Highway Traffic State Estimation and Short-term Prediction

Andreas Allström

Department of Science and Technology
Linköping University, Sweden
Linköping 2016
Highway Traffic State Estimation and Short-term Prediction

©Andreas Allström, 2016

Printed in Sweden by LiU-Tryck, Linköping, Sweden, 2016

ISBN: 978-91-7685-757-1
ISSN: 0280-7971
ABSTRACT

Traffic congestion is increasing in almost all large cities, leading to a number of negative effects such as pollution and delays. However, building new roads is not a feasible solution. Instead, the use of the existing road network has to be optimized, together with a shift towards more sustainable transport modes. In order to achieve this there are several challenges that needs to be addressed. One challenge is the ability to provide accurate information about the current and future traffic state. This information is an essential input to the traffic management center and can be used to influence the choices made by the travelers. Accurate information about the traffic state on highways, where the potential to manage and control the traffic in general is very high, would be of great significance for the traffic managers. It would help the traffic managers to take action before the system reaches congestion and limit the effects of it. At the same time, the collection of traffic data is slowly shifting from fixed sensors to more probe based data collection. This requires an adaptation and further development of the traditional traffic models in order for them to handle and take advantage of the characteristics of all types of data, not just data from the traditionally used fixed sensors.

The objective of this thesis is to contribute to the development and implementation of a model for estimation and prediction of the current and future traffic state and to facilitate an adaptation of the model to the conditions of the highway in Stockholm. The model used is a version of the Cell Transmission Model (CTM-v) where the velocity is used as the state variable. Thus, together with an Ensemble Kalman Filter (EnKF) it can be used to fuse different types of point speed measurements. The model is developed to run in real-time for a large network. Furthermore, a two-stage process used to calibrate the model is implemented. The results from the calibration and validation show that once the model is calibrated, the estimated travel times corresponds well with the ground truth travel times collected from Bluetooth sensors.

In order to produce accurate short-term predictions for various networks and conditions it is vital to combine different methods. We have implemented and evaluated a hybrid prediction approach that assimilates parametric and non-parametric short-term traffic state prediction. To predict mainline sensor data we use a neural network, while the CTM-v is ran forward in time in order to predict future traffic states. The results show that both the hybrid approach and the CTM-v prediction without the additional predicted mainline sensor data is superior to a naïve prediction method for longer prediction horizons.
ACKNOWLEDGEMENT

First, I would like thank my supervisors Jan Lundgren and Clas Rydergren at Linköping University, Department of Science and Technology (ITN). They have provided valuable feedback and support throughout the work with this thesis.

I would also like to thank my colleagues at Linköping University and Sweco. Without you, this thesis would not have been possible and there are some people that I would like to mention in particular. First of all, I would like to mention Magnus Fransson, Mats Sandin, Joakim Ekström, Rasmus Ringdahl and Viktor Bernhardsson, who to various extent have been involved in the work presented in this thesis. I also would like to show my appreciation to David Gundlegård at Linköping University who has been a great colleague and mentor throughout the work with this thesis, and with whom I have had many interesting and challenging discussions. Furthermore, I want to mention my group managers at Sweco, Jeffery Archer, Jenny Widell and Thomas Sjöström, who have supported me and encouraged me during the work with this thesis.

Another group of people that I would like to thank for their support and feedback on my research is the members of the Swedish ITS Postgraduate School.

Furthermore, I would like to express my gratitude towards Alexandre M. Bayen and Joe Butler for welcoming me to UC Berkeley and CCIT/PATH. The trips I have made to Berkeley during the work with this thesis has been a great source of inspiration and motivation. I also have to thank the Swedish Transport Administration and Tomas Julner in particular, who has funded most of the work presented in this thesis.

Finally, I would like to thank friends and family for always supporting me. In particular Sandra, Teodor and Isak, who brings so much love and happiness to my life.

Andreas Allström
Stockholm, May 2016
TABLE OF CONTENTS

Abstract .......................................................................................................................................................... 1
Acknowledgement ........................................................................................................................................ 3
Table of contents .......................................................................................................................................... 5
1 Introduction .......................................................................................................................................... 9
   1.1 Mobile Millennium Stockholm ................................................................................................ 10
   1.2 Objective and contribution ...................................................................................................... 11
   1.3 Outline ......................................................................................................................................... 12
   1.4 Publications ................................................................................................................................. 12
2 Macroscopic traffic flow theory ....................................................................................................... 14
   2.1 Introduction to macroscopic traffic flow theory .................................................................. 14
   2.2 Macroscopic variables ............................................................................................................... 15
      2.2.1 Flow ..................................................................................................................................... 15
      2.2.2 Speed ................................................................................................................................... 15
      2.2.3 Density ................................................................................................................................ 16
   2.3 The fundamental diagram ......................................................................................................... 16
   2.4 Macroscopic traffic flow models ............................................................................................. 18
   2.5 Vehicle trajectories and space-time speed contour plot ....................................................... 20
3 Traffic data collection ........................................................................................................................ 22
   3.1 Traditional methods for real-time traffic data collection ..................................................... 22
   3.2 Emerging and non-traditional methods for traffic data collection ..................................... 22
      3.2.1 Bluetooth ............................................................................................................................ 22
      3.2.2 GNSS-equipped vehicles .................................................................................................. 24
      3.2.3 WiFi ..................................................................................................................................... 25
      3.2.4 Mobile network data ......................................................................................................... 27
      3.2.5 Traffic signal detectors ..................................................................................................... 28
   3.3 Real-time traffic data collected in Stockholm today ............................................................. 28
      3.3.1 Radar detectors .................................................................................................................. 28
      3.3.2 GNSS-equipped vehicles ................................................................................................. 29
      3.3.3 License plate recognition cameras .................................................................................. 30
      3.3.4 Data processing ................................................................................................................. 31
   3.4 Evaluation of travel times collected with Bluetooth in Stockholm .................................... 32
1 INTRODUCTION

As a consequence of the ongoing urbanization around the world the problems with congestion are increasing in almost all large cities. Research has shown that it is not possible to solve the problem by just building new roads since new roads increase the demand and creates new traffic, see for example Hills (1996), Goodwin (1996) and Litman (2004). Building new roads is also a very expensive solution, in particular in urban areas where the amount of land available is very limited and tunnels are the only option. Instead, the use of the existing road network has to be optimized. Besides improving the condition of the available road infrastructure there exists several methods that can be used to achieve a more effective use of the existing road network. These methods are in general referred to as mobility management, which includes various strategies and policies used to influence the choices made by the traveler. Several of the strategies used include some kind of Intelligent Transport Systems (ITS) service. One method aimed to influence the choices made by the travelers is to inform them of the current and future traffic state and alternative modes of transport. Based on this information the traveler can choose to travel at another time, drive another route, use another mode of transport or not travel at all. However, for the provided information to have any impact on the choices made by the traveler, it has to be accurate and relevant. The information can, for example, be distributed to the travelers through variable message signs along the road, through in-vehicle navigation equipment or through a website or a mobile application. All these information services can be labelled as ITS services. Furthermore, in modern traffic management the aim is to be more proactive and an essential part of this is the access to accurate estimations of the current and future traffic state for a large part of the network.

In the ITS Handbook (Miles and Chen, 2004) ITS is defined as “a generic term for the integrated application of communications, control and information technologies to the transport system”. ITS services can, in general, be described as an information chain, as presented in Figure 1. The information chain illustrates the process from collection of data through data processing and all the way to the users. Basically all ITS services require that some kind of data is collected, this process is covered by the first box in the ITS information chain, Data Acquisition. It should be noted that also data from external factors such as weather and events are included in the information chain presented in Figure 1. Once the data is collected, it is transmitted to a data center where it is processed. In the information chain, this process is handled in the box Data Processing and can involve fusion of different data sources and extraction of the requested information. This information is transmitted to the Information Distribution where the extracted information is distributed to the users through an app, a website or variable message signs along the road. The information can also be distributed to a traffic management center and be used as decision support and for traffic control, in the information chain this is covered by the box Information Utilisation. This thesis covers activities related to the boxes Data Acquisition and Data Processing. The boxes Information Distribution and Information Utilisation include indeed very important aspects but they are outside the scope of this thesis.
In parallel with an increasing demand for accurate traffic information with large coverage, more and more traffic data is collected from a variety of sources. Technologies like Bluetooth and WiFi together with GPS-devices, smart phones, mobile network data and more traditional data sources like radar and loop detectors have the potential to create a highly comprehensive traffic database. However, the data available will not automatically improve the accuracy of traffic state estimations since the temporal and spatial resolution as well as the aggregation, accuracy and precision differ substantially. Therefore, models and algorithms that combine heterogeneous data are necessary in order to produce accurate and reliable estimates and predictions of the current and future traffic state.

1.1 Mobile Millennium Stockholm

A large part of the research presented in this thesis has been carried out within the project Mobile Millennium Stockholm. The Swedish Transport Administration aims to create a system that can produce travel time estimations and predictions for the major cities in Sweden and as a part of this ongoing work the Mobile Millennium Stockholm project was initiated in 2010. The project is a collaboration between the Swedish Transport Administration, the Swedish organizations Linköping University, the Royal Institute of Technology in Stockholm and Sweco Society, and the University of California, Berkeley in the United States. The purpose of the Mobile Millennium Stockholm project is to assimilate the knowledge gained from the UC Berkeley Mobile Millennium project and develop new methods for data fusion that can facilitate an adaptation of the system that meets Swedish requirements, see Allström et al. (2011).

The starting point for the Mobile Millennium project at UC Berkeley was the Mobile Century field trial, carried out in February 2008. The objectives of the field trial were to collect data for future research and demonstrate the potential of online real-time data processing, privacy-preservation and data-handling efficiency, see Amin et al. (2010) and Herrera (2009).

The Mobile Century field trial later evolved into the Mobile Millennium project, see Bayen, Butler and Patire (2011). Mobile Millennium was a research project focusing on the design, implementation, and operational deployment of novel algorithms and innovative techniques to address current road traffic challenges. It includes a traffic monitoring system that uses the GPS in mobile phones to gather traffic data, process it, and distribute the traffic state back to the phones.
in real time. The Mobile Millennium project officially ended in December 2010 and evolved into a more production-like system that fuse data from different sources in real-time and estimates the current traffic state for selected parts of the road network in California. The production system runs in real-time while a research/development system is running in parallel. During 2010 and 2011 the Mobile Millennium system was adapted to the traffic data made available in Stockholm and models for estimation of the traffic state on both highways and arterials has been implemented for Stockholm. A model based on kinematic wave theory that captures the backpropagation of traffic jams is used for highways. The model is denoted as the Cell Transmission Model for velocities, the CTM-v, and together with an Ensemble Kalman filter it can combine various point speed measurements and estimate the traffic state. This model is referred to as the MMS-model in this thesis.

1.2 Objective and Contribution

The overall aim of the work presented in this thesis is to improve traffic management and the traffic information provided to the traveler, and thereby optimize the use of the existing road network. In order to achieve this there are several challenges that needs to be addressed, one being the estimation and prediction of the current and future traffic state. The objective of this thesis is to contribute to the development and implementation of the MMS-model and to facilitate an adaptation of the MMS-model to the conditions of the highways in Stockholm. The focus in the first part of the thesis is on the verification, calibration and validation of the MMS-model for a road stretch on the highway in Stockholm. The purpose is to implement a method that can be used to calibrate a traffic model such as the MMS-model and demonstrate which estimation results that are achievable. In the second part of the thesis the focus is on the extension of the MMS-model to also predict the future traffic state. The purpose is to propose and evaluate a hybrid approach for short-term traffic state prediction where the MMS-model is combined with a neural network.

The contributions of this thesis are:

- An overview of available methods for traffic data collection, with a focus on emerging technologies. The strengths and weaknesses of the different methods are described.
- An evaluation of travel times collected with Bluetooth in Stockholm.
- A summary of previous research on assimilation and fusion of traffic data for traffic state estimation.
- A description of the work performed in order to adapt the MMS-model to the conditions of the highway in Stockholm.
- An initial verification of the MMS-model based on other web based traffic information services and collected GPS-data.
- Implementation of a two-stage process for calibrating the parameters related to the implemented MMS-model model and the fundamental diagram. The parameters related to the CTM-v and the EnKF are calibrated using an implemented calibration framework while the parameters related to the fundamental diagram are calibrated using the compass search method.
Calibration of the MMS-model for parts of a highway in Stockholm using the implemented two-stage calibration process. The calibrated model is also validated against travel times collected with Bluetooth.

A survey on previous research on methods for short-term traffic state prediction and how the predicted traffic state can be used.

Implementation and evaluation of a new hybrid approach for short-term traffic state prediction. In the proposed hybrid approach the MMS-model is used to predict the future state based on input from predicted boundary flows and mainline sensor data predicted with a neural network.

1.3 OUTLINE
The outline of the thesis is as follows. Chapter 2 begins with an overview of characteristics of highway traffic and the relationships between the macroscopic state variables speed, flow and density. This is followed by Chapter 3 where traditional methods for collecting traffic data is briefly described followed by a more extensive presentation of emerging and non-traditional methods for traffic data collection and their strengths and weaknesses. Finally, there is an overview of the current situation in Stockholm when it comes to traffic data collection.

In Chapter 4 the focus lies on traffic data processing and in particular data fusion. First, a summary on previous research on fusion of traffic data is presented followed by a more detailed description of the MMS-model. The adaptations of the MMS-model that has been made in the implementation for Stockholm is also presented. Chapter 5 introduces the calibration and validation concept and both the initial verification of the implemented MMS-model and the more extensive two-stage calibration process that has been implemented is described. Finally, the results from the calibration and validation of the implemented MMS-model are presented.

Chapter 6 begins with an overview of methods for short-term traffic state prediction. This is followed by a more detailed description of the method used where the MMS-model and a neural network are combined in order to produce accurate short-term travel time and traffic state predictions. The thesis ends with Chapter 7 where the conclusions are presented together with a discussion on future research.

1.4 PUBLICATIONS
Parts of this thesis have previously been published in the following publications:


2 MACROSCOPIC TRAFFIC FLOW THEORY

In this chapter, macroscopic traffic flow theory is introduced together with the macroscopic traffic variables and their relationship. The chapter includes a brief overview of macroscopic flow models used to model highway traffic. For a more in depth description of macroscopic traffic flow theory the reader is referred to Treiber and Kesting (2013) or Elefteriadou (2014).

2.1 INTRODUCTION TO MACROSCOPIC TRAFFIC FLOW THEORY

Traffic flow theory ties together variables describing the traffic state and the related analytical models. The models and the theory can be categorized based on a number of different perspectives with one of the most common being the aggregation level. Based on aggregation levels there are three main categories of traffic flow models: microscopic, macroscopic and mesoscopic. In the microscopic models the behavior of each driver is modelled individually and influenced by the surrounding traffic. In the macroscopic models aggregated variables are used and the traffic flow is modelled like fluids or gases in motion. Mesoscopic models are hybrids of the microscopic and macroscopic models where the macroscopic modeling approach with microscopic variables is used in some cases, and the other way around for other cases.

The macroscopic traffic flow theory was first introduced in the 1930s by the American professor Bruce D. Greenshields, see Greenshields et al. (1933). Greenshields measured traffic flow, density and speed using a 16 mm Simplex movie camera, see Figure 2. Based on the measurements, a linear relationship between speed and density and the associated parabolic relationship between speed and traffic flow was presented, see Greenshields (1935). More than 80 years has passed since then and during these years the model and the relationship between the parameters have been further developed. During the last decades, when more traffic data and computer power has been made available, a large amount of research has been invested in traffic modeling.

Figure 2: Greenshields with his camera and the results from the measurement in the form of a speed-density diagram (Greenshields, 1935).
2.2 MACROSCOPIC VARIABLES

The basic macroscopic traffic variables, flow, density and speed, reflects the average state of the traffic. The macroscopic data consists of aggregated microscopic data, i.e. aggregated individual vehicle data. The data is in general aggregated over fixed time intervals with 60 seconds or five minutes being the most common. However, some detector systems have the functionality to deliver vehicle-by-vehicle data, which is to prefer since some information might be lost in the aggregation process. In the remainder of this chapter, the macroscopic traffic variables and their relationships are further described together with an overview of different macroscopic flow models.

2.2.1 Flow

The flow $Q$ is the number of vehicles $n$ that pass a point $x$ during a specified time interval $\Delta t$ starting at time $t$. This is expressed as

$$Q(x,t) = \frac{n}{\Delta t},$$  \hspace{1cm} (2.1)

In order to use this expression, the flow must be measured over time at a specific point and cannot be obtained from a single snapshot of a certain length of a road.

For real-time applications the flow data is often aggregated over one or a couple of minutes before they are transmitted to the traffic management center. For offline planning purposes, the daily flow or the flow during a specific hour is normally used. The maximum flow rate of a road is denoted as the capacity of that road. Depending on road design, number of lanes, speed limit, vehicle composition and other parameters the capacity of a motorway normally lies between 1800 and 2300 vehicles per hour per lane, see Trafikverket (2013).

2.2.2 Speed

The average speed can be computed in two different ways; as time mean speed (arithmetic mean speed) or space mean speed (harmonic mean speed).

The time mean speed $V_t$ is the average speed of the vehicles that passes a point during a specific time interval

$$V_t(x,t) = \frac{1}{n} \sum_{i=1}^{n} v_i,$$  \hspace{1cm} (2.2)

where $v_i$ is the speed of vehicle $i$, $n$ is the number of vehicles and $x$ and $t$ are the position and time for which the speed is calculated.

The space mean speed $V_s$ is based on the average time it takes to travel a given distance or the average speed at a specific time instant, which makes it very difficult to measure. However, Wardrop (1952) and others have shown that the space mean speed is equivalent to using the harmonic mean of the individual vehicle speeds $v_i$ for $n$ vehicles that passes the point $x$ during a specific time interval.
Furthermore, in Van Lindt (2004) the relationship between time mean speed and space mean speed is defined as

\[ V_t = \frac{\sigma_s^2}{V_s} + V_s, \]  

(2.4)

where \( \sigma_s \) is the variance of the space mean speed. It should be noted that it is not possible to calculate the space mean speed from time speed measurements based on this relationship since the variance of the space mean speed in general is unknown. However, according to Equation 2.4 is \( V_s \leq V_t \) and empirical studies has shown that the difference can sometimes be up to a factor four. \( V_s = V_t \) only holds when all individual speeds are equal. During free flow, when the speed of the vehicles in general is homogenous this difference is small. But since the time mean speed (arithmetic mean) overestimates the influence of faster vehicles the difference between the two mean speeds are larger when there is a large variability of speeds, for example when the traffic changes state from free flow to congestion. In the continuation of this report, speed will be denoted \( V \) and it refer to the space mean speed.

2.2.3 Density

The density \( \rho \), also described as concentration, is the third of the basic macroscopic variables. It is defined as the number of vehicles \( n \) occupying a specific length of the road \( \Delta x \) at a certain time instant \( t \). The density can also be defined using the headways of the vehicles at the specific road stretch according to

\[ \rho(x, t) = \frac{n}{\Delta x} = \frac{1}{\bar{h}}, \]  

(2.5)

where \( h \) is the average headway. In contrary to flow and speed, the density is an instantaneous variable and computed at a certain time instant, not over a specific time interval. This makes it difficult to measure since it requires observation of the entire road stretch at a certain time instant. Instead it is often estimated using the hydrodynamic relationship, also known as the continuity equation, discovered by Greenshields (1935). The continuity equation states that density equals flow divided by speed and it is used to associate the instantaneous variable density with flow and speed according to

\[ \rho(x, t) = \frac{Q(x, t)}{V(x, t)}. \]  

(2.6)

2.3 The fundamental diagram

The graph describing the relationship between the three macroscopic variables flow, speed and density is called the fundamental diagram. From this diagram a number of different quantities illustrating the traffic state can be derived.
As mentioned in Section 2.1, Greenshields (1935) presented a linear relationship between speed and density and the associated parabolic relationship between speed and traffic flow that is valid for traffic at highways. This relationship is also applicable for fluids, gas and other matter moving in a closed environment.

Looking at the quantities that can be derived from the fundamental diagram of Greenshields we find that the speed at maximum capacity is half the maximum speed and the density at maximum capacity is half the maximum density, see Diagram I in Figure 3. Even though Greenshields only based this relationship on seven observations and it is a simplification of observed traffic behavior, it is still used because of its simplicity. The relationship between the variables still holds, at least under certain conditions, while the shape of the fundamental diagram has been developed during the years. One of the most used formulations is the triangular diagram introduced by Daganzo (1994) and Newell (1993). In the triangular fundamental diagram the mean speed equals the maximum speed for all traffic states that have densities smaller than the critical density, see Diagram II in Figure 3.

In Figure 3 the fundamental diagram of Greenshields and the triangular fundamental diagram are presented together with empirical data from a highway section in Stockholm. As can be seen the measurements are noisy, in particular the flow measurements during congestion. The measurements could include systematic errors, but it could also be a case of a large variation in time-gaps and non-equilibrium traffic dynamics which occurs during congestion. Thus, it is important to distinguish between measured data and the fundamental diagram which is a theoretical representation used for traffic modeling.

![Figure 3: The fundamental diagram of Greenshields (I) and Daganzo-Newell (II) together with empirical data from a section of a highway in Stockholm. Density on the x-axis and flow (bottom) and speed (top) on the y-axis.](image-url)

Figure 3 presents the most important characteristics of the fundamental diagram. Those are the capacity $q_{\text{max}}$, the critical density $\rho_{\text{cr}}$, the critical speed $v_{\text{cr}}$, the jam density $\rho_{\text{max}}$ and the free flow...
speed $v_f$. The slope of the curve in the flow-density diagram where the density is higher than the critical density corresponds to the backward propagating speed $w_f$. The backward propagating speed $w_f$ describes at what speed a traffic jam propagates backwards in the network. The figure also visualizes the difference between the free flow state ($\rho < \rho_c$) and the congested state ($\rho > \rho_c$). Given the relationship between the three variables it is also possible to plot the flow versus the speed. However, this is not as demonstrative and fundamental for traffic flow modeling as the flow-density and speed-density diagram.

The fundamental diagram assumes that, under similar conditions, drivers will behave in a similar way. However, the behavior is of course dependent on speed limit, road characteristics, weather and so on. Hence, using different fundamental diagrams for different external conditions will improve the correspondence between the fundamental diagram and measured data.

