

Traffic Speed Estimation and Prediction Using Floating Car Data

Felix Rempe

Vollständiger Abdruck der von der Fakultät für Bauingenieurwesen und
Umweltwissenschaften der Universität der Bundeswehr München zur Erlangung des
akademischen Grades eines

Doktors der Ingenieurwissenschaften
genehmigten Dissertation.

Gutachter:

Prof. Dr.-Ing. Klaus Bogenberger

Prof. Dr. rer. nat. Peter Wagner

Prof. Dr. Boris Kerner

Die Dissertation wurde am 08.05.2018 bei der Universität der Bundeswehr München
eingereicht und durch die Fakultät für Bauingenieurwesen und Umweltwissenschaften am
22.11.2018 angenommen. Die mündliche Prüfung fand am 04.12.2018 statt.

BUNDESWEHR UNIVERSITY MÜNCHEN

DOCTORAL THESIS

Traffic Speed Estimation and Prediction Using Floating Car Data

by

Felix REMPE

*A thesis submitted to the Bundeswehr University München
for the degree of Doktor der Ingenieurwissenschaften (Dr.-Ing.)*

in the

Faculty of Civil Engineering and Environmental Sciences
Institute for Transport and Regional Planning
Department of Traffic Engineering

First advisor: Prof. Dr. Klaus BOGENBERGER

Second advisor: Prof. Dr. Peter WAGNER

Third advisor: Prof. Dr. Boris S. KERNER

Munich, January 2019

Acknowledgment

The present thesis concludes the results of my studies conducted to gain new insights into the dynamics of traffic flow and to contribute to the improvement of traffic systems. In this preface I like to express my thanks to all the great people who supported me in reaching this goal.

My special thanks go to Professor Bogenberger. He always took the time to discuss ongoing projects, generated great ideas how to go on and gave me a lot of freedom to develop and follow own ideas. At the same time, his remarkable guidance pushed me forward and allowed me to reach challenging, but not impossible goals. Besides Klaus, I had two excellent supervisors in the traffic & routing team at BMW Group, Philipp and Ulrich. The regular and fruitful discussions resulted in many approaches that were tried out, continuously improved, and finally some of them turned into successful solutions. From their indispensable input and feedback I learned so much which made this time as a doctoral student very rewarding. Apart from technical and academic considerations, the colleagues at BMW as well as at the chair created a very positive atmosphere. This made me enjoy the work and helped to get through occasional challenging times. My sincerest thanks go to my family, friends and especially my partner Katrin. They gave me security and strength, allowed me to relax and regain energy off work.

Finally, I like to thank Professor Wagner and Professor Kerner for their valuable input, all the colleagues at university for proof-reading and all other people not mentioned by name, who contributed in one or another way to this thesis!

Executive Summary

This thesis proposes novel methods to use Floating Car Data (FCD) for applications in traffic speed estimation and prediction. Three approaches are developed and evaluated using real FCD collected by a large fleet of vehicles.

The first method targets traffic speed estimation on freeways. It describes how to process raw and sparse trajectory data using empirical traffic features described in the Three-Phase theory to compute a continuous traffic speed estimate in space-time. Therefore, first the three traffic phases are reconstructed, and second, traffic velocities inside each phase domain are estimated. In an evaluation with 101 congestion patterns the method achieves higher accuracies than comparable state-of-the-art approaches. An efficient implementation using the Fourier transform as well as a high degree of flexibility and robustness contribute to its practical utilization.

The second method seeks to provide short-term congestion front forecasts. Continuous speed information is analyzed for current hazardous congestion fronts. Flow data and speed information in the proximity of the fronts are fused and processed with an analytical front propagation model. The results of a comparison of several variants of the method and a naive model show that one of the proposed model variants forecasts more accurately in a 10-minute horizon than all others.

The third method focuses congestion in urban road networks. Using one year of FCD, a small number of subnets showing regular congestion are extracted. A statistical analysis of the congestion level inside these subnets reveals patterns of spatio-temporal congestion. Based on these patterns, a network-wide congestion forecast method is developed and applied. Its higher accuracy compared to typical time series forecasts indicate that these subnets serve as valuable features for prediction models to reflect the network-wide status of congestion.

Kurzzusammenfassung

In der vorliegenden Dissertation werden neue Verfahren für den Einsatz von FCD in Anwendungen zur Verkehrslagenschätzung und -prognose beschrieben. Insgesamt drei Ansätze werden entwickelt und mit den FCD einer großen Fahrzeugflotte evaluiert.

