attribute in comparison with random utility models. For example neural network models are able to detect non, compensatory relationships or synergy effects in the data, however main difficulty is that the results from these models are difficult to interpret and to apply in practice.

In the past thirty to forty years methods concerning Decision Tree Learning (DTL) algorithms, which fall in the category of knowledge discovery and data mining, have been developed and tested. A decision tree is a decision support tool that uses a tree-like graph to model decisions and their possible outcomes. The main advantage is that these tools allow the analyst to gain a clear insight on the structure of the revealed behavior. A first application of decision tree techniques in the field of traffic and transport can be found in Wets et al. (2000). This paper applied the algorithm to develop a model to represent mode choice. Furthermore Yamamoto et al. (2002) attempted to induce the mechanism of drivers' route choice from empirical data.

The amount of data that is available in our society is exploding. The so-called 'big data' is regarded as a powerful resource that is able to accelerate the development of improved services, customer awareness and productivity. Large companies such as Google already rely on large scale data mining applications for 15 years. Due to emergence of 'connected' devices together with the increasing computing power, the use of large-scale real time data becomes accessible in the field of traffic and transport.

The application of large scale data sources in traffic was recognized in 2007 by Park, Bell, Kaparias and Bogenberger. These authors stated that, due to more widespread usage of personal navigation devices, incorporating user preference within the route guidance is one of the most desired features to improve the user satisfaction towards navigation systems. By suggesting a learning model that employs the C4.5 DTL learning algorithm, the driver route choice behavior with respect to driver preferences is represented. C4.5 is an algorithm used to generate a decision tree and is developed by Quinlan (1933). The algorithm itself will be discussed in a future part of this report. However, since no real world data was available during the time of the research the authors relied on experiments based on simulation data that was derived from the simulation software suite ICNavs.

During the development and implementation of this this, no actual implementation of the decision tree learning algorithm has been implemented and evaluated to model the route choice in the field of traffic and transport. This master thesis aims to go beyond the scope and results from prior work by including non-simulated user data to examine the regularities in user preferences within routing behavior. The required revealed preference data will be extracted from the KATE mobile data acquisition platform. The regularities in user preferences will be incorporated within a user model based on the C4.5 decision tree learning algorithm. This results in an individual user model that can implicitly and automatically adapt its model output to the personal attributes of a specific user. By comparing the predictive performance of the adaptive model with two traditional (un-adaptable) routing algorithms this study aims to induce the mechanisms of the driver's route choice from empirical data without presupposing a strict and inflexible theoretical construct. By not defining an initial construct a broad perspective can be utilized which may lead to new insights that perhaps can be utilized to improve the theory based models.

1.3 RESEARCH RELEVANCE

As stated previously the main aim of this thesis is to implement an automated method that implicitly incorporates the regularities in user preference within a 'Personalized Adaptive Routing Algorithm'. Main element of the master thesis is the incorporation of a decision tree learning algorithm that induces the mechanism of drivers' route choice from revealed behavior. The relevance of this thesis can be approximated from two perspectives.

Scientific relevance

Data mining is a process which extracts implicit, previously unknown and potentially useful information from a large database (Shi, 2002). While the emphasis in transportation system analysis has shifted from aggregated models that describe large capital decision to disaggregated models of individual decision making that determine the transportation demand and supply, great efforts have been made to capture the structural and often causal relations that are inherent in behavior at the individual level. Discrete choice models that are used to describe, explain and predict choices between two or more discrete alternatives, have been developed to examine the behavior of individual decision makers that can be described as ‘facing a choice set which is finite, mutually exclusive and exhaustive’. Based on the theory of micro-economics the decision maker would obtain some relative utility from each alternative and ultimately would choose the alternative with the highest utility. Discrete choice models are powerful but complex. The art of finding an appropriate model for a particular application requires close familiarity with the phenomenon that is being studied and a strong understanding of the methodological and theoretical background of the model.

On the other hand, a decision tree represents the choice behavior as sequential examinations of attributes. Main advantage is that the analyst can gain a clear insight in the structure of the choice behavior being examined. The decision tree can for instance easily be converted to a set of production rules that represent the choice behavior by a set of if-then rules which determine the choice according to the conditions indicated by the sub-sets of attributes.

The structure of results derived from knowledge discovery and data mining methods is more flexible to represent the relationship between the attributes of the alternatives and the choice than (traditional) random utility based models. Due to this fundamental difference results derived from decision tree algorithms can potentially offer insights in route choice behavior which random utility models may be unable to reveal. These insights can subsequently be utilized to improve the knowledge concerning route choice and individual personal attributes which can be applied to improve the traffic management policies.

Societal relevance

From the societal perspective the added value of the incorporation of user preferences can be approximated by means of the concept of ‘user satisfaction’.

Since the introduction of route planners, an increasing amount of people have relied upon these applications for finding their way to local businesses and friends and to plan large distance trips. Although the available planners are becoming very reliable in terms of their input data in terms of infrastructural characteristics and information concerning the current traffic situation, they all mostly rely on fixed assumptions in terms of user characteristics and preferences.

In reality the assumption that every driver is ‘universal’ does not match the real world which is represented by a relatively low predictive performance of route choice algorithms compared to the revealed route (Park, Kaparias, & Bogenberger, 2007). Drivers may choose from a variety of routes between their origin and destination. Differences in knowledge and user preference may influence the distribution of users over a variety of possible routes.

Based on the discussion above it can be concluded that a spectrum of factors influences the drivers’ route choice, these factors are currently not yet adequately included in route planning applications. By incorporating and combining these considerations the predictive accuracy of routing algorithms can be improved which positively affects the usability and user-friendliness of personal navigation devices. Based on the assumption that the predictive performance of the route algorithm is the determinant factor for the user satisfaction, it can be concluded that an improvement of the predictive performance, expressed as the amount of route suggestions that match the observed route, will ultimately be reflected within the user satisfaction towards the route suggestion.

1.4 Objective and research question

The objective of this study is to contribute to the understanding of route choice modeling by implementing and evaluating a real-life framework that includes the C4.5 decision tree learning algorithm to integrate user preference in a so called 'Personalized Adaptive Routing Algorithm'. Due to the inclusion of the user preference it is envisioned that route suggestions more closely match the routes that are actually taken by the user. In this context 'more closely match' is defined as the performance of the adaptive routing algorithm in comparison with the traditional shortest path routing algorithms.

The personalized adaptive routing algorithm will be implemented based on revealed behavioral data (GPS logging data). The performance, expressed as the predictive accuracy of both the original shortest path routing algorithm and personalized routing algorithm, will be assessed and compared. This predictive accuracy is quantified by the proportion of trips in which the proposed route matches the revealed route.

Main aim of the adaptive routing algorithm itself is to aggregate criteria values to determine relative weight of route attributes such as directness, familiarity and travel distance. The algorithm will deduct these attributes from the historic observed route choice behavior. A set of possible routes between an origin and destination will be evaluated based on the relative weight of the criteria values and ultimately the route that best suits the user characteristics is selected.

The objective above can be translated into the following main research question:

How can the C4.5 decision tree learning algorithm be applied and utilized to identify and integrate individual preference mechanisms within route choice algorithms and how does this algorithm perform in relation to traditional routing algorithms?

This main research question is further differentiated in the following components:

1. How can route choice be described and what knowledge is currently available that describes the influence of personal factors and preferences within route choice?

   To fully comprehend the impact of user preferences in route choice and to understand how learning algorithm can be applied to support routing algorithms it is important to explore the possibilities, relevance and performance of the already available (traditional) routing models and algorithms.

2. Which previous initiatives have been undertaken to implement data mining algorithms in the field of transportation that aim to describe route choice?

   In this study the current and past initiatives that describe the implementation of DTL algorithms (both C4.5 based and broader) in route choice will be examined. Main aim is to derive the objectives and focus of these projects. The literature study within this thesis is not intended to be exhaustive but aims to gain insight in the opportunities, challenges and solutions that have been suggested. Based on this analysis we seek an overview of 'good practices'.

3. What input information, data sources and techniques are necessary to translate GPS derived floating car data towards a real life test implementation and evaluation of the personalized routing algorithm based on the C4.5 DTL algorithm?

   In the past a small number of researches, which will be further discussed in chapter two, have explored the application of learning algorithms in transportation, most of these researches approached the subject from a literature point of view or employ simulations to test the performance. Within this study a real life implementation is envisioned. To facilitate this implementation the theoretically oriented principles from previous studies should be translated towards a practically feasible implementation during this thesis.

4. How can the predictive performance of the DTL based adaptive routing algorithm be measured and how does the adaptive routing algorithm perform in comparison to two traditional routing algorithms (shortest path and multi-attribute utility maximization)?

   This thesis envisions to describe and demonstrate the added value of a personal (individual) approach within the generation of route proposals. However to prove the added value an evaluation framework is necessary to assess the performance. This evaluation framework should be able to quantify the difference between the adaptive and two traditional algorithms. One of the traditional routing algorithms will be based on the shortest path algorithm (travel time based) and the other will employ a multi-attribute utility maximization function. Based on this evaluation framework each routing algorithm will be evaluated and the differences in performance will be examined.

5. What are the future opportunities and challenges when applying large scale (big) data sources to investigate and explain personal route choice behavior?

   Although investigating and evaluating the performance of the Decision Tree Learning algorithm is useful, the true importance of this thesis lies in its possible added value in terms of scientific knowledge within route choice behavior. This thesis is one of the first that attempts to utilize high quality, large scale, GPS derived data to describe the behavior of people in traffic. It is expected that large scale data sources are becoming increasingly important, experiences and lessons learned from this thesis can be of added value for future initiatives.

1.5 Outline thesis

This report consists of 8 chapters. The past chapter introduced the research subject and presented the research objective and research questions. **Chapter two** will provide a state of art concerning the background of routing behavior, routing algorithms, learning algorithms and will more deeply discuss the past attempts to link routing behavior towards learning algorithms. The **third chapter** will discuss the experimental design by further discussing the data sources, data gathering procedures and the translation towards a system architecture that support the decision tree learning algorithm. **Chapter four** will discuss the data gathering and processing procedures. There are a number of steps that need to be taken to translate the GPS locational data towards data that describes the travel behavior which includes trip-, route- and travel mode information. Moreover the efficient handling of the large amounts of data within this study offers challenges in terms of processing time and complexity. **Chapter five** will discuss the general data analysis, main aim of this thesis is to utilize the GPS data to implement a decision tree learning algorithm. However to evaluate the value of the data-set it is important to also evaluate the general characteristics such as for example the number of trips, trip distribution and trip length distribution. Furthermore the demographic characteristics of the user group will be investigated. Within the **sixth chapter** the system architecture of the decision tree learning algorithm will be implemented and discussed, main element of this chapter is the comparison of the 'Personalized Adaptive Routing Algorithm' and the traditional shortest paths algorithm. This chapter will conclude with a general analysis in which all separate results will be integrated. **Chapter seven** will discuss the opportunities and challenges that were encountered during this master thesis program, main aim is to define relevant issues that are relevant for future applications of the decision tree learning algorithm and moreover this chapter provides a connection towards the conclusions and recommendations of this study. **Chapter eight** will discuss the results for the sub-research questions and will moreover discuss the results in respect to the main research question. Moreover this chapter will contain the main recommendations for future research.

2. BACKGROUND

Based on the previous introduction it is possible to define three major topics that have been discussed briefly in the introduction but require some additional explanation; the expanding role of technology in traffic, route choice behavior and identifying decisions structures underlying route choice behavior patterns. This second chapter will provide a 'state of art' based on these three topics.

One major trend that is currently dominating developments in the automotive sector is 'connected mobility'. Due to the rapid development and increasing penetration of handheld or in-car connected devices the technology within cars has created a growing market. The first section of this chapter will further investigate the role of technology in traffic by describing Intelligent Transport Systems and Traveler Information Systems.

The second topic that was briefly mentioned in the first chapter, and which is more complex to interpret, is the route choice behavior. Many aspects of travel behavior are of interest for behavioral analysis. General questions such as 'why do people travel' and 'where do people go' are critical for understanding the factors that affect the demand and which locations are affected by the demand. Other questions such as 'at what time do people depart', 'which route do people take' and 'which modality do people use' are related to trip specific information and allow us to analyze the specific infrastructural sections which are affected and moreover describe the effect of transport over a certain time. Route choice models play an important role in many transport applications, for example, intelligent transport systems, GPS navigation and transportation planning (Frejinger, Route Choice Analysis: Data, Models, Algorithms and Applications, 2008). What makes the analysis of travel behavior highly complex is that all questions above are interrelated. The second section of this chapter will focus on route choice behavior and aims to identify the aspects and modeling approaches that describe which route a given traveler would take to move from location A to location B in a given infrastructural network.

Route choice models are often based on the foundation of micro-economic theory; linear-in-parameters utility functions are applied within discrete-choice models. The assumptions underlying the random utility models are however often critically appraised by behavioral scientist (Garling, Kwan, & Golledge, 1994). In comparison to route choice models knowledge discovery and data mining methods have more flexible structures to represent the relationship between the attributes of a set of alternative routes and the revealed choice. Due to this characteristic learning models can potentially provide better behavioral insights that traditional models are unable to reveal. Based on this statement it is worthwhile to investigate the algorithms that are able to derive, validate and test the rules that describe choice mechanisms that generate observed activity patterns.

2.1 THE EXPANDING ROLE OF TECHNOLOGY IN TRAFFIC

INTELLIGENT TRANSPORT SYSTEMS

It is not necessarily cost-effective and moreover often physically impossible to increase the capacity of the available infrastructure to facilitate the demand during peak periods. Within the structural outline Infrastructure and Spatial Planning, in Dutch known as the 'Structuurvisie Infrastructuur en Ruimte', the Dutch government expressed the desire to achieve an equal balance between the infrastructural supply and the traffic demand (Ministerie van Infrastructuur en Milieu, 2012). Main aim is to improve the effective and efficient use of the available infrastructural capacity during the full day. It is assumed that a relative small reduction of the peak traffic load at specific traffic corridors can cause a significant improvement of the traffic flow and the perceived comfort.

The objective to jointly develop solutions and policy measures which aim to achieve a balance between the infrastructural supply and demand is one of the leading principles of the Intelligent Transport Systems (ITS). Main aim of ITS application is to apply computer, communication, information and vehicle-sensing technologies to coordinate transportation systems efficient and safely. ITS applications target transit systems as well as private transportation and the intended benefits of ITS systems are improved safety, improved traffic efficiency, reduced congestion, improved environmental quality, energy efficiency and improved economic productivity (Kumar & Singh, 2005).

TRAVELER INFORMATION SYSTEMS

Traveler Information Systems (TIS) are an integral component of the concept of ITS, these information systems are developed to enhance the personal mobility, safety and productivity of transportation (Mouskos, Greenfeld, & Pignataro, 1996). It is envisioned that travelers should be able to compare available transportation modes for a particular trip based on factors such as travel time, trip distance and costs. Moreover the service should be able to function as a clearinghouse for information about existing travel conditions such as road maintenance, congestion and the impact of incidents. Main aim is to provide a reliable source of both static and dynamic traveler information and to assist the individual traveler in being able to undertake and complete the journey while preserving the user satisfaction.

Although the definition above seems to indicate a desire for an integral information service, in practice it seems more difficult to develop a single data source or application that combines information services for route and modality choice. This separation can also be recognized within the available scientific literature. Many scientific papers which describe TIS implementations only provide an isolated view on a single means of transportation and especially the management of road traffic was pre-dominant in previous works. The research of Lyons (2006) shares this opinion and state that especially in the United States of America the word ‘traveler’ is synonymous with the term ‘driver’. The separation of mode choice and route choice is also reflected in the differences between the various definitions for the term ATIS and ITIS; several papers discuss Advanced Traveler Information Systems and some papers discuss Intelligent Traffic Information Systems. Different opinions exist concerning Integrated Traveler Information Systems and Intelligent Traveler Information Systems; the various terms such as traffic, traveler, advanced and integrated all illustrate the lack of a uniform detailed definition and implementation.

When traveler information systems were first developed the business model mostly relied on public sector agencies which took responsibility of all the aspects concerning the data collection. Within this period the information was mainly disseminated through public-sector owned devices such as information panels above and alongside the infrastructure. Moreover other mass media (both public and commercial) such as television and radio played an important role in relaying the information towards users.

During the development of the internet, online trip planners became widely available. These online planners are primarily designed to generate a proposed route based on the origin and destination which are manually selected by the users. These route planners provide only limited functionality in terms of their decision rules, most planners have a single attribute optimization function (shortest path or shortest distance) or use a rather limited decision rule involving user-defined criteria (avoid highways, toll roads and ferryboats). Moreover most of these route planners only incorporate static information in terms of travel time.

Within the last decade technology has become an integral component of our everyday life, this development also affected the transportation sector. Although one major manufacturer, TomTom NV, only introduced the first portable device aimed at the consumer market in the early months of 2004, today the navigation devices are a central part of our vehicles. The advantages towards users are evident. Users are able to reach their destination by means of the shortest en fastest route which results in less stress and exposure in traffic.

The functionalities of personal navigation devices (PND's) are continuing to develop; between 2005 and 2008 the next generation 'adaptive' navigation devices were developed. Based on the Traffic Message Channel (TMC) on the FM radio broadcasting bandwidth the devices were able to receive and process traffic information to determine the optimal route with more realistic travel time estimates. Main disadvantage of this technology is that information could only be supplied to the user group as a whole. Secondly the data stream only allowed one-way communication from the provider towards the user, this implies that is was not possible to examine the response of the user. Lastly also the available bandwidth was limited; in practice this meant that only periodic snapshots of the high level motorways were supplied.

Today consumers are uninterruptedly connected to the internet and currently manufacturers of personal navigation devices are working on integrating the connectivity within PND's to provide users with real-time information. The content of this information is often highly dynamic and its validity may change rapidly. In the past the data reliability, available bandwidth and technical aspects (e.g. processing power) were often limiting factors. Recent developments offer new possibilities to improve the usefulness and effectiveness of real time information within routing navigation devices. For example TomTom N.V. introduced the premium service 'HD-traffic' in 2008. Also Google has recently applied 'traffic' overlays in their map applications to visualize the dynamic traffic information. Although the penetration rates of these technologies is increasing the majority of users still relies on the traditional 'static' navigation devices.

2.2 Route choice behavior

One of the main issues that engaged traffic engineers and scientists for a long time was the question how to get and provide insight in the route choice behavior and the closely connected effects on the traffic flow patterns and costs on the network-level.

There have been many efforts to investigate the manner in which travelers decide which routes to consider and ultimately which route to use. Main focus of the available literature is to understand the decision mechanism that underlies the route choice behavior and to establish an appropriate modeling theory and modeling form. Two assumptions are recurring frequently in literature. The first assumptions states that route choice is often regarded as a two stage process. The first stage consists of a process that generates a 'choice set' in which the feasible alternatives are determined which are known and considered by the decision maker. Subsequently, as a second step, the decision maker adopts a choice criterion that eliminates the inferior alternatives until the best alternative is identified (Bekhor, Ben-Akiva, & Ramming, 2006). The other frequently recurring assumption is that, for simplicity and convenience, travelers choose the route that offers the lowest travel time or travel distance from a set of alternative routes. According to the research of Volpe, Lappin, Bottom, &

Gardner (2002) it is however not difficult to conclude that this is usually only an approximation of a more complex decision-making process.

Some researchers have put effort in obtaining a more detailed understanding of the method in which travelers consider an alternative set of routes and afterwards select one to travel from their origin towards the destination. For example Huchingons, McNees et al. (1977) and Ratcliffe (1972) assume that the choice is based on a set of coherent interrelated criteria. These authors deduct the route choice behavior by means of a statistical data analysis in which observed data is used to deduct the relevant choice criteria and their relative importance. Wachs (1967) employed a principal component factor analysis to determine the different reasons in which individuals explain their route choices. Additionally Heathington, Worall and Hoff (1971) estimated the probability of drivers that would divert to an alternative route when these drivers were supplied with information of the current traffic condition. This research was based on a questionnaire which was presented to 732 drivers. The responses were related to perceived attitudes toward diversion and were in part based on past behavior. The authors concluded that drivers are more likely to avoid delays on an outward journey from home to work instead of an inward journey from work to home. The respondents moreover stated that the motive for the route diversion was to avoid delays rather than to save travel time. Based on these results, quantified as the probability of diversion for drivers receiving information on traffic conditions, the paper concluded that a unique and worthwhile service could be provided by an information system operation on motorways. Pedersen (1998) employed a principal component analysis to deduct the influencing factors within the route choice which resulted in four orthogonal factors in the route choice of car users: safety, interest, purpose and hindrance. Moreover user profile analysis showed that the results for men and woman were not significantly different.

Several studies have aimed to model the route choice as a continuous variable in a variety of ways. Duffell and Kalombaris (1998) studied a variety of cases in Hertfordshire looking at how drivers choose either a ‘rat run’ or a main route on a network. In this context a rat run was defined as the usage of a minor road route as an alternative to a major road route. The authors concluded that travel time is the single most important criterion affecting driver route choice in networks where there is a viable alternative to the main route. Moreover the results indicate that drivers are willing to travel an increased distance if it will reduce their travel time provided that the travel distance is not doubled or the alternative is tortuous.

The research of Duffell and Kalombaris (1998) only approaches the route choice behavior at an aggregated level in which travel behavior is approximated integrally. Other models also exist in which the disaggregate choice analysis is approached by means of random utility models. Random utility models assume that drivers are utility maximizers and disaggregate the choice in various choice components. Each of the individual components is subsequently individually approached and operationalized. The methodology most widely used to operationalize random utility theory is discrete choice modeling which assumes that the probability of choosing a specific alternative is equal to the probability that the utility of that alternative is greater than or equal to the utilities of all other alternatives. The most important pillars in discrete choice models are multinomial and nested logit models which rely on simplistic assumptions. Main disadvantage of the random utility models is the ‘independence from irrelevant alternatives’ property which hampers the applicability to general route choice analysis and particular in urban road networks. The independence of irrelevant alternatives is a property which implies that the relative odds between two alternatives are the same no matter what other alternatives are available

A number of modifications to the basis multinomial logit specification have been proposed in alternative models which aim to address this problem, for example the nested logit model is developed for route choice modeling which avoids the independence for irrelevant alternatives while maintaining the relative simplicity.

One of the main complicating factors to apply discrete choice modeling methods is the very large number of practically feasible routes between most origins and destinations and moreover the complex overlapping characteristics of these routes. This characteristic is recognized by Ben-Akiva, Bergman et al. (1984) who proposed a two stage model structure. The first stage includes a labeling approach which aims to reduce the number of potential routes in which each of the routes reflects a criterion that might be relevant to route choice. These criteria are called "labels" and include: minimize time, minimize distance, maximize scenery along routes, etc. A generalized independence function is defined for each label which allows a network minimum path algorithm to build trees that are minimal with respect to the criterion. Within the second stage, a choice model from the set of labels is applied to predict the chosen route by means of a nested logit model.

From a cognitive point of view, anomalies have been discovered that appear to violate the crucial microeconomic underpinnings of discrete choice models (Walker, 2001). For example, and more towards the subject of this final thesis, the research of Volpe, Lappin, Bottom and Gardner (2002) discussed that, within the study of individual route choice behavior, it is important to capture the heterogeneity in drivers' preferences. This study emphasizes the significance of amorphous influences on behavior such as choice, knowledge and attitude. According to Volpe, Lappin et al. (2002) individual preferences across individuals result in differences regarding their response to alternative attributes. Furthermore the behavior of an individual can also manifest temporal variations. A logit model with fixed coefficients is not suitable to account for these variations. The authors state that, to accurately model the individual route choice behavior it is necessary to capture differences in intrinsic preferences and subjective evaluation of alternative attributes.

To include both individual and temporal variations the mixed multinomial logit (MNL) model is developed which captures heterogeneity among individuals and allows correlation over alternatives and time. According to the research of Volpe, Lappin et al. (2002) this generality comes at a cost; the choice probabilities cannot be computed analytically which translates in a requirement for large amounts of resources in terms of processing power. One MNL model was tested by Han, Algers et al. (2001) which proofed that a significant improvement in the statistical performance of the models was found by allowing the coefficients of observed variables to vary randomly across individuals.

Typology of users preferences

From a formal point of view and from common sense, in the context of unimodal point-to-point route planning queries, three families of user preferences may be distinguished:

- Spatial preferences,
- Spatio-temporal preferences,
- Global preferences.

Spatial preferences represent the largest class of preferences in terms of diversity. Every entity of the road network and moreover the environment can be concerned with such a preference, a preference is denoted as spatial when it concerns a geographic entity which has spatial coordinates. According to the research of Hadjali, Mokhtari and Pivert (2012) two approaches exists which designate road network elements concerned by a spatial preference. The first, the so called explicit approach, designates the different entities concerned by the preference by means of their inherent individual properties. An example of an explicit spatial preference is to avoid secondary or toll roads. The second implicit approach uses references to other spatial entities; by means of spatial relations the favored network elements are identified. For example the preference for a road which has multiple adjacent petrol stations is implicit. This implicit approach however needs further specification in terms of the zone in which a spatial entity is included or excluded, for example the distance between the road and petrol stations should be defined in which both are being considered 'adjacent'.

Time is an important factor in route planning because the duration of a trip is often a major factor to the satisfaction of the users. Therefore it is essential to the user to express the preferences regarding the temporal aspect. The spatio-temporal aspects both involve spatial and temporal entities, therefore this factor is rather complex. The research of Hadjali, Mokhtari and Pivert (2012) stated that the spatio-temporal preference P_{st} is formed based on a spatial preference P_s and a temporal preference P_t . In the context of a spatio-temporal preference this temporal preference component expresses the validity period of a spatial preference. This implies that the term expresses the duration or the timing of their event.

There are two types of temporal preferences involved in a route planning challenge: quantitative and qualitative. Quantitative preferences express absolute boundaries or restrict the temporal distance between two instances. For example the route between the original and destination should be less than 2 hours. A qualitative preference will provide a means to specify the relative position of an interval of two temporal entities. An example of a qualitative spatio-temporal preference is to arrive at the destination before nightfall. To process such a preference it is necessary to specify the term "night".

Contrary to spatial and temporal preferences which are specific and quantitative the global preferences characterize the route as a whole. This involves qualitative criteria such as cheapest, fastest, and safest. There are also other properties which can be of interest in specific contexts, for example emergency services can include robustness and freight services could require routes with a minimum dimension in terms of road-width and clearance-heights.

INCORPORATING USER PREFERENCE WITHIN ROUTE CHOICE

If all available literature concerning route choice is examined it becomes apparent that gaining insight in the influencing personal factors within route choice is difficult by user modeling due to the sheer amount of variables and influencing factors. This statement was also supported by the paper written by Park, Bell, Kaparias and Bogenberger (2007) which introduced a methodology to incorporate user preferences and attributes within the routing algorithms of navigation systems by means of a learning algorithm.

According to Park et al. (2007) a method utilizing a learning algorithm, which observes and processes the observed repeated route choice behavior, is a sensible way to acquire knowledge on users' preferences; due to the implicit approach the need for users interaction is reduced which minimizes any inconvenience.

2.3 Earlier attempts that studied the relation between user preference and route choice

Drivers generally rely on car navigation systems when they are not fully aware of the road information for their destinations. Based on an origin and destination, there are many possible and feasible routes which all have different characteristics. Depending on the trip motive and the user preference, the desired route can vary. Most of the currently available personal navigation devices present a single route based on static evaluation criteria, calculated in a predefined manner on the basis of mainly travel distance, travel cost, traffic information etc. Since 2008 TomTom N.V. distributes real time traffic information based on aggregated mobile phone data and data that flows back from their devices, however the market penetration of this technology is increasing but still limited compared to the full population of PND's. In 2010 the predictions for the American market were that 20% of the in-vehicle navigation systems sold in 2010 would include connectivity and that 6% of the handheld navigation systems would by connected to the internet by either an embedded modem or a tethered mobile device (Isupply Market Research, 2010). The suggested routes derived from static devices might not represent the user expectations and therefore be less attractive. Perhaps instead of presenting a single route, the presentation or processing of a set of possible routes with different tune-able characteristics, could be a desirable feature of a modern car navigation system.

There is one important research that will be providing the theoretical foundation for this thesis. This research, written by Park, Bell, Kaparias and Bogenberger (2007) introduced a methodology to incorporate user preferences and attributes within the routing algorithms of navigation systems by means of a learning algorithm. Main aim was to give the algorithm a scientific underpinning and moreover to test the methodology in a simulator study. Upon closer examination it became apparent that the assumptions of this research fitted well inside the original aim of this thesis due to the fact that it offered a specific and focused research subject.

The paper of Park et al. (2007) will play an important role in this thesis, many of the assumptions and principles that are made and discussed in this report are derived, translated and applied based on the methodology, results and recommendations of Park et al. (2007). To keep this report readable, compact and to prevent repetitions it was decided not to include a full summary of the paper of Park et al. (2007) within the main text of this report but to include it as appendix 1. In addition, below a brief overall framework will be sketched of other studies that have contributed to the scientific knowledge concerning personal characteristics in relation to the route choice behavior of drivers.

Both prior to and subsequent to the research of Park et al. (2007) other authors and agencies attempted to study the impact of user preferences in route guidance systems. The text below is derived from the research of Nadi and Delavar (2011) which conducted a large scale literature study.

There is a large variety of subjects and methodologies. Some researches, for instance Duckham and Kulik (2003), only modeled particular parts of the user behavior such as the desire to choose of simplest route with the fewest number of turns. A research to gain insight in the desire for a more reliable route-navigation with the least semantically equivalent junctions was described by Haque, Kulik and Klippel (2007). Some other researchers employed multi-criteria techniques to model a combination of user desires, for example

Sadeghi Niaraki (2008), Sadeghi Niaraki and Kim (2009) and lastly Stephanov and Smith (2009) used the Analytical Network Process (ANP) or the Analytical Hierarchical Process (AHP). The research Sedeghi Niaraki (2008) used an ANP model to calculate the relative influence among groups of road segment criteria. Within the paper of Sadeghi Niaraki and Kim (2009) a user-centric impedance model of each road segment was developed. The paper discussed the various efficient criteria that affected the personalized impedance modeling and then combined the criteria into a single impedance function that covered situational and personal user preferences in the route finding process.

Similar to the research of Park et al. (2007) also other researches utilized machine learning techniques to develop a personalized route planning model, for instance Rogers and Langley (1998) Rogers, Fiechter and Langley (1999) and Choi, Kim, Kang and Jeon (2008). Furthermore some studies found that asking the users about their preferences concerning route choice is the most direct methodology (Akasaka and Onisawa (2008), Volkel and Weber (2008) and Sadeghi Niaraki and Kim (2009)).

Golledge (1995) analyzed the important route selection criteria of humans during various experiments. Jozefowiez et al. (2000) studied the routing problems and paid special attention to the definitions, objectives and multi-objective algorithms proposed for solving these problems. Eksioglu, Vural and Reisman (2009) studied the taxonomic classification of vehicle routing literatures based on the type of study, scenario, physical, information and data characteristics.

