TU/e

MEASUREMENT AND EVALUATION OF TRANSPORT NETWORK PERFORMANCE USING GPS DATA

X. E. K. Ezechiëls

2015/ 2016

Student# 0918085
Construction Management and Engineering
Faculty of the Built Environment
TU/e

Colophon

This document is the master thesis of Xaief Ezechiels. This thesis is created in fulfilment of the requirements for the Master of Science degree in Construction Management and Engineering at the Technical University of Eindhoven.

Thesis submission	Friday August 19th 2016
Thesis defense (Final colloquium)	Wednesday August 24th 2016
Place	Faculty of Built Environment of the Technical University of Eindhoven (Vertigo)

Author	Xaief Ezechiels xaief_ezechiels@hotmail.com

Graduation committee

Chairman	**prof.dr.ir. B. de Vries**	TU/e
1st supervisor	**dr. Q. Han**	TU/e
2nd supervisor	**dr. T. Feng**	TU/e
3rd supervisor	**ir. A.J. Jessurun**	TU/e

Keywords

Transport network performance; travel time reliability; circuity; GPS data; performance measures; commuter; spatial partition; temporal partition; transportation mode

Document information

Pages	95
Words	26041
Chapters	7
Tables	23
Figures	43
Formulas	9
Appendices	11

A copy of this document is available on: http://repository.tue.nl/

Preface

During my two year Master track of Construction Management and Engineering the word "optimization" sparked my interest a lot. And being inspired by the never ending road construction work in the Netherlands I knew anything good can be made better. Continues optimization is progress. This led me to make it a hobby to sometimes just sit, look around, identify problems and come up with creative sustainable solutions.

When the time came to choose a graduation topic, I gravitated towards optimizing the intercity rail network as overcrowded trains and delays where becoming a big problem in the Netherlands. Unfortunately due to privacy issues the necessary resources for the study were unavailable. Luckily I also realized that for continues optimization network performance evaluation is very important. For this reason I choose to write this thesis on road network performance measurement and evaluation.

The execution of this study and the creation of this thesis document would not have been possible without the guidance and expertise of my graduation committee: Mrs. Han, Mr. Feng and Mr. Jessurun. Their feedback and expertise always inspired me to work hard, be very thorough and deliver quality work. Thank you to Mrs. Han for your patience and strict management, to Mr. Feng for your assistance and support during the entire graduation period. And to Mr. Jessurun for the last minute feedback, which helped a great deal in the last phases of the thesis.

Changing the world has always been my dream, and at an early age I knew that in order for me to chase my dream I had to set goals bigger than myself and have patience. At that time I did not realize this also required making large sacrifices. Not only by me, but also by everyone around me. My journey up to this point was long and hard, but it was always in good company. For this reason I would like to extend my gratitude and appreciation to everyone in my circle for their love and support. Also I would like to dedicate this thesis to my mother, who always put my dreams before her own. "Look mom, WE did it!!"

Xaief Ezechiels
Eindhoven, August 2016

Summary

With the increasingly growing population urban areas are becoming denser. A direct consequence is increasing traffic congestion. For this reason continuous optimization of the transport network is vital. However, for continues optimization, the transport network performance first needs to be measured.

Problem description

The increasingly growing population will have very serious consequences in the future. Especially for small a country with an already enormous population, like the Netherlands. Apart from available resources, over-populating in an already "dense" country will also cause social problems. Two social problems that already exist are the cost of housing and job availability. When these two cross paths, even more problems arise. These include long travel times and traffic congestions. These problems have a negative effect on the productivity and the wellbeing of people and the companies they work for.

This thesis proposes an entirely different way of creating a solution for this growing problem. By proposing a method with which any transport network can easily and at low cost be evaluated the government can optimize its public transport system and its city planning. The problem areas can be identified and targeted. Additionally the method can be used to simulate future scenarios. This thesis describes a way to evaluate the performance of a transport network by using performance measures which relate to traffic congestion. The proposed method will provide decision makers with a better understanding of current travel demand and transport infrastructure problems as well as help them in creating effective solutions.

Research motivation

The research idea originated in wanting to create a method to optimize the transportation network regarding railway transportation. The research would include using smart card data from OV chip cards to build OD matrices and eventually create a model, which policy makers could use to analyses the transportation network and optimize it if needed. Unfortunately due to unavailability of OV-chip card data, the original study was altered to investigate road transportation network instead of rail transportation network.

Over the years different solutions have been applied to ease traffic congestion. As a solution the Dutch government has invested in their public transport system for years and encouraging its citizens to use it. But with a growing population the capacity of this system will reach its limits at some point.

This thesis proposes an entire new way of creating a solution for this growing problem. By proposing a method with which any transport network can easily and at low cost be evaluated the government can optimize its public transport system and its city planning. The problem areas can be identified and targeted. Additionally the method can be used to simulate future scenarios.

Methodology and data

In the literature different approaches were identified. To filter the best fitting approaches they are assessed on two criteria. These include applicability and feasibility. The applicability of each method highly depends on its relevance to the objective of this thesis. Additionally on the aspect of feasibility of the methodology depends on the available data and how it can be utilized. From the identified approaches the most relevant ones are chosen. These contain important features that are then used to form the methodology of this thesis.

The methodology of this thesis is divided in five main steps. These include:

1) Processing GPS data
2) Modelling region wide travel patterns
3) Generating performance measures
4) Detect regions with serious mismatch problem
5) Method evaluation

Data

For this study GPS data, collected by the Technical University of Eindhoven, was used. The data is for the period March until May of the year 2014. This collected data was recorded in the Noord Brabant region of the Netherlands, and is part of a massive data collection which was held the entire year of 2014. The year was split into four parts and in each part, respondents were recruited to join the survey for three consecutive months.

The GPS data consists of location information which uses the geographic coordinate system. Here each recorded location is described as a set of coordinates: (longitude, latitude). Other important information the data contains includes timestamp information and the method of transportation.

Methodology

The analysis starts by utilizing and cleaning the raw GPS data. The available data consisted of three parts, which include journey data, stage data and GPS trace data. Each part contains important information for the analysis method and therefore were combined to one data table. Generally GPS data contains random errors due, which needed to be filtered out before continuing. Within this context data with longitude and latitude zero were removed. Additionally all data with transportation method other than car was filtered out. The actual distance and speed are calculated and as a very last filter all data with speed higher than 120 km/h was removed. These commuters do not experience any serious problems in the transport network and are therefore left out of the scope of this thesis.

Once the data is utilized trip characteristics can be extracted. These include the distance from origin to destination, the distance from each recorded location to the following recorded location and of course the speed and route directness between each recorded locations.

After the characteristics are extracted the performance measures can be generated by applying spatial and temporal travel partition for modelling commuter travel patterns. Spatial partition includes applying a grid based method which splits the study area into smaller study regions. Temporal partition includes making each trip subject to the time of day and the type of day, e.g. weekday or weekend day. After this the performance measures V, R and TT can be generated.

With the generated measures the problem regions can easily be detected by comparing the measures to their respective threshold values. For example, any speed lower than a threshold of 20 km/h indicates that the capacity of the routes within this region may not be sufficient to meet the current travel demand. A lower speed could also imply that people have no other choices than to take these inefficient routes for traveling between the two regions (Cui, et al., and 2016-B). Within this context, by comparing each of the three measures, which include V, R and TT, with a threshold value, the regions experiencing these problems are detected. Additionally regions with serious congestion problems are detected by combining the three different measures into one, this forms the problem indicator: V < TV and R > TR and TT < TTT and M > TM (combination of temporal, spatial and travel time reliability performance measures).

The evaluation of the methodology consists of a comparison of the proposed approach in this thesis with the approach found in literature (Cui, et al., 2016-A). Here taxi data is generated from a fraction of the raw GPS data, approximately one week data, which was used in the approach of this paper. After the measures are generated for both approaches, the problem regions are identified using a variation of threshold values. In this way each approach will identify a variety of regions with serious problems. The results are then statistically analyzed and the quality of each approach can be derived.

Results

The analysis method

Regarding the data processing, in this thesis GPS data from three months is used. This data was first utilized before it could be analyzed and for the very first step in utilizing the GPS data the software MS Excel and MS SQL were used. This step mostly consisted of merging tables containing vital information.

Regarding travel pattern modelling, the first step in doing so is the spatial and temporal partition of the trips. For the spatial partition a grid was created for the Netherlands using MapInfo Professional and generating a total of 90117 cells or regions with a size of 1 km^2. After this, the coordinates were imported and each GPS point was linked to a region. This was done by joining the grid table with the data tables for March, April and May 2014. After this the temporal partition was executed by dividing the trips according to the time of day and the type of day. For this study the focus is on weekdays which are equivalent to business days, because this is when most people travel to work and when poor network performance has the worst implications. Additionally the chosen time of day is during the morning peak, between 7:00 AM – 9:00 AM.

After the temporal and spatial partition the performance measures were generated for all three months. These include the average speed per trip V, the average route directness per trip R, the average amount of trips and the average travel time per kilometer per trip TT. Here the performance measures were analyzed for each month on business days during the morning peak.

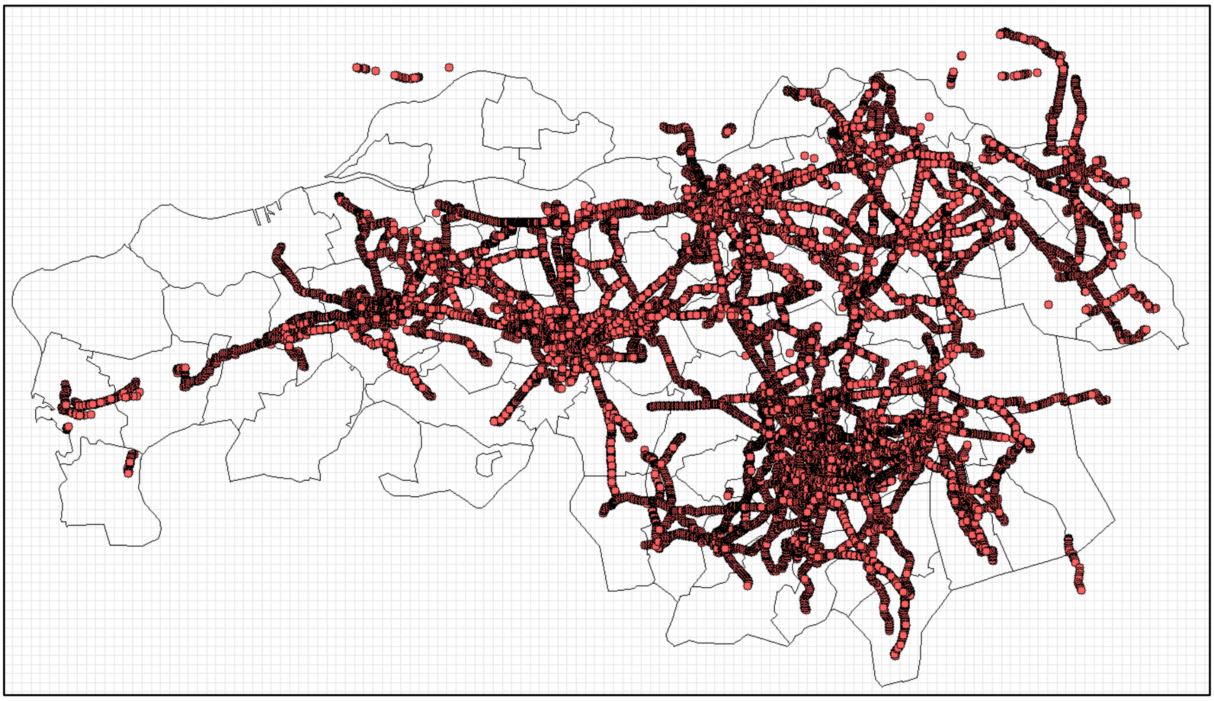


Figure 0-1. Travel pattern April 2014

After this the generated performance measures were used model the travel pattern of the commuters for each of the three months in 2014. Figure 0-1 present the travel pattern of April 2014.

Identifying serious mismatch regions

After the performance measures are generated, the next step is to set to identify the regions with serious mismatch problems. This step begins by extracting the threshold values for the performance measures from literature. These include TV, TR and TM. The threshold value for travel time reliability, TTT, is estimated using the percentile statistical method. This method estimates the probability of the worst travel time that commuters may experience once per month in percentages. This estimation is done using IBM SPSS Statistics 22.

Finally the problem regions are identified using five indicators, which include: VM, RM, VRM and VRTTM. These indicators are used to detect the areas in which certain problems occur. T

Figure 0-2 presents the amount of regions detected by each indicator. Further in the thesis thematic maps are presented which show exactly in which part of Noord Brabant each problem occurs.

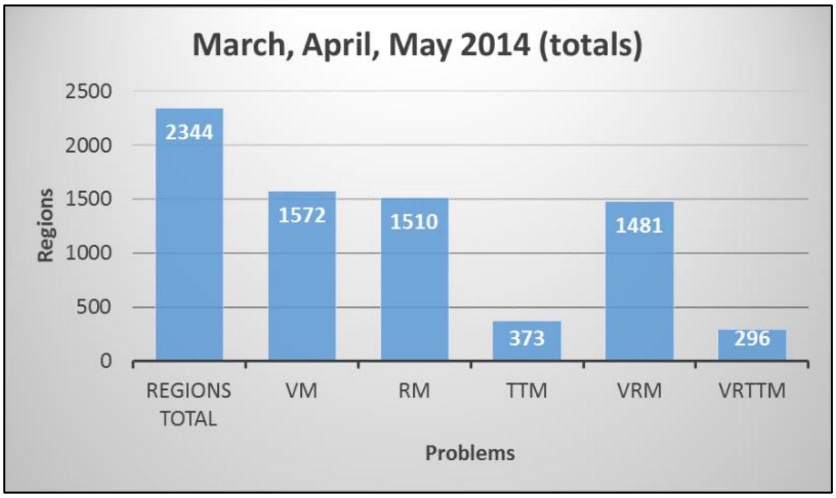


Figure 0-2. Results identified regions

Method validation

With regards to the method validation, the thesis approach was compared with an approach from literature. Here the same data was used and analyzed using both methods/ approaches. The results were then statistically analyzed and compared. All in all, the statistical analysis revealed that the thesis approach yields more accurate and more reliable results than the approach from the literature. With this it can the validation of the method is concluded.

Conclusions

This thesis provides a reasonably easy way to measure transport network performance as it proposes an approach to evaluate network performance by identifying regions with a serious mismatch problem between travel demand and transport network services.

The approach is validated with the most effective validation method. The approach is compared to the original approach it is based on from literature. The validation procedure revealed that the thesis approach produces better and more accurate results than the approach from the literature. This is proven by the fact that the statistical analysis of standard deviation and the coefficient variation of the identified regions using the variation of average speed, travel time, and the amount of trips yield lower values for the thesis approach.

All in all this thesis can serve as a guide for measuring performance using easily obtained GPS data. The thesis approach provides policy makers ad traffic managers with a way to evaluate transport network easily and at low cost. Additionally the data collection method is not labor intensive. The approach can be used strategically in solving congestion related problems in different phases. In this sense policy makers can use different problem indicators to identify regions with specific problems, which they can solve faster. Additionally, due to the fact that many regions experience more than one problems, policy makers and traffic managers can strategically identify regions with experiencing multiple problems and find one solution to solve them both.

Recommendations

The analysis method can be used to detect in which parts of the transportation network certain problems occur and at what time. The recommended next step is to examine the problem area and make a route cause analysis for the problem. After the cause or causes for the problem is identified a sustainable solution can be tailored for the problem. It is also recommended that the seriousness of a problem in a certain area should be taken into account for tailoring sustainable. The seriousness of the problem can be measured by calculating the occurrence of a specific problem in a certain area with probabilities. By determining the probability of a problem occurring more times than a certain threshold amount, traffic managers and policy makers can strategically target these areas first. It is also recommended to study the side-effects of targeting these areas first, as this could cause the probability of certain problems occurring to decrease or even disappear in other areas.

In this thesis the travel time reliability is estimated using the 95 percentile method. However a more accurate way of determining the travel time reliability would be to calculate it. By using traffic flow data the travel time reliability can easily be calculated. This calculated value of the travel time reliability will provide a more accurate threshold reference point as it takes a lot more aspects of the transport network and the commuter into consideration. This includes stopping times etc. However it should be noted that traffic flow information is very hard to obtain. Most traffic flow information is only available in real time. Creating a database which stores this information is also highly recommended.

A large portion of the analysis described in this thesis was executed using MS Excel and MS SQL. Although these tools did get the job done, the GPS data was almost too much to process. The aim was to do the analysis in this study in a way that everybody could understand and replicate without buying expensive special software tools. However these special tools are more advanced and created to process large portions of data easily. For this reason it is recommended that for future work other, more advanced tools are used which can process bulk data in a shorter amount of time. The recommended tool to use for follow up research is FME desktop. FME is a software tool which has the ability to process bulk data easily and is basically an integrated collection of Spatial ETL tools for data transformation and data translation and contains a library of over 5,000 coordinate systems and provides native support for location data. <u>It is highly recommended that FME be used for follow up research on this topic in the Netherlands</u>. Apart from the fact that the GPS data can be processed very fast, this data can also be translated from the basic geographical coordinate system into data which is coherent to the RD-New coordinate system of the Netherlands. This will lead to higher accuracy of the results overall. As the FME software can translate GPS data from one coordinate system to almost any coordinate system designed for a given country, it is also highly recommended to use this tool for follow up research for evaluating transport network performance in other countries.

Abstract

This thesis proposes an approach to measure transport network performance by identifying the mismatch between travel demand and transport network services. For this analysis three performance measures are used, which include temporal performance measures, spatial performance measures and travel time reliability performance measures. These measures are generated by analyzing individual commuter travel information in the form of GPS data. The generated measures are then used to construct problem indicators for detecting regions where commuters experience serious mismatch problems. These problems in turn relate to the measures and include poor connectivity, slow driving speed and longer travel times. The results show the validity of the proposed approach and that it produces accurate and reliable results. The thesis also proved the effectiveness of using GPS data to measure transport network performance. The approach also proved to have social benefits, as it provides commuters, policy makers and traffic managers with a modern, easy and relatively cheap way to identify problem areas in a transportation network.

Table of Contents

1. Introduction

1.1 Problem definition

The Netherlands is a small country with an enormous population, which is growing fast. This will have very serious consequences in the future. Apart from available resources, over-populating in an already "dense" country will also cause social problems. Two social problems that already exist are the cost of housing and job availability. When these two cross paths, even more problems arise. These include long travel times and traffic congestions. These problems have a negative effect on the productivity and the wellbeing of people and the companies they work for. The two afore mentioned issues relate to poor transportation network performance. A transportation network can be described as the foundation of any kind of movement or flow from one location to another. This includes road networks, rail networks, bicycle lanes, etc. For this reason, any bottlenecks in the network today will surely result in even worse problems in the near future.

Network performance can be analyzed in different ways, and from different perspectives. With the help of emerging technology, new data collection methods and availability of open data source information new opportunities are presented to better understand and to improve the network performance analysis. In the last few year's commuters, traffic managers and policy makers have also showed much interest in the reliability of travel time, making it a very important performance indicator (Mahmassani, Hou, & Saberi, 2013).

To optimize the transportation network, it first needs to be analyzed. This part starts by identifying the bottlenecks in the network. To do so information about the network is necessary, and depending on the type, the amount and the quality of data the bottlenecks can be identified easily. The difficult part is collecting and processing the data. There are different types of travel data and for each there are different ways of collecting them.

GPS data has gained a lot of popularity in the field of transportation, in recent years due to its high accuracy, high resolution of spatial and temporal information and light burden for respondents. By utilizing GPS devices, the movements of a person or vehicles can automatically be recorded with high accuracy. This information can then be used to construct Origin-Destination (OD) matrices. OD matrices are a way of representing travel patterns across a network and can be seen as a fundamental method for describing how commuters move through a network. An OD matrix is built with different information. This includes the type of day, the time of day etc. Additionally GPS data can also provide information needed for determining travel time reliability.

1.2 Research motivation

The research idea originated in wanting to create a method to optimize the transportation network regarding railway transportation. Personal observation of longer than a year shows that during the morning and evening peak trains are full and traffic congestion on roads is normal. Another observation shows that the majority of the commuters, using the train, are of a certain age category. For this reason it was be assumed that these commuters are living and working or going to school in different cities, and that the main consequence is traffic congestion during morning and evening peaks. The research would include using smart card data from OV chip cards to build OD matrices and eventually create a model, which policy makers could use to analyses the transportation network and optimize it if needed. Unfortunately due to unavailability of OV-chip card data, the original study was altered to investigate road transportation network instead of rail transportation network.

Over the years different solutions have been applied to ease traffic congestion. As a solution the Dutch government has invested in their public transport system for years and encouraging its citizens to use it. But with a growing population the capacity of this system will reach its limits at some point.

This thesis proposes an entire new way of creating a solution for this growing problem. By proposing a method with which any transport network can easily and at low cost be evaluated the government can optimize its public transport system and its city planning. The problem areas can be identified and targeted. Additionally the method can be used to simulate future scenarios.

1.3 Objectives and aims

Research objectives
The main objective for this study is to develop a method to evaluate the performance of a transport network by using performance measures which relate to traffic congestion. The proposed method will provide decision makers with a better understanding of current travel demand and transport infrastructure problems as well as help them in creating effective solutions.

Aims

- The method should be clear and the indicators should be extractable from travel information
- The travel information should be available or easy to obtain and at low cost
- The evaluation method should benefit commuters, traffic managers and policy makers
- The method should provide a way to detect congestion areas
- The approach should be applicable for different types of transport networks and for different modes of transportation

Research limitations
The research boundaries are presented in table 1 and are chosen based on the outcome of the literature study in chapter 3 and the available data.

Table 1. Research limitations

Data type	GPS data (journey, stage and traces)
Study area	Regional scale: Noord- Brabant
Study period	March, April and May of 2014 (three months)
Mode of transportation	Car

Time of day	Morning peak (7AM- 9AM)

1.4 Research questions

In this paper the following research question is answered to ultimately achieve the main objective:
"To what extent can commuters' travel demand, derived from GPS data, be analyzed with the transport network services and travel time reliability, in order to evaluate the performance of a transportation network in the Netherlands?"

In order to preserve the quality of the study, the main research question is divided into the three sub-questions:

- Which measures are needed to accurately analyze network performance?
- To what extent can travel time reliability of the transportation network be generated from commuters travel data, specifically GPS data?
- In what way does the travel time reliability relate to the overall performance of the network?

1.5 Research design

1.5.1 Research approach

The literature study on the performance of transport network and the applicability of modern data sources within this field is the bases of this research. It provided the study with possible methodologies, a guide for what kind of data was needed and from what sources it could be obtained. Additionally it contributed a lot into the overall execution of the study.

The research approach is divided in three steps

1. Utilize GPS data
2. Extend the approach for performance analysis and implement this method
3. Evaluate proposed approach

1.5.2 Research model

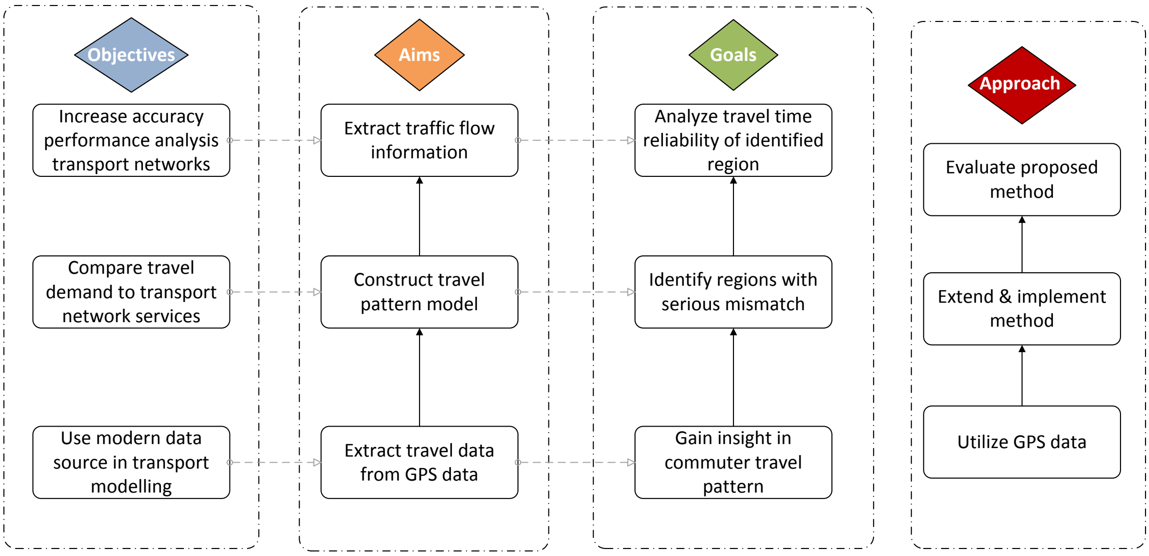


Figure 1-1. Research model

1.6 Expected results

The expected results of the developed method should describe the objective and answer all research questions. Within this context, the results will:

- Show policy makers detailed information about the effect of current travel demand on the transport network performance.
- The results should also give an insight on the performance of the transport network, with regards to identifying and displaying poor problem areas and areas with poor travel time reliability.
- Overall this paper should act as a guide for policy makers or traffic managers to accurately analyze transportation networks and identify problem areas. In this way it will contribute to designing effective solutions to current problems within the transport network.

1.7 Validation

As this paper actually proposes an extension on an existing approach, the validation is very important. For this part, a sample data set is created to represent taxi trip data. A random experimental week will be chosen. The data from this weak will be processed not only with the proposed extended method and the existing method, but also with combinations of both. Here after the results and the methods will be evaluated statically on their sensitivity and reliability

1.8 Report structure

The remainder of this thesis report is structured as follows:

- Chapter 2 describes the glossary, and includes the list of tables, figures etc.
- Chapter 3 describes the literature review results. Relevant references from the literature are discussed regarding network performance, travel time reliability and the use of GPS data in travel demand modelling. The focus here is on identifying approaches, used in previous studies, to measure network performance;

- Chapter 4 contains the methodology used to answer the research questions and process the data;
- Chapter 4 comprises the results of the performance analysis, regarding travel time reliability and network mismatches related to travel demand and transport network services. Furthermore, traffic flow information is extracted from the GPS data for the purpose of measuring travel time reliability on network level;
- Chapter 6 presents the conclusions from the research and considers the implications of the results. Each conclusion describes an answer to each research questions. Ultimately some additional conclusions are added;
- Chapter 7, the final chapter, presents the recommendations based on the formulated conclusions. These recommendations also include advice on utilization of the methodology and results regarding further research.

2 Glossary

2.1 Tables

2.2 Figures

2.3 Formulas

2.4 Glossary of Terms

Circuity / route directness	The circuity or route directness of travel paths, describe the ratio between the actual travel distance between a pair of origin and destination in the network and the distance of these two location.
Coefficient of variation	The coefficient of variation (CV) is a standardized measure of dispersion of a probability distribution or frequency distribution.