Die erste Methode zielt auf die Verkehrslageschätzung auf Autobahnen ab. Sie basiert auf einer Verarbeitung von Rohdaten mit Hilfe von charakteristischen Verkehrseigenschaften in eine raumzeitlich kontinuierliche Repräsentation der durchschnittlichen Verkehrsgeschwindigkeit. Dafür werden zuerst die drei Verkehrsphasen in Raum und Zeit rekonstruiert und, in einem zweiten Schritt, die Geschwindigkeiten innerhalb der Phasen abgeschätzt. In einer Evaluation mit über 101 historischen Stausituationen resultiert das Verfahren in höheren Genauigkeiten als vergleichbare Methoden. Sowohl eine effiziente Implementierung als auch ein hoher Grad an Flexibilität und Robustheit tragen zur praktischen Anwendung des neuen Verfahrens bei.

Die zweite Methode zielt auf die Kurzfristprognose von Staufronten ab. Zuerst werden raumzeit-kontinuierliche Geschwindigkeitsinformationen auf gefährliche Staufronten untersucht. Danach werden Fluss und Geschwindigkeitsdaten in der Nähe dieser Fronten mit einem analytischen Modell verarbeitet, um die Position der Staufronten für einen kurzen Zeithorizont numerisch zu bestimmen. In einem Vergleich mit mehreren Varianten zur Formulierung des Modells kristallisiert sich eine Variante als vielversprechendste für einen bis zu 10-minütigen Prognosehorizont heraus.

Die dritte Methode fokussiert Stau in urbanen Netzen. Zuerst werden auf Basis von einem Jahr FCD Subnetze identifiziert, die sich regelmäßig stauen. In einer statistischen Stauanalyse werden die Stauzustände innerhalb dieser Subnetze auf Beziehungen und Muster untersucht. Basierend auf den resultierenden Muster wird ein netzweites Prognosemodell entwickelt. Eine Evaluation zeigt, dass die Berücksichtigung des netzweiten Stauzustandes genauere Prognosen ermöglicht als eine rein lokale Betrachtung. Das motiviert die Verwendung von speziellen Merkmalen, wie beispielsweise die Subnetze, für datengetriebene Verkehrsprognosen in urbanen Netzen.

Related Publications

Journals:

- Felix Rempe, Philipp Franeck, Ulrich Fastenrath and Klaus Bogenberger. “A phase-based smoothing method for accurate traffic speed estimation with floating car data”. In: Transportation Research Part C: Emerging Technologies 85 (2017), pp. 644–663.

Conference Proceedings:

- Felix Rempe, Lisa Kessler and Klaus Bogenberger. “Fusing probe speed and flow data for robust short-term congestion front forecasts”. In: 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) (2017), pp. 31–36.
- Felix Rempe, Philipp Franeck, Ulrich Fastenrath and Klaus Bogenberger. “Online Freeway Traffic Estimation with Real Floating Car Data”. In: Intelligent Transportation Systems (ITSC) (2016), pp. 1838–1843.
- Felix Rempe, Gerhard Huber and Klaus Bogenberger. “Travel time prediction in partitioned road networks based on floating car data”. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC) (2016), pp. 1982– 1987.
- Felix Rempe, Gerhard Huber and Klaus Bogenberger. “Spatio-Temporal Congestion Patterns in Urban Traffic Networks”. In: Transportation Research Procedia 15 (2016), pp. 513–524.

Conference Presentations:

- Felix Rempe, Klaus Bogenberger. “Feature Engineering for data-driven traffic state forecast in urban road networks”. In: 98th Annual Meeting of the Transportation Research Board (2019).
- Felix Rempe, Philipp Franeck, Ulrich Fastenrath and Klaus Bogenberger. “A Phase-Based Smoothing Method for Accurate Traffic Estimation with Floating Car Data”. In: 96th Annual Meeting of the Transportation Research Board (2017), #05542.
- Felix Rempe and Klaus Bogenberger. ”A comparison of traffic estimation algorithms based on floating car data”. Presented at 1st Symposium on Management of Future motorway and urban Traffic Systems. (2016), Chania, Greece.