Richter (2009) regionalized the environment by considering the strategies for each of the emerging regions to cognitively minimize the complexity of traveled routes and moreover presented a process to define context specific route direction that are memorable and include the spatial situations that the user can encounter during a route (Richter, Context-specific route directions: generation of cognitively motivated wayfinding instructions. Ph.D. Dissertation, 2007)

An ontology for modeling user and context using a vector of criteria and a methodology to calculate an overall preference of users in a tour planning application was described by Zipf and Jost (2006).

If all available studies, as described above, are combined and compared it becomes apparent that all studies can be differentiated in two categories. The first mainstream modeling approach aims to describe the route choice behavior from a theoretical perspective, by making assumptions a model is erected which is subsequently tested. The other mainstream of researches aims to gain insight in the influencing factors by means of revealed behavioral data or data derived from questionnaires (data driven-approach).

The data driven models can be separated in stated preference models, in which users are questioned or interviewed about the factors within their personal route choice, or revealed preference models in which the behavior of a certain subject is implicitly examined during a longer time period. Both stated and revealed preference methods have advantages and disadvantages, stated preference studies are often criticized because the behavior they depict is not observed and thus generally fail to take certain constraints into account. Moreover stated preference models are often hindered by the difference between observed and perceived attribute values. On the other hand revealed preference studies have high ‘face validity’ because the data reflects real choices and implicitly takes the constraints into account but on the other hand fail to correctly predict or model an attribute that is not in included in the dataset or out of range.

Main advantage of the first type of researches, theory driven approaches, is that the hypotheses are grounded in theory and that the certainty of the influence of each affecting variable is often higher. However data-driven techniques, which utilize real-life data related to the traffic systems, are appropriate when the understanding of the first principles of system operations is not comprehensive or when the system is too complex that developing an accurate model is prohibitively expensive. In these cases data-driven approaches are often quicker and cheaper.

The study of Zhang and Levinson (2008) described that, to operationalize the proposed theory of route choice, the perception and cognition processes for learning routes in a network, and route attributes must be explicitly modeled. The authors have derived a flow diagram that shows how a traveler makes a route choice decision given actual attributes of one or more routes. This flow diagram, as depicted in FIGURE 1, offers a good summary to combine all theoretical information that was proposes in this section.

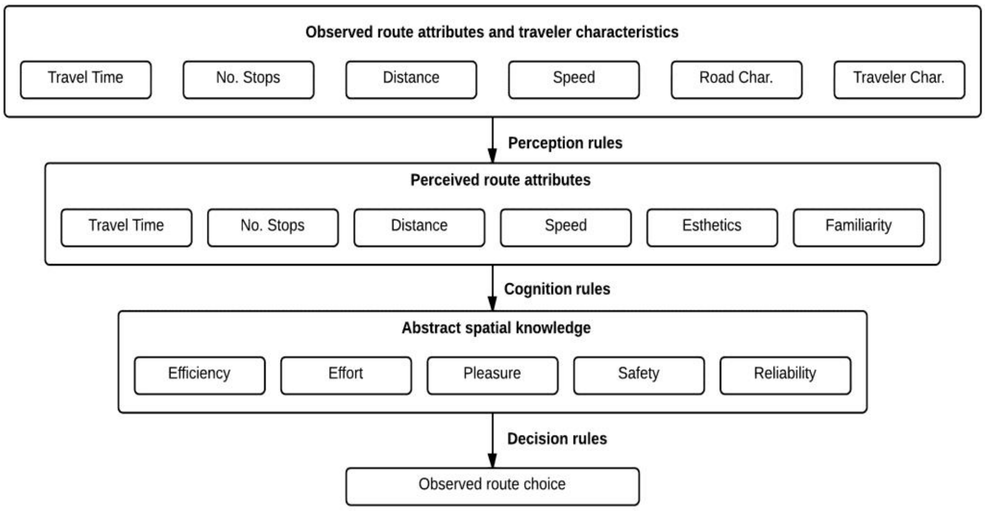


Figure 1: Route perception and cognition (ZHANG & LEVINSON, 2008)

Based on the theory described above, we can position both the data driven and theoretical driven approaches within the flow diagram. The revealed preference data models are located on top and describes the observed route attributes and traveler statistics, however these methods tend to relate the observed attributes directly to the final choice and ignore the perception and cognition processes. The stated preference data models are located in the middle of the figure, these models focus on the perceived influence of factors.

On the other hand the theory driven approach are located in the lower part of the flow diagram and aim to describe and quantify the 'efficiency' and 'utility' of a route. However these methods tend to focus on economic assumptions while disregarding the trip differentiating factors and personal characteristics

2.4 IDENTIFYING DECISION STRUCTURES UNDERLYING REVEALED ROUTE BEHAVIOR PATTERNS

Over the past 15 years, activity based transportation models have been an important instrument for modeling travel behavior. Most models are based on the assumption that travel demand is derived from the activities that individuals and households need or wish to perform, this implies that travel is not approached from an isolated point of view but that travel is a way to participate in activities and to realize goals in life.

Great progress has been made in developing frameworks that aim to explore, understand, analyze and predict the individual choice behavior; this resulted in the operationalization of several models within the field of traffic and transport. The multitude of these modeling attempts can be divided into two approaches. Firstly we can identify the discrete choice utility-maximizing models that were discussed in the previous section. Based on the statement that utility-maximization do not necessarily reflect the behavioral mechanisms underlying travel decisions, an alternative approach has been developed which is based on the desire for rule-based computational process models.

Secondly; data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can for instance include statistical models, mathematical algorithms and machine learning methods. Although the term suggest otherwise data mining goes beyond the collection and managing of data, it also includes the analysis and prediction.

When relating to travel, and more specifically routing behavior, the focus on patterns and relationships derived from revealed behavioral databases may provide new insights and relations. The choice to determine whether a route is to be advised to a user can be regarded as a binary classification; a route can be labeled either positive (attractive) or negative (not attractive).

A classification model is useful in two purposes, firstly a classification model can serve as an explanatory tool to distinguish between objects of different classes (descriptive model). Secondly a classification model can be used to predict the class label of unknown records, which is particularly interesting within this master thesis. A classification model is particularly suited for predicting or describing data that is based on a binary or nominal trainings set. The techniques are less effective for ordinal categories (large, medium or low) and furthermore relationships among categories are ignored.

Chen and Mizoguchi (1999) emphasize that a learning system is to be considered ‘intelligent’ if the learning system can adapt its task to the learning content based on a learner model, so this learning model can be regarded as a very important part in intelligent learning algorithms. The learning model can be regarded as a function from the input features to the target features. Many of the learning methods differ in what representations are considered for representing the function. This section will consider the decision tree learning algorithm. This algorithm is one of the most widely used models which is still fundamentally important for the discipline of machine learning.

DECISION TREE LEARNING

Decision tree learning is one of the most widely used and practical methods for indicative inference (Mitchell, 1997). The method can be used for approximation of discrete-valued functions that is robust towards noisy data and that is moreover capable of learning disjunctive expressions.

A decision tree can be described as a simple representation for classifying examples. The decision tree methodology is one of the most popular applied techniques for supervised classification learning (Mitchell, 1997). The main goal of decision tree modeling techniques is to create a model that predicts the value of a target variable based on several input variables. The model can be visualized as a tree in which each internal (non-leaf) node is labeled with an attribute. The arcs coming from a node labeled with an attribute are labeled with each of the possible values of the attribute. Moreover each leaf of the tree is labeled with a class or a probability distribution over the classes.

The majority of this section is presented as a summary of two extensive book chapters, firstly chapter three 'Decision Tree Learning' from the book 'Machine Learning' that is written by Mitchell (1997) and chapter four 'Classification: Basic Concepts, Decision Trees and Model Evaluation' from the 'Data Mining' that is written by Tan, Steinbach and Kumar (2006).

DECISION TREE REPRESENTATION

Decision trees classify instances by sorting them down the tree from the root to a leaf node, this leaf node provides the classification of the instance. Each node in the tree specifies a test of some attribute, and each branch descending from that node corresponds to one of the possible values for this attribute.

FIGURE 2 shows an example of a typical decision tree for the choice of a random commuter to either travel to work by car or bicycle. As seen in the picture the first question that may be asked is the travel distance from home to work. If it is larger than 10 KM the bicycle is no longer regarded as a feasible alternative. Otherwise, when the distance is smaller than 10 KM, a follow up question should be posed: what is the expected precipitation? When the weather forecasts do no not predict any rain the commuter will travel to work by bicycle and when rain is expected the commuter will use his car.

The example illustrates how a classification problem can be addressed; it is based on a set of carefully developed questions about the attributes of a training set. Each time an answer is gathered a follow up question is asked until a conclusion is reached concerning the class label of the record.

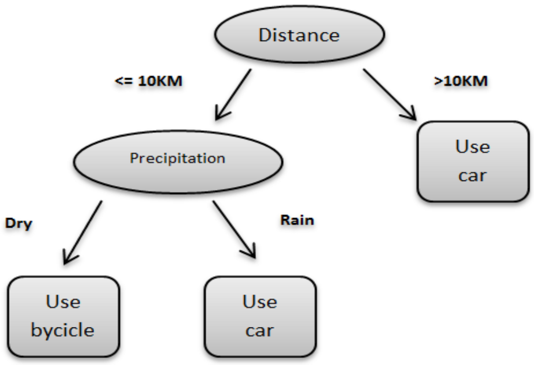


Figure 2: Example of decision tree for the choice of modality for travelling to work

Three specific elements of the decision tree can be appointed, firstly the **root node** which has no incoming arcs and zero or more outgoing arcs, in FIGURE 2 the node 'Distance' can be defined as the root node. Secondly the example in FIGURE 2 contains an **internal node**, which has exactly one incoming arc and two or more outgoing arcs; in FIGURE 2 the node 'Precipitation' can be classified as an internal node. Lastly the decision tree contains **leaf** or **terminal nodes**, which has exactly one incoming arc and no outgoing arcs. Within FIGURE 2 the node 'Use bicycle' can be classified as a terminal node.

Each leaf node of the decision tree is assigned a class label, the non-terminal nodes contain attribute test conditions to separate records that have different characteristics. Classifying a test record is straightforward as soon as the decision tree is constructed, starting from the root node the test conditions can be applied to follow the branch based on the outputs of the tests. This will ultimately lead to a terminal node which specifies the class label that will be assigned to the record.

Building a decision tree

From a theoretical point of view the amount of decision trees that can build given a set of attributes is exponential, while one tree is more accurate than others finding the optimal tree is often computationally infeasible. Some efficient algorithms have been developed to generate a reasonable accurate, although still suboptimal, decision tree in a relative short amount of time. Main element of these algorithms is a greedy strategy that grows the decision tree by making a series of locally optimum decisions about which attribute to use for partitioning the data.

A typically method to construct a decision tree is the 'divide-and-conquer' approach in which a first attribute is selected to place at the root node off the tree. Subsequently this root node splits and divides the dataset in different subsets. Based on these internal nodes the process becomes recursive because the splitting can be repeated for every branch of the tree. To determine which attribute to split on, given a set of data-examples with different classes, different algorithms can be adopted.

One of the most widely used decision tree learning algorithms is the C4.5 algorithm which can be classified as a supervised learning algorithm, this algorithm finds segments in the sample data set that are internally homogeneous in their choice (Li & Limsoon, 2003).

The C4.5 algorithm

As stated earlier, the C4.5 algorithm can be defined as a supervised learning algorithm that finds segments in the sample that are internally homogeneous. The decision tree is constructed in a recursive fashion by partitioning the training records into successively purer subsets. Suppose D_t to be a set of training records that are associated with node t and y $=\{y_1,y_2,\ldots y_{end}\}$ to be the class labels. The following steps will be taken:

- **Step 1** – If all records in D_t belong to the same class y_t than t is a leaf node labeled as y_t
- **Step 2** – If D_t contains record that belong to more than one class an attribute test condition is selected to partition the records from D_t into smaller subsets. A child node is created for each outcome of the test condition and the records of D_t are distributed among the child nodes based on the outcome. This algorithm is than recursively applied to each child node.

At each state of the subdivision a segment is divided according to an explanatory variable that is selected based on the gain ratio which will be described later on. The explanatory value that produces the highest gain ratio is chosen from all explanatory variables at each division and this continues until all cases in each segment describe the same choice.

Difficulty when segmenting the data is the variety in categories in the dataset, when an explanatory value is categorical all possible combinations of categories will be compared and the combination that describes the highest gain ratio is chosen as the preferred division for the variable. If the variable is a continuous variable all cases in the segment will be rank-ordered to the case value of the variable and the threshold that describes the highest gain ratio is chosen. The division for non-ordered discrete values can produce two or more segments while ordered discrete and continuous only lead to binary segmentations. For all potential attributes the variable which produces the highest gain ratio is selected to divide the segment.

MATHEMATICAL INTERPRETATION

One of the subjects that has been described above is the threshold that is applied to split determine whether to split the attribute and if yes; when. Main aim in the DTL algorithm is to select the attribute that is most useful for classifying examples. To quantify the usefulness of each attribute the statistical property 'information gain' is defined which describes how well a given attribute separates the training examples according to their target classification.

To further define 'information gain' in detail we start with discussing and defining the measure that is often used in information theory. The term 'Entropy' describes the impurity of an collection of records (Tan, Steinbach, & Kumar, 2006). Let p(i|t) denote the fraction of records belonging to class i at a given node t. Let's for example suppose a two-class problem, in this case the class distribution can be written as ($p_{0,}p_1$) where $p_{1\,=}$ 1-p_0. Within splitting the data main aim is to find a distribution in which the class distribution of the attributes is unevenly distributed which will lead to a more pure partitioning. The measures developed for selecting the best split are often based on the degree of impurity of the child nodes. The smaller the degree of impurity, the more skewed the class distribution. The specific formula for Entropy(t) can be defined as:

$$I = \ Entropy(t) = \ -\sum_{i=\ 0}^{c-1} p(i|t)log_2p(i|t)$$

For each attribute the best possible (lowest) Entropy is determined, this implies testing multiple split criteria for each attribute. Subsequently the attribute with the lowest Entropy is selected to split the records. The formula for Entropy is based on the paper of Shannon (1948), although its origin goes back to Pauli and von Neumann (1932). Given a collection t, containing positive and negative examples of a target concept, the Entropy for t relative to the Boolean classification is:

$$I = \ Entropy(t) = \ -p_{\oplus}log_2p_{\ominus} - p_{\ominus}log_2p_{\oplus}$$

Where $p_{\oplus}$ is the proportion of positive examples in t and $p_{\ominus}$ is the proportion of negative examples in t. To illustrate, suppose t is a collection of 25 examples, including 15 positive and 10 negative examples [15+, 10-]. Then the entropy of t relative to this classification is:

$$-\left(\frac{15}{25}\right) log_2\left(\frac{15}{25}\right) - \left(\frac{10}{25}\right) log_2\left(\frac{10}{25}\right) = \ 0{,}970$$

Notice that the Entropy is 0 if all the examples within a collection belong to the same class and that the entropy is maximized when the collection contains an equal number of positive and negative examples. If the collection contains a unequal number of positive and negative examples the Entropy can vary between 0 and 1. FIGURE 3 visualizes the shape of the entropy function relative to a binary classification.

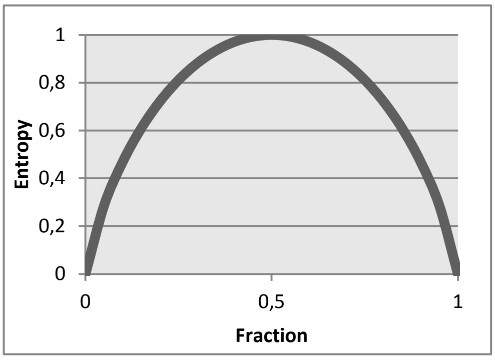


Figure 3: Entropy relative to the proportion of binary positive examples (THOMAS & JOY, 1991)

It is possible to interpret the entropy from information theory as the minimum amount of bits that is necessary to encode the classification of an arbitrary member of t (Mitchell, 1997). For example when all examples are positive, if $p_{\oplus} = 1$, the receiver already knows the drawn example and it is not necessary to send a message. If $p_{\oplus}$ is equal to 0,5 one bit is required to indicate whether the drawn example is either positive or negative.

To determine how well a test condition performs, it is necessary to compare the degree of impurity of the parent node (before splitting) with the degree if impurity of the child nodes (after splitting). The larger their difference, the better the test condition. The gain, Δ, is therefore the criterion that can be used to determine the goodness of the split.

$$\Delta\text{info} = gain\ (\Delta) = I(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

Where I() is the impurity measure of a given node, N is the total number of records at the parent node, k is the number of attribute values and lastly $N(v_j)$ is the number of records associated with the child node v_j. Decision trees often utilize the test condition in which the gain is maximized, however because I(parent) is equal for all test conditions, maximizing the gain is similar to minimizing the weighted average impurity measures of the child nodes. If the entropy is used as the impurity measure the difference in entropy is known as the information gain Δ_{info}.

One difficulty with using a splitting measure such as entropy is that it tends to favor attributes that have a large amount of distinct values. For example take a set of instances which are all described by an unique variable (in example date). Within the DTL algorithm the set cannot be grouped in purer partitions since every instance will become a separate partition (if every entry is made on a separate day). These separate partitions do not have any predictive value. Furthermore this statement above is also applicable in less extreme situations, a test condition that results in a large number of outcomes may not be desirable because the number of records associated with each partition is too small to enable any reliable predictions. There are two ways for overcoming this problem, firstly it is possible to restrict the test conditions to binary splits only. Furthermore one strategy is to modify the splitting criterion to take into account the number of outcomes produced by the attribute test condition. Within the C4.5 DTL algorithm, a splitting criterion known as the gain ratio is used to determine the goodness of a split. The gain ratio penalizes attributes (such as date) by incorporating a term called 'Split Information', that is sensitive to how broadly and uniformly the attribute splits the data:

$$Split\ information(S, A) = -\sum_{i=1}^{c} \frac{|S_i|}{S} log_2 \frac{|S_i|}{S}$$

The gain ratio can subsequently be defined as:

$$Gain\ ratio = \frac{\Delta_{info}}{Split\ info}$$

Where split info is actually the entropy of the node in respect to the relevant attribute. If we analyze the equation it becomes evident that the gain ratio becomes larger as the proportion of the additional information given by the subdivision is larger relative to the potential information that can be generated by the subdivision.

PRUNING THE DECISION TREE

Errors that are originating from the classification model can be divided into two types: training errors and generalization errors. Training errors describe the amount of misclassification errors that are committed by the mode, generalization error is the expected error of the model on previously unseen errors.

As discussed previously a decision tree learning model should both be able to fit the training data well and also must be able to accurately classify records that the model has not seen before. This implies that a good model must both minimize the training errors as well as the generalization error. Especially this last criteria is important, a model that fits the training data too well can have a poorer generalization error than a model with a higher training error. This last situation is known as overfitting.

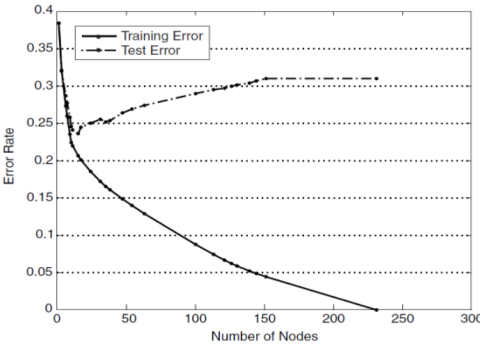


Figure 4: Training and test error rates (TAN, STEINBACH, & KUMAR, 2006)

To further investigate overfitting the model we should start with noting that, within DTL algorithms, the training error of a model can be reduced by increasing the model complexity. The leaf nodes of the tree can be expanded until it perfectly fits the training data, however in this case the test error can increase because the tree may contain nodes that fit incorrect or incomplete training data. These nodes reduce the performance of the tree because they do not generalize well to the test data-set.

FIGURE 4, derived from chapter four of the book 'Introduction to Data Mining' visualizes the error rate for both for an example decision tree by visualizing the error rates from the training and test data set with different number of nodes (Tan, Steinbach, & Kumar, 2006). What becomes apparent is that when the algorithm adds new nodes to grow the tree, the amount of errors reduces monotonically. However, when measured over the test examples, the error rates first decreases but then increases again. It is important to notice that the training and test errors are both very large when the size of the decision tree is very small, this is known as 'model underfitting' which occurs when the model still has to learn the structure of the tree. Once the tree becomes too large the test error begins to increase while the training error still decreases. This situation is known as 'model overfitting'.

There are two methods to avoid overfitting, the first approach is prepruning by means of an early stopping rule, within this methodology the algorithm is halted before a fully grown tree is build that perfectly fits the training data. This avoids the generation of overly complex subtrees that overfit the training data. In practice it is difficult to choose the exact right threshold for early termination, if this threshold is to high the model will be underfitted and when the threshold is too low the method will not successfully overcome the overfitting problem. Another method is post-pruning in which the decision tree is initially grown to its maximum, and perhaps overfitted, size. After the algorithm finishes a post-pruning step is applied which trims the fully grown tree in a bottom up fashion. Two rather different operations can be considered for post-pruning: subtree replacement and subtree raising (Soman, Diwakar, & Ajay, 2006). At each node within the decision tree a learning scheme might decide whether it performs a subtree replacement, subtree raising or leave the tree as it is (unpruned).

Post-pruning tends to produce better results than pre-pruning, main reason is that the post-pruning makes decisions based on a fully grown tree, the pre-pruning process can suffer from premature termination which hampers the reliability of the final decision tree.

Subtree replacement is the primary pruning operation, this methodology selects a number of subtrees and replaces these with single leaves. By replacing the subtrees in leafs the accuracy on the training set will decrease if the original tree was produced by means of an algorithm that continued to build the tree until all leaf nodes were pure. However, it may increase the accuracy on an independent test data-set.

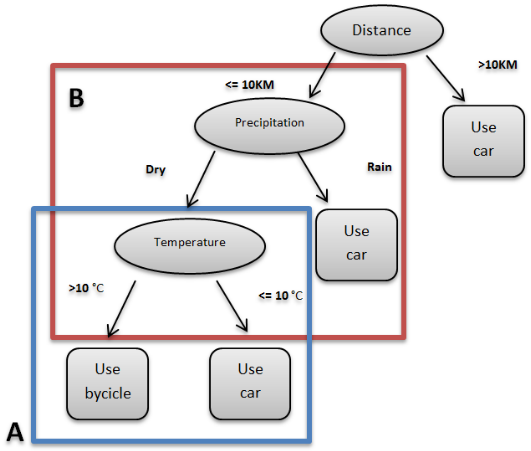


Figure 6: Tree pruning mechanisms

If subtree replacement is implemented, it starts from the leaves and works back up toward the root. See Figure 6 and assume that we evaluate the subtree within the blue marked box 'A'. If there were 19 instances within the subtree that were classified as 'bicycle' and if there was only 1 instance classified as 'car' we could replace the who subtree in the leaf 'bicycle'.

The second pruning operation, subtree raising, is more complex. See Figure 6, and consider the red marked box 'B'. If there only are very small amount of instances in which rain fell (classified as 'car) the pruning mechanism could raise the subtree in the blue box to substitute the subtree in the red box.

To decide whether to apply subtree replacement or subtree raising operation it is necessary to estimate the error rate that would be expected at a particular internal nodes as well at the leaf nodes. Based on this estimate it is possible to replace or raise a subtree by comparing the estimated error of the subtree with that of its proposed replacement.

The trainings dataset cannot be used directly as the test error estimate because the original decision tree is specifically constructed to perfectly represent that dataset. The first available method is the 'standard verification technique' in which a part of the original dataset is cut from the trainings dataset before implementing the model and which could be used as an independent test to estimate the error at each node.

The alternative is to derive the estimate of error based on the trainings dataset itself. The C4.5 DTL algorithm considers a set of instances that reach each node and deducts the majority class that represents that node. Based on this assumption it is possible to deduct the number of 'errors', *E*, out of the total number of instances, *N*. The errors can be converted to a probability of error at node *q* and that the N instances are generated by a Bernoulli process with parameter *q* in which *E* represent the errors.

Within this methodology a upper confidence limit is used, given a confidence *c* it is possible to deduct the confidence limits *z* such that (Witten & Frank, 2005):

$$Pr\left[\frac{f-q}{\sqrt{q(q-q)/N}} > z\right] = c$$

Where *N* describes the number of samples, *f*=E/N is the observed error rate and q is the true error rate. This leads to an upper confidence limit for *q*, this upper limit can be used to derive the follow pessimistic estimate for the error *e* at the node (Witten & Frank, 2005):

$$e = \frac{f + \frac{z^2}{2N} + \sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}}$$

The positive sign before the square root is in the numerator is implemented to obtain the upper confidence limit. In the function *z* represents the number of standard deviations corresponding to the confidence interval c which can be derived from the normal distribution.

The decision to apply pruning is made based on the comparison of the upper limits of the confidence intervals for the two trees. Firstly the expected error is approximated when the pruning step is applied, subsequently the backed-up error is computed from the children assuming we do not apply pruning. If the expected error is less than the backed-up error the tree is pruned.

3. METHODOLOGICAL APPROACH

This chapter presents the research methodology that is employed within this thesis, or more specifically; it will describe how the C4.5 decision tree learning algorithm is employed to predict and validate individual route choice decisions. Within the approach five stages can be identified. These stages include the collection and assimilation of scientific literature, a large scale data gathering and data processing activity, the implementation of the C4.5 decision tree learning algorithm and lastly the evaluation and reflection in which the results from the experiment will be discussed.

Section 3.1 focusses on the approach and methodology by describing the five major stages in this research. The research will be characterized in section 3.2 and subsequently the data collection methods will be discussed in section 3.3. Section 3.4 will describe the data processing methods and lastly section 3.5 will describe the main assumptions that are employed within this thesis and lastly section 3.6 will provide the outline of the envisioned system architecture to implement the 'Personalized Adaptive Routing Algorithm' and subsequently the main components of this system architecture will be discussed in detail within section 3.7.

3.1 THE RESEARCH STRATEGY

The research strategy comprises of five key stages. The aim of the **first stage** is to gain further insight in the theoretical backgrounds of travel behavior, and especially route choice decisions. Moreover it aims to explore and elaborate the methodological aspects of learning algorithms. This should result in a methodology in which the Decision Tree Learning algorithm can be applied to induce the mechanism of driver's route choice without presupposing any theoretical constructs. The **second stage** will aim to generate a large scale trip database in which for multiple users an array of trips will be gathered which are described by the origin, destination, timestamps and the revealed route. This database is generated based on the results of a field operational test in which a global positioning system based data acquisition platform. In the **third stage** the data gathered from the second stage will be processed, main aim is to process the raw data from the data gathering phase towards a structure that supports the implementation of the 'Personalized Adaptive Routing Algorithm'. During these data processing activities we hope to identify the opportunities, challenges and possible issues that arise during the implementation of the DTL algorithm. The aim of the **fourth stage** is to actually implement the decision tree algorithm based on the data structure that is derived from the third phase. The **fifth stage** will particularly focus on the comparison of the performance of the DTL based personalized adaptive routing algorithm with the traditional shortest path and multi-attribute routing algorithm. Main aim is to ascertain whether the 'Personalized Adaptive Routing Algorithm' is able to learn the user preference through implicit interaction with the driver. Moreover it aims to research if the algorithm is able to plan better routes for that user in future operations (learning behavior). The major results of this phase is the answering and discussion of the five sub-questions, and moreover the central research question answered.

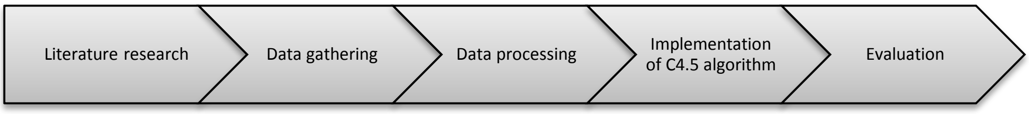


Figure 7: Research strategy

3.2 Research typology

The purpose of this research is to implement and evaluate a real-life implementation of a 'Personal Adaptive Routing Algorithm' based on the C4.5 DTL decision tree learning algorithm. Because this study simultaneously discusses the implementation and evaluation of the proposed algorithm it is possible to define the research typology from this master thesis as 'exploratory research'. Exploratory research is a form of research conducted for a problem of which the cause and possible solutions not yet have been defined. It aims to determine the best research design, data collection method and the selection of subjects. Especially when the purpose of the research is to gain familiarity with a phenomenon or to acquire new insight into a specific object to an exploratory research is often useful. Exploratory research often relies on secondary research such as reviewing available literature and data or more formal approaches through in-depth case studies and pilots.

3.3 Data collection

This section will discuss the data collection within this thesis, firstly the data collection methods will be discussed. Secondly the case study, which lies at the root of this graduation project, will be elaborated.

Data collection methods

The aim of this thesis is twofold. On one hand a qualitative assessment of the underlying theories concerning the application of the C4.5 DTL algorithm is applied, and on the other hand the actual implementation and evaluation of the algorithm within a quantitative pilot study. These different components require different data collection methods to be employed.

Literature review: Exploration of the underlying theories and assumptions

To understand the underlying principles of the C4.5 DTL algorithm this thesis includes a review of the international qualitative research literature on the subject of decision tree learning. The literature review was undertaken in order to better understand the available dimensions, processes and application of decision tree learning algorithms. To the best of this authors knowledge no real life implementation of the DTL algorithms have yet been applied in the field of traffic engineering, however a large amount of knowledge has been compiled while discussing the theoretical implications of DTL algorithms.

The literature review for this thesis was carried out to provide information relating to the historic and general background and contexts of decision tree learning algorithms and more specifically the C4.5 algorithm. The study focused on both national and international literature and more specifically focused at the following subjects:

- Sharpen the theoretical framework concerning decision tree learning algorithms,
- Become familiar with the latest developments,
- Identify gaps in knowledge and weaknesses in previous publications,
- Study the definitions from previous publications as well as the research settings of these publications,
- Inspect the research methods from methodologies used by others and to possibly improve them in the current research,
- Identify the requirements that are applicable for attribute classes that are included.

Case study: Implementing the decision tree learning algorithm

The relevance and performance of the C4.5 decision tree algorithm was explored by means of Global Positioning System (GPS) based disaggregate trip data. The data is collected within the overarching project KATE of the Netherlands Organization for Applied Scientific Research (TNO). Main aim of the project is to develop a 'research toolbox' in which up-to-the-minute travel times and travel behavior can be predicted. Main premise of the KATE platform is to utilize the capabilities of modern smartphones.

Individual travel information has been of significant interest to researchers. This interest arises based on the fact that trips can place great strain on the transportation networks. While past researches have investigated travel distance, travel times, model and route choices associated less attention has been paid to investigate the day-to-day dynamics of travel behavior. While repetitive trips (such as commuting) are often assumed to be static and therefore highly predictable, research results indicate that users often change departure times, routes and also often use trip chaining that significantly affect the departure time and route choice behavior (Hainan, Randall, Ogle, & Wang, 2004).