Commuter	Individual travelling from one place to another.
Destination	The final location of a trip, where the next activity is performed.
Euclidean distance	The distance between two locations measured by a straight line between them.
Evening peak	The peak period of travel demand during the evening. In this research the evening peak is defined from 4 pm to 6 pm.
Hiversine distance	The distance between two points, based on the law of Haversine states.
Holidays	Any national holiday in the Netherlands.
Imputated/ inferred GPS data	GPS data where missing (raw) data was replaced with substituted values.
Journey sequence	A journey is comprised of multiple journey sequences and each journey sequence contains one or more stages/ trips.
Morning peak	The peak period of travel demand during the morning. In this research the morning peak is defined from 7 am to 9 am.
Origin	The start location of a trip.
Performance measures	A quantifiable indicator used to assess the performance of a transportation network (how well a transportation network is achieving its desired objectives).
Raw GPS data	Raw data collected with GPS devices. This data mostly contains irrelevant information as well as errors and recording time gaps.
Region	Study area divided in smaller parts or "regions".
Spatial partition	Division of the study area into smaller areas using a grid based method
Standard deviation	A quantity expressing by how much the members of a group differ from the mean value for the group.
Temporal partition	Division of each trip into different time periods of a day and different type of days.
Thematic map	A thematic map displays information regarding a specific data set, such as the average speed with which commuters move in a certain area.
Time of day	A specific period during the day.
Trace data	The trace data is basically the raw GPS data of each commuter, and contains all location data from a commuter from when it starts recording until it stops recording.
Transport network	A transport network, or transportation network is a realization of a spatial network, describing a structure which permits either vehicular movement or flow of some commodity.
Transportation mode	The type of transportations, relating to the infrastructure and vehicles used.
Travel pattern	The common movement of people between different locations.
Travel purpose	The reason for travelling, related to the activity that is performed.
Travel time reliability	A transport network performance measure which describes the consistency or dependability in travel times, as measured from day to day or across different times of day.
Trip	The travel of an individual from one activity location to another, which may consist of one or more trip legs.
Trip duration	The amount of time between the start and end of the trip
Type of day	Weekday, Holiday or Weekend day.
Weekdays	Monday to Friday excluding holidays.
Weekend days	Saturday and Sunday excluding holidays.

3 Literature review

The third chapter of this paper describes the findings of the literature study, which took place during the first phases of the research.

This literature study is the bases of the research and was performed having the goals:

- Firstly the literature study must provide all necessary background information. This information provides a guide to understanding the applied data, methods and limitations of the study.
- During the literature study different research methods should be assessed with the goal of finding the most fitting methodology options. The results from using this method should provide realistic answers to the research questions.
- The literature study should provide the necessary information, which is important for the assessment framework. This includes formulating a comprehensive and consistent framework of assessment criteria for the methodology choice.
- Lastly the literature study is very important as it also registers previous papers, which are relevant to this paper and ultimately used as reference. The goal is to compare this study with previous studies, to get a more insight on the research topics, and to extract important information from these studies to finally create a new or combine discovered methodologies applicable for this paper. The literature study also serves as the foundation for corroborating conclusions and recommendations made in this paper.

A transportation network is the foundation for moving from one place to another. A transport network is a complex network that usually includes road networks, rail networks and bicycle lanes (Buhl, et al., 2016). Understandably any problems in the network today will have severe consequences in the near future. For this reason it is important for policy makers, traffic managers and even commuters that there is a reliable method to evaluate transport networks on performance aspects easy, fast and at low cost. Within this context the main focus is on traffic congestion, because congestion is a growing issue for almost any transport network anywhere in the world (Tsekeris & Geroliminis, 2013). Network performance can be analyzed in different ways, and from different perspectives. With the help of emerging technology, new data collection methods and availability of open data source information new opportunities are presented to better understand and to improve the network performance analysis. In this literature study previous studies on this topic are reviewed and the necessary information is gathered to form a methodology with which an attempt to evaluate the performance of a transportation network in the Netherlands will be made.

The remainder of this chapter is structured as followes:

- First transport network performance is discussed
- After this the performance measures are identified in the literature
- Then these measures are used to identify different methods
- After the previous step the discovered methods are discussed
- Then the necessary data is defined using the findings from the literature and the availability of data

- In the end the different methods are assessed on different criteria and an outline of this thesis is presented

Important to mention is that from here after **"this thesis"** describes this thesis, for which the literature review was conducted.

3.1 Introduction: Transportation Network Performance

A transport network can be described as the basis of all movement within a certain area, and is basically a network of infrastructure designed to realize vehicular movement or flow of some commodity from one location to another.

The objective of **this thesis paper** is to measure and evaluate the performance of a transportation network. For this reason various literature was reviewed to firstly get more insight on what network performance actually entails. After this it was important to identify what studies have already been conducted on this topic, the methodology used in these studies and lastly the findings of these studies. Analyzing the results was equally important, because they revealed the research limitations, the gaps and how their proposed method could or could not fit into the context of **this thesis paper**.

In recent years transport network performance has gained much attention as it is a key aspect in traffic planning and management. Transport networks can be measured in different ways depending on different characteristics of the network itself. A report by the Board of Transportation Research (Research Board of Transportation, 2010) states that "network performance measurement is an attempt to evaluate the transportation system as a whole, considering all modes of transportation, all potential strategies (e.g., capital versus operational investments), and all jurisdictions (e.g., state, regional, and local) ".The different methods of evaluating network performance are called transport network performance strategies. The indicators to assess these strategies are called transport network performance measures.

To strategically identify information relevant to the topic of **this thesis** a suitable network performance analysis process was developed based on the information found in the previously mentioned report (Research Board of Transportation, 2010). This process is divided into several steps, which include:

- Identifying performance measures
- Identifying strategies
- Assess strategies using measurements
- Defining the necessary data

3.2 Identifying network performance measures

As mentioned earlier, network performance has been a hot topic for some time now. In this sense numerous studies have been conducted to better understand transportation networks and evaluate their performance. These include studies by (Wu, Thai, Yadlowsky, Pozdnoukhov, & Bayen, 2015), (Amini, Peiravian, Mojarradi, & Derrible, 2015), (Gillen & Hasheminia, 2016)

and (Mishra, Welch, & Jha, 2012). In these studies authors have identified and validated performance measures which can effectively asses the performance of a network. Depending on the objective of the research and on the network which was assessed during the research different performance measures were found in the literature. By linking the objective of the previous studies to the objective of **this thesis** several relevant studies were filtered and investigated further.
In the paper of (Cui, et al., 2016-A) three particular performance measures were found. These include:

1) **Temporal travel efficiency measures**. These measures focus on the temporal aspect of the travel in the network and includes travel speed, travel times and congestion situations.
2) **Spatial travel efficiency measures**. These measures include route directness or circuity, which compares the ratio between the actual travel distance between a pair of origin and destination in the network and the Euclidean distance of these two locations.
3) **Accessibility measures**, which take into account both the spatial and temporal travel efficiencies as well as the distribution of land-use and activity locations across the urban area. These measures investigate the ease and extent to which the combination of land-use and transport systems enable commuters to reach activities and destinations by means of certain transport modes.

Additionally the literature also revealed the increasing importance of **travel time reliability** as a network performance descriptor for both travelling public and traffic managers and policy makers (Mahmassani, Hou, & Saberi, 2013). Travel time describes the time it takes an average commuter to move from one location to the other location (Elefteriadou & Cui, 2005). Travel time reliability is a performance measure which indicates how dependable the travel time on a given transport network is (Lyman, 2007). Although the definition of this measures can vary in different contexts, each definition is closely related to the variation of travel time.

3.3 Identifying network performance strategies

Transport network performance measurement studies have been conducted numerous times in the past. Each study was executed from a different perspective and with a different objective. For this reason, most studies have used different transport network performance strategies. A transport network performance strategy represents the methodology which researchers have used to evaluate the transport network performance. In this part of the literature study the identified network performance measures are used to identify transportation strategies that are aligned with the aim of **this thesis**.

Within this context two main strategies or methods were identified in the literature. These include:

1) **Evaluating transport network conditions**
Transportation networks are designed for commuting between two or more locations. Therefore the performance of these networks is also dependent on the mobility demand of commuters. This method focusses on the macroscopic characteristics of a transportation network and the travel demand of its users. This method is essentially

designed for detecting congestion areas within the network. As this strategy can be assessed with different measures, the detection process will vary depending on these applied measures. In the literature three studies were identified which contained three different assessments of this strategy. (See Table 2)

2) **Evaluating transport network reliability**
In transportation networks in large cities the race for space is becoming increasingly intense, and this issue effects both experts and residents who often face congestion (Oliveira, Portugal, & Junior, 2016). In this sense the reliability of such a network represents an aspect of congestion that is a consequence of unpredictable travel conditions, rather than everyday delays. The main performance measure which can be used to asses this strategy includes the travel time reliability (Uchida, 2014). In the literature three studies were identified in which the transport network reliability was evaluated on the basis of travel time reliability (see Table 2).

Although the two strategies use completely different performance measures and investigate different aspects of the transport network, ultimately both approaches were proposed to solve different congestion issues. The first strategy was created to detect congestion areas within the network and the second one was created to detect the possible level of congestion a transportation network. In this sense the two strategies are perfect ingredients for developing a suitable methodology for this research paper.

Based on the performance measures different assessments of the strategies were identified in the literature. These are presented in Table 2.

Table 2. Identified performance strategies and measures

Performance strategy	Performance measures	Assessment (found in literature)
Transport network conditions	Temporal efficiency measures	Origin-Destination (OD) prediction (Lu & Li, 2014)
	Temporal and spatial efficiency measures	Identifying mismatch between travel demand and transport network services (Cui, et al., 2016-A)
	Temporal, spatial and accessibility efficiency measures	Detecting network accessibility (Cui, et al., 2016-B)
Transport network reliability	Travel time reliability	Assessing transport network reliability and vulnerability (Oliveira, Portugal, & Junior, 2016)
	Travel time reliability	Two fluid model strategy (Hong S. , Lee, Bong, & Kho, 2005)
	Travel time reliability	Connecting travel time reliability and the Network Fundamental Diagram of Traffic Flow (NFD) (Mahmassani, Hou, & Saberi, 2013)

The next part of this chapter will describe the assessment of the performance strategies. Additionally the limitations, the gaps and how each particular strategy could or could not fit into the context of **this thesis** will be described.

3.3.1 Assessment network performance strategies

"O-D prediction"
In (Lu & Li, 2014) the authors used temporal travel efficiency measures to build a travel demand model directly based on an enormous amount of GPS data. The objective of this paper was to propose a method which could predict travel demand up to several hours later.

For the foundation of this approach authors used temporal efficiency measures to build a time dependent Origin Destination matrix. After the origin and destination position were obtained, they were matched into statistical sectors to form this time-dependent OD travel demand matrix. The input data for this approach was collected in February of the year 2010 in Singapore, and consisted of an enormous amount of taxi trip data.

Although the objective of this 2014 study is very interesting, it is not aligned with the one of being conducted for **this thesis**. The approach only takes one type of performance measures into consideration and thus making the OD matrix only time dependent. For this reason this approach has room for improvement. Combining or adding more performance measures should make the approach more accurate and more reliable.

Comparatively, **this thesis** also aims to identify and calculate performance measures. The difference is that the method should include not one, but multiple performance measures. For this reason the next strategies were identified.

"Identifying mismatch between travel demand and transport network services"
In 2016 (Cui, et al., 2016-A) sought to measure network performance by identifying mismatches between travel demand and transportation network services. In this study the authors used the previously mentioned approach proposed by (Lu & Li, 2014) as foundation and extend it by assessing the network performance with two performance measures, instead of one. These performance measures include temporal efficiency measures and spatial travel efficiency measures.

The study was conducted in a Chinese city Harbin and focused on using travel data from taxi trips to model the travel behavior within the network. For this study the aim was to also use a modern data source, with the intention of making the proposed method easily transferrable to other cities. The author's perspective on gathering the necessary data was inspired by the most common technique in the realm of transport modelling. This includes the Floating Car Data (FCD) method, which uses a batch of sampled vehicles equipped with navigation systems or any type of GPS technology, as moving sensors to build Intelligent Transportation Systems (ITS). The popularity of this method is due to its ability to monitor real time traffic conditions along road links and predict short term traffic. In earlier studies, this method already proved to be very effective.

By extending the original approach with a second type of performance measures, the spatial travel efficiencies measures, the evaluation method improved greatly as it was able to systematically identify region pairs with poor transport performance on a city scale.
In the latter this approach includes multiple performance measures to assess the transport network performance, and thus making it an applicable candidate for creating the methodology of **this thesis**. On the other hand, similar to the original approach, the input information was collected in the form of taxi trip data. And although taxi trip data can provide accurate travel information, in general taxis do not behave the same way common commuters do, in a network. Often the travel demand of a commuter taking a taxi differs from one using a private vehicle. Additionally taxi trip data contains large time gaps, which make it hard to determine the exact path the vehicle followed from origin to destination. For this reason the study does have room for improvement.

“Detecting network accessibility”
After the authors completed the previously described research paper, they sought to improve their method further by incorporating the third type of performance measures in their approach (Cui, et al., 2016-B). Within this context the authors proposed an approach which could measure transport network performance based on temporal travel efficiency measures, spatial efficiency measures and accessibility measures. The aim was to evaluate the current urban land use and road network conditions, and to also identify poorly connected (accessible) regions.

The input data for this study consisted of the same taxi trip data used in the study on identifying mismatches between travel demand and transport network services. The results revealed the potential and effectiveness of the proposed method in identifying accessibility problems related to mobility by car. However, it can be argued that the accessibility measures only play a vital role when the transport network is being measured on very large scale.

Ultimately the improved method shows much potential in evaluating transport network performance and it uses multiple performance measures. On the other hand, the main focus of this approach is to determine network performance on the basis of how accessible it is. Unfortunately this does not align with the research objective of **this thesis**. However it does provide more insight on how different measures can complement one another by combining them.

“Assessing transport network reliability and vulnerability”
In (Oliveira, Portugal, & Junior, 2016), authors conducted a study on two road network performance attributes. These include network reliability and vulnerability. For each of the two, indicators or measures where used to develop their approach.

The objective of this study is to compare the network reliability and vulnerability in complex networks to ultimately get more insight on the way they are related to each other. Within the area of transport network reliability the authors used travel time reliability measures. The results revealed no correlation between reliability and congestion. Instead it was discovered that reliability is related to increasing road saturation or level of congestion.

In the latter, the study revealed the nature of travel time reliability. In this way more insight was obtained on how travel time reliability could be applied as a performance measure for the assessment of the transport network performance within the context of **this thesis.**

“Two fluid model”
In (Hong S. , Lee, Bong, & Kho, 2005) researchers sought to analyze the reliability of transportation networks with the two flood model, proposed by Herman and Prigogine 1979. As most reliability studies, this model used travel time reliability measures which included travel time and stop- and running time per unit distance. To obtain values of these measures researchers proposed to build the input data for the two-fluid model using GPS devices. The obtained data was used in the model, after which the results were compared with previous work for validation purposes. The results of this study proved that GPS data could indeed be used to utilize the two fluid model and provide reliable results, when implemented in the

model, for analyzing urban transport network performance. Additional findings also revealed the high level of importance the cut-off speed threshold has within this model.

The findings of the two fluid model showed that the values of the travel time reliability measures could be extracted from GPS data and it revealed a different way of collecting travel information. Here the data was obtained using individual GPS devices. In comparison to taxi trip data, this data could be collected from any type of commuter. Additionally the time gaps will be smaller and the exact path could be identified. This contributes enormously to development of the methodology of **this thesis**, in particular the applicable performance measures and the data collection method.

"Connecting travel time reliability and the Network Fundamental Diagram of Traffic Flow (NFD)"
In 2013 the authors of (Mahmassani, Hou, & Saberi, 2013) presented a research method which could describe network reliability using the Network Fundamental Diagram (NFD) of traffic flow. By linking NFD and travel time reliability performance measures, the authors sought to first establish a bridge between network traffic flow theory and travel time reliability and then to use this bridge to extend travel time variability models from the link and path levels to the network level.

The approach was tested on simulated data and later validated with actual real world data from New York City. Here taxi trip data from one weekday was used to extract trajectory data. The results of this validation were positive and also showed that travel time variability increases with network density and flow rate.

The findings of this paper show that NFD and travel time reliability can be used for the assessment of performance measures to improve the reliability of travel in a network. This reflects the aim of **this thesis** perfectly, which makes the approach an applicable candidate to contribute in the development of the methodology of **this thesis.**

Additionally, this paper provides a macro level analysis of travel time reliability with regards to transport network reliability. However only data from a single day was used in the method. Different days in different months or different seasons could yield different results. Never the less this study gives more insight and contributes a lot to choosing length of the period for which data will be collected for **this thesis**.

3.3.2 Performance assessment input data

3.3.2.1 Identified data type

Within the context of this study various measures were identified in the literature which can be used to evaluate network performance. Each measure is generated with different kind of information. For the temporal measures which include travel time, travel date and type of day, timestamp information is essential. For spatial efficiency measures location information is important. This includes information on the place of origin (start of the trip) and the place of destination (end of the trip). Additionally information on the study area is also necessary. With this information the route directness can be derived. The travel time reliability measures are much similar to the temporal efficiency measures. However the difference is that travel time reliability measures focus on the variation of travel time rather than travel time itself.

The type of data necessary to extract the performance measures can be obtained using different methods. These include traditional methods and modern methods. The traditional methods basically include conducting surveys and interviews. Although these methods cost a lot of time, money and energy, they have proven to be very effective. On the other hand this method does not always deliver the most accurate results. The modern methods focus on digital information from the Global Positioning System (Storey & Holtom, 2003). This type of data can be collected very quickly, in huge amounts and at low cost. However GPS data does not contain all necessary information the performance measures require. For this reason the necessary data should be “extended” GPS data. This can be achieved with imputation of the GPS data. Also it should be noted that the distribution function of travel time for the travel time reliability measures should, ideally, be estimated using historical data. However, due to unavailability or incompleteness of this data the distribution of travel time can be approximated if some of the statistics are known. These include the mean and the standard deviation.

One of the aims of **this thesis** is to develop a method, with which policy makers can analyses performance of a transportation network easily and at low cost. For this reason and in accordance with what the literature revealed, GPS data will be used as the main source of information. The application of GPS data has gained a lot of popularity in recent years, which has made obtaining data from this source type increasingly easier. Within the context of this thesis, GPS data is the optimum choice to generate the performance measures.

3.3.2.2 Data collection methods

GPS data is information collected from the Global Positioning System. This system was created to accurately measure specific positions and time at a specific location on earth. Each position on earth or geographical position can be described in coordinates. The most common choice of coordinates is one which consist of a longitude and a latitude (Vuren & Carey, 2011). In present times GPS is applied in almost every field of technology, but it is most commonly used in navigation systems of vehicles. Here, the output data is mainly used for real time processing.

The advancement of the Global Positioning System (GPS) creates the opportunity to use this technology as a new data collection method to overcome the lack of reliable travel data in transportation research (Castro, Zhang, & Li, 2012). There are multiple ways of collecting GPS data, depending on the quality and amount of data needed. The literature revealed two GPS data collection devices, which include GPS devices mounted in vehicles (cars, busses, taxis, etc.) and standalone GPS devices issued to individual commuters.

Collection from vehicles
In 2011 (Vuren & Carey, 2011) used data from a GPS management system for goods vehicles in his research paper. Here researchers fitted 15,000 different heavy goods vehicles (HGV’s) across England with the specific GPS devices. Similar approaches have been used in recent studies, in which GPS data from taxis were collected. Most taxis are equipped with GPS

devices and make daily trips through entire network. For this reason the collection method provided a relatively cheap solution with very accurate results.

One of the first times this method was implemented was in Singapore in 2014 (Lu & Li, 2014). Two years later the same method was used in China in China, where researchers built a travel demand model based on massive taxi GPS data collected for the city of Harbin (Cui, et al., 2016-A). Generally this collection method has a unique take on measuring transport network performance as it focusses on both spatial and temporal transport efficiencies.

Individual collection
The method of travel data collection has changed a lot in the past years. As (Ortúzar & Willumsen, 2011, p. 55) explains in his paper, this change is due to the major advances in telecommunications. Now general use of personal GPS units offer specific advantages in tracking movement over longer periods of time. Individual GPS collection can be realized in a number of ways. These include the use of special GPS devices issued to individuals.

(Feng & Timmermans, Using Recurrent Spatio-Temporal Profiles in GPS Panel Data for Enhancing Imputation of Activity Type, 2014) issued GPS devices to a large number of respondents. With their help activity-travel diaries for several months could be created and stored. This data contains not only origin- destination information of each individual but also individual GPS traces of each trip. Additionally, with the help of the respondents, the collected data was transformed into complete travel diaries.

From the literature a summary was be made containing the advantages and limitations of using GPS data as a data source, which are relevant to this paper (Vuren & Carey, 2011). The results are presented in Table 3.

Table 3. Advantages and Limitations of using GPS as a data source

GPS data source	
Advantages	**Limitations**
The collection process has no effect on the road network or congestion	GPS data has a limited sample size
During the collection process no permits are needed and there is no need to ask permission or assistance from the police	Depending on the data collection method or device vehicle type could play a role in the uniformity of the data. Different size vehicles move differently through a network
The collection process does not involve any health or safety risks	The accuracy of the origin is highly dependent on when the device is first turned on
Privacy issues by respondents are irrelevant and so the quality of data remains high. Additionally data collection does not involve interviewing respondents. In this way the obtained data is not subject to mood, humor or misunderstanding.	GPS data is comprised of a significant amount of individual location information, and it much virtual memory to be stored. For this reason there are software and hardware requirements to host and analyze the large volumes of data
Data can be collected for long periods of time, depending on the amount of data needed. Additionally recording time-gaps can be set e.g. every other minute or random.	From GPS data is hard to distinguish and when some, and the reason why, breaks or stoppages occur.
Post processing of the data enables non-uniform days in terms of traffic can be removed e.g. local train services not working, poor weather conditions etc.	For GPS data to be applicable additional information has to be inferred. For example, GPS data does not contain any specific information on vehicle type, driver behavior or route choice. This information can be added, if the respondents complete a travel diary.
GPS data provided origin- destination data for the whole network and are not limited to survey locations. Additionally even the traces are recorded easily	As the quality of GPS data is linked to the amount of data that is collected, processing it is very time consuming
Collecting or obtaining GPS data is relatively cheap and almost effortless compared to conventional survey methods	In some countries GPS data is hard to collect and interpret. Mostly because the information on transportation system is out of date.

3.4 Conclusions from the literature study

This part of the chapter describes the findings of the literature study. First all findings are listed and discussed, then an overview on the relevant methodologies is given. Then the best fitting approaches are filtered, based on their applicability and feasibility for this thesis. And at the end of this chapter a suitable approach is formulated.

3.4.1 Overall conclusions

From the literature study it was concluded that transport network performance has been studied various times, by different people and using different measures. Each approach has been able to analyze the performance in one way or the other. The objective of this literature study was to examine previous research and use it as a guide to develop a methodology for the remainder of **this thesis**. Some important and relevant findings include:

1) Identified performance measures include: temporal efficiency measures, spatial efficiency measures, accessibility efficiency measures and travel time reliability measures. In previous research these measures have been used individually and in combination. Previous studies reveals that the performance of a transport networks can be measured on different aspects (Oliveira, Portugal, & Junior, 2016) (Cui, et al., 2016-B). By doing this a better understanding can be derived for how the network works and how it performs. In this sense different measures can also be used in combination with each other.

2) Identified strategies or methods include: evaluating transport network conditions, evaluating transport network reliability and evaluating transport network vulnerability. Previous work (Oliveira, Portugal, & Junior, 2016) proved the possibility of evaluating transport network performance using a fusion of network reliability and network vulnerability. Although vulnerability is not an aspect this thesis aims to stud, the concept of fusing transport network conditions and transport network reliability strategies would be great contribution for forming the methodology of this thesis. At the end of the day both strategies are assessed with the relevant performance measures.

3) The link between evaluating transport network conditions and evaluating transport network reliability is that both strategies evaluate the transport network performance in terms of congestion. By evaluating transport network conditions areas with severe congestion are identified and evaluated, and by evaluating the travel time reliability the level of congestion can be determined and evaluated.

4) The performance measures can be generated with travel information collected with various methods. In this thesis the focus is on using GPS data, which can be collected with GPS devices mounted in vehicles or standalone GPS devices issued to individual commuters.

   In general commercial vehicles (taxis and busses) do not behave the same way normal commuters do in a network, because the travel demand of a commuter taking a taxi or bus often differs from one using a private vehicle. And although taxi trip data can provide accurate travel information it contains large time gaps, which make it hard to determine the exact path the vehicle followed from origin to destination. For this reason input information collected in the form of taxi trip data is not recommended.

3.4.2 Overview methodology and choice

In the literature different approaches were identified. To filter the best fitting approaches they are assessed using two criteria. These include applicability and feasibility. The applicability of each method highly depends on its relevance to the objective of this thesis. Additionally on the aspect of feasibility of the methodology depends on the available data and how it can be utilized. Table 4 lists all the approaches found in the literature and assesses each on the basis of the two criteria. The goal is to filter approaches which contain important features that can be used to form the methodology of this thesis. At the end of the assessment an outline of the methodology is presented.

Table 4. Criteria quality methods

Criteria / **Approach**	**Applicability**	**Feasibility**
A1 by (Lu & Li, 2014) Focus: Network performance by building OD matrix Data: GPS taxi trip information	Score: **Moderate** • Approach based on temporal performance measures • Time based OD matrix • Approach has room for improvement	Score: **Moderate** • Indicators can be generated using GPS data • However the available data through university differs from the data used in this approach
A2 by (Cui, et al., 2016-A) Focus: identifying mismatch between travel demand and transport network services. Data: GPS taxi trip information	Score: **High** • Approach for analyzing transport network performance based on temporal and spatial performance measures	Score: **Moderate** • Indicators can be generated using GPS data • However the available data through university differs from the data used in this approach
A3 by (Cui, et al., 2016-B) Focus: Accessibility as performance attribute Data: GPS Taxi trip information Land use data: activity location information	Score: **High** • Approach for analyzing transport network performance based on temporal, spatial and accessibility performance measures • Accessibility focusses on circuity within the study area and to directly outside of the study area. However this paper only aims on the study area • As an extended method it shows how a previous approach can be extended and optimized	Score: **Moderate** • Indicators can be generated using GPS data • However the available data through university differs from the data used in this approach
A4 by (Oliveira, Portugal, & Junior, 2016) Focus: evaluating network performance based on network reliability and network vulnerability Data: simulated data	Score: **Moderate** • Uses two strategies for performance assessment • Insinuates that network reliability is mostly based on travel time reliability	Score: **Moderate** • Method shows how simulated data can be used for extracting measures
A5 by (Hong S. Y., Lee, Chung, & Kho, 2005) Focus: Application of two fluid model for reliability analysis **Data: GPS test vehicle information**	Score: **High** • Approach for analyzing transport network performance based on travel time performance measures	Score: **High** • Indicators can be generated using GPS data • The available data through university differs slightly from the data used in this approach
A6 by (Mahmassani, Hou, & Saberi, 2013) Focus: Application NFD for reliability analysis Data: Simulated data and GPS taxi trip information	Score: **Moderate** • data alone • Moderate implementation • More for theoretical analysis	Score: **Moderate** • Indicators can be generated using GPS data • However the available data through university differs from the data used in this approach (some indicators e.g. network density hard to extract from GPS)

Based on the assessment of the approaches using the two criteria an outline of the methodology can be formed. This includes:

1) Methodology based on a combination of network reliability and network conditions strategies.
2) The performance measures include temporal, spatial and travel time measures.
3) The data which will be used is individual inferred GPS data obtained from the Technical University of Eindhoven.

Generally the methodology will be an extension of the recent analysis (Cui, et al., 2016-A) involving the performance analysis of the road network of the Chinese city Harbin. Essentially it will complement the previous study by using more accurate data, more accurate calculation methods and by extending the analysis further with a travel time reliability analysis in the very end.

The next chapter focusses on the applied methodology. The chapter starts with a description of the identified problem and the importance of developing a solution for it.

4 Methodology and data

This chapter described the methodology to ultimately answer the research questions. In the literature study different methodologies were presented and finally the most relevant ones were highlighted. In this chapter the applied methodology is divided into different parts, which are described as the research outline. Additionally the reasoning for each step is also included. The chapter begins with the description of the data that was used during the course of this study. The following part contains the research outline. Then the analysis framework is presented. The chapter ends with a set of conclusions regarding the methodology.