Table of Contents

| | |
|--|------------|
| Acknowledgment | iii |
| Executive Summary | v |
| 1 Introduction and Motivation | 1 |
| 1.1 Research Objectives | 3 |
| 1.2 Outline of the Dissertation | 4 |
| 2 State of the Art | 6 |
| 2.1 Traffic Flow Theory | 6 |
| 2.1.1 Introduction | 6 |
| 2.1.2 Three-Phase Traffic Theory | 9 |
| 2.2 Sensing Traffic Conditions | 13 |
| 2.3 Traffic State Estimation and Prediction | 17 |
| 3 Data | 23 |
| 3.1 Sampling and Collecting Data | 23 |
| 3.2 Preprocessing | 24 |
| 3.3 Statistical Exploration of Available Data | 25 |
| 4 Phase-based Traffic Speed Estimation with FCD | 28 |
| 4.1 Motivation | 28 |
| 4.2 Related Work | 29 |
| 4.3 Solution Approach | 31 |
| 4.3.1 Representation of FCD in Space-Time | 33 |
| 4.3.2 Traffic-Motivated Data Smoothing | 35 |
| 4.3.3 Modeling Phase Probabilities | 37 |
| 4.3.4 Estimating Phase Velocities | 46 |
| 4.3.5 Aggregating Probabilities and Velocities | 47 |
| 4.4 Evaluation | 48 |
| 4.4.1 Setting Parameters | 48 |
| 4.4.2 Implementation | 51 |
| 4.4.3 Qualitative Evaluation | 56 |
| 4.4.4 Quantitative Assessment of Estimation Accuracy | 62 |
| 4.4.5 Reconstruction of Mega-jams | 84 |
| 4.4.6 Sensitivity Analysis | 87 |
| 4.4.7 Run-time Analysis | 89 |
| 4.5 Conclusion and Outlook | 91 |

| | |
|--|------------|
| 5 Forecasting Congestion Fronts using FCD and Flow Data | 93 |
| 5.1 Related Work | 94 |
| 5.2 Prediction Model | 95 |
| 5.2.1 Phase Front Propagation | 96 |
| 5.2.2 Estimating Phase Flows | 98 |
| 5.2.3 Estimating Phase Densities | 99 |
| 5.3 Evaluation | 101 |
| 5.3.1 Accuracy Assessment | 101 |
| 5.3.2 Results | 103 |
| 5.3.3 Discussion | 106 |
| 5.4 Conclusion and Outlook | 107 |
| 6 Congestion Analysis and Prediction in Urban Road Networks | 109 |
| 6.1 Related Work and Solution Approach | 110 |
| 6.2 Definition of Congestion Clusters | 112 |
| 6.2.1 Dynamic Congestion Pockets | 112 |
| 6.2.2 Static Congestion Clusters | 114 |
| 6.3 Data-Driven Congestion Prediction in a Clustered Network | 116 |
| 6.3.1 Overview of the Machine Learning Pipeline | 117 |
| 6.3.2 KNN Travel Time Predictor | 118 |
| 6.4 Evaluation | 121 |
| 6.4.1 Data Preparation | 121 |
| 6.4.2 Cluster Metrics | 122 |
| 6.4.3 Static Clusters in the Munich Road Network | 126 |
| 6.4.4 Congestion Pattern Analysis | 128 |
| 6.4.5 Cluster-Based Congestion Prediction | 139 |
| 6.5 Conclusion and Discussion | 144 |
| 7 Conclusion | 147 |
| 7.1 Summary | 147 |
| 7.2 Outlook | 148 |
| A Appendix | 152 |
| Abbreviations | 161 |
| List of Symbols | 163 |
| List of Figures | 166 |
| List of Tables | 169 |
| Bibliography | 170 |

Chapter 1

Introduction and Motivation

Traffic congestion is a major problem for transportation networks all over the world. The additional required travel time for road users and the need for buffer times due to the unreliability of estimated arrival times constitute a significant waste of resources. Moreover, due to an overall increase of travel time and acceleration processes, traffic on severely congested roads is prone to cause higher emissions and to reduce air quality.