The current principles of today's transportation policies move away from expanding the infrastructure to meet the unrestricted growth of the traffic demand. New applications and technologies will be implemented that manage the travel demand more efficient and understanding the variability of demand is key to the success of these measures. Especially the day-to-day dynamics of (repetitive) travel behavior can significantly influence the development, analysis and performance of traffic management measures.

In the past research in the field of travel dynamics were impeded by the complex nature of gathering and subsequently analyzing observations, one of the major impediments to develop a larger body of knowledge is the lack of sufficient data at detailed levels. In the past this data was often gathered by self-reported travel diaries.

Advancements in GPS technology facilitate automated, implicit and multiday data collection applications. The KATE platform is developed to provide a programming structure for mobile applications. The platform consists of various components to develop applications for mobility services such as dynamic route advises or applications to detect travel patterns. 'ReisAlarm' is a first realization of a smartphone application that is currently being developed based on the KATE platform.

The main components of the 'ReisAlarm' application consist of a link to the professional and private calendar of the user and an intelligent multi-modal travel time calculation model. Based on this calendar the application deducts the origin and destination and predicts the travel time for the various available travel modalities. This travel time prediction does not only include the historically observed traffic situation on the route, but also takes the current traffic situations into account. The application monitors weather circumstances, incidents or road maintenance works and subsequently uses this information to adjust the route advice. Based on the available information the application suggests the user to depart early, to take an alternative route or to choose an alternate travel modality. When, due to critical incidents, no feasible arrangements or alternatives exists it is envisioned that the application advices the user to cancel the trip and moreover supplies the contact details of the participants concerned in the calendar entry.

One major component of the ´ReisAlarm´ application, and critical for this study, is the positional monitoring algorithm. This algorithm detects if the phone is moving and moreover periodically determines the exact position of the smartphone. This location tracking service is integrated within the background services of the operating system of the phone and automatically starts when the phone is switched on. Based on these location traces and a sophisticated data processing algorithm the travel behavior of the user is deducted and stored on a server.

In order to respond to the dynamic characteristics of travel behavior a rule-based programming mechanism has been integrated within the ReisAlarm application, for example specific questionnaires can be pushed to the user when the application detects that a user installed the application or when the location tracing algorithm detects when a user has made a trip.

The 'ReisAlarm' application that is applied in this study is designed for the Android smartphone operating system. Although multiple 'visual' differing applications are developed, mainly two differing versions can be defined. The first only includes the location monitoring algorithms (0-measurement) and the second version actively supplies multi-model travel advice that is facilitated by a link between the application and the calendar that is stored on the device (1-measurement).

PROJECTS IN WHICH THE REISALARM APPLICATION IS APPLIED

The 'ReisAlarm' application has been employed in three separate projects within TNO. The results from all three projects will be combined in one data set to which the C4.5 DTL algorithm will be applied. The text below will further discuss the main principles and characteristics of these projects.

ENABLING TECHNOLOGY PROGRAM

Within the Enabling Technology Program (ETP) TNO develops methods for monitoring innovation and behavioral changes. The developed methods are subsequently applied on innovations, interventions and collaborations in several contexts. In one of the seven sub-projects, the ETP Mobility Behaviour project, knowledge and researches are gathered from the field of traffic and transport in which measures are actually implemented. By a combination of small and large scale researches in combination with interviews, workshops and sessions TNO aims to unravel the behavioral determinants within travel behavior.

Between the end of September 2012 and the end of April 2013 47 TNO users were invited to participate in the 'ReisAlarm' experiment of which 43 users actually installed the application. The invited users were mainly TNO employees of which 25 users were located in Soesterberg and 18 users were located in Delft. Furthermore 4 users were invited that are not employed by TNO.

To gain further insight in the demographic, temporal and situational influencing factors the specific ETP ReisAlarm application includes a survey module which can be used to deploy and process predetermined surveys. Within the project two types of surveys were employed; after installing the ReisAlarm application the participant received a first use questionnaire in which the demographic and personal characteristics such as gender, car ownership, preferred mode of transport etc. Furthermore the application was configured to push post-trip surveys in which the information concerning the trip specific circumstances such as trip motive, expected delay, observed delay and obtained traffic information can be gathered.

Project characteristics

- Lead time: started on the 24th of September 2012 and still ongoing,
- Large scale deployment on the 15th of October 2012,
- Included first use and trip-end surveys,
- 0-measurement application without travel and route advices,
- Data between the 24th of September 2012 and 31th of March 2013 used in thesis.

Sensor City Mobility project

The Province of Drenthe and the municipality of Assen desire to create an environment for research concerning intelligent traffic management and intelligent travel information. By means of a broad consortium of public authorities, research institutes and private businesses an extensive sensor network is developed. With this sensor network all important traffic flows can be measured in an around Assen. Main aim is to study the feasibility and possibility to create new sources of data and new algorithms to predict how much traffic will travel to the city center at an early stage.

The 'ReisAlarm' application was employed within one of the five use cases in the Sensor City Mobility (SCM) project, main aim is to support the user in his decision making process for departure time and travel modality. To test the 'ReisAlarm' application in preparation for the large scale implementation the 0-measurement version of the application was distributed. Between September 2012 and December 2012 4 users had installed the application for testing and developing purposes. From December 2012 onwards the application was supplied to a broader group which also includes a number of external test participants within Assen. In total the application was supplied to 28 users of which 24 users installed the application.

Project characteristics

- Lead time: started on the 24th of September 2012 and still ongoing,
- Large scale deployment on the 14th of October 2012 to the members within the consortium and on the 21st of December 2012 to the external test participants,
- 0-measurement application without travel and route advices,
- Did not include any questionnaires,
- Data between the 24th of September 2012 and 31th of March 2013 used in thesis.

TNO internal ReisAlarm trial

Based on the initial experiences within the ETP Behaviour project it was decided to further investigate the possibilities and performance of the ReisAlarm application prior to the public distribution of the application within the Sensor City Mobility project in Assen. Due to the fact that the tests of the application in Assen still focused on the 0-measurement version (without route advices) a full version of the application, independent from the project in Assen, is distributed organization-wide within TNO. Main aim was to collect the first experiences concerning the route advice module and to research the cross-platform and multi device stability.

During the last weeks of December 2012 the internal test of the 'ReisAlarm' application has been communicated by means of flyers, emails and word of mouth publicity. Potential participants could register by means of a registration portal on the internet and within the second week of January 2013 the application was distributed among 40 participants. From these 40 participants 25 users have installed the application.

Project characteristics

- Lead time: started on the 24th of September and still ongoing,
- Large scale deployment on the 12th of February 2013,
- 1-measurement application that includes travel and route advices,
- Included the first-use questionnaire,
- Data between the 24th of September 2012 and 31th of March 2013 used in thesis.

Privacy statement

Prior to distributing of the ETP Mobility Behaviour application the participants were verbally informed that the application monitors their exact location, moreover the email by which the application was distributed repeated this message. Participants in the TravelAlert internal test project were invited by means of flyers and had to register by means of an on-line registration procedure. Within this procedure the participants had to agree to the terms and conditions which indicated that the application contained a location monitoring algorithm. Lastly participants in the Sensor City Mobility had to sign a statement which included the terms and conditions. In general all participants were informed that the data derived from their smartphone could be used for scientific purposes, moreover all applications included an icon that was added to the left corner of the Android status bar to emphasize that the application was active.

3.4 Tools for the data analysis

Qualitative literature review

Literature reviews are important as research tools, especially in theoretical emerging areas in which large scale operational tests are unavailable or unfeasible. Literature reviews are of value in situations where an overload of information is available and when it is literally impossible to process all available literature.

Identifying potential topics of interest is typically the beginning of a literature review, discovering general information concerning the topic and to discover where this information fits within broader and narrower subject categories. Based on this information the major ideas, issues, controversies and prominent researches can be identified. Within this research the second chapter, already investigated.

Based on the preliminary literature research the possible opportunities and gaps within the available literature were defined, main aim was to develop a framework to understand where the available literature was positioned. This eventually led to a research proposal that is subsequently translated to the thesis outline as described in the first chapter of this report.

In this thesis the literature framework mainly functions within a supporting role to describe and substantiate the choices and assumption that were made during this research.

Case study data analysis

One of the main characteristics that differentiates this thesis in respect to past researches is that the proposed decision tree learning algorithm is actually implemented and tested based on real life GPS data from 94 participants.

Main point of departure for this case study is the literature review that was described previously. In the past a number of researches described the implementation of a decision tree algorithm to predict routing behavior based on stated trip information (based on trip diaries) or simulated travel behavior. Due to the inherent complexity of gathering and subsequently analyzing observations of dynamic travel behavior data past researches were unable to utilize detailed input data (Mahmassani, Hatcher, & Caplice, 1997).

One of the main advantages of the KATE mobile data acquisition platform is that it is able to detect and process travel data at a very detailed level. Based on the GPS derived data from the KATE platform it is possible to better capture travel behavior during a long uninterrupted period of time while minimizing the explicit user interaction (naturalistic travel data).

Main input element for the previous attempts to include decision tree learning algorithms within route choice was a set of simulated trips and their routes. If we consider the output information from the KATE platform it becomes apparent that the data is very suitable as a starting point for testing the real life performance of 'Personalized Adaptive Routing Algorithm'. It is for instance very well possible to derive both the trips and routes based on a set of GPS traces given that there are enough locations with an acceptable accuracy within each trip.

One essential element that is required to implement the decision tree algorithm is the indicator framework that aims to describes the personal factors within route choice; this indicator framework supplied the attributes by which the dataset was divided within the DTL algorithm. To facilitate a comparison of the results of this thesis and the results of Park et al. (2007) the indicator framework that was suggested and implemented in 2007 is implemented in this study. Aim is to define and elaborate the indicators that can be included with relative ease without major adjustments to the output data. However, before assuming the framework of Park et al. (2007) as leading a literature study has been conducted which examined other researches that applied, suggested or recommended an additional and more detailed attributes. Within this literature review no alternative or extended attributes were found.

Based on the indicator framework the locational data from the KATE platform will be translated to represent the trips, routes and necessary information to evaluate the attributes from the indicator framework. Main challenge are the necessary data processing steps to segment and align the raw GPS traces into sequences that describe trip information. For example all locational traces should be separated in different sets of measurements which each describe a separate journey. Furthermore the separate traces should be mapped to the underlying infrastructural network, due to the noisiness of the GPS traces this mapping is challenging. The data processing challenges will further be discussed in a future section of this report.

Even without incorporating the personal preferences within route choice, the data from the KATE mobile acquisition platform (trips, timestamps and routes) supply a great amount of information that can be utilized to study the general characteristics of travel behavior. Both for the full population and for specific user groups the revealed trip database have been used to determine the average number of trips, average trip length and the distribution of these variables. The user groups have been determined by means of the in-app first use questionnaire in which information concerning age en gender is collected. By differentiating trips that are made by (private) car and train the modality choice of travelers can be examined. When the information above is combined with the user specific information derived from the first use questionnaire it is possible to examine the general travel behavior of specific user groups.

One method to incorporate driver preference is to identify the personal influential factors by means of explicit interaction with the user, for example by utilizing a questionnaire. On the other hand a more simple approach can be utilized by learning and manipulating user preference implicitly by not modeling the factors affecting preference but by treating each journey (and the corresponding route) from the past as a statement of preference. Based on a specific origin and destination a set of proposed routes can be calculated and based on the historic revealed behavior in terms of for example directness and familiarity a route can be proposed that matches the drivers personal expectations.

Due to the fact that the 'Personalized Adaptive Routing Algorithm' uses revealed preference as input, the trip data from the KATE platform can effectively be utilized to test the performance of the algorithm. By firstly adopting a training set of revealed routes as 'training' data the prediction model can be adapted to the user. By subsequently testing the remaining set of trips as 'test' data it becomes possible to test the accuracy of the adaptive personalized routing algorithm. For example, given the fact that a set of routes between a specific origin and destination is available, and given the fact that the algorithm has defined one of these routes as 'most' likely it is possible to compare the predicted and revealed route. If the performance of both the personalized and traditional route algorithms are compared the added value of the personalized algorithm is demonstrated.

A second indicator that is applied to test the adaptive routing algorithm is the 'learning' curve of the algorithm. As previously stated, a subset of the revealed behavior is used as training data to calibrate the route prediction model. This does not necessarily mean that the algorithm is performing optimally. During the evaluation the trips from the test data-set the previous trips will automatically be added to the training set. By including the additional information to the training set the predictive performance of the personalized routing algorithm will improve over time. Although by separately processing each individual users independently the amount of data is limited, especially if the data is divided in a trainings and test set. However it is assumed that the extensive data gathering time will have supplied enough information per user to test the personalized routing algorithm. A future chapter will further discuss the evaluation framework to compare the performance of the DTL algorithm with the traditional algorithms, moreover a future section of this chapter will further discuss the system architecture in detail.

3.5 Main assumptions and theoretical principles

Two major assumptions are inherent to the research methodology discussed earlier, these assumptions are:

- Drivers prefer the attributes and characteristics of routes and roads that they have taken before,
- Drivers are assumed to make informed decisions, trips and routes are not extended or changed out due to a lack of information.

The first assumption is straight forward, if the revealed behavior is assumed to be a 'statement of preference' the combination of attributes of this specific route is preferred by that specific route. The decision tree learning algorithm will deduct these attributes and uses these to classify future unseen classifications. The second assumption is somewhat more difficult because it does not fully represent reality in which drivers are not always fully informed. Drivers can for instance choose one specific route because they are not aware of the existence of an alternative, but this situation is indistinguishable from the data set. It is however expected that, although this potential problem can arise, the proposed personalized route advice may more accurately match the route that the driver would choose.

Based on a preliminary review of the available decision tree learning algorithms (as described in chapter 2.4) it is expected that the C4.5 algorithm (when applied with default parameters) tries to find a small (and therefore simple) decision tree. Main influencing factor in the tree size is the default post pruning mechanism. An important element that distinguishes the C4.5 algorithm from all other decision tree learning algorithms is the ability to handle both continuous and discrete attributes while improving the error handling. The C4.5 algorithm can create relatively understandable decision trees while preserving the computational efficiency of the methodology. Therefore the focus within this thesis will lie on implementing the C4.5 algorithm, however the other algorithms have been included within the literature study.

The indicators that will be included as attributes within the decision tree learning algorithm will be mainly based on the previous research of Park et al. (2007) that discussed the implementation of the C4.5 DTL algorithm in traffic. The attributes described in this research will be translated towards measureable output from the KATE platform. As stated before previous and other studies that described and proposed alternative or improved indicators have been studied, however based on the results of this literature review and the data that is currently available from the KATE platform no additional attributes have been applied during this study.

3.6 Overview of proposed system architecture

Following the rapid expansion and widespread utilization of database technology, there is a growing interest in developing and utilizing techniques for extracting knowledge from data. More particularly there is a growing interest in techniques that go beyond the traditional statistical analysis and produce symbolic rather than numerical data description. Such methods should be able to discover ‘conceptual’ patterns in the data, suggest explanations for them and generate plausible predictions.

The first sections of this chapter described the proposed research strategy. Based on this strategy and the theoretical delimitation from chapter two this section will present the proposed system architecture that aims to integrate the C4.5 DTL algorithm within route choice to generate a personalized adaptive routing algorithm.

Outline system architecture

This thesis aims to describe the experiment in which the C4.5 decision tree learning algorithm is applied to analyze its applicability, learning ability and performance in relation to route choice behavior. Key element in this experiment is the collection, processing and enrichment of floating car data to describe the actual revealed route choice. This data should describe the route attributes and driver choice. Within this experiment the dataset derived from the KATE platform that employs GPS and internet connected Android smartphones to continuously monitor the location of the participants.

The proposed system architecture is visualized in the flow diagram depicted in Figure 8, moreover it has been added as appendix 2. The text below will further describe the main outline.

Based on the revealed trips that will be derived from the GPS dataset the necessary information such as origin, destination and revealed route are extracted. Subsequently for each revealed trip a set of maximally disjoint path sets between the given origin and destination will be computed based on a k-shortest path algorithm. This algorithm does not only find the shortest path, but also fourteen other paths in order of increasing costs (travel time based). The total number of 15 paths is assumed to represent a wide variety of routes while remaining computationally feasible. To prevent the generation of paths that are not significantly different the path generation algorithm includes a Monte Carlo analysis which randomly increases the costs of specific parts of the previous paths to ensure the generation of significantly differing paths.

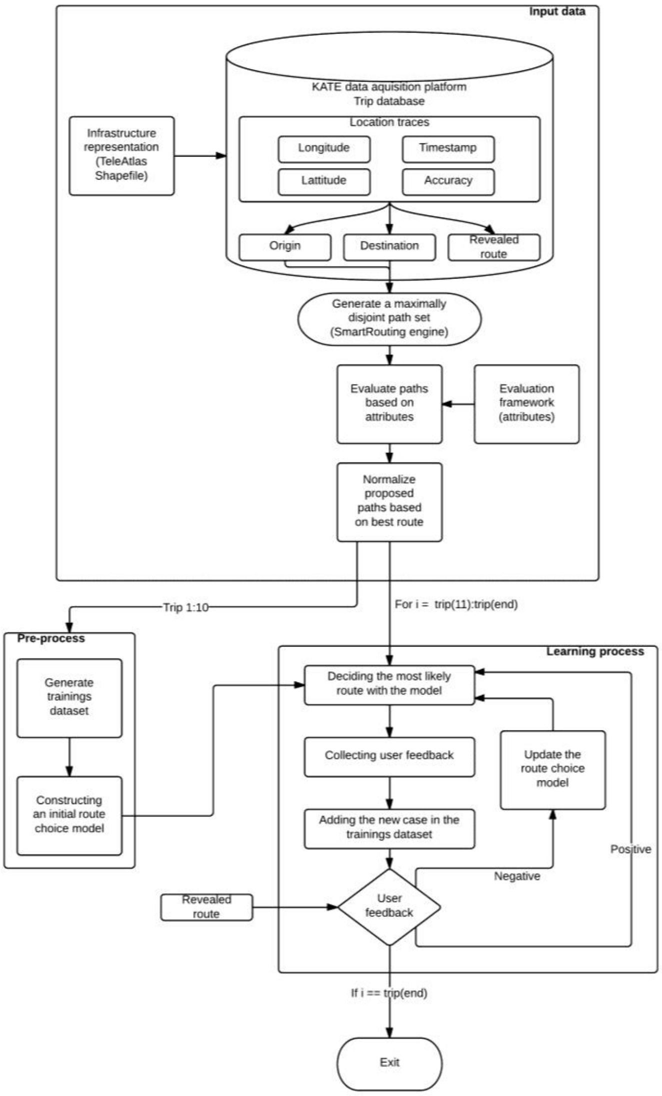

Input data
KATE data aquisition platform
Trip database
Infrastructure representation (TeleAtlas Shapefile)
Location traces
Longitude
Timestamp
Lattitude
Accuracy
Origin
Destination
Revealed route
Generate a maximally disjoint path set (SmartRouting engine)
Evaluate paths based on attributes
Evaluation framework (attributes)
Normalize proposed paths based on best route
Trip 1:10
For i = trip(11):trip(end)
Pre-process
Generate trainings dataset
Constructing an initial route choice model
Learning process
Deciding the most likely route with the model
Collecting user feedback
Update the route choice model
Adding the new case in the trainings dataset
Negative
Positive
Revealed route
User feedback
If i == trip(end)
Exit

Figure 8: Proposed DTL system architecture

After generating a set of possible paths the individual paths will be evaluated and described based on an evaluation framework that consists of the attributes which will also be used within the decision tree learning algorithm. As discussed previously this thesis will employ the original attributes from the research of Park et al. (2007). Because each individual trip within the personal trip database is not similar in terms of distance, and therefore not comparable with other trips, all paths will be normalized based on the shortest route (travel time based) that is available within the set of paths.

After generating, evaluating and normalizing the set of possible paths for each revealed trip the first 10 trips will be forwarded in the pre-process component. Based on the first 10 trips the initial decision tree model will be generated. Main input for the initial model is the set of paths for each trip and moreover the revealed route of the user.

After the preliminary calibration of the model the remaining set of routes between an OD pair is input to the learning process. Within the learning model all paths of each remaining trip will be classified by the DTL model. Based on these classification the most likely route will be chosen. This predicted route is fictively suggested to the user and the user feedback is collected by comparing the revealed route of the driver with the personalized predicted route. If the feedback is negative (the prediction was incorrect) a new case is added to the data set and the model is updated, when the feedback is positive (the prediction was correct) the new case is still added but the model is not updated because the model is still valid.

The process of classifying the set of possible paths and comparing the predicted route will continue until all trips for that specific users have been exhausted. Subsequently the whole process will start with the next user.

In order to assess the performance of the DTL route choice algorithm the predicted routes of the two traditional algorithms and the 'Personalized Adaptive Routing algorithm' route will be assessed and compared with the revealed route. These two traditional algorithms are both utility based, one of the algorithm utilizes selects the shortest path (travel time based) as predicted route and within the second algorithm a multi attribute utility function is employed that utilizes the same attributes as the DTL route choice algorithm. Both traditional algorithms will be further discussed in a future part of this report. Based on the predictive accuracy of each algorithm the added value of the 'Personalized Adaptive Routing algorithm' is analyzed.

The proposed system architecture is visualized in the flow diagram depicted in FIGURE 8; as discussed the initial data, route attributes and choice results are gathered in the pre-process section. Moreover an initial route choice model is generated by means of the first 10 revealed trips. Subsequently the loop in which the remaining trips are classified and where the model is adaptively updated is depicted in the learning process.

Based on the main outline above it is possible to define a number of key components of the system architecture which will further be discussed in the next section.

3.7 DETAILED DESCRIPTION OF THE SYSTEM ARCHITECTURE

The last section discussed the main outline of the system architecture. Although the main components already have been introduced within the main outline of the architecture this section will further elaborate the main components. Within FIGURE 8 three main 'blocks' have been marked; input data, pre-process and learning process. Each of these blocks will be represented by one subheading in this section.

INPUT DATA

The first major component of the system architecture, as depicted in FIGURE 8, is the input data. This component consists of six major elements. Each element will be discussed in the text below.

TRIP DATABASE

As discussed previously the GPS data set is derived from the TNO KATE data acquisition platform. By actively tracking 94 users knowledge is being generated in the individual travel choices that are made pre-trip and en-route. All these users were equipped with an Android smartphone on which the location monitoring application ('ReisAlarm') is installed.

The application collects the location of a device every 5 minutes which are stored as location traces, however to improve the data quality while travelling the update interval is reduced to 1 minute when the application detects that the phone is moving. Main reason to utilize a longer update interval when the device is stationary to preserve the battery.

Within each measurement the timestamp, longitude, latitude and accuracy is collected and transmitted to a central server by means of the mobile internet connection. When for some reason the internet connection is unavailable the device temporary stores the data locally and transmits the cached locations when the connection to the server is restored.

Chapter four will further discuss the technical implementation of the KATE data gathering processes to support the trip structure needed for the 'Personalized Adaptive Routing algorithm'.

Infrastructure representation

By only gathering the GPS location traces it is not possible to directly determine the revealed route, to further investigate the travel behavior it is necessary to link the GPS traces to the available infrastructural network. It is for example not possible to derive the distance travelled based on the raw GPS locations due to the fact that road section not necessarily correspond to the aerial distance between the GPS locations (for example bends etc.). Within a future section of this report the methodology to MapMatch the locational traces to the infrastructural network is described, however in respect to this architecture it is important to mention from which database the infrastructural characteristics were deducted. In this master thesis the TeleAtlas shape file, developed by TomTom N.V. is utilized. The reference year for this shape file is 2011. This shape file is supplied in the context of the Sensor City Mobility project in Assen. The TeleAtlas shape file represents the road network by means of a graph, each node in the graph represents an intersection and the links represent the roads themselves. The representation of a single road often requires many links, since each road segment (the unbroken piece of road between two intersections) is represented by a single link. Each link and node is linked to a database which contains all the individual information (e.g. road type, longitudinal and lateral location, maximum speed).

Building a set of maximally disjoint paths

Within the proposed 'Personalized Adaptive Routing Algorithm' a set of possible paths is evaluated, main aim is to select the 'best' route that represents the preferences of a given user. This methodology implies that a method is required to generate this set of routes based on a given origin and destination. Main challenge in building this choice set is that a separate set has to be made for every unique trip within the GPS data set which is computationally expensive, especially when a large network is utilized. Moreover, a second challenge is that each route in the set of paths should be significantly different, it should be avoided that two routes may only differ by 0.1% due to the addition of a minor link. A set of significantly different routes is essential to the performance algorithm, if the set of paths consists of one possible route with fourteen minimal variations of this route the chance of correctly predicting the revealed route are minimal.

The choice set within this master thesis is generated by means of a modified version of the Smart Routing engine. This engine is developed by TNO within the Sensor City Mobility project and aims to spread the connected users on several routes to maximally utilize the available infrastructural capacity. The routing algorithm within the SmartRouting engine determines the K-shortest paths based on a shortest path utility optimization function (Yen, 1971). To prevent the algorithm to generate non-differing routes the algorithm includes repeated Monte Carlo simulations. These Monte Carlo iterations simulate major incidents on the routes by increasing the link travel times on specific links which negatively influence the utility of the path.

The SmartRouting engine is programmed in Java code and runs separately from Matlab in which all the other data processing functions are applied. The communication is facilitated by a UDP/TCP connection. After a request for a set of routes has been send to the SmartRouting engine the SmartRouting engine starts calculating a maximally disjoint path set and returns the paths to the Matlab client.

Although the structure between SmartRouting and Matlab a is kept separate, a number of significant changes have been applied to the SmartRouting algorithm. Firstly the structure and response of the SmartRouting engine had to be changed, where the original program only returned the path with the 'best' score this thesis required the whole set of paths. Therefore the SmartRouting engine not only transmitted one path but a set of paths.

Moreover it was not directly possible to receive the full set of link identification codes (linkid's) that represented a path. The SmartRouting was only able to provide one waypoint coordinate (latitude and longitude) for every 5 kilometers of a path, to convert these waypoints to linkid's that represent the complete path the waypoints are MapMatched to the infrastructural network. The methodology to MapMatch these point is similar to the methodology to MapMatch the GPS traces that represent the revealed route which will be discussed in a future chapter. Lastly the original SmartRouting implementation only utilized the detailed infrastructural network in and around Assen; all low level roads outside a predefined perimeter were deleted. Within this research the whole network from the TeleAtlas shapefile is utilized.

During this thesis it became apparent that while generating the set of possible path, the algorithm provided less 'unique' paths than initially expected. Based on visual controls it became apparent that the SmartRouting algorithm creates paths that are very similar, for example links on highways were avoided by using the access and egress lanes of the motorway. To increase the chance that the algorithm generates significantly different routes the amount of routes that was deducted from the SmartRouting algorithm was increased to 50. Due to the increased amount of Monte Carlo iterations more significantly routes were deducted because the amount of delay within the network increases during each Monte Carlo iteration. Based on post-processing in Matlab, the 50 routes were analyzed and the 15 most significantly different routes were recorded in the set of possible paths.

Evaluation framework

Fundamentally, building a decision tree entails determining the attributes that can partition the training examples most effectively so that the data in the same segments are as homogeneous as possible. In the past the research of Park et al. (2007) researched the applicability of DTL algorithms on routing and moreover tested the methodology based on simulation data. As input for the simulations, the researchers have compiled a list of indicators. This list was based on the availability of data in the digital maps of the utilized simulation model.

To research the effects of the real-life implementation of the decision tree learning algorithm, and to facilitate a comparison between the results of the research of Park et al. (2007) the indicator framework of the research of Park et al. (2007) is used as a starting point in this research.

Within this research an indicator set of seven attributes has been utilized, table 1 summarizes the attributes, moreover this table describes the quantification and the range of values will be indicated in which the values may differ. As discussed previously these attributes are deducted from the paper of Park et al. (2007) to allow a comparison of the current and past results. However the methodology to quantify the attributes is reassessed to facilitate the revealed trip data from the KATE mobile data acquisition platform.

Attributes	**Quantified by**	**Range of values**
Travel Distance (td)	∑ Travel distance	td ≥ 1
Travel time (tt)	∑ Travel time	tt ≥ 1
Aversion (ave)	∑ of links on each road type	ave ≥ 1
Complexity (comp)	∑ of number of left, right and U turns	comp ≥ 1
Travel time reliability (rel)	∏ Travel time reliability	0 < rel≤ 1
Directeness (dir)	Arial distance / actual distance	0 < dir ≤ 1
Familiarity (fam)	No. of times of visited links / total no. Links	0 < fam ≤ 1

Table 1: Route choice attributes

EVALUATING THE POSSIBLE PATHS BASED ON ATTRIBUTES

Based on the set of proposed routes for each individual trip, the value of each attributes for each of the possible paths is computed. The indicators that have been introduced above is further described in the text below.

The **travel distance** is quantified by means of the sum of the link distances within a path, similar the total **travel time** will be computed by summing the free-flow travel time of each link on each route. In an ideal situation it would have been of added value to also include historic link travel times to enrich the input data of the model. However during this thesis a reliable database representing the historic link based travel times was not available.

The type of roads has an impact on the route choice depending on the situation and individual preference. Where a random driver may prefer the high level network because these roads have fewer intersection where intersecting traffic flows converge another driver (i.e. elderly) dislike these roads because they are not confident at driving at high speeds. Within this research it is assumed that provincial and minor roads are relatively less attractive, in this context the road type and the corresponding link length can be considered as the extent of the perceived **aversion** to a certain road type. The aversion will be calculated by weighing the links distances on all types of roads within a path. Major access roads (gebiedsontsluitings) will be weighted by a value of 1, minor roads (erftoegangswegen) will be weighted with a value of 2 and all other classes will receive a value of zero. If the total length of a path is 15 kilometer of which 2,5 km is travelled on minor roads, 5 km is travelled on major access roads and if the remaining 7,5 kilometers is travelled on highways the aversion score is $2.5\times2+5\times1+7.5\times0=10$.

The number of turns during a route can be interpreted as the complexity of a route, a turning movement requires more attention of the driver compared to straight road section. This implies that a route with more turns can be considered as a more complex route. Furthermore the difficulty in driving differs from the directions of movements (right turn, left turn and a u-turn). Within this research it is assumed that drivers are likely to choose a route which has fewer turns. Moreover the type of turn is important, for example a left turn is considered more difficult because the driver should cross the oncoming traffic. The **complexity** of a route is represented by the number of intersections and the layouts of these intersections. The difficulty in quantifying this attribute is that each node within the TeleAtlas database is not necessarily an intersection in reality. Large roads are often divided in various links by means of intermediate nodes. It is however possible to pinpoint the real intersection by comparing the number of links that are attached to a node. If the number of links is equal to 2 (one access and one egress link) it is assumed that this node is not an intersection in reality. All the remaining nodes are further analyzed by calculating whether a turn movement should be made at this and if yes; which type of turn movement. This differentiation will be based on the angle between the incoming and exit link. This angle is calculated based on the GPS coordinates of the three nodes that are involved. Intersections in which a driver should turn are given a relative higher weight within the calculation of the route complexity. The type of movement, straight-on, left turn, right turn or u-turn, will be deducted by means the angle between the links. If the angle between the incoming and exit link is between 350 to 360 degrees or between 0 and 10 degrees the movement is defined as 'straight-on', when the angle is between 10 and 150 degrees the turn is defined as 'right'. When the turn angle is between 150 and 210 degrees the turn movement is defined as a 'u-turn' and lastly an angle between 210 and 350 degrees is defined as a left turn. The final score is computed by summing the number of turns. Straight movements will be weighted with a value of zero, right turns are weighted with a score of 1 and u-turns and left turns are weighted with a score of 2. For more information concerning the type of turn and the accompanying angle see Figure 10. It should be noted that this figure is indicative and that the angels itself are however out of scale.