4.1 Data description and collection procedure

For this study GPS data, collected by the Technical University of Eindhoven, was used. The data is for the period March until May of the year 2014. This collected data was recorded in the Noord Brabant region of the Netherlands. In 2016 (Feng & Timmermans, Integrated imputation of activity-travel diaries incorporating the measurement of uncertainty measurement of uncertainty, 2016) researched the imputation of activity- travel diaries. For their work they collected GPS data from different years. This data was retrieved and used for this thesis. The next paragraphs describe the entire data collection procedure.

The data used in this study is part of a massive data collection which was held the entire year of 2014. The year was split into four parts and in each part, respondents were recruited to join the survey for three consecutive months. They were each issued unique login information for the web survey webpage. With this information they had access to a personal account, where they could manage their activity-travel data. Each respondent was compensated around 40 euros after the three month participation. The exact amount of the compensation relied heavily on their consistency in validating the diaries and the condition of the GPS devices after three months.

The activity-travel data for day was collected with the help of portable GPS devices, model 747pro, and a web survey system. The devices were issued to a set of respondents, who were instructed to keep them active while commuting. Additionally it was requested to upload the recorded GPS traces using software that was provided to them, and eventually verify and correct the imputed GPS traces with a web survey on a dedicated website.

After the GPS traces are uploaded they are automatically transferred into a sequence of annotated activities and trips using a data processing program. These traces were processed with an embedded processing application called Trace Annotator, which ran in the background of the Web system. The application uses a merge procedure to divide the activities and trips at an aggregated level and an imputation model at the epoch level.

After the data is processed and transformed into imputed daily activity-travel diaries, respondents were requested to validate these diaries, via a prompted recall survey instrument. Important to mention is that prompted recall instruments are not necessarily error free.

Additionally respondents were also given the opportunity to provide some additional information that could not be derived from the GPS traces. The main idea here was to obtain the so-called ground truth data of activities and trips. This includes information such as specific travel party information, trip related financial information and opinions on travel experiences.

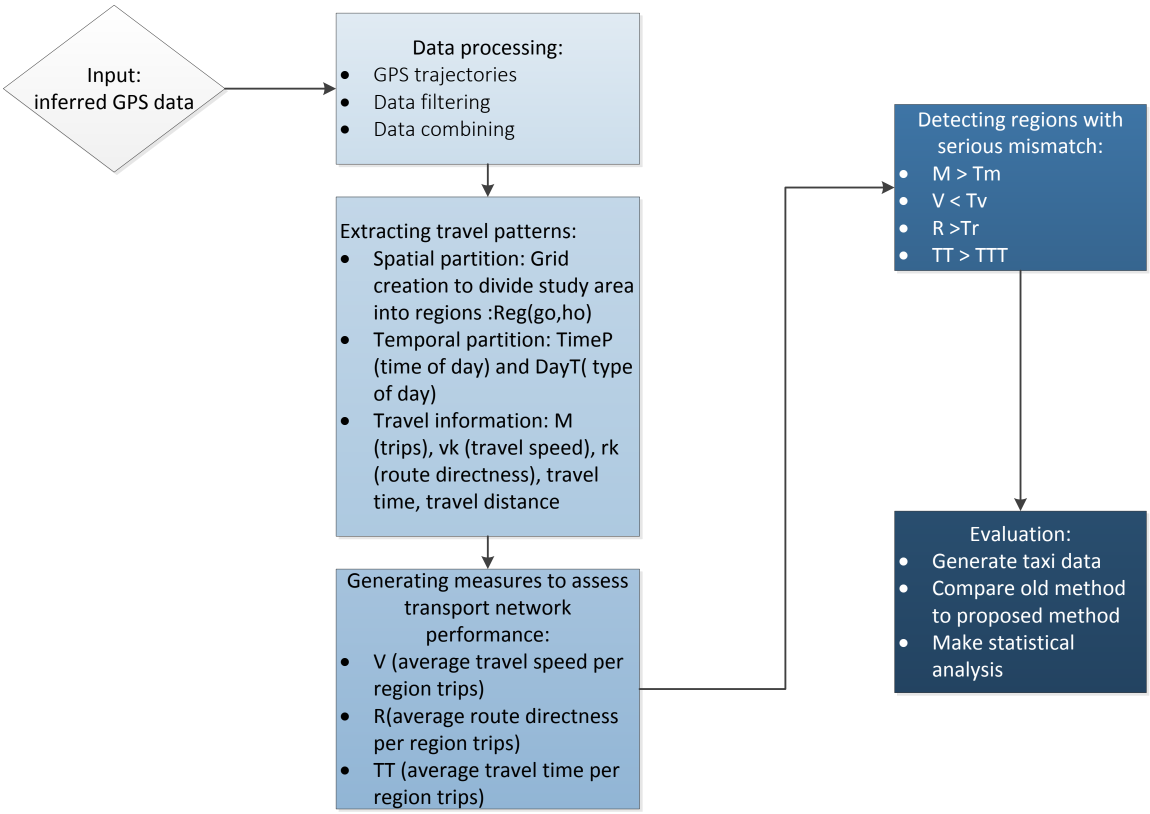
4.2 Analysis procedure

The research method is divided in five main steps. These include:

1) Processing GPS data
2) Modelling region wide travel patterns
3) Generating performance measures
4) Detect regions with serious mismatch problem
5) Method evaluation

Figure 4-1 presents the overall structure of the approach, and is discussed in full detail in the next sub chapter.

Figure 4-1. Analysis procedure



4.3 Analysis framework

4.3.1 GPS Data Processing

Before the data can be used it first needs to be processed. This step is necessary for the utilization of the GPS data for further analysis. This subchapter includes the trajectory description of the data and utilization.

The inferred GPS data was available in three parts, which include journey data, stage data and GPS trace data. The trace data is basically the raw GPS data of each commuter. In this sense a GPS trace contains all location data from a commuter from when it starts recording until it stops recording. In this thesis and these two are represented by the origin and destination and for every individual day during the study period of three months.

A journey is comprised of multiple journey sequences and each journey sequence contains one or more **stages/ trips** (see figure 4-2). Unfortunately the journey data only describes origin and destination of each journey sequence. The trace data on the other hand was split in separate user files. Each user file contained each and every GPS trace of that person during a months' time. A GPS trace is basically one bundle of recorded information, which contains a timestamp, location coordinates, etc. Such a bundle of information is recorded almost every other second. All GPS data was available in CSV file format, which means it could easily be imported in MS Excel or any other type of data base.

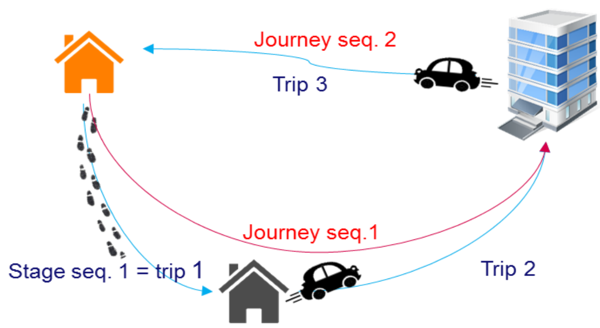


Figure 4-2. Journey sequence vs. stage sequence/ trip

In figure 4-2 a hypothetical journey from home to work is described. The person walks to the house of a co-worker. From there they drive together to work. This describes the first journey sequence and first two stage sequences or trips. After work the co-worker drops the person off at home. This describes the second journey sequence and the third stage sequence (or third trip) Important to mention is that journey sequence is based on destination location and each stage is based on the method of transportation

Figure 4-3 illustrates the journey data. Each person is given an ID, the "persno", with which they could be distinguished. The columns M, N, O and P illustrate the origin and destination (OD) coordinates, respectively. From the figure it can be concluded that the journey data only has OD information of each journey. Information on mode of transportation and each step in the journey are missing. This information was extracted from the stage data and trace data, respectively. The location information in this data type is in the form of a geospatial coordinate set$\boldsymbol{l_i} = \{\boldsymbol{x_i}, \boldsymbol{y_i}\}$, which represent the longitude (lon) and latitude (lat) of a point.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	persno	travday	jourseq	nstages	tleft	tarr	purfrom	purto	minsfrom	minsto	totkm	totmins	lonfrom	latfrom	lonto	latto
2	2397	1-4-2014	1	1	1-4-2014 08:56	1-4-2014 09:06	home	betaald we	0	392	10	2461.87	4.790042	51.60176	4.773091	51.58819
3	2397	1-4-2014	2	1	1-4-2014 15:39	1-4-2014 15:48	betaald we	dagelijke b	392	19	9	674.28	4.773091	51.58819	4.785136	51.58738
4	2397	1-4-2014	3	1	1-4-2014 16:08	1-4-2014 17:19	dagelijke b	dagelijke b	19	32	71	76894.9	4.785136	51.58738	5.133561	52.07459
5	2397	1-4-2014	4	3	1-4-2014 17:51	1-4-2014 18:11	dagelijke b	vrije tijd	32	299	19	11657.42	5.133561	52.07459	5.288088	52.05151
6	2397	1-4-2014	5	1	1-4-2014 23:11	1-4-2014 23:14	vrije tijd	home	299	345	2	225.99	5.288088	52.05151	5.288101	52.05207
7	2397	2-4-2014	1	1	2-4-2014 07:48	2-4-2014 08:36	home	betaald we	168	509	48	61874.59	5.14014	52.04311	4.772703	51.58635

Figure 4-3. Journey data

Figure 4-4 illustrates the stage data. Each day is comprised of journey sequences. And each sequence is comprised of stages. A sequence is based on the travel purpose and travel method of a commuter. If, for example, a commuter travels to work on a certain day by first walking to the train station and then taking the train this first journey sequence is split into two stages. The first stage is characterized by walking as the method of transportation and the second stage by the train as method of transportation. The second journey sequence would be when the commuter would go grocery shopping by bike later that day. All in all the stage data table contains the first piece of the missing information which includes the mode of transportation. Because this thesis focusses on commuters travelling by car, the missing information had to be added to the journey data. By combining the two tables on the basis of the “persno” and “travday” a new table was created with which the individuals, the days on which, and the time at which a car was used could be extracted.

	A	B	C	D	E	F	G
1	persno	travday	jourseq	stageseq	method	travmins	travkm
2	2397	1-4-2014	1	1	EBIKE	10	2461.87
3	2397	1-4-2014	2	1	WALK	9	674.28
4	2397	1-4-2014	3	1	CAR	71	76894.9
5	2397	1-4-2014	4	1	TRAIN	5	8439.89
6	2397	1-4-2014	4	2	CAR	9	2769.56
7	2397	1-4-2014	4	3	WALK	5	447.97

Figure 4-4. Stage data

Figure 4-5 shows the trace data, which includes the timestamp and coordinates of the individual every other second. In this figure the “userid” is the same as “persno” from the previous one. Although the time gaps do vary sometimes, a visualization of the data would show the entire path of each trip for each journey the individual went on each day of the months March, April and May of the year 2014.

By comparing the combined table of journey and stage data with trace data, information could be extracted which contains a person’s ID, travel day, location at each point in time for when a car was used. This data is used for further analysis in the thesis. An impression of the merged table is presented in **chapter 5**.

	A	B	C	D	E	F	G	H
1	userid	routeid	timestamp	x	y	accuracy	speed	heading
2	2397	608152	1-4-2014 08:56	4.789835	51.60246	48	0	0
3	2397	608152	1-4-2014 08:56	4.790042	51.60176	32	0	0
4	2397	608152	1-4-2014 08:56	4.790042	51.60176	24	0	0
5	2397	608152	1-4-2014 08:56	4.789897	51.60185	8	0	0
6	2397	608152	1-4-2014 08:56	4.789897	51.60185	6	0	0
7	2397	608152	1-4-2014 08:56	4.790046	51.6016	4	15.22354	151

Figure 4-5. Trace Data

After the GPS data is utilized, the driving speed at each point, represented by **speed$_i$**, can be calculated by dividing the direct distance of two following points with the time duration of moving between these two points (see formula 1). Calculating this **speed$_i$** is important as it provides the first data filter. Data (cars) with a speed higher 120 km/h are filtered out. The reason for this, is that the speed limit in the Netherlands is 120 km/h. records showing a higher speed are therefore considered as invalid or errors.

$$speed_i = \frac{HD(l_{i+1,}l_i)}{t_{i+1} - t_i}$$

Where,

$HD_{(l_{i+1}+l_i)}$ = direct distance two following points

$t_{i+1} - t_i$ = the time it took to move from the first point to the following one

Formula 1. speed$_i$

The direct distance of two following points can be calculated in numerous ways. In past literature the Euclidean distance was used, because of its easy implementation and straightforwardness (Cui, et al., 2016-A). However, this distance represents the distance of the straight line between two coordinates also known as the “shortest path” distance, and not the real distance. For this reason the real distance is **also** used in this study, which is extracted from a Haversine equation and will be represented as the Haversine Distance (HD). This formula was extracted from (Feng, Moiseeva, & Timmermans, Processing of National Travel Survey GPS Pilot Data, 2011) and is illustrated in formula 2.

$$HD(l_{i+1,}l_i) = 2 \times R \times (\sqrt{(\sin\left(\frac{y_{i+1} - y_i}{2}\right)^2) + cos(y_i) \times \cos(y_{i+1}) \times (\sin\left(\frac{x_{i+1} - x_i}{2}\right)^2)}$$

Where,

R= the radius of the sphere (earth) = 6731 km

$\boldsymbol{l_i} = \{\boldsymbol{x_i}, \boldsymbol{y_i}\}$ = coordinates first point = $\mathrm{l_{origin}} = \{\mathrm{lonfrom}, \mathrm{latfrom}\}$

$\boldsymbol{l_{i+1}} = \{\boldsymbol{x_{i+1}}, \boldsymbol{y_{i+1}}\}$ = coordinates following point = $\mathrm{l_{destination}} = \{\mathrm{lonto}, \mathrm{latto}\}$

Formula 2. Haversine distance

Apart from calculating “HD”, “speedi” and removing data with a higher $speed_i$ than 120 km/h, it is also important to clean the GPS data in this first step of the analysis. Raw GPS data usually contains random errors. These errors mostly occur when less than four satellites are in range of the GPS device, for sending signals, while the data is being recorded. It is important that all data is compatible for further analysis. In the latter good quality data will give the best and most accurate results. Within this context the data is cleaned by removing data with latitude and/ or longitude zero.

4.3.2 Extracting travel patterns

For the construction of the travel pattern model the trip characteristics travel speed and route directness are needed. After this the study area is divided into smaller study cells and each day is divided and categorized. By doing this each part of the model is labelled, which makes it easy to focus on specific segments. Within this context chapter 4.3.2 starts with determining the trip indicators, followed by spatial partition of the study area and temporal partition of the trips in accordance with the study period.

4.3.2.1 Determining trip indicators

In order to analyze passenger travel patterns two indicators/ characteristics are needed. The first one describes the actual travel speed. This is calculated by dividing the actual distance of the journey by the travel time (See formula 3).

$$v_k = \frac{Link(l_o, l_d)}{(t_d - t_o)}$$

Where,

v_k= travel speed each trip k

$(t_d - t_o)$ = travel time origin to destination

$Link(l_o, l_d) = \sum HD(l_{i+1,} l_i)$ represents sum of the actual travel distance of each all trips, and it is estimated with the sum of the Haversine distance between each two following points along the path.

Formula 3. Average speed at each point

The second indicator represents the route directness or circuity. The circuity or route directness of travel paths, describe the ratio between shortest (the straight line) distance between a pair of origin and destination in the network and the Euclidean distance of these two location. The equation of route directness gives a clear visualization of this description (see formula 4).

$$r_k = \frac{Link(l_o, l_d)}{ED(l_o, l_d)} = \frac{\sum ED(l_{i+1,} l_i)}{ED(l_o, l_d)}$$

Where,

r_k[1]= route directness each trip k

$ED(l_o, l_d)$ = represents the Euclidean distance between the origin and destination positions

Formula 4. Route directness at each point

$$ED(l_{i+1,} l_i) = \sqrt{(y_{i+1} - y_i)^2 + (x_{i+1} - x_i)^2}$$

Where,

$\boldsymbol{l_i} = \{\boldsymbol{x_i}, \boldsymbol{y_i}\}$ = coordinates first point = $\mathrm{l_{origin}} = \{\mathrm{lonfrom, latfrom}\}$

$\boldsymbol{l_{i+1}} = \{\boldsymbol{x_{i+1}}, \boldsymbol{y_{i+1}}\}$ = coordinates following point = $\mathrm{l_{destination}} = \{\mathrm{lonto, latto}\}$

Formula 5. Euclidean Distance

[1] A high value for r_k indicates that a certain region is poorly connected. Commuters will have to takes detours when travelling.

These indicators are very important as they are used in the last step of travel pattern modelling to generate two important performance measures. These include the average speed per trip, represented by ‘V”, and the average route directness per trip, represented by “R”. These measures are described in the following parts of this chapter.

4.3.2.2 Spatial partition and temporal partition

Spatial partition

Using a grid-based method the study area is divided into smaller areas (Liu, Gong, Gong, & Liu, 2015). These areas are called **regions** and each region is represented with a **cell ID or region ID**, which more or less consists of letters and numbers e.g. “**CD231”.** The Grid is generated on **MapInfo** as a layer and each region is set at 1 km^2 in size. This layer is set on top of another layer, which illustrates the study area: **Noord Brabant**. By importing the GPS data in MapInfo each coordinate aligns with a cell or region of the grid. The region ID can then be extracted and added to the information of the specific GPS point. After this each location can also be represented with its region ID. By doing this the followed path of each person can be followed by region. This part of the analysis method is fundamental for detecting the regions with serious problems as the regions where user experience certain problems can be identified by their region ID.

Temporal partition

Apart from the spatial partition, each trip is also divided into different time periods of a day. This is represented by **TimeP**. In the analysis a certain time period for **TimeP** is chosen in which traffic congestion commonly occurs. Additionally, as suggested by the literature (Cui, et al., 2016-A), the type of day is also distinguished. Previous studies revealed that travel behavior and travel patterns generally differ across various time periods of a day and also across different type of days. Travel demand during the week differs from travel demand on weekends and holidays. The different type of day includes a weekend-day, weekday or holiday, and represented with **DayT**. Each **DayT** can be extracted from any calendar of the year 2014 **(see appendix 1)**.

4.3.2.3 Model construction

After the two partitions the start times of each trip and the origin and destination locations are estimated in the study area and corresponding time periods according to their region ID. In this way a commuter travel pattern model is described as: OD (Reg (x_o, y_o), Reg (x_d, y_d), TimeP, Day, DayT). Each model represents all the trips that originate from region Reg (x_o, y_o) and end in region Reg (x_d, y_d). With this matrix two transport performance measures are derived. These include:

1) The average travel speed per trip, represented by **V** (see formula 5)
2) The average route directness per trip, represented by **R** (see formula 6)

Table 5. Performance indicators

Measure	Poor performance	Good performance
Speed (V)	Low value	High value
Route directness (R)	High value	Low value
Travel time (TT)	High value	Low value

Table 6. Network performance based on combinations V and R

		R	
		High	**Low**
V	**High**	Region is connected to long high speed detours. Commuters have to take a detour through a highway. The longer travel time and greater distance indicate poor network performance.	Region is well connected and commuters can travel at normal to high speed. This indicates the best network performance
	Low	Commuters have to take detours when travelling, and still suffer heavy congestion along these detours (slow speeds). This indicates the worst network performance.	Region is well connected with direct routes, but the travel demand exceeds the capacity of these routes and commuters have no choice but to take these routes

4.3.2.3.1 Generating measures

V and R are performance measures for spatial and temporal travel efficiencies. Additionally the ratio of both, $\frac{V}{R}$, could be used as a third measure. This measure assesses transport performance by taking both temporal and spatial travel efficiency measures at the same time. Within this context a low value for this measure would mean low travel speed and poor connection between two locations. However due to the fact that it is not used to detect problem regions it is left out of the analysis.

The following formulas were used for generating the measures:

$$\boldsymbol{V = V(Reg(x, y_o), Reg(x_d, y_d), TimeP, Day, DayT) = \frac{\sum_{k=1}^{M} v_k}{M}}$$

Where,

V= Average travel speed over all trips

k= 1 = for every single trip separately

M= total number of trips to and from a certain region

$\boldsymbol{Reg(x_o, y_o), Reg(x_d, y_d)}$ = the origin and destination within the region

$\boldsymbol{TimeP}$ = time of the day

$\boldsymbol{Day}$ = date

$\boldsymbol{DayT}$ = type of day (weekend day, weekday or holiday)

Formula 6. Average speed over all trips to and from a specific region

$$\boldsymbol{R = R(Reg(x_o, y_o), Reg(x_d, y_d), TimeP, Day, DayT) = \frac{\sum_{k=1}^{M} r_k}{M}}$$

Where,

R = Average route directness over all trips
r_k = route directness (circuity)

Formula 7. Average route directness over all trips to or from a certain region

4.3.2.3.2 Travel time reliability

The last performance measure is the travel time reliability. For travel time reliability, ideally, the distribution function of travel time should be estimated using historical data. However, due to the fact that GPS data was used, the travel times were easily extracted. Travel time is measured in (travel) time per unit travelled distance. For further analysis the travel time is represented by **TT** and is calculated with formula 7.

$$TT = \frac{\boldsymbol{t_d - t_o}}{\boldsymbol{HD(l_o, l_d)}}$$

Where,

TT= travel time per unit travelled distance between origin and destination

Formula 8. Travel time reliability measure

4.3.3 Detecting problem regions

To detect the daily transport problems, the performance measures V, R and TT are tested against threshold values. The first step in doing so includes identifying regions with higher volumes of traffic demand. In this sense the amount of trips (M) to and from a certain region need to be higher than a certain amount. This amount is represented by T_M.

A higher volume of travel demand in the morning, represents regions in residential or employment areas, which make them more likely candidates for traffic congestion. Additionally the large number of passenger trips increases the accuracy of the estimated parameters V, R and TT. In this way these parameters can better represent general traffic conditions.

After the regions with high travel demand are detected V and R are used to measure the temporal and spatial travel efficiencies within these regions. Here, regions with R higher than another threshold (T_R) or V lower than a threshold (T_V), are defined as problematic regions with temporal or spatial travel efficiency problems. Additionally TT is also implemented to measure travel time reliability within the regions with high volume travel demand. The regions with a higher travel time per unit distance than a threshold T_{TT} are also defined as problematic regions. The regions with V lower than T_V, R higher than T_R and TT higher than T_{TT}, represent regions experiencing the worst problems. Within these regions commuters are forced to take detours for travel, while still moving at slow speed due to heavy traffic along these detours and taking a significantly longer time to reach their destination due to a high level of congestion (Cui, et al., 2016-A).

The indicators used to identify problem regions include:
1) VM: $V < T_V$ and $M > T_M$ (temporal performance measure)
2) RM: $R > T_R$ and $M > T_M$ (spatial performance measure)
3) TTM: $TT < T_{TT}$ and $M > T_M$ (travel time reliability performance measure)

The worst regions are described as regions experiencing all three problems at once. Within these regions commuters are forced to take detours, where they also experience high level of traffic congestion. These regions are identified with:
4) VRM: $V < T_V$ and $R > T_R$ and $M > T_M$ (temporal and spatial performance measures)
5) VRTTM: $V < T_V$ and $R > T_R$ and $TT < T_{TT}$ and $M > T_M$ (temporal, spatial and travel time reliability performance measures)

4.3.4 Method validation

The proposed approach is validated comparing the results with results derived using an approach from the literature. The method validation consists of four steps, which include:
1) Generate taxi trip data
2) Implement both approaches
3) Make statistical analysis
4) Compare results

For the method validation only a fraction of the entire data set is used. Here a random week in one of the three months is chosen. Only this week will be used for further analysis.

4.3.4.1 Generate taxi data

As explained at the end of the literature review, the approach in this paper is an improved and extended version of the method created by (Cui, et al., 2016-A). For this reason, the most effective validation is to compare the improved and extended approach proposed in this paper to the original one from literature. The three key differences between both approaches are presented in the table below

Table 7. Differences in approaches

Difference	Approach Thesis	Approach literature
Input data	Individual commuter data	Taxi trip data
Performance measures	1. Speed 2. Route directness 3. Travel time reliability	1. Speed 2. Route directness

To recreate the method from literature, first taxi trip data needs to be generated from the original raw GPS data. From literature it was revealed that taxi trip data includes large time gaps between recorded locations. For this reason the large portions of the raw data were removed, creating a data set that looked very similar to actual taxi trip data.

4.3.4.2 Identify most problematic regions using both approaches

After creating the taxi data, both the approach from the literature and the approach proposed in this paper are applied using the test data. After all measures are generated for both approaches the problem regions are detected. For the validation not only the empirical threshold values are used to detect the regions, but a variation of these values. By doing this the quality of each method can be estimated using a statistical analysis on the results.

The variation of the threshold values is based on their empirical values found in the literature (Cui, et al., 2016-A), which include:

- T_V = 20 km/h
- T_R = 1.5
- T_M = 20 trips
- T_{TT} => Compared to the previous threshold values, this value was not directly derived from literature, but is estimated using a statistical method. The literature suggest using the 95th percentile method, which estimates the 95% probability of the worst travel time that commuters may experience once per month, namely one of 20 week days. The value for this threshold is estimated in **chapter 5**.

After each method is applied, with each threshold a number of problem regions is detected. With these numbers a descriptive analysis is performed using SPSS to finally extract the **standard deviation** and **the mean.** With these variables **the coefficient of variation** can be determined.

The coefficient of variation (C_V) is used to measure the consistency of the results across all validation experiments. The C_V is calculated by dividing the standard deviation by the mean value. A high C_V value reflects inconsistency among the results within the group of a specific validation method. In turn a lower C_V value reflects a higher consistency in the results and thus indicate the most reliable method which yields the most accurate and reliable results. In other words, by determining the coefficient of variation for each method, the approach producing the best quality result is revealed.

$$c_v = \frac{\sigma}{\mu}$$

Where,

C_V = The coefficient of variation

σ = The standard deviation

μ = The mean

Formula 9. The coefficient of variation

4.4 Conclusions regarding the methodology and data

The analysis starts by utilizing and cleaning the raw GPS data. The available data consisted of three parts, which include journey data, stage data and GPS trace data. Each part contains important information for the analysis method and therefore were combined to one data table. Generally GPS data contains random errors due, which needed to be filtered out before continuing. Within this context data with longitude and latitude zero were removed. Additionally all data with transportation method other than car were filtered out. The actual distance and speed are calculated and as a very last filter all data with speed higher than 120 km/h was removed. These commuters do not experience any serious problems in the transport network and are therefore left out of the scope of this thesis.

Once the data is utilized trip characteristics can be extracted. These include the distance from origin to destination, the distance from each recorded location to the following recorded location and of course the speed and route directness between each recorded locations.

After the characteristics are extracted the performance measures can be generated by applying spatial and temporal travel partition for modelling commuter travel patterns. Spatial partition includes applying a grid based method which splits the study area into smaller study regions. Temporal partition includes making each trip subject trips to the time of day and the type of day, e.g. weekday or weekend day. After this a travel pattern model is created with which the performance measures V, R and TT can be generated.