2.4 MACROSCOPIC TRAFFIC FLOW MODELS

The continuity equation (Equation 2.6) is the foundation for all macroscopic traffic flow models. From the continuity equation and the fundamental diagram, we can derive the traffic state at a given time and calculate the remaining variables from a given value of one of the macroscopic variables. Also, if traffic would be stationary and homogenous these values would be valid for the entire road segment and for a certain time period. Unfortunately, traffic is not stationary or homogenous, it is indeed very dynamic. If we want to model how the traffic evolves over time and space, we have to use a traffic flow model.

In the 1950s, Lighthill and Whitham (1955) and Richards (1956) proposed a macroscopic traffic model based on the conservation of vehicles. Fundamental in the theory of Lighthill-Whitham-Richards (LWR) is a partial differential equation known as the LWR PDE. The LWR PDE expresses how the density $\rho$, given an initial condition and boundary conditions, evolves over a road stretch with a certain length $L$ and over a certain time period $T$ as:

$$
\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial Q(\rho(x,t))}{\partial x} = 0, \quad (x,t)\in (0,L) \times (0,T),
$$

$$
\rho(x,0) = \rho_0(x), \quad \rho(0,t) = \rho_L(x), \quad \rho(L,t) = \rho_r(x).
$$

A basic condition for the LWR PDE is the static relationship, presented earlier as the fundamental diagram, which states that the flow $Q$ can be expressed as a function of the density

$$
Q(\rho) = \rho V(\rho).
$$

During the years, a number of different models based on this theory has been proposed with different fundamental diagrams and mathematical representation. This collection of models is referred to as LWR models, but since they only include one dynamic equation and variable they are also referred to as first-order traffic models.

One of the most used first-order models is based on the triangular fundamental diagram presented previously. This model can be formulated both as a continuous and a discrete model. The continuous version (Equation 2.7) is difficult to solve analytically, especially if the modelled network is large and consist of several links. For larger networks and real-time applications a
discrete model, where both space and time are discretized, are used. Space is discretized into cells with space step \( \Delta x \) and indexed by \( i \) while time is discretized into time steps \( \Delta t \) indexed by \( n \). The discretization of the LWR PDE was first introduced by Daganzo (1994) and was named the Cell Transmission Model (CTM). In the CTM the discretization can be done using a Godunov scheme presented by Godunov (1959) as

\[
\rho_i^{n+1} = \rho_i^n - \frac{\Delta t}{\Delta x} \left( G(\rho_i^n, \rho_{i+1}^n) - G(\rho_{i-1}^n, \rho_i^n) \right),
\]

(2.9)

where \( G(\rho_1, \rho_2) = \begin{cases} Q(\rho_2) & \text{if } \rho_{cr} \leq \rho_2 \leq \rho_1, \\ Q(\rho_{cr}) & \text{if } \rho_2 \leq \rho_{cr} \leq \rho_1, \\ \min\left(Q(\rho_1), Q(\rho_2)\right) & \text{if } \rho_1 \leq \rho_2 \leq \rho_{cr}. \end{cases} \)

(2.10)

\( G(\cdot) \) denotes the Godunov scheme, the function \( Q(\cdot) \) denotes the fundamental diagram and \( \rho_1, \rho_2 \) and \( \rho_c \) denotes the upstream density, the downstream density and the critical density respectively. The cell for which the density is calculated in Equation 2.9 and the cell downstream and upstream are denoted \( i, i+1 \) and \( i-1 \) respectively, see Figure 4.

Figure 4: Graphical representation of the discretized network where \( I \) denotes the cell for which the density is calculated.

Equation 2.10 can be graphically represented as shown in Figure 5.

Figure 5: Graphical representation of Equation 2.10.
To ensure numerical stability, the time step $\Delta t$ and the cell size $\Delta x$ has to be chosen so that the condition

$$v_{\text{max}} \frac{\Delta t}{\Delta x} \leq 1,$$

is satisfied.

The first-order models are well suited for modeling of congestion due to bottlenecks with insufficient capacity and the propagation of this congestion. However, due to the static relationships between the speed and density/flow these models have trouble capturing traffic instabilities and capacity drop phenomena that can occur at traffic breakdown. To solve this problem higher order models with two or more dynamic variables has been proposed. See for example Payne (1971) where two state variables, density and speed, were proposed and Helbing (1996) where a third state variable, the variance of the speed, was introduced. However, these models also have their drawbacks and Kerner (2004) has shown that there sometimes occur situations that are difficult to capture with the fundamental diagram. As a solution to this, Kerner (2004) introduced the three-phase-traffic theory where the three phases are: free flow, synchronized flow and congestion. Microscopic traffic models have been developed on this theory although it has been shown that it also sometimes fails to fit and explain empirical data, see for example Schönhof and Helbing (2009).

There is an ongoing discussion in the traffic flow theory community on which model that is the best and most suitable to use for different applications and scenarios.

2.5 VEHICLE TRAJECTORIES AND SPACE-TIME SPEED CONTOUR PLOT

So far, only the macroscopic variables and their relationship have been presented. Traditionally, the macroscopic traffic flow theory has been based on point measurements of the macroscopic variables and their static relationship. However, there is a trend within traffic data collection towards more vehicle based data collection. Hence, an introduction to microscopic variables and vehicle trajectories provide a relevant context for the continuation of this thesis.

With the introduction of the GPS in the 1990s and other positioning methods used to measure and monitor the traffic state a new type of traffic data has been introduced. These data are often denoted as microscopic variables and include detailed information like the headway between vehicles, acceleration and deceleration, lane changes and speed profiles of individual vehicles.

The position of a vehicle measured over time is called a trajectory and is normally visualized in a space-time diagram, see Figure 6. If one would for example collect point measurements at location A and/or B this would only provide information about the traffic state at these certain positions along the y-axis. In contrary, the trajectory provides detailed information about a certain vehicle along a studied road stretch. From the slope at a certain position and time we get the instantaneous speed and the travel time between two points easily can be calculated.
A space-time speed contour plot is also a convenient way to visualize how the traffic state evolves over time. In a space-time speed contour plot the studied time horizon is discretized into certain time steps and the road stretch into cells. The aggregated speed in each cell for each time step is visualized using a color scheme. In Figure 7 green indicates speeds closer to the free flow speed while red indicates congestion and low speeds. In such a plot one can for example see how a traffic jam propagates backwards over time by studying the shape of the red area. The slope of the upstream front of the red area corresponds to the backward propagating speed $w_f$. 
3 Traffic data collection

This chapter begins with a brief overview of traditional methods for collecting traffic data. This is followed by a more detailed presentation of emerging techniques for collecting traffic data and the current status in Stockholm when it comes to traffic data collection and data processing.

3.1 Traditional methods for real-time traffic data collection

As mentioned in the previous chapter, the traditional macroscopic traffic flow theory is mainly based on speed and/or flow data collected at certain positions. The methods used to collect such data are mainly inductive loops, magnetometers and radar detectors. Both inducted loops and magnetometers are embedded in the roadway while the radar detectors are mounted over or next to the roadway. These methods detect basically all passing vehicles and are in most cases installed to deliver aggregated data in real-time. As concluded by, for example, Marti et al. (2014) the accuracy of these methods is, in general, very good. The disadvantage is that they only collect data at the positions where they are installed; they provide no knowledge of the traffic state between the sensors. In addition, they are in general expensive to install, maintain and have difficulties detecting vehicles when speeds are low during periods of heavy congestion.

For the collection of travel times License Plate Recognition cameras (LPR), also known as Automatic Number Plate Recognition cameras (ANPR), have been used for a number of years. It is a method where cameras that capture the license plate of passing vehicles are installed and it is widely used in different types of toll systems. If a license plate is captured by two or more cameras the time stamp could be used to determine the travel times between the locations where the cameras are installed. If correctly installed and calibrated this method captures the travel time of the vast majority of the vehicles. However, the number of outliers, i.e. measurements that for various reasons should not be included in the aggregation, is in general high.

3.2 Emerging and non-traditional methods for traffic data collection

With the ongoing digitalization of the society and in particular the introduction of smartphones, the area of real-time traffic data collection has been revolutionized. Wireless communication often includes some kind of positioning technique or at least the possibility to detect the presence of a unit used for some kind of wireless communication. Since these methods often use existing systems and infrastructure they are in general cost effective and easier to implement than traditional methods. However, unlike the traditional methods they are not originally developed to collect traffic data and the collected data often has to be filtered.

In this section a number of emerging and non-traditional methods for traffic data collection is presented.

3.2.1 Bluetooth

Bluetooth is a global standard protocol used for exchanging data wirelessly over short distances. Bluetooth is, for example, used in hand free devices and for communication between different devices in vehicles. The standard is based on an electronic identifier in each device called a Media Access Control (MAC) address. Each MAC address is unique and serves as an electronic nickname so that electronic devices can keep track of who is who during data communications.
The MAC address consists of a public part that can be captured and used for estimating travel times. The public part of the MAC address consists of three parts. The first and second part is allocated to the manufacturer (Sony, Apple, Garmin, etc.) and the type of device (smart phone, hands free, navigation device etc.). While the last part is a unique 48-bit address that makes it possible to match consecutive captures of the MAC address and estimate travel times between two Bluetooth sensor locations. Imagine, for example, that a vehicle equipped with an active Bluetooth device is driving along a road and is detected by a sensor at location A at time \( t_1 \). After driving a certain distance the device is detected again by a sensor at location B at time \( t_2 \). The travel time for the specific route can then easily be calculated, given that the route is known. As with all type of data collected by re-identification, the collected data has to be processed before being used and outliers have to be removed.

The individual user cannot be traced using the MAC address since neither the manufacturer nor the retailer keep records or a database linking a particular MAC address to a specific user. Thus, the only way a device can be tracked to a specific person is to have previous knowledge of whose device it is and be allowed to view their MAC address. Furthermore, the MAC address is normally encrypted by the Bluetooth sensors and this encryption is changed every day which prevent the possibility to track a device over consecutive days.

Not all vehicles are equipped with active Bluetooth devices which mean that only the travel time of a certain share of the vehicle fleet can be captured. In the literature this figure varies between 2 % and 20 %. In the US, where this technology is used to collect travel times at a number of different places in a number of different cities, the penetrations rates vary between 2 % and 10 %, see for example Haghani et al. (2010), Porter et al. (2011) and Malinovskiy et al. (2010). In Europe, conducted field trials indicate a higher penetration rate, between 20 % and 30 %, see Barcelo et al. (2010) and Lahrmann et al. (2010).

One thing that might affect the travel time estimation is the fact that the process of capturing the MAC address takes about 5 seconds on average but may take up to 10 seconds in some extreme cases, as described by Huang and Rudolph (2007). However, a number of different field trials have reached the conclusion that travel times estimated from Bluetooth sensors deployed along a motorway or arterial are comparable to those estimated from GPS and toll tag readers, see for example Haghani et al. (2010), Porter et al. (2010) and Kim et al. (2011). Malinovskiy et al. (2010) compared Bluetooth data with data from a license plate recognition system and discovered that despite that the sample size obtained from Bluetooth is significantly smaller, the estimated travel times are still representative of the actual conditions. As described by Porter et al. (2010), it is possible to adjust the area where Bluetooth sensors can capture active Bluetooth devices. A smaller area, which also means a lower number of captured vehicles, decreases the error in the travel time estimate.

A field trial conducted in Copenhagen by Lahrmann et al. (2010) showed that around 20 % of the vehicles were equipped with a Bluetooth device. Furthermore, it was demonstrated that one Bluetooth sensor could cover a motorway with two lanes in each direction which makes it a very cost effective way of collecting data. However, the travel times estimated from Bluetooth were around 10 % higher than those estimated from GPS data, though it is not clear if any filter was used.
To summarize the Bluetooth technology, we can conclude that it is a cost effective method compared to traditional methods used for collecting travel times, like ANPR. The Bluetooth sensors are easy to install and even though not all vehicles are detected the estimated travel times are in most cases representative of the actual conditions when outliers have been removed. The results from a field trial conducted in Sweden are presented in Section 3.4.

3.2.2 GNSS-EQUIPPED VEHICLES

Using satellites for positioning was first introduced in the 1990s but the use of this technique has increased rapidly with the introduction of the smartphone.

As of today, there exists two operational Global Navigation Satellite Systems (GNSS): the American system GPS and the Russian GLONASS. GPS, or Global Positioning System, is the most used and has been somewhat synonym to GNSS in general. The regional Chinese GNSS Beidou is planned to be global by 2020 while the European GNSS Galileo is in the first deployment phase and expect to be operational at the earliest by 2020. For a detailed description of the technical aspects of GNSS positioning see Küpper (2005).

For traffic data collection GNSS based positioning are primarily used in two different ways. It could either be used for positioning in a smartphone application or in a dedicated navigation device. In the literature, a vehicle whose position can be tracked is called a vehicle probe and the data collected is denoted as floating car data (FCD). The basic data that is collected from a GNSS-equipped device is positions with a time stamp and in most cases the data also include id, speed and direction. However, the quality of the positioning data from a GNSS-equipped device can vary a lot and depends on four main properties:

- Penetration level
- Sampling strategy
- Measurement type
- Measurement accuracy

The penetration level is the number of equipped vehicles compared to the total number of vehicles in the area of interest. The sampling strategy refers to the frequency with which the position of the probe is logged and how often these are sent to the traffic information server. The measurement type is mainly related to the type of sensors that the probe is equipped with, e.g. GPS or accelerometer, but also what kind of data that is supported by the transmission protocol. The measurement accuracy refers to which accuracy that can be achieved in each measurement, e.g. positioning accuracy. The positioning accuracy from GNSS is lower in urban areas with tall buildings and the system only works well in open air. In tunnels the GNSS will not provide a useful position estimate. The importance of the different properties depends on what type of traffic state variable it is used to estimate, but also what kind of traffic application that it is intended for.

The device type, i.e. whether the data is collected in a fleet operational system, a public smartphone app or some other device, will also affect the characteristics of the data that is collected and might introduce bias in the traffic estimations. It is, for example, very likely that a navigation device is located in a vehicle, which is not necessarily true for a smartphone. Also, some devices have poor GNSS-units and the location accuracy reported from these devices can potentially affect the results.
of traffic state estimations. Furthermore, some probe client types, e.g. taxis, can use dedicated bus/taxi lanes and hence indicate lower travel times than for regular vehicles. Another possible problem is that some client types, e.g. heavy vehicle fleet management clients and buses, are more restricted in terms of speed limit.

As presented earlier the very basic information provided by a GNSS-unit is the position and a time stamp. Given that an id is also collected and maintained this data can be used to estimate travel times. If the time between two consecutive logged positions for a specific probe is long, the path inference problem becomes more difficult, this is especially problematic in dense urban areas with a large number of possible routes. Map matching and solutions to the path inference problem for devices with a low sampling frequency are further described in Hunter et al. (2012) and Rahmani and Koutsopoulos (2013). Furthermore, a low sampling frequency also introduces a delay for when travel times can be calculated.

There is no doubt that GNSS-equipped devices can provide accurate positioning data that can be used for traffic state estimations. In many research studies this data is used as ground truth, but the data availability can be a problem. Floating car data is mainly collected by commercial actors like haulage contractors or other professional drivers. Historically these companies have been reluctant to share this data, partly because the data contain sensitive information about the routes used and their customers. However, the public authorities have lately realized the potential of this data and are now willing to pay money for the data. This has of course been a good incentive for the commercial actors to process the data and make it available and there is now GNSS based traffic data available on the market. In Sweden there is an ongoing evaluation of GNSS based travel times provided by a number of different companies.

The introduction of GNSS-equipped mobile phones has opened up a new and potentially much larger group of probe vehicles. Today the vast majority of the population owns a GNSS-equipped smartphone and several applications for collecting traffic data is available. The concept of using data collected by the population is called crowdsourcing and one of the most well-known examples of a crowdsourcing based system is Waze. Waze is a free GPS based turn-by-turn navigation application for smartphones where the users both build the digital network, share their GPS-data and report traffic problems. It was released in 2009 and has grown rapidly since then. In July 2012 the company announced that they had passed 20 million users worldwide and when Google bought Waze in November 2013 the number of user was almost 50 million. How many of these that are collecting data in Sweden is not official but the congestion that occurs in the big cities is captured in the real-time information provided by Waze. The crowdsourcing based systems might have a larger penetration rate than the data from commercial actors for parts of the network, but the data is in general not available for researchers or governments.

### 3.2.3 WiFi

WiFi is a wireless networking technology that uses radio waves to provide wireless high-speed Internet and network connections. The WiFi technology can be used for collecting traffic data in two different ways and both travel times and positions of individual devices can be derived. The communication standard for WiFi is, just like for Bluetooth communication, based on the MAC address. Hence, fixed sensors that detect passing devices with active WiFi can be installed along the road and used to collect travel times, see Musa and Eriksson (2012). Abbott-Jard et al. (2013)
concluded that by combining sensors that detect the MAC-address of both Bluetooth and WiFi devices the sample rate can be increased and the travel time estimation become more accurate.

Given that a device has an active WiFi connection the positioning of the device could be done using WiFi, if the WiFi hotspot or access point that the device is connected to have been mapped. In the US for example, there are companies that sell large databases of WiFi access points and their location. Such a database can rather easily be created from people moving around in the city with a smartphone application. This data may then be used to estimate a user's position. However, the database has to be maintained an updated regularly and the positioning does of course not work when the device is out of range of WiFi signals. When map matched, the estimated positions could be used to estimate travel times. In contrary to the method where a fixed sensor detecting the MAC address is used, travel times from WiFi positioning can only be calculated within a device, a smartphone for example. The consequence of this is that the generated travel times will be spread randomly over the network and not only over certain sections where sensors has been installed. The availability of this data is also an issue. Just like position data collected with GNSS-equipped devices this position data is only available for the user, unless it is shared through some application.

The major advantage with WiFi positioning is the low energy consumption which extends the battery time of the smart phone compared to the GNSS positioning that previously often has been used. According to Thiagarajan et al. (2010) sampling GPS consumes up to 20 times more energy than sampling WiFi. The coverage of where WiFi positioning could be used is generally very good in urban areas but not as good in more rural areas. WiFi typically producing location estimates within a mean radius of about 40 meters of the true location. However, in urban areas where the access points are densely spaced and overlapping each other, a better accuracy can be obtained. If the position is map matched and possibly combined with mobile network positioning, described in Section 3.2.4, and sparse GNSS data the accuracy can be improved even further. Another advantage with WiFi positioning is that it works well in a wide range of conditions, including tall buildings and indoors, when GNSS signals may be weak.

Thiagarajan et al. (2010) showed that it is possible to achieve good position accuracy using WiFi localization. A Hidden Markov Model (HMM)-based map matching scheme was used together with a travel time estimation method that interpolates sparse data to identify the most probable road segments driven by the user and to attribute travel times to those segments. In the conducted field trial in Boston, WiFi performs better than GPS sampled every 60 seconds and worse than GPS sampled every 30 second.

Athanasiou et al. (2009) demonstrates that for high sampling frequencies, travel times derived from WiFi positioning are comparable to travel times derived from GPS in absolute terms. Data was collected through wardriving in an urban area in Athens where 2184 unique WiFi access points were discovered, which corresponds to 2.1 access points per 100 m$^2$. Combining WiFi positioning with map matching reduced the average error to approximately 20 m which is very close to the average error of GPS in urban environments (5-15 meters). The evaluation demonstrated that the produced travel times are practically identical to the ones derived from GPS data. Furthermore, when applying a typical speed profile classification on travel times, even for sampling rates of up to 30 seconds, the produced travel times are still of respectable quality.
To summarize this section on WiFi, it can be concluded that it is a widely spread technique that can be used both for positioning and travel time estimation from fixed sensors. Used for positioning it requires knowledge about the location of hotspots but it doesn’t drain the battery as much as the GPS and the accuracy of the positions are comparable to those of GPS in urban areas.

3.2.4 MOBILE NETWORK DATA

The mobile operators always have to keep track of where each individual mobile phone is located to be able to connect phone calls and optimize the use of their network. This type of data is of course interesting for traffic planners to use. The first project where positioning data from the mobile network was used for traffic estimations was the CAPTIAL project conducted by University of Maryland Transportation Studies Center (1997). The results from that project were not that successful due to poor location accuracy but since then a lot has happened. Both the available data, the methods to process the data and the data power available have developed fast and today there are companies that offer real-time travel times and travel pattern analysis based on positioning data from the mobile network. Several research projects have also demonstrated the potential for this data source, see for example Steenbruggen et al. (2011) and Valerio et al. (2009). Both Bar-Gera et al. (2007) and Gundlegård and Karlsson (2009) have showed that positioning data from the mobile network can be used for travel time estimation, both in urban areas and along motorways. Research has also been done on travel pattern analysis and OD-matrix estimation based on mobile network data with promising results; see for example Caceres et al. (2007) and Wang et al. (2012).

The accuracy of the positioning data obtained from the mobile network depends on where in the network the data is captured, how the mobile phone itself is used and the design of the network. A mobile phone that is active in a phone call or transmitting or receiving data generates the most detailed positioning data. However, mobile phones that are not actively used also generate positioning data in the mobile network but with lower temporal and spatial accuracy. A more detailed description on the data available in mobile networks for travel time and travel pattern analysis can be found in Küpper (2005) and Gundlegård and Karlsson (2009).

A lot of research in this area has been based on the most basic positioning data in the mobile network, that is the position and time stamp from when a call has been placed or a SMS or MMS has been sent i.e. data that is used for billing purposes in the operator’s network. The mobile network operator Orange has during both 2013 and 2014 made such a data set available for researchers. This has further increased the research activity in this area since the data availability has been the major issue for many research organizations.

There is no doubt that mobile network data has great potential, mainly for OD-matrix estimation but also for travel time estimation. However, there are some issues that need to be resolved before this potential can be fulfilled. The most obvious obstacle is the integrity aspect. Similar to many other data sources there exist a trade-off between accuracy and integrity and the challenge is to find the right level of aggregation, besides the standard data protection. Another obstacle has been the data availability, which is partially linked to the integrity aspect. Even though there exist commercial actors that use positioning data from mobile networks in some countries, data from Sweden has just recently been made available for research.
3.2.5 Traffic Signal Detectors

Basically all modern traffic signals in Sweden are equipped with a detector. The traditional type of detector is an inductive loop embedded in the roadway while the new generation of detectors consists of small wireless magnetometers. Theoretically, it would be possible to extract traffic flow, speed and travel time from these detectors. Inductive loops are used to collect speed and flow on highways in many cities around the world, in particular in the US. Unfortunately, the inductive loops connected to the traffic signals and the overall traffic signal system is in general not designed to collect this type of data. Only in Stockholm there are almost 600 traffic signals and the potential when it comes to data collection from traffic signals is very large. As a consequence of this, a lot of research and development has been carried out during the last years, especially when it comes to travel time estimation. When a vehicle traverses an inductive loop or a magnetometer the inductance changes, also referred to as the magnitude, produce an analog waveform output called the vehicle signature. The signature depends on the composition of the vehicle and by matching signatures from two detectors travel times can be estimated. Research has shown that, depending on the driving environment and type of detector, matching rates between 60 % and 100 % can be achieved, see Blokpoel (2009), Ndoye et al. (2011) and Jeng et al. (2010). Since acceleration and retardation affects the signature the matching process is more difficult for urban areas with a lot of stop and start movement.