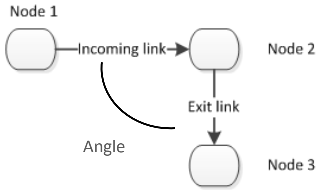


Figure 9: Representation of route complexity

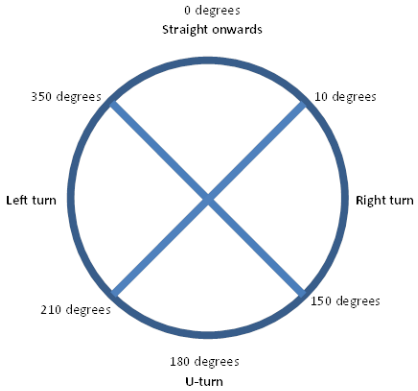


Figure 10: Visualization of type of turn

Travel time **reliability** is defined as the consistency of a given trip's travel time. It is possible to consider reliability in the historical sense such that a distribution of travel times is examined, and specific statistics can be reported, such as the mean, median, standard deviation, variance, etc. Thus, the degree of historical variability of travel times can be reported as the reliability of a particular repeated route. The reliability of the route is computed based on the floating car data from the KATE platform. Every trip in the system is subsequently matched to the infrastructural network. Based on the revealed route, the free flow travel time and the location timestamps of each separate trace the revealed travel time on each link of the route are calculated. These revealed link travel times were be stored in a database for off- and on-peak measurements. Based on the separate unique measurements the variation of the average link travel time is computed. The sum of the variation for all links on a specific route is used as an indicator for the travel time reliability. It should be noted that this methodology assumes that all individual link travel times are independent experiments which may not fully represent reality. Due to the limited penetration rate of the 'ReisAlarm' application and the limited amount of data that is currently available it should be noted that this attribute does not provide a full coverage of the complete infrastructural network. However due to the unavailability of other data sources that provided an alternative with a higher penetration rate it was decided that, although the information is not completely reliable, this attempt could support a 'proof of concept'. Based on all revealed trips from the KATE platform the reliability database covers 11,7% of the total infrastructural network. If links within the network are not represented in the database an estimate is generated based on the average link reliability for that specific link category which is also corrected for the link distance.

The revealed link travel times are explicitly not included in the travel time indicator. Due to the importance of the travel time attribute within the whole system architecture it was decided that utilizing an 'experimental' indicator with a limited coverage of the infrastructural network is not advisable.

The computation of the **directness** of each proposed route is fairly straightforward, the Euclidean (aerial) distance between the origin and destination coordinates will be divided by the total travel distance of each path. The total distance of a path cannot be smaller than the Euclidian distance, this results in a value that ranges between 0 and 1 and the route with a value close to 1 is considered to be more attractive. For example a route with an Euclidian distance of 17 kilometers and a total distance of 28 kilometer results in a score of 17/28 = 0.607.

The **familiarity** with the network can be calculated from the number of times a driver previously travelled on a specific link, an individual database for each user was generated that included all links and the number of times the driver has travelled on each specific link. After processing each trip this database is updated based on the revealed route. The familiarity of a route is subsequently computed by dividing the total travel-distance on familiar roads by the total travel-distance of the proposed path.

NORMALIZING THE POSSIBLE PATHS

It is not possible to use the specifically observed values for each attribute within the learning model. Main reason is that the routes within the dataset for the DTL algorithm will consist of multiple trips with significantly varying characteristics in terms of trip length and trip duration. In the learning model therefore relative values of the attributes over a reference route of each origin destination pair are used. Often in reality a driver also generally selects a superior route among a set of feasible routes based on comparison and within this thesis the reference route is selected based on the travel time. For example TABLE 2 visualizes the observed values for a trip between Delft and Leiden, based on the observed values route 5 is selected as the reference route. Based on the attributes the relative values, as shown in TABLE 3, can be deducted.

The disadvantage of the structure in which all attributes are relative is that it is difficult to translate to attributes to measureable values after the learning process, is is difficult to interpret the results of one specific trip after the normalization step. Moreover the selection of a 'best' route before initiating the learning process introduces a bias within the learning algorithm towards the travel time.

	Travel time	Travel distance	Aversion	Complexity	Reliabilty	Directness	Familiarity
Route 1	32,159	50,999	30,6	33	8,516	0,644	0,978
Route 2	32,191	49,492	30,0	31	8,234	0,664	0,956
Route 3	32,643	53,111	23,0	34	8,322	0,619	0,8
Route 4	32,191	49,492	30,0	31	8,423	0,664	0,956
Route 5	31,951	51,008	21,3	34	8,428	0,644	0,8

Table 2: Observed attribute values for a route set

	Travel time	Travel distance	Aversion	Complexity	Reliabilty	Directness	Familiarity
Route 1	1,00650997	0,99982	1,43662	0,97059	1,01044	1	1,2225
Route 2	1,0075115	0,97028	1,40845	0,91176	0,97698	1,03106	1,195
Route 3	1,02165816	1,04123	1,07981	1	0,98742	0,96118	1
Route 4	1,0075115	0,97028	1,40845	0,91176	0,99941	1,03106	1,195
Route 5	1	1	1	1	1	1	1

Table 3: Relative attribute values for each path set

PRE-PROCESS

At the beginning of the learning process an initial model is necessary for classifying the first set of routes. Within this thesis a ‘fixed’ predetermined model is not suitable since the main aim of this thesis is to integrate the individual attributes and preferences within the route choice algorithm. As an alternative it is possible to build the trainings dataset based on the observations of a pilot study or by selecting a small part of the original database as a preliminary training dataset. The latter methodology of selecting a section or the original dataset will be applied within this thesis

According to the research of Park et al. (2007) the minimum size of data required for building the initial decision tree depends on the number of attributes, since the decision tree algorithms cannot be conducted until the number of trainings examples is larger than the number of attributes. Although, from a theoretical point of view, a larger trainings dataset will produce a more reliable model this research utilize the first 10 trips as trainings dataset for the initial model. Main reason for this decision is that a larger amount of trips will further reduce the amount of trips that can be utilized as test trips. Furthermore, because of the learning process, the trainings dataset will automatically be increased when the model is incorrect. When the model is not able to produce a reliable model based on an insufficient amount of trainings data, the predictive performance should show a steep learning curve that reflects the underfitting of the model.

It was a deliberate choice to use the first 10 trips of a user for the trainings dataset instead of a random selection of trips. When the algorithm is implemented in reality also only the historically available data is available to generate the initial model.

IMPLEMENTING THE LEARNING PROCESS

As previously stated the initial model will be built based on the first 10 trips. After building the initial model the remaining trips will be used to test and update the model. For each trip within the test-dataset the possible paths are classified. Within this classification each path is tested by comparing the attributes from the path with the decision rules within the decision tree. As a result every path is classified by either a '0' or '1' (binary classification). The '1' indicates that the path corresponds with the user model and is therefore 'attractive'. The output '0' indicates that the path does not correspond with the user model and is therefore classified as 'unattractive' .

After individually classifying all possible paths the classifications are evaluated. Ideally one path within the set of proposed paths should be evaluated as attractive but because all classifications are performed independently also zero or multiple paths can be evaluated as attractive. If one path is selected to be attractive this path is automatically labeled as the 'predicted route'. Subsequently the model continues with comparing the predicted path with the revealed route. If multiple paths are selected to be attractive the model classifies the path with the shortest travel as the 'predicted route' and compares this predicted path with the revealed route. If zero paths are attractive the test is classified as a failure and the model will automatically continue with adding the test trip to the trainings data and rebuilding the model.

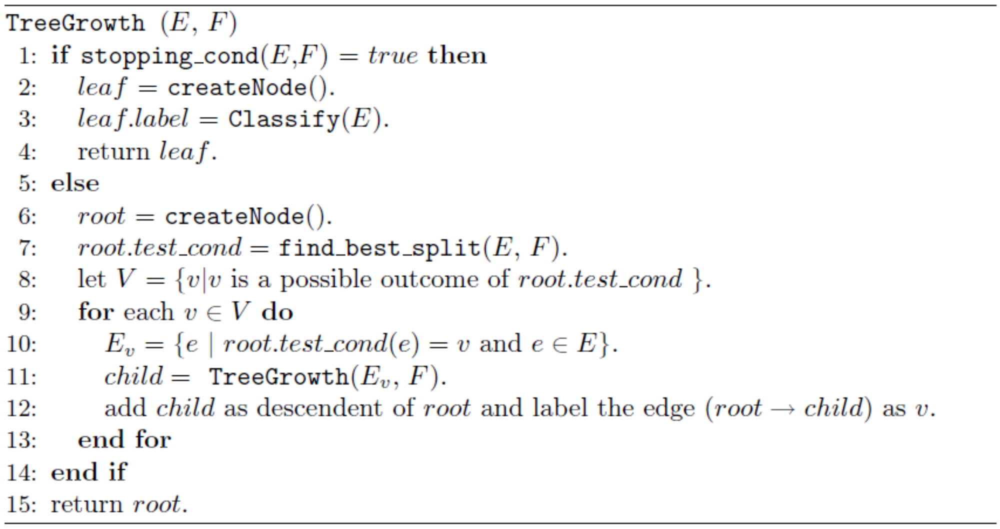
After classifying the possible routes the user feedback is collected. When one of the paths is classified as the predicted route the user feedback is collected. This is done by comparing the revealed route and the predicted route. When the predicted route corresponds with the revealed route the test is classified as a success, subsequently the test trip is added to the trainings dataset but the model is not rebuild. Main reason is that, due to the correct prediction, the model is still valid. If the predicted route however does not correspond with the predicted route the test is classified as a failure. Based on the previous trainings database, to which the data from the unsuccessful test route will be added, the model is recalculated.

As depicted in FIGURE 8 this process above continues until all the route choice data for a specific user has been used and the model than continues with the next user.

The proposed routes for each origin/destination relation does not necessarily include the revealed route. Based on the link identification numbers of each proposed route, the link identification numbers of the revealed and the corresponding link lengths (in meters) it is possible to calculate the overlap between the possible paths and the revealed route. Based on this overlap it is possible to determine whether the revealed route is represented within the proposed routes. When the revealed route is not available in the trainings dataset the revealed route will be added to the dataset. During the tests the revealed route is intentionally not added; main reason is that in reality the algorithm is applied before making the actual trip and the revealed route will be not available at that moment. Although described earlier, this same methodology will also be applied when generating the initial model.

Although the previous text elaborately described the process and steps within the decision tree learning algorithm, it did not describe the code-based implementation of the algorithm. Based on the theories and mathematical interpretation described in chapter 2.4 the decision tree learning algorithm can be translated to the pseudo code which has been described in TABLE 4. This broad skeleton for the creation of the decision tree is visualized in the book

‘Introduction to Data Mining’ that is written by Tan, Steinbach and Kumar (2006) which introduced a decision tree induction algorithm called ‘TreeGrowth’. The input to the algorithm consists of the training records E and the attribute set F. The algorithm recursively selects the best attribute to split the data and expands the leaf nodes of the tree until a stopping criterion is met.



```
TreeGrowth $(E, F)$
if stopping_cond$(E,F)$ $= true$ then
    $leaf$ = createNode().
    $leaf.label$ = Classify$(E)$.
    return $leaf$.
else
    $root$ = createNode().
    $root.test\_cond$ = find_best_split$(E, F)$.
    let $V = \{v | v$ is a possible outcome of $root.test\_cond$ $\}$.
    for each $v \in V$ do
        $E_v = \{e \mid root.test\_cond(e) = v$ and $e \in E\}$.
        $child$ = TreeGrowth$(E_v, F)$.
        add $child$ as descendent of $root$ and label the edge $(root \rightarrow child)$ as $v$.
    end for
end if
return $root$.
```

Table 4: Pseudo code for implementing the decision tree (Tan, Steinbach, & Kumar, 2006)

Within the pseudo code above some references to external functions are made. The ‘createNode()’ function extends the decision tree by creating a new node. A node in the decision tree has either a test condition, described as node.test_cond or a class label noted as node.label. The function ‘find_best_split’ determines which attribute should be selected as the test condition for splitting the training records. The choice of the test condition depends on the impurity to determine the goodness of fit. A measure for the impurity is the previously described entropy function which is described in chapter 2.4. The Classify function determines the class label to be assigned to a leaf node. Each leaf node *t* the fraction can be denoted as P(i|t) from class I associated with node t. In the majority of cases, the leaf node is assigned to the class that has the majority number of training records:

$$leaf.label = \underset{i}{argmax}\ p(i|t)$$

Where the argmax operator return the argument I that maximizes the expression *p(i|t).* Besides providing information that is required to determine the class label of a leaf node, the fraction *p(i|t)* can also be used to estimate the probability that a record assigned to leaf node t belongs to class *i*. Lastly the stopping_cond function executes a stopping command by testing whether all labels have the same class label or the same attribute values.

After building the decision tree, a tree pruning algorithm can be applied to reduce the complexity of the tree. For more information concerning pruning please refer to chapter 2.4.

The actual implementation of the C4.5 algorithm in this thesis is facilitated by the open source Data Mining Software WEKA, which is developed by the University of Waikato. This software package contains a package of machine learning algorithms for data mining tasks (Hall, Eibe, Holmes, Pfahringer, Reutemann, & Witten, 2009). The algorithms are available in either a dedicated graphical user interface or can be accessed directly from the Java code.

Within this research the J48 classifier from the WEKA software suite has been integrated within the Matlab code. This Java class based algorithm can generate an unpruned and pruned C4.5 based decision trees. The J48 classifier has been developed based on the original publication of Quinlan (1933) in which the original C4.5 outline has been described.

4. Processing the GPS-based data

The proposed system architecture from section 3.6 described the envisioned implementation of the 'Personalized Adaptive Routing Algorithm'. One of the main input elements of the algorithm is the trip database derived from the KATE mobile data acquisition platform. As described the main output from the application are provided by the location traces which include the longitudinal position, lateral position, timestamp and the accuracy of the measurements. As depicted in the overall system architecture these location traces have to be translated towards trip information that include the origin, destination and the revealed route.

One of the main challenges within this project was the data processing activity in which the raw locational traces were translated to specific trip information. Important factors were the scale and subsequent complexity of the data-set that originated from the data acquisition platform. Data from multiple devices were logged and needed to be stored in a robust system that is both efficient and insensitive to errors. To achieve this it was necessary to pay attention to data reduction, data compression, the distribution of data towards multiple instances and the efficiency of the data analysis itself.

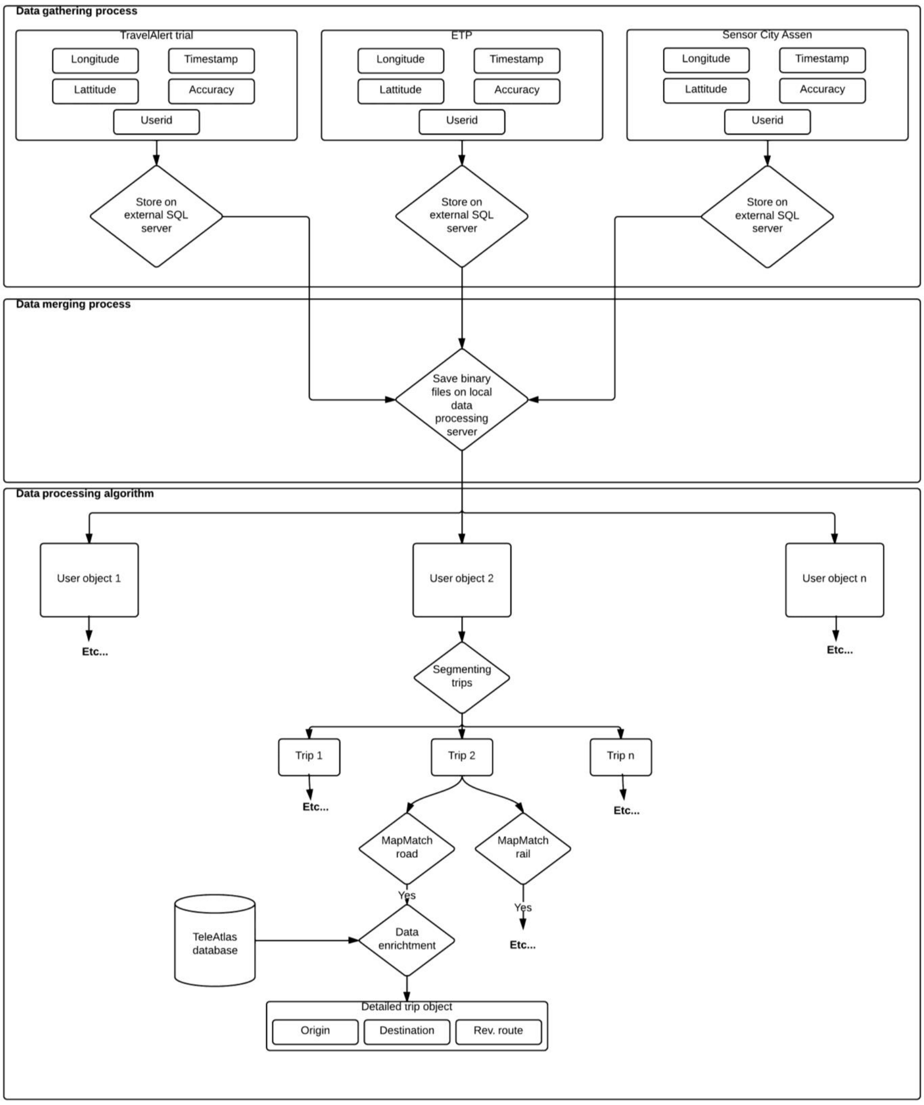


Figure 11: Data processing structure

Main focus was to design an efficient data processing script which aimed to gather, merge and process the raw locational traces from the KATE platform. These three steps will be further discussed below. Main aim is to explain the trip representation and to give the reader a sense of the magnitude of the data gathering and implementation processes.

FIGURE 11 visualizes the data processing structure, a larger copy of this figure has been added as appendix 3. As can be seen in the picture the data processing has been separated in three major parts; the data gathering, the data merging and the data processing. The text below will discuss each of these elements.

4.1 DATA GATHERING PROCESS

As discussed previously in this report, the KATE mobile data acquisition platform is utilized to differentiate trips and the revealed routes of the subjects who were participating in one of the tree projects that employed the 'ReisAlarm' application.

Data for this thesis was derived from all three projects in which the 'ReisAlarm' application was used, the data gathering started at the 24th of September 2012 and ended at the 1st of April 2013. Within this period, in total over all three projects, the application was installed on 149 unique devices. After installation, the location traces algorithm was automatically activated.

During the data collection the application determines the location of the device in a five minute interval, this measurement is preferably based on a location derived from a network triangulation to preserve the battery of the device. Only when the network triangulation fails the device falls back on a GPS based location.

Based on the distance travelled between each successive measurement the application determines whether the device is stationary or moving. The configuration of the KATE platform is flexible to accommodate the specific characteristics of the project in which it is applied. During this master the a specific configuration is applied to determine the status of the participants. The criteria in which the application determines if the device is moving is based on the aerial distance and time between two subsequent measurements. If the aerial distance between two subsequent measurements (thus 3 consecutive locations traces) are both larger than 500 meters the application switches to the 'on-trip' mode in which the time interval between measurements is reduced to one minute. Furthermore the application will mainly rely on the GPS chipset for its locations while on-trip.

If the location of a device does not change for three successive measurements the application will decrease its update interval back to one measurement every five minutes and switches to the 'post-trip' mode. If the location of the device subsequently remains similar for the next 30 minutes the application will change its status to stationary. If the application however detects that the device is moving again during the first 30 minutes the application will directly switch back to the 'on-trip' mode without any delay.

During the data collection subjects were asked not to alter their behavior in any way. The location tracing algorithm was started together with the Android operating system and functions completely autonomous, besides installing and registering the application the subjects did not have to attend to the application in any specific way.

The application of the ETP Behavioral study and the application utilized on the Sensor City Mobility project only consisted of the zero-measurement functionalities and did not include any route advices. These application therefore only recorded the subjects natural driving routines. The validity of the data of the TravelAlert internal test is more complex because this application actively supplied a route advice based on calendar items. Within this thesis only a certain amount of data of the TravelAlert test is included. To preserve the data quality (i.e. all data that is not influenced by travel and routing advices) all trips in which the user has opened the main user interface of the application prior (3 hours) or during the trip are deleted. This split is possible because the app logs all the user activities within the application.

All locational traces were stored at an external SQL server; the communication is facilitated by means of an Open Database Connectivity application programming interface. To preserve the privacy of the users the external server is password protected and additionally the access to the off-site server is controlled by only allowing specific physical addresses (IP-addresses) of the machine(s) that access the data on the server.

Main advantages of employing a SQL database server is that all available data (location traces, questionnaires, etc.) is centralized and organized in a well-structured fashion which increases the flexibility and user-friendliness of these systems. A user can pull up the required information with a query that is based on specific keywords.

In total, between the 24th of September 2012 and the 1st of April 2013, 1.581.006 individual GPS measurements have been collected over the three projects in which the 'ReisAlarm' application is employed.

4.2 DATA MERGING PROCESS

One disadvantage of the SQL database configuration that was employed in this study is its limited efficiency and speed, the database structure enables the researcher to quickly access all GPS data logs of one specific users. However, when traces from multiple days are requested, or worse when all data for multiple users over multiple days is requested the database becomes slow and unresponsive.

The multi role, multi attribute and flexibility characteristics of the database server hindered the performance and efficiency of the data processing scripts employed within this thesis. Especially in combination with the latency due to the physical separation of the database server and processing server a middleware solution was required which provided a link between the two components.

To process the data from the KATE platform all the primary information from the location traces (timestamps, latitudes, longitudes, accuracy) are stored locally on the processing server in binary files. By defining recurring structures, deleting the non-critical data and by storing the data in binary files locally on the processing server the efficiency and speed of the data processing scripts is maximized.

To preserve the privacy of the participants no information is stored on the local processing server that directly relates to personal attributes of the user, to differentiate the data for an individual user an anonymous user identification code is utilized that is based on 72 random characters.

The file structure described above represents the main input for the processing scripts to derive the trips and the associated revealed routes for each user. All additional information, such as the questionnaires and the in-app users activities are collected ad-hoc by accessing the database server.

4.3 Data processing

Chapter three discussed the methodology employed in this research, this methodology purely focused on the implementation of the C4.5 decision tree learning algorithm. The main elements for this master thesis are the GPS data logs that describe the travel behavior and the TeleAtlas shapefile that describes the infrastructure. However to be able to leverage this data it is necessary to identify the individual trips and the route that has been travelled within each journey. To further investigate the travel behavior of each user it is necessary to link all the on-trip locational traces to the available infrastructural network.

Segmenting the traces into separate trips

In the period prior to this research a large database with existing data processing scripts was available within TNO. Within this thesis the existing code is integrated and applied. Based on two subsequent locational traces it is possible to determine the status of the driver, by calculating the distance between the two locations and the time between the two measurements. A radius of 50 meters is assumed for spatial extension of a single stationary location, movements that stay within this radius are not considered significant and the corresponding state will be labeled as 'stationary'. If the radius is larger than 50 meters and the speed between the measurements is larger than 10 km/h the measurement is labeled as 'moving'.

It can occur that a user makes a stop within a trip, for example the participant may drop off its children at school or may stop to refuel. To investigate the travel behavior it is important to incorporate these events as stops and not as multiple separate trips. Based on this statement it is assumed that if the time between to traces that are defined 'moving' is less than 30 minutes the user has made a stop. The separate elements (legs) are processed but the trip is summarized as a whole.

In reality the intermediate destinations can be a decisive factor within the route choice and should preferably be included in the path generation algorithm as waypoints or within the evaluation by means of a separate indicator in the evaluation framework. However during this study it was not possible to apply both options due to the design choices within the system architecture of the KATE platform. Another possibility was to delete all the trips in which intermediate stops have been detected to further clean up the trip database, however due to noise in the data (in which for examples traffic jams and traffic light waiting times were detected as intermediate stops) this would have let to a significant reduction of the data which would have hampered the implementation of the learning algorithm.

MapMatching the GPS traces to the infrastructural network

The alignment of the GPS traces to the infrastructural network, described by the TeleAtlas Shapefile, is a challenge. Main difficulty is that each locational trace contains inaccuracy. The exact inaccuracy is dependent on the source from the locational trace. If the trace is based on a mobile network triangulation the inaccuracy is often around 1000 meter. When the trace is based on GPS positioning the accuracy is approximately 10 meters but this inaccuracy may increase significantly in urban areas, at bridges and tunnels where the performance of GPS in terms of signal strength is limited (the device requires a direct line of sight towards the satellites).

An additional problem is the accuracy of the underlying infrastructural network described in the TeleAtlas database, the representation of individual links and nodes can be inaccurate due to recent infrastructural or spatial changes. Especially the representation of the lower level roads within the TeleAtlas database is known to have small inaccuracies.

Due to the noise within the positioning algorithm and the inaccuracy of the underlying network nearly all GPS traces fall onto off-road zones. This implies that the locational traces need to be snapped to the network nearest to it and to connect all the on-road locational traces by means of an algorithm that calculates the shortest route between the traces. To match the locations to the available infrastructural network the MapMatcher application of TNO is utilized. This application, which is available through a webserver, considers all locational traces, the individual accuracy and time of each trace and calculates the most probable route.

IDENTIFYING MULTIMODAL TRAVEL BEHAVIOR

One of the difficult items in using mobile smartphones to gather locational traces is that, based on the locational traces, it is not possible to assume that the user has travelled by means of private transport. Where some previous researches used in-car technology that were activated and deactivated based on the car's ignition power the data from the KATE platform should be post-processed to exclude trips in which the public transport was utilized. Within the public transport modalities it is possible to deduct two major classes of transport, the first is the public transport that utilizes a separate infrastructural network (e.g. trains) and the public transport that shares its network with private transport (busses and trams).

Trips that are made by means of the first class of public transport can be deducted by MapMatching the locational traces on their respective infrastructural network, within this master thesis the locational traces will be MapMatched to both the road- and rail network. If the results indicate that the majority of the locational traces within one separate leg of the trip can be explained by means of the rail network the trip is not included as input for the decision tree learning algorithm. If more than 80% of the locational traces can positively be linked to the rail network and when the distance travelled within this leg is more than 2 kilometer the trip is omitted as input for this master thesis.

The second class of public transport is very difficult to deduct, due to the shared network the only methodology to deduct these transport modes is to combine the stops that are recognized en-route with the scheduled timetables. However this requires a detailed data quality, for example the on-trip update interval that was utilized does not allow the researchers to trace stops to load and unload passengers. Moreover the timetables of the lines were not available within this thesis. Due to these factors it has currently not been possible to eliminate trips that were made by public transport that utilizes the same infrastructural network as to car from the trip database.

Due to the minimum speed of 10 km/h that is necessary between two subsequent measurements, trips that are made by foot are not recorded by the location tracing algorithm. On the other hand cyclist can have a speed higher than 10 km/h and moreover share the same network as private motorized transport. This implies that trips by means of the bicycle cannot be deducted from the trip database.

That either trips by bus, metro, and cyclists cannot be directly deducted from the trip database does not imply that there is no method to filter these trips to ensure the data quality during the implementation of the DTL algorithm. This study contains two separate analyses based on the data derived from the KATE platform. The first will generally describe the dataset and the travel behavior of the participants, within this analysis the trips made by these modes of transportation will be kept inside the trip database. During the second analysis, which will further discuss the implementation and evaluation of the personalized routing algorithm, care should be taken only to include trips made by car. Within this analysis a number of filters have been applied to clear out the trips made by other modes than train. The method to filter these trips will discussed in a future part, chapter 6.1, of this report.

DATA ENRICHMENT

Based on the steps described in the previous section a database of separate trips is derived and the locational traces within these trips are coupled with the infrastructural network to deduct the revealed route. However by only investigating the origin, destination, timestamps and identification code of each trip no valuable conclusions can be made concerning the travel behavior. To improve the usability of the dataset it is necessary to further enrich the data to enhance, refine and improve the raw data.

One of the data enrichments that is made during this research is the utilization of additional attributes of the TeleAtlas shapefile. Main output from the MapMatching application are the link identification codes (linkid's) of the most probable route. Based on these linkid's, that directly relate to the unique identification codes of the TeleAtlas shapefile, we are able to further investigate the characteristics of the revealed route. Within the TeleAtlas shapefile an extensive range of attributes is included which describe each link within the network. For example the link length, free flow travel time, road class type and road name can be de deducted based on the linkid.

The data enrichment within this master thesis is focused on deriving the attributes that will be utilized within the implementation of the decision tree learning algorithm. These attributes have been described in the outline of the system architecture in section 3.6.

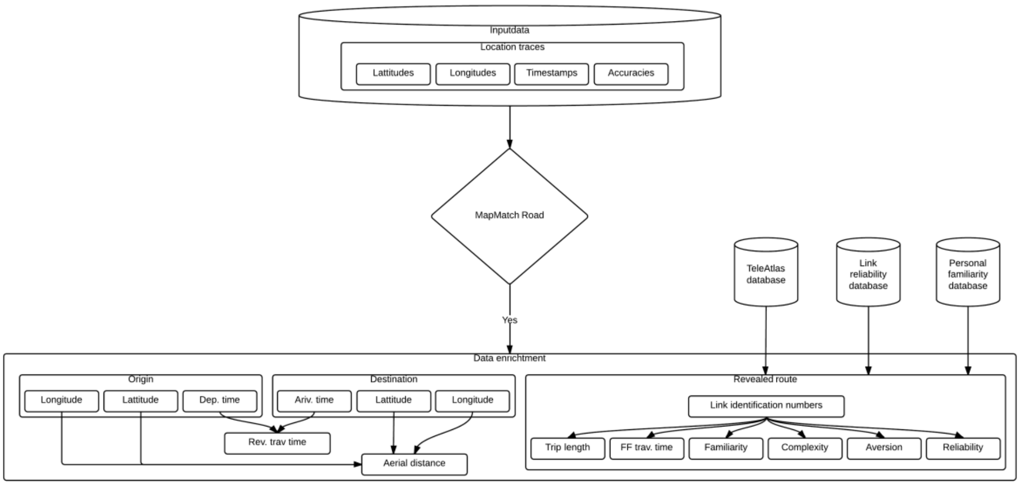


Figure 12: Data enrichment structure

A more detailed structure for the data enrichment is described FIGURE 12 moreover a large version of this flow chart has been added as appendix 4. As depicted in the picture the origin, destination and link identification numbers are deducted from the MapMatch component. Based on the coordinates of the origin and destination we are able to deduct the aerial distance and moreover both the departure time and arrival time can be translated into the revealed travel time. If the link identification numbers are subsequently coupled with the TeleAtlas infrastructural database the trip length in kilometers, the free flow travel time, the complexity and aversion can be derived.