Therefore, it is of utmost importance to develop strategies that target the reduction of traffic congestion. One strategy is to increase the road capacity by constructing new roads, adding lanes to existing roads or re-designing given roads. While this is an effective approach to relax traffic problems locally, it is costly, requires free space and often happens at the expense of the environment. Another strategy is to use existing roads more efficiently, i.e. to apply control measures that increase the throughput of the traffic network. In order to do so, a variety of control tools such as Variable Speed Limits (VSLs), dynamic traffic signal timing, dynamic shoulder use etc. are utilized. A third promising approach is to distribute traffic demand over the network. Therefore, current and predictive traffic state information or individual route recommendations harmonized with the expected road utilization are delivered to the road users. Optimally, a part of the road users reroutes, which relieves known bottlenecks. Despite its potential, the effectiveness of this approach depends on several factors such as the driver's willingness to reroute, the quality of the traffic state predictions as well as the route recommendations.

These strategies face a similar fundamental problem: They require accurate information about traffic conditions on the road in order to be effective. E.g. the construction of new or broader roads needs to be based on the long-term analysis of traffic conditions in order to enhance the capacity of the most severe bottleneck. Strategies that are supposed to act in real-time require exact information about current and future traffic conditions in order to perform well. Controlled traffic signals at intersections reduce overall waiting

times most effectively if they know about the current queue lengths. A ramp-metering system at freeways requires accurate short-term traffic state predictions in order to dose the in-flow optimally. Furthermore, road users and traffic managers require traffic speed predictions in order to estimate travel times along potential routes and determine routes which optimize individual or collective goals.

Providing accurate traffic information is challenging. Up until the last years the most common traffic sensors were stationary detectors such as inductive loops, which are integrated into the road pavement. They sense passing vehicles and report high quality data at short time intervals. Though, due to high costs of installation and maintenance usually only main roads are equipped and distances between two sensors can exceed several kilometers. In urban regions, the coverage with detectors and the access to data is usually even more limited and varies from city to city. This sparsity of conventional data and the resulting lack of accurate traffic information constitute a vast limitation for many applications.

The development of cheap and efficient electronics during the last decades nowadays also affects traffic engineering. New vehicles on the road are equipped with a multitude of sensors and electronic components. They are globally connected using mobile phone networks and, in the near future, will be connected via local data networks (Vehicle-to-Everything (V2X)). These new technologies enable to collect sensor data, also called FCD, from fleets of vehicles and utilize them for traffic engineering applications. Due to ever-decreasing costs, the amount of collected and transmitted data is increasing every year. Compared to stationary sensors such as inductive loops this new type of data has several advantages. It is able to cover the entire road network instead of just pre-defined locations. Furthermore, while detectors are usually configured to provide average traffic quantities, data of individual vehicles allows to determine traffic conditions with a higher accuracy than ever before. Therefore, the use of vehicle-generated data provides huge potential for traffic engineering applications.

On the other hand, this type of data challenges existing advances in traffic state estimation and prediction. Since (for now) only data of a few vehicles is reported, the usually considered macroscopic traffic quantities such as flow and density are not available. Moreover, the number of reporting vehicles on a road segment varies over time and place. This demands a traffic state estimation method to be highly flexible with respect to given sparse data.

Due to these reasons novel methods for traffic state estimation and prediction using FCD need to be developed. Additionally, the potential of this technology for, yet, non-monitored road networks needs to be assessed. This thesis is dedicated to these objectives. Specifically, novel methods are presented that exploit the strengths of FCD

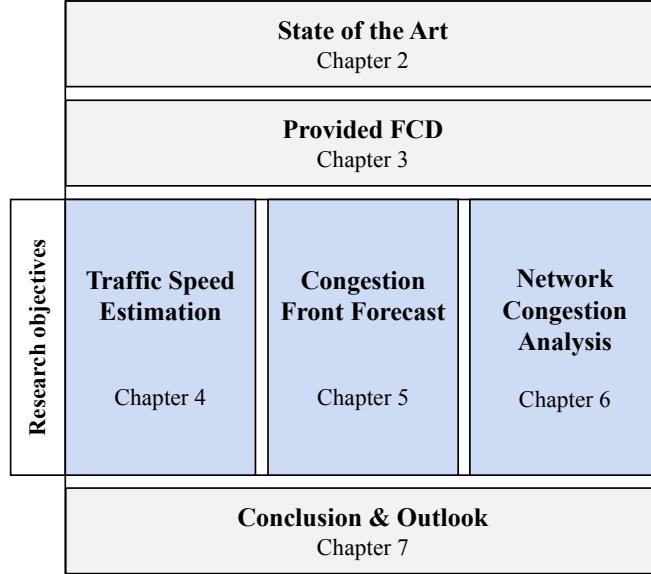


Figure 1.1. Structure of this thesis

for various applications which enable traffic managers, control algorithms and travelers to make better decisions and, finally, are supposed to contribute to the avoidance of traffic congestion.