Using existing loop detectors and matching the vehicle signature has great potential to provide the traffic analyst with travel times and possibly also flow measurements. This method is not based on wireless communication like the other methods presented, but it is still a very cost-effective method that would provide valuable traffic data.

3.3 Real-time Traffic Data Collected in Stockholm Today

The main traffic data sources in Stockholm are traditional radar detectors and GNSS-equipped vehicles, mainly taxis, together with a system of license plate cameras used to estimate travel times for a number of routes around Stockholm. In this section the characteristics of the data sources and how the data is processed are briefly described. For more detailed information about the data processing see Strömgren et al. (2004) and Vägverket (2006).

3.3.1 Radar Detectors

There are around 1000 radar detectors in the Stockholm network, primarily located on highways. The radar detectors are an integral part of a Motorway Control System (MCS) on the main highway that passes through Stockholm, see Figure 8. These radar detectors are situated approximately every 500 meters and collect data specific to each traffic lane. They collect speed and traffic flow data that is aggregated over one-minute periods. The MCS has recently been extended to the motorway going south from Stockholm to Södertälje and further extension of the system will follow in the coming years.
3.3.2 GNSS-EQUIPPED VEHICLES

There are around 1500 taxis operating as probe vehicles in the Stockholm area. The data reported includes an anonymous id, time-stamped location (latitude and longitude) and taxi-status (free or occupied). Different time-based or distance-based events trigger a transmission of the current position of a taxi to the dispatch central. If a taxi is caught in congestion and moves slowly the generation of position data is time-based, while it is distance-based if it is moves faster. The sampling frequency is approximately 110 seconds on average, but varies a lot. Even though the taxis are supposed to report their location at least every minute, the actual frequency deviates from what is expected, see Allström et al. (2011).

The instant speed and heading of a vehicle is not included in the data reported by the taxis. The reason for this is that the fleet dispatching systems is not dependent of this data and by excluding it, the cost of data transmission is reduced. Due to this, the initial processing of this data includes a path inference filter, besides the map matching and filtering of outliers. A long time interval between consecutive probe updates complicates the path inference process and reduces the quality of the filtered data. For more details on the path inference problem see Rahmani and Koutsopoulos (2013) and Hunter et al. (2012).

In Figure 9 all reported locations from taxis in Stockholm during a five-minute period are visualized. From the figure it can be noticed that the majority of the taxis are located in the central part of the city and that the roads going to the two airports north and northwest of the city center is well covered by probe data from taxis.
3.3.3 License Plate Recognition Cameras

Until the later part of 2015, travel times derived from license plate recognition (LPR) cameras were collected for approximately 100 routes in and around Stockholm, see Figure 10. The cameras were installed at locations on important routes throughout the city, and the data was reported for each individual vehicle.

A problem with the LPR data is that the travel time data is reported after passing the camera at the end of a measured route. Thus, the data is already old making it hard to detect congestion situations in real-time. In addition, the actual route is unknown, unless the route between the cameras does not include any intersecting streets. Nevertheless, this data has been extremely useful for validation purposes and for long-term evaluations of how the travel times evolve over time.
There is an ongoing discussion on how the terminated system with LPR-cameras should be replaced. Bluetooth sensors and aggregated travel times from floating car data bought from commercial actors are the two most probable alternatives. Travel times from Bluetooth sensors have successfully been evaluated in Stockholm (see Section 3.4) and are currently used for parts of the highway network where the MCS-system has been temporarily taken down during construction work. A system where travel times based on floating car data from commercial actors is aggregated and delivered to the traffic management center is currently being evaluated in Gothenburg and Stockholm.

3.3.4 DATA PROCESSING
The data processing carried out in the current travel time system implemented at the traffic management center in Stockholm is relatively unsophisticated. A simple data fusion algorithm is applied for links where travel times are calculated from two or more data sources. In the algorithm, a quality value allocated to each measurement is used for weighting purposes. For links where no real-time data is available, the travel time is estimated from historical data as well as real-time data from neighboring links, see Vägverket (2004). The links for which the traffic state is estimated is generally quite long and the estimations are generated only once every five minutes. Currently, there are no predictions of travel time in Stockholm.

Besides the real-time calculations of travel time, travel time calculations are also made by services that do not operate in real-time. These services make use of a historical database with indicators.
that describe the congestion level in Stockholm over a longer time-period. The historical database contains all traffic data collected since 2004.

3.4 Evaluation of travel times collected with Bluetooth in Stockholm

As presented in Section 3.2.1, Bluetooth is used for collecting travel times at many places around the world. However, prior to the evaluation we present here there had been no trials with Bluetooth sensors made in Sweden. The purpose of the field trial presented here, was to determine the penetration rate of Bluetooth devices and the validity of the travel times collected with Bluetooth sensors. The results have previously been presented in Allström et al. (2012c). In the field trial, two Bluetooth sensors that detects all active Bluetooth devices that passes the between Alviksplan and Brommaplan in Stockholm has been installed. The route for which travel times has been collected is presented in Figure 11. The route is 1.7 km long, and has two intersections with traffic lights. The posted speed limit varies between 50 km/h and 70 km/h and the number of lanes are two in each direction.

![Figure 11: The road stretch where travel times collected with Bluetooth has been evaluated. © Lantmäteriet/ Metria.](image)

On this route, permanent license plate recognition cameras (LPR) is already installed and continuously collect travel times. However, the LPR cameras only cover one lane in each direction. To be able to determine the penetration rate of Bluetooth devices portable LPR cameras which covered both lanes in the west-bound direction were mounted in the beginning and end of the road stretch. These portable LPR cameras was used to both count the number of vehicles and collect ground truth travel times. This data is in the continuation of this chapter denoted as evaluation data.

The analysis of the collected data showed that the penetration rate of the Bluetooth sensors were around 15 % while the permanent LPR-cameras captured around 45 % of the vehicles passing the two sensor locations, see Table 1. However, the permanent LPR-cameras only covered one lane while the Bluetooth sensors covered both lanes. All travel times longer than 17 min were removed.
Even though the penetration rate is interesting and a valuable input to the analysis of the collected data, the accuracy of the collected travel times is what really matters. In Figure 12 and Figure 13, each travel time collected with Bluetooth and the permanent LPR cameras respectively, has been plotted together with the evaluation data. The time of day given in the figure is when a vehicle exits the studied road stretch and a travel time can be calculated. Figure 12 shows the afternoon peak between 16 and 17 when the average travel time is around 8-9 minutes while Figure 13 shows the collected travel times between 13 and 14 when there is less traffic and the average travel time is around 3 minutes. From the figures, it can be concluded that the travel times collected with Bluetooth correspond well with the evaluation data and the travel times from the permanent LPR cameras. However, in both figures, it is noticeable that outliers exist and the collected data needs to be processed.

Figure 12: All travel times collected with each method between 16 and 17.

Table 1: Penetration rate for each data collection for two different time periods.

<table>
<thead>
<tr>
<th></th>
<th>13-14</th>
<th>16-17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Permanent LPR</td>
<td>45 %</td>
<td>44 %</td>
</tr>
<tr>
<td>Bluetooth</td>
<td>14 %</td>
<td>16 %</td>
</tr>
</tbody>
</table>
Besides looking at the raw data an aggregation of the collected travel times has been made. To aggregate the data is a simple process to reduce the impact of outliers. A harmonic mean value over five-minute periods has been calculated and is presented in Figure 14 and Figure 15.
Figure 15: Travel times collected with each method between 13 and 14 aggregated over five-minute periods.

In Figure 12 there are two outliers in the Bluetooth data around 16.30 and 16.35 which has a large impact when the data is aggregated, see Figure 14. This further confirms that the lower penetration rate that Bluetooth sensors implies requires some kind filtering process.

From the figures it can further be concluded that the aggregated travel times collected with each method follows the same pattern. However, it can also be noticed that the aggregated travel times collected with Bluetooth in general is a bit lower than the evaluation data. This is even more obvious when the RMSE (Root Mean Square Error) and MPE (Mean Percentage Error) is studied in Table 2. The RMSE, which is an aggregate measure of the accuracy, is calculated as

$$RMSE = \sqrt{\frac{1}{k} \sum_{i=1}^{k} (\hat{y}_i - y_i)^2},$$

(3.1)

where $k$ is the number of time periods, $\hat{y}$ is the aggregated travel time from Bluetooth or the permanent LPR cameras and $y$ is the ground truth travel time from the evaluation data.

The MPE, which indicates structural bias, is defined as

$$MPE = \frac{1}{k} \sum_{i=1}^{k} \frac{y_i - \hat{y}_i}{y_i}.$$

(3.2)

The bias that can be seen in the MPE is most likely explained by the location of the detectors. The different types of detectors are probably not situated in the exact same place. Furthermore, the location of the detectors is close to stop lines in both ends which might have an affect the results.
Table 2: Root Mean Square Error and Mean Percentage Error for the aggregated travel times collected with Bluetooth and permanent LPR.

<table>
<thead>
<tr>
<th></th>
<th>13-14</th>
<th></th>
<th>16-17</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bluetooth</td>
<td>Permanent LPR</td>
<td>Bluetooth</td>
<td>Permanent LPR</td>
</tr>
<tr>
<td><strong>RMSE (s)</strong></td>
<td>19.5</td>
<td>17.7</td>
<td>37.7</td>
<td>25.0</td>
</tr>
<tr>
<td><strong>MPE (%)</strong></td>
<td>-5.6</td>
<td>6.0</td>
<td>-5.6</td>
<td>-0.7</td>
</tr>
</tbody>
</table>

The conclusion from this rather straightforward evaluation is that Bluetooth is a suitable method for collecting travel times. The penetration rate is fairly high and the collected travel times correspond well with the evaluation data that has been collected. However, the exact location of the sensors has to be known, and the collected data has to be processed in order to remove outliers and a more sophisticated processing than just aggregating the data is recommended.

This evaluation was done in 2012. After that, Bluetooth sensors has been used to collect travel times on parts of the highway in Stockholm when the radar detectors temporarily had to be removed due to construction works.

3.5 SUMMARY AND CONCLUSIONS

There is no doubt that the possibility to collect detailed traffic data in a cost-effective way has increased during the last decade, in particular after the introduction of the smartphone. Many of the new emerging data sources are probe based and provide large coverage without any major infrastructure deployment. They will however not replace the traditional ones in a foreseeable future; each category of traffic data collection has its own strengths and weaknesses. The traditional ones are superior when it comes to flow data collection, which is essential for some applications, while the new data sources in general are more suitable for collecting travel time and trajectory data. Unlike most traditional fixed sensors many of these new data sources only provide data from a subset of the vehicles. In some cases, this is not an issue since a penetration rate of 3 % to 5 % is enough to provide reliable speed estimates, see for example Srinivasan and Jovanis (1995) or Turner and Holdener (1996). However, using just a subset of the vehicles might introduce bias in the data. For example, professional drivers with good knowledge of the traffic system tend to avoid parts of the road network with recurring congestion which might cause an underestimation of the traffic congestion. At the same time crowdsourced data might also underestimate the traffic congestion since the users of these systems in general are well informed of the current traffic state.

Another important difference between the traditional fixed sensors and many of the new emerging traffic data sources is the access to data. For the new probe based data we generally cannot control when and where we get access to data. For applications like incident management this can be problematic. Fixed sensors on the other hand can be placed where we identify a need for data and we get constant high quality measurements in these locations.

The introduction of new data sources certainly opens up new opportunities in the field of traveler information and traffic management. But it also creates new challenges like protection of the integrity of the drivers and data availability. Furthermore, it requires an adaptation and further development of the traffic models that are used so that they can handle all types of data, not just
data from the traditionally used fixed sensors. For some applications and at some places the traffic models might be obsolete due to the abundance of data and models that filter out outliers are sufficient. However, in most cases and for most applications it is necessary to utilize the characteristics of both traditional fixed sensors and new probe based sensors.

In Stockholm, the investment and maintenance costs for the fixed detectors are considered very high. For this reason, the Swedish Transport Administration is trying to find ways to optimize the use and minimize the number of fixed sensors. An important part in this work is to use advanced traffic models and data fusion algorithms to combine the data from fixed detectors with data from other sources, such as GNSS-equipped vehicles and travel times from Bluetooth, LPR-cameras or probe vehicles.
4 TRAFFIC STATE ESTIMATION USING MULTIPLE DATA SOURCES

This chapter covers traffic data fusion in general and the model chosen for further experiments on the Stockholm highway network in particular. The chapter begins with an introduction to the topic and a summary of previous research on fusion of traffic data, i.e. methods for combing different types of traffic data. This is followed by a more detailed presentation of the model we have used for estimation of the traffic state on the highways around Stockholm, the CTM-v and the Ensemble Kalman filter, here denoted the MMS-model. Finally, the adaptations necessary for the chosen model to estimate the traffic state on the highways in Stockholm is presented.

4.1 INTRODUCTION TO TRAFFIC DATA FUSION

As presented in Chapter 2, the traditional macroscopic traffic models used for estimating the traffic state on highways were developed based on data from fixed sensors. The introduction of new methods that collect point speeds and trajectory data from individual vehicles and travel times based on a subset of all vehicles, forms a new prerequisite for traffic modelers. Each data source has its strengths and weaknesses and in order to utilize the strength of each data source the traditional traffic models have to be modified and new models for fusing different data sources have to be developed. The characteristics of the different types of traffic state measurements can vary both in time and space. One way to categorize the different data types is presented in Table 3.

Table 3: Description of different measurement types, their characteristics and the data collection method used.

<table>
<thead>
<tr>
<th>Measurement type</th>
<th>Characteristics</th>
<th>Collection method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static point measurements</td>
<td>Measurements of flow and/or speed collected at a fixed position. Detects basically all passing vehicles. The data is in general aggregated over one or five-minute periods.</td>
<td>Fixed sensors (radar, inductive loops and magnetometers).</td>
</tr>
<tr>
<td>Dynamic point measurements</td>
<td>Individual speed measurement from probe vehicles. The measurement contains position and instant speed (in some cases also heading).</td>
<td>GNSS equipped probe vehicles where the speed is included in the measurement.</td>
</tr>
<tr>
<td>Static travel times</td>
<td>Travel times collected with re-identification using fixed sensors at two or more positions. In general, only a subset of the vehicles is detected, how many depends on the method used. The actual route has to be estimated.</td>
<td>Bluetooth, WiFi, LPR, Toll tag, vehicle signature matching and other methods where re-identification is used.</td>
</tr>
<tr>
<td>Dynamic travel times</td>
<td>Id, position and timestamp of consecutive measurements from a certain vehicle are used to generate travel times. The data has to be map matched and the used route has to be estimated.</td>
<td>GNSS (or WiFi) equipped vehicles/ devices where only id is collected.</td>
</tr>
</tbody>
</table>

As described in Section 3.3, the data in Stockholm mainly consist of static point measurements (radar detectors), static travel times (LPR, Bluetooth), and dynamic travel times (GNSS-equipped
taxis). However, there is an ambition to also collect dynamic point measurements. Hence, a model that can combine all types of measurements would be ideal. However, this is not at all trivial. The spatial and temporal difference between the data types makes it a very challenging task.

The purpose of the data fusion model in this context is to, based on different types of measurements that vary in time, space and quality, estimate the current traffic state and possibly also predict the future traffic state. See Figure 16 for a graphical representation of the process where the four data types presented in Table 3 are combined. However, combining measurements that vary in both time, space and quality is not straightforward. For example, imagine that the available data consist of radar measurements, GNSS-data and travel times from Bluetooth. The radar delivers speed and flow measurement at fixed positions, the sensors detect all vehicles but the data is aggregated over one-minute periods before it is transmitted to the traffic management center. Add to this that the quality of this data is poor at low speeds. The GNSS-data consist of point speed measurements from individual vehicles collected where vehicles with such a device happen to be located. Hence, the traffic management center cannot control when and where such data is collected and it only delivers the speed of an individual vehicle, no flows. The third type of data available in this fictive example is travel times from Bluetooth. This data consists of aggregated travel times between fixed positions, the positions of the Bluetooth sensors. The travel times are based on a subset of the vehicles since not all vehicles carry an active Bluetooth device and the data might be biased. Furthermore, since the positions of the sensors are fixed, an incident between these two positions will not be detected until the first vehicle affected by the incident passes the second sensor. Even though this particular example is fictive, it illustrates the reality for many traffic management centers. By using a model that can fuse these different types of measurements, both the quality of the estimation and the size of the network on which the traffic state can be estimated, can increase.

Figure 16: Illustration of the data fusion process where the different types of measurements are combined in order to estimate the current traffic state.
Mitchell (2007) defines data fusion as “The theory, techniques and tools which are used for combining sensors data, or data derived from sensor data, into a common representational format. In performing sensor fusion our aim is to improve the quality of the information, so that it is, in some sense, better than would be possible if the data sources would be used individually”.

Data fusion is a well-established research topic within the military and aircraft sector, see for example Gustafsson (2010) and Hall and Llinas (1997). Within the transport sector the interest for data fusion has, for natural reasons, increased as new types of data has been made available. The rapid increase of new data sources during the last decade has also made the task more complex and consequently further research is required.

In a survey on data fusion models and methods in Intelligent Transportation Systems (ITS) El Fazoui et al. (2011) concluded that there are several models and methods available that can be used to fuse heterogeneous data. They range from more naïve statistical approaches as different weighting methods to more complex methods as neural networks and Kalman filters. El Fazoui et al. (2011) further concluded that ITS is one of the areas where data fusion is most relevant, but also the most challenging.

4.2 PREVIOUS WORK ON TRAFFIC DATA FUSION

The research on traffic data fusion has escalated during the last 20 years. However, ITS and the traffic field consist of a variety of different applications where data fusion can be used, such as incident detection and traffic state estimation and prediction. The focus of this survey is on previous work on data fusion for traffic state estimation and prediction on highways. Furthermore, only data fusion models where the output is an estimate of a specific traffic variable, in most cases speed, has been studied.

The survey by El Fazoui et al. (2011) shows that for traffic state estimation and prediction a large variety of methods and models have been evaluated. Autoregressive models, Bayesian frameworks, Kalman filters and neural networks are some of the methods that are mentioned. However, the conclusion is that none of the proposed methods produce accurate estimations and predictions except for some special conditions. The explanation for this is that the dynamics in traffic cannot be formalized by a single procedure. In order to achieve accurate results for a variety of network configurations and data sets available a combination of different approaches is recommended. One solution is to assimilate estimations of the same quantity from different estimation and prediction models and methods. It is noted that various Kalman filters and particle filters are well established methods used for data assimilation.

The Kalman filter was first presented by Kalman (1960) and Kalman and Bucy (1961). It is a recursive algorithm that estimates the state variables in a noisy linear dynamical system. This is done continuously as new observations are added and as the system evolves over time. Each update provides an unbiased estimate of the state variables together with a covariance matrix that describe the uncertainty of the estimates. The process is relatively easy to compute which makes the Kalman filter suitable for real time implementations. Different extensions and generalizations of the Kalman filter have been developed over the years. A more detailed description of the Kalman filter in general and the Ensemble Kalman filter in particular can be found in Section 4.3.3. For a more
in depth description of Kalman filtering in traffic management and control the reader is referred to Van Lint and Djukic (2012) and Antoniou et al. (2010).

Byon et al. (2010) used a Single-Constraint-At-A-Time (SCAAT) Kalman filter that uses the single most recent speed measurement from any available sensor, in this case loop detectors and GPS probe data. The state is updated based on the characteristics of that particular sensor (i.e. the noise and variance of the measurement) and the accumulated state estimation from the previous step. The results are promising but the limitation is that only speed measurements can be used.

Bachmann et al. (2013) compared different methods for fusing speed from loop detector data with travel times collected with Bluetooth. This was done both on collected data and on simulated data. Neural network, measurement fusion Kalman filter, SCAAT Kalman filter, fuzzy integral, ordered weighted averaging and a simple convex combination of data, were some of the evaluated methods. The results show that most data fusion techniques improve the accuracy over single type sensor approaches. The measurement fusion Kalman filter and the simple convex combination perform well in all scenarios and often significantly improve the accuracy of the estimations. The authors also conclude that the measurement fusion Kalman filter would be particularly useful if flow and density also would be available. The measurement fusion Kalman filter is a multi-sensor multi-temporal Kalman filter presented by Mitchell (2007). The classical Kalman filter is suitable for fusion of multi-temporal data where a sequence of data from one type of sensor is available, while the measurement fusion Kalman filter considers several sequences of measurements from different types of sensors. However, this version of the Kalman filter requires measurement from all sensors to produce a fused estimate.

Qing et al. (2010) proposed a method called TravRes that can be used to reconstruct the shape of trajectories but also to fuse different types of traffic data. TravRes uses an iterative approach and is based on a method called PISCIT. The method assumes that each cell in the space-time speed contour plot has homogenous traffic conditions and that the floating car data is consistent and can be exactly reconstructed from the space-time speeds. TravRes consist of an inner and an outer iteration loop where the inner loop reconstructs the trajectories and the outer loop reconstructs the space-time speed contour plot. The two loops are iteratively ran until a certain threshold is reached and the given floating car data is consistent enough with the space-time speeds. The TravRes method can also be used to fuse floating car data with local speed measurements or travel times and the validation of the method shows that only a few iterations often are enough to improve the estimation error.

Treiber and Helbing (2002) introduced the adaptive smoothing method which is a non-linear spatio-temporal low pass filter that uses the information from loop detectors to reconstruct the spatio-temporal traffic dynamics for a certain road network. The proposed method is based on a number of parameters such as the propagation velocity in congestion and free flow. In accordance with the kinematic wave theory, the method assumes that the information travels in the direction of traffic flow when in free flow, while the information travels against the direction of traffic while in congestion. Each data point is assigned a weight which is based on the spatiotemporal distance from the point where the traffic state is estimated. This method was later extended and generalized by Van Lindt and Hoogendoorn (2010) to also handle other types of data. The so-called Extended Generalized Treiber-Helbing filter (EGTF) can handle all kinds of data, no matter their spatial and
temporal resolution, as long as the data provides a mean to distinguish between free flow and congestion. The original filter proposed by Treiber and Helbing (2002) assumed that the data was collected at fixed points and for fixed time interval. However, in the EGFT the data can stem from any position and time. Two weights are introduced where the first one is a data specific weight which can be interpreted as the reliability of that particular data source. The second weight is data source and data point dependent and reflect the fact that data from different sources will result in different distinctions between free flow and congestion. The EGFT is evaluated both on simulated data in Van Lint and Hoogendoorn (2010) and on empirical data in Van Lint (2010) with promising results. In Van Lint (2010) the method is compared with piecewise, linear and quadratic speed-based interpolation methods and as earlier concluded by Van Hinsbergen (2008) using kinematic wave theory in combination with filtering technique results in more accurate travel time estimations in comparison with interpolation methods.