The second enrichment of the data is based on the external link reliability database, which describes the historic revealed travel times for all links, from all users, that were measured during one of the projects in which the KATE platform was employed. This database separates both peak hour and non-peak hour measurements. This database contains separate revealed link travel times for each specific link identification number for each period. To represent the reliability all measurements are combined to calculate the variance of the revealed travel time on that specific link. After a trip is processed the external database is automatically updated to include the revealed link travel times of the traversed links. It is however important to note that, as discussed before in chapter 3.7, the coverage of this database is limited; currently 11.7% of the total network is available.

The last enrichment that is performed represent the familiarity of a route. Based on the individual familiarity database, which describes the links that were previously traversed by this specific user and moreover contains how often each link was traversed, we can calculate the familiarity of all links that were traversed during the trip.

5. GENERAL ANALYSIS OF THE TRIP DATABASE

The preceding part of this report focused on the subject, system architecture, underlying principles and the proposed methodology. As mentioned earlier the data derived from a mobile data acquisition platform will be deployed to study the predictive performance of an 'Personalized Adaptive Routing Algorithm'.

The KATE mobile data acquisition platform utilizes the technology within smartphones to reveal the travel behavior of a group participants. As discussed in chapter three the platform has been used in three different projects within TNO. Although small elements of the projects differ, it can be said that the essential key features off all projects are similar.

Within this master thesis the data of the three projects is integrated to generate one major trip database, this integration has been done by first processing all the projects at the individual user level by generating a database of trips that were made by each individual user.

Subsequently the results from each project were integrated based on the hardware id (IMEI number) of the smartphone. When separate sets of trips were found within multiple projects, and if the hardware id was equal, the two sets were combined in one user database. If, during the test, the participant installed the experiment on a new telephone this resulted in separate database entries.

After merging the project databases to one combined database, the database was further post processed and structured. All trips in which the MapMatcher was unable to deduct the revealed route or when the MapMatcher detected a trip by train were deleted. Furthermore all the users in which only a few locational traces were detected but showed no actual trips were deleted from the database.

In total between the 24th of September 2012 and the 31st of March 2013 1.581.006 locational traces were detected. The total number of trips that was detected was 11.490 of which 1172 trips (just above 10%) were made with train. In total 101 unique devices were used in the three experiments and these devices were registered to 95 unique users, of which 78 were male and 18 were female .

The total distance travelled within the experiments was 277.021 kilometers, the distance travelled each day is visualized is FIGURE 13. Within this figure it becomes especially apparent that the distance travelled each day gradually increases between September 2012 and the beginning of December 2012, in this period the majority of the trips were gathered within the ETP project. Between December 2012 and the January 2013 the distance travelled each day decreases because most of the participants de-installed the application. In the middle of January 2013 the 'ReisAlarm internal test' application was distributed which significantly improved the amount of data that was gathered. A high resolution and larger copy of FIGURE 13 is added as appendix 5.

The average distance travelled on each trip is 26,8 kilometer. To put this number in perspective it is important to investigate the trip length distribution. FIGURE 14 visualizes the number of trips that fall within a specific distance interval, what especially becomes apparent is that the majority of the trips is shorter than 10 kilometers. Between 0 and 50 kilometers a steep descent is apparent, however above 50 kilometers the number of trips in each interval is fluctuating without showing a significant trend. The magnitude of the last distance bin (100 km and larger) is obvious due to the fact that this bin included all remaining trips.

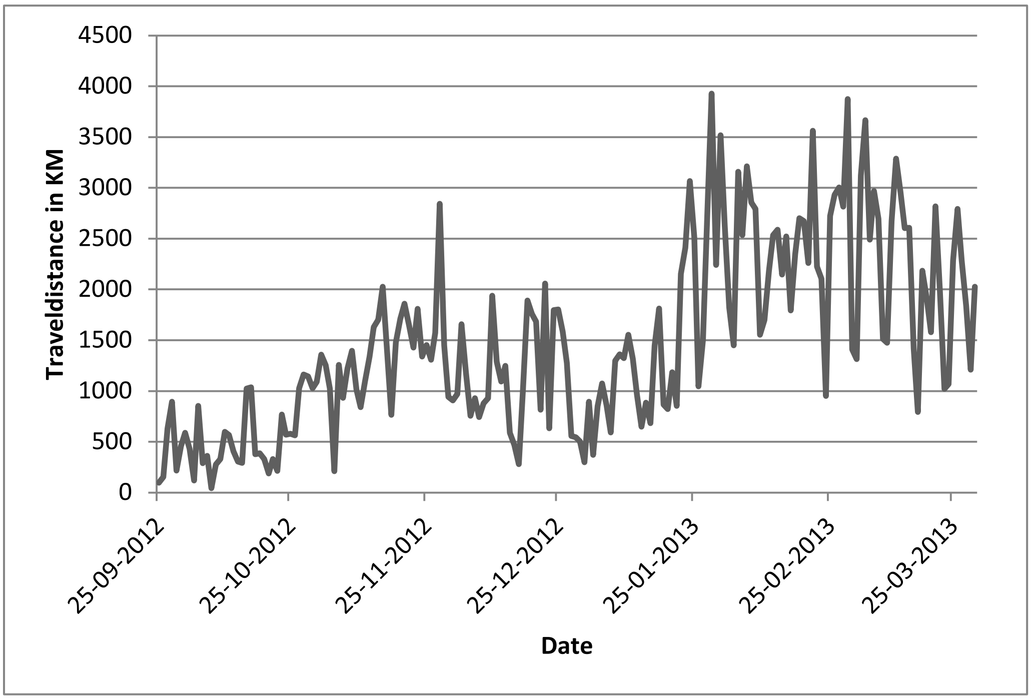


Figure 13: Daily distance travelled

Often the distribution of trips among the week, and the subsequent mean travel distance on each day of the week, is interesting to analyze the travel behavior for different motives. For example most trips during the working days are work related and most trips in the weekends are leisure related. FIGURE 15 describes the number of trips that were registered on each day of the week. What becomes apparent is that although the number of trips on Saturday and Sunday are lower than the other days, the difference is not shockingly big.

FIGURE 16 also visualizes the average trip distance of each day of the week, the difference between the days is less obvious than the figure that described the number of trips. Apparently, although the number of trips on Sunday is lower the trip distance is higher than for example on Saturday. Apparently travelers reserve their Sunday for long distance trips and stay closer to home on Saturday.

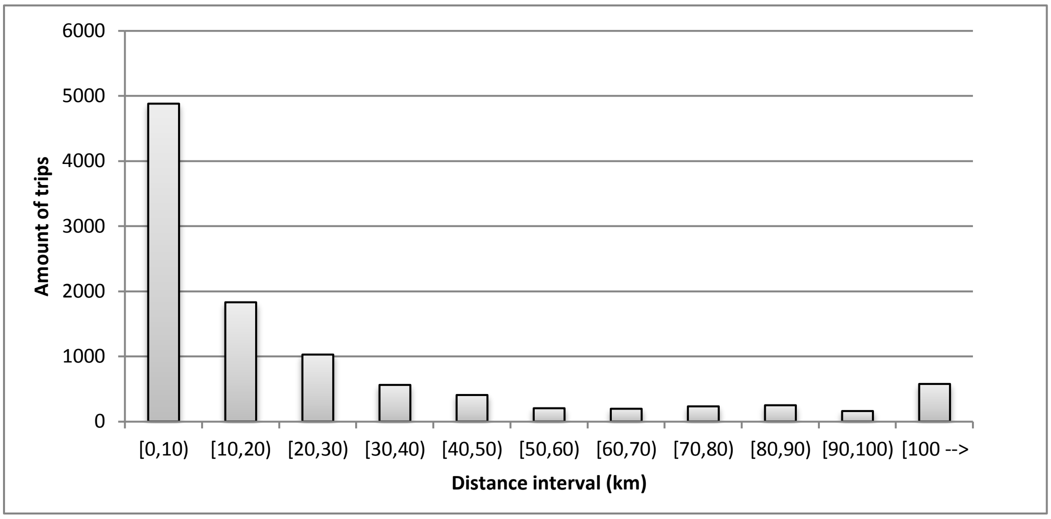


Figure 14: Number of trips per distance interval

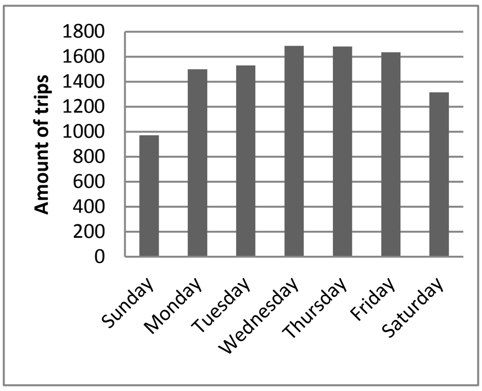


Figure 15: Average number of trips for each day

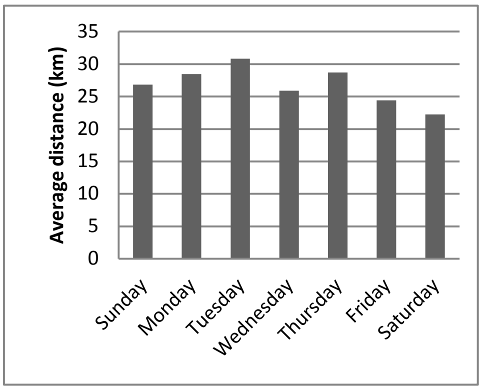


Figure 16: Average trip distance for each day

Although the total number of trips during weekdays and weekend is only slightly different the variation in departure times during weekdays is significantly different in comparison to the departure times during the weekend. FIGURE 17 visualizes the average number of trips that have been started during each hour of the day. Moreover this figure differentiates the weekdays and the weekends. What becomes clearly apparent is that especially during the weekdays the departure times are concentrating between 0800 and 09.00 in the morning and between 17.00 and 18.00 in the evening which correspond to the 'normal' office hours. In the weekends however, the trips are more spread during the full day with its peak between 15.00 and 16.00.

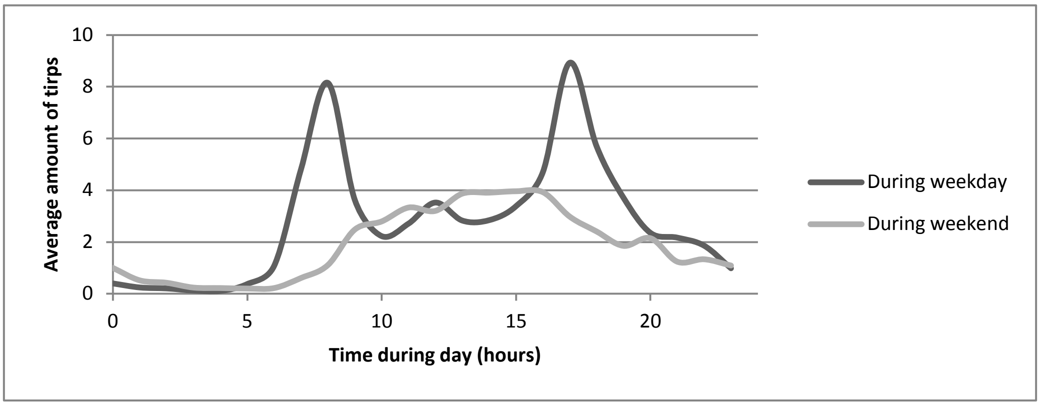


Figure 17: Distribution of trips during day

Although this thesis did not aim to obtain a representative sample of the Dutch population, it is important to compare the results of the sample with the general Dutch mobility data to assess the quality of the data derived from the KATE platform.

Statistics Netherlands, in Dutch the 'Centraal Bureau voor de Statistitiek', is responsible for collecting and processing data in order to publish statistics to be used in practice, by policymakers and for scientific research. Continuous research on travel behavior of people in the Netherlands has been carried out since the beginning of the 1980s. Data on the mobility of individuals is collected by means of a survey of the Dutch households. Based on these survey results, estimates of the mobility of individual residents of and travelling within the Netherlands are made by means of an increment and weighting process (Centraal Bureau voor de Statistiek, 2013). The results of the surveys facilitate the comparison of travel behavior derived from the KATE platform.

TABLE 5 is derived from the survey of 2011 (Centraal Bureau voor de Statistiek, CBS Statline - Mobiliteit in Nederland; mobiliteitskenmerken en motieven, 2011) and describes the main travel time and travel distance per day in general and for each specific gender.

Data derived from the 'Mobiliteit in Nederland' survey			
	Number of trips	Distance travelled	Travel time
Male and female	2,62	26,78	55,71
Male	2,54	31,31	58,52
Female	2,69	22,3	52,94

Table 5: Key mobility indicators from the 'Mobiliteit in Nederland' survey

Based on the data from the KATE platform, the average number of trips, trip distance and travel time for the full sample and the specific gender is visualized in TABLE 6.

Data derived from the 'KATE' platform			
	Number of trips	Distance travelled	Travel time
Male and female (n = 96)	2,51	67,51	108,95
Male (n = 78)	2,47	71,52	110,34
Female (n = 18)	2,64	53,29	104,15

Table 6: Key mobility indicators from the KATE platform

If both tables are compared it becomes apparent that the results from the sample utilized in this experiment does not represent the population of the Netherlands. Especially the trip distance and average travel time derived from the GPS database are significantly higher than the results from the CBS. However, this difference can be explained. Both the ETP behavioral projects and the TravelAlert internal tests aimed to include employees of TNO and moreover the participants in the Sensor City Mobility experiment were approached through their employers. There is a clear bias towards employed users while participants of the CBS survey are randomly selected.

One interesting analysis is to investigate how the trips and the subsequent locational traces deducted from the GPS data-set are spread throughout the Netherlands. Within FIGURE 18 the surface of the Netherlands is divided in blocks of 20 square kilometer. All the locational traces from the trips that were registered were combined and all the traces within an individual box were counted. Based on these counts a color legend was created that describes the number of traces that were detected within a specific box. Based on the figure we are able to derive three 'hot spots', firstly the surrounding of Delft and Rotterdam clearly stands out. In addition, many data points are concentrated in the vicinity of Utrecht and lastly Assen can be recognized. A fourth, but less obvious, data point lies near Eindhoven and Helmond. These data points can be explained by the fact that an important branch of the TNO Automotive department is located in Helmond. Furthermore the company that was involved as the software developer in the ReisAlarm experiment is located in the Dutch town Nuenen.

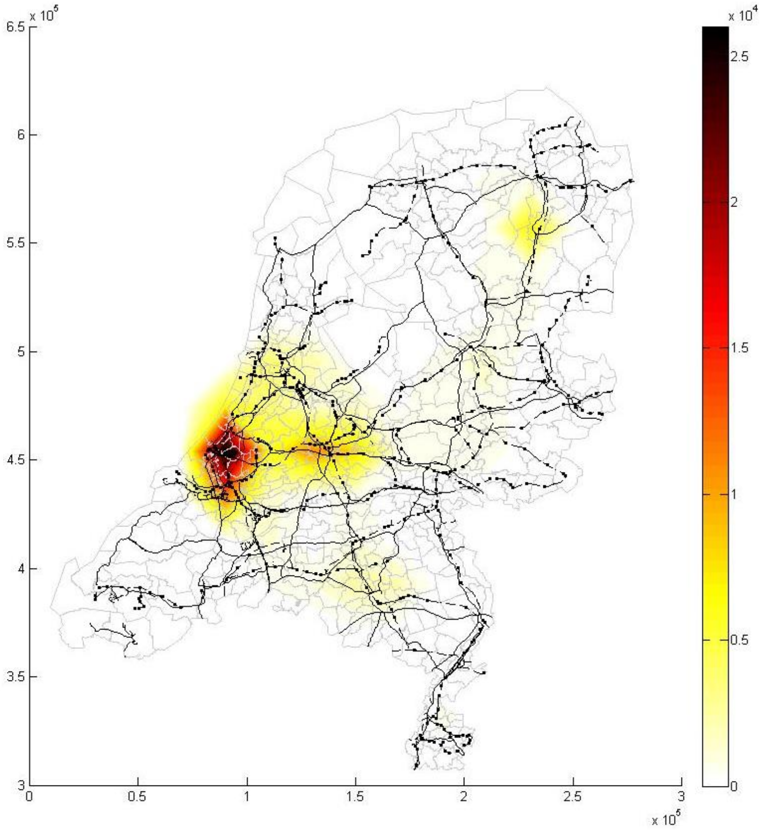


Figure 18: Location heat map

One variable that is collected within the first use questionnaire of the application is the age of the participants. Although not all participants completed this survey it is worthwhile to research the main characteristics in terms of the mobility of specific groups.

Table 7 visualizes the average trip distance and average trip duration for each user group, differentiated by age. This table also displays the number of users that is detected within the specific age interval. Although at first glance it seems that the average trip distance and average trip duration considerable varies, it is difficult to draw conclusions due to the oblique distribution of the number of users within each interval.

Average mobility characteristics of specific age intervals			
Age category	No. of users	Avg. trip distance (km)	Avg. trip duration (min.)
0-17	0	0	0
18-24	8	26,6	50,6
25-49	54	29,2	46,4
50-64	13	26,3	36,8
65+	0	0	0
Unknown	26	29,9	52,2

Table 7: Mobility characteristics of user group

6. Incorporating personalization into routing

Any driver knows that some roads are more and some are less attractive and that specific roads may take longer to traverse in the peak hours than they do at night. Most of this knowledge is gained by experience from earlier trips which subsequently influences the route choice behavior of future trips.

The dynamic aspect of travel behavior described above is not incorporated by the traditional route planners which ignore the identity and personal characteristics of a driver for whom the route is intended.

This chapter describes the implementation and evaluation of the personalized adaptive routing algorithm that utilizes the prior trips (and the associated routes) to improve the route suggestions in the future.

6.1 Refining the input trip database

The trip database from the KATE mobile data acquisition platform is used as the input database for the personalized adaptive routing algorithm. However not all detected trips are suitable to be utilized in the final dataset for the C4.5 learning process. For example trips that are not usable as input are the trips in which the origin is equal to the destination. Due to the delay between subsequent GPS locations it is possible for a traveler to drop-off an object or person before the application detects that the user is stationary. At the same time the user is probably already on its way back. The set of proposed routes subsequently do not correspond with the revealed route, therefore it is impossible to produce a matching route advice in these specific circumstances. To filter these trips, all trips where the postal code of the departure location was similar to the postal code of the destination location were removed from the dataset.

One other factor that can significantly influence the performance of the routing algorithms is the modality by which the trip has been made. For example the route choice of a cyclist in an urban area will be different than the route choice of a car driver. Furthermore the other modes of public transportation such as bus, tram and metro are mostly used on short distance urban trips. To facilitate a fair comparison between the various routing algorithms the dataset is filtered to only include long distance trips (a travel distance larger than 10km) in the final dataset. It is assumed that for trips above this distance both private motorized transport and the train are the only feasible travel alternatives. Due to missing data (i.e. timetables and routes) the trips made by regional public transport (bus and tram) is regarded as a car trip, because these modes are utilizing the same infrastructural network it is currently unfeasible to detect these trips and filter these from the database.

One of the main underlying principles of the C4.5 based Personalized Routing algorithm is that the revealed behavior is assumed to be a ‘statement of preferences’. However this implies that the algorithm requires a significant number of trips for all individual users to train and test the performance of the algorithm. Based on this requirement all users which had registered 10 or fewer trips were deleted from the database.

During the translation of the original trip original dataset towards the database that will be classified by the decision tree algorithm the number of trips has been reduced from 10318 car trips to 4345 trips. Main reason for the sharp decrease (58%) is the requirement that the trip distance should be 10 km or longer, this resulted in 4879 trips that were deleted.

The 4879 remaining trips were registered to 77 unique devices and 73 unique users which include 66 males and 17 females. The total distance travelled within these trips was 210.432 km and the time spend in traffic was 4127 hours. The average trip distance was 48.4 km and the standard deviation of the mean travel distance was 49,2 km which indicates that the distances are distributed widely.

6.2 Implementing the personalized routing algorithm

Section 3.6 described the system architecture that is applied to implement the learning algorithm. For the revealed origin and destination pair of each entry of the trip database a set of alternative routes is generated. Every individual alternative route is subsequently processed by computing the attribute values for the indicators travel time, travel distance, directness, complexity, reliability, aversion and familiarity.

As discussed within the architecture the observed values for each attribute are described for each possible path, these values are changed towards relative values in respect to the reference path (shortest path in time).

To generate the initial model the first set of 10 routes are selected which will be used to build an initial decision tree. As described in chapter four the first 10 routes were intentionally selected instead of a random selection to represent a real life implementation of the adaptive algorithm. Subsequently the remaining trips will be used to test and update the model. Each of the proposed paths is classified, when one of the predicted routes is classified as attractive the predicted route is compared with the revealed route to verify it the test was successful or not.

Main aim of this thesis is to compare the performance of the adaptive routing algorithm with the traditional routing algorithm. To facilitate this comparison also two route advices are generated that are based on static (non-adaptive) routing algorithms. The first traditional route proposal is generated based on the assumption that drivers prefer the route with the shortest path. This advice is generated by comparing the travel time of all proposed routes, the route with the shortest path is classified as 'attractive'. The second traditional route proposal is generated based on the assumption that a driver prefers the route that generates the maximal utility. This utility can be expressed as a single value based on the linear function with the same attributes that are included in the learning algorithm. The utility of route I, U_i is calculated as the following utility function:

$$U_i = 0.5 \times td - 0.5 \times tt - 0.2 \times ave - 0.2 \times comp + 1.5 \times rel + 2.5 \times dir + 3.5 \times fam$$

This utility function, and the relative weights, are deducted from the research of Park et al. (2007). Main reason to retain this utility function is the relative good performance in the research of Park et al., moreover by keeping the original function we are able to compare the results.

From an evaluative perspective, a new estimate of the function and the relative weights based on the revealed data would have been of added value for this research. However within the limited timeframe of this study, it was not possible to re-evaluate the utility function.

6.3 Evaluation framework

The comparison of the predictive accuracy when comparing the route suggestions with the revealed route will function as the main performance measure to assess the performance of the adaptive routing algorithm.

An essential characteristic of a routing algorithm is how accurately it predicts the driver's choice behavior. An efficient way to represent the predictive accuracy of an adaptive model is to calculate the percentage of predictions that correspond to actual choices made by the user over the period of time that the system has used.

However by only comparing the predictive performance of the three algorithms only part of the objectives of this research will be achieved. Despite the focus on performance oriented indicators it is also important to consider the process and intermediate results. These results can be used to analyze the applicability of DTL algorithms and specifically to analyze the learning ability of the learning algorithm in the context of route choice. By comparing the predictive accuracy of the learning algorithm at various moments the learning behavior can be investigated. Based on the learning behavior and the increasing amount of training data while testing, it is expected that the performance will improve over time. One of the advantages of the DTL algorithm is that the structure based on nodes and leafs can be easily interpreted by humans. The interpretation of the model structure is important in giving insight into choice mechanism and confidence in validity of the traditional routing models.

It is expected that the performance of the decision tree learning algorithm will fluctuate, in practice the model will be able to classify certain trips better than others. By determining the cases in which the model functions correctly and by comparing the trip characteristics of these trips with the characteristics of the trips in which the model fails more in-depth information on the performance of the DTL algorithm is acquired. For example if the algorithm is able to predict certain users better than other users it is important to gain further insight in the factors that induce these differences.

Lastly, characteristics that are inherent to the application of the C4.5 Decision Tree Learning algorithm should be further investigated. As discussed in the in-depth analysis of the C4.5 DTL algorithm the performance of the algorithm is dependent on the amount of pruning that is applied. When applying the C4.5 algorithm the amount of pruning is dependent of the confidence interval that is applied. By employing a sensitivity analysis, in which the desired confidence interval is varied within a certain interval the phenomenon of over- and under fitting can be investigated. Similarly the C4.5 is dependent on the various attributes that are applied within the model. To investigate the effect of the effect of the attributes within this master thesis a second sensitivity analysis is applied in which the model is ran while one of the attributes travel distance, complexity, aversion, complexity, familiarity and reliability is omitted. Due to the fact that the attribute travel time is required to calculate the relative values of the attributes over the reference route of each OD-pair the attribute travel time is not included in the sensitivity analysis.

6.4 Results

Predictive performance

Key characteristics of the routing algorithm is its ability to predict the driver's choice behavior. Based on the 4345 trips that were gathered the predictive performance of the C4.5 based adaptive routing and the traditional shortest path and maximized utility algorithm are compared. In total 3407 trips were tested. The difference between the 4345 towards 3407 trips can be explained by the utilization of 10 trips for each user as a trainings set. Furthermore the algorithm failed to execute the prediction algorithm for 168 trips. The reason for these failures are divers but the main reason is that there were inconsistencies during the process to generate and MapMatch the possible paths for an OD-relation which resulted in 0 possible paths.

One important precondition for successfully comparing the performance of the routing algorithms in terms of their predictive performance is the representation of the revealed route within the set of possible routes. If the revealed route is not included in the set of possible routes the algorithm will automatically fail to predict the correct route. From the 3407 trips within the database the revealed route was available within the set of possible routes in 1360 of the trips, within 2047 trips the revealed route was not within the set of possible routes. These results significantly influence the outcomes of this study; because the representation of the revealed route within the set is a prerequisite for a successful route prediction the best possible performance of each of the algorithms is 1360 out of 3407 trips. This is similar to a maximum predictive performance of 40%.

As discussed earlier the results for the C4.5 DTL based 'Personalized Adaptive Routing Algorithm' will be compared with two traditional routing algorithms. For each trip within the trip database a route prediction was generated based on each of the three algorithms, subsequently the predicted route was compared with the revealed route of the user. When the revealed route matched the predicted route the test was classified 'successful' and when the algorithm failed to predict the revealed route the test was classified as 'failure'. FIGURE 19 shows the number of correct predictions after implementing each of the algorithms. In total 3407 tests were done for each algorithm, the adaptive personalized routing algorithm successfully predicted 121 of the trips, whereas the number of correct predictions for the Shortest Path (SP) and Utility Maximization (UM) were 647 and 0. It is very striking that the utility maximization is not able to successfully predict any of the trips during the testing of the system. Apparently the attributes and the relative weights within the multi-attribute utility function are not able to represent the route choice behavior of the drivers.

	Successful tests	**Failed tests**	**Total number of tests**
C4.5 learning algorithm	121 (4%)	3204 (96%)	3407
Traditional SP algorithm	647 (19%)	2760 (81%)	3407
Traditional UM algorithm	0 (0%)	3407 (100%)	3407

Figure 19: Main prediction results

Based on the results described above it becomes apparent that the predictive performance of all algorithms is limited, the traditional shortest path algorithm achieves the best relative performance of 19%. Based on the correct predictions and the associated trip information the mean distance of the correctly predicted routes can be deducted. The mean distance of the trips in which the C4.5 based algorithm correctly predicted the revealed route was 35,9 km and the mean trip distance in which the traditional shortest path algorithm correctly predicted the correct route was 55,7 km.

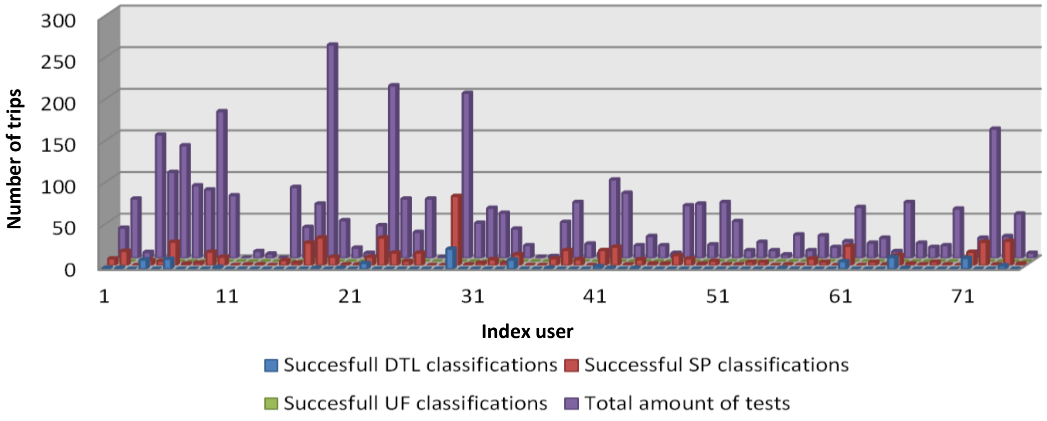


Figure 20: Distribution of correct predictions among users

Within this study the data of multiple users has been used to which all algorithms have been tested individually. It is worthwhile to investigate if the correct classifications have been concentrated in one or multiple users. The distribution of correct predictions has been visualized in FIGURE 20. The total number of tests for each user has been visualized by the purple bar in the background. The number of successful classifications for the DTL based 'Personalized Adaptive Routing Algorithm', the traditional shortest path algorithm and the multi-attribute utility functions have been visualized by the blue, red and green bars. Based on this figure it becomes apparent that none of the individual experiments significantly performed better than the others on specific users. Although at first glance it seems that an increasing amount of data (represented by the total number of tests) increases the chance for a successful prediction, this only applies to the shortest path prediction. For example the test subject with the highest amount of trips (user number 16) does not have any correct DTL classifications.

If we however compare the results in terms of the predictive performance with the number of trips that could still be predicted correctly (in the most favorable case 1360 out of 3407 trips), the results can be approximated from another perspective. Assumed that the we would only test the algorithms on the trips in which the revealed route was available in the set of possible routes (1360 trips), the predictive performance of the C4.5 based algorithm is 9% and the predictive performance of the traditional shortest path algorithm would be 48%. Based on these results there is still room for improvement, but the results do not directly provide a cause to completely repeal the DTL based and traditional distance based shortest path algorithm.

EVALUATING THE DECISION TREE LEARNING ALGORITHM

Based on the unique devices that were registered 77 individual decision tree models were determined. The initial model was based on the first 10 trips of each user, subsequently the remaining trips were used as test data. The model is updated when the revealed behavior does not correspond with the predicted.

Due to the limited number of trips that was available within the trip database for each user not every individual experiment of the DTL algorithm had the same number of updates. Every DTL algorithm was trained based on the first 10 revealed trips and subsequently the model was tested until the trips for that specific user were exhausted.