With the generated measures the problem regions can easily be detected by comparing the measures to their respective threshold values. Within this context the regions where commuters experience problems, are identified by comparing the generated measures to their respective threshold values. The thresholds for V, R, M and TT are defined as T_V, T_R, T_M and T_{TT}, respectively. The values of these parameters are empirical values found in the literature (Cui, et al., 2016-A). Here it was revealed that V lower than its threshold value or R higher than T_R, are identified as problematic regions with temporal or spatial travel efficiency problems. Additionally a TT higher than T_{TT} would also describe a problematic region. The three problems are detected with three indicators based on the three performance measures. For example, any speed lower than a threshold of 20 km/h indicates that the capacity of the routes within this region may not be sufficient to meet the current travel demand. A lower speed could also imply that people have no other choices than to take these inefficient routes for traveling between the two regions (Cui, et al., 2016-B). Within this context, by comparing each of the three measures, which include V, R and TT, with a threshold value regions experiencing these problems are detected. Additionally regions with serious congestion problems are detected by combining the three different measures into one, this forms the problem indicator: $V < T_V$ and $R > T_R$ and $TT < T_{TT}$ and $M > T_M$ (combination of temporal, spatial and travel time reliability performance measures).

The evaluation of the methodology consists of a comparison of the proposed approach in this thesis with the approach found in literature (Cui, et al., 2016-A). Here taxi data is generated from a fraction, approximately one week data, of the raw GPS data which was used in the approach of this paper. After the measures are generated for both approaches, the problem regions are identified using a variation of threshold values. In this way each approach will identify a variety of regions with serious problems. The results are then statistically analyzed, after which the quality of each approach can be derived.

5 Results

Chapter 5 presents the results of the approach applied in this study. This chapter is structured similar to the previous one. Here each of the steps described in the methodology is followed and the results are presented. Additionally at the end of each part the results are interpreted. The chapter begins with the data processing step where a description of the distribution of transportation methods used during the study period March 2014 until May 2014 is given.

5.1 Data processing

The data processing step is divided into two parts, which include a description of the raw data, specifically the distribution of the transport methods in the network, and the utilization of this data. Here the results are not only shown, but each process is also explained. This includes the used software tools and the reasoning.

5.1.1 Description raw data: Distribution transport network

The data used for the analysis is collected for three months. For each month three tables were available. This included a journey, stage and trace data table. In this study the focus is on the transportation mode "car". To understand to what extent focusing on the transport method "car" will yield enough data for good quality results, it was first necessary to investigate the distribution of the different modes of transportation in the network.

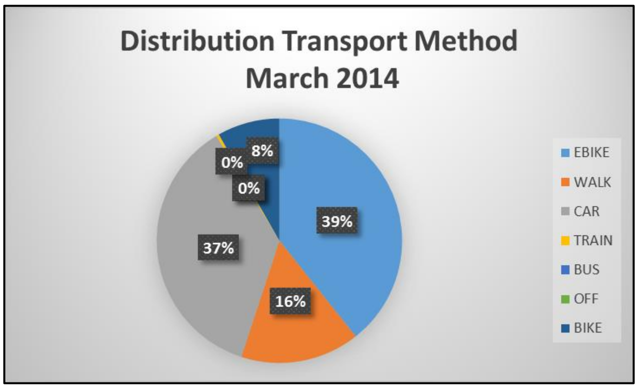


Figure 5-1. Distribution transport method: percentages

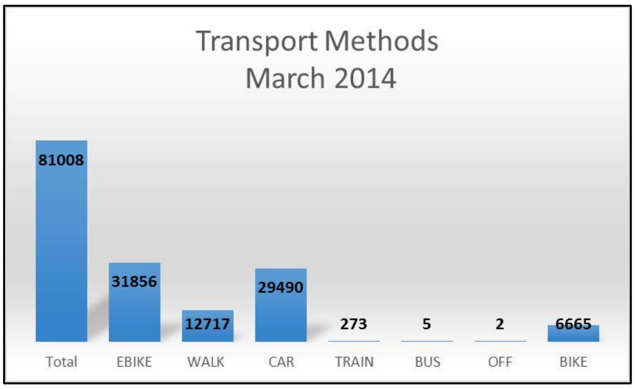


Figure 5-2. Distribution transport method: numbers

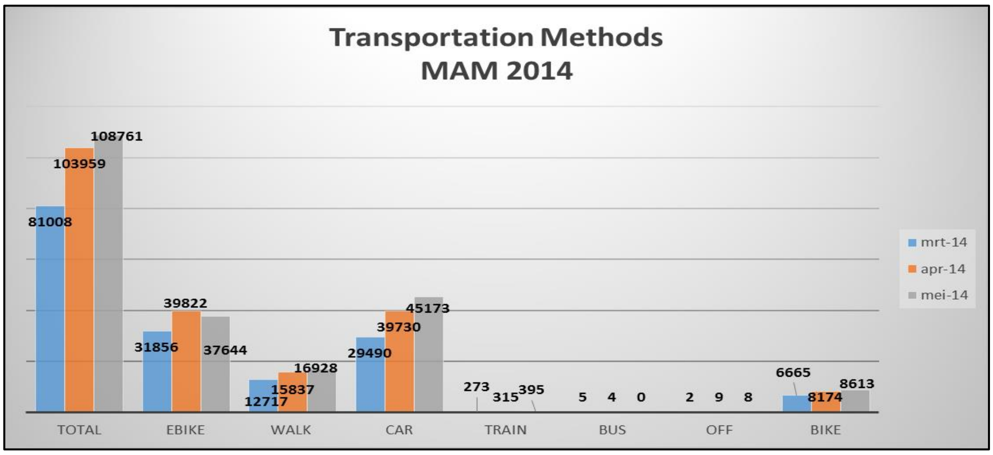


Figure 5-2. Distribution transport methods March April May 2014: numbers

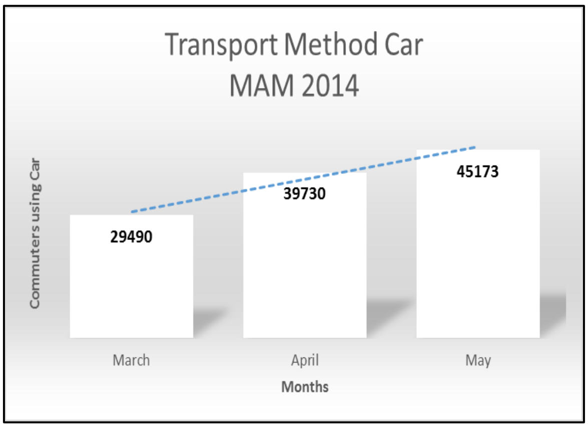


Figure 5-3. Transport method focus: Car March April May 2014

The figures 5-1 until 5-4 illustrate the distribution of transportation methods used in the three month study period. From the figures 5-3 and 5-4 it can be seen that car usage accounts for almost half of the used transportation methods. It can also be seen that the number of commuters using cars is increasing every month. Of course it should be pointed out that the data is generated from a standard set of commuters who most likely do not use public transportation. However, this has no implication on further analysis as the thesis focusses on commuters using a car for transportation. And thus concluding that there is enough "car" data to produce accurate results.

5.1.2 Utilization GPS data

For the very first step in utilizing the GPS data the software MS Excel and MS SQL were used. Here the journey data and stage data were merged into one "journey-stage" data table, using MS Excel Query. The two tables were joined based on the person ID "persno", the date "travday" and journey sequence "jourseq". After this all records containing "Car" as method of transportation were filtered. The filtered table was then exported to CV file format and imported as a table in MS SQL database.

In the second part, each of trace data tables recorded from individual commuters was merged into one enormous trace data table. This table contains the trace information of every commuter per month. The tables were merged using a bulk fuse code for CSV files (see figure 5-5). After the combined file was imported as a table in MS SQL the excess string data, which was a result of combining the CSV files, was deleted. Then the data type of the different columns was enhanced (altered) so the journey stage data table and the trace data table could be compared to each other, and eventually merged into one much larger data table which was used for further analysis.

```
copy *.csv bulkfusetables.csv
```

Figure 5-4. Bulk fuse code CSV

The raw trace data consist of separate tables for each commuter, as illustrated in figure 5-6. As mentioned before, these individual tables, were fused into one enormous table. By doing this all data could be analyzed at the same time. Figures 5-7 illustrates a part of the merged tables.

OSDisk (C:) ▸ Users ▸ s141445 ▸ Documents ▸ TUe ▸ S3Q4 ▸ Data ▸ Data- Usage ▸ 1

Name	Date modified	Type	Size
2397.csv	12-11-2014 09:33	Microsoft Excel Co...	398 KB
2401.csv	12-11-2014 09:33	Microsoft Excel Co...	224 KB
2409.csv	12-11-2014 09:33	Microsoft Excel Co...	1,690 KB
2411.csv	12-11-2014 09:33	Microsoft Excel Co...	1,115 KB
2413.csv	12-11-2014 09:33	Microsoft Excel Co...	2,008 KB
2420.csv	12-11-2014 09:33	Microsoft Excel Co...	1,389 KB
2421.csv	12-11-2014 09:33	Microsoft Excel Co...	562 KB
2423.csv	12-11-2014 09:33	Microsoft Excel Co...	518 KB
2424.csv	12-11-2014 09:33	Microsoft Excel Co...	300 KB
2426.csv	12-11-2014 09:33	Microsoft Excel Co...	741 KB
2427.csv	12-11-2014 09:33	Microsoft Excel Co...	260 KB
2428.csv	12-11-2014 09:33	Microsoft Excel Co...	438 KB
2429.csv	12-11-2014 09:33	Microsoft Excel Co...	539 KB
2433.csv	12-11-2014 09:33	Microsoft Excel Co...	846 KB
2435.csv	12-11-2014 09:33	Microsoft Excel Co...	884 KB
2443.csv	12-11-2014 09:33	Microsoft Excel Co...	1,137 KB
2444.csv	12-11-2014 09:33	Microsoft Excel Co...	145 KB
2449.csv	12-11-2014 09:33	Microsoft Excel Co...	209 KB
2451.csv	12-11-2014 09:33	Microsoft Excel Co...	891 KB
2452.csv	12-11-2014 09:33	Microsoft Excel Co...	842 KB

	A	B	C	D	E	F	G	H
1	userid	routeid	timestamp	x	y	accuracy	speed	heading
2	2397	495744	3-3-2014 14:59	4.79121	51.60021	108	0	0
3	2397	495744	3-3-2014 14:59	4.790953	51.5991	24	0	0
4	2397	495744	3-3-2014 14:59	4.791055	51.59888	8	20.66129	0
5	2397	495744	3-3-2014 14:59	4.79116	51.59869	6	20.35158	0
6	2397	495744	3-3-2014 15:00	4.791496	51.59787	4	16.13826	169

4386	2397	601344	30-3-2014 18:19	4.790262	51.60708	8	15.68175	195
4387	2397	601344	30-3-2014 18:19	4.790186	51.60691	6	16.10616	190
4388	2397	601344	30-3-2014 18:19	4.788736	51.6059	8	13.38043	188
4389	2397	601344	30-3-2014 18:19	4.788695	51.60581	6	13.20755	194
4390	2397	601344	30-3-2014 18:20	4.788471	51.60445	6	15.44296	173
4391								
4392								

Figure 5-5. Bulk fuse code CSV

4388	2397	601344	30-3-2014 18:19	4.788736	51.6059	8	13.38043	188
4389	2397	601344	30-3-2014 18:19	4.788695	51.60581	6	13.20755	194
4390	2397	601344	30-3-2014 18:20	4.788471	51.60445	6	15.44296	173
4391	2401	568595	22-3-2014 16:41	4.95775	51.62631	61	0	0
4392	2401	568595	22-3-2014 16:41	4.954737	51.62911	10	44.13528	351
4393	2401	568595	22-3-2014 16:41	4.953506	51.63026	10	45.93672	309
4394	2401	568595	22-3-2014 16:41	4.951603	51.63108	10	37.60506	303

6472	2401	601475	30-3-2014 19:07	4.960485	51.62365	10	0	294
6473	2401	601475	30-3-2014 19:11	4.96011	51.62331	5	0	132
6474	2409	494073	2-3-2014 21:23	4.783601	51.58477	65	0	-1
6475	2409	494073	2-3-2014 21:23	4.783728	51.5849	10	0	-1
6476	2409	494073	2-3-2014 21:23	4.784383	51.58529	5	21.132	31
6477	2409	494073	2-3-2014 21:24	4.785554	51.58646	5	20.34	32

Figure 5-6. Trace data after fusion

5.1.3 Data cleaning and filtering

Before the journey-stage data and the trace data are merged, journey- stage data table is cleaned and filtered. This table was used as a lookup table to extract trace data from the trace data table, for this reason it was important that this lookup table (journey-stage data table) did not have any errors.

1) **First all data with longitude and/ or latitude zero where removed.** For the March, April and May the amount of 69931, 2257, 3730 records were removed. (see table 8)

Table 8. Records after filter 1

Month	Records original	Records removed	New total
March 2014	81009	69931	11078
April 2014	103959	2257	101703
May 2014	108761	3730	105031

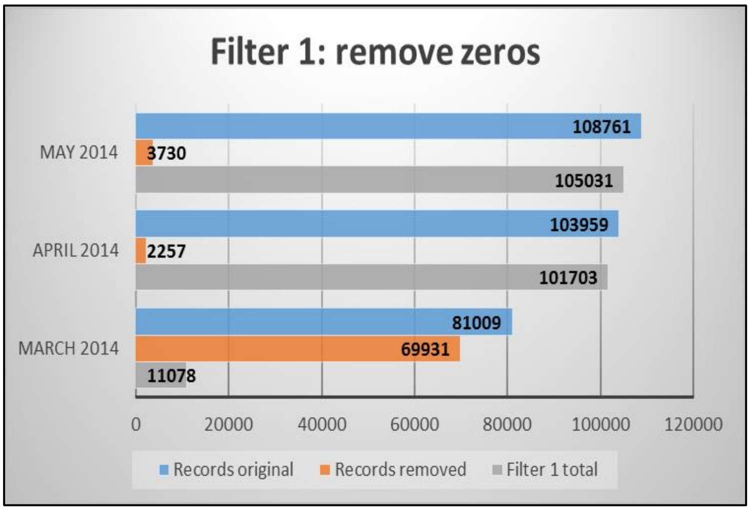


Figure 5-7. Records after filter 1

Figure 5-8 presents a visual representation of the records before and after the first filter, which included the removal of all records with a longitude/ latitude zero. As can be seen from the chart, for the month March 2014 approximately 85% of the records were removed. Fortunately this had no implication on the study as the table had yet to be merged with the individual trace data.

2) **After the first filter the records with transportation method other than "car" were removed.** For the months of March, April and May 2014 the amount of 6971, 63051, 61509 records were removed, respectfully. See table below.

Table 9. Records after filter 2

Month	Records after first filter	Records removed	New total
March 2014	11078	6971	4107
April 2014	101703	63051	38652
May 2014	105031	61509	43522

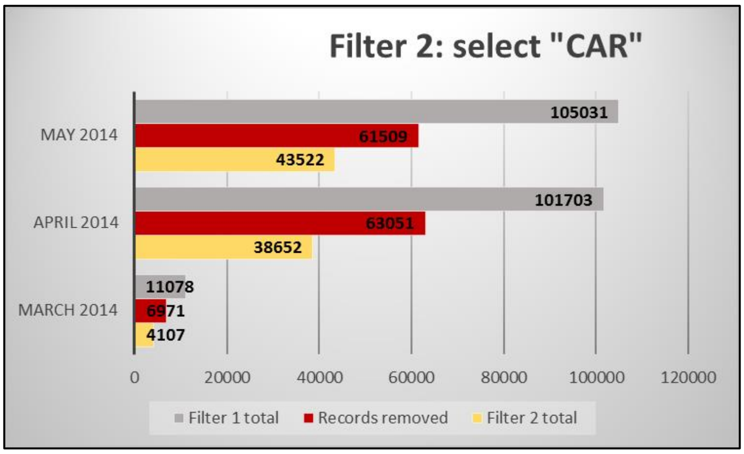


Figure 5-8. Records after filter 2

As previously presented the transportation method car counts for almost half of the transport methods used in the network. For this reason a little more than 50% of the records (after filter 1) were removed. See figure 5-9.

After filter 2 the table was converted to CSV file format and imported to MS SQL.

Figures 5-10 and 5-11 present the table before and after filter 1 and 2.

persno	travday	jours	nstag	stages	method	travmins	travm	tleft	tarr	purfrom	purto	minsf	minsto	totmins	totm	lonfrom	latfrom	lonto	latto
2397	3-3-2014	1	1	1	BIKE	4.0	978.99	2014-03-03 15:05:32.00	2014-03-03 15:09:59.000	home	sociaal	5	143	4	978.99	4.79152	51.59651	4.7729195	51.588
2397	3-3-2014	2	1	1	WALK	16.0	825.71	2014-03-03 17:33:55.00	2014-03-03 17:50:04.000	sociaal	recreatie(	143	5	16	825.71	4.77292	51.588	4.779610438	51.59516
2397	3-3-2014	3	2	1	CAR	2.0	322.13	2014-03-03 17:55:07.00	2014-03-03 17:59:50.000	recreatie(park wandeler	recreatie(	5	5	4	6337.96	4.77961	51.59516	5.0659456	51.56657
2397	3-3-2014	3	2	2	TRAIN	2.0	6015.83	2014-03-03 17:55:07.00	2014-03-03 17:59:50.000	recreatie(park wandeler	recreatie(	5	5	4	6337.96	4.77961	51.59516	5.0659456	51.56657
2397	3-3-2014	4	1	1	WALK	11.0	741.03	2014-03-03 18:05:00.00	2014-03-03 18:16:00.000	recreatie(park wandeler	sociaal	5	252	11	741.03	5.06595	51.56657	5.08873102	51.5586
2397	3-3-2014	5	1	1	WALK	5.0	444.97	2014-03-03 22:28:19.00	2014-03-03 22:33:51.000	sociaal	recreatie(	252	16	5	444.97	5.08873	51.5586	4.780350447	51.59544
2397	3-3-2014	6	1	1	WALK	17.0	1466.67	2014-03-03 22:50:00.00	2014-03-03 23:07:34.000	recreatie(park wandeler	home	16	352	17	1466.67	4.78035	51.59544	4.78866373	51.60405
2397	4-3-2014	1	1	1	EBIKE	10.0	2642.58	2014-03-04 08:50:42.00	2014-03-04 09:01:21.000	home	betaald w	230	524	10	2642.58	4.79016	51.60183	4.784436122	51.59032
2397	4-3-2014	2	1	1	EBIKE	7.0	2089.14	2014-03-04 17:46:14.00	2014-03-04 17:53:26.000	betaald werk	home	524	666	7	2089.14	4.78444	51.59032	4.78859279	51.60369
2397	5-3-2014	1	1	1	BIKE	7.0	1436.94	2014-03-05 08:54:09.00	2014-03-05 09:01:22.000	home	betaald w	234	547	7	1436.94	4.78976	51.5937	4.771866196	51.58621
2397	5-3-2014	2	1	1	EBIKE	7.0	1946.55	2014-03-05 18:08:38.00	2014-03-05 18:16:05.000	betaald werk	home	547	643	7	1946.55	4.77187	51.58621	4.788521399	51.60405
2397	6-3-2014	1	1	1	EBIKE	9.0	2615.09	2014-03-06 13:10:37.00	2014-03-06 13:19:49.000	home	sociaal	490	323	9	2615.09	4.78852	51.60405	4.792006969	51.60774
2397	6-3-2014	2	1	1	EBIKE	19.0	5267.68	2014-03-06 18:43:26.00	2014-03-06 19:03:20.000	sociaal	sociaal	323	153	19	5267.68	4.79201	51.60774	4.852069093	51.6301
2397	6-3-2014	3	1	1	BIKE	21.0	5187.29	2014-03-06 21:36:50.00	2014-03-06 21:58:21.000	sociaal	home	153	421	21	5187.29	4.85207	51.6301	4.788563195	51.60436
2397	7-3-2014	1	1	1	WALK	20.0	1466.94	2014-03-07 08:30:03.00	2014-03-07 08:50:10.000	home	betaald w	210	22	20	1466.94	4.78774	51.6029	4.779218486	51.59548
2397	7-3-2014	2	2	1	CAR	28.0	45279.23	2014-03-07 09:12:30.00	2014-03-07 09:48:34.000	betaald werk	dagelijke l	22	29	35	61940.19	4.77922	51.59548	4.5682092	52.16997
2397	7-3-2014	2	2	2	TRAIN	7.0	16660.96	2014-03-07 09:12:30.00	2014-03-07 09:48:34.000	betaald werk	dagelijke l	22	29	35	61940.19	4.77922	51.59548	4.5682092	52.16997

Figure 5-9. Journey-stage before filtering and cleaning

1	persno	travday	jou	nsta	sta	method	travmins	travm	tleft	tarr	purfrom	purto	minsfrom	minsto	totmins	totm	lonfrom	latfrom
48	2397	13-3-2014	3	1	1	CAR	13.0	4423.54	2014-03-13 19:50:41.000	2014-03-13 20:04:13.000	sociaal	recreatie(p	255	85	13	4423.54	4.789098	51.60434
83	2401	23-3-2014	4	1	1	CAR	13.0	4252.34	2014-03-23 19:01:53.000	2014-03-23 19:15:03.000	sociaal	home	69	584	13	4252.34	4.919407	51.62909
144	2409	5-3-2014	2	1	1	CAR	19.0	6248.42	2014-03-05 08:23:18.000	2014-03-05 08:43:17.000	betaald we	betaald we	15	9	19	6248.42	4.65504	51.69928
171	2409	9-3-2014	3	1	1	CAR	7.0	2891.28	2014-03-09 12:19:22.000	2014-03-09 12:26:57.000	niet-dagelij	dagelijke b	60	20	7	2891.28	4.781006	51.58311
197	2409	13-3-2014	5	1	1	CAR	30.0	5651.17	2014-03-13 19:00:43.000	2014-03-13 19:30:45.000	dagelijke bc	home	12	569	30	5651.17	4.780792	51.58293
213	2409	20-3-2014	5	1	1	CAR	10.0	1427.71	2014-03-20 08:25:13.000	2014-03-20 08:35:32.000	betaald we	betaald we	3	4	10	1427.71	4.370018	51.97675
225	2409	22-3-2014	1	1	1	CAR	7.0	2795.77	2014-03-22 08:02:03.000	2014-03-22 08:09:28.000	home	betaald we	182	220	7	2795.77	4.722057	51.60210
251	2409	28-3-2014	5	1	1	CAR	15.0	6110.75	2014-03-28 12:24:18.000	2014-03-28 12:39:39.000	dagelijke bc	dagelijke b	15	20	15	6110.75	4.791098	51.53921
253	2409	28-3-2014	7	1	1	CAR	15.0	5648.75	2014-03-28 16:52:06.000	2014-03-28 17:07:14.000	sociaal	sociaal	223	56	15	5648.75	4.775827	51.58167
260	2409	29-3-2014	3	1	1	CAR	12.0	4892.40	2014-03-29 10:01:59.000	2014-03-29 10:13:59.000	dagelijke bc	dagelijke b	17	56	12	4892.4	4.757111	51.60022
277	2411	6-3-2014	3	1	1	CAR	24.0	9414.05	2014-03-06 17:17:02.000	2014-03-06 17:41:50.000	betaald we	home	264	678	24	9414.05	4.926222	52.38668
285	2411	8-3-2014	1	1	1	CAR	8.0	3138.35	2014-03-08 13:49:26.000	2014-03-08 13:57:36.000	home	dagelijke b	529	21	8	3138.35	5.001638	52.35283
329	2411	18-3-2014	1	1	1	CAR	2.0	559.86	2014-03-18 09:06:37.000	2014-03-18 09:08:52.000	home	sociaal	246	206	2	559.86	4.929961	52.39609
337	2411	21-3-2014	2	1	1	CAR	33.0	9879.44	2014-03-21 14:19:17.000	2014-03-21 14:52:28.000	betaald we	niet-dageli	320	10	33	9879.44	4.927085	52.38634
369	2411	30-3-2014	3	1	1	CAR	8.0	3056.57	2014-03-30 15:10:55.000	2014-03-30 15:19:47.000	dagelijke bc	dagelijke b	144	43	8	3056.57	5.010195	52.35194
418	2413	12-3-2014	5	1	1	CAR	9.0	3478.55	2014-03-12 13:15:26.000	2014-03-12 13:25:03.000	dagelijke bc	vrije tijd	32	70	9	3478.55	5.873575	51.80430
551	2420	13-3-2014	5	1	1	CAR	59.0	7489.66	2014-03-13 16:04:28.000	2014-03-13 17:03:30.000	dagelijke bc	vrije tijd	13	98	59	7489.66	4.960089	51.6235
557	2420	14-3-2014	3	1	1	CAR	6.0	2304.62	2014-03-14 16:42:01.000	2014-03-14 16:48:27.000	dagelijke bc	recreatie(p	38	132	6	2304.62	4.940628	51.63433
562	2420	16-3-2014	1	1	1	CAR	107.0	12561.23	2014-03-16 09:47:31.000	2014-03-16 11:35:27.000	home	dagelijke b	287	35	107	12561.23	4.957629	51.62666
566	2420	17-3-2014	3	1	1	CAR	8.0	3020.93	2014-03-17 13:49:52.000	2014-03-17 13:58:30.000	betaald we	vrije tijd	230	144	8	3020.93	4.786539	51.58730
567	2420	17-3-2014	4	1	1	CAR	8.0	2556.13	2014-03-17 16:22:52.000	2014-03-17 16:31:12.000	vrije tijd	sociaal	144	91	8	2556.13	4.7836	51.59878

Figure 5-10. Journey-stage table after first and second filter

In the very last part of the data utilization, the filtered journey- stage table was merged with the trace data table in MS SQL using a query which identifies the coordinates and timestamps in the trace data containing "car" as mode of transportation. Additionally the time of leaving (journey start) and the arrival time (journey end), from the journey data, were used to locate and display each trace which took place between the two moments in time. After this the filtered and fused table is exported to CSV file format and imported in MS Excel (see figure 5-12).

THis.sql - S141445\...(TUE\s141445 (53))

```sql
select * from [dbo].[March2014]
where [userid] in
(select [persno] from [dbo].[March filter1_correct]
where [timestamp] between [tleft] and [tarr])
order by [userid] asc
```

Figure 5-11. Trace identifier

After the journey-stage-trace data table was created, the Haversine distance, HD (l_i, l_{i+1}), between each two following recorded locations was calculated along with the **speed$_i$**. After this all records with speed$_i$ higher than 120 km/h were removed. These records are above the speed limit in the Netherlands and are most likely errors.

5.2 Extracting travel patterns

After the data is utilized, cleaned and filtered it can be used for further analysis. The next step is to focus on certain aspects of each trip. The first step in doing so is the spatial and temporal partition of the trips.

5.2.1 Spatial partition

The grid was created for the Netherlands using MapInfo Professional and generating a total of 90117 cells or regions with a size of 1 km^2. After this, the coordinates were imported and each GPS point was linked to a region. This was done by joining the grid table with the data tables for March, April and May 2014. As mentioned in the chapter 4, this step is very important because by assigning each coordinate to a region the problematic regions can be identified based on the analysis of the information these coordinates contain. In the end the merged tables were exported to CSV file format and opened in MS Excel for further analysis.

Figure 5-13 illustrates the map of the Netherlands in QGIS to get an idea of the study area.