4.3 Traffic state estimation using the MMS-model

We have chosen to work with a modified version of the Cell Transmission Model named the Cell Transmission Model for velocities (CTM-v) that, together with an Ensemble Kalman filter (EnKF), can be used to fuse speed measurements from loop detectors and probe data.

This method, presented by Work et al. (2010) and Bayen et al. (2011), was developed within the Mobile Millennium project to run in real-time and estimate the traffic state for the highway network around San Francisco. Mazare et al. (2012) evaluated almost 1000 scenarios where the CTM-v and an Ensemble Kalman filter was used to fuse probe and loop detector data and estimate travel times. The results clearly showed that when complementing loop detector data with probe vehicle data, better estimates for travel times were obtained.

The data fusion based on the CTM-v and the EnKF is appealing for many reasons. First of all it is developed to run in real-time and for a large network. Furthermore, it can fuse different types of point speed measurements but the Kalman filter also enables the possibility to include travel time measurements. There are also many aspects within the real-time implementation that can be improved and developed further such as the calibration of the parameters and introducing more dynamic parameters. Finally, running the CTM-v model forward in time and use it for prediction is a relatively unexplored area which will be analyzed further in this thesis.

The model is based on the kinematic wave theory and the LWR PDE and its discretization, the Cell Transmission Model, which was previously described in Section 2.4. Traditionally, these models use density as a measure of state. However, due to the characteristics of the data available in the Mobile Millennium project, new models were developed that use velocity as the state. The new models are referred to as the LWR-v and the CTM-v. The velocity state estimation problem is solved using the EnKF. A hypothetic data flow in the real-time system is presented in Figure 17. In the figure, data of four different types, the ones presented in Table 3, are pre-processed and transferred to the EnKF. In the EnKF the pre-processed data assimilated with the estimation from the CTM-v model. The output from the EnKF is a traffic state estimation that can be used in various applications and is also used as input in the next estimate by the CTM-v model.
Figure 17: The figure illustrates a hypothetic data flow in the implemented data fusion system. The MMS-model is found within the dashed line.

In the following sections we describe the CTM-v and the EnKF. We also present how the extension of the CTM-v from individual links to a larger network with on- and off-ramps is done. For a more in depth description of the models the reader is referred to Work et al. (2008) and Work et al. (2010).

4.3.1 THE CELL TRANSMISSION MODEL FOR VELOCITIES

The previously introduced LWR-v is consistent with the classical LWR PDE, given by Equation 2.7. The conservation law for velocities with initial conditions is formulated as

\[
\frac{\partial v(x,t)}{\partial t} + \frac{\partial V(v(x,t))}{\partial x} = 0 \quad v(x,0) = v_0(x),
\]

where \( V(\cdot) \) is a function, a fundamental diagram, where the relationship between the speed and density is affine, as it is in Greenshields fundamental diagram presented in Section 2.3. For velocity functions that are not affine there is no equivalent conservation law, see Work et al. (2010) for proof.

Hence, if we want to use a fundamental diagram that corresponds better with measured data we have to discretize the LWR PDE with the Godunov scheme (Equation 2.10) before the inversion to the velocity domain is made. The formulation of the CTM-v requires that the velocity function is strictly decreasing and invertible. Thus, the Daganzo-Newell fundamental diagram presented in Section 2.3 cannot be used since it is not strictly decreasing in free flow, i.e. the section for densities lower than the critical density, see Diagram II in Figure 18. Instead a hyperbolic-linear velocity function \( V_{HL} \) with a linear expression in free flow and a hyperbolic expression in congestion is used according to
where $\rho_{\text{max}}$ is the jam density, $v_f$ the free flow speed, $w_f$ the backward propagating shock wave speed and $\rho_{\text{cr}}$ the critical density. In Figure 18 the hyperbolic-linear fundamental diagram (III) is illustrated together with the fundamental diagrams of Greenshields (I) and Daganzo-Newell (II).

Since the fundamental diagram has to be invertible, continuity is required at the breakpoint where $\rho = \rho_{\text{cr}}$ and the relation

$$\frac{\rho_{\text{cr}}}{\rho_{\text{max}}} = \frac{w_f}{v_f},$$

must be satisfied.

The inverted version of the hyperbolic-linear fundamental diagram can be expressed as

$$\rho = V_{HL}^{-1}(v) = \begin{cases} \rho_{\text{max}} \left(1 - \frac{v}{v_f}\right) & \text{if } v \leq v_{\text{cr}}, \\ \rho_{\text{max}} \left(1 + \frac{v}{w_f}\right) & \text{otherwise}, \end{cases}$$

where the density now is expressed as a function of the speed.
Furthermore, Equation 2.8 together with the fact that the hyperbolic-linear fundamental diagram \( V_{HL}(\cdot) \) is invertible gives the relationship
\[
\tilde{Q}(\rho) = \tilde{Q}(v) = Q\left(V_{HL}^{-1}(v)\right) = V_{HL}^{-1}(v)v,
\]
where the flow \( \tilde{Q}(v) \) is expressed as a function of the inverted velocity function where \( V_{HL}^{-1}(v) \) gives the density \( \rho \), which multiplied with the speed \( v \) gives the flow.

In the discrete case the estimated velocity in the next time step \( n + 1 \) is expressed as the inversion of Equation 2.9 and 2.10 formulated as
\[
v_{i+1}^{n+1} = V_{HL} \left( V_{HL}^{-1}(v_{i}^{n}) - \frac{\Delta t}{\Delta x} \left( \tilde{G}(v_{i}^{n}, v_{i+1}^{n}) - \tilde{G}(v_{i+1}^{n}, v_{i-1}^{n}) \right) \right),
\]
which is the Cell Transmission for velocities, the CTM-v, where \( i, i+1 \) and \( i-1 \) denotes the cell for which the velocity is predicted and the cell downstream and upstream, respectively, see Figure 4 in Section 2.4. \( \tilde{G} \) denotes the transformed Godunov flow in the velocity domain which is formulated as
\[
\tilde{G}(v_{1}, v_{2}) = \begin{cases} 
\tilde{Q}(v_{2}) & \text{if } v_{cr} \geq v_{2} \geq v_{1}, \\
\tilde{Q}(v_{cr}) & \text{if } v_{2} \geq v_{cr} \geq v_{1}, \\
\tilde{Q}(v_{1}) & \text{if } v_{2} \geq v_{1} \geq v_{cr}, \\
\min\left(\tilde{Q}(v_{1}), \tilde{Q}(v_{2})\right) & \text{if } v_{1} \geq v_{2},
\end{cases}
\]
where \( v_{1}, v_{2} \) and \( v_{cr} \) denotes the upstream velocity, the downstream velocity and the critical velocity. Equation 4.7 can be graphically represented as shown in Figure 19.

Figure 19: Graphical representation of Equation 4.7, where \( v_{1} \) and \( v_{2} \) denotes the upstream velocity and the downstream velocity respectively.
Using the inverted hyperbolic-linear fundamental diagram in Equation 4.4 we get the Hyperbolic-linear model

\[
\hat{G}(v_1, v_2) = \begin{cases} 
  v_2 \rho_{\text{max}} \left( \frac{1}{1 + \frac{v_2}{v_f}} \right) & \text{if } v_c \geq v_2 \geq v_1, \\
  v_c \rho_{\text{max}} \left( 1 - \frac{v_c}{v_f} \right) & \text{if } v_2 \geq v_c \geq v_1, \\
  v_1 \rho_{\text{max}} \left( 1 - \frac{v_1}{v_f} \right) & \text{if } v_2 \geq v_1 \geq v_c, \\
  \min(V_{\text{in}}(v_1), V_{\text{in}}(v_2)) & \text{if } v_1 \geq v_2.
\end{cases}
\] (4.8)

Equations 4.6 and 4.8 defines the evolution of the velocity field for each cell except for the first and last cell of the link. For the boundaries the equations

\[
v_{0}^{n+1} = V_{\text{HL}} \left( V_{\text{HL}}^{-1}(v_{0}^{n}) - \frac{\Delta t}{\Delta x} \left( \hat{G}(v_{0}^{n}, v_{1}^{n}) - \hat{G}(v_{1}^{n-1}, v_{1}^{n}) \right) \right),
\] (4.9)

\[
v_{\text{max}}^{n+1} = V_{\text{HL}} \left( V_{\text{HL}}^{-1}(v_{\text{max}}^{n}) - \frac{\Delta t}{\Delta x} \left( \hat{G}(v_{\text{max}}^{n}, v_{\text{max}}^{n+1}) - \hat{G}(v_{\text{max}}^{n-1}, v_{\text{max}}^{n}) \right) \right).
\]

are used. These equations include references to points outside the modelled link \( v_{0}^{n} \) and \( v_{\text{max}}^{n} \).

4.3.2 EXTENDING THE MODEL TO A NETWORK WITH ON- AND OFF-RAMPS

The CTM-v estimates the velocity evolution on a link which can consist of several cells. In the implementation of the model, the highway network is modelled as a directed graph where the links are divided into cells that are approximately 300 meters in length. However, the highway network also contains on-ramps and off-ramps and to be able to model how the velocity field evolves over time in a network a CTM-v network update algorithm has been formulated. The algorithm sequentially applies the CTM-v for each link in the network and models the flow distribution at each junction.

The modeling of the flow distribution at a junction in the MMS-model is dependent on three conditions. The first condition, the conservation of vehicles across the junction, assumes that the junction has no storage, this implies that the flow into the junction must equals the flow out of the junction. The second condition gives that the total volume entering the junction from an incoming link is distributed on the outgoing links according to allocation parameters, also denoted split ratios. An allocation matrix gives the aggregated information about the route distribution across the junction.

The first two conditions create an upper bound on the flows on each link into and out of the junction. However, the first two conditions cannot always be fulfilled since they combined create strong boundary conditions, for proof see Work et al. (2010). Hence, a third condition that says that the flow across the junction should be maximized is introduced. A linear programming problem is used to calculate the flow on the incoming links to the junction and the strong boundary
conditions from the first two conditions provides upper bounds for this problem. For a more in-dept description of the linear programming problem the user is referred to Work et al. (2010).

In the CTM-v network update algorithm the velocity field for each cell \( i \in \{0, \ldots, I_{\text{max}}\} \) and all links \( l \in \{l_0, \ldots, l_k\} \) of the network is written as

\[
v^n := [v_{i,0}^n, \ldots, v_{i,\text{max}}^n, \ldots, v_{l,0}^n, \ldots, v_{l,\text{max}}^n], \tag{4.10}
\]

and the velocity at the next time step \( t = (n + 1) \) is given by

\[
v^{n+1} = M[v^n], \tag{4.11}
\]

where \( M \) denotes the CTM-v network update algorithm presented by Work et al. (2010). The algorithm consists of two steps:

1. For all junctions in the network compute the maximum incoming and outgoing allowed flow and solve the linear programming problem for each junction.
2. For all links in the network compute the velocity field according to the CTM-v using Equations 4.6 and 4.9.

### 4.3.3 The Ensemble Kalman Filter

The basic principle of the Kalman filter is to predict a new state and its uncertainty and in the next step correct the prediction based on new measurements.

The Kalman filter model assumes that the true state \( x \) at time \( n \) is evolved from the state at \( n-1 \) according to two basic equations; the linear state equation

\[
\hat{x}^n = A\hat{x}^{n-1} + \epsilon^{n-1}, \tag{4.12}
\]

and the measurement equation

\[
y^n = H\hat{x}^n + e^n, \tag{4.13}
\]

where \( \epsilon \) and \( e \) represent the Gaussian noise terms for the state and the measurement, \( A \) is the state transition matrix and \( H \) is the measurement matrix. The state transition matrix \( A \) relates the state \( \hat{x} \) at the predicted time step with the state in the current time step while the measurement matrix \( H \) relates the state \( \hat{x} \) to the measurement \( y \). These matrices could either change at each time step or, like in this example, be static. The state equation can also include a control input if the modelled system contains such an input.

The iterative Kalman filter process consist of two stages, the time update (prediction) and the measurement update (correction), see Figure 20. The subscript \( p \) represents that the variable is predicted in the time update.
Figure 20: The recursive Kalman filter consists of two steps, the time update and the measurement update.

As presented in Figure 20, in the time update the state estimation \( \hat{x}_n \) is first predicted based on the state estimation from the previous time step and the state transition matrix \( A \). This is followed by a calculation of the predicted estimate error covariance matrix \( P_p \), based on the error covariance matrix \( P \) from the previous time step, the state transition matrix \( A \) and the state noise covariance matrix \( Q \) which describes the variance of \( \varepsilon \).

The measurement update consists of three steps. First the Kalman gain \( G \) is calculated based on the predicted error covariance \( P_p \), the measurement matrix \( H \) and the measurement noise covariance matrix \( R \) which describes the variance of \( \varepsilon \). This is followed by an update of the state estimation \( \hat{x} \) from the predicted state estimation \( \hat{x}_p \), the Kalman gain \( G \), the measurement \( y \) and the measurement matrix \( H \). Finally the error covariance matrix \( P \) is updated based on the predicted estimate error covariance \( P_p \), the Kalman gain \( G \) and the measurement matrix \( H \).

The Kalman gain \( G \) is describing the certainty of the measurements and the current state estimation and it can be tuned to minimize the error covariance matrix \( P \). In the extreme case with a gain of one the filter ignores the predicted state while the filter will ignore the measurement if the gain is zero.

For examples and a more thoroughly description of the classic Kalman filter and its characteristics the reader is referred to Faragher (2012) and Welch and Bishop (2001).

A number of different adaptions and extensions to the Kalman filter have been proposed over the years. The basic Kalman filter is designed for linear problems but the extension of the CTM-v to networks implies a non-linear and non-differentiable state equation. However, there exist different extensions to the Kalman filter that can manage non-linearity. One commonly used example is the Extended Kalman filter that has been used for traffic estimation by Wang and Papageorgiou (2005) and others. Though, in the Extended Kalman filter all functions have to be differentiable, which is not the case in the CTM-v, see Bayen et al. (2011) for proof. An alternative to the extended Kalman filter is the Ensemble Kalman filter. The Ensemble Kalman filter (EnKF), first presented by Evensen (2003), is an extension of the classical Kalman filter that can handle both non-linearity and non-differentiability. Furthermore, in the EnKF, the state estimate distribution is represented by an ensemble, a collection of state vectors. The error covariance matrix \( P \) is a sample covariance computed from the ensemble. This makes the EnKF suitable for problems with a large number of
variables, such as traffic state estimation for large networks. The ensemble of model states is propagated forward in time which makes it possible to calculate the mean and covariance of the error needed in the measurement update step. For a detailed description of the EnKF the reader is referred to Evensen (2003).

Given the CTM-v network update algorithm described earlier the state equation is

\[ v^n = M[v^{n-1}] + e^n, \]

which is Equation 4.11 complemented by the Gaussian white zero-mean state noise \( e^n \) with covariance matrix \( Q \).

The network measurement equation is given by

\[ y^n = H^n v^n + e^n, \]

where \( y^n \) is the measurement, \( H^n \) is the measurement matrix and \( e^n \) is the white, zero mean measurement noise with covariance matrix \( R \). The measurement matrix \( H \) is a linear matrix describing the discrete cells on the highway for which measurement is observed during each time step \( n \).

The implemented Ensemble Kalman filter presented by Work et al. (2010) can be summarized in four steps:

1. Initialization: An ensemble of size \( K \) of model states \( v^0(k) \), where \( k \in \{1, \cdots, K\} \), is generated from a process with mean speed \( \bar{v}^0(k) \) and covariance matrix \( P^0 \). This sample represents our initial state.

2. Time update: In the time update, the predicted state \( v^p_n(k) \) of each of the \( K \) ensemble members is calculated using the CTM-v model, i.e. the state equation (Equation 4.14)

\[ v^p_n(k) = M[v^{n-1}(k)] + e^n(k). \]

Then the ensemble mean \( \bar{v}^p_n(k) \) and the covariance matrix \( P^p_n \) of the predicted mean is updated using

\[ \bar{v}^p_n(k) = \frac{1}{K} \sum_{k=1}^{K} v^p_n(k), \]

\[ P^p_n = \frac{1}{K-1} \sum_{k=1}^{K} (v^p_n(k) - \bar{v}^p_n)(v^p_n(k) - \bar{v}^p_n)^T. \]

3. Measurement update: In the measurement update the Kalman gain \( G^n \) is calculated

\[ G^n = \frac{P^p_n (H^n)^T}{H^n P^p_n (H^n)^T + R^n}, \]

and the state estimation \( v^n(k) \) is updated based on new measurements \( y^n \).
\[ v^n(k) = v^n_{p}(k) + G^n \left( y^n(k) - H^n v^n_{p}(k) + e^n(k) \right). \] (4.20)

4. Return to Step 2 which is repeated until new measurements are available.

In the implemented model the CTM-v network algorithm is initiated every six second, i.e. the linear problem at each junction is solved together with the CTM-v which gives a velocity prediction for the next time step for each cell. Available pre-processed point speed measurements are assimilated into the model every 60 seconds. Given the measurements and the prediction from the traffic model the traffic state, the velocity field is then computed using the Ensemble Kalman filter.

Worth noticing is that each measurement vector \( y^n \) is represented by an ensemble that has the actual measurement as mean and the variance of the ensemble is used to represent the measurement errors. This ensures that the updated ensemble has the correct covariance.

Since the state covariance is represented by a limited number of ensemble members, non-physical correlations may arise. In practice this means that correlations between parts in the network that do not correlate in practice might occur. Furthermore, the presented Ensemble Kalman filter requires that the predicted error covariance \( P_p \) is calculated for the entire network in order to calculate the Kalman gain \( G \) in the measurement update. This might create problems when the state on a large network is estimated. Work et al. (2010) has solved this by implementing a covariance localization method. This limits the correlation between links in the network and only nearby links can be correlated. The method also provides an efficient way to update the state variables in the measurement update. The computation of the covariance matrix in Equation 4.18 is transformed into computation of small localized covariance matrices for each link in the network, given nearby links on which the correlation is relevant.

4.4 NETWORK CREATION

To be able to run the MMS-model we need a network for which the traffic state shall be estimated. Due to the nature of the CTM-v, the network has to be represented in a certain way. The process of creating this network consists of two parts. Firstly, a mathematical representation of links and nodes is created based on a set of shape-files, in this case taken from a street database, a digital map, from Navteq. This representation is called the model graph. Secondly, the network for which the traffic state is to be estimated is selected as a subset of the model graph. The created model graph will also be used for estimation of the traffic state on arterials, not only highways that are the focus of this work. Hence, also links outside the highway network are included.

The shape files contain a lot of data about the links, most of which can be ignored. For this context, information such as the speed limits, number of lanes and the function class of the link and are the most relevant. The function class of a link is based on a hierarchical classification of the road network. Furthermore, information about restricted driving maneuvers and special driving conditions are also used.

The process of creating a model graph consists of several steps that are briefly described below.
1. **Link removal**

   All small residential streets are removed since they are not considered to be of interest for the state estimation. Only links with function class 1-4 are used which means that all of the grey links in Figure 21 where the network in Stockholm is showed, are removed.

2. **Link merging and node removal**

   The Navteq map consists of a lot of short links that are merged together in this step. As a consequence, a number of nodes are also removed. However, the links that are merged together must have the same speed limit and the same number of lanes. Furthermore, a node that is removed must have exactly two adjacent links that are of the same type, bi- or unidirectional.

3. **Composite intersection consolidation**

   At some places in the digital map from Navteq “phantom” links are included that create composite intersections. A composite intersection is defined as an intersection of doubly digitized roads. In this step, “phantom” links are removed and composite intersections are reduced to a single node in the model graph.

4. **Unidirectional links**

   All bidirectional links are transformed into unidirectional links and each bidirectional link is replaced by two unidirectional links. The two unidirectional links have reverse geometries and opposite start and end nodes.

5. **Categorization**

   Links are categorized as highways, arterials or ramps. The categorization is primarily based on the link attribute function class. All links with function class 1 or 2 are categorized as highways, and links with a function class 3 or 4 are categorized as arterials. Ramps are identified through the link attribute ramp. In Figure 21 the highways are colored blue and the arterials green.

![Figure 21: Categorization of highways and arterials in Stockholm. Highways are colored blue and arterials green.](image-url)
Since the highway model is suitable for larger networks while each link is modelled independently in the arterial model, manual adjustments to the model graph are necessary. First of all, to create a connected highway network, ramps that connect two highways are categorized as highway links. There are also some parts of the highway network that have traffic signals, these routes are treated as arterials.

6. Link Shortening
Highway links that are too long and exceed a certain length threshold are divided into as few links as possible. Each link is kept under the threshold value while also maintaining link breaks at the boundary of Navteq links. The value for this threshold is 1500 meters.

4.5 Real-time implementation for Stockholm
The described model for data fusion and traffic state estimation was developed and evaluated on data collected in the San Francisco Bay Area. The data collected in Stockholm is similar to this data but there are many subtle differences in the characteristics of the data which required adjustments to the pre-processing filters and models used for it to work properly for Stockholm. Furthermore, a network for Stockholm that the model can run on had to be created according to the process described in Section 4.4. Due to different speed categories and speed units in the digital map from Navteq over San Francisco Bay Area and Stockholm, adaptions to the database conversion process had to be made. The speed unit in the Navteq street database is mph or km/h but the model graph and the models use m/s (meters per second).

As presented in Section 3.3 the main data source on the highway network in Stockholm is radar detectors that collect point speed and flow aggregated over one-minute intervals. The sensors are an essential component of the motorway control system on the highway around Stockholm and they are located approximately every 500 m. The collected radar data first has to be pre-processed. The purpose of the pre-processing is to aggregate data collected for several lanes and to filter out high speeds. Since the highway model is not lane dependent, an aggregation of data is necessary where data is available for several lanes. For each lane where a measured speed and flow is available, a density is calculated by dividing the flow with the speed. This is done in accordance with the macroscopic traffic theory presented in Chapter 2. The calculated density is used as a weight when an aggregated speed is calculated for all the lanes covered by the detector site. The filtering of measured speeds that are higher than the speed limit is carried out due to the fact that very high speeds can create problems for the model. As presented previously the model use the relationship between flow, density and velocity and due to the characteristics of the fundamental diagram very high speeds can result in negative flows.