The average number of updates that was applied to the algorithm was 42 updates, however the corresponding standard deviation of 48.5 indicates that the distribution of this average value is wide. FIGURE 21 visualizes the cumulative number of individual decision tree models that reached a certain number of updates. For example all decision trees were submitted to one update and 25 of the 77 were submitted to 50 updates. This figure indicates that, as the number of updates increases, the sample to which the results are related decreases significantly. Due to the decreasing sample size care should be taken when assessing the results of decision trees that are in an advanced (fully grown) stage.

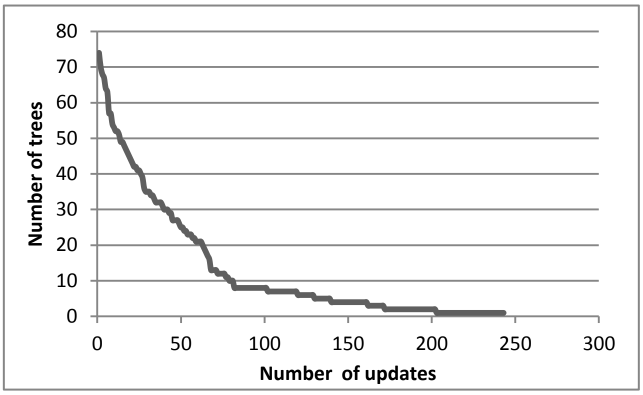


Figure 21: Distribution of the number of updates

During the process of testing and updating the model the 'training error' of the model is determined at each iteration. This error rate is computed by applying the training data to the model and by comparing the predicted classifications derived from the model with the observed classification. The error is quantified by dividing the number of incorrect classifications by the size of the trainings dataset. The error of the initial tree, averaged over all 77 users, is 4.99%. An essential characteristics of a learning model is its tendency to improve the predictive accuracy when the model is updated and functions over a longer period of time. FIGURE 22 visualizes the learning behavior of the model. The average training error at each update has been visualized on the x-axis. Moreover the standard deviation of the error has been visualized as error bars within the figure. Based on this figure we can deduct that the average and standard deviation of the trainings error marginally reduces when the period of system usage increases. Based on the standard deviation of the error it becomes apparent that the data points tend to get closer to the mean value. However, from 50 days and onwards the reduction diminishes and moreover the standard deviation fluctuates. These effects can be explained by the sample size, since the majority of the trees consists of 50 and less testing trips the data density at an extensive time of system usage is limited. Based on these results we can conclude that the predictability of the models improves marginally as more data on route choices is collected.

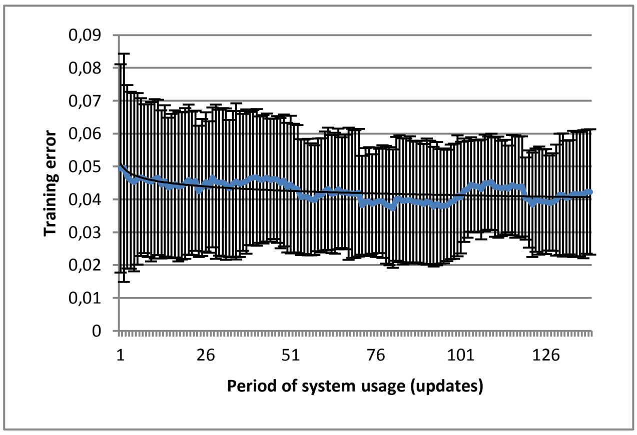


Figure 22: Average and distribution of learning error during the system usage

The marginal decrease of the training error when using the model is especially striking when compared with the results of Park et al. (2007). During the experiments in the latter research a clear improvement of the predicted performance was detected, the training error decreased from 8% in the initial tree towards 1.6% in the final tree. However to further explain the differences between these two studies it is important to further investigate the inner structure of the decision trees. If for example, the pruning mechanisms prohibited a tree size above a certain number of nodes the training size could be minimized on a certain threshold (underfitting of the model).

	Average tree size	Standard deviation of the tree size
Initial tree	4,2	2,59
Intermediate tree	9,2	8,01
Final tree	9,2	7,95

Table 8: Tree size for the various tree building stages

If the tree size of the model is examined it becomes apparent that average tree size of the initial model over all users is 4.2 nodes with a standard deviation of 2.59 nodes. The average size of the intermediate tree, when the first half of the number of test trips for a user is used, is 9.2 nodes with a standard deviation of 8.01 nodes. Lastly the average size of the final tree, when all test trips for a specific participant are used, is 9.2 nodes with a standard deviation of 7.95 nodes. If we however compare the average size of the decision trees with the results from Park et al. (2007) it becomes apparent that the trees of the current implementations are much more compact. During the research of Park et al. (2007) the average size of the final tree was 64 nodes. One of the influencing factors within the differences between the current research and the paper of Park et al. (2007) could be the pruning mechanisms that are included in the DTL algorithm, however the paper does not explicitly describe the pruning algorithm and threshold which hampers the comparison of the results. In a future section of this chapter a sensitivity analysis is conducted to gain further insight in the effect of various pruning thresholds on the current results.

However, based on the number of nodes in the various states of the tree building process it can be concluded that the structures of the initial decision trees are simpler than the trees that were built in later iterations. Since the number of training examples is relatively small to classify the instances with one or two attributes and their threshold values. However as the size of the trainings dataset increases (due to updates) the tree structure tends to become more complex to allow the model to accommodate the various cases within the trainings dataset.

Another method to gain insight in the processes and strategies of the Decision Tree learning algorithm is the tree structure, and more specifically the node labels of the first three nodes. Since the DTL algorithms splits the instances top down in a top down fashion while using the information gain as the primary indicator the model positions the ‘significantly influential’ nodes at the top of the decision tree. Please refer to the legend above to determine the classification of a specific color.

FIGURE 23 visualizes the structure of the initial decision tree. The x-axis describes how often each attribute is found as a first, second or third node in the individual decision trees of the 77 available users. It becomes apparent that the attribute travel distance is indicated as the most crucial attribute in the decision trees.

Similarly FIGURE 24 visualizes the structure of the intermediate tree, again the travel distance is regarded as the most crucial attribute in the splitting process. If we compare the structure of the intitial and intermediate decision trees the most obvious changes are apparent within the second and third node. Due to their limited tree size the intital decision trees do not contain any third nodes, within the intermediate trees almost all trees contain a third node.

Lastly the structure of the final decision tree is visualized in Figure 25. Again the attribute travel distance significantly stands out as the most crucial attribute within the first node.

Based on the structure of the three phases in the tree building process we can conclude that especially the structure of the second and third are likely to modify. It seems that within the intermediate tree the relative differences in the second and third node are more pronounced than within the final tree. The attributes in the latter are more evenly distributed. However the recurrence of travel distance as the main first node can explain the relatively good performance of the shortest path time based traditional routing algorithm.

INFLUENCE OF THE INDIVIDUAL ATTRIBUTES

The attributes implemented within the DTL algorithm are crucial to the functioning of DTL algorithm since a tree can only be ‘learned’ by splitting the trainings dataset into subsets based on an attribute value test. Although the C4.5 algorithm is robust in the presence of noise, the performance of the algorithm can significantly diminish when there are structural issues in one or multiple input-attributes.

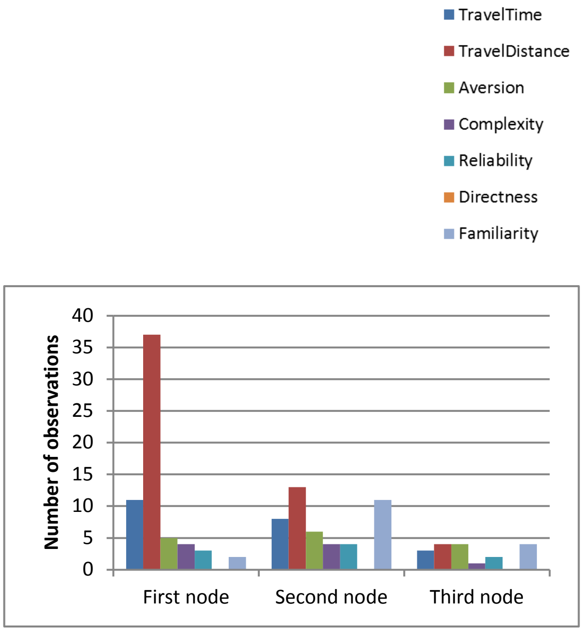


Figure 23: Structure of the initial decision tree

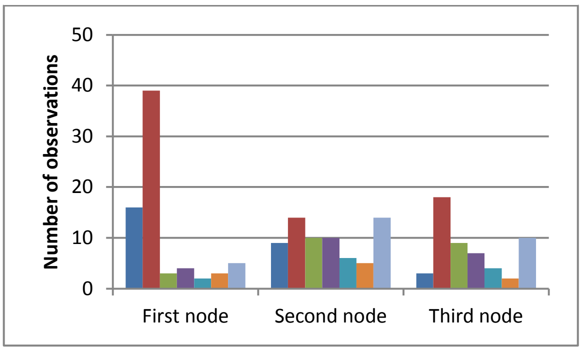


Figure 24: Structure of the intermediate decision tree

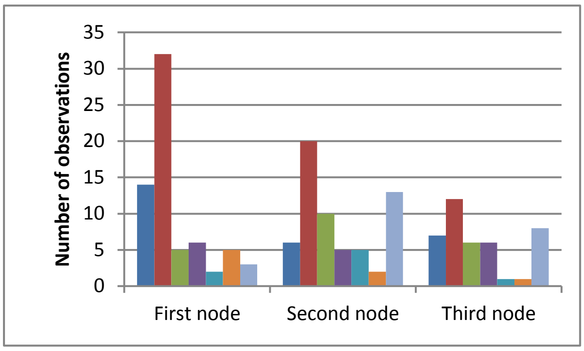


Figure 25: Structure of the final decision tree

Since the results described earlier indicate that the predictive performance of the DTL algorithm is significantly worse than the performance of the traditional shortest path algorithm, it is worthwhile to apply a sensitivity analysis to test the performance of the DTL when applying different sets of attributes. Within this sensitivity analysis the model is tested in multiple iterations in which one input attribute is excluded from the dataset. By retesting the predictive performance and by comparing the results with the primary results described before we can gain insight in the effect of the absence of one specific attribute.

In the sensitivity analysis six tests have been conducted, in every test one of the following attributes has excluded from the dataset; travel distance, aversion, complexity, reliability, directness and familiarity. The attribute travel time has been excluded from the sensitivity analysis because this variable is essential to compare the relative values of each attribute in relation to the 'best' route.

After deleting the information for one specific attribute the remaining attributes are classified by the DTL algorithm, again the first 10 trips are used to build an initial tree and successively all other trips are applied as 'tests'. The results of the iterations of the sensitivity analysis will be mutually compared, the results are not compared with the traditional algorithms.

The number of successful and failed tests have been visualized in TABLE 9.

Test number	Absent variable	Success DTL	Failure DTL	Total number of tests
1	Travel Distance	126 (3,7%)	3281 (96,3%)	3407
2	Aversion	118 (3,5%)	3289 (96,5%)	3407
3	Complexity	117 (3,4%)	3290 (96,6%)	3407
4	Directness	124 (3,6%)	3283 (96,4%)	3407
5	Familiarity	101 (3,0%)	3306 (97,0%)	3407
6	Reliability	124 (3,6%)	3283 (96,4%)	3407

Table 9: Results predictive performance attribute sensitivity analysis

If we examine the results above it becomes apparent that the relative results differ significantly, but the variation of the results in relation to the total number of tests are minimal. At first sight it seems that the number of successful DTL classifications decreases when the attribute familiarity is removed, however it is difficult to determine whether the difference is caused by a direct relation of the attribute or noise in the data.

In addition to the predictive performance of each separate iteration also the tree size of the initial, intermediate and final tree have been recorded. TABLE 10 describes both the average tree size and the standard deviation of the average tree size.

Test number	Absent variable	Initial Tree	Intermediate Tree	Final Tree
1	Travel Distance	4,3 (2,5)	9,4 (8,1)	12,2 (9,3)
2	Aversion	4,3 (2,6)	8,7 (7,6)	11,1 (8,7)
3	Complexity	4,1 (2,5)	8,3 (6,6)	10,5 (8,7)
4	Directness	4.3 (2,6)	9,4 (8,0)	12,5 (10,0)
5	Familiarity	4,2 (2,2)	8,5 (6,8)	11,3 (8,2)
6	Reliability	4,2 (2,5)	8,7 (6,5)	11,5 (8,4)

Table 10: Results average tree size attribute sensitivity analysis

Lastly the trainings error during the various model updates has been examined to gain insight in the learning behavior of the DTL model. FIGURE 26 visualizes the training error during the period in which the algorithm was tested.

If we analyze the combined results of the sensitivity analysis it becomes apparent that the decision tree learning algorithm is not susceptible to the omission of one specific attribute. It is noteworthy that the various tests within FIGURE 26 share similar local maxima and minima. Especially in the beginning of the tree growing procedures the differences between the several iterations are minimal. While proceeding with the tree growing procedures the differences between the iterations increase, as for example can be seen in FIGURE 26. However also the limited sample size at these updates could have been influential; the trainings error is averaged over all users and when the number of users decreases the variation is likely to increase due to the lower sample size. All of these considerations imply that not the available attributes but other processes are more significant in the performance of the DTL algorithm.

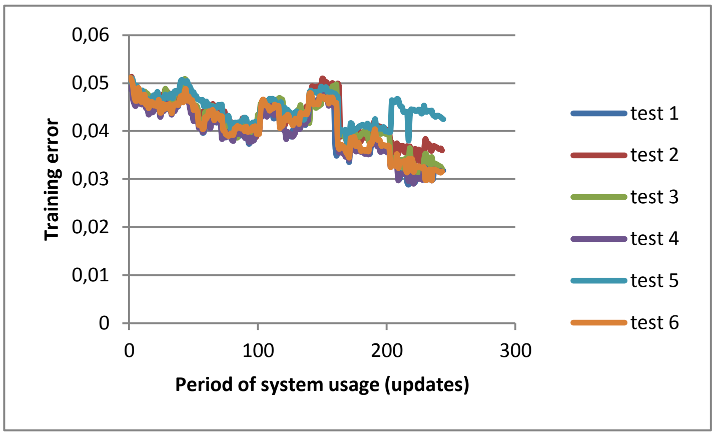


Figure 26: training error attribute sensitivity analysis

INFLUENCE OF THE PRUNING THRESHOLD

One of the other influential mechanisms, as also described in the theoretical elaboration of the DTL algorithm in chapter 3.3 that can strongly influence the performance of the DTL algorithm is the pruning threshold that is applied. Originally the C4.5 DTL algorithm firstly builds a decision tree and then prune subtrees from the tree in a subsequent pruning phase to improve the accuracy and prevent 'overfitting'.

The C4.5 algorithm uses a pruning mechanism that replaces subtrees with the best possible leaf which is the majority of the leaf classifications. If the error rate of the new tree would be equal or smaller than the original tree and when the subtree does not contain subtrees with the same property the subtree is replaced by the leaf. Main element of this methodology is the error rate calculation of the decision tree.

The main method for pruning within the C4.5, as described earlier is the Pessimistic Error Pruning method which uses a continuity correction for the binomial distribution to generate an error rate, this methodology requires a confidence threshold, a smaller confidence factor incurs less pruning of the tree. The confidence interval within the C4.5 algorithm should be larger than 0 and smaller than 1, by default a confidence factor of 0.25 is applied.

To gain insight in the processes behind the pruning mechanisms a sensitivity analysis has been conducted in which 10 different values between 0 and 0.5 have been applied and evaluated. Based on the predictive performance, the tree size and training error more knowledge will be gained to hopefully improve the performance of the 'Personalized Adaptive Routing Algorithm'.

TABLE 11 describes the predictive performance during the various tests. From the results it becomes apparent that the performance of the algorithm improves as the confidence factor is increased. The algorithm successfully predicted 96 of the routes at a confidence factor of 0.05 and this increased to 136 successful tests at a confidence factor of 0.5, based on these figures we can conclude that the performance improves while increasing the confidence factor. If we however compare the results with the total number of tests the relative improvement between the various confidence factors is limited the predictive performance at a confidence factor of 0.05 is equal to 2.8% and the predictive performance at a confidence factor of 0.5 the predictive performance is equal to 4.0%. However while interpreting these results it is important to realize that, due to the inconsistencies between the revealed route and proposed routes, only a maximum predictive performance of 40% could be achieved.

Test number	Confidence factor	Success DTL	Failure DTL	Total number of tests
1	Z = 0.05	96 (2,8%)	3311 (97,2%)	3407
2	Z = 0.1	104 (3,1%)	3303 (96,9)	3407
3	Z = 0.15	109 (3,2%)	3298 (96,8)	3407
4	Z = 0.2	116 (3,4%)	3291 (96,6%)	3407
5	Z = 0.25	121 (3,6%)	3286 (96,4%)	3407
6	Z = 0.3	127 (3,7%)	3280 (96,3%)	3407
7	Z = 0.35	127 (3,7%)	3280 (96,3%)	3407
8	Z = 0.4	132 (3,9%)	3275 (96,1%)	3407
9	Z = 0.45	134 (3,9%)	3273 (96,1%)	3407
10	Z = 0.5	136 (4,0%)	3271 (96,0%)	3407

Table 11: Results predictive performance pruning sensitivity analysis

The pruning mechanisms strongly influence the size of the created decision tree model because non-significant subtrees are replaced by either terminal leafs or lower subtrees. Therefore we should clearly see the effects of the confidence factors for pruning in the average tree sizes. Table 12 visualizes the average tree size at the various tree growing stages.

Test number	Sensitivity threshold	Initial Tree	Intermediate Tree	Final Tree
1	Z = 0.05	3,0 (1,9)	6,5 (5,4)	8,4 (7,2)
2	Z = 0.1	3,0 (1,9)	7,0 (5,6)	9,6 (7,3)
3	Z = 0.15	3,6 (2.0)	7,8 (6,6)	10,3 (7,7)
4	Z = 0.2	3,9 (2,3)	8,8 (7,4)	10,8 (7,8)
5	Z = 0.25	4,3 (2,6)	9,2 (8,0)	11,8 (8,9)
6	Z = 0.3	4,4 (2,5)	9,7 (7,9)	11,7 (8,4)
7	Z = 0.35	4,5 (2,7)	10,1 (8,3)	12,2 (8,6)
8	Z = 0.4	4,6 (2,7)	10,7 (9,6)	12,6 (9,0)
9	Z = 0.45	4,6 (2,7)	11,1 (9,5)	12,6 (8,8)
10	Z = 0.5	4,6 (2,7)	11,5 (9,4)	13,0 (9,3)

Table 12: average tree size pruning sensitivity analysis

The pruning mechanisms strongly influence the size of the created decision tree model because non-significant subtrees are replaced by either terminal leafs or lower subtrees. Therefore we should clearly see the effects of the confidence factors for pruning in the average tree sizes.

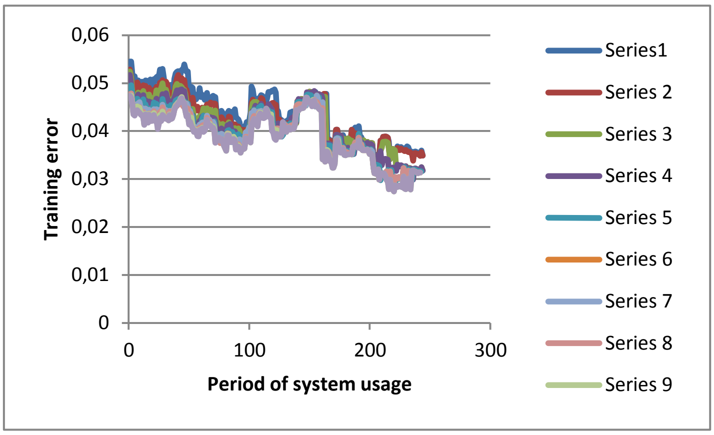


Figure 27: Training error pruning sensitivity analysis

Based on the results above it clearly becomes apparent that a smaller confidence factor incurs less pruning of the tree. However, similar to the results of the analysis of the predictive performance the results are less significantly as expected. Especially when we take the results of Park et al. (2007) into consideration, which recorded an average tree size of 66 nodes for the final tree, the results of the pruning sensitivity analysis are smaller than initially expected.

To conclude the pruning sensitivity analysis lastly the training errors of the various sensitivity iterations have been visualized in FIGURE 27. Based on this figure we can deduct that a lower sensitivity threshold, which implies more pruning, has a negative influence on the training error. This observation is logical, since pruning a model that is perfectly trained to match the trainings-data is simplified.

6.5 ANALYSIS OF THE RESULTS

The results of the implementation and evaluation of the 'Personalized Adaptive Routing Algorithm' illustrate an apparent contradiction, on one hand it seems that the chosen methodology and architecture effectively describes the past (historic) behavior of the users but on the other hand the model fails to predict the future behavior of the users. During the tests of the C4.5 based DTL algorithm and the traditional algorithms it became apparent that the traditional distance based shortest path algorithm was more successful in predicting the route choice of the test subjects than the adaptive algorithm.

Where does this contradiction originate and what does it imply? A possible first step is to revert to the original results that were described in the framework of Park et al. (2007). If the test errors are compared it becomes apparent that the error rates described in Park et al (2007) are very similar to the results above. The average error of the initial decision tree in this report of 4.99% is similar to the results of Park et al. (2007) in which the average error rate of the initial trees varied between 1.4% and 8.3%. Moreover both this thesis and the research of Park et al. (2007) identified a learning pattern of the algorithm; as the model was updated the average learning became lower.

One major impediment of the methodology that is applied in this study are the discrepancies between the revealed route of the user and the set of proposed route based on the origin/destination relation. In almost 60% of all the tests in which the proposed adaptive routing algorithm was compared with the traditional algorithm the revealed route was not part of the set of possible routes. This implied that the route prediction algorithm failed to produce a route prediction even before one of the routing algorithms were applied.

One of the main differences between this thesis and the results from Park et al. (2007) is the methodology to derive the proposed and revealed route. Within the research of Park et al. (2007) the software ICNavS was utilized to derive the set of proposed routes and moreover to simulate the revealed route. Due to this architecture the possibilities of having a set of possible routes that do not correspond to the revealed route is reduced because these routes are not made by human choices but by route selection rules which were predefined to be as similar to human reasoning as possible.

Another differentiating factor between this thesis and the research of Park et al. (2007) is the type of trips that have been applied to the decision tree learning algorithm. Within the latter research a network with a length of approximately 2.5 km and width of 1.5 km was applied in which all origin and destination pairs were modeled. These OD pairs were located at least 1 km apart from each other. This implies that the research of Park et al. strongly focused on short distance trips while this research, due to the technical limitations of the KATE data acquisition platform, focused on long distance strips. In real life routing behavior is especially dependent on the first section of the trip (from the origin towards the entry point of the high level road network) and the last section of the trip (from the exit point of the high level road network towards the destination). In the intermediate section, for example on the highway the number of realistic examples is very limited which implies that the mid-piece of a set of proposed routes is often very similar. Especially since the ratio of distance on underlying and high-level roads are often disproportionately, the differences in terms of routes scores were averaged out as noise.

Another factor that could explain the variations between proposed and revealed routes is the architecture of the study network, within the research of Park et al. (2007) the network was based on the underlying road network of the South Kensington area in West London. The high level roads (motorways) were excluded from the study network. Within this thesis the full network of the Netherlands was applied, based on visual inspections it became apparent that the route generation algorithm strongly focused on the shortest path in terms of travel distance. This sometimes resulted in route suggestions that ran straight through high density urban areas. In reality a flanking route based on the high level network would seem be more realistic. It seems that the route generation algorithm underestimated the travel time in urban areas, this can be explained by the omission of intersection delays in the urban environment. Efforts were made to compare the database travel times with the revealed travel times but no relation between road category and travel times were found.

One other difference is the number of routes that is used to feed the model, in the research of Park et al. (2007) an input set of 675 routes that represents one individual driver has been used. This dataset represented a time series data (journey records) of a driver that has travelled in the research area for 675 times and every trip had a unique OD-part. Based on the information from chapter five, which indicated that the average number of trips per day for a random user is 2,6. This implies that the research of Park simulated almost one year of data to implement and test the model. Moreover, in reality users mainly travel on regular routes. In practice it will be difficult to derive a set of 675 unique routes for a specific user within a reasonable amount of time to train the model.

The results of both the DTL based algorithm and the utility maximization function both score low in terms of predictive performance. These functions have one major component in common; the indicators that are used. Although the variable sensitivity analysis did not show that one individual indicator is impeding the results, it is possible that combinations of variables or perhaps all variables are impeding the results. Perhaps an recommendation for future research is a more detailed analysis of the individual attributes.

One difficulty that was encountered during the implementation of the DTL algorithm was the combination of multiple data sources and algorithms, during the implementation of the DTL algorithm all individual programming pieces must coincide and function as whole. For example, based on GPS noise in the revealed MapMatch output the revealed route occasionally includes ‘loops’. Especially when these loops include multiple one-way links (with unique link identification numbers) it is difficult to post process these routes and compare them with a set of proposed routes.

From a theoretical point of view, two complications often arise when applying machine learning to real world classification problems. One of these problems is the imbalance of classes, for example when one class occurs much more often than the other and the other is asymmetric misclassification costs which occurs when the costs of misclassifying an example from one class is much larger than the other class. Traditional learning algorithms treat positive and negative examples as equally important and therefore do not always produce a satisfactory classifier under these conditions. Within this thesis, when 14 negative and 1 positive route are classified by the algorithm this imbalance can negatively affect the performance. A possible method to counteract the imbalance is to raise the cost of misclassifying the minority class or the make the algorithm cost sensitive to the intentionally imbalanced training set.

From another perspective, although the predictive performance is limited, the interpretability can be of added value for researches that aim to investigate the factors that influence of attributes on travel behavior. Decision tree algorithms offer added value due to their interpretability, and this property facilitates the use of decision trees to extract useful knowledge about the optimal decision processes within the traffic and transport domain. The main argument on the interpretability of decision trees rests on the fact that the contribution of each node can be understood in isolation, summing the contributions generates a prediction and classification. As depicted in FIGURE 28, from considering in isolation each of the decision nodes it can be observed that a route with a reliability higher than the shortest path and a travel time lower than 1.17 times the shortest path are both predictors for a viable route. Both attributes generate a positive contribution to the prediction sum. Apparently the user considers to trade-off some travel time for a more reliable route. It is easy to interpret the meaning of all the decision nodes in a decision tree in a similar way.

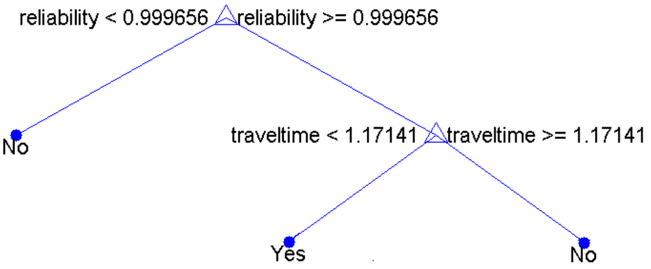


Figure 28: Interpretability of decision tree

7. Opportunities and threats

The previous parts of this report has elaborated the backgrounds, theoretical underpinning and also researched the translation of individual trip data towards a 'Personalized Adaptive Routing Algorithm'. Main aim throughout this research was to utilize the data derived from the KATE mobile data acquisition platform to implement a decision tree inductive learning algorithm to derive and validate the rules that describe individual preference mechanisms within route choice.

The previous chapters have mainly focused on the quantitative analysis, but in the quantitative analysis alone does not lie the greatest value of this research in terms of scientific knowledge. This study distinguishes itself from prior scientific literature in two perspectives; this study practically applies and evaluates the knowledge on inductive learning algorithms, in the past these methodologies were only tested in simulated environments. But perhaps more important is the way in which this practical implementation is facilitated; the trip data that is used as input for the DTL model is obtained through high quality (in terms of accuracy, density and reliability) floating car data based on private mobile phones.

This chapter will discuss the experiences and the lessons learned during this thesis. Main aim is to discuss a variety of topics that have not yet received attention earlier in this thesis but which do deserve some coverage due to their significant effect in the course of this thesis. This chapter consists of two paragraphs, the first will describe the main opportunities and the second describes the major threats.

7.1 Opportunities

This first section will elaborate the opportunities derived from the decision tree learning based personalized routing algorithm and the usage of location-based floating car data.

Personalized routing

The previously described system architecture and implementation of the personalized routing algorithm was mainly focused on the user's perspective. The main challenge was to integrate the personal factors of a specific driver within a route advise to generate a route advice that lies close to the expectation of this specific users. But besides the predictive accuracy of the routing algorithm there are many more opportunities when user profiling is applied to route planners. For example, although the familiarity is already integrated in the route proposal, the route familiarity could also be applied during the trip itself to compress the route guidance. For example, a personal navigation device should perhaps not give continuous route directions but instead combine familiar segments into contextualized steps. A route for example may contain directions from an origin to a local highway, if the user frequently travels to that highway the routing device can remove the steps from home to work and replace them with one message 'Go to the local motorway'. This significantly reduces the cognitive load that both printed directions or in-car navigation systems induce, causing users to lose focus in eyes-busy driving environments (Patel, Chen, Smith, & Landay, 2006).

The starting point within this master thesis was to primarily focus on the user optimum in terms of the route proposal, the effect of the route choice on the flow has been disregarded. As discussed in the second chapter, a relative small reduction of the peak traffic load at specific corridors can already cause a significant improvement in the overall traffic flow. Currently initiatives are being developed in which the routing of traffic will be centralized, advantage is that this enables the traffic manager to optimally utilize the available capacity of the infrastructure. However the success of these measures is dependent on the extend in which the end-user complies to the route proposal, especially when the route advice may not be fully optimal for an individual user the user may disregard the advice. Including personal attributes within the SmartRouting algorithm may potentially improve the user acceptance, every route proposal can be tailored to balance both the system and user optimum.

KATE AS RESEARCH PLATFORM

Between the start of this graduation project and the delivery of the present report, the development of the KATE mobile data acquisition platform has gained momentum. This thesis described one of the first attempts in which an automated method was employed to investigate the travel behavior of travelers. Especially the combination of the location tracing algorithm and the survey module offers opportunities to gather high quality information concerning the pre-trip and en-trip choices. In the Netherlands data on mobility is gathered by means of household surveys, every member of the household asked to register which trips are made during the day. For every trip the participant is requested to provide the motive, origin, destination, travel modality and travel distance of each trip. Discussions concerning the validity of the results from these household surveys remain, especially short distance trips and specific ethnic groups are underrepresented.

As demonstrated in this thesis the KATE platform is able to accurately register the origin, destination and travel distance. The trip motive and the trip modality is more difficult to deduct, however by means of the questionnaire module this additional data can be collected. During the course of the case study the performance of the rules that activated these questionnaire fluctuated, often unnecessary surveys were transmitted. However if improved, the technology is usable as an alternative or complement to the traditional household survey. Due to the reduced need for user input the data quality can be improved.