1.1 Research Objectives

Traffic state estimation and prediction is a broad field of research in traffic engineering including a multitude of requirements and challenges for various scenarios. To deal with all problems is out of scope of one dissertation. In this section three specific research objectives are posed, which are highly relevant for traffic engineering, and which benefit significantly from the usage of FCD (see Figure 1.1):

1. In comparison to minor roads, a congested freeway affects many travelers. Furthermore, due to potentially high velocities many severe accidents occur on freeways. Accurate traffic information on freeways enable numerous applications such as VSLs, ramp metering, congestion warnings and travel time estimates which improve traffic safety, comfort and efficiency. Though, current estimation algorithms are challenged by the sparsity of FCD and high dynamics of vehicle flow in dense traffic conditions in time and space. The first objective is to develop a practice-ready method that combines collected high-resolution FCD with current traffic flow theory in order to **estimate traffic speeds**. It is supposed to be more

accurate than existing methods and consider practical issues such as efficiency, flexibility and robustness.

2. For drivers, upcoming congestion fronts are dangerous hazards. In order to alert the driver and increase their attention, (in-vehicle) congestion front warnings are valuable information which enhance traffic safety and, due to fewer accidents, improve traffic efficiency. Though, the provision of reliable warnings at the right point in time is challenging. Reasons are the non-stationary character of congestion fronts, the sparsity of data as well as the time that is required to acquire and process measurements. In order to improve the position accuracy and provide warnings ahead of time, short-term congestion front forecasts are required. Based on a fusion of speed and flow data, the objective is to develop and evaluate a method for **short-term congestion front forecasts**.
3. Due to the increasing urbanization of society the severity of traffic congestion in urban road networks is getting increasingly relevant. Traffic flow dynamics in urban networks differs substantially from dynamics on freeways such that many existing methods can not be transferred. Challenging is that an urban network consists of thousands of mutually connected (signalized) road segments, and until recently, sensor data was vastly limited. Nowadays, with FCD, large-scale data for urban road networks is available. The third objective is to study **network-wide congestion** and to develop new **tools for congestion pattern analysis and prediction** in urban networks.

1.2 Outline of the Dissertation

This work is structured in the following way. Section 2.1 presents the state of the art in traffic flow theory on freeways and summarizes the road to modern traffic flow theory. Next, an overview of current sensor technology including its strengths and limitations is given in section 2.2. Section 2.3 reviews the state-of-the-art literature about traffic speed estimation and prediction on freeways and urban networks and identifies research gaps. Subsequently, the contributions of this dissertation are put in context with existing works. Next, the available FCD collected from a fleet of vehicles which is used for method development and evaluation throughout this dissertation is introduced in chapter 3.

The subsequent three chapters deal with the three research objectives. Chapter 4 describes a novel approach to combine the Three-Phase traffic theory with sparse FCD in order to estimate freeway traffic speeds. After a summary of related approaches and their limitations when applied to FCD (section 4.2), the solution approach is developed

in section 4.3. Section 4.4 presents the results of an extensive evaluation that compares the accuracy of the developed method with other state-of-the-art approaches. In addition, its computational efficiency and sensitivity to parameter changes are analyzed. The conclusion in section 4.5 summarizes the chapter, discusses open issues and proposes future directions.

Chapter 5 presents a model that provides short-term congestion front forecasts given FCD and flow data. Variants of an analytical forecast model are motivated in section 5.2 and their performance is compared in section 5.3. Section 5.4 summarizes the contributions of the method and gives an outlook.

Next, the potential of FCD for urban traffic speed estimation and prediction is focused. Chapter 6 develops an approach that enables network-wide congestion pattern analysis and forecast. Frequently congested regions of a network are identified using a novel clustering approach (section 6.2). Based on these so-called congestion clusters, a methodology to perform a data-driven network-wide congestion prediction is proposed (section 6.3). Subsequently, the clustering approach is applied to one year of FCD collected in the congestion-prone road network of Munich city. First, statistical tools are applied in order to identify spatio-temporal congestion patterns using the clusters (section 6.4.4). The effectiveness of the prediction method is presented in section 6.4.5. Section 6.5 reviews the proposed and evaluated method and discusses open issues.