The measurements used in Stockholm are loaded from different feeds with different types of sensors compared to those used in the original system in California. This has required some changes to the interface where the measurements are loaded.

Before the radar measurement can be used, each sensor has to be map matched to the network. Since the used digital map is not so detailed, especially around the ramps to the highway, the map matching of the radar detectors is difficult. The process is also complicated by the fact that the highway is just one link in each direction, no matter how many lanes it consists of, while a detector site consists of several unique radar detectors. As a result, it is not possible to map match the
sensors to the closest link. The problem is illustrated in Figure 22 where a section of the highway north of Stockholm is illustrated. The Northbound section, the link to the right in the figure, has an on-ramp while the southbound section, the link to the left in the figure, has an off-ramp. The dots represent the individual radar detectors mounted over each lane. In reality, all sensors on the northbound direction are located on the mainline of the highway while the sensor marked in the southbound direction is located on the ramp and the rest are located on the mainline of the highway. Hence, it is not possible to automatically map match the sensors. To overcome this problem, all radar detectors close to a ramp had to be corrected manually using the street view function in Google Maps.

Figure 22: The figure illustrates a part of the highway north of Stockholm where the dots are the individual radar detectors mounted over the highway. The link to the right in the figure illustrates the northbound direction with an on-ramp while the link to the left illustrates the southbound direction with an off-ramp.
Before a traffic model can be used for analysis or real-time estimations of the traffic state it must be proven that the model is a good enough representation of the infrastructure and the traffic dynamics. An important step towards a valid model is model calibration. In the literature, the definition of calibration varies but in general it can be described as an iterative process where the model parameters are adjusted in such a way that the model output matches field observations. In our case it could mean that the estimated travel times are compared with measured travel times. If possible, it would be preferable if this iterative change in model parameters was managed by an automated algorithm as pointed out by Ngoduy and Maher (2012), Munoz et al. (2004) and Cremer and Papageorgiou (1981).

Once the model is calibrated it has to be validated. The definition of model validation differs but most of the studied papers state that validation, in this context, is a process where the output from the calibrated model is compared to field observations. See for example Braban-Ledoux (2000), Ngoduy and Maher (2012) and Cremer and Papageorgiou (1981). However, it is important that the validation should be performed for datasets spanning other time periods or for other output data than those used during the calibration process.

Prior to the calibration and validation of the model, the model can be verified in order to determine if it produces the desired output for a given set of input data. According to Rakha et al. (1996), the verification of a traffic simulation model has two purposes: to ensure that the output is consistent with the logic on which the model is based and to verify that outputs are consistent over the range of typical input.

In this chapter the verification, calibration and validation of the implemented MMS-model is presented. This process is carried out in several steps and the outline of the chapter follows the chronological order of the different steps. This order follows, to a large extent, the order in which data for verification, calibration and validation has been made available. To get a first view of how the uncalibrated MMS-model performs an initial visual verification against a number of traffic information services is presented in Section 5.1. In the uncalibrated MMS-model all parameters are manually set based on experience from the Mobile Millennium project at UC Berkeley. The initial visual verification is followed by a verification of the uncalibrated MMS-model output against travel times collected with GPS, presented in Section 5.2. Finally, the implemented calibration process is described together with the results from the calibration and validation of the model parameters against travel times collected with Bluetooth, see Section 5.3.

5.1 INITIAL VISUAL VERIFICATION

The first initial verification of the uncalibrated MMS-model included a visual comparison with other web based traffic information providers. In the verification, the real-time colored map that describes the current traffic state generated by the MMS-model was compared against similar maps from other traffic information providers. These providers included:

- Trafiken.nu - The official website for traffic information in Sweden. The colored map for Stockholm is mainly based on data from the radar detectors mounted over the motorway.
- Google Maps - A real-time traffic information service available as a layer in Google Maps.
- Trelucity/DN - A real-time traffic information service previously available at [dn.se/ sthlm/ trafiklaget](dn.se/ sthlm/ trafiklaget)
- TomTom - A real-time traffic information service available at [tomtom.com/ livetraffic](tomtom.com/ livetraffic)

As the above traffic information services each have different link lengths, different thresholds for the color coding, and also different delays, i.e. how often the information is updated, it is difficult to perform a robust verification. Furthermore, the estimations of traffic state are based on fixed sensors in some cases and probe vehicles in others. It is understood that TomTom and Google at the time for this verification only used probe vehicle data. This means that they only displayed information at locations where there are probes available, and also that these locations vary over time. Factors such as this make the information service unsuitable as an evident representation of ground truth for verification or validation purposes. Taken together, the different maps can be considered as a good indication of the current traffic state and therefore can be compared with the real-time output from the uncalibrated MMS-model as an initial form of verification.

During the morning peak on the 24th of November 2011 screenshots were recorded every 10 minute of the service providers mentioned above and of the visualization of the estimated traffic state for Stockholm generated from the MMS-model, denoted Mobile Millennium Stockholm in Figure 23.

![Figure 23: Screenshots from five different traffic information services, the output from the uncalibrated model (Mobile Millennium Stockholm) is in the middle.](image-url)

From the screenshots it could be concluded that the real-time output from the uncalibrated MMS-model was relatively similar to the real-time information provided by the other services. The congestion shown on the trafiken.nu map is in general the same as that indicated by the MMS-model, although there are differences in terms of exact location and time. These differences are most probably due to different aggregation levels and variations in link attributes (lengths, lane
division, etc.) among the various traffic information services. This makes it almost impossible to establish a ground truth for comparison from these data sources.

5.2 Verification against travel times collected with GPS

In order to get a more comprehensive verification of the model, the estimation output from the uncalibrated MMS-model is compared with GPS-data collected in an earlier pilot project, see Allström et al. (2012a). Furthermore, the estimation from the MMS-model is also compared with travel times estimated by the current system used by the traffic management center in Stockholm, described by Strömgren et al. (2004). Finally, we compare the estimation output from a scenario where only every other sensor is used to a scenario where data from all sensors is used. The radar detectors are expensive to install and maintain and this analysis is useful to determine the possibility to reduce the number of radar detector sites. In this section the results from these initial verifications and estimation comparisons are presented. The results have previously been presented in Allström et al. (2012b).

5.2.1 Experimental setup

In 2010, a small field test involving ten probe vehicles with GPS-equipped mobile phones was carried out in Stockholm, see Allström et al. (2012a). The field test took place on the main highway passing through Stockholm and the vehicles drove back and forth on a specified road stretch during morning and afternoon peaks on the 16th and 23rd of March. The GPS-data from the field trial has been map matched and travel times and trajectories have been extracted. In this verification of the estimated traffic state, attention has been on the northbound traffic during the morning peak period between 7:30 and 9:30 and in the afternoon between 16:00 and 17:00 for the 16th of March. During the morning peak, there was some heavy congestion while the afternoon data showed only moderate levels of congestion. The studied road stretch is approximately 4.5 km and included 15 radar detector sites, see Figure 24. The spacing between the detector sites on this stretch varies between 150 and 450 meters. The posted speed limit at the studied road stretch is 70 km/h and it has four on-ramps and four off-ramps.
5.2.2 VERIFICATION OF THE ESTIMATED SPACE-TIME SPEED CONTOUR PLOT

In this step of the verification, the GPS-trajectories and the estimated traffic state is compared, see Figure 25. The number of trajectories is limited, approximately 20 during a two hour period, making it difficult to draw any definitive conclusions. It is, however, possible to identify a number of phenomena that appears in both the estimated space-time speed contour plot and the plot of the trajectories. The trajectories are consistently smoother during the afternoon due to the free flow conditions. During the morning peak however, the trajectories are more deviant illustrating the stop and go conditions that are consistent with congestion.

In Figure 25c there is a green area in the space-time speed contour plot between 9:00 and 9:30 for the first 2.5 km. This indicates a situation where the queue dissolves and the velocities are higher and more stable. This situation is also evident in Figure 25a where the inclination of the trajectories becomes steeper suggesting higher velocity. Furthermore, in the top left hand corner of Figure 25b and Figure 25d there are corresponding areas in both figures where congestion occurs and causes a reduction in velocity. Even though the number of GPS-trajectories is limited, the data is correspondent with the characteristics of the space-time speed contour plot of estimated velocities.
The fixed sensors are quite densely spaced on the studied road stretch allowing the possibility to eliminate a number of sensors and instead use these as a de facto ground truth. In Figure 26, the results from two scenarios with different numbers of sensors are compared. Figure 26a visualizes the estimated traffic state using all sensors, while Figure 26b visualizes a scenario where the estimated traffic state is based on every other sensor. As expected, the figures indicate that as long as the traffic is in the same state (either congested or free flow) the estimations are very similar. However, when the congestion dissolves the scenario with fewer sensors takes considerably longer time to identify the change in state. This is not surprising since the estimation is only based on static point speed measurements and the model is not calibrated. Hence, in most cases the model is dependent on measurements in order to change state.

Figure 26c illustrates the absolute difference between Figure 26a and Figure 26b, i.e. the difference between the scenario where all sensors are used and the one where only every other sensor is used. In all cells that are black the difference in estimated velocity is larger than 10 km/h. As observed earlier, great differences occur when the traffic state changes at a position of approximately 1.5 km and time 9.00. There is also one cell located around position 3.5 km where the difference between the two scenarios is large.
Figure 26d visualizes the absolute difference between the measurements and the estimated velocities from the scenario where only every other sensor is used. All cells where no measurements are available are colored white in the figure. The measured velocities are assigned to the corresponding network cell containing the detector site. Each cell is approximately 150 meters long throughout the studied road stretch due to several on and off-ramps and densely spaced detector sites. For some sections of the studied road stretch the sensors are so densely spaced that most cells include a radar detector site. For cells where measurements have been used in the estimation there are only minor differences between measured and estimated velocities. However, the differences between estimated and measured velocity for cells that have not been included in the estimation vary considerably. This implies that the uncalibrated model is very dependent on measurements. The differences are generally smaller than 10 km/h but there are some areas where the difference is much larger. These areas are basically the same as those in Figure 26c and can be explained by the change in traffic state which the uncalibrated MMS-model has trouble catching.
5.2.3 VERIFICATION OF ESTIMATED TRAVEL TIMES

While the space-time speed contour plots are interesting to study for a comparison of estimated traffic state for different scenarios, the most important output from the model is the estimated travel time. In Figure 27, the estimated travel time from the uncalibrated MMS-model for three different scenarios has been compared with the travel time estimated by the current system implemented at the traffic management center in Stockholm, as well as with measured travel times from the field test described previously. The estimations generated by the current traffic management system are based on the same data as the estimations generated by the implemented MMS-model, i.e. only data from the radar detectors. The method used is a smoothing procedure and an arithmetic mean is used to aggregate travel times over five minute intervals, see Strömgren et al. (2004).

![Figure 27: Comparison of estimated travel times from the uncalibrated model, measured travel times and estimations made by the current system.](image)

From the figure it can be concluded that, irrespective of whether all sensors or only one sensor per kilometer is used, there is little difference in the estimated travel times. Furthermore, the estimated travel times from the MMS-model follow the same trend as the measured travel times from the field test. It is also noted that the estimations made by the current system at the traffic management center in Stockholm react more slowly to changes in state. The fact that the estimations generated by the current system at the traffic management center in Stockholm react more slowly to changes in state can be explained by the use of a smoothing procedure and that an arithmetic mean is used to aggregate travel times and that it is updated only once every five minutes. Furthermore, in the current system used at the traffic management center, travel times on a specific link are estimated by calculating an arithmetic mean from the available measurements for that specific link. For the studied road section there are 19 links and the link length varies between 20 meters and 1 km. In contrast, the implemented MMS-model calculates a space mean speed for each of the 35 cells that represents the studied road stretch. This makes the MMS-model more detailed.
and better equipped to follow trends compared to the current traffic management system given the same data sampling frequency.

5.2.4 CONCLUSIONS FROM THE VERIFICATION AGAINST GPS-DATA
According to the results from the verification of the estimated traffic state, the MMS-model has a good potential to improve the estimation of travel times in the greater Stockholm area. Given that the MMS-model used has not undergone a thorough calibration against the available data in Stockholm, the results are interpreted as very promising. The estimated travel times generated by the MMS-model correspond relatively well to the measured travel times of the earlier field test, even when the number of sensors is reduced. This suggests that the model is robust with regard to the underlying data sources used.

However, even though the initial results are promising they also indicate that there is a potential to improve the estimation. The model appears to react a bit slow to changes in traffic state, something that probably is due to the current parameter setting in the MMS-model. There are parameters related to the fundamental diagram as well as the boundary flows and the Ensemble Kalman Filter that needs to be calibrated for Stockholm conditions.

5.3 CALIBRATION AND VALIDATION OF THE MMS-MODEL
In order to improve the accuracy of the estimations and the robustness of the model, the MMS-model has to be calibrated and validated. Here, we present a method for calibrating the parameters related to the fundamental diagram and a calibration framework where the model parameters are adjusted iteratively. The results from the calibration of the MMS-model for a road stretch in Stockholm are also presented. The calibration framework was first developed on data from California and presented in Fransson and Sandin (2012). Finally, the results from the validation of the calibrated model are presented.

The MMS-model contains several parameters that can be adjusted in the calibration process. The pre-processing filter (briefly described in Section 4.5), the CTM-v (described in Section 4.3.1), and the EnKF (described in Section 4.3.3) all include parameters that needs to be calibrated. The pre-processing filter and the CTM-v both include link specific parameters related to the hyperbolic-linear version of the fundamental diagram. The CTM-v also contains parameters describing the boundary flows, i.e. the capacity of the sinks, the demand of the sources and allocation parameters for each junction, also entitled split ratios. Finally, the EnKF includes parameters describing the noise and error from the model and the measurements. In Table 4 all parameters that are calibrated are presented.
Table 4: Parameters that are calibrated and how they are grouped in this thesis.

<table>
<thead>
<tr>
<th>Parameter set</th>
<th>Parameter</th>
</tr>
</thead>
</table>
| Fundamental diagram | Free flow speed, $v_f$ (m/s)  
|                 | Critical speed, $v_c$ (m/s)       
|                 | Shockwave speed, $w_f$ (m/s)      
|                 | Jam density, $\rho_{\text{max}}$ (veh/mlane) |
| Boundary flow  | Capacity of sinks (veh/s)         
|                | Demand of sources (veh/s)         
|                | Split ratios                      |
| EnKF           | EnKF Model noise mean (m/s)       
|                | EnKF Model noise St Dev (m/s)     
|                | EnKF Measurement noise Mean (m/s) 
|                | EnKF Measurement noise St Dev (m/s) |

The developed calibration framework, used for calibration of the boundary flow and EnKF parameters, enables calibration of parameters in a traffic model based on travel time measurements. The main motive for using travel time measurements in the calibration process is that there are several cost efficient options to measure travel times available, e.g. mobile Bluetooth detectors and floating car data. Another motive is that travel times can fully capture the aggregated traffic state over a given spatial domain. The evaluation of Bluetooth travel times presented in Section 3.4 further confirmed that Bluetooth is a valid method for collecting travel times. However, a conclusion from this evaluation was that the collected travel times has to be filtered before it is aggregated. Hence, an interquartile range filter (IQR-filter) has been used in order to filter the Bluetooth data that the model is calibrated against. In the IQR-filter, all measurements within a five-minute period are grouped together and all measurements that does not lie within the posted (interquartile) range are considered as an outlier. For a more detailed description of the IQR-filter, see Upton and Cook (1996). Figure 28 shows the resulting filtered and aggregated travel times for one of the road stretches in the test site used in the calibration.

Figure 28: Filtered and aggregated travel times from Bluetooth for one day for one of the road stretches used.
It is possible to calibrate all parameters of the MMS-model using only travel times. However, to reduce the risk of ending up with parameters that has little connection to the physical system and to reduce the number of parameters to calibrate in each process, the calibration process has been divided into two phases: calibration of fundamental diagrams and a calibration framework where the boundary flows and EnKF-parameters are calibrated. The first phase, described further in Section 5.3.1, is performed using measurements from stationary sensors while the second phase, described in Section 5.3.2, is performed using the travel time measurements from Bluetooth.

5.3.1 Calibration of the Fundamental Diagram Parameters
Each link in the network is assigned a fundamental diagram and each fundamental diagram is defined by the four parameters free flow speed \( \nu_f \), critical speed \( \nu_c \), shockwave speed \( \nu_s \) and jam density \( \rho_{\text{max}} \). However, the critical speed is calculated as the free flow speed minus the shockwave speed and is therefore not considered in the calibration process. The parameters in each of the fundamental diagram are individually calibrated using one month of speed and flow data from radar detectors from March 2013. Optimal parameter values are found using the Compass Search method (Kolda, Lewis and Torczon, 2003) with the objective function

\[
\min \left\{ \sum_{i=1}^{n} (\rho_i - \bar{\rho}_i)^2 + \sum_{i=1}^{n} (\nu_i - \bar{\nu}_i)^2 \right\}, \tag{5.1}
\]

where \( \rho_i \) is the density for radar station \( i \) computed from the speed and flow radar measurements, \( \nu_i \) is the observed speed for radar station \( i \), and \( \bar{\rho}_i \) and \( \bar{\nu}_i \) are the corresponding values computed from the fundamental diagram, implicitly given by the parameter values.

The link specific relation between the speed \( \nu \) and the density \( \rho \) in the hyperbolic-linear fundamental diagram are given by Equation 4.2 and the continuity requirement in Equation 4.3, both in Section 4.3.1.

For links where no radar measurement to calibrate the fundamental diagram parameters against is available, the parameters of a nearby and similar link with respect to number of lanes, speed limit and closeness to on- and off-ramps are used. For links with several radar stations a mean value based on all radar measurements on that specific link has been calculated.

5.3.2 Calibration of Boundary Flows and EnKF Parameters
Besides the fundamental diagram parameters the parameters related to the boundary flows and the EnKF needs to be calibrated. The iterative calibration framework used to calibrate the boundary flows and EnKF parameters is illustrated in Figure 29. In the figure the calibration process can be found within the dashed line while the previously described MMS-model is outside the dashed line. The figure consists of three main parts: the measurement loader, the MMS-model (denoted Highway model in the figure) and the calibration framework. In the measurement loader the collected radar data and travel times from Bluetooth are filtered. The filtered radar and Bluetooth data are later fed to the MMS-model and the calibration process respectively. The output from the MMS-model is aggregated in time and space related to the placement of the Bluetooth detectors and then compared with the aggregated Bluetooth travel times in the calibration process. If the
convergence criteria is met the iteration is stopped, otherwise the MMS-model parameters are adjusted and a new iteration is performed.

In Figure 30 and Figure 31, the process of calibrating the boundary flow and EnKF-parameters is illustrated in another way. In Figure 30 the spatiotemporal aggregation process is described and the top layer represent the radar measurements (and in this example also point speed measurements from probe vehicles) available at different space-time coordinates. The middle layer represent the output from the MMS-model based on the measurements in the top layer, while the bottom layer represents the aggregated output from the MMS-model where the space mean speeds corresponding to the collected Bluetooth travel times are calculated. The travel times from the MMS-model are computed by simulating fleets of vehicles between the locations of each Bluetooth sensor pair.

In Figure 31 the final step of a single iteration of the calibration loop is illustrated where the current parameter set is evaluated by comparing the aggregated output from the MMS-model (top layer in Figure 31 and bottom layer in Figure 30) to the ground truth from Bluetooth measurement (middle layer). The bottom layer symbolizes the difference between the two top most layers and a performance metric that expresses this difference is computed.
Figure 30: Depiction of the increased level of processing made during one calibration step (Fransson and Sandin, 2012).

Figure 31: The final step of a single iteration of the calibration loop where the current parameter set is evaluated by comparing the aggregated model velocity field (top layer) to the ground truth (middle layer) (Fransson and Sandin, 2012).
There are several methods available that can be used to calibrate a set of parameters. In the framework presented here we use Complex method for the process where the MMS-model output is compared with the measured travel times and the parameters are adjusted if the convergence criteria is not met. However, the method used for this process is exchangeable in the presented calibration framework. The Complex method is a gradient free search algorithm presented by Box (1965) and Andersson (2001). In the literature the Complex method is one of the methods that have been used for solving parameter estimation problems related to macroscopic traffic modeling, see for instance Cremer and Papageorgiou (1981) and Kotsialos et al. (2002). In the presented calibration framework the Complex method is used to calibrate the parameters related to the boundary flows (sinks, sources and split ratios) and EnKF. Sources and sinks are considered to be ghost cells that set the boundary conditions at the network edges. The sources states the number of vehicles entering the network at each source, while the sinks state the number of vehicles that can leave the network at each sink for a given time period. They are both given in vehicles per second per lane. At each off-ramp a split ratio needs to be set, it is a parameter used to determine the proportion of vehicles of the major stream that wishes to leave the mainline via the off-ramp, and how many that wishes to continue forward. Finally, there is a set of parameters that are part of the EnKF-filter, the model and measurement noise mean and standard deviation described in Section 4.3.3. Thus, the number of parameters depends on network size (the number of on- and off-ramps) and the number of data sources.

The Complex method is used to solve problems on the general form

$$\min_{\mathbf{x}} \ f(\mathbf{x})$$

$$\text{s.t.} \quad x^l \leq x_i \leq x^u, \quad i = 1, \ldots, N,$$

(5.2)

where $x^l$ and $x^u$ denote the lower and upper bound for variable $x_i$, respectively. In our case the variable $x_i$ represent the $N$ parameters that are adjusted and calibrated. When the search algorithm is initiated, a complex defined by at least $K$ points of dimension $N$ is created and each point represent a set of parameter values. The purpose of choosing a complex with $K > N + 1$ points is to ensure that the complex keeps its dimensionality when the points are moved around in the $N$-dimensional space during the iterations. If the complex has too few members it could flatten itself against the first constraint that is reached during the movement in the $N$-dimensional space, see Box (1965). The objective function $f$ is evaluated for each of the $K$ points. In relation to the model, this equals running the MMS-model, computing travel times and compare the estimated travel times to measured and filtered travel times from Bluetooth. In the objective function the estimated and measured travel times are compared by the mean absolute percentage error (MAPE) defined as

$$f(x) = \frac{1}{|\mathcal{T}|} \sum_{j \in \mathcal{T}} \sum_{t \in \mathcal{I}} \left| \frac{y_{jt} - \hat{y}_{jt}(x)}{y_{jt}} \right|,$$

(5.3)

where $\mathcal{T}$ is the set of Bluetooth routes, $\mathcal{I}$ is the set of time intervals over which the travel times are aggregated, $y_{jt}$ is the observed and aggregated travel time over Bluetooth route $j$ at time interval $t$ and $\hat{y}_{jt}$ is the estimated travel time, dependent on the parameters $x$. The difference between the best and the worst point is then computed and compared to a convergence criteria. If the
convergence criteria is satisfied, the search is completed and the point with the best performance is chosen as the model parameter set. But as long as the criterion is unsatisfied the worst point, i.e. the worst set of parameters, will be moved over the centroid of all the other points in the domain, where it is once again evaluated. The search can also be completed if the number of iterations has reached a maximum value specified by the user. The flow chart of the Complex method is presented in Figure 32. For a more in depth description of the calibration framework and the Complex method the reader is referred to Fransson and Sandin (2012) and Box (1965), respectively.