As the number of vehicles on our infrastructure increases it is important to develop high quality monitoring and control capabilities. Currently the Dutch road authorities mainly relies on video and infrastructure based sensors (e.g. measurement loops) however these systems are limited in terms of information accuracy. A promising and viable approach to improve the accuracy and timeliness of the monitoring is to utilize the vehicles and the sensor within the vehicle to generate up-to-date information. Although the electronic control unit (ECU) of the vehicle may provide a large amount of data concerning conditions in the vicinity (weather), travelling speed or context information (status of warning lights) the automotive industry is reluctant to release the information from the ECU. Meanwhile, due to the technological developments, more sensors are included in the devices in our day-to-day life. Smartphones contain a number of sensors (GPS and accelerometer) that can be employed to chart the dynamics in traffic. Although this thesis mainly focused on the general trip related information, also some effort was being put in the utilization of floating car data. By using the various en-route waypoints and the associated timestamps the revealed link travel times were estimated. However further developments, tests and data can underpin and improve the reliability of these measures. Moreover the data quality will improve by decreasing the interval in which the location tracing algorithm receives a location fix.

7.2 THREATS

The development and implementation of a ‘Personalized Adaptive Routing Algorithm’ also possesses threats and risks. Main subjects that require attention are the privacy and user collaboration which hamper the functioning of an “ideal” monitoring system.

USER ACCEPTANCE AND PRIVACY

User acceptance is often a crucial factor determining the success or failure of implementing an ITS system. Especially when utilizing a data acquisition platform such as KATE, when the boundaries of the individual privacy of users are moved, the discussion concerning user acceptance becomes even more delicate and complex.

Within the earlier part of this report the subject of privacy was intentionally not brought to the attention. It is assumed that the issue privacy is more related to the political and social debate than to the technical and feasibility in terms of the implementation and evaulation which is discussed in this report. The complete omission of the subject privacy will however hamper the completeness of this report, for this reason the subject is briefly discussed.

The discussion related to privacy is not directly related to the personalized routing algorithm but to the data that is needed to train and apply the decision tree learning model. As a first step it is important to define the term “privacy”. Within the context of this report the term privacy refers to the right that individuals are entitled to freedom of movement and expression, freedom from harassments and indignities, and freedom from invasions of their personal level (Wright, 1995). Perhaps the greatest threat to privacy posed by the data gathering by means of the KATE platform lies in its ability to conduct unwarranted and unwelcome surveillance on specific individuals; the application can monitor a person’s movement and behavior without explicitly notifying the user.

Beside the ability to track and monitor, the KATE platform collects and stores a large amount of sensitive information of the users. This information can be used to create personal, travel and behavioral profiles. It is important to realize that there is an area of tension between the objectives of the KATE platform and the alternative purposes for which the gathered information can be utilized.

During the course of this research much attention has been paid to the security and privacy of the participants, it was repeatedly communicated to the participants that the application monitors their location. Moreover the group of participants in the TravelAlert internal test program have agreed to the terms by means of internet survey. Moreover it was communicated that the information could be used for scientific studies and that the data would not be shared with third parties.

From a technology point of view, all efforts were made to anonymously store the information. One external server contains all the complete databases and that also operates and communicates with the client applications. To preserve the privacy of the participants the access to these servers was protected. A 32 digit user identification code was functioning as a primary key between all databases. For development purposes, for example to allow communication with the participants in case of problems, the names and email addresses were stored.

The processing servers at TNO operated based on a 72 digit unique userid’s, no personal information that could link the data to a specific user was stored in these server.

However, by only limiting the amount of information that can link a device to a specific person the privacy of participants is not yet guaranteed. Only based on the location traces it is straightforward to deduct person’s home location, work location etc. and it is possible to specifically pinpoint the location of a random user. It is possible to cut of specific parts of the locations, for example the last kilometer of a trip but that would also significantly reduce the data quality which is also undesirable.

USER PARTICIPATION

To train and implement a decision tree learning model in route choice it is crucial to gain enough information on each specific user. One of the lessons we can learn, based on the recent experiments, is that the cooperation of users is crucial, but also very fragile. The KATE platform was rapidly developing during the first months of this research, due to the recent rapid development and steep learning curve it was expected that some (minor) bugs and issues would arise. Although the final version of the application was 'frozen' and extensively tested still a number of users had to deal with technical difficulties. Most reported problems were related to the stability and battery usage of the application. Due to these difficulties a number of rapidly succeeding versions of the application had te be released. Especially due to the fact that the ReisAlarm application in the baseline (zero-measurement version), had no direct added value from a user's perspective the user participation decreased rapidly.

Most issues concerning the stability of the application were closely related to the design of the Android operating system. Due to its open-source architecture there is a wide range of devices and furthermore multiple versions of each device are distributed. In combination with the diversity in Android versions (v2.2, v2.3 v4.0 and v4.1) it has proven to be difficult to remedy problems prior to distributing the application. To avoid these problems there are actually two options, the first is to support only one single device, and perhaps the application should be distributed by means of a device that can be loaned to the participants during the experiment. Main advantage in this case is the possibility to configure these devices specifically for the experiment by disabling automatic updates etc. On the other hand a wide range of popular smartphones can be supported, but this requires extensive testing of each of these devices on all available Android versions.

Another point of interest is the perceived demand of the application on the smartphone, especially the battery usage of the application should be carefully monitored. There is a continues trade-off between data quality and battery usage. Although the battery usage of the application was limited a number of participants commented concerning the battery usage. Especially recent smartphones are extremely powerful in terms of specifications, and contrary to the increase in computing power the batteries did not (much) improve. Most of these modern smartphone have just enough battery power to reach the end of the day. When an additional application is installed the battery life decreases below an acceptable threshold for the user and the application is subsequently perceived as the main reason.

Lastly, to continuously keep users involved in the experiment it is important to include a clear added value. Either this added value can be included in the application, or the added value can be external by rewarding the participation. Especially when trip end surveys are included, which implies that the users should actively participate by filling in these surveys, a motivation is required to maintain the participation of the users.

8. Conclusions and recommendations

This chapter will, based on the research questions described in chapter two, discuss the main results of this thesis and will moreover present the conclusions. By critically reflecting on the research design, methodology and results the added value of this thesis to the current scientific domain will be made transparent. This chapter concludes with a number of recommendation and directions for future work.

8.1 Conclusions in regard to the sub-research questions

Earlier within this final report the following main research question was described:

How can the C4.5 decision tree learning algorithm be applied and utilized to identify and integrate individual preference mechanisms within route choice algorithms and how does this algorithm perform in relation to traditional routing algorithms?

This main research question is further differentiated in the following components. Each sub question will be described and answered in the text below.

1. How can route choice behavior be described and what knowledge is currently available that describes the influence of personal factors and preferences within route choice?

Route choice behavior can be regarded as a two stage process. Within the first state the users generates a 'choice set' of feasible alternatives, within the second stage a choice criteria is adopted that eliminates all the inferior alternatives until the best alternative is identified.

Two main currents within the route choice modeling can be defined. Within the first group, the theory driven models, a discrete choice model is often used due to the fact that there is a clear set of discrete alternatives within route choice. A traveler is assumed to use a random utility based function to 'objectively' assess the alternatives. Two important pillars in discrete choice models are multinomial and nested logit models, however these have received comments concerning their validity due to their simplistic assumptions.

The second main current consists of the data driven models which can be separated in stated preferences models and revealed preference models. Both have advantages and disadvantages; stated preference studies are often criticized because they fail to take certain behavioral constraints into account. On the other hand revealed preference fail to correctly predict or model an attribute that is not in included in the dataset or out of range.

In traffic and transport studies have historically relied on theoretical or data-driven stated preference models, however due the technological developments revealed preference models have gained popularity. Data mining represents the computer assisted processing of large sets of data and aims to extract the meaning of the data. Its advantage is that data mining can predict behaviors and trends, but more importantly it is able to find hidden patterns or information which was previously missed or outside the expectation.

The main added value of data mining, and therefore this thesis, lies in its flexible combination of using revealed data attributes in a model that use a very broad, limited and flexible theoretical underpinning. In the decision tree learning algorithm, as employed within this study, the revealed behavioral data is employed in a structure which only assumed that travelers make conscious and well-informed decisions. Especially in route choice, and its inherent complex decision making process, this methodology may provide future insights in the route choice behavior of drivers in traffic.

2. *Which past initiatives have been conducted that aim to implement data mining algorithms in the field of traffic and transportation?*

Although data mining, and more specifically decision tree learning algorithms, are widely used in the medicine and the information and communication industry, the application of data mining methodologies in the field of traffic and transport and specifically route choice are limited. Within the last decade four studies have been conducted that apply decision tree learning algorithm as an alternative to random utility models to model travel choice behavior. Three of these studies were based on stated preference questionnaires ((Wets, Vanhoof, Arentze, & Timmermans, 2000) (Yamamoto, Kitamura, & Fujii, An analysis of drivers' route choice behavior by data mining algorithms, 2002) and (Arentze & Timmermans, 2002)

The research of Park, Kaparias and Bogenberger (2007) was the first research to suggest the utilization of revealed behavioral data as input for the decision tree learning algorithm. This study introduced an implementation structure and tested the methodology based on data derived from the simulation program ICNavs. Currently no real life implementation of the decision tree learning algorithm has been implemented and evaluated. This study therefore went beyond the scope and results from prior work by including real life (non-simulated) user data within the implementation of the DTL algorithm.

3. *What input information, data sources and techniques are necessary to translate GPS derived floating car data towards a real life test implementation and evaluation of the C4.5 DTL based personalized adaptive routing algorithm?*

Main aim of the decision tree learning algorithm is to provide a representation of a decision procedure for determining the class of a given instance. The algorithm partitions a dataset (described as attributes and an instance class) into disjoint regions by means of a divide-and-conquer strategy.

Based on the definition above we can deduct one major requirement while implementing a C4.5 learning model that should be able to represent the drivers individual attributes in route choice behavior; large volumes of detailed data concerning the trips, and corresponding routes, are required to provide input for the model.

This report described and integrated a methodology to derive and analyze travel behavior based on GPS enabled smartphones. The KATE data acquisition platform, developed by TNO, contains a location tracing algorithm which employs various sensors and update intervals to efficiently map the travel behavior of a group participants. Between the 25th of September 2012 and the 31st of March 2013 95 users were equipped with a smartphone application that is based on the KATE platform. The total number of trips that were detected was 11.490 of which 1172 trips were made with train. The total distance travelled by was 277.021 kilometers.

Main output from the KATE platform are the location traces (latitude, longitude, accuracy and timestamp). To leverage the data from the KATE platform it is necessary to identify individual trips within the traces and infer the sequence of roads (expressed as links and their infrastructural characteristics) that a driver traversed.

The 'Personalized Adaptive Routing Algorithm' has been applied by implementing a decision tree algorithm that aims to represent the route choice behavior with respect to the individual preference and characteristics. For each individual OD-relation from the KATE trip database a set of possible paths has been generated based on a k-shortest path route generation algorithm in combination with a set of Monte Carlo iterations. All paths were subsequently evaluated in terms of travel time, travel distance, directness, complexity, travel time reliability, familiarity and aversion. However because the observed attribute values cannot be compared between trips, relative values of the attributes over the best path between each origin/destination pair are used. This relative attribute values are the main input for the learning process within the 'Adaptive Personalized Routing Algorithm'.

For each individual user, based on data from the KATE platform, an initial training dataset was generated that consisted of 10 trips, this training dataset was utilized to generate a preliminary DTL model. Subsequently all remaining trips were individually used as a test dataset, during the learning process the trainings set was automatically extended and if necessary the model was updated during the test iterations.

Based on a C4.5 DTL learning algorithm, the routes from the past were employed to improve the route predictions in the future. By comparing the predicted route derived from the 'Personalized Adaptive Routing Algorithm' and the route actually taken by the drivers more information concerning the performance of learning algorithms in route choice is gathered.

4. *How can the predictive performance of the DTL based adaptive routing algorithm be measured and how does the adaptive routing algorithm perform in comparison to two traditional routing algorithms (shortest path and multi-attribute utility maximization)?*

The main indicator to assess the performance of the DTL algorithm is the predictive performance which is represented by the amount of correctly predicted routes. To place the results of the DTL based algorithm in perspective, adjacent to the DTL route prediction also two traditional route prediction models were applied which were based on the traditional distance based shortest path (SP) and a traditional multi-attribute utility function (UM). Subsequently the predicted routes from these algorithms were compared with the revealed route.

Table 13 represents the main results for the adaptive and traditional routing algorithms. Based on the results it becomes apparent that the predictive performance of all algorithms is limited, the traditional shortest path algorithm achieves the best relative performance. Main reason for the limited performance are inconsistencies between the representation of the revealed route within the set of possible routes, if the revealed route is not included in the set of possible routes the algorithm will automatically fail to predict the correct route.

From the 3407 trips within the trip database the revealed route was available within the set of possible routes in 1360 of the trips. This result significantly influences the outcomes of this study, in the most favorable case every single algorithms can achieve a predictive performance of 40%.

	Successful tests	**Failed tests**	**Total number of tests**
C4.5 learning algorithm	121 (4%)	3204 (96%)	3407
Traditional SP algorithm	647 (19%)	2760 (81%)	3407
Traditional UM algorithm	0 (0%)	3407 (100%)	3407

Table 13: Main predictive performance of the various algorithms

The results in terms of the quantitative predictive performance do not directly substantiate an actual implementation of the personalized C4.5 based routing algorithm. However to gain further insight in the processes within the DTL algorithm several tests have been applied to gain insight in the learning behavior of the 'Personalized Adaptive Routing Algorithm'. To examine the learning behavior the training error was calculated during the time in which the system was applied. The results indicated that the algorithm adjusts itself to the user in which it was applied, however the learning effect was less significant than the previous DTL implementation (Park et al. (2007)) suggested. Moreover the tree size during the various experiments was investigated; these results indicated that the average size of the tree increases to accommodate the individual user characteristics. However similar to the results towards the training error, the current results are less significant than the results of previous research of Park et al. (2007) described.

The attributes included in the DTL algorithm and the amount of pruning that is applied to the decision tree can significantly affect the results. To map these effects two separate sensitivity analyses have been applied in this thesis. The first analysis iteratively discarded one of the attributes and the second analysis tested the algorithm when applying various pruning confidence intervals. The results proved that these attributes influence the predictive performance, however in relation to the number of failed tests and the performance of the traditional shortest path algorithm the results did not reveal insights that could prove the added value of the 'Personalized Adaptive Routing Algorithm' in terms of predictive performance. However the results did prove that none of the individual attributes or the applied pruning mechanism significantly impeded the performance of the algorithm. To achieve a significant improvement an analysis of other data-mining algorithms is suggested rather than fine-tuning the current C4.5 implementation.

5. What are the future opportunities and challenges when applying large scale (big) data sources to investigate and explain personal route choice behavior?

This thesis aimed to take two steps within the domain of transportation, for the first time a large scale mobile data acquisition platform was applied in the Netherlands to implicitly investigate the travel behavior of users. Moreover this data and the subsequent results have also been applied in practice to implement an adaptive personalized routing algorithm.

Although the results of the comparison the algorithms did not demonstrate an added value to implement the personalized routing algorithm in practice, this research did reveal that the usage of floating car data offers perspectives and opportunities to gain further insight in the performance of the infrastructure. Moreover this thesis demonstrated the aspects of the individual user that utilizes the infrastructure. Especially because a smartphone application can function as both a monitoring- and communication tool the application offers possibilities to gather implicit information such as the trip information (both spatial and temporal) and explicit user information such as his personal, social, economic and demographic attributes by means of the questionnaire module.

Especially when the data quality is further improved by tuning and adapting the update interval, the individual vehicles can become cost-effective sensors to gain insight in the traffic flow.

The evidential outcomes of this thesis can serve as feedback information for scientists and road authorities to shade light on the decision making processes and the reliance on navigation systems. However, it has to be pointed out that the data used in this thesis is not perfect. More observed variables should be investigated and future work should provide more results to strengthen the implications of this thesis.

8.2 Conclusions with regard to the main research question

The previous chapter described the answers to the main research questions, based on the overall results of this thesis the following conclusion can be drawn concerning the main research question:

How can the C4.5 decision tree learning algorithm be applied and utilized to identify and integrate individual preference mechanisms within route choice algorithms and how does this algorithm perform in relation to traditional routing algorithms?

In this report, we have described the mean features and implementation of an approach that aims to include personal attributes within route choice by implementing the C4.5 decision tree learning algorithm within a so called 'Personalized Adaptive Routing Algorithm'. The main framework for this thesis was supplied by the paper of Park et al. (2007) that theoretically described the methodology and moreover employed simulation experiments to investigate the applicability of the learning model to adaptive rout planning. This current thesis has gone beyond the scope and results from prior work.

Based on the paper of Park et al. (2007) and the revealed data we have been able to generate a methodology that supports a real-life implementation of the C4.5 decision tree algorithm by developing a process and programming structure which translate the raw data from a GPS based data acquisition platform (in this study the KATE platform which is developed by TNO) towards the necessary input for the learning algorithm.

The predictive performance of the adaptive and traditional routing algorithms were compared, the results point out that, in its current implementation, the 'Personalized Adaptive Routing Algorithm' is not able to achieve a higher predictive performance than the traditional routing algorithms. Based on 3407 trips the traditional shortest path algorithm achieves a 19% predictive performance while the DTL based algorithm achieves a predictive performance of 4%. The multi-attribute utility function scored a predictive performance of 0%.

The results of the implementation and evaluation of the decision tree learning algorithm illustrated an apparent contradiction, on one hand it seems that the chosen methodology and architecture effectively describe the past (historic) behavior of the users but on the other hand the model fails to predict the future behavior of the users.

Although the results in terms of predictive performance do not support a future actual implementation of a decision tree learning based routing algorithm. The results do support the statement in the beginning of this thesis concerning the complexity of human decision making and its current limited application in the traditional routing algorithms. Despite the predictive performance of traditional shortest path algorithm is better than the predictive performance of the adaptive personalized routing algorithm, there is clearly room for future improvements.

One of the main advantages in the methodology applied in this research is the ability of C4.5 algorithm to resemble the various components within decision making process. Especially its ability to represent the conditional interaction between route attributes within the hierarchical structure of the decision tree are elements that behavioral scientist can be exploited in the research to identify the determinants in travel behavior.

The results of this study do not correspond to the results of the previous attempts, and more specifically the research of Park et al. (2007), to implement decision tree learning algorithms in route choice. One of the main factors that impeded the results of this study was the coherence between the revealed route and the set of possible routes based on the origin and destination of the trip. Within almost 60% of the tests the routing algorithm failed because none of the proposed routes was similar to the revealed route. Several attempts have been made to improve the route generation algorithm but none significantly improved the performance.

Several tests have been performed to further investigate the learning processes within the decision tree learning algorithm, experiments were carried out to test the influence of individual attributes within the algorithm and moreover various pruning thresholds have been applied. During these tests no significant improvements in terms of predictive performance have been achieved. Based on these results we can conclude that further researches in alternative methodologies are perhaps more successful than optimization efforts of the current C4.5 DTL based methodology.

One of the correlations that can further be researched in the future is the limited performance of both the DTL based and the Utility Maximization (UM) based algorithm, it is striking that the UM algorithm was not able to generate any correct prediction. Main relationship between these both the DTL based and UM algorithm are the route attributes. Although the sensitivity analysis did not suggest that one single attribute significantly negatively influenced the results; a certain combination or group of attributes can be influential in the performance.

However the discussion to improve the algorithm will be further discussed in the next section concerning the recommendations for future researches.

8.3 Recommendations for future research

Despite the limitations in terms of the results of this thesis, its true added value lies in the experiences and knowledge that is obtained while carrying out this research. This section will further discuss the recommendations for future research.

During the discussion of the results it became apparent that a number of differences exist between the current implementation and the historic framework of Park et al. (2007). While the study of Park et al. (2007) focused primarily on a relatively small research area (short distance trips) this thesis focused on large distance trips. If we assume that changes in route choice behavior mainly take place in the first and last part of the route, which is often location on the underlying road network, it is worthwhile to further focus the adaptive personalized on these specific network segments. By splitting the route prediction in, for example, three separate elements which each resemble an important part of the route (for example origin towards motorway, the high level network and the part from the motorway towards the destination) the performance of the ‘Personalized Adaptive Routing Algorithm’ can possibly be improved. By breaking up the route prediction two possible issues are resolved; firstly when the elements on the underlying network are individually predicted the scores of these sections are not averaged out by the distance on the motorways. Especially motorway sections do not allow a large amount of variation in terms of possible routes, in combination with the disproportionate distance at these links can negatively influence the performance of the algorithm. Secondly, by breaking up the route predictions and excluding the motorway elements, the road-categories within each section of the route become more evenly distributed. However, major difficulty when implementing these changes is that a route may consists of a combination of paths, furthermore often multiple entry and exit points towards the motorway can be determined. Both these characteristics significantly increase the required computing power to generate and process the set of possible routes.

Within this thesis, the data from the KATE platform determined the main starting point for implementing the 'Personalize Adaptive Routing Algorithm'. During the development of the platform a balance was apparent between data quality and battery usage. Due to a design choice to limit the battery usage the application did not perform optimally in registering short distance trips. To improve the data derived from the KATE platform it is advised to further research the possibilities to improve the data quality in urban trips by improving the speed in which the application detects the movement of a device. Currently there is a delay between the moment a user departs and when the application decreases the update interval to one minute to generate high quality data. This delay negatively influences the data quality in two perspectives, on the one hand the amount of location measurements within the first 10 to 15 minutes is often limited. This leads to more uncertainty in terms of the route that the user actually took because there are often multiple routes between these two points. Especially because this first part, often on the underlying road network, is a determining factor in the route choice the algorithm can significantly benefit from an improved data quality. On the other hand the application is often not able to register short distance trips because the user already arrived at his (or her) destination before the application started to regularly monitor the location of the user, therefore short distance trips with a travel time of less than 10 to 15 minutes are often not or poorly registered. To gain a better and more complete picture of the travel behavior of the participants a more responsive application in terms of the movement and trip detection is advised.

Within this research it was infeasible to deduct trips that were made by either cyclist or urban public transport (bus, tram and metro) from the GPS data. Although an improved data quality, as discussed above, will simplify the detection of these trips also more and better algorithms are required to detect and process these trips. To deduct the trips made by bicycle and urban public transport more separate MapMatching algorithms are required and both require a specific representation of the infrastructural network. Furthermore, to detect public transport, also the timetables and the stops need to be represented within the MapMatching algorithm.

A link between the route generation algorithm and the decision tree learning can improve the performance. Currently the decision tree learning algorithm is dependent on the route set that is provided. During this thesis it became apparent that the route set does often not include the revealed route, this can be prevented by also including the personal characteristics in the route generation step. It may also be worthwhile to include locations or waypoints that are often chosen by the user within the algorithm that generates the possible routes. This recommendation can very well coincide with the recommendation to break up the route prediction algorithm in separate sections, the access and exit points to and from the high level network can be chosen based on the historic data from the specific user.

One of the main elements that can improve the adaptive algorithm is the explicit support of trip chaining. The algorithm currently generates a route based on the origin and destination, in reality the assumption that each trip only has an origin and destination may not represent reality because some users may desire a stopover during the trip to drop off their children or to pick someone up. The addition of via points within the route algorithm may facilitate these specific circumstances and improve the satisfaction towards the system.

As discussed in chapter 6.5 the sensitivity analysis that described the individual performance of each of the attributes did not demonstrate that one single attribute significantly reduced the performance of the personalized adaptive routing algorithm. However, because both algorithms that applied the individual attributes (DTL based and multi-attribute utility functions) were performing worse than the single attribute shortest path algorithm it is advised to further examine the path evaluation framework. Firstly the addition of a real-life travel time estimate, instead of the free-flow travel time, may improve the performance of the algorithm. Especially because all possible paths between an original destination are relative to the path with the shortest travel time this indicator significantly affects the performance of the algorithm. Especially on the lower level road network, which is crucial to the DTL based algorithm, dynamic factors such as intersections delays are an important factor which have currently not been integrated in the indicator framework. Although realistic travel times, based on floating car data, are currently implicitly applied in the travel time reliability indicator, it should be noted that the travel time reliability (the variation between measurements) is fundamentally different than the average real-time travel time (the average over separate measurements). To include the real-life travel times the reliability and coverage of the revealed link travel times should be improved. This can be accomplished by decreasing the update interval of the location tracing algorithm but especially more data is required by further distributing the application to more participants. Also other data sources can be employed, for example TomTom N.V. is known to generate a reliable database of revealed link travel times based on data that is retrieved from their navigation systems.

In this research the route attributes that were used were quantify the proposed path were based on the framework of Park et al. (2007). Main motivation was that, by employing a similar quantification, the results of both the previous and present research could be compared. However some of the quantifications were quite arbitrary, for example the aversion was quantified by multiplying the distance on certain road types by a value ranging from 0 to 2. However the study of Park et al. (2007) covered a more limited and similar research area (e.g. the motorways were not included), due to the more extensive research area in this research a more detailed quantification of the indicator is advised. For example a distinction can be made of road categories inside and outside the built-up area which better resembles the road categorization (Duurzaam Veilig) that is applied in the Netherlands.

Besides improving the existing indicator framework also alternative and additional attributes may be examined. This research mainly included quantitative attributes, however qualitative attributes such as aesthetics and driving comfort also influence route choice behavior. Moreover information about the surrounding of the infrastructural network (e.g. shopping centers or petrol stations) are able to provide added value to the predictive accuracy of the algorithm. A link with Geographic Information System (GIS) database that describe the land use functions of the surrounding landscape can be integrated within the indicator framework to quantify these qualitative indicators.

Another aspect of the implementation of the DTL algorithm that can be examined to further analyze the results is the composition of the user group. One of the differentiating factors between this study and the paper of Park et al. (2007) is the amount of data that was available for a specific user, where Park et al. (2007) applied the DTL algorithm to one specific user with 675 trips, the present research included the data from 73 separate users which had an average of 67 trips per user. To better recreate the circumstances from the paper of Park et al. (2007) ideally more travel data for each individual user is required. However, if we take into account that the average daily number of trips per user is 2,6 the time required to generate the dataset is long. Another possibility to recreate a more extensive dataset is to combine the information of multiple users to one combined dataset; however this requires a more equal user group instead of the more diverse group (in terms of social characteristics, daily mileage and geographic dispersion) that was monitored within the KATE experiments.

Furthermore the performance of the algorithms can be improved by developing methodologies to train the model on drivers with little or no driving data. For example clustering can be applied to generate an initial model, this would allow users with sparse data to be linked with other users. This would allow the system to utilize data from other similar users to refine the predictions for new and relatively unknown users.

Lastly it is recommended to investigate the extend in which the adaptive routing can be applied in practice to support further testing of the ‘Personalized Adaptive Routing Algorithm’ when the test results are re-evaluated, a connection should be made with the personal navigation device to visualize and integrate the personal route advice.

BIBLIOGRAPHY

Akasaka, Y., & Onisawa, T. (2008). Personalized pedestrian navigation system with subjective preference based route selection. In D. Ruan, F. Hardeman, & K. Meer, *Intelligent Decision and Policy Making Support Systems, Studies in Computational Intelligence, vol. 117.* (pp. 73–91). Berlin-Heidelberg: Springer.

Arentze, T., & Timmermans, H. (2002). Parametric action decision trees: incorporating continuous attribute variables into rule-based models of activity-travel behavior. *Proc. 84th TRB annual Meeting*.

Bekhor, S., Ben-Akiva, M., & Ramming, S. (2006). Evaluation of choice set generation algorithms for route choice models. *Ann. Oper. Res*, 144, 235-247.

Ben-Akiva, M., Bergman, M., Daly, A., & R., R. (1984). Modelling Inter Urban Route Choice Behaviour. In J. Volmuller, & R. Hamerslag, *Proceedings of the 9th International Symposium on* (pp. 299-330). Utrecht: VNU Press.

Centraal Bureau voor de Statistiek. (2011). *CBS Statline - Mobiliteit in Nederland; mobiliteitskenmerken en motieven*. Opgeroepen op 05 12, 2013, van Centraal Bureau voor de statistiek: http://statline.cbs.nl

Centraal Bureau voor de Statistiek. (2013). *Onderzoek verplaatsingsgedrag en Mobiliteitsonderzoek Nederland - Korte onderzoeksbeschrijving*. Opgeroepen op 05 13, 2013, van Centraal Bureau voor de Statistiek: http://www.cbs.nl/nl-NL/menu/themas/verkeer-vervoer/methoden/dataverzameling/korte-onderzoeksbeschrijvingen/onderzoek-verplaatsingsgedrag.htm

Chan, W., & Mizoguchi, R. (1999). Communication Content Ontology for Learner Model Agent in Multi-agent Architecture. *In*, 95-102.

Choi, W., Kim, S., Kang, T., & Jeon, H. (2008). Study on method of route choice problem based on user preference. In J. Carbonell, & J. Siekmann, *Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, vol. 4694.* (pp. 645-652). Berlin-Heidelberg: Springer.

Duckham, M., & Kulik, L. (2003). Simplest paths: automated route selection for navigation. In W. Kuhn, M. Worboys, & S. Timpf, *Spatial Information Theory* (pp. 16-185). Berlin-Heidelberg: Springer.

Duffel, J., & Kalombaris, A. (1998). Empirical Studies of Car Driver Route Choice in Hertfordshire. *Traffic Engineering and Control*, 398-408.

Eksioglu, B., Vural, A., & Reisman, A. (2009). The vehicle routing problem: A taxonomic review. *Computers & Industrial Engineering 57*, 1472-1483.

Frejinger, E. (2008). *Route Choice Analysis: Data, Models, Algorithms and Applications.* Lausanne: ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE.

Frejinger, E., & Bierlaire, M. (2007). Capturing correlation with subnetworks in route choice models. *Transp. Res. Part B 41*, 363-378.

Garling, T., Kwan, M.-P., & Golledge, R. (1994). Computational-process modeling of household travel activity scheduling. *Transportation Research 25B*, 355-364.

Golledge, R. (1995). Path selection and route preference in human navigation: a progress report. In A. K. Frank, *Spatial Information Theory: A Theoretical Basis for GIS, Lecture Notes in Computer Science, vol. 988.* (pp. pp. 207–222.). Berlin-Heidelberg: Springer.

Hadjali, A., Mokhtari, A., & Pivert, O. (2012). Expressing and processing complex preferences in route planning queries: Twoards a fuzzy-set-based approach. *Fuzzy Sets and Systems*, 82-104.

Hainan, L., Randall, G., Ogle, J., & Wang, J. (2004). Using Global Positioning System Data to Understand Day-to-Day Dynamics of Morning Commute Behavior. *Transportation Research Record: Journal of the Transportation Research Board, No. 1895*, 78–84.

Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, H. (2009). *The WEKA Data Mining Software: An Update.* SIGKDD Explorations, Volume 11, Issue 1.

Han, B., Algers, S., & Engelson, L. (2001). Accommodating Drivers' Taste Variation and Repeated Choice Correlation in Route Choice Modeling by Using the Mixed Logit Model. *Transportation Reseach Board 80th Annual Meeting*.