To conclude the thesis, chapter 7 summarizes the contributions of the dissertation to the state of the art in traffic speed estimation and prediction using FCD. The final outlook gives a vision of future traffic systems and describes potential ways in order to advance to these visions.

Chapter 2

State of the Art

This chapter first gives a short introduction to traffic flow modeling and, in particular, the Three-Phase traffic theory. Second, an overview of current sensor technology applied to observe traffic conditions is presented. Finally, the need for traffic state estimation and prediction algorithms is motivated and the corresponding state of the art is elaborated.

2.1 Traffic Flow Theory

2.1.1 Introduction

The purpose of traffic flow theory is the detailed understanding of the spatio-temporal dynamics of traffic flow. As such, it is fundamental for all types of applications in transportation engineering. Advances in traffic flow theory enable the enhancement of design, operation and development of Intelligent Transportation Systems (ITS).

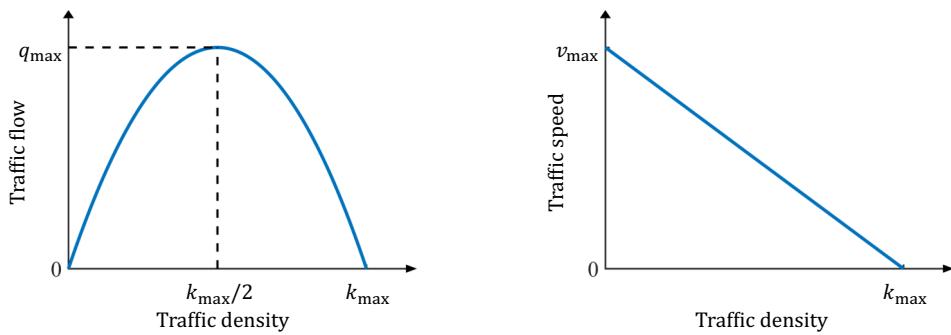


Figure 2.1. Fundamental diagram according to Greenshields' studies

Traffic engineering is said to be born with the empirical studies of Greenshields in 1935 (Greenshields, 1935). In measurements of vehicle speed and vehicle density he identified a linear relationship between these two variables. Consequently, flow and density of traffic form a parabolic relationship (see Figure 2.1). This first type of Fundamental Diagram (FD) was the beginning of 80 years of empirical and theoretical studies conducted by a broad community of civil engineers, mathematicians, physicians and computer scientists. (Greenberg, 1959), for example, noticed that an exponentially decreasing function fits traffic speed and traffic density better. As a consequence, also the relation between flow and density is an exponential function. Up to now, many more functions have been proposed that claim to fit empirical data best.

Besides the development of FDs, space-time traffic flow models have been developed. They consider also the temporal evolution of traffic conditions. Typically, these models are categorized into macroscopic and microscopic ones. Macroscopic models consider traffic as a flow of particles that can be described sufficiently well with macroscopic variables such as flow, density and speed. Most famous is the Lighthill-Whitham-Richards (LWR) model which applies the conservation law of fluid to traffic (Lighthill and Whitham, 1955; Richards, 1956). It assumes a function $q(k)$ (i.e. a FD) between flow q and density k :

$$\frac{\partial k(t, x)}{\partial t} + \frac{\partial q(k(t, x))}{\partial x} = 0. \quad (2.1)$$

Given simple boundary conditions of a space-time domain, this model can be solved analytically, which results in the kinematic wave equations of traffic (Richards, 1956). However, applied to real sensor data no analytical solution exists or is challenging to determine. In this case, the usual approach is to set up a numerical simulation in form of a Cell Transmission Model (CTM) using the Godunov discretization scheme (Daganzo, 1994; Daganzo, 1995; Lebacque, 1996). In doing so, many models have been developed which differ in the assumed FDs, additional (stochastic) terms, the modeling of in- and outflows on freeways etc. Since models of this type are based only on the continuity equation, they are also referred to as *first-order* models. First-order models are limited with respect to their modeling capabilities as they do not allow to integrate further traffic dynamics. Therefore, some higher-order models have been proposed that pose additional equations. The first, on which many other models are based, is the Payne model (Payne, 1971). Besides continuity, it models the influence of the vehicles' surroundings on their velocities. Famous other higher-order models include the Kerner-Konhäuser model (Kerner and Konhäuser, 1994), the Gas-kinetic-based model (Treiber et al., 1999) or the model proposed by (Aw and Rascle, 2000).