Figure 32: The Complex method presented in the form of a flow-chart (Fransson and Sandin (2012)).
It has been noted by Andersson (2001) that this kind of search algorithm can get stuck around a local optimum. The algorithm was implemented in a way that moves the same point gradually as long as it continues to be the worst point. A random value is added to the point in order to avoid a situation where the worst point ends up too close to another point. This works well for this application of the Complex method and also increases the search effort compared to a reflection through the centroid.

5.3.3 EXPERIMENTAL SETUP

To analyze the impact of the calibration framework and different parameter settings, data collected along the main highway in Stockholm has been used. In this section, the available data and the test site are presented together with the setup of the analyzed calibration scenarios.

The test site is a 5 km long segment of the southbound part of the highway just north of Stockholm city. This particular site was chosen since both point measurements from radar detectors and travel times from Bluetooth sensors were collected here during one week in March 2013. See Figure 33 for an overview of the placement of the two detector types. The studied road stretch is divided into ten links, as marked in the figure.

Figure 33: Each red dot in the figure represents a radar detector; the red dots marked with a cross represent radar detectors that did not deliver any data during the field trial. Each Bluetooth sensor is illustrated with an antenna symbol.
The radar detectors are mounted over each lane of the highway and collect speed and traffic flow aggregated at one-minute intervals. The Bluetooth sensors were placed every 500-1000 m and collected travel times from all active Bluetooth devices that passed two sensors. However, not all radar stations were active during the field trial, the seven radar stations marked with a cross in Figure 33 did not deliver data during the field trial. In the middle of the test site the speed limit changes from 90 km/h to 70 km/h while the number of lanes varies between two, three and four along the road stretch. There are three on-ramps and three off-ramps on the road segment and adding the start and end of the network there are in total four sinks and four sources in the network.

As mentioned in Section 5.3, three different sets of parameters that can be calibrated have been identified, see Table 4. To analyze which sets of parameters that are most critical to calibrate, eight different scenarios has been set up, see Table 5.

Table 5: Description of analyzed scenarios.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Time period (21st of March 2013)</th>
<th>Fundamental diagram</th>
<th>Boundary flows</th>
<th>EnKF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sinks</td>
<td>Sources</td>
</tr>
<tr>
<td>1</td>
<td>06-22</td>
<td>Default values</td>
<td>Default values</td>
<td>Default values</td>
</tr>
<tr>
<td>2</td>
<td>06-22</td>
<td>Calibrated</td>
<td>Default values</td>
<td>Default values</td>
</tr>
<tr>
<td>3</td>
<td>06-22</td>
<td>Default values</td>
<td>Default values</td>
<td>Calibrated</td>
</tr>
<tr>
<td>4</td>
<td>06-22</td>
<td>Calibrated</td>
<td>Default values</td>
<td>Default values</td>
</tr>
<tr>
<td>5</td>
<td>07-09</td>
<td>Default values</td>
<td>Default values</td>
<td>Default values</td>
</tr>
<tr>
<td>6</td>
<td>07-09</td>
<td>Calibrated</td>
<td>Default values</td>
<td>Default values</td>
</tr>
<tr>
<td>7</td>
<td>07-09</td>
<td>Default values</td>
<td>Default values</td>
<td>Calibrated</td>
</tr>
<tr>
<td>8</td>
<td>07-09</td>
<td>Calibrated</td>
<td>Default values</td>
<td>Calibrated</td>
</tr>
</tbody>
</table>

Two of the scenarios, Scenario 1 for the time interval of 06:00 to 22:00, and Scenario 5 for 07:00 to 09:00, corresponds to "default values". Here, default values correspond to the values used in the uncalibrated model, i.e. parameters set manually based on experience mainly from California. Scenario 1 and 5 are used as base scenarios in order to analyze the improvement of the traffic state estimation that the calibration process result in. In Scenario 2 and Scenario 6 only the parameters for the link specific fundamental diagrams are calibrated. In scenario 3 and 7 the sources, split ratios and EnKF parameters has been calibrated, while the default values have been used for the fundamental diagram parameters and the sinks. Scenario 4 and Scenario 8 corresponds to the situation where all available parameters are calibrated.

All parameters are static in the calibrated implementation of the model. In this context, static imply that the values of the parameters are always the same and do not depend on the time of the day. To analyze the possible improvement that a more dynamic set of parameters could infer, these parameters has been calibrated both against one day of data, but also against data only from the morning peak, 7am to 9am.
5.3.4 Calibration Results

In this section the results from each scenario is presented together with the best parameter setting that the calibration process resulted in. The results have previously been presented in Allström et al. (2014).

Figure 34 exemplifies the results of the calibration of a fundamental diagram. The data points (+) in the figure are speed observations and the calculated density for one radar station for the whole month of March 2013, for the times 06:00 to 23:00. The density is calculated using the relationship in Equation 2.6. The green line, made up by circles, is the calibrated hyperbolic-linear fundamental diagram for this specific radar station.

In Table 6, the default and calibrated values for all fundamental diagram parameters for all links in the studied network are presented. As can be seen in the table the calibration increases the free flow speed $v_f$ and consequently the capacity, i.e. the critical speed $v_{cr}$, since the critical speed is calculated as

$$v_{cr} = v_f - w_f,$$

where $w_f$ is the shockwave speed. In the default case the free flow speed equals the posted speed limit. The calibration process also decreases the shockwave speed for all links, meaning that the shockwaves propagate backwards slower than in the scenario with default values. Finally, the jam density is increased for some links and decreased for other.

Link 4, 9 and 10 do not have any radar detectors, hence, the value for these links are taken from nearby similar links with respect to number of lanes, speed limit and closeness to on- and off-ramps. The default values are used in scenario 1, 3, 5 and 7 while the calibrated values are used in scenario 2, 4, 6 and 8.
Table 6: Default and calibrated fundamental diagram parameters for all links in the studied network.

<table>
<thead>
<tr>
<th>Link</th>
<th>Free flow speed (m/s)</th>
<th>Critical speed (m/s)</th>
<th>Shockwave speed (m/s)</th>
<th>Jam density (veh/m/lane)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Default</td>
<td>Calibrated</td>
<td>Default</td>
<td>Calibrated</td>
</tr>
<tr>
<td>1</td>
<td>26.34</td>
<td>27.09</td>
<td>20.98</td>
<td>23.43</td>
</tr>
<tr>
<td>2</td>
<td>26.34</td>
<td>27.32</td>
<td>20.98</td>
<td>23.81</td>
</tr>
<tr>
<td>3</td>
<td>26.34</td>
<td>27.91</td>
<td>20.98</td>
<td>23.62</td>
</tr>
<tr>
<td>4</td>
<td>20.79</td>
<td>23.30</td>
<td>15.42</td>
<td>20.03</td>
</tr>
<tr>
<td>5</td>
<td>20.79</td>
<td>22.37</td>
<td>15.42</td>
<td>18.67</td>
</tr>
<tr>
<td>6</td>
<td>20.79</td>
<td>21.76</td>
<td>15.42</td>
<td>19.71</td>
</tr>
<tr>
<td>7</td>
<td>20.79</td>
<td>20.97</td>
<td>15.42</td>
<td>18.85</td>
</tr>
<tr>
<td>8</td>
<td>20.79</td>
<td>23.30</td>
<td>15.42</td>
<td>20.03</td>
</tr>
<tr>
<td>9</td>
<td>20.79</td>
<td>21.76</td>
<td>15.42</td>
<td>19.71</td>
</tr>
<tr>
<td>10</td>
<td>20.79</td>
<td>21.76</td>
<td>15.42</td>
<td>19.71</td>
</tr>
</tbody>
</table>

Table 7 shows the default values and the calibrated values for scenario 3, 4, 7 and 8 for the parameters describing the boundary flows and the mean and standard deviation for the measurement and model noise in the EnKF filter. From the table it can be concluded that the default values in many cases are very rough estimates with little or no relation to the reality. For example, the split ratios are set to 1 for all off-ramps which imply that no vehicles take off at the off-ramps. Furthermore, the demand at the on-ramps are set to zero for all on-ramps except for the link in the beginning of the network. The reason for this rough parameter setting is probably that in the original implementation of the model the estimation was based on both fixed sensors and an abundance of probe data that heavily reduces the importance of the parameter setting. The calibration process naturally increases the on-ramp flow for all on-ramps besides the first link in the network where the demand is reduced. Furthermore, the calibration lowers the sink capacity that in the default case is set to five veh/s for all off-ramps. However, the sink capacity has proven to have little effect on the quality of the estimation and is excluded from the calibration process in scenario 3 and 7 in order to reduce the number of parameters to be calibrated. The calibration process also adjusts the model and measurement noise in the EnKF, however there is no obvious pattern in how the parameters are adjusted. As mentioned previously, in the default case no vehicles are assumed to take off at the off-ramps. The calibration adjusts the split ratios for all off-ramps to around 0.6-0.9, which means that between 10 % and 40 % of the vehicles takes off at each off-ramp.
Table 7: Default parameter values and calibrated values for scenario 3, 4, 7 and 8 for the boundary flow and EnKF parameters.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Boundary flow</th>
<th>EnKF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sinks</td>
<td>Sources</td>
</tr>
<tr>
<td>Default values</td>
<td>5, 5, 5</td>
<td>1.6, 0, 0, 0</td>
</tr>
<tr>
<td>Calibrated values Scenario 3</td>
<td>0.1, 1.5</td>
<td>0.4, 1.3</td>
</tr>
<tr>
<td>Calibrated values Scenario 4</td>
<td>1.1, 1.0</td>
<td>0.7, 1.2</td>
</tr>
<tr>
<td>Calibrated values Scenario 7</td>
<td>1.4, 0.4</td>
<td>1.0, 0.9</td>
</tr>
<tr>
<td>Calibrated values Scenario 8</td>
<td>2.2, 2.6</td>
<td>2.9, 3.3</td>
</tr>
</tbody>
</table>

In scenario 3 and 7, when the default fundamental diagram parameters are used, the calibration process tries to compensate the low free flow speed with a reduction of vehicles in the network in order to increase the travel times. This result in split ratios where a larger number of vehicles exit the highway and the demand at the sources are in general lower than in the scenarios with calibrated fundamental diagram parameters. That the free flow speed is too low when the default parameter setting is used is obvious in Figure 35, where the resulting travel time estimation during one day (06:00 to 22:00) is illustrated together with Bluetooth travel time measurements. In scenario 1, when the default parameters are used, the estimated travel times are continuously too high compared to the Bluetooth travel times, especially in free flow. In scenario 3, when the default fundamental diagram parameters and sink capacities are used and all other parameters are calibrated the travel time is too high in free flow but too low in congestion. A consequence of the fact that the calibration process reduces the vehicles in the network in order to compensate for the low free flow value but of this is that the travel time is too low in congestion because the network is too empty. In scenario 2, when only the fundamental diagram parameters are calibrated, the estimated travel times follows the Bluetooth travel times well in free flow but overestimates the travel times in congestion. This is probably due to the high split ratios in the default case, which implies that no vehicles exit the highway. In scenario 4, when all parameters are calibrated, the estimated travel times corresponds well to the measured travel times from Bluetooth. A number of estimation performance metrics confirming this is presented later on.
Another way to evaluate the calibration process is to study the space-time speed contour plot. In Figure 36 and Figure 37, the space-time speed contour plot of scenario 1 and scenario 4 is presented. From the figures it can be concluded that for scenario 1, when the default values are used, the space-time speed contour plot looks very strange. The red parts illustrating congestion, which is expected to be coherent, is split in an unrealistic way. Furthermore, it is obvious that the uncalibrated model cannot estimate the state between the measurements in a credible way. In Figure 37, where the space-time speed contour plot of the fully calibrated scenario 4 is presented, the shape of the congestion looks more realistic and the estimated speeds between the measurements are as expected.
The fact that the implemented model uses static parameters limits the possibility for the calibration to adapt to changes in traffic demand. In Figure 38 the measured flow during one day for one source is shown. From the figure it can be noted that the demand is not static, actually it varies a lot during the day, which indicate that the model might be improved if dynamic modeling of the demand is used. This is further analyzed by evaluating the model for a shorter time interval. Figure 39 shows the result during the morning peak hour for the calibration scenarios 5-8 and the corresponding Bluetooth travel time measurements.
Figure 38: Observed flow for one radar detector for one day.

Figure 39: Estimated travel times from the MMS-model for scenario 5-8 and travel times from Bluetooth detectors.

From Figure 39 it can be concluded that the estimation in scenario 8 where all parameters are calibrated follows the measured travel times well. However, in scenario 7, when the default fundamental diagram parameters and sink capacities are used and all other parameters are calibrated, the estimated travel times also correspond well with the measured Bluetooth travel times. In scenario 5 and 6, where default values has been used with exception of scenario 6 where calibrated fundamental diagram parameters has been used, the travel times are overestimated. This can be compared to the corresponding scenario 1 and 2 where the travel times also are
overestimated. Hence, one conclusion is that if only the fundamental diagram parameters are calibrated the travel times in congestion are overestimated.

In Table 8 three performance metrics are presented for each calibration scenario. The studied performance metrics are the Mean Absolute Percentage Error (MAPE), the Mean Percentage Error (MPE) and the Root Mean Square Error (RMSE). The MAPE, which provides an indication of the bias and variance of the relative error is defined as

\[
MAPE = \frac{100 \%}{|T|} \sum_{t \in T} \left| \frac{y_t - \hat{y}_t}{y_t} \right|.
\]  

(5.5)

where \(y_t\) is the average filtered measured travel time from Bluetooth during aggregation period \(t\), \(\hat{y}_t\) is the average estimated travel time from the calibrated MMS-model during aggregation period \(t\) and \(T\) is the set of time intervals over which the travel times are aggregated. It should be noted that this is not the same MAPE value as the MAPE value used in the calibration process. In the calibration process each route between consecutive Bluetooth sensors, i.e. each subsection, is analyzed, here the travel time for the whole road stretch is analyzed. In Table 8, the travel times for the whole road stretch is found in the row called MAPE, while the MAPE value from the calibration is denoted MAPE\(_C\).

The MPE, which indicates structural bias, is defined as

\[
MPE = \frac{100 \%}{|T|} \sum_{t \in T} \frac{y_t - \hat{y}_t}{y_t}.
\]  

(5.6)

Finally, the RMSE, which is an aggregate measure of the model accuracy, is defined as

\[
RMSE = \sqrt{\frac{1}{|T|} \left( \sum_{t \in T} (\hat{y}_t - y_t)^2 \right)}.
\]  

(5.7)

From the table it can be concluded that the calibration process increases the accuracy of the estimation. Although, given that the accuracy of estimation based on the default parameters was rather low, to increase the accuracy is relatively straightforward. However, the resulting MAPE in scenario 4, 7 and 8, which is under 5 \%, demonstrate a high accuracy for the estimations based on the calibrated parameters. This can be compared with the results from the evaluation of travel times from Bluetooth sensors in Section 3.4 where the MAPE-value describing the difference between the studied data collection methods also was around 5 \%. It is interesting to note that the RMSE and MPE value for scenario 3, where only sources, split ratios and the EnKF parameters are calibrated, is similar to the RMSE and MPE value for scenario 4 when all parameters are calibrated. However, there is a large difference between the MAPE values for the two scenarios. The reason for this can be seen in Figure 35 where scenario 3 is overestimating the travel time in free flow and underestimating it in congestion. The difference between the Bluetooth travel times and scenario 3 is captured by the MAPE but for RMSE and MPE the overestimation in free flow and underestimation in congestion cancel each other out. The MAPE\(_C\), the MAPE value from the calibration process, where each section between the Bluetooth sensors are analyzed individually, is consistently higher than the MAPE value when the travel time for the whole road stretch is
analyzed. The reason for this is that in the calibration process the travel time is underestimated for some sections and overestimated for others. When looking at the whole road stretch these differences cancel each other out.

Table 8: The calculated performance values for the eight calibration scenarios.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
<th>Scenario 5</th>
<th>Scenario 6</th>
<th>Scenario 7</th>
<th>Scenario 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPE (%)</td>
<td>52.4</td>
<td>9.1</td>
<td>11.0</td>
<td>4.9</td>
<td>27.7</td>
<td>16.6</td>
<td>3.9</td>
</tr>
<tr>
<td>MAPLe (%)</td>
<td>34.0</td>
<td>12.6</td>
<td>14.6</td>
<td>10.4</td>
<td>30.5</td>
<td>19.9</td>
<td>11.5</td>
</tr>
<tr>
<td>MPE (%)</td>
<td>33.2</td>
<td>6.7</td>
<td>1.8</td>
<td>1.3</td>
<td>20.7</td>
<td>11.7</td>
<td>-1.6</td>
</tr>
<tr>
<td>RMSE (s)</td>
<td>171</td>
<td>87</td>
<td>53</td>
<td>58</td>
<td>151</td>
<td>95</td>
<td>25</td>
</tr>
</tbody>
</table>

5.3.5 Validation of the calibrated model

Since there are several parameters to adjust in the model, it is not surprising that it is possible to get a good fit with the Bluetooth travel times. However, to make sure that the calibration does not over-fit the model a validation of the calibrated parameters has been made where travel times with the calibrated parameters has been estimated using data from another date than the one used in the calibration. In the calibration we used data from the 21st of March while data from the 25th of March is used for validation.

If the calibrated parameters from scenario 4 is used to estimate the travel times from 6.00-22.00 on the 25th of March a MAPE value of 4.6 % is achieved, see Figure 40. For the other scenarios the MAPE value from the validation are also very similar to the result from the calibration. Even though the traffic situation is similar during the day used for calibration compared to the day used for validation, this further validates that the calibrated model is suitable to use for estimations on the highway network and that a high accuracy can be achieved for the estimations.
5.4 Conclusions from the Calibration and Validation

A process for calibrating the MMS-model based on two stages has been developed. In the first stage the parameters related to fundamental diagrams are calibrated, and in the second stage a calibration framework based on a search method is applied to the problem of finding the best possible parameters for the boundary flows and the EnKF. The results from the calibration and validation show that once the model is calibrated, the estimated travel times correspond well with the ground truth travel times collected from Bluetooth sensors. To achieve a good accuracy all set of parameters has to be calibrated, the parameters related to the fundamental diagram as well as the parameters related to the boundary flows and the EnKF. The capacity of the link, i.e. the critical speed, is integral in order to capture the shockwaves correctly and naturally, the free flow speed is crucial when the traffic is in free flow. For the estimation problem, when calibrating the parameters in a two-stage process, errors in the fundamental diagram parameters can to a certain extent be compensated with fictive changes in the demand. This is not a major problem for travel time estimation, but becomes much more important when using the model for travel time prediction. When travel times are estimated the model is fed with continuous measurements that compensates for model errors but when travel times are predicted by running the CTM-v forward no measurements are available. The prediction problem will be analyzed further in the next chapter.
6 SHORT-TERM TRAFFIC STATE PREDICTION

This chapter begins with Section 6.1 and Section 6.2, where short-term traffic state prediction is introduced and previous work on short-term traffic state prediction is presented. This is followed by Section 6.3 where the modifications we have made that were necessary for the MMS-model to not only estimate the current traffic state but also predict the future traffic state are described. In Section 6.4 and Section 6.5 the experimental setup and the initial results from when we have used the MMS-model together with a neural network to predict the traffic state are presented. The chapter is concluded in Section 6.6.

6.1 INTRODUCTION TO SHORT-TERM TRAFFIC STATE PREDICTION

A prediction of the traffic state is useful for the traffic management center when they decide on which information to give to the travelers and how to manage the traffic. The current trend is that the traffic management aim at working more proactive to avoid congestion, however, for this to be feasible a predicted traffic state is essential. For the traveler the predicted experienced travel time, i.e. the time it will take between two points given that the trip starts at a given time, is more relevant than a travel time only based on the estimated current traffic state. However, a prediction of the future traffic state is in general based upon the latest traffic state estimate. The prediction of the traffic state is usually considered to be a more complex problem than estimating the traffic state, since the future is unknown. Short-term is in this thesis considered as prediction horizons up to 30 minutes in to the future. Although it should be noted that since we predict the experienced travel time we have to predict the future traffic state up to 60 minutes ahead when we predict the travel times for journeys taking place 30 minutes in to the future.

The literature consist of several examples of traffic state prediction applications. Primarily, it is a critical component for real-time traffic control and management, see for example Bajwa et al. (2005), Chen and Grant-Muller (2001), Van Lint et al. (2005) and Wang et al. (2006). The ability to accurately predict the traffic state could increase the ability of the traffic managers to take action before the system reaches congestion and at least limit the effects of it.

Accurate travel time predictions are also useful to travelers when they decide their departure time and/or route and transport mode. It is a general assumption that accurate travel time information can be supplied to users so that they can make an adequate route choice and choice of departure time. Such an ability would result in a better experience for the individual and more reliable travel times, see Van Lint (2006). Travel-time prediction is also a useful performance indicator of the traffic system that is easy to understand, see for example Bajwa et al. (2005), Billings and Yang (2006) and Chen and Grant-Muller (2001).

A traffic state prediction system would also change a traveler's information system from being reactive to being proactive. A reactive system informs travelers about the current state of the road or the historical situation, whereas a proactive system state what the actual traveler experience will be given their departure time. If the travelers' information systems are transformed from reactive to proactive it would increase the confidence in the system among the travelers as well as traffic managers, see for example Bajwa et al. (2005), Van Lint et al. (2005) and Van Lint (2006).
6.2 Previous Work on Short-Term Traffic State Prediction

An abundance of models used for short-term traffic state prediction can be found in the literature. For example, Van Hinsbergen et al. (2007) identified 45 different kinds of short-term prediction models (short-term being between 5 and 180 minutes in most cases) from more than 130 studied references.