Haque, S., Kulik, L., & Klippel, A. (2007). Algorithms for reliable navigation and wayfinding. In J. Carbonell, & J. Siekmann, *Spatial Cognition V: Reasoning, Action,* (pp. 308–326). Berlin-Heidelberg: Springer.

Heathington, K., Worall, R., & Hoff, G. (1971). Attitudes and Behavior of Drivers Regarding Route Diversion. *Highway Research Record*, 18-26.

Hoogendoorn-Lanser. (2005). *Modelling travel behaviour in multi-modal networks.* Netherlands TRAIL Research School.

Huchingson, R., McNees, R., & Dudek, C. (1977). Survey of Motorist Route-Selection Criteria. *Transportation Research Record 643*, 45-48.

Isupply Market Research. (2010, October 07). *Automotive Navigation Heads into the Cloud* . Opgeroepen op June 17, 2013, van Press Release: http://www.isuppli.com/Automotive-Infotainment-and-Telematics/News/Pages/Automotive-Navigation-Heads-into-the-Cloud.aspx

Jozefowiez, N., Semet, F., & Talbi. (2000). Multi-objective vehicle routing problem. *Eur. J. Oper. Res. 189,*, 293–309.

Kumar, P., & Singh, P. (2005). Advanced traveler information system for Hyderabad City. *IEEE Transactions on Intelligent Transportion System, vol 6*, 26-37.

Li, J., & Limsoon, W. (2003). *Using Rules to Analyse Bio-medical Data: AComparison between C4.5 and PCL.* Berlin Heidelberg: Springer.

Lyons, G. (2006). The role of information in decision-making with regard to travel. *IEE Proceedings: Intelligent Transport Systems Vol 153 No. 3*, 199-212.

Mahmassani, H., Hatcher, S., & Caplice, C. (1997). Daily Variation of Trip Chaining, Scheduling and Path Selection Behavior of Work Commuters. In P. Stopher, & M. Lee-Gosselin, *Understanding Travel Behaviour in an Era of Change* (pp. 351–380). Oxfor, United Kingdom: Pergamon-Elsevier Science.

Ministerie van Infrastructuur en Milieu. (2012). *Structuurvisie Infrastructuur en Ruimte.* Den Haag: Drukkerij Ando.

Mitchell, T. (1997). *Machine Learning.* Columbus: McGraw-Hill.

Mouskos, K., Greenfeld, J., & Pignataro, L. (1996). Towards a Multi-Modal Advanced Traveler Information System. *NJIT Research, Vol. 4*.

Nadi, S., & Delavar, S. (2011). Multi-criteria, personalized route planning using quantifier-guided ordered weighted averaging operators. *International Journal of Applied Earth Observation and Geoinformation 13*, 322-335.

Park, K., Kaparias, I., & Bogenberger, K. (2007). Learning user preferences of route choice behavior for adaptive route guidance. *Special Issue: Selected papers from the 13th World Congress on Intelligent Transport Systems and Services* (pp. 159-166). Londen: The Institution of Engineering and Technology.

Patel, K., Chen, M., Smith, I., & Landay, J. (2006). Personalizing Routes.

Pedersen, D. (1998). Factors in Route Selection. *Perceptual and Motor Skills 86*, 999-1006.

Quinlan, R. (1933). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann Publishers.

Ratcliffe, E. (1972). A Comparison of Drivers' Route Choice Criteria dn Those Used in Curent Assignment Processes. *Traffic Engineering and Control 13*, 526-529.

Richter, K. (2007). *Context-specific route directions: generation of cognitively motivated wayfinding instructions. Ph.D. Dissertation.* Bremen: Universität Bremen.

Richter, K. (2009). Adaptable path planning in regionalized environments. In K. Hornsby, C. D. Claramunt, & L. G., *Spatial Information Theory, Lecture Notes in Computer Science, vol. 5756.* (pp. 453-470). Berlin: Springer.

Robinson, D., & Reed, V. (1998). *The A-Z of social research jargon.* Ashgate/AREMA.

Rodrigue, J., Comtois, C., & Slack, B. (2009). *The Geography of Transport Systems.* New York: Routledge.

Rogers, S., & Langley. (1998). Personalized driving route recommendations. *Proceedings of the AAAI Workshop on Recommender Systems* (pp. 96–100.). Madison: WI, USA,.

Rogers, S., Fiechter, C., & Langley, P. (1999). A route advice agent that models driver. *Proceedings of the AAAI Spring Symposium on Agents with* (pp. 106–113). CA, USA,: Stanford.

Sadeghi Niaraki, A. (2008). *Ontology based geospatial model for personalize route. Ph.D. Dissertation.* Inha University.

Sadeghi Niaraki, A., & Kim, K. (2009). Ontology based personalized route planning system. *Expert Syst. Appl. 36 (2, Part1)*, 2250–2259.

Scriven, M. (1991). *Evaluation Thesaurus.* Newbury Park, CA: Sage Publications.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 379–423.

Shi, Z. Z. (2002). *Knowledge Discovery.* Beijing: University Press.

Soman, K., Diwakar, S., & Ajay, V. (2006). *Insight into Data Mining.* New Delhi: Prentice Hall of India Private Limited.

Stepanov, A., & Smith, J. (2009). Multi-objective evacuation routing in transportation. *Eur. J. Oper. Res. 198*, 435–446.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Boston: Addison-Wesley.

The Slovene Society Informatika. (2005). *Informatica, An International Journal of Computing and Informatics.* Ljubljana: The Slovene Society Informatika.

Thomas, M., & Joy, A. (1991). *Elements of Information Theory.* John Wiley & Sons, Inc.

Volkel, T., & Weber. (2008). RouteCheckr: personalized multicriteria routing for mobility impaired pedestrians. *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility, October 13–15 ,*, (pp. 185–192). Halifax, Nova Scotia, Canada,.

Volpe, J., Lappin, J., Bottom, B., & Gardner, B. (2002). *Understanding and Predicting Traveler Response to Information: A Literature Review Prepeared for Office of Metropolitan Planning and Programs.*

von Neumann, J. (1932). *Mathematical Foundations of Quantum Mechanics.* Princeton: Princeton University Press.

Wachs, M. (1967). Relationships Between Drivers' Attitudes Toward Alternate Routes and Driver and Route Characteristics. *Highway Research Record 197*, 70-87.

Wachs, M. (2002). Social trends and research needs in transport and environmental planning. *Social Change and Sustainable Transport*, 17-26.

Walker, J. (2001). *Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures and Latent Variables.* Massachusetts: Massachusetts Institute of Technology.

Wets, G., Vanhoof, K., Arentze, K., & Timmermans, H. (2000). Identifying decision structures underluing activity patterns: an exploration of data mining algorithms. *Proc. 79th TRB Annual Meeting*.

Witten, I., & Frank, E. (2005). *Data Mining - Practical Machine Learning Tools and Techniques.* San Fransisco: Elsevier Inc.

Xu, L., & Guojun, M. (2007). *An Algorithm to Approximately Mine Frequent Closed Itemsets from Data Streams.* Acta Electronica Sinica.

Yamamoto, T., Kitamura, R., & Fujii, J. (2002). An analysis of drivers' route choice behavior by data mining algorithms. *Proc. 81st TRB annual Meeting*.

Yamamoto, T., R., K., & H., F. (2001). The effects of a periodic vehicle inspection program on household vehicle transactions behavior. *Journal of Infrastructure Planning and Management*, 137-146.

Yen, J. (1971). Finding the K Shortest Loopless Paths in a Network. *Management Science, Vol. 17, No. 11,*, 712-716.

Zadeh, L. (1998). Roles of soft computing and fuzzy logic in the conception, design and deployment of information/intelligent systems. *Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Applications*, 1-9.

Zhang, L., & Levinson, D. (2008). Determinants of Route Choice and Calue of Traveler Information - A field experiment. *Journal of the Transportation Research Board*, 81-92.

Zipf, A., & Jost, M. (2006). Implementing adaptive mobile GI services based on ontologies: examples from pedestrian navigation support. *Comput. Environ. Urban Syst. 30*, 784-798.

APPENDIX 1 - PERSONALIZED ROUTING ALGORITHM ACCORDING TO PARK ET AL. (2007)

Existing navigation systems help drivers to navigate through traffic and hence improve driving comfort. Past researches have concluded that the satisfaction level of the systems appears to be relatively low which can be related to the level of network familiarity. A reasonable approach to improve the user satisfaction is to personalize navigation systems by developing a route planning process which would be able to learn user preferences through interaction with the driver and enabling it to plan better routes for that driver in future operations. Within this interaction it is important to acquire knowledge on the preferences whilst minimizing efforts on the users, for example by discovering regularities when observing repeated route choice behavior.

The paper of Park et al. (2007) discusses various methods which could be applied to establish a user model for routing algorithms. It compares the discrete choice analysis models, based on utility maximization, with the non-parametric methods like data mining. Main characteristic of the latter is that the non-parametric methods do not require estimating any parameter by describing the distribution of variables and furthermore do not assume any particular form of functions.

Within the family of non-parametric models, two distinct algorithms can be distinguished which can be used to accommodate route choice rules discovered from the observed (revealed) travel behavior data. Firstly the Artificial Neural Network (ANN) which provides models of data relationships through highly interconnected, simulated “neurons”. These “neurons” accept inputs, apply weighting coefficients and feed their output to other “neurons” which continue the process through the network to the eventual output. The alternative, the Decision Tree Learning (DTL) algorithm uses a decision tree as a predictive model which maps observations about an item to predict the future target value of an item. The authors choose to apply the DTL algorithm although the ANN algorithm outperforms other machine learning methods. Reason is that the ANN methods utilizes a ‘black box’ and supplies output that is very difficult to interpret. Opposed to the ANN method, the DTL algorithm provides a more comprehensible model structure which simplifies the interpretation of the modeling results.

Main aim of the paper of Park et al. (2007) was to explore the advantages of the DTL algorithm for developing adaptive route guidance, within the study a decision tree is constructed by making use of the C4.5 algorithm. This specific algorithm has been widely applied, the classifying is straightforward and it is fairly simple to apply.

System architecture

Two route guidance architectures can be identified, firstly the ‘autonomous navigation’ in which the car functions largely independently. The car receives traffic information which should subsequently be evaluated by the on-board navigation device. The second architecture, the so called ‘supported navigation’, consists of navigation devices which receive pre-determined candidate paths which are generated by the traffic information centers and afterwards downloaded to the subscribed cars. To simplify the research the authors of Park et al. (2007) only focused on the route advices based on the autonomous system architecture.

The system design for adaptive personalized routing, as proposed in the research or Park et al. (2007), initiates with the generation of a set of feasible routes based on the available network data and traffic information. Subsequently the predictive model assesses all candidate paths and selects the route which fits the user. This route can be subsequently presented to the driver and after reaching its destination the system determines the implicit reaction of the user by comparing the observed route with the suggested route. This approach, in which the passive user feedback (expressed as acceptance, rejection of deviation on-route) is used, enables the system to acquire knowledge of user preferences effectively without asking the drivers explicitly. If the user feedback is negative, the learning process is executed, leading to the updating of route selection routes.

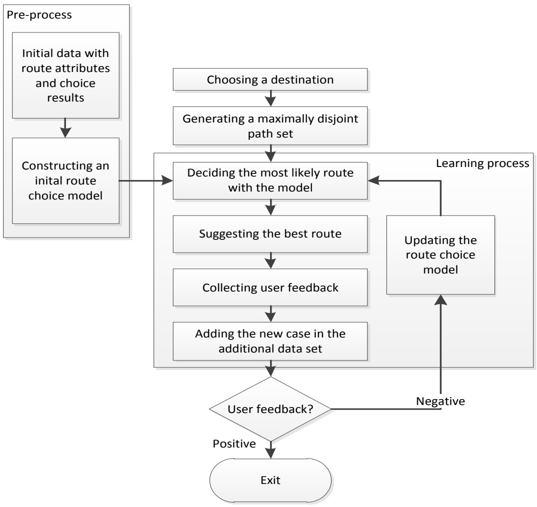

Pre-process
Initial data with route attributes and choice results
Constructing an inital route choice model
Choosing a destination
Generating a maximally disjoint path set
Learning process
Deciding the most likely route with the model
Suggesting the best route
Collecting user feedback
Adding the new case in the additional data set
Updating the route choice model
User feedback?
Negative
Positive
Exit

FIGURE 29: ADAPTIVE ROUTING SYSTEM ARCHITECTURE

The learning algorithm can be defined as 'deducting knowledge from experiences with respect to some classes of tasks and performance measure'; with respect to adaptive route guidance the following three elements can be described as:

- A task is the search for a route which best fits driver's preference using a decision tree;
- A performance measure is the extent to which the route corresponds to the driver's preference;
- An experience comprises of personal, route and trip characteristics along with the route choice.

Experimental design

Within the paper of Park et al. (2007) experiments with the learning model were carried out to analyze its application and learning ability in the context of route choice. During the research it was difficult to observe actual route choice data, at the time the research was executed no methodology and resources were available to develop a framework to follow a large user group. As an alternative the route choice experiments were generated using the simulation program 'ICNavS'. The first major step was to collect suitable data on route attributes and driver choices. Subsequently the best routes are determined by lexicographical rules and utility maximization rules described in Peeta and Yu (2005). After this pre-processing, a set of routes between the origin and destination is input to the learning model and a route is chosen by initial decision tree. If the result is different from the observed (revealed) choice, the decision tree is updated. In order to gain more accurate results of the adaptive learning algorithm and to facilitate the comparison of the results, three different sets of route selection routes were applied; two sets of lexicographic rules and one set of utility maximization.

Results

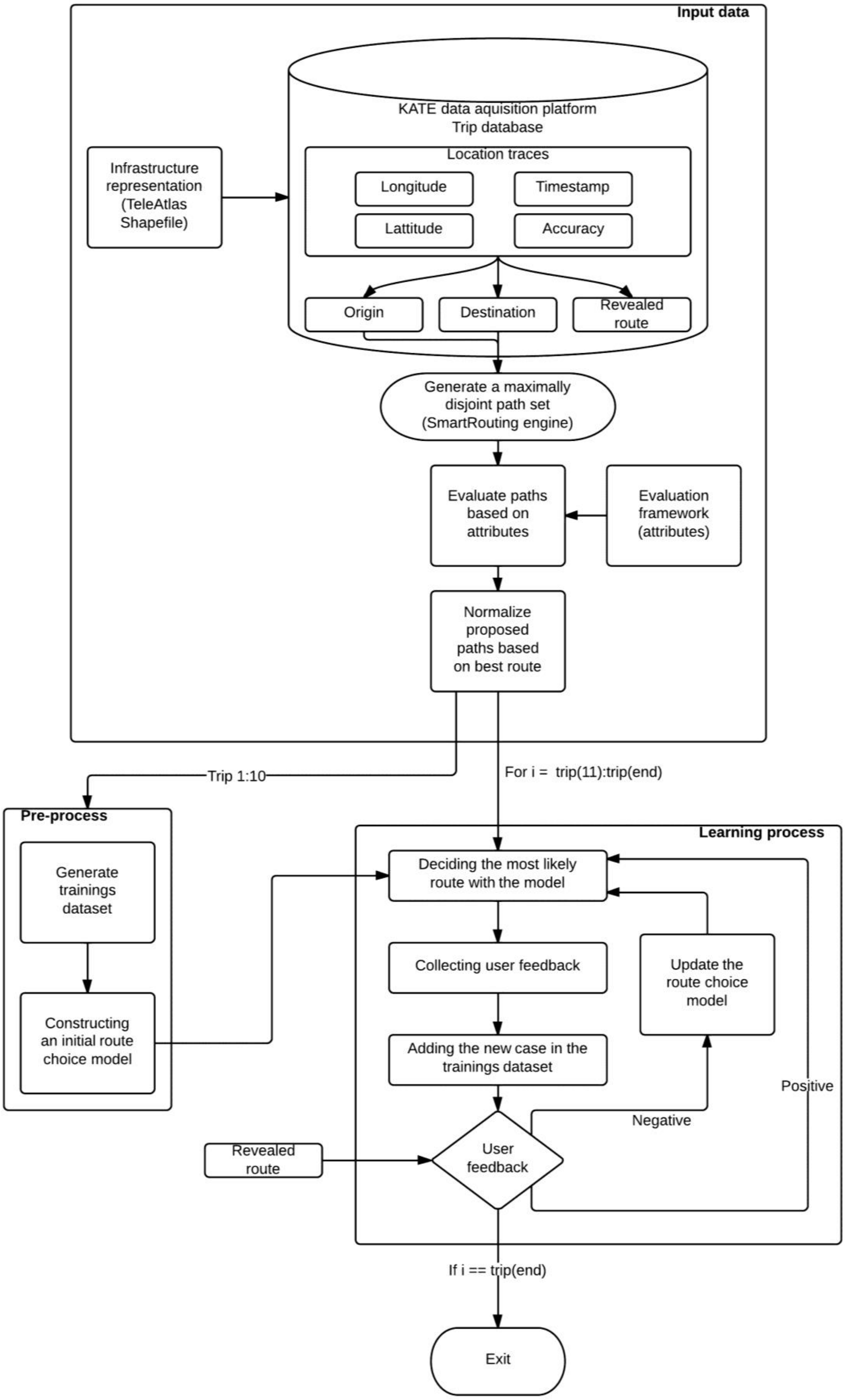
Various indicators were used to assess the performance of the various route choice algorithms. The first indicator was the number of updates of the data set, since the decision trees are updated whenever the actual choice of the driver is different from the predicted choice this measure is valid. A second efficient method to represent the predictive accuracy of the various algorithms was to plot the percentage of predictions that correspond to actual choices made by the user over the period of time that the system has used. For the purpose of analyzing how well the models accommodate the route selection rules, error rates of each tree were computed by dividing the number of incorrect choice decisions predicted by the current tree by the size of the data set for constructing the current model so far. This result shows how well the adaptive routing algorithm is learning from past errors.

From all the combined results it became apparent that the DTL approach seemed to outperform the UM approach in terms of human interpretability and predictability, after the training period the average number of updates decreased which indicated that the proposed routes were matching the observed routes. There is however no evidence that indicated that the decision tree models are superior to MNL models in terms of predictability, the predictive accuracy of the decision tree models does show a consistent trend over the three tests while the traditional models demonstrated variable predictive accuracy depending on the data used.

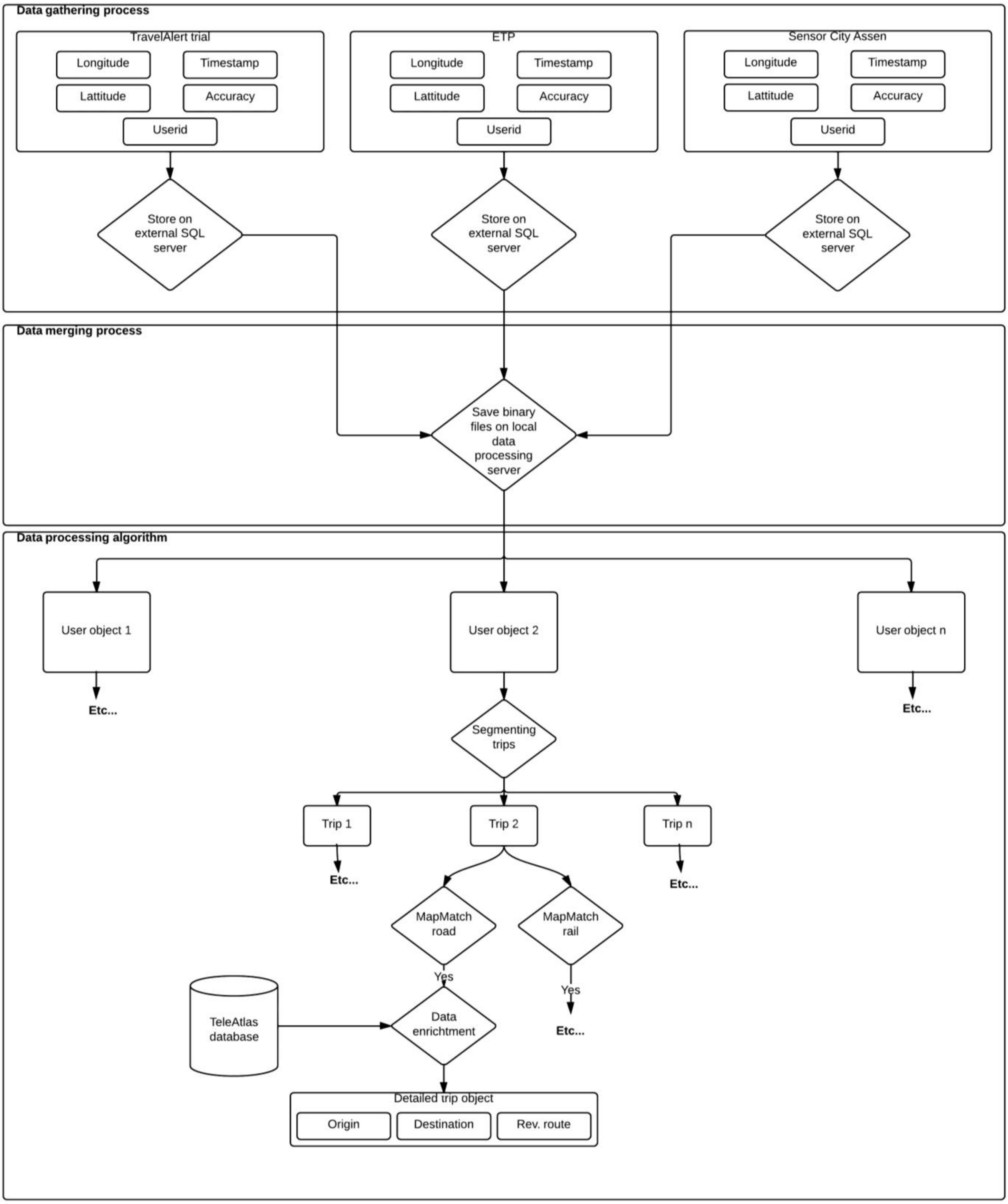
Directions for future work

From the results, as described above, a number of directions for future works became apparent with which route guidance could be applied in real life situations. Firstly it was advised to develop a user interface to provide routing information and collect user feedback. Furthermore additional attributes, which were excluded due to the lack of information in the simulation study, about the transport network and the surrounding area should be considered. Also, it was advised to device a method to deal with qualitative attributes such as aesthetics (e.g. distance through natural environments) or driving comfort (number of stops and turns, which also affects the route choice of drivers. Lastly additional process of incorporating user preferences in route generation should be considered, for example by weighting certain attributes related to the users preferences in the process of computing generalized link costs.

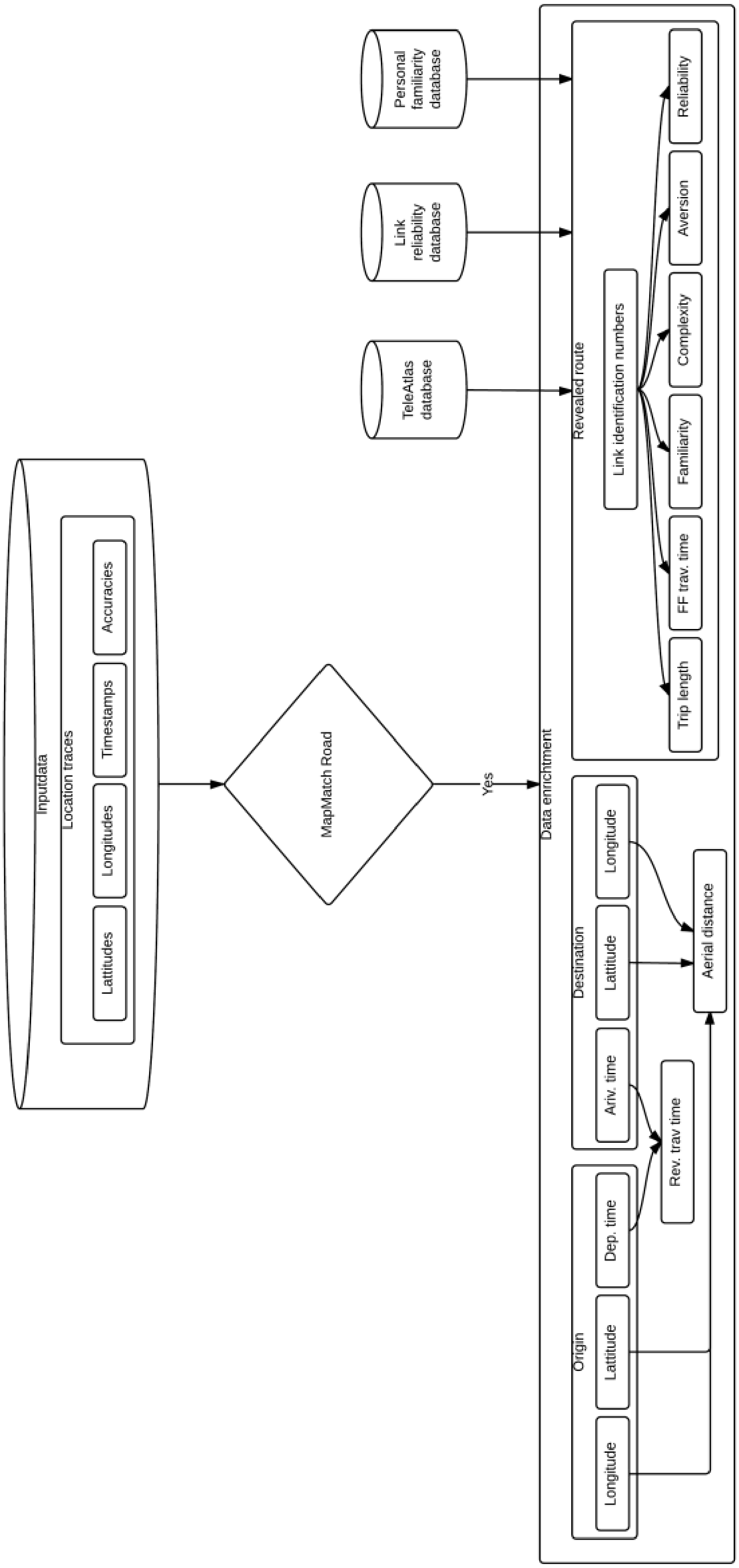
APPENDIX 2 – PERSONALIZED ROUTING ARCHITECTURE


Input data
KATE data aquisition platform
Trip database
Infrastructure representation (TeleAtlas Shapefile)
Location traces
Longitude
Timestamp
Lattitude
Accuracy
Origin
Destination
Revealed route
Generate a maximally disjoint path set (SmartRouting engine)
Evaluate paths based on attributes
Evaluation framework (attributes)
Normalize proposed paths based on best route
Trip 1:10
For i = trip(11):trip(end)
Pre-process
Generate trainings dataset
Constructing an initial route choice model
Learning process
Deciding the most likely route with the model
Collecting user feedback
Update the route choice model
Adding the new case in the trainings dataset
Positive
Negative
Revealed route
User feedback
If i == trip(end)
Exit

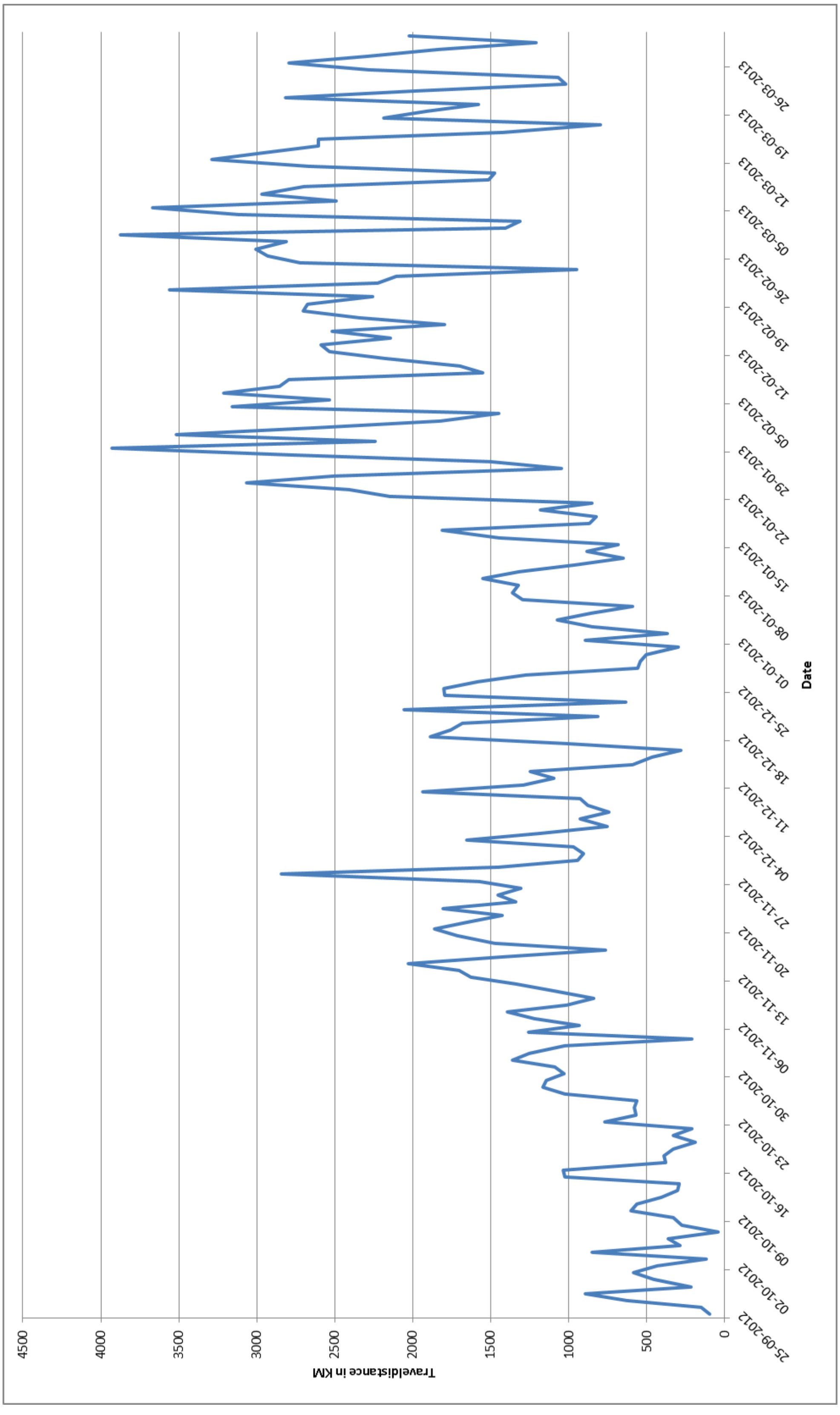
APPENDIX 3 – DATA PROCESSING ARCHITECTURE



Appendix 4 – Data enrichment procedure



APPENDIX 5 – DAILY DISTANCE TRAVELLED


Traveldistance in KM
Date