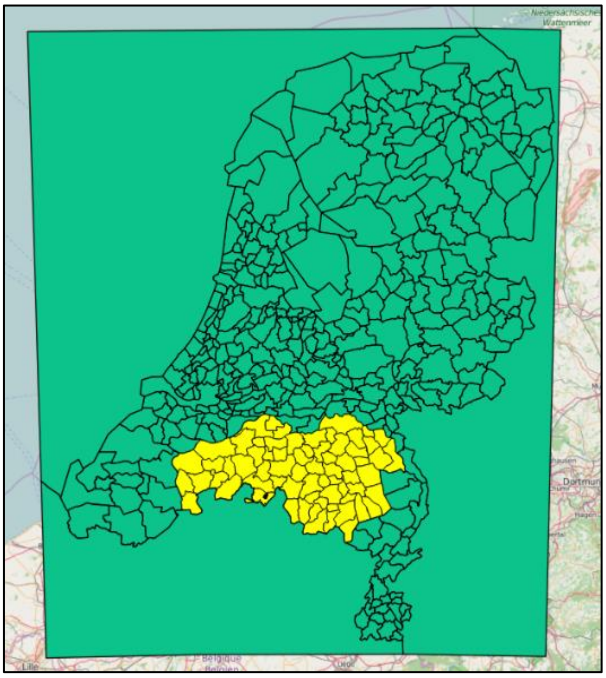


Figure 5-12. Map of the Netherlands (via QGIS)

Figure 5-13. Map of the Netherlands with GRID (via MapInfo)

Figure 5-14 presents the map in MapInfo with the grid with the focus on the study area: Noord Brabant.

The travel patterns of the commuters in each of the three months are presented in figures 5-16 until 5-18.

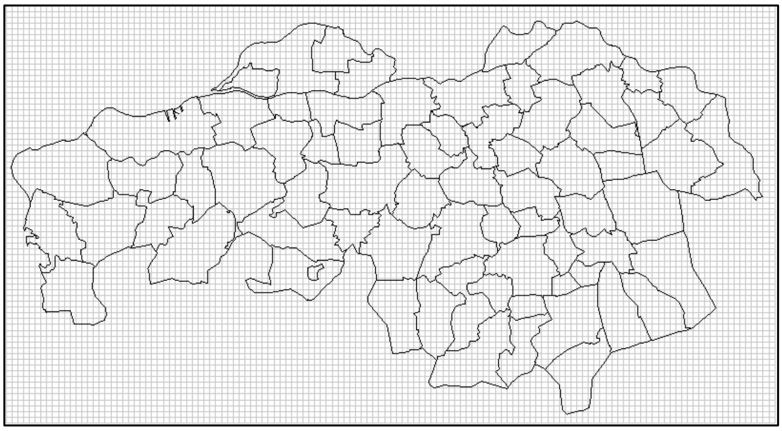


Figure 5-14. Noord-Brabant map with grid

Important to mention is that the part of the grid covering Noord- Brabant, as seen in figure 15-5, counts 9432 cells/ regions.

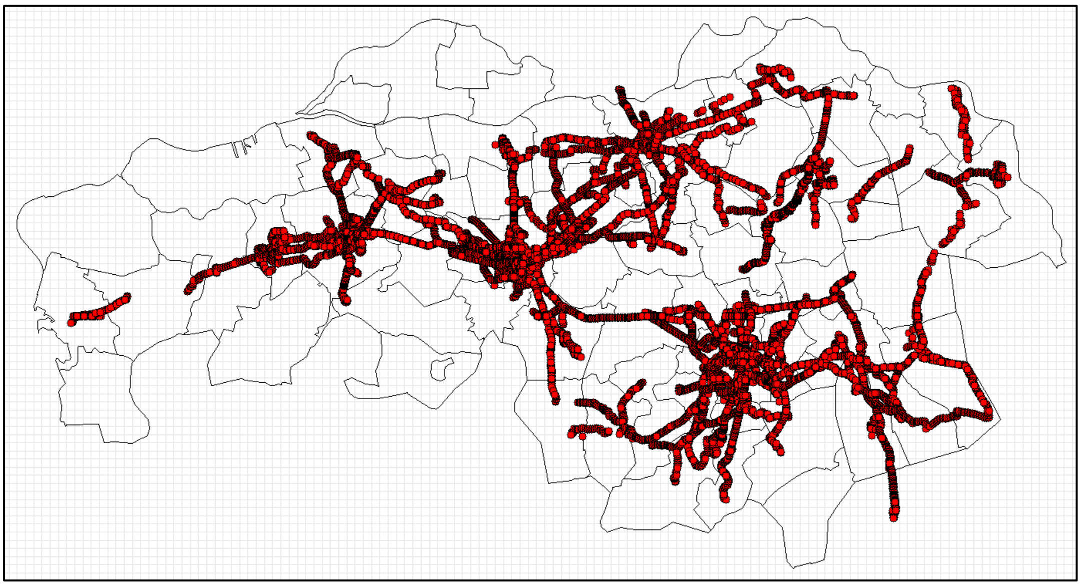


Figure 5-15. Commuter travel patterns March 2014

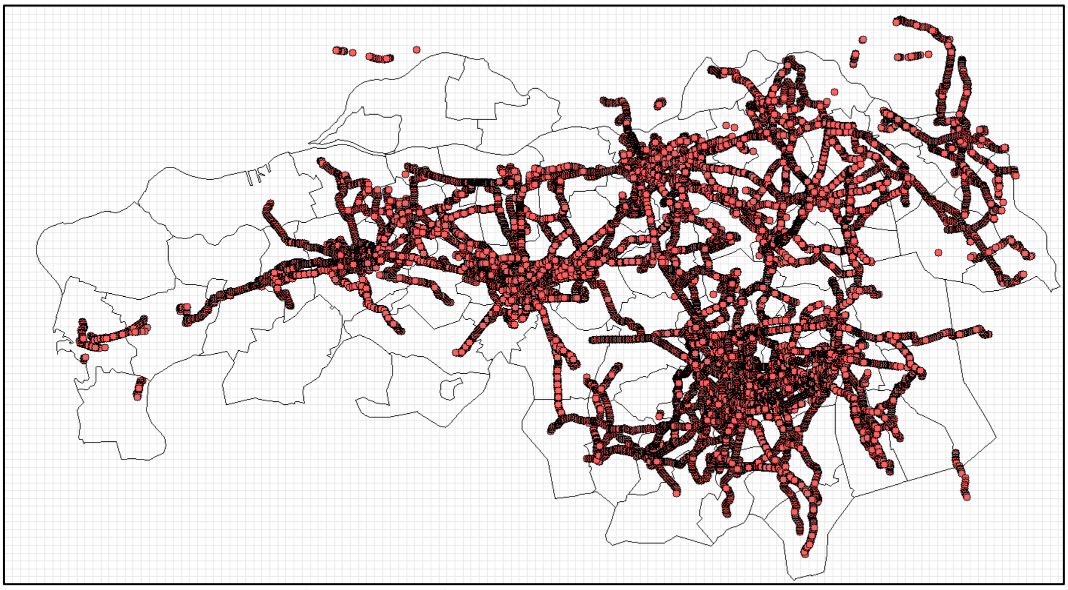


Figure 5-16. Commuter travel patterns April 2014

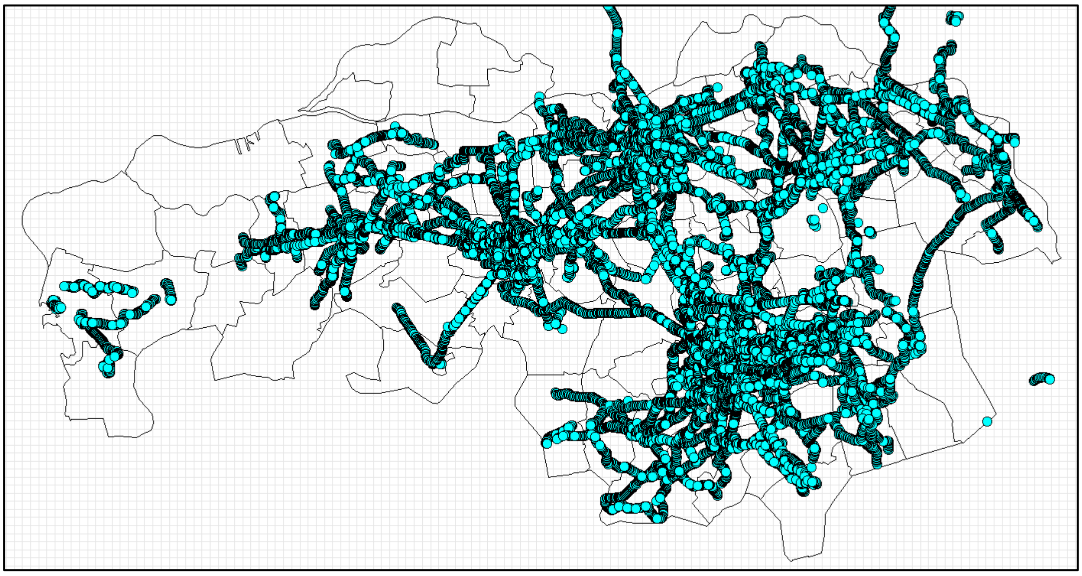


Figure 5-17. Commuter travel patterns May 2014

5.2.2 Temporal partition

The temporal partition is based on dividing the trips according to the time of day TimeP and the type of day DayT.

The type of day DayT is divided into:

- Weekend day
- Week day
- Holiday

For this study the focus is on weekdays which are equivalent to business days, because this is when most people travel to work and when poor network performance has the worst implications. Table 10 presents the type of days and the amount of these days each of the three months has.

Table 10. Type of days in March April May 2014

Month	Weekdays	Weekend days	Holidays
March 2014	21 days	10 days	2 days (both also weekend days)
April 2014	20 days	8 days	4 days (2 also weekend days)
May 2014	19 days	9 days	5 days (2 also weekend days)

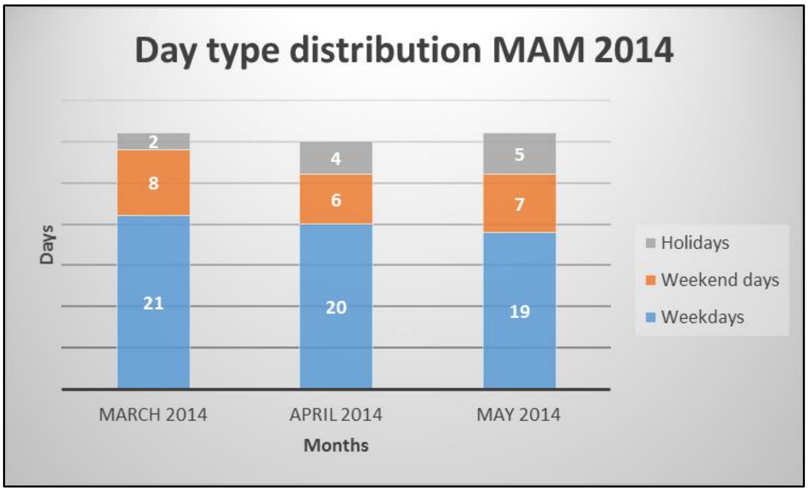


Figure 5-18. Types of days in May April March 2014

As can be seen from figure 5-19, each month has around the same amount of week days. March had no holidays during the week, and thus will most likely have generated the most data (usable and unusable data)

For the Time of day **TimeP** the study aims to capture the time of day, not only when the congestion level is high but also when it has the most implication. From the literature two options were identified. The first includes the morning peak, when most people travel to work or educational institutions. The second option includes the evening peak, when commuters travel home. Based on the two criteria the chosen TimeP is during the morning peak. The exact chosen time is between **7:00 AM – 9:00 AM.**

5.2.3 Generating measures

After the temporal and spatial partition the performance measures were generated for all three months. These include the average speed per trip V, the average route directness per trip R, the average amount of trips and the average travel time per kilometer per trip TT. Figure 5-20 until 5-23 present the results.

Figure 5-20 presents the average speed V per trip during 7:00 AM and 9:00 AM for all three months in 2014. Here it can be seen the average speed is very low during the morning peak. It also shows that in April and May the average speed between 7:00 AM and 8:00 AM is almost zero. This could be caused by traffic jams occurring in some parts of the study area.

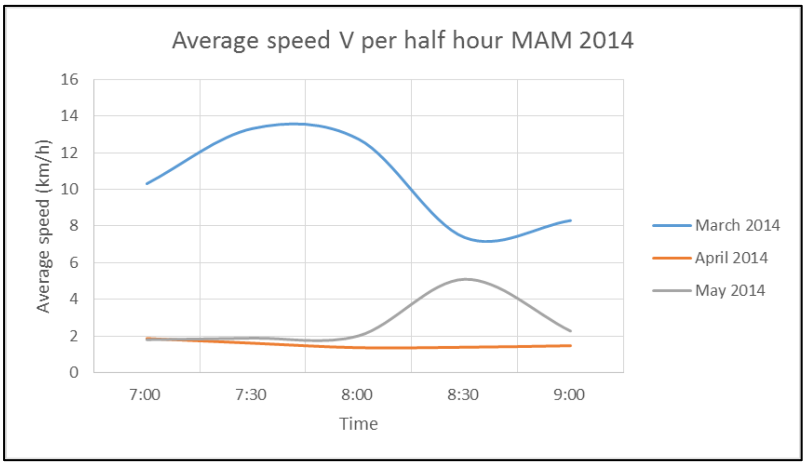


Figure 5-19. Average speed V per half hour for March April May 2014

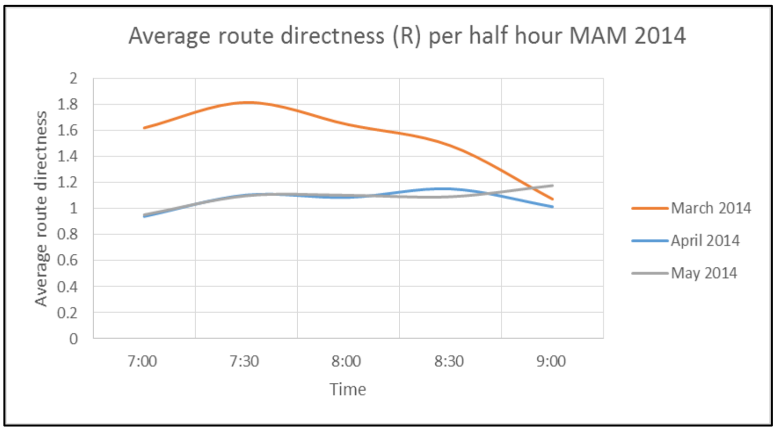


Figure 5-20. Average route directness R per half hour for March April May 2014

Figure 5-21 presents the average route directness R per trip for the study period. From this figure it can be seen that the route starts increasing until approximately 7:30AM and then stays constant or decreases towards the end of the morning peak 9:00AM.

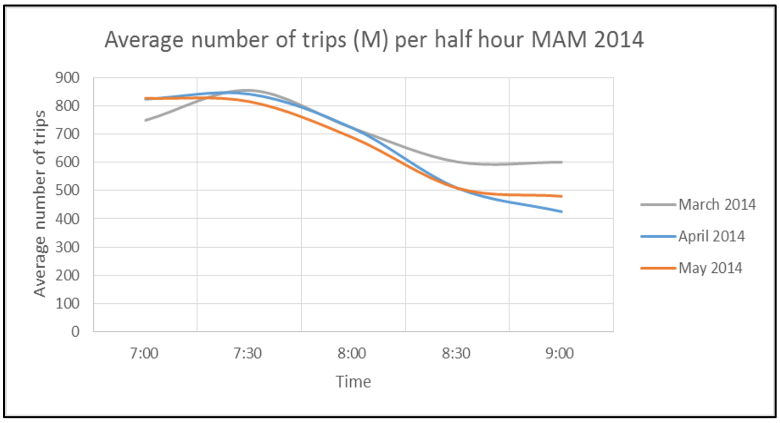


Figure 5-21. Average number of trips M per half hour for March April May 2014

Figure 5-22 presents the average number of trips during the study period. Here it can be seen that, similar to the route directness, the amount of trips increase until 7:30 and then decrease towards the end. Between 7:00AM and 7:30 AM commuters start traveling.

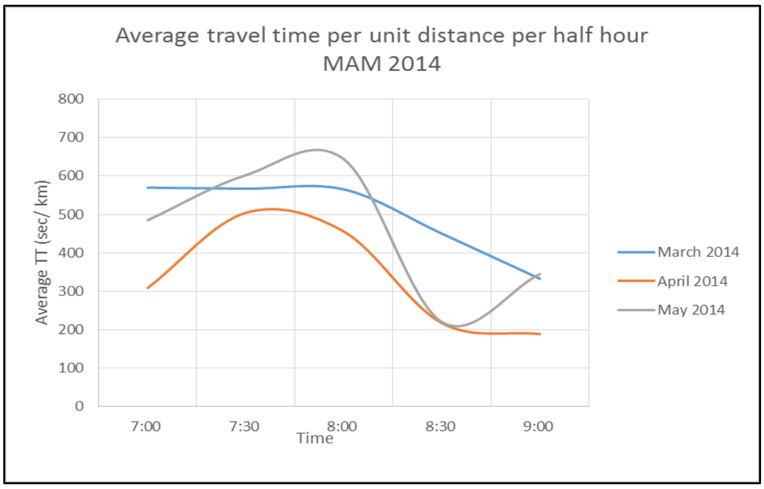


Figure 5-22. Average TT per half hour for March April May 2014

Figure 5-23 presents the average travel time per km per trip. Here it can be seen that during the morning peak the travel times increase until 7:45 AM- 8:05 AM and then start decreasing towards the end of the morning peak. Within this time commuters experience worst network performance

After the performance measures are generated, the next step is to set to identify the regions with serious mismatch problems. This step begins by choosing the threshold values.

Setting threshold values

The empirical values derived from literature include:

- T_V = 20 km/h
- T_R = 1.5
- T_M = 20 trips

T_{TT} is estimated using the percentile statistical method which estimates the probability of the worst travel time that commuters may experience once per month in percentages. This estimation is done using **IBM SPSS Statistics 22**. Here all travel times between origin and destination are extracted and imported in SPSS. After this the frequency analysis is applied.

In the literature 95 percentile is common choice, but for this thesis the 75% and 90% were also estimated. The reason for this is that the 95th percentile describes a kind of worst case scenario as it only shows the 95% probability of commuters experiencing the worst travel time **once** per month. And although the 95% is a good measure, it can be seen from the output table (see table 11) that the 90% is much closer than mean value.

For this reason the threshold T_{TT} will be set at 500 seconds per km (see figure 5-24). This translates into an approximate TT of 8 min/ km. This means that during this time commuters have driving speed of 7.5 km/h, which makes this TT of 500 seconds per kilometer a good threshold value.

Table 11. Frequency analysis output table

Statistics

		March2014	April2014	May2014
N	Valid	82871	118945	117076
	Missing	36074	0	1869
Mean		657,2045	393,2144	522,6763
Variance		1859860,185	6952602,676	13328352,475
Minimum		,00	,00	,00
Maximum		8701,00	61595,00	87542,00
Percentiles	75	507,0000	108,0000	109,0000
	90	2297,0000	481,0000	468,0000
	95	3231,0000	1330,0000	1559,0000

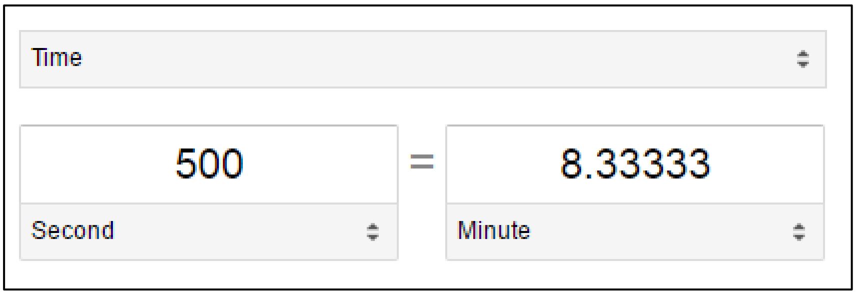


Figure 5-23. TTT from seconds to minutes

5.3 Identifying regions with serious mismatch problems

Figures 5-25 until 5-30 describe the results and present the number of detected regions which experience the predefined problem indicators in values and percentages. Here the total number of regions, which are being used for traveling, is represented by **“Regions total”**

The 5 problem indicators include:

- The three problems are detected with:
    1. $V < T_V$ and $M > T_M$, represented with VM
    2. $R > T_R$ and $M > T_M$, represented with RM
    3. $TT > T_{TT}$ and $M > T_M$, represented with TTM

- The worst regions are described as regions experiencing all three problems at once. Within these regions commuters are forced to take detours, where they also experience some level of traffic congestion. These regions are identified with:
    4. $V < T_V$ and $R > T_R$ and $M > T_M$, represented with VRM
    5. $V < T_V$ and $R > T_R$ and $TT > T_{TT}$ and $M > T_M$, represented with VRTTM

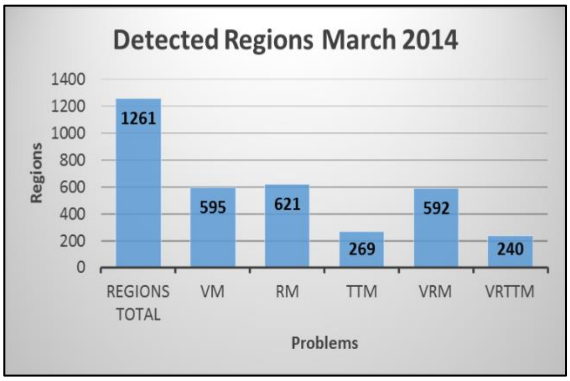


Figure 5-25. Detecting problem regions March 2014

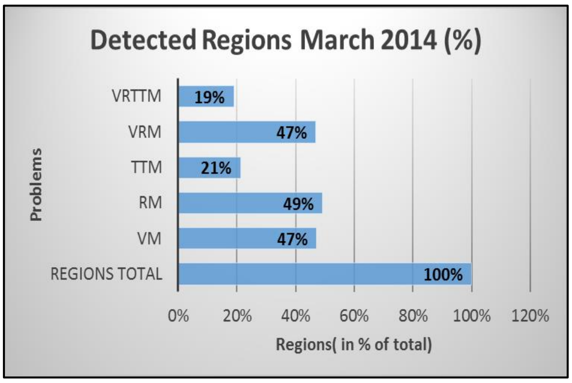


Figure 5-24. Detected problem regions March 2014 (each in % of the total)

Figure 5-25 and 5-26 illustrate the amount of detected regions. Here it can be seen that 1261 regions were used for traveling this month. From these 1261 regions almost half experiences mismatch between travel demand and the ability to actually travel (transport network services). In fact a total of 592 regions experience all problems. This is 47% of all the regions commuters travelled from and to. Combined with the travel time reliability only 240 regions were detected, which counts for 19% of all the regions. The latter suggests that although 49% experience serious mismatch problems, only 19% of the total or 41% of the mismatch regions experience the mismatch problems and have a poor network reliability. Important to mention is that some regions experience multiple problems, and are therefore detected separately for each problem. Figure 5- 26 shows the % amount of regions detected by each problem indicator related to the total of 100% or total regions detected in March 2014.

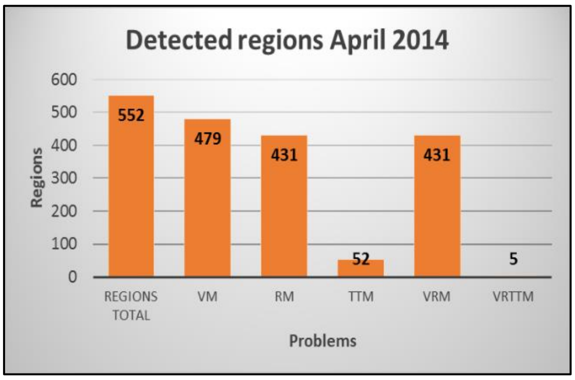


Figure 5-27. Detected problem regions April 2014

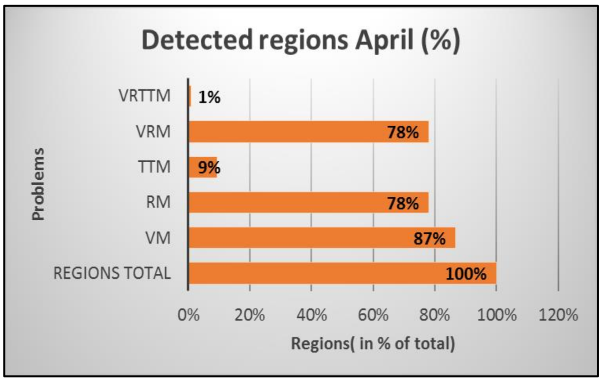


Figure 5-26. Detecting problem regions April 2014 (%)

From figure 5-27 to figure 5-30 it can be seen that the amount of regions being used, in comparison with March 2014, is reduced to more than half in April and May 2014. However around the same amount of regions experience the mismatch problems. It can also be seen that the amount of regions experiencing poor network reliability is reduced considerably. Within the context it can be concluded that during March 2014 most probably road construction was being performed, which led to commuters choosing alternate routes. This explains the higher amount of used regions and poor network travel time reliability. It can also be concluded that although the road construction has finished, around the same amount of regions still experience problems between travel demand and the ability to travel freely.

In this sense the prosed approach can not only be used to measure transport network performance, but also to evaluate the effect of changes (congestion solutions) implemented in the network. Traffic managers and policy makers can implement short term solutions in a network, monitor them by collecting GPS data and evaluate the effectiveness of the solution with the approach proposed in this theory. In this way traffic managers and policy makers can test solutions without spending enormous amounts of money. In turn the short term solution can become a long term solution or it can be removed completely.

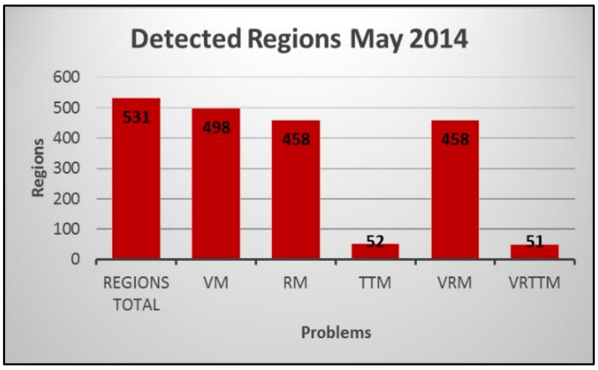


Figure 5-28. Detected problem regions May 2014

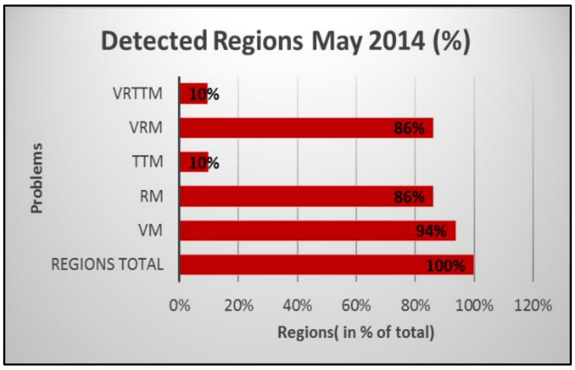


Figure 5-29. Detected problem regions May 2014 (in %)

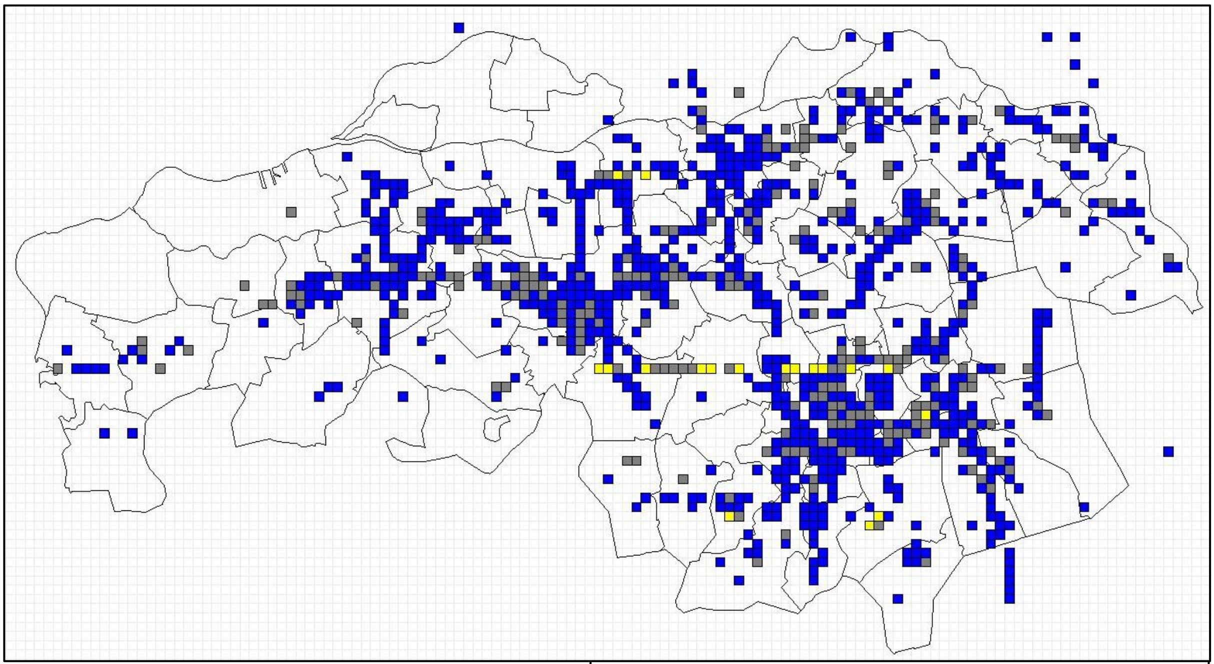


Figure 5-30. Thematic Map April 2014: V < T_V

Figures 5-31 until 5-33 illustrate thematic maps, for March, April and May 2014, of the detected regions based on the first three problem indicators. Understandably the last two indicators can be visualized by combining these maps. These thematic maps present a visualization of the problem regions in the study area. Separate thematic maps for each of the three months of 2014 can be found in the appendix 3- 11.