The different approaches for short-term traffic state prediction can be categorized based on which family of models it is based on. Van Lint and Van Hinsbergen (2012) used three main families to categorize short-term traffic state prediction models: naïve, parametric and non-parametric. Naïve methods include using the latest known value, a historic average or a simple clustering method, and are not further studied in this survey. From the literature it can be concluded that the accuracy of these models is usually low and basically any parametric or non-parametric method have a higher accuracy than these methods.

In the following sections we give a short description of the most commonly used models that are classified as parametric or non-parametric models.

6.2.1 Parametric Models

Parametric models are characterized by the presence of model parameters with a predetermined structure, in general based on observed behavior. These parameters need to be calibrated according to empirical data. Traffic simulation models (micro-, meso- and macroscopic) are the main family of models within the parametric models for short-term traffic state prediction.

As previously briefly described in Section 2.1, traffic simulation models are usually divided into three groups; micro-, meso- and macroscopic. The names of these groups give an indication of the level of detail in the simulations. Microscopic models are the most detailed, where individual cars interact with other vehicles. The microscopic models are also the most computationally demanding and are in general only used for smaller areas. Macroscopic models assign a given Origin/Destination (OD)-matrix to a given network and simulate aggregated traffic variables such as speed, density or flow. The macroscopic models can be either static, with no time dependent variation, or dynamic, when the model or the input changes over time.

In the literature there are several examples of traffic state prediction using traffic simulation software and traffic modeling, see for example BOSS-METANET (Papageorgiou et al., 2010), DynaMIT (Ben-Akiva, et al., 2001), DynaSMART-X (Mahmassani, 2005), AIMSUN On-line (Torday, 2010) and PTV Optima (Meschini and Gentile, 2010).

Traffic simulation models are useful since they are based on assumptions of the actual traffic behavior and they can therefore model and estimate the traffic state that an incident or lane closure causes. They are also less dependent on data compared to non-parametric models.

The more noticeable downsides of traffic simulation models is their efficiency, they generally demand a lot of computational capacity, in particular the microscopic models. Model parameters also need to be calibrated in order to ensure the model’s accuracy and its ability to represent the infrastructure. The traffic simulation models are also dependent on predicted boundary flows.
6.2.2 NON-PARAMETRIC MODELS

Non-parametric models are a group of models that include a flexible number of parameters. They differ greatly from the parametric type since the parameters, as well as the structure of the model, need to be determined from data. Non-parametric models are created from a large amount of historical data. The advantage in relation to parametric models is that they can capture the traffic dynamics even though no knowledge of the traffic behavior as such is needed. However, non-parametric models are unable to capture the dynamics of unforeseen events, e.g., traffic accidents. Although such events are impossible to predict per se, some parametric models enable the effects of unforeseen events to be modelled, e.g., with increased queuing and by the spreading of shockwaves in time and space. A short description of the most common non-parametric models used for short-term traffic state prediction is found below.

**Linear time series**

Time series analysis is a general term for models that estimates the future state of the system based on past states, observations and error terms. No effort is taken in trying to capture the highly dynamic nature of a traffic system and the behavior of the drivers. Instead, the seasonality (the time variation) is modelled with statistical functions.

There are a number of different models that are categorized as time series models and that has been used for traffic state prediction. The most common ones are:

- Linear regression based on the covariates of the prediction function, see for example Kwon et al. (2000) and Rice and Zwet (2004).
- AutoRegressive Integrated Moving Average (ARIMA), a Box-Jenkins model that is a common statistical technique that can be used for prediction, see for example Chen et al. (2011), Lee and Fambro (1999) and Zhang et al. (2011).
- Kalman filtering where the prediction of the future state is based on the state at the previous time step and on measurements available at that time step, see for example Ben-Akiva et al. (1992), Chen and Grant-Muller (2001) and Guo and Williams (2010).

**$k$-Nearest Neighbor**

The $k$-Nearest Neighbor ($k$-NN) model monitors the current state of the road and the future state is predicted by matching the current state with the $k$ closest “neighbors” in a historical database. A neighbor is usually a point in a time series where the time, e.g., time of day, is equal to the time of the day of the prediction. The prediction is based on the average value of the $k$ nearest neighbors, see for example Clark (2003) and Bustillos and Chiu (2011).

**Locally weighted regression**

In a locally weighted regression model the prediction residual of each data point is weighted proportionally to its proximity to the current measurement. This method has been used by for example Nikovski (2005) and Zhong et al. (2005).
**Fuzzy logic**

Fuzzy logic is based on a set of rules that are created manually or automatically. A traffic situation that occurs corresponds to one or more rules and based on the degree of correspondence and the current rules a prediction is made. The method has been used by for example Zhang and Ye (2008) and Coufal and Turunen (2004).

**Bayesian networks**

Bayesian networks are directed probabilistic graphical models that consider neighboring links as influential over the link currently having its state predicted. The method has been used for predicting traffic flow by Castillo et al. (2008) and Shiliang et al. (2006)

**Neural networks**

Neural networks (NN) is a collective name of a vast range of different kinds of models. They are, according to Van Hinsbergen et al. (2007), the most common way of predicting traffic. Most likely because they have the ability to model both non-linear and time dynamic processes. For examples of traffic state prediction based on neural networks see for example Zhang (2000), Van Hinsbergen et al. (2009) and Vlahogianni (2007).

A neural network can be conceptualized as a network with nodes and directed links. The nodes are ordered by an input layer, at least one hidden layer and an output layer, see Figure 41 for an example.

![Figure 41: Schematic view of a fictive neural network with four input nodes, one hidden layer with two nodes and one output node.](image)

The structure of the network, the number of nodes and hidden layers, is not fixed but is adaptive during the training phase were the neural network is trained to react in a certain way to some input and thus responding with a desired output. The neurons of the neural networks, the nodes of the network, have a function that produces this reaction to the input. Furthermore, the connectors of the networks have weights that further manipulate the data being sent between the neurons.

There exists a number of different extensions to the neural networks and Van Hinsbergen et al. (2007) found almost twenty different types of neural networks in his survey. The standard version is a back propagation neural network (BPNN) where the input is fed to the hidden layer and an output is produced. This output is then compared with the desired output and the difference is
propagated backwards through the network. An example of an extension is the evolutionary neural networks (ENN) that, according to Van Hinsbergen et al. (2007), has shown very positive results when used for traffic state prediction. In the ENN, which is inspired by the theory of evolution, different neural network settings or “individuals” compete with other individuals and the one that produces the most accurate predictions over time will be chosen as the predictor, see for example Abdulhai et al. (1999).

6.2.3 COMPARISONS OF MODELS FOR SHORT-TERM TRAFFIC STATE PREDICTION

From the studied literature we can see that many comparisons of different types of models for short-term traffic state prediction have been made through the years. Van Hinsbergen et al. (2007) has studied the comparisons in detail and found that under specific conditions, linear regression, the locally weighted regression model and evolutionary neural networks seem to perform well in relation to other models. However, no comparison between traffic simulation models and other types of models has been found. Furthermore, very few studies exists where the studied models has been applied for a large network and for both highway and urban environment. Van Hinsbergen et al. (2007) further conclude that since no method has been proven to be the best method in any situation, methods that combine or select the most appropriate model for a specific situation are interesting to study further.

Another extensive survey on short-term traffic state prediction is the one by Vlahogianni et al. (2014). From around 170 articles, mainly from 2004 and onwards, it was concluded that the research has moved from a classical statistical perspective to neural networks and more data driven approaches. Vlahogianni et al. (2014) identifies ten challenges within the area of short-term traffic state prediction and presented related directions for further research, see Table 9. The proposed hybrid prediction approach presented in the coming section is in line with many of the identified future research directions.

Table 9: Ten identified challenges and directions for further research within the field of short-term traffic state prediction according to Vlahogianni et al. (2014).

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Further research directions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Developing responsive algorithms and prediction schemes</td>
<td>Weather and incident responsive algorithms, enhancing the efficiency of online computations using artificial intelligence, standardizing the requirements with regard to the spatial and temporal data coverage</td>
</tr>
<tr>
<td>2. Freeway, arterial and network traffic predictions</td>
<td>Focus on network level predictions, synergy with traffic flow theory and models</td>
</tr>
<tr>
<td>3. Short-term predictions: from volume to travel time</td>
<td>Producing existing or novel measures of traffic performance using data from multiple sources or using novel technologies for collecting and fusing data</td>
</tr>
<tr>
<td>4. Data resolution, aggregation and quality</td>
<td>Determining the optimal degree of aggregation in relation to the short-term forecasting application, Quality of probe data</td>
</tr>
</tbody>
</table>
5. Using new technologies for collecting and fusing data

Testing the efficiency of new technologies for collecting traffic data, reliability under all types of traffic flow (constrained, unconstrained), market penetration, standardization, cost, privacy issues, Effectiveness of fusing strategies

6. Temporal characteristics and spatial dependencies

Focus on network level spatio-temporal approaches, fusing modeling and data-driven algorithms

7. Model selection and testing

Establishing synergies with statistics for estimating model specification and fit.

8. Compare models or combine forecasts?

Introducing combinations of forecasts for multiple steps ahead predictions, testing the reliability of combinations of forecasts over single model predictions

9. Explanatory power, associations and causality

Synergy with statistics and computationally intelligent algorithms to enhance the transparency of data-driven approaches

10. Realizing the full potential of artificial intelligence

Introducing intelligence to data collection and storage, traffic analysis, optimization modeling and decision making

Judging by the papers presented here, no available method/model/approach for short-term traffic state prediction is better than the rest for all types of situations or network types. It seems vital to combine different methods or find a method that chooses the best method given the situation for which the traffic state is to be predicted for. Furthermore, as stated in Table 9, it is important to focus on methods that can handle large networks and new type of data sources.

6.2.4 SHORT-TERM PREDICTION USING THE CELL TRANSMISSION MODEL

One conclusion from the extensive survey on short-term traffic state prediction models by Vlahogianni et al. (2014) was that there are few examples of when analytical modeling approaches, such as the cell transmission model, has been combined with non-parametric models to forecast traffic state and replicate traffic dynamics. One of the features of the CTM, which also holds for the CTM-v, is the possibility to run the model forward in time and use it for prediction. Below is a short presentation of four previous cases when the cell transmission model, in combination with other models, has been used for short-term traffic state prediction.

Xiong et al. (2014) presents a three-stage framework for short-term travel time prediction on highways. An ARIMA model is used to, based on historical data, predict the demand that in the next step is used as input to a CTM. The CTM is then used to predict the resulting traffic flow from the previously predicted demand. In the final stage travel times are extracted from the predicted traffic flows using N-curve analysis. The framework was evaluated on empirical data from a short stretch of motorway in Australia without any on- and off-ramps. The results are promising, but the authors conclude that the model should be developed to also handle more complicated road infrastructures, e.g. on- and off-ramps and situations with non-recurrent incidents.
Pan et al. (2013) used an extension of the stochastic CTM (SCTM) to perform short-term prediction of the density. The extension imply that the spatial-temporal correlation of traffic flow also was considered. The boundary flows of the SCTM was predicted from historical and real-time measurements using a multivariate normal distribution based best linear predictor. The model was evaluated on empirical data from a freeway in California and the results show that the model performs well for the analyzed scenarios. However, the changes in traffic state is a bit delayed in the prediction compared to the measurements. According to the authors this is an effect of the reactive nature of the algorithm and the fact that the model cannot capture a moving bottleneck.

Szeto et al. (2009) developed a model where a seasonal ARIMA (SARIMA) model and the CTM was used to predict traffic flow. The SARIMA model is a multistep forecasting model that captures the daily, weekly and other seasonal effects of traffic flows. It was used to predict the boundary flows that are used as input to the CTM. The model was evaluated on empirical data from the city center in Dubin. The results suggest that the model are useful for short-term prediction of traffic flow, in particular in congested states with frequent queue spillback occurrence.

Tampère and Immers (2007) used the CTM together with an Extend Kalman filter that combines the output from the CTM with the measurements. The model was used to predict the traffic state on a highway in Brussels and the results shows that the model tend to overestimate the travel times during congestion. Furthermore, the predicted travel times react slow to changes and seem to be delayed. According to the authors this is due to the naïve method used to predict the boundary flow.

From the literature we can conclude that previous examples where a CTM-model has been used for prediction have been promising, although there exist several areas where the used approaches can be improved.

6.3 THE MMS-MODEL FOR PREDICTION

Based on the findings in the presented literature survey, we have implemented and evaluated a hybrid short-term prediction approach based on the MMS-model together with a non-parametric prediction of sensor data. In the MMS-model the CTM-v always predict the velocities one time step ahead, but it can also predict several time steps ahead when continuously fed with the velocities predicted by the CTM-v in the previous time step and predicted boundary conditions. However, to get the previously described MMS-model to produce accurate short-term traffic state predictions it is vital to make the model more dynamic. As mentioned in Section 5.3.3 and Section 5.3.4, all model parameters that were calibrated in the previous chapter are static, i.e. the same value is used for the whole day and for all types of days, both weekdays and weekends. When the MMS-model is used to estimate the traffic state the model is continuously fed with new measurements that compensate for the static boundary flows, i.e. the static sources, split ratios and sinks. However, if the CTM-v model is ran forward in time and used to predict the future traffic state, there are no real-time measurements available that can compensate for a static parameter setting. Hence, when used for prediction the MMS-model is very dependent on the boundary flows, in particular the split ratios and the sources. An example of the differences between a static parameter value, the collected measurement and a dynamic value for one on-ramp is illustrated in Figure 42. In the figure, the measurements are represented by the blue dots, the static value used in the estimation
model is the green line and the red line illustrates a dynamic value where the measurements has been aggregated over a 15-minute period.

Figure 42 illustrates a fictive example where the dynamic value are based on the collected measurements for that particular day. In a real-time implementation the dynamic boundary flows have to be predicted by either using non-parametric methods, as is done in Pan et al. (2013), Calvert et al. (2010) and Wu et al. (2014), or by simple averaging of historical data. It should be noted that accurate measurements of both inflow rates and split ratios can be difficult to obtain during congested periods. Curb side driving together with blocking back of either the mainline or the ramps may for example result in measurements which does not represent the actual inflow and split ratio profiles.

The data flow when the MMS-model is used for prediction can be seen in Figure 43. Starting from a known traffic state (given by the most recent traffic state estimation) the future traffic state is predicted based on predicted boundary flows (sources and split ratios) and the CTM-v network update algorithm that is ran forward in time until the desired prediction horizon is reached. When the CTM-v network update algorithm is ran forward in time it is continuously fed with the prediction from the previous time step. If the prediction is done without any additional predictions of mainline sensor data we do not need to use the EnKF in the prediction process.
Previous research has shown that it is easier to predict the mainline sensor data than the boundary flows, since the variation is larger on the on- and off-ramps than on the mainline, see for example Wu et al. (2014). As is pointed out by Wu et al. (2014), mainline flows are actually aggregations of upstream on-ramp flows, and mainline sensor measurements thus have a reduced noise. This makes non-parametric prediction of mainline sensor data more accurate in comparison to predictions of on-ramp flows. Hence, one possible extension to the prediction approach, presented in Figure 43, is to take advantage of the data fusion capabilities of the EnKF and add non-parametric predictions of the mainline sensor data to the prediction model. This results in an alternative hybrid traffic state prediction approach where both a parametric and a non-parametric prediction method are used. The data flow in the proposed approach is presented in Figure 44.

We have implemented and evaluated both the prediction approach presented in Figure 43, denoted CTM-v prediction in the results, and the one in Figure 44, denoted Hybrid prediction in the results. The results from the hybrid short-term prediction using the MMS-model has previously been published in Allström et al. (2016).
6.4 EXPERIMENTAL SETUP

To analyze the accuracy of the proposed short-term traffic state prediction approach a section of the highway just north of central Stockholm has been studied. It is at large the same road stretch that was used for the calibration in Section 5.3. The road stretch is 7 km long and part of the southbound Stockholm ring road. Data from January to March 2013 is used, and for this period radar detector data is available. During March, travel time data from Bluetooth sensor data is also available. The radar detectors collect speed and flow for each lane, aggregated at one-minute intervals. Bluetooth sensors are placed every 500-1000 meters and collect travel times from all active Bluetooth units that pass two consecutive sensors. The sections with Bluetooth measurements are marked in Figure 45 with BT1 up to BT7, together with the length of each section. There are three on-ramps and three off-ramps within the studied section, and the number of lanes varies between two and four. In the middle of the test site the speed limit changes from 90 km/h to 70 km/h. Bluetooth data is only available for a limited number of days and has only been used for calibration of EnKF parameters and for validation of the presented results.

In Figure 45, radar detectors with available data are marked with “+”. For sensors that are used for the non-parametric prediction the sensor-ids are included for future references. The test site is divided into nine links.

Figure 45: The test site in Stockholm used for evaluation of the short-term prediction. Not drawn according to scale.

6.4.1 SET-UP AND CALIBRATION OF THE MMS-MODEL

In order to produce reliable short-term predictions of the traffic state using the MMS-model we have implemented a more dynamic version of the model. In the dynamic version of the model, the boundary flows are updated every 15 min, instead of using a static value. As a consequence of this, parts of the calibration presented in Chapter 5 has been updated. Speed and flow measurements from 23 days (only Mondays to Fridays are selected) have been used for a revised calibration of the fundamental diagrams as well as the sources, the inflow at on-ramps, and split ratios at diverging nodes.

All fundamental diagrams are specified using the same jam density, which is measured from aerial photos over the Stockholm highway system, see Strömgren (2011). Free flow speed has been measured using data from Bluetooth sensors. The most downstream section of the test site is a bottleneck, which usually is activated both during the morning and afternoon peak period. Capacity measurements has been used for calibrating the shockwave speed (given jam densities and free flow travel times) for the bottleneck. For the remaining sections standard capacities from Olstam
et al. (2013) has been applied. For cells within each of the marked links in Figure 45, the same fundamental diagram is used.

Profiles with daily variations of sources and split ratios are constructed by taking mean values for 15-minute periods from the 23 days. For the on and off-ramp, where no sensors are mounted, this is done by comparing the flow before and after the on/off-ramp respectively. For the remaining on-ramps the inflows are measured on the ramp itself. For the split ratios, flow measurements from the mainline and off-ramp sensors are used to calculate the share of vehicles that diverge from the mainline. However, in the presence of queues, this computation of inflow and split ratios introduces potential errors. The measured split ratios may be an effect of the mainline functioning as an extended ramp and the inflows may be restricted by blocking back of queues from the mainline. However, with no other data available this is the best method we can use. The dynamic boundary flows are predicted with a naïve method where we simply use historical profiles constructed by taking mean values for 15-minute periods. The historical data consisted of radar data from 23 weekdays prior to the predicted day.

For the studied highway section, restricted sink capacities are only set for the outflow at the most downstream subsection. The reason for this is that blocking back is mainly an issue in this subsection. Furthermore, the sink capacity is only lower than the bottleneck capacity at the most downstream section during the afternoon. A sink capacity, as well as start and end time of the reduced capacity period, has been calibrated based on the 23 calibration days. Length of the afternoon congestion period and the maximum queue length have been the main comparisons used for this calibration. Overall, the afternoon is more difficult to predict, mainly due to blocking back from the surrounding traffic network.

The parameters related to the Ensemble Kalman Filter has also been revised from the ones presented in Chapter 5. The measurement noise is assumed to have zero mean and the standard deviation has been estimated to 1 m/s from measurements. The model noise is assumed to have zero mean and standard deviation 0.8 m/s. The model noise standard deviation has been calibrated by evaluating MAPE values calculated from estimated travel times and Bluetooth travel times for a number of different parameter settings.

Since we also use predicted mainline sensor data we have to set the uncertainty for these predictions in the EnKF. The uncertainty for the predicted mainline sensor data is assumed to have zero mean and our initial guess of standard deviation was to use the same as for the measurement noise. There is, however, an increased uncertainty in the predicted measurements and the best result has been obtained with a standard deviation twice the one of the measurement noise (2 m/s).

Furthermore, the linear programming problem used to allocate the flow across a junction, which was briefly introduced in Section 4.3.2, is replaced with a proportional distribution of the flow across the junction. The reason for this is that the linear programming problem does not give a unique solution, instead we got an extreme point dedicating all the capacity to one of the incoming links, in our case the on-ramp. The proportional distribution of the flows keep the flow out of the junction to the total incoming flow if the incoming flow is lower than the capacity, or to the capacity if the incoming flow is equal or higher than the capacity.
The validation of the estimation of the updated more dynamic MMS-model has been done with data from the 21st of March, and the resulting mean average percentage error (MAPE) based on total and subsection travel times, are presented in Table 10. These results can be compared to the accuracy of the calibrated static MMS-model. As can be seen the accuracy of the dynamic and static MMS-model are very similar, in particular for total travel times.

Table 10: Accuracy of the estimation of the calibrated dynamic MMS-model compared with the accuracy of the calibrated static MMS-model.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Dynamic</th>
<th>Static</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPE based on subsection travel times (%)</td>
<td>11.5</td>
<td>10.4</td>
</tr>
<tr>
<td>MAPE based on total subsection travel times (%)</td>
<td>4.8</td>
<td>4.9</td>
</tr>
</tbody>
</table>

6.4.2 Prediction of mainline sensor data

The prediction of mainline sensor data can be done with a number of different non-parametric methods. Here, a non-linear autoregressive neural network with exogenous input (NARX) model has been used. It is a recurrent dynamic neural network and is based on the linear ARX model. NARX is commonly used for short-term predictions based on time series analysis. In Zeng and Zhang (2013) such a model is successfully applied for travel time predictions on a freeway. A similar approach is the state space neural network, applied in Van Lint et al. (2005) for prediction of travel times using travel times and flow data from fixed traffic sensors as input.

The NARX-model can be summarized with

\[ y(t) = f(y(t-1), y(t-2), \ldots, y(t-n_y), x(t-1), x(t-2), \ldots, x(t-n_x)), \]

where \( f \) is a non-linear function and the value \( y(t) \) in time period \( t \) is dependent on the values of \( y \) in the previous time periods until time period \( n_y \) and previous values of an independent (exogenous) input signal \( x \) until time period \( n_x \). Hence, the input for predicting speeds at a specific sensor is previous speed measurements at the specific mainline sensor location, as well as previous speed measurements from the surrounding sensors. The NARX-model include time delays of inputs as well as time delayed feedback of outputs. Thus, the predictions is based not only on the most recent measurements, but on the trend from several recent measurements. Another factor that can influence the prediction is time of day and day of week, given that the training data is clustered. Outputs from the prediction are the speeds for coming time periods at the sensor locations. For the experiment presented here, the NARX-model in the Neural Network Times Series Tool in MATLAB has been used. Hence, the main contribution of this work is not in the development of a non-parametric model for prediction, but in the way they is combined with the MMS-model in the hybrid prediction approach.