On the other hand, microscopic models seek to describe the motion of each vehicle individually depending on conditions in the proximity of the vehicle. Variables such

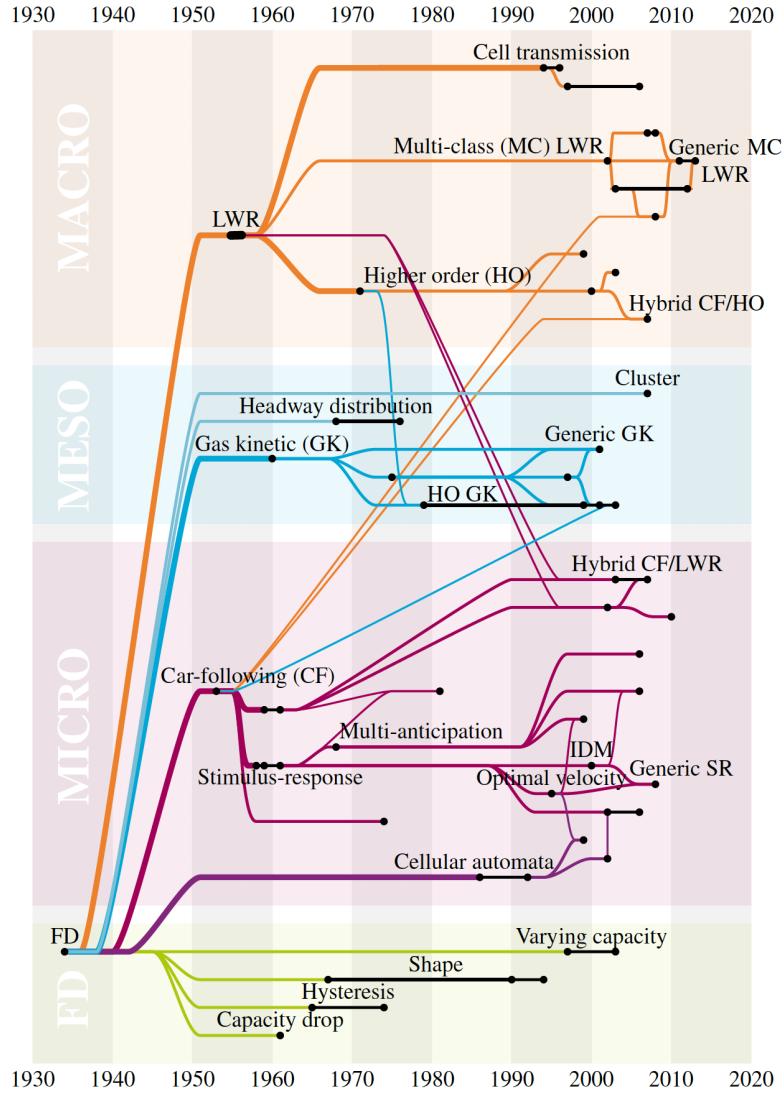


Figure 2.2. Overview of traffic models (Wageningen-Kessels et al., 2015)

as the individual velocity, position and velocity difference to the preceding vehicle may influence its acceleration. The first model was published in (Reuschel, 1950) which assumes that a vehicle acceleration only depends on the speed difference and distance gap. Among many others, more sophisticated models that have been developed during the last decades are the Intelligent Driver Model (IDM) (Treiber et al., 2000), Newell's car-following model (Newell, 2002), the psychophysical Wiedemann model (Wiedemann, 1974) or the stochastic Gipps model (Gipps, 1981). In order to overcome computational issues solving the equations for a huge number of vehicles, microscopic cellular automata have been proposed that simplify the representation and simulation of traffic in greater networks (Nagel and Schreckenberg, 1992).

Besides micro- and macroscopic formulations, some mesoscopic models were published.

Some describe the behavior of vehicles in aggregate terms such as probability distributions. For instance, (Buckley, 1968) propose a model for time headway distributions, (Mahnke and Kühne, 2007) study jam formation by modeling clusters of cars and (Hoogendoorn and Bovy, 2001) develop a gas-kinetic traffic model. Others describe the movement of individual vehicles based on macroscopic traffic conditions, e.g. the MATsim (Horni et al., 2016) or MESO project (integrated into SUMO) (Krajzewicz et al., 2012; Eissfeldt