By combining the thematic maps the areas with serious mismatch problems (areas detecting using the last two indicators) can be extracted. These include the areas with more than one problem.

Figure 5-31 present the thematic map of areas where the average speed is smaller than 20 km/h. The yellow parts show areas were commuters drive at an average speed between 7 and 20 km/h. These areas are most likely the congested areas. The blue and grey parts show areas where cars move at very slow pace. For these areas further investigation is recommended. As these slow speeds could be caused by commuters having to decrease speed for traffic lights. However the grey parts could also be areas where severe congestion occurs on weekdays between 7:00 AM – 9:00 AM.

Figure 5-32 shows the thematic map of areas where the route directness is higher than 1.5. As mentioned before, this indicates that commuters are forced to take detours due to the congestion. From this map it can be concluded that the blue and grey areas suffer from congestion on weekdays between 7:00 AM- 9:00 AM. For this thematic map it is recommended to not take the yellow parts into consideration. The cause for the extremely high values for route directness could also be traffic light stops related. However in this case the values seem too high for it to be caused by severe congestion.

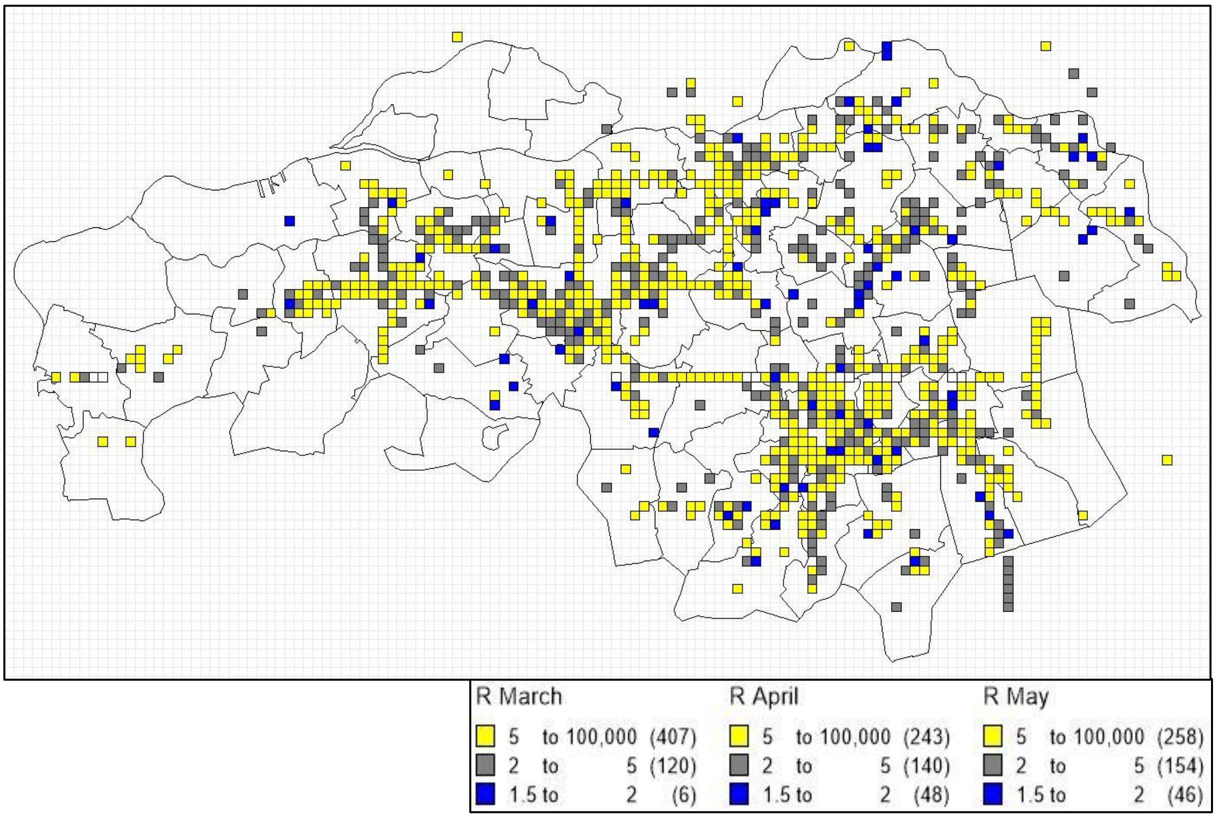


Figure 5-31. Thematic Map April 2014: R > T_R

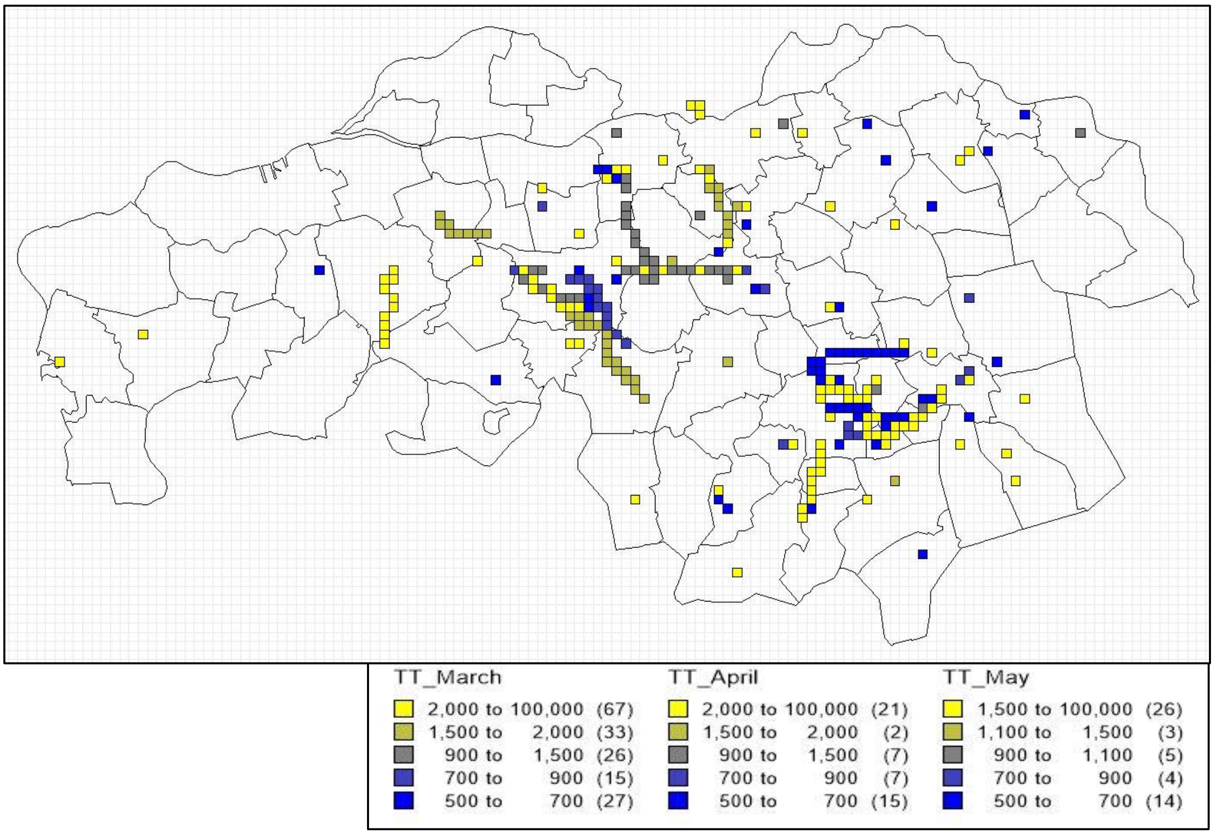


Figure 5-32. Thematic Map April 2014: TT > T_{TT}

Figure 5-33 present the thematic map of areas where it takes commuters more than 500 second to move 1 km (8 min/km). From this map it can be concluded that areas with a higher travel time per unit distance than 900 sec/km should not be considered. These areas need further investigation and also caused by traffic light stops.

The blue and dark blue parts should be the focus. These areas are most likely caused by traffic congestion. In these areas it takes commuters more than 8 minutes to move 1 km. The blue areas are most likely areas with severe traffic congestion and the dark blue areas normal congestion on weekdays between 7:00 AM – 9:00 AM.

5.4 Method validation

In the previous chapters an approach was proposed to evaluate the network performance. As explained at the end of the literature review, this approach from this paper is an improved and extended version of the method created by (Cui, et al., 2016-A). For this reason, the most effective validation is to compare the improved and extended approach proposed in this thesis to the original one from literature.

To avoid confusion the approach in this thesis will be referred to as **thesis approach** from now on and the base approach will be referred to as **base approach.**

Table 12 shows the three key differences between the approaches:

Table 12. Comparison two approaches

Difference	Thesis approach	Base approach
Input data	Individual commuter data	Taxi trip data
Performance measures	1. Speed 2. Route directness 3. Travel time reliability	1. Speed 2. Route directness

To conduct the validation as accurate as possible the following tasks were completed:

1. **From the entire data set one week was chosen for further analysis. This included the week of 8th April until 15th of April 2014 (eight days).**

For the method validation only a small sample of the data is sufficient. For this reason a random week in a random year was chosen. This week consisted of eight days, two of which were weekend days and six week days. Within this week data recorded in the morning peak, between 7:00 AM and 9:00 AM, was used for further analysis.

2. **For the base approach taxi trip data was generated from the one week original data described above.**

Literature revealed that taxi trip data was very similar to the individual commuter data used in this thesis. The key difference is the fact that taxi trip data contains large time gaps between records. Within this context all OD GPS coordinates of every trip, from the original data, were highlighted. After this nine in every ten records, between the origin and destination records, were removed. From the original data set, which included 34866 records, a total of 31174 records were removed. And the new table counted 3692 records. The result showed a data table very similar to that of actual taxi trips (see figure 5-34 and 5-35). After the filter process, the column "Taxi Check" was removed.

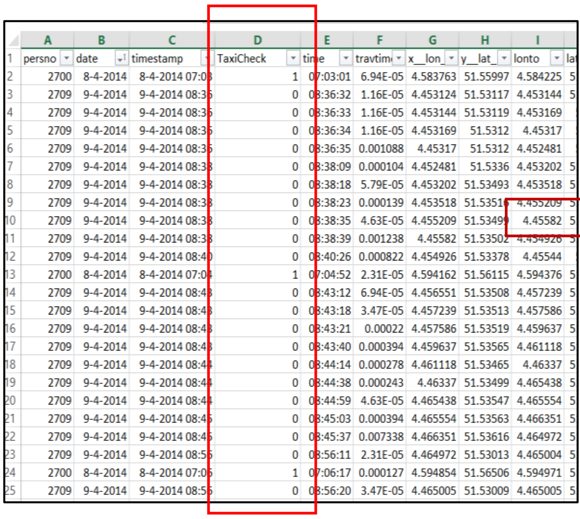


Figure 5-33. Generating taxi data: creating time gaps

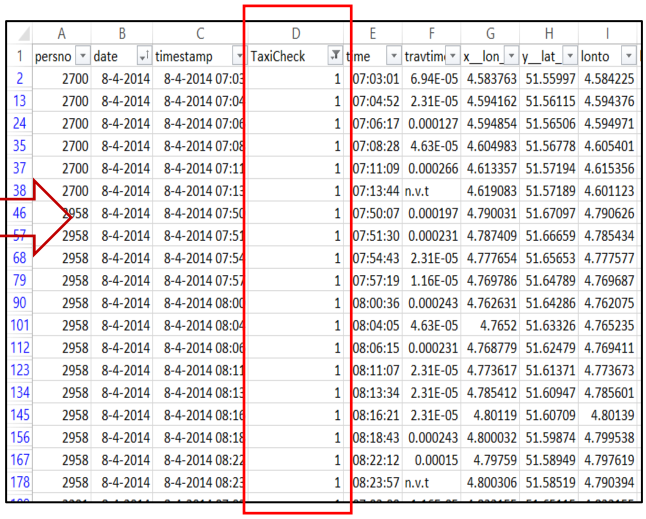


Figure 5-34. Generating taxi data: filter process

3. **For the validation, regions with serious mismatch were identified using both approaches. Here the different thresholds were varied.**

In this step the regions which experience the three problems are identified. As the two methods are constantly compared the validation measure included varying the threshold values and detecting the region for each threshold. The results are then statistically analyzed to calculate the standard deviation. With this value the sensitivity and quality of each method is determined.

The validation method consists of detecting regions using elements of both approaches. These include:

i. **Base approach**
   a. $V < T_V$ and $R > T_R$ and $M > T_M$

This was performed three times. Each of which one threshold varies and the two others stay constant (empirical values). The threshold variation chart is presented in table 13. Each value is basically a variation of the empirical value (either smaller or larger).

Table 13. Validation variation

$V < T_V$ and $R > T_R$ and $M > T_M$	
Measures	Value variation
T_V	(15, 18, 19, **20**, 21, 22, 25)
T_R	(0.5, 0.7, 1.1, **1.5**, 1.7, 1.9, 2.5)
T_M	(15, 18, 19, **20**, 21, 22, 25)

ii. **Thesis approach**

a. V < T_V and R > T_R and TT < T_{TT} and M > T_M

b. TT < T_{TT} and M < T_M

A bit similar to the previous validation measure, here **only T_{TT}** is varied. The other threshold are constant (empirical values). T_{TT} is varied over (200, 400, 450, **500**, 550, 600, 1000), with 500 being the empirical/ estimated value for T_{TT}.

In the next part of this chapter the results of the validation are presented.

5.4.1 Variate the speed threshold

Firstly the regions are identified based on the average speed per trip. The indicator is represented by: V < T_V and R > T_R and M > T_M, and the threshold values include:

- T_V = (15, 18, 19, **20**, 21, 22, 25)
- T_R = 1.5
- T_M = 20

The amount of identified regions is presented in table 14. Here we see that for both approaches the same amount of regions were identified for each variation. However the thesis approach does identify three times the amount of regions compared to the base approach.

Table 14. Detected regions: variation T_V

T_V	Thesis approach	Base approach
15	166	55
18	166	55
19	166	55
20	166	55
21	166	55
22	166	55
25	166	55

Table 15. Validation: Variate T_V

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	c_v
VariateTV_Thesis_approach	7	166,00	166,00	166,0000	,00000	,000	**0,00**
VariateTV_Baseapproach	7	55,00	55,00	55,0000	,00000	,000	**0,00**
Valid N (listwise)	7						

Table 15 shows that both approaches have a standard deviation of 0, and thus c_v of 0. This indicates that both methods produce the same quality results. However the approach of this thesis does identify more regions than the base approach. In the latter policy makers or traffic managers could still use both approaches to identify the bottlenecks in the network based on the average travel speed of commuters. The approach from this thesis would be more thorough and most likely be used for creating long-term solutions. The base approach, on the other hand, would provide enough information to create short term solutions.

5.4.2 Variate the circuity threshold

In the second part of the method validation regions are detected by varying the threshold for route directness and setting the other threshold to their empirical values. The indicator used to identify the problems is: V < TV and R > TR and M > TM. The thresholds include:

- T_V = 20 km/h
- T_R = (0.5, 0.7, 1.1, **1.5**, 1.7, 1.9, 2.5)
- T_M = 20

The amount of identified regions is presented in table 16. Here it can be seen that the thesis approach identifies more regions than the base approach. However, the identified regions per variation calculated with the base approach does estimate more constant results.

Table 16. Detected regions: variation T_R

T_R	Thesis approach	Base approach
0.5	196	57
0.9	186	57
1.1	179	57
1.5	166	55
1.7	157	55
1.9	142	53
2.5	127	47

Table 17. . Validation: Variate T_R

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	c_v
VariateTR_Thesis_approach	7	127,00	196,00	164,7143	24,61513	605,905	**0,15**
VariateTR_Base_approach	7	47,00	57,00	54,4286	3,59894	12,952	**0,07**
Valid N (listwise)	7						

Table 17 shows that the thesis approach has a much higher standard deviation and a c_v twice as big as the base approach, which indicates that the base approach produces much better results for identifying problem regions based on the route directness. The reason for this could be because of the gaps in the taxi trip data. These large gaps in data result into shorter OD destination as this is calculated by the Euclidean distance (shortest path). Understandably shorter distances between origin and destination translate into better route directness. However, this does mean that the base approach produces more accurate results, but not necessarily more reliable ones.

5.4.3 Variate the trip threshold

Similar to the two previous region identification procedures, the indicator: $V < T_V$ and $R > T_R$ and $M > T_M$ is used. Here T_M is varied and the two others are set at threshold values as shown below:

- $T_V = 20$
- $T_R = 1.5$
- T_M = (15, 18, 19, **20**, 21, 22, 25)

The amount of identified regions is presented in table 18. Here it can be seen that, again, the approach in this thesis identifies more regions than the base approach. Also the thesis approach does identify a more constant amount of regions for each variation of the trip threshold.

Table 18. Detected regions: variation T_M

T_M	Thesis approach	Base approach
15	172	79
18	168	64
19	168	58
20	166	55
21	166	52
22	165	48
25	163	40

Table 19. . Validation: Variate T_M

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation		c_v
VariateTM_Thesis_approach	7	163,00	172,00	166,8571	2,85357	8,143	**0,017**
VariateTM_Base_approach	7	40,00	79,00	56,5714	12,46137	155,286	**0,220**
Valid N (listwise)	7						

Table 19 shows that the thesis approach has a smaller c_v than the base approach. This indicates that the thesis approach produces better quality results than the base approach, when the trip threshold values are varied. In the latter the derived conclusion is understandable as the thesis method, which uses trace data, has significantly more trip counts than the base approach which uses taxi data.

5.4.4 Validation travel time reliability measure

The travel time reliability is part of the thesis approach. For this reason the validation was done separately. This validation is divided into two parts. The first includes identifying regions only based only the travel time reliability problem and second one includes identifying problem regions where all three problems occur simultaneously.

5.4.4.1 Compare travel time problems

In this first part regions are identified where only the only high travel times per unit distance occur. Here the indicator TT > T_{TT} and M > T_M is used. For which T_M is set at the empirical value of 20 trips and T_{TT} is varied based on the estimated value of 500 seconds per km (see chapter 5.3). T_{TT} = (200, 400, 450, **500**, 550, 600, 1000)

The amount of identified regions is presented in table 20. Here it can be seen that almost an even amount of regions are detected with both approaches. As the travel time reliability measure is part of the thesis approach, this result proves its validity as a performance measure. Also the thesis approach does identify a slightly more constant amount of regions for each variation of the travel time threshold.

Table 20. Detected regions_1: variation T_{TT}

T_{TT}_1	Thesis approach	Base approach
100	109	48
400	30	17
450	27	16
500	23	16
550	20	13
600	20	13
900	12	7

Table 21. Validation: Travel Time reliability

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	c_v
VariateTTTM_Thesis_approach	7	11,00	31,00	18,4286	6,21442	38,619	**0,337**
VariateTTTM_Base_approach	7	8,00	27,00	15,7143	6,15668	37,905	**0,397**
Valid N (listwise)	7						

Table 21 shows that both approaches have almost the same c_v. However the thesis approach does yield a slightly smaller result. This indicates that the approach of this thesis is better for detecting regions based on the travel time reliability. Also important is the fact that both approaches produced results very close to each other, proving the significance and good quality of the travel time reliability in performance measurement.

5.4.4.2 Variate travel time threshold

In this second part regions are identified which experience multiple problems at the same time. For this reason the indicator TT > T_{TT} and V < T_V and R > T_R and M > T_M is used. Here the only the T_{TT} threshold is varied based on the estimated value of 500 seconds per km, while the others are set at empirical values as shown below:

- T_{TT} = (200, 400, 450, **500**, 550, 600, 1000)
- T_V = 20 km/h
- T_R = 1.5
- T_M = 20 trips

The amount of identified regions is presented in table 22. Here it can be seen that the amount of regions detected with each approach are very close. In fact almost the same amounts of regions are detected with the previous problem indicator. As this indicator identifies the worst problematic regions it can be argued that regions with long TT generally also experience all other problems (V < T_V, R > T_R and M > T_M).

This has significant meaning for traffic managers and policy makers as it proves that travel time reliability is one of the most effective performance measures.

Table 22. Detected regions_2: variation T_{TT}

T_{TT}_2	Thesis approach	Base approach
100	86	46
400	20	15
450	17	14
500	14	14
550	11	11
600	11	11
900	6	6

Table 23. Validation: Variate T_{TT}

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	c_v
VariateTTT_VRM_Thesis_approach	7	10,00	30,00	17,4286	6,21442	38,619	**0,357**
VariateTTT_VRM_Base_approach	7	8,00	27,00	15,7143	6,15668	37,905	**0,392**
Valid N (listwise)	7						

Table 23 shows a very similar result to the previous part. The reason for this is that the same regions which experience the travel time problem also experience all the other problems. This conclusion is applicable for both approaches. For policy makers and traffic managers these findings can be very important as it reveals the regions with the worst network performance.

By finding solutions for the regions identified with this indicator.

5.5 Discussion and conclusions regarding obtained results

The majority of the commuters used either cars or bikes for traveling during the study period. However, from the GPS data it could be seen that there was a small but noticeable increase in car usage each month.

During the filtering and cleaning of the GPS data, almost half of the raw journey-stage data was removed. However this did not have any impact on further analysis, because this data was yet to be fused with the trace data. Once the tables were fused there was enough travel information to conduct the analysis correctly

Within the three months the average travel speed during the morning peak was considerably low, especially in April and May 2014. The reason for this is most likely the occurrence of traffic jams during this time. Between 7:00 AM and 8:00 AM most people started travelling, which resulted in longer travel times per km and higher values for route directness during this hour.

The approach successfully identified problem regions, based on the five problem indicators. From the three thematic maps previously shown as figure 5-31, figure 5-32 and figure 5-33 it can be concluded that the transport network of the region Noord-Brabant performs good. The average amount of detected regions is around 500. This accounts for 5% of the total amount of regions Noord- Brabant is split into to. Additionally it is also important to mention that the majority of the identified problem regions require further investigation. All in all only in a small portion of the identified regions commuters experience traffic congestion during weekdays during the morning peak.

The approach of this thesis produces more accurate and more reliable results than the approach from the literature. This is proven by the fact that the statistical analysis of standard deviation and the coefficient variation of the identified regions using the variation of average speed, travel time, and the amount of trips, yields smaller values for the thesis approach.

With regards to the travel time reliability, both approaches showed similar results. This proves the validity and quality of using the travel time reliability as a performance measure.

6 Conclusions

Chapter 6 presents the scientific relevance and the societal relevance of this thesis. Additionally it describes the answers to the main and sub research questions.

6.1 Scientific relevance

The thesis approach provides a reasonably easy way to measure transport network performance. In comparison to previous studies, this thesis evaluates the network performance based on three performance measures. These include temporal performance measures, spatial performance measures and travel time reliability performance measures. Here, a framework is designed to identify regions with serious mismatch between travel demand and transportation network services. Generally the approach proposed in this paper is an extended and improved version of another approach proposed in a previous study (Cui, et al., 2016-A). By using GPS trace data instead of taxi trip data, the Hiversine distance instead of the Euclidean distance and by extending the approach with travel time reliability as a performance measure more accurate results can be obtained more easily. The validation of the approach done by comparing the outcome of both the thesis approach and the base approach. For the validation taxi trip data was generated from the trace data. Data from 8th of April 2014 until the 15th of April 2014 was extracted and analyzed using both approaches. After the performance measures were generated regions with serious mismatch problems were identified using three problem indicators from both approaches, which include:

1) $V < T_V$ and $R > T_R$ and $M > T_M$ (literature and this thesis)
2) $V < T_V$ and $R > T_R$ and $TT > T_{TT}$ and $M > T_M$ (thesis)
3) $TT > T_{TT}$ and $M > T_M$ (thesis)

For the first problem the threshold values T_V, T_R, T_M were varied one by one, while the two others were set at empirical value. For the second and third problem indicator T_{TT} was varied, while all others were set at empirical value. In total five analysis were conducted with seven variations of one of the thresholds each time, for each approach. The reason for this is that the detected regions from analysis of each approach was statistically analyzed using SPSS, to finally extract the standard deviation and coefficient variation. In the latter the results of the statistical analysis proved the validity of the thesis approach and that it produces more accurate results than the base approach.

At the beginning of the analysis a grid was created over the Netherlands. From this grid Noord-Brabant counted a total of 9432 regions. Of these 9432 an average of 500 regions experience serious problems. This accounts for only 5% of the 9432 regions. With that being said, and based on the thematic maps: figure 5-31, figure 5-32 and figure 5-33, it can be concluded that based on the analysis the road transportation network in the Noord- Brabant region performs good overall. It is also important to note that the majority of the identified problem regions require further investigation (as stated in chapter 5.3). All in all only in a small portion of the identified regions commuters experience traffic congestion during weekdays during the morning peak. However these areas should be investigated to determine the severity or seriousness of the problems.

6.2 Answers to the research questions

6.2.1 Answers to the sub questions

"Which measures are needed to accurately analyze network performance?"

In the first phase of the thesis a literature study was conducted. During this study different research papers were reviewed. The goal of this study was to collect as much relevant information as possible to create the foundation of this thesis. The reviewed literature revealed that transport network performance has been studied many times, by different people and using different measures. Each approach has been able to analyze the performance in one way or the other. During the literature study three relevant performance measures were identified. These include temporal efficiency measures, spatial efficiency measures and travel time reliability measures. In previous research these measures have been used individually and sometimes in combination. More recent studies reveal that the performance of a transport networks should be measured on different aspects (Oliveira, Portugal, & Junior, 2016) (Cui, et al., 2016-B). By doing this a better understanding can be derived for how the network works and how it performs. In this sense different measures can also be used in combination with each other.

In the latter it can be concluded that the temporal efficiency measures, spatial efficiency measures and travel time reliability measures are three performance measures fit to measure network performance individually and/ or in combination with each other.

"To what extent can travel time reliability of the transportation network be generated from commuters travel data, specifically GPS data?"

Travel time reliability is described as the travel time per unit distance. Ideally this measure is derived from historic data. However as this data is scarce and hard to obtain, other methods to determine travel time reliability were researched. In the literature the 95th percentile method was identified. This included statistically estimating the travel time reliability rather than calculating it or deriving it from historic data.

From the GPS data both the travel time between origin and destination locations per trip and the distance between origin and destination of each trip were extracted. And by dividing the **travel time between origin and destination locations of all trips made by an individual commuter (t_d-t_o)** by the **distance between origin and destination of all trips travelled by an individual commuter (HD (l_d-l_o)),** the travel time per kilometer was calculated. The values were then exported to SPSS and the percentiles were estimated. For research purposes the 75th, 90th and 95th percentile were estimated with frequency analysis. The results showed that the 90th percentile value of around 500 seconds per km was the closest to the mean value.
This value was later used as threshold value for the travel time reliability performance measure.

In the latter it can be concluded that the travel time reliability of a transportation network can be generated from GPS data by doing a frequency analysis of results from the **travel time between origin and destination locations of all trips made by an individual commuter (t_d-t_o)** divided by the **distance between origin and destination of all trips travelled by an individual commuter (HD (l_d-l_o)).**

In what way does the travel time reliability relate to the overall performance of the network?

Evaluating the performance network based on travel time reliability is executed in this thesis by detecting regions with serious mismatch problems with the indicators:

1) $V < T_V$ and $R > T_R$ and $TT > T_{TT}$ and $M > T_M$
2) $TT > T_{TT}$ and $M > T_M$

The first indicator is a combination of temporal, spatial and travel time performance measures and the detected regions represent the regions with the most serious mismatch problem between travel demand and transport network services. The second indicator only uses travel time reliability as a performance measure to identify the problem regions. Understandably neither indicator identified many regions. However each indicator did identify around the same amount of regions. This, along with the validation of the approach, does prove that travel time reliability is an excellent overall performance measure.

From the results of the generated measures it could also be seen that high values of travel times occur around the same time as higher route directness values and higher trip counts. This indicates, just as literature suggests, that travel time reliability is not necessarily related to congestion itself but to the level of congestion.

In the latter it could be concluded that travel time reliability relates to overall network performance in the way that it measures network performance based on the level of congestion.

6.2.2 Answer to the main research question

"To what extent can commuters' travel demand, derived from GPS data, be analyzed with the transport network services and travel time reliability, in order to evaluate the performance of a transportation network in the Netherlands?"

This thesis proposes an approach to evaluate network performance by identifying regions with a serious mismatch problem between travel demand and transport network services. Here three performance measures are used which include temporal performance measures, spatial performance measures and travel time reliability measures. These measures are generated from travel information in the form of GPS data. This data is collected from individual commuters who travel daily in Noord Brabant during March, April and May of the year 2014.

The majority of these commuters either used a car or a bicycle for traveling during the study period. However, from the travel data it could be seen that there was a small but noticeable increase in car usage each month. This could have an effect on road network performance.

Within the three months the average travel speed during the morning peak was considerably low, especially in April and May 2014. The reason for this is most likely the occurrence of traffic jams during this time. Between 7:00 AM and 8:00 AM most people started travelling, which resulted in longer travel times per km and higher values for route directness during this hour.

The problem regions are identified by using 5 problem indicators, which include:

- Detection of regions with individual problems:
    1. $V < T_V$ and $M > T_M$, represented with VM
    2. $R > T_R$ and $M > T_M$, represented with RM
    3. $TT < T_{TT}$ and $M < T_M$, represented with TTM

- The worst regions are described as regions experiencing all three problems at once. Within these regions commuters are forced to take detours, where they also experience some level of traffic congestion. These regions are identified with:
    4. $V < T_V$ and $R > T_R$ and $M > T_M$, represented with VRM
    5. $V < T_V$ and $R > T_R$ and $TT > T_{TT}$ and $M > T_M$, represented with VRTTM

The approach is validated with the most effective validation method. The approach is compared to the original approach it is based on from literature.

The validation procedure revealed that the thesis approach produces better and more accurate results than the approach from the literature. This is proven by the fact that the statistical analysis of standard deviation and the coefficient variation of the identified regions using the variation of average speed, travel time, and the amount of trips yield lower values for the thesis approach.

6.3 Societal relevance

Apart from the scientific relevance, this study also provides social benefits. These include:

1) This thesis can serve as a guide for measuring performance using easily obtained GPS data. The thesis approach provides policy makers ad traffic managers with a way to evaluate transport network easily and at low cost. Additionally the data collection method is not labor intensive.

2) In this thesis the chosen study area was Noord Brabant. However, the proposed approach is created for universal use. The approach can not only be used for other areas, but also to evaluate the performance of different transport networks.

3) The prosed approach can not only be used to measure transport network performance, but also to evaluate the effect of changes (congestion solutions) implemented in the network. Traffic manager and policy can implement short term solutions in a network, monitor them by collecting GPS data and evaluate the effectiveness of the solution with the approach proposed in this theory. In this way traffic managers and policy makers can test solutions without spending enormous amounts of money. In turn the short term solution can become a long term solution or it can be removed completely.

4) The approach proved that regions with longer travel times per unit distance generally also experience all other problems ($V < T_V$, $R > T_R$ and $M > T_M$). This has significant meaning for traffic managers and policy makers as it proves that travel time reliability is one of the most effective and important performance measures. When evaluating a network this aspect should not be ignored.

5) The approach can be used strategically in solving congestion related problems in different phases. In this sense policy makers can use different problem indicators to identify regions with specific problems, which they can solve faster. Additionally, due to the fact that many regions experience more than one problems, policy makers and traffic managers can strategically identify regions with experiencing multiple problems and find one solution to solve them both.

6) Apart from the social relevance to policy makers and traffic managers, the approach is also beneficial for individual commuters. By recording their travel time and travel distance for one or two weeks, commuters can easily calculate the travel time per unit distance of all alternative routes they can take to work. Then the travel time reliability of the network can be estimated using SPSS. The estimated value can then be used it as threshold value. Commuters can stop using routes that took longer than the threshold value.

7 Recommendations and future work

This chapters discusses the recommendations to all parties this study concerns. The chapter also includes recommendations for future work.

7.1 General recommendations

The analysis method created in this thesis is a problem detection method. In this sense the method can be used to detect in which parts of the transportation network certain problems occur and at what time. The recommended next step is to examine the problem area and make a route cause analysis for the problem. With this method the cause or causes for the problem can be identified. From here a sustainable solution can be tailored for the problem.

A recommended aspect which should be taken into account for tailoring sustainable solutions for specific problems is the seriousness of these problems. This seriousness of the problem can be measured by calculating the occurrence of a specific problem in a certain area with probabilities. By determining the probability of a problem occurring more times than a certain threshold amount, traffic managers and policy makers can strategically target these areas first. It is also recommended to study the side-effects of targeting these areas first, as this could cause the probability of certain problems occurring to decrease or even disappear in other areas.

As explained in the conclusions the approach can and should also be used to monitor the network performance, especially after a solution for a problem has been implemented. This will create a feedback loop which traffic managers can use in the future to optimize the network. By doing so policy makers can make small changes in certain areas at the end of a certain period of time. Large investments in infrastructural works can be avoided along with long construction periods and frustrated commuters.

7.2 Recommendations for future work

Travel time reliability

In this thesis the travel time reliability is estimated using the 95 percentile method. However a more accurate way of determining the travel time reliability would be to calculate it. By using traffic flow data the travel time reliability can easily be calculated. This calculated value of the travel time reliability will provide a more accurate threshold reference point as it takes a lot more aspects of the transport network and the commuter into consideration. This includes stopping times etc. However it should be noted that traffic flow information is very hard to obtain. Most traffic flow information is only available in real time. Creating a database which stores this information is also highly recommended.

Alternative software tools

A large portion of the analysis described in this thesis was executed using MS Excel and MS SQL. Although these tools did get the job done, the GPS data was almost too much to process. The aim was to do the analysis in this study in a way that everybody could understand and replicate without buying expensive special software tools.

However these special tools are more advanced and created to process large portions of data easily. For this reason it is recommended that for future work other, more advanced tools are used which can process bulk data in a shorter amount of time. The recommended tool to use for follow up research is FME desktop.

FME software tool (FME desktop, FME server)

FME is a software tool which was formerly referred to as the Feature Manipulation Engine. This tool is produced by Safe Software Inc. and is basically an integrated collection of Spatial ETL tools for data transformation and data translation.

FME can be used in combination with GIS, CAD, and BIM. The software also contains a library of over 5000 coordinate systems and provides native support for location data. This tool also has the ability to process bulk data easily.

For the above mentioned reasons it is highly recommended that FME be used for follow up research on this topic in the Netherlands. Apart from the fact that the GPS data can be processed very fast, this data can also be translated into data which is coherent to the RD-New coordinate system of the Netherlands. The coordinate system used in this thesis is the **geospatial coordinate system** (longitude, latitude), but by using RD- New system (Rijks Driehoek coördinaten system), which is especially designed for the Netherlands, the accuracy of the calculated distances will be much more accurate. This will lead to higher accuracy of the results overall.

As the FME software can translate GPS data from one coordinate system to almost any coordinate system designed for a given country, it is also highly recommended to use this tool for follow up research for evaluating transport network performance in other countries.

References

Amini, B., Peiravian, F., Mojarradi, M., & Derrible, S. (2015). *Comparative Traffic Performance Analysis of Urban Transportation Network Structures.* University of Illinois at Chicago, Civil Engineering Department. Retrieved May 25, 2016, from http://arxiv.org/abs/1507.03612

Buhl, J., Gautrais, J., Reeves, N., Sol´e, R., Valverde, S., Kuntz, P., & Theraulaz, G. (2016). Topological patterns in street networks of self-organized urban settlements. *European Physical Journal, B 49*, 513–522.

Castro, P., Zhang, D., & Li, S. (2012). Urban traffic modelling and prediction using large scale GPS. *10th International Conference on Pervasive Computing.* Newcastle; United Kingdom: Pervasive 2012. doi:Code 91424

Cui, J., Liu, F., Hu, J., Janssens, D., Wets, G., & Cools, M. (2016-A). Identifying mismatch between urban travel demand and transport network services using GPS data: A casestudy in the fast growing Chinese city of Harbin. *Neurocomputing*(181), 4- 18.

Cui, J., Liu, F., Janssens, D., Ana, S., GeertWets, & Cools, M. (2016-B). Detecting urban road network accessibility problems using taxi GPS data. *Journal of Transport Geography*(51), 147–157.

Elefteriadou, L., & Cui, X. (2005). *Review of Definitions of Travel Time Reliability.* Draft. Retrieved June 2016, from http://www.dot.state.fl.us/planning/statistics/mobilitymeasures/define-ttr.pdf

Feng, T., & Timmermans, H. J. (2014). *Using Recurrent Spatio-Temporal Profiles in GPS Panel Data for Enhancing Imputation of Activity Type.* Eindhoven: Urban Planning Group, Department of the Built Environment Eindhoven University of Technology.

Feng, T., & Timmermans, H. J. (2016). Integrated imputation of activity-travel diaries incorporating the measurement of uncertainty measurement of uncertainty. *Procedia World Conference on Transport Research.* Shanghai.

Feng, T., Moiseeva, A., & Timmermans, H. (2011). *Processing of National Travel Survey GPS Pilot Data.* Technical report , Technical University of Eindhoven, the Department for Transport.

Gillen, D., & Hasheminia, H. (2016). Measuring reliability of transportation networks using snapshots of movements in the network –An analytical and empirical study. *Transportation Research Part B*(000), 1–17.

Hong, S. Y., Lee, C., Chung, S. B., & Kho, S. Y. (2005). Analysis of Two-fluid Model Using GPS Data. *Journal of the Eastern Asia Society for Transportation Studies, 66*, 560 - 572,.

Hong, S., Lee, C., Bong, C. S., & Kho, S. (2005). Analysis of two-fluid model using GPS data. *Journal of the Eastern Asia Society for Transportation Studies, 6*, 560 - 572.

Liu, X., Gong, L., Gong, Y., & Liu, Y. (2015). Revealing travel patters and city structure with taxi trip data. *Journal of transport Geography*(43), 78-90.

Lu, Y., & Li, S. (2014). An empirical study of with-in day OD prediction using taxi GPS data in Singapore. *Proceedings of Transportation Research Board 93rd Annual Meeting.*

Lyman, K. (2007). *Travel Time Reliability in Regional Transportation Planning.* Field Area Paper.

Mahmassani, H. S., Hou, T., & Saberi, M. (2013). Connecting Networkwide Travel Time Reliability and the Network Fundamental Diagram of Traffic Flow. *Journal of the Transportation Research Board*(2391), 80-91.

Mishra, S., Welch, T. F., & Jha, M. K. (2012). Comparative Traffic Performance Analysis of transportation networks. *Transportation Research Part A*(46), 1066-1085.

Oliveira, E. L., Portugal, L. d., & Junior, W. P. (2016). Indicators of reliability and vulnerability: Similarities and differences in ranking links of a complex road system. *Transportation Research Part A*(88), 195–208.

Ortúzar, J. d., & Willumsen, L. G. (2011). *Modeling Transport* (4 ed.). London: John Wiley and Sons, Ltd.

Research Board of Transportation. (2010). *NCHRP Report 664.* CAMBRIDGE SYSTEMATICS, INC., UNIVERSITY OF MARYLAND, CENTER FOR ADVANCED TRANSPORTATION TECHNOLOGY, RESOURCE SYSTEMS GROUP, National Cooperative Highway Research Program.

Snellen, D., Borgers, A., & Timmermans, H. (2002). Urban form, road network type, and mode choice for frequently conducted activities: a multilevel analysis using quasi-experimental design data. *Environment and Planning A, 34*, 1207-1220.

Storey, B., & Holtom, R. (2003, November). The use of historic GPS data in transport and traffic monitoring. *Traffic Engineering and Control, 44*(10), 376-379.

Tsekeris, T., & Geroliminis, N. (2013). City size, network structure and traffic congestion. *Journal of Urban Economics*(76), 1–14.

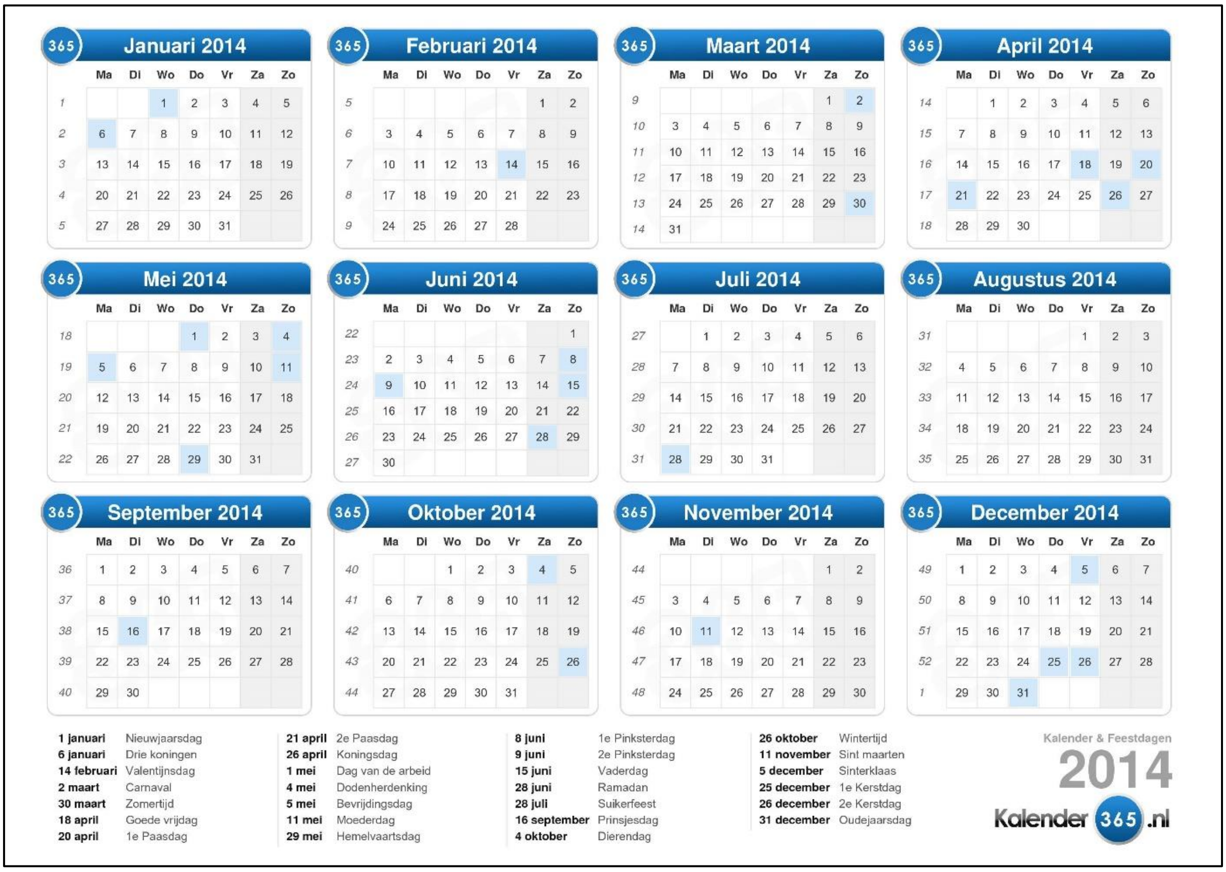
Uchida, K. (2014). Estimating the value of travel time and of travel time reliability in road networks. *Transportation Research Part B*(66), 129–147.

Vuren, T. v., & Carey, C. (2011). Building Practical Origin-Destination (OD/Trip) Matrices from Automatically Collected GPS data. *European Transport Conference*. Retrieved May 31, 2016, from http://.etcproceedings.org

Wu, C., Thai, J., Yadlowsky, S., Pozdnoukhov, A., & Bayen, A. (2015). Cellpath: fusion of cellular and traffic sensor data for route flow estimation via convex optimization. *Transportation Research Procedia*(7), 212 – 232.

Appendix 1

Appendix 1. Calendar 2014



Januari 2014

	Ma	Di	Wo	Do	Vr	Za	Zo
1			1	2	3	4	5
2	6	7	8	9	10	11	12
3	13	14	15	16	17	18	19
4	20	21	22	23	24	25	26
5	27	28	29	30	31		

Februari 2014

	Ma	Di	Wo	Do	Vr	Za	Zo
5						1	2
6	3	4	5	6	7	8	9
7	10	11	12	13	14	15	16
8	17	18	19	20	21	22	23
9	24	25	26	27	28		

Maart 2014

	Ma	Di	Wo	Do	Vr	Za	Zo
9						1	2
10	3	4	5	6	7	8	9
11	10	11	12	13	14	15	16
12	17	18	19	20	21	22	23
13	24	25	26	27	28	29	30
14	31						

April 2014

	Ma	Di	Wo	Do	Vr	Za	Zo
14		1	2	3	4	5	6
15	7	8	9	10	11	12	13
16	14	15	16	17	18	19	20
17	21	22	23	24	25	26	27
18	28	29	30				

Mei 2014

	Ma	Di	Wo	Do	Vr	Za	Zo
18				1	2	3	4
19	5	6	7	8	9	10	11
20	12	13	14	15	16	17	18
21	19	20	21	22	23	24	25
22	26	27	28	29	30	31	

Juni 2014

	Ma	Di	Wo	Do	Vr	Za	Zo
22							1
23	2	3	4	5	6	7	8
24	9	10	11	12	13	14	15
25	16	17	18	19	20	21	22
26	23	24	25	26	27	28	29
27	30						

Juli 2014

	Ma	Di	Wo	Do	Vr	Za	Zo
27		1	2	3	4	5	6
28	7	8	9	10	11	12	13
29	14	15	16	17	18	19	20
30	21	22	23	24	25	26	27
31	28	29	30	31			

Augustus 2014

	Ma	Di	Wo	Do	Vr	Za	Zo
31					1	2	3
32	4	5	6	7	8	9	10
33	11	12	13	14	15	16	17
34	18	19	20	21	22	23	24
35	25	26	27	28	29	30	31

September 2014

	Ma	Di	Wo	Do	Vr	Za	Zo
36	1	2	3	4	5	6	7
37	8	9	10	11	12	13	14
38	15	16	17	18	19	20	21
39	22	23	24	25	26	27	28
40	29	30					

Oktober 2014

	Ma	Di	Wo	Do	Vr	Za	Zo
40			1	2	3	4	5
41	6	7	8	9	10	11	12
42	13	14	15	16	17	18	19
43	20	21	22	23	24	25	26
44	27	28	29	30	31		

November 2014

	Ma	Di	Wo	Do	Vr	Za	Zo
44						1	2
45	3	4	5	6	7	8	9
46	10	11	12	13	14	15	16
47	17	18	19	20	21	22	23
48	24	25	26	27	28	29	30

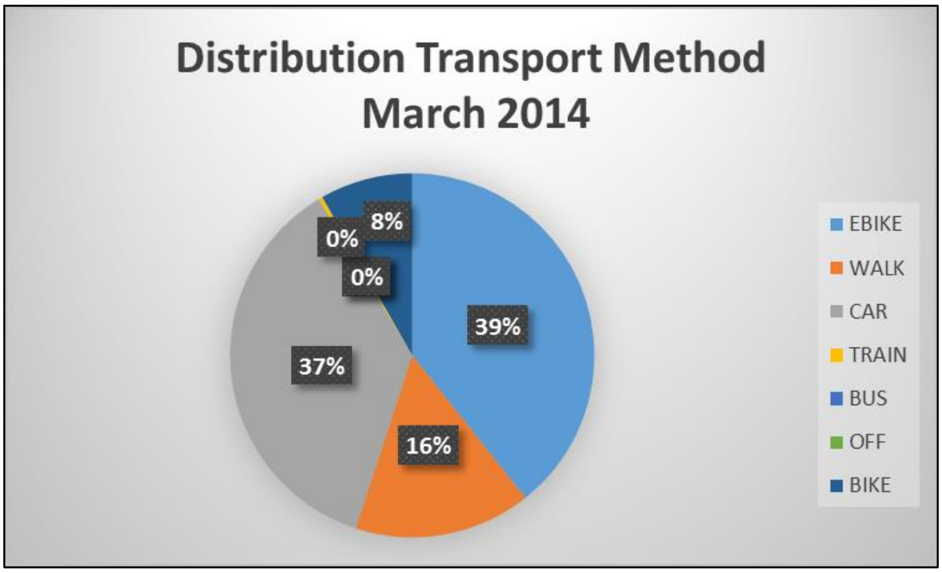
December 2014

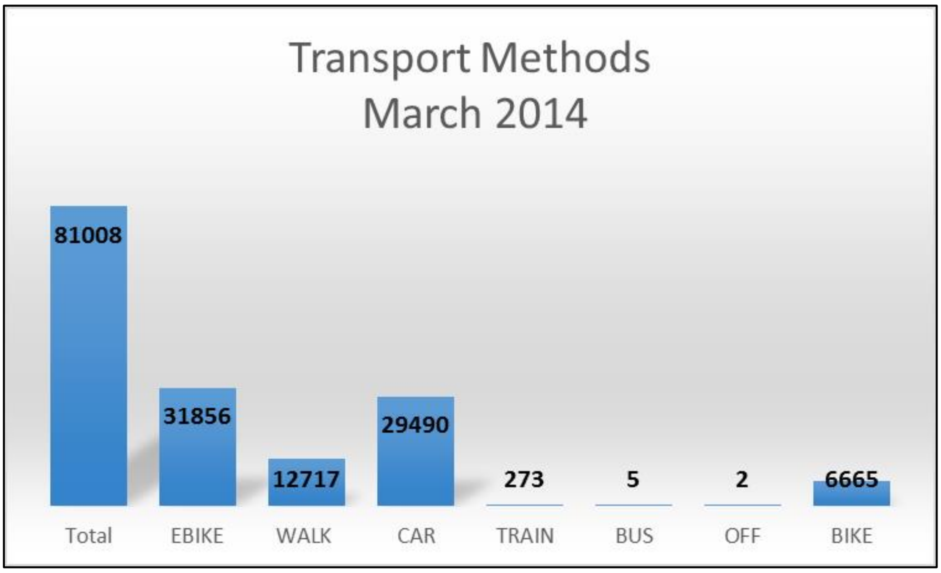
	Ma	Di	Wo	Do	Vr	Za	Zo
49	1	2	3	4	5	6	7
50	8	9	10	11	12	13	14
51	15	16	17	18	19	20	21
52	22	23	24	25	26	27	28
1	29	30	31				

1 januari	Nieuwjaarsdag
6 januari	Drie koningen
14 februari	Valentijnsdag
2 maart	Carnaval
30 maart	Zomertijd
18 april	Goede vrijdag
20 april	1e Paasdag
21 april	2e Paasdag
26 april	Koningsdag
1 mei	Dag van de arbeid
4 mei	Dodenherdenking
5 mei	Bevrijdingsdag
11 mei	Moederdag
29 mei	Hemelvaartsdag
8 juni	1e Pinksterdag
9 juni	2e Pinksterdag
15 juni	Vaderdag
28 juni	Ramadan
28 juli	Suikerfeest
16 september	Prinsjesdag
4 oktober	Dierendag
26 oktober	Wintertijd
11 november	Sint maarten
5 december	Sinterklaas
25 december	1e Kerstdag
26 december	2e Kerstdag
31 december	Oudejaarsdag

Appendix 2-A

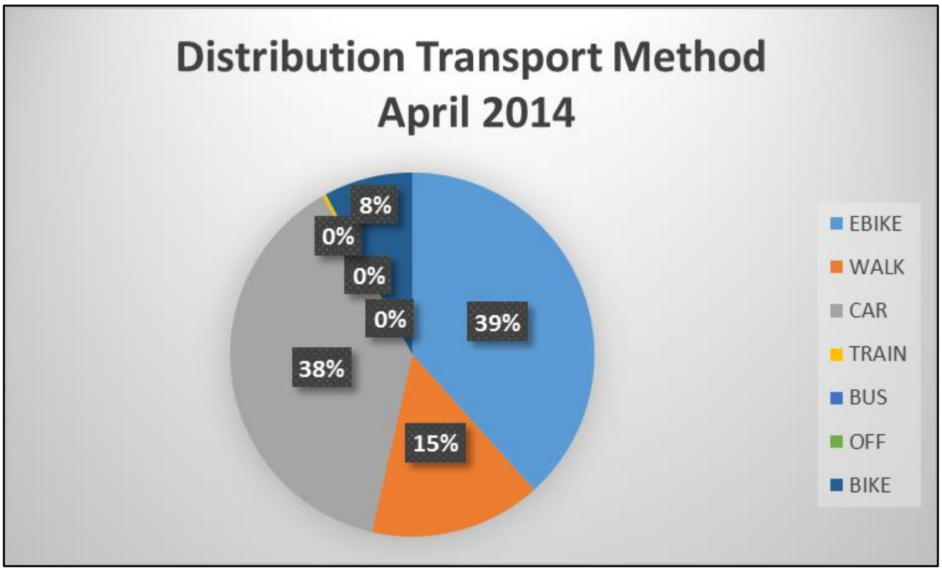
Appendix 2-A. Distribution TM March 2014

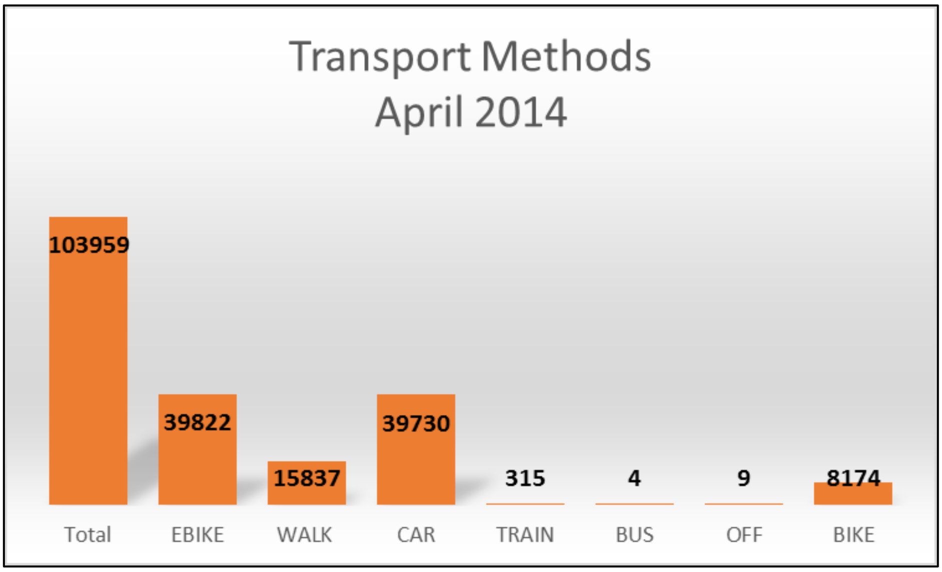




Appendix 2-B

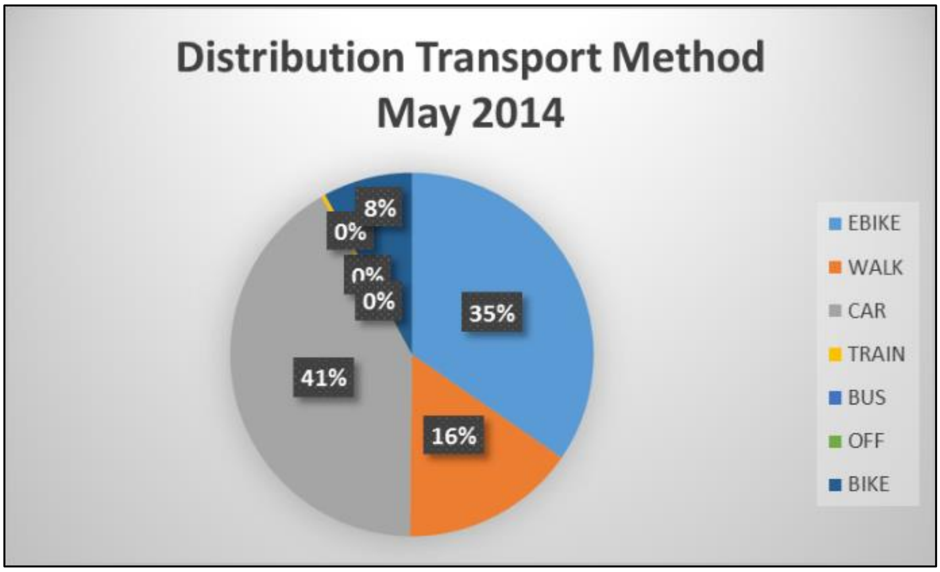
Appendix 2-B. Distribution TM April 2014

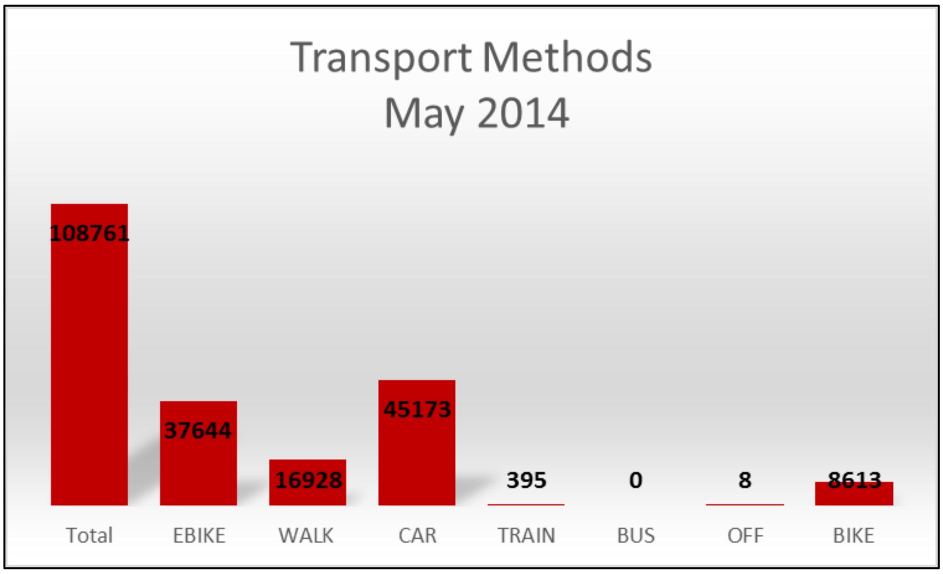




Appendix 2-C

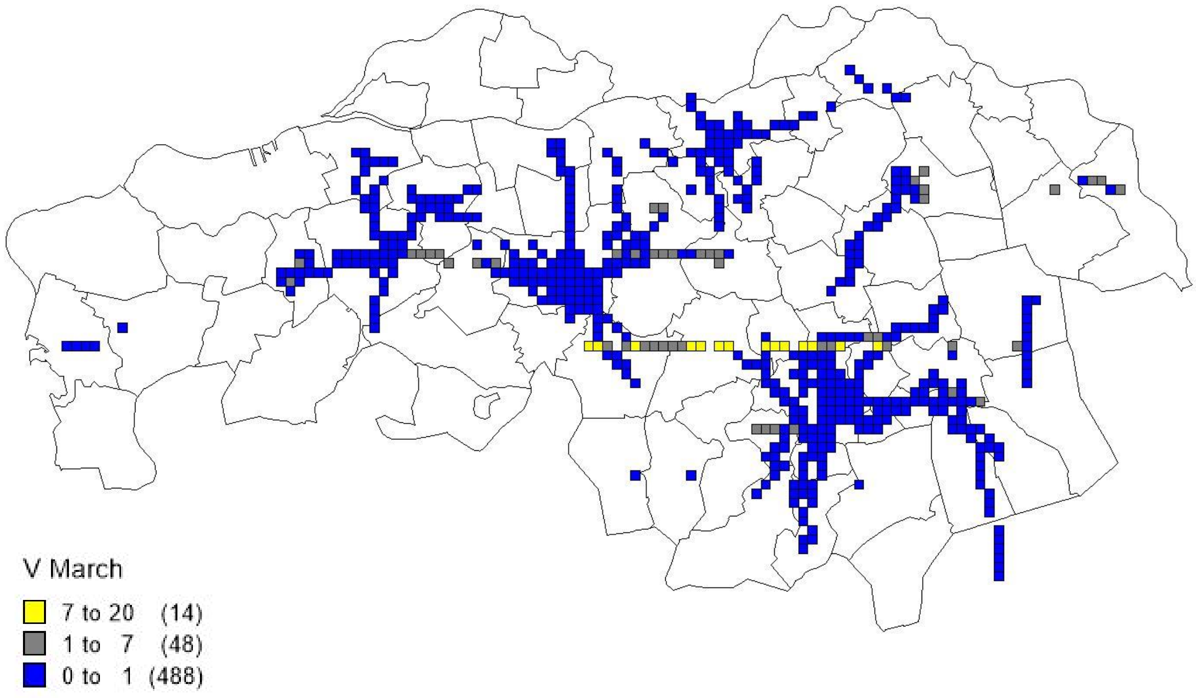
Appendix 2-C. Distribution TM May 2014



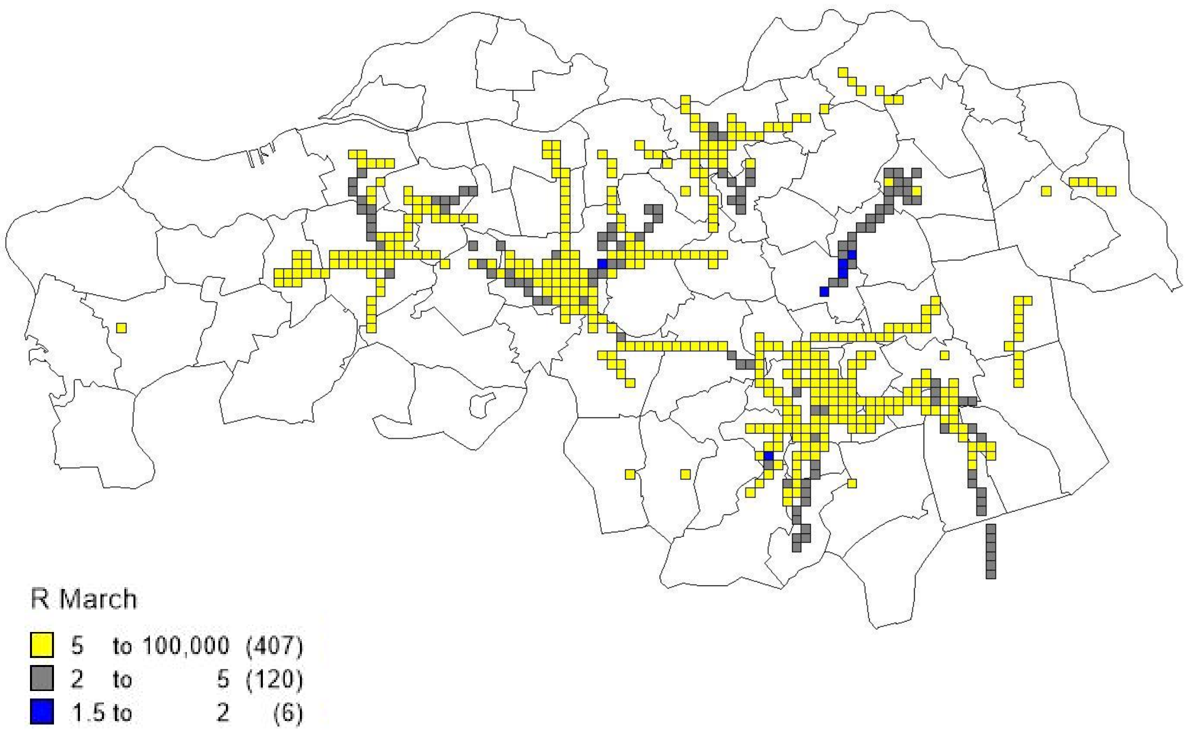


Appendix 3-11

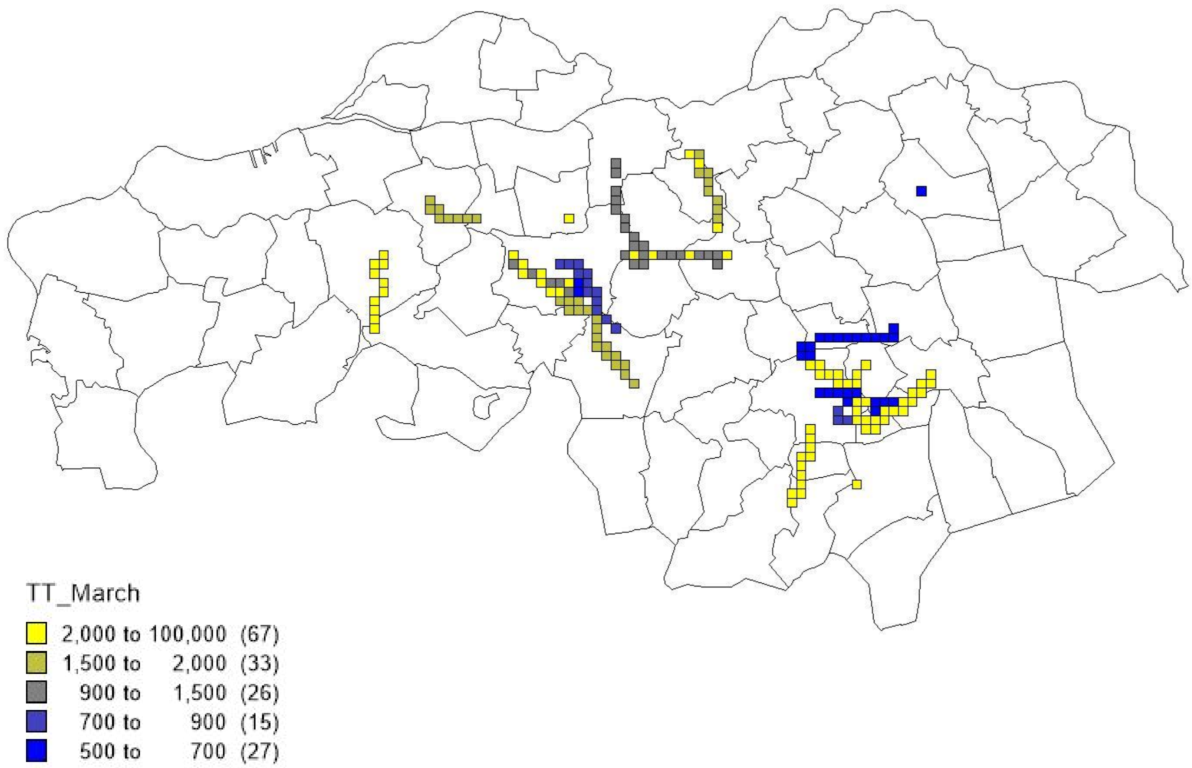
Appendix 3. Thematic Map March 2014: V < TV



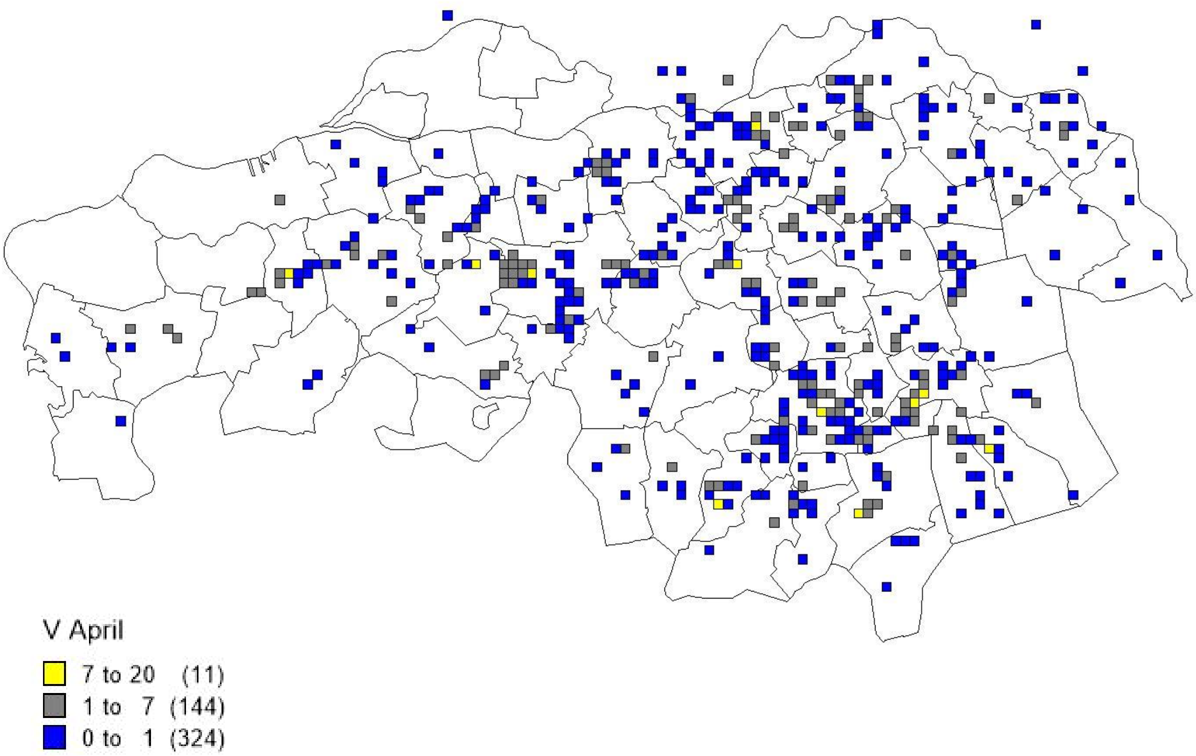
Appendix 4. Thematic map March 2014: R > TR



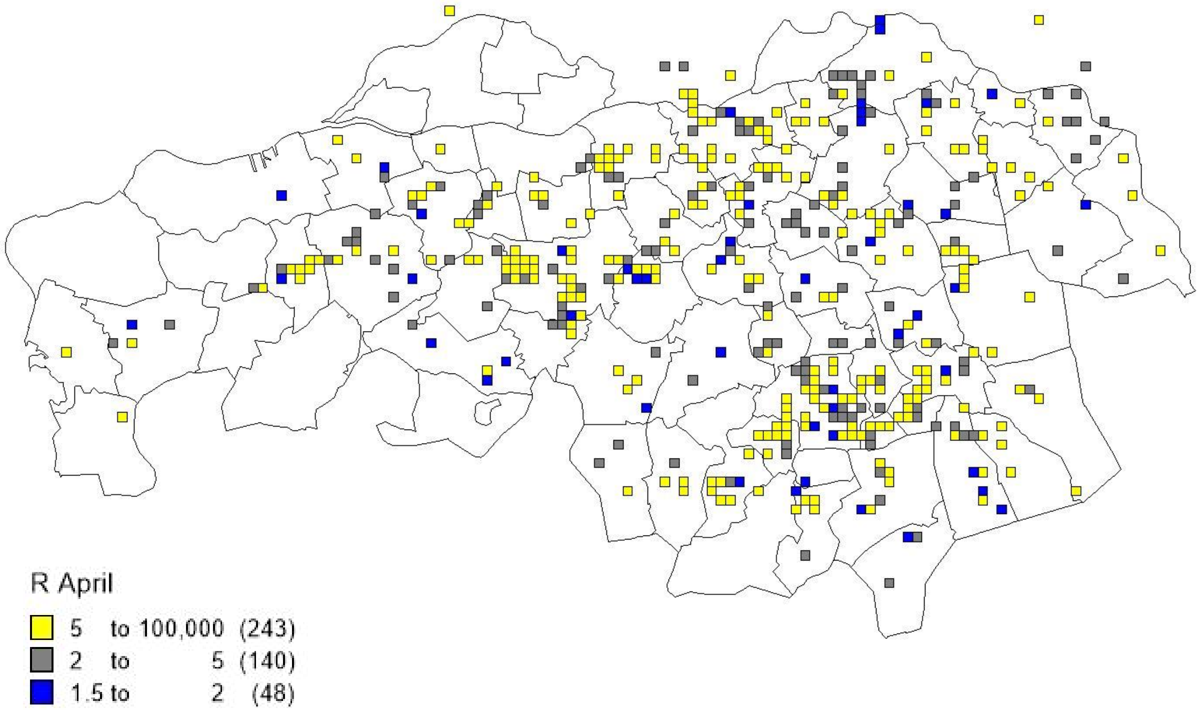
Appendix 5. Thematic map March 2014: TT > TTT



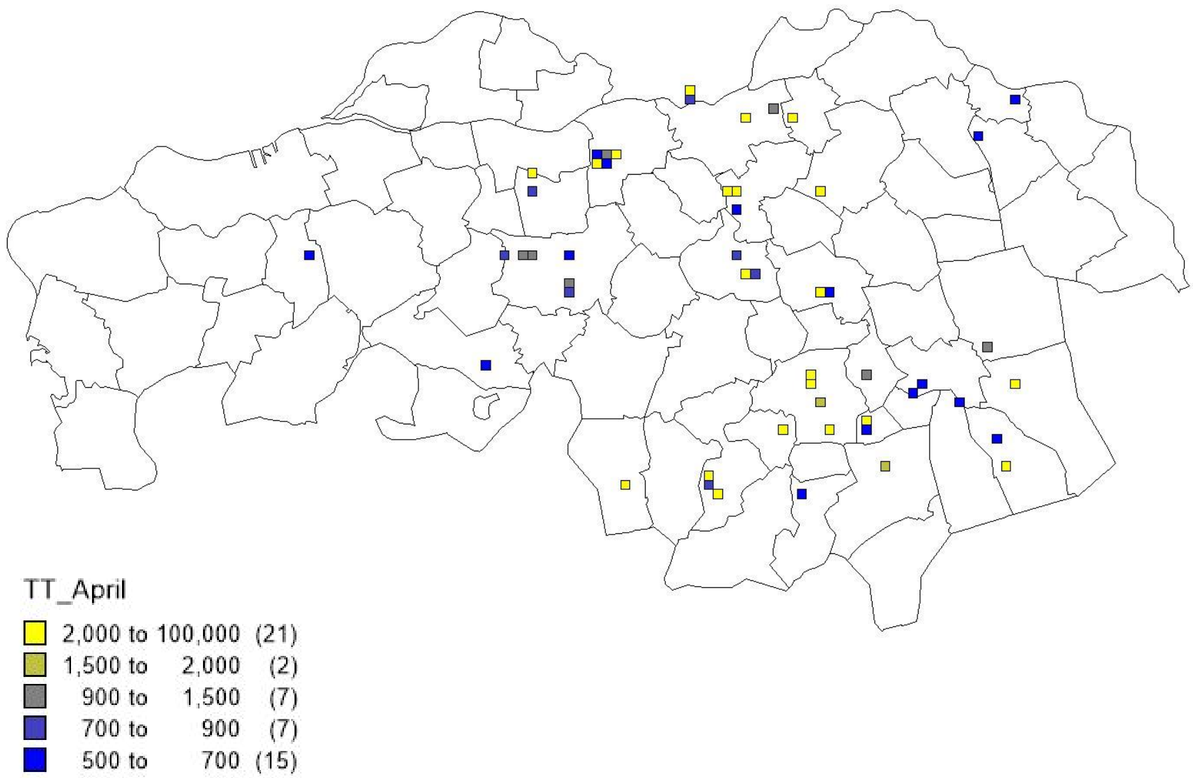
Appendix 6. Thematic map April 2014: V < TV



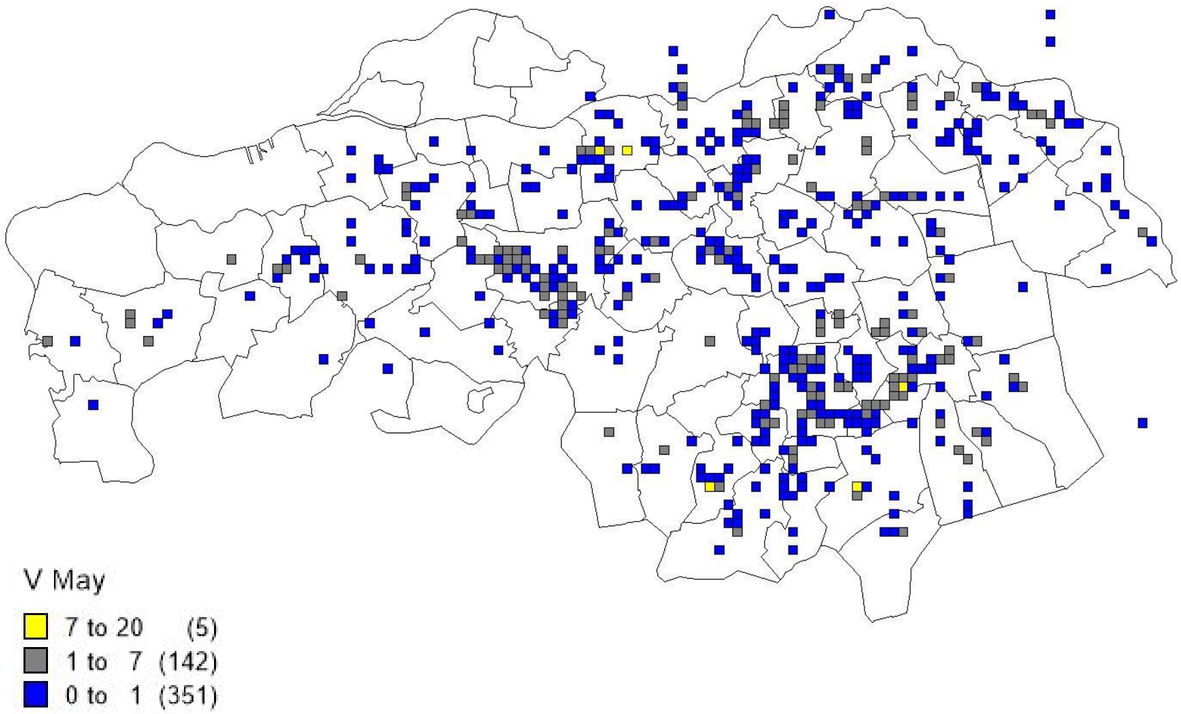
Appendix 7. Thematic map April 2014: R > TR



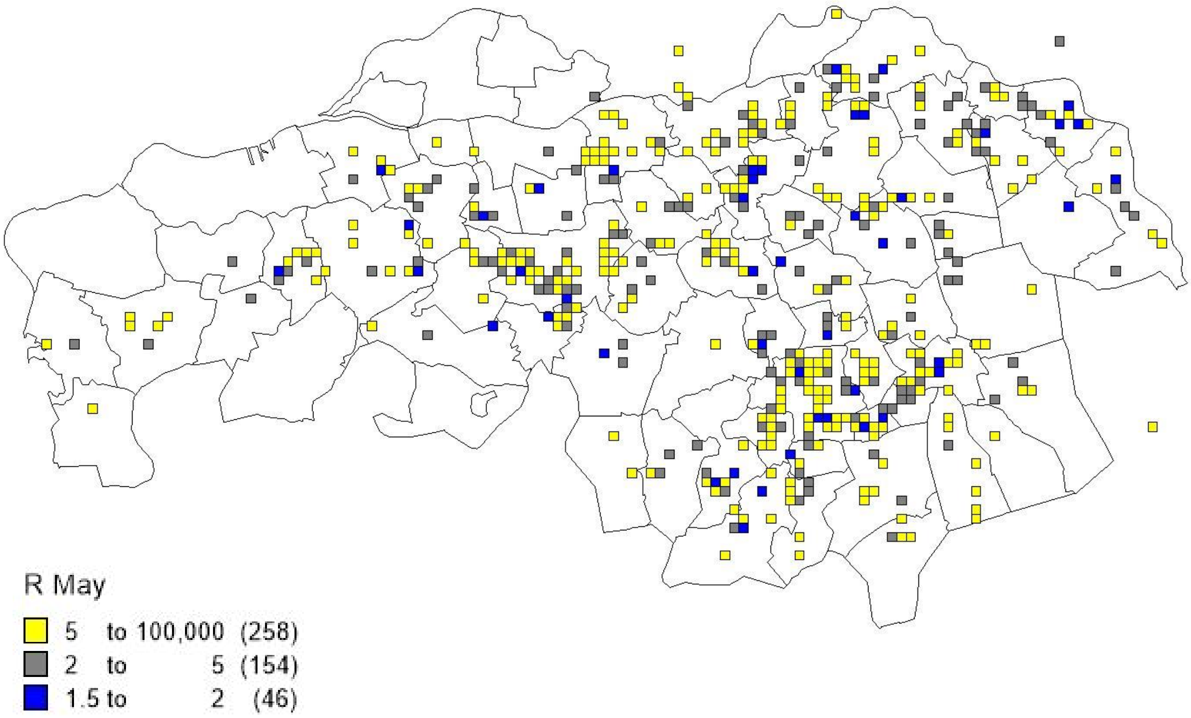
Appendix 8. Thematic map April 2014: TT > TTT



Appendix 9. Thematic map May 2014: V < TV



Appendix 10. Thematic map May 2014: R > TR



Appendix 11. Thematic map May 2014: TT > TTT