The prediction of speeds at mainline radar detectors locations is done for sensor 230, 231, 235, 236, 238, 239 and 244, see Figure 45. It would of course be ideal to predict the speed for all sensors but the NARX-model demands complete data sets for training and prediction, i.e. no missing values, which only was available for the sensors listed. When predicting the speed at one of the sensors all other sensors listed above were used as exogenous input. In order for the NARX-model to capture changes in the traffic state it is important to find the appropriate time resolution of the predicted mainline sensor data. Here, a prediction horizon of 60 minutes is used and for the first
In 10 minutes, measurements for 2-minute periods are predicted. For predictions 10 to 30 minutes ahead in time we use a time resolution of 5-minute periods and for 30 to 60 minutes ahead in the future we use 10-minute periods. Using a time delay of five periods in the NARX neural network we look 10 minutes back in time for prediction 10 minutes ahead, 25 minutes back when predicting 10 to 20 minutes ahead in time and 50 minutes back when predicting 30 to 60 minutes ahead in time. The NARX neural network is set-up with an output feedback loop of the two most recent time periods and one hidden layer with ten neurons. Each time period result in a neural network to train and in total we train 12 neural networks for each sensor location, using hour of day and day of week (grouped into Monday-Thursday, Fridays and Saturday-Sunday), and speed measurements from all available sensors. The training is done using 53 days during January, February and March 2013. These settings where chosen based on previous experience which has shown that an increased number of neurons and/or time period results in overfitting of the model for this data set.

In Table 11 the MAPE values for predicted mainline sensor data on the 21st of March 2013 are presented. The MAPE values are based on the difference between predicted and a measured mean value for 2-, 5- and 10-minute periods, depending on for how far ahead in the future the prediction is made. While the MAPE values not always increases when we predict one time period further ahead in time, the overall trend is that it is an increased uncertainty when predicting measurements further ahead in time. Sensor 244 stands out with the lowest MAPE value, and this is because there are very few time periods of a day when the congestion has reached this sensor, and predicting speeds during non-congested traffic states has shown to be much easier than to predict speeds during congested periods.

<table>
<thead>
<tr>
<th>Sensor</th>
<th>1-2</th>
<th>3-4</th>
<th>5-6</th>
<th>7-8</th>
<th>9-10</th>
<th>11-15</th>
<th>16-20</th>
<th>21-25</th>
<th>26-30</th>
<th>31-40</th>
<th>41-50</th>
<th>51-60</th>
</tr>
</thead>
<tbody>
<tr>
<td>230</td>
<td>4.41</td>
<td>4.51</td>
<td>4.77</td>
<td>4.60</td>
<td>4.76</td>
<td>5.61</td>
<td>5.12</td>
<td>5.44</td>
<td>5.42</td>
<td>7.10</td>
<td>7.35</td>
<td>7.48</td>
</tr>
<tr>
<td>231</td>
<td>6.05</td>
<td>6.17</td>
<td>6.37</td>
<td>6.55</td>
<td>6.50</td>
<td>7.41</td>
<td>7.62</td>
<td>7.65</td>
<td>7.98</td>
<td>8.46</td>
<td>9.23</td>
<td>10.23</td>
</tr>
<tr>
<td>235</td>
<td>6.15</td>
<td>5.65</td>
<td>6.21</td>
<td>6.15</td>
<td>6.09</td>
<td>6.71</td>
<td>7.11</td>
<td>7.00</td>
<td>6.36</td>
<td>11.51</td>
<td>13.01</td>
<td>11.42</td>
</tr>
<tr>
<td>236</td>
<td>4.91</td>
<td>4.99</td>
<td>6.38</td>
<td>5.03</td>
<td>6.08</td>
<td>6.82</td>
<td>5.93</td>
<td>6.20</td>
<td>5.97</td>
<td>11.88</td>
<td>13.59</td>
<td>11.89</td>
</tr>
<tr>
<td>238</td>
<td>4.70</td>
<td>4.27</td>
<td>5.11</td>
<td>4.88</td>
<td>5.23</td>
<td>7.52</td>
<td>7.01</td>
<td>7.28</td>
<td>7.08</td>
<td>14.69</td>
<td>16.56</td>
<td>15.26</td>
</tr>
<tr>
<td>239</td>
<td>4.46</td>
<td>4.35</td>
<td>4.80</td>
<td>4.54</td>
<td>4.67</td>
<td>5.69</td>
<td>6.17</td>
<td>5.97</td>
<td>6.89</td>
<td>9.97</td>
<td>10.07</td>
<td>9.18</td>
</tr>
<tr>
<td>244</td>
<td>2.73</td>
<td>2.84</td>
<td>2.92</td>
<td>2.78</td>
<td>2.79</td>
<td>3.29</td>
<td>3.28</td>
<td>3.18</td>
<td>3.06</td>
<td>4.56</td>
<td>4.86</td>
<td>4.98</td>
</tr>
</tbody>
</table>

### 6.4.3 Evaluation of the Proposed Hybrid Prediction Approach

For evaluation of the hybrid prediction framework we have used data from the 21st of March 2013. This data has not been used for either calibration of the dynamic MMS-model or training purposes.
Travel times are predicted for a car entering the highway at time \( t \), and calculated by simulating a car passing through the space-time speed contour plot corresponding to the prediction done at time \( t \). Thus this travel time prediction makes use of predicted traffic states ranging from 0 up to around 20 minutes ahead in the future, depending on how long time takes to drive through the evaluated highway section (i.e. depending on the level of congestion). Similarly, a travel time prediction for a car starting 30 minutes from now make use of information from a prediction horizon ranging from 30 up to around 50 minutes into the future. As comparison, a naïve prediction where it is assumed that the current traffic conditions will prevail during the next 60 minutes, is computed (resulting in an instantaneous travel time). Note that the naïve prediction only makes use of the most recent traffic state estimation. The estimated travel time used for comparison is obtained from simulating a car passing through the estimated space-time speed contour plot.

To evaluate the accuracy of the predicted and estimated travel times, MAPE values of predicted and estimated travel times versus measured travel times for five minute averages, using Bluetooth sensors, are computed for seven subsections for which there exists Bluetooth data. The seven subsections cover all but the last section of the highway in Figure 45, and two different MAPE values are computed based on the measurements. The first one is the mean error for all sections, denoted MAPE\(^c\) in Chapter 5 and given by Equation 5.3. For the other one we sum up the subsection travel times into a total travel time across all subsections, denoted MAPE in Chapter 5 and given by Equation 5.5. For the total travel times we also provide mean and maximum absolute errors. Note that all errors computed between predicted and Bluetooth travel times will be based on cars starting at different locations (the beginning of each Bluetooth segment) of the highway at the same time. Each subsection is, however, short and thus predictions at most 35 minutes ahead into the future will be used, which is the maximum observed travel time.

Space-time speed contour plot are used for comparing predictions of the traffic state, with and without predicted mainline sensor data, with estimation results. Here the estimation based on all available sensor data and CTM-v output is used as reference, since it is the best available information.

6.5 RESULTS FROM THE HYBRID SHORT-TERM PREDICTION APPROACH

In Table 12, the mean absolute percentage error (MAPE) as well as mean and maximum absolute error value for the predicted travel times in relation to the measured travel times, is presented. For comparison, we also provide the accuracy for the estimation and for the predicted travel times computed by running only the CTM-v with dynamic boundary flows but not using the EnKF model for assimilating the output with predicted mainline sensor data. This prediction can be done off-line and is therefore a simple approach for predicting the travel times. The entry marked Estimate is based on the estimated travel times as given by the MMS-model and only used for comparison. Similarly, CTM-v refers to the case when using the CTM-v with historical boundary flows only. In the other three methods we always start our prediction from the current estimation (denoted Naïve prediction, CTM-v prediction and Hybrid prediction). The error values related to the Estimate corresponding to an estimation done after the car has passed the road stretch. Thus, the estimation makes use of speed measurements which were not available at the time the car entered the section, which of course is not possible in real-time. In comparison, the Naïve prediction makes
use of the most recent estimation, and assumes prevailing traffic conditions during the time it takes to drive through the section.

The three prediction methods which are based on the most recent estimation plus a prediction show similar results for a car starting its journey at the time of the prediction time, i.e. when the prediction horizon is 0 minutes, when compared with the Bluetooth reference travel times. For a journey taking place 15 and 30 minutes ahead in time, the Hybrid prediction produce the most accurate prediction considering subsection MAPE values, although the differences are rather small when comparing with CTM-v prediction. For the case of a prediction 30 minutes ahead in time, the Naïve prediction is performing worse than the CTM-v. Considering the total travel time in Table 12, the Hybrid prediction results in lowest maximum absolute error values for all prediction horizons, but for predictions even more far ahead into the future it tends to produce the same value as the CTM-v prediction. This could be related to the quality of predicted mainline sensor data being less reliable for predictions further ahead in time. For shorter prediction horizons, the use of predicted mainline sensor data clearly reduces the maximum error, in comparison to the other approaches.

Table 12: Comparison of predicted and estimated travel times with measured travel times.

<table>
<thead>
<tr>
<th>Prediction/ estimation method</th>
<th>Time horizon</th>
<th>0</th>
<th>5</th>
<th>15</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAPE based on subsection travel times</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimation</td>
<td>11.5</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTM-v</td>
<td>19.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve prediction</td>
<td>13.8</td>
<td>14.7</td>
<td>18.5</td>
<td>27.0</td>
<td></td>
</tr>
<tr>
<td>CTM-v prediction</td>
<td>13.0</td>
<td>17.6</td>
<td>16.9</td>
<td>16.9</td>
<td></td>
</tr>
<tr>
<td>Hybrid prediction</td>
<td>13.0</td>
<td>15.7</td>
<td>15.3</td>
<td>16.9</td>
<td></td>
</tr>
<tr>
<td><strong>MAPE based on total subsection travel times (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimation</td>
<td>4.6</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTM-v</td>
<td>12.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve prediction</td>
<td>5.3</td>
<td>6.6</td>
<td>19.5</td>
<td>19.7</td>
<td></td>
</tr>
<tr>
<td>CTM-v prediction</td>
<td>5.1</td>
<td>9.2</td>
<td>10.1</td>
<td>10.9</td>
<td></td>
</tr>
<tr>
<td>Hybrid prediction</td>
<td>5.5</td>
<td>6.8</td>
<td>7.7</td>
<td>7.9</td>
<td></td>
</tr>
<tr>
<td><strong>Mean absolute error in seconds based on total subsection travel times (s)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimation</td>
<td>22</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTM-v</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve prediction</td>
<td>28</td>
<td>32</td>
<td>50</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>CTM-v prediction</td>
<td>25</td>
<td>45</td>
<td>48</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>Hybrid prediction</td>
<td>28</td>
<td>32</td>
<td>39</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td><strong>Maximum absolute error in seconds based on total subsection travel times (s)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimation</td>
<td>205</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTM-v</td>
<td>378</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve prediction</td>
<td>245</td>
<td>342</td>
<td>397</td>
<td>571</td>
<td></td>
</tr>
<tr>
<td>CTM-v prediction</td>
<td>304</td>
<td>423</td>
<td>445</td>
<td>378</td>
<td></td>
</tr>
<tr>
<td>Hybrid prediction</td>
<td>213</td>
<td>237</td>
<td>367</td>
<td>374</td>
<td></td>
</tr>
</tbody>
</table>
In Figure 46 the estimated and predicted travel times are shown for a journey starting at the time of the prediction, and journeys starting 5, 15 and 30 minutes ahead into the future. For a journey starting at the time of the prediction (Figure 46a), the Naïve prediction is very similar to the estimated travel time. For journeys starting 5, 15 and 30 minutes into the future, the Naïve prediction will simply be offset by 5, 15 and 30 minutes respectively. Thus, since we are comparing with estimated travel times, the height of each peak will be correctly predicted by the Naïve prediction but it will be offset in time. During the morning peak period, both the CTM-v prediction and Hybrid predictions manage to follow the estimation well, but for the afternoon peak period the Hybrid prediction performs better, both in terms of capturing the start and end of the congested period, as well as the length of the travel time.

![Figure 46: Travel time for a journey beginning in 0 (a), 5 (b), 15 (c) and 30 (d) minutes. Showing Estimation (green), Naïve prediction (black), only CTM-v prediction (blue) and Hybrid prediction (red).](image)

In Figure 47 the space-time speed contour plot for the CTM-v, CTM-v prediction and Hybrid prediction, starting in a known estimation for a 30-minute prediction horizon is presented. For comparison the space-time speed contour plot for the Estimation is also included in the figure. The upper part of the space-time speed contour plot corresponds to the end of the highway section illustrated in Figure 45, and vice versa for the lower part. First of all, one can notice that the morning peak period is predicted rather well by the CTM-v, the CTM-v prediction and the Hybrid prediction. The afternoon peak period is, however, more difficult. For the afternoon, it is only the Hybrid prediction...
that captures the last part of the congestion period. Overall, the afternoon is more challenging to predict since there is a blocking back from outside the modelled network. Clearly, the predicted measurements in the Hybrid prediction approach captures this better.

6.6 CONCLUSIONS FROM HYBRID SHORT-TERM PREDICTION USING THE MMS-MODEL

The main conclusion from the survey on previous work on short-term traffic state prediction is that, in order to produce accurate short-term predictions for various types of networks and conditions, it seem vital to combine different methods. Furthermore, there are few examples of when analytical modeling approaches, such as the CTM-v, has been combined with non-parametric models to predict the traffic state.

Motivated by these findings we have implemented and evaluated a hybrid prediction approach for assimilating parametric and non-parametric short-term traffic state prediction. The previously presented MMS-model has been modified to handle more dynamic boundary flows and the NARX-model has been used to predict mainline sensor data. The proposed short-term prediction approach has been evaluated for a highway section in Stockholm. The results show that both the

Figure 47: Space-time speed contour plot for Estimation (a) and CTM-v (b). 30-minute predictions are presented for CTM-v prediction (c), and Hybrid prediction (d). The legend shows velocity in km/h.
hybrid approach and the CTM-v prediction without the additional predicted mainline sensor data outperform the naïve prediction for longer prediction horizons, but between the two approaches, there are smaller differences. However, the hybrid prediction is an improvement in comparison to the CTM-v prediction for all prediction horizons. In particular for the afternoon peak, for which there are large uncertainties in the input flows, split ratios, and capacities, used within the CTM-v model. In terms of worst-case performance, the hybrid prediction is performing better than the CTM-v prediction when comparing with Bluetooth measurements.
7 CONCLUSIONS AND FUTURE WORK

For a more proactive and effective traffic management to be feasible, accurate estimations and short-term predictions of the traffic state is essential. This thesis has demonstrated how the CTM-v and an Ensemble Kalman filter can be used to estimate the traffic state for the highways in and around Stockholm. The verification of the MMS-model indicate that the model can produce traffic state estimations in real-time for a large network. Furthermore, the link specific fundamental diagram parameters has been calibrated and a calibration framework that enables an iterative calibration of the model parameters against measured travel times has been implemented. The results from the validation show that once the model is calibrated the estimated travel times corresponds well with the ground truth travel times collected from Bluetooth sensors. To achieve a good accuracy all parameters has to be calibrated, the parameters related to the fundamental diagram as well as the boundary flows and the EnKF noise parameters. When a traffic state is estimated the model is fed with continuous measurements that compensates for calibration and model errors. However, when future traffic states are predicted by running the CTM-v forward no measurements are available. Hence, the model has been modified to use a more dynamic set of boundary flow parameters and extended to also be able to predict the future traffic state.

The survey on previous work on short-term traffic state prediction shows that it is vital to combine different methods in order to produce accurate short-term predictions for various types of networks and conditions. Furthermore, there are few examples of when analytical modeling approaches, such as the CTM-v, has been combined with non-parametric models to predict the traffic state. Hence, to further improve the traffic state predictions from the MMS-model, a hybrid prediction approach where the CTM-v and the EnKF are used together with a neural network has been implemented and evaluated. The neural network, a NARX-model, is used to predict mainline sensor data which compensate for the fact that no measurements are available when the model is used for prediction. The results are promising and the hybrid short-term prediction approach is an improvement in comparison to when the CTM-v is used for predictions without the additional predicted mainline sensor data for all analyzed prediction horizons. In particular for the afternoon peak, for which there are large uncertainties in the boundary flows and capacities. Furthermore, the proposed prediction approach is well suited to handle several of the challenges within the field of short-term traffic state prediction identified by Vlahogianni et al. (2014) and presented in Table 9.

It should be noted that in both the estimation and prediction experiment presented in this thesis only data from fixed radar detectors has been used. Unfortunately, no dynamic point speed data from probe vehicles has been available. The main motive to use the CTM-v and the EnKF for traffic estimation and prediction is the fact that it enables different types of speed data to be fused. Hence, to fulfil the potential of the model it should be evaluated using data from different sources. The CTM-v model together with the EnKF can incorporate point speed measurements from both fixed sensors and probe vehicles. In Patire et al. (2015), the traffic state has been estimated with the CTM-v and EnKF using both loop detector data and point speed from probe vehicles. The results show that when data from multiple sources are fused, superior results are obtained. Even very sparse probe data is useful and improve the results. Furthermore, in Gundlegård et al. (2015), the model has been further developed to also handle travel time measurements as input in the estimation process.
Besides using other types of data than fixed sensor data in the estimation and prediction process, there are several areas where the proposed estimation and hybrid prediction approach can be further developed. The calibration of the fundamental diagrams and the prediction of both boundary flows and mainline sensor data can be improved to also take into consideration the weather conditions. Previous research show that adverse weather conditions have an affect both on the free flow speed and the capacity as the drivers reduce their speed and also the time headways and spacing, see Billot et al. (2009). Hence, if the fundamental diagrams and the prediction of boundary flows and mainline sensor data does not take this into consideration it will of course have a considerable effect on the accuracy of the predictions. Another obvious improvement of the proposed hybrid prediction approach is to implement a more sophisticated method for the prediction of boundary flows, and possibly use shorter time intervals for when it is updated. In addition, it would be interesting to compare the accuracy of the proposed hybrid prediction approach with a strict non-parametric approach for predicting travel time during different conditions.

A known problem is that fixed sensors occasionally fails to report data. Given that the used method for prediction of mainline sensor data demand that data is available for all periods this is a problem. Using a method that could handle missing data would improve the predictions in two ways. First of all, a longer time period could be used as training data which presumably would improve the predictions. However, more importantly, a prediction of mainline sensor data could be made for a larger number of sensors, which we expect would have a significant impact on the accuracy of the predicted traffic state, in particular for longer time horizons.

Another possible way to improve the accuracy and robustness of the proposed hybrid prediction approach is to better utilize the advantages of the non-parametric and the parametric prediction methods. The non-parametric neural network is more suited for prediction of recurring congestion. At the same time the parametric model is superior during non-recurring incidents, which can be modelled as reduced capacities in the CTM-v. Using EnKF parameters to govern the influence of each method, either the parametric or non-parametric prediction, can be trusted more or less depending on the current traffic situation. This, however, introduce additional filter parameters related to predicted measurement uncertainty, which need to be calibrated.

Even though there are many possible ways to improve the accuracy of the MMS-model, an important question to raise is which accuracy is sufficient for the user and the traffic management center. In a White Paper published by FHWA, see Ahn et al. (2008), data quality measures within the context of real-time travel information applications are investigated. The conclusions in the paper are based on a number of previous studies and reports. For travel time data collection or estimation an error range of 10-17 % has been recommended while the recommended error range for speed data collection or estimation is 5-20 %. Among the papers studied is a paper by Toppen and Wunderlich (2003) where the authors concluded that at the highest levels of accuracy, little is gained by making further improvements and if the accuracy error is below 5 %, it makes little sense to invest in improved accuracy. In another of the studied papers, written by Tarnoff et al (2008), it is concluded that if the error exceeds 20 %, the public lose confidence in the information source, undermining the support and usefulness of the system. The accuracy of the estimations performed with the MMS-model lies between 5 and 12 % depending on how the MAPE values are calculated, for the whole route or for each individual section separately. Hence, the accuracy of the MMS-
model is acceptable, at least for the studied scenario. To further confirm that the MMS-model is suitable for real time estimations and predictions it should be evaluated for other scenarios where, for example, an incident occurs and the space-time speed contour plot deviates from the normal daily patterns. The hybrid short-term prediction approach has successfully been tested for a 30-minute prediction horizon for the whole highway network in and around Stockholm but the delay and scalability of the model should also be studied further.
8 References


Chen H. and Grant-Muller S. (2001), Use of Sequential Learning for Short-Term Traffic Flow Forecasting, Transportation Research Part C.


Kwon J., Coifman B. and Bickel P. (2000), Day-to-day travel time trends and travel time prediction from loop detector data, Transportation Research Record, 2000. 1717: p. 120-129.


Mahmassani H. S., Fei X., Eisenman S., Zhou, X. and Qin, X. (2005), DYNASMART-X evaluation for real-time TMC application: CHART test bed, Maryland Transportation Initiative, University of Maryland, College Park, Maryland.


Meschini L. and Gentile G. (2010), Real-time traffic monitoring and forecast through OPTIMA – Optimal Path Travel Information for Mobility Actions, in Proceedings of Models and


Schönhof M. and Helbing D. (2009), Criticism of three-phase traffic theory, Transportation Research Part B: Methodological, 43(7), 784-797.


Torday A. (2010), Simulation-based decision support system for real time traffic management, In 89th Transportation Research Board Annual Meeting 2010 (No. 10 - 2120), Washington, D.C.

Trafikverket (2013), TRVMB Kapacitet och framkomlighetseffekter - Trafikverkets metodbeskrivning för beräkning av kapacitet och framkomlighetseffekter i vägtrafikanläggningar, TRV 2013:64343.

Treiber M. and Helbing D. (2002), Reconstructing the spatio-temporal traffic dynamics from stationary detector data, Cooper@tive Tr@ns@tion Dyn@rics 1 3.1-3.24.


Van Hinsbergen C.P.I.J., Van Lint J.W.C. and Van Zuylen H.J. (2009), Bayesian training and committees of state-space neural networks for online travel time prediction, Transportation Research Record: Journal of the Transportation Research Board 2105, 118-126.


Van Lint J.W.C., Hoogendoorn S.P. and Van Zuylen H.J. (2005), Accurate Freeway Travel Time Prediction with State-Space Neural Networks under missing Data, Transportation Research Part C.

Vlahogianni E.I., Karlaftis M.G. and Golias J.C. (2014), Short-term traffic forecasting: Where we are and where were going, Transportation Research Part C: Emerging Technologies, 2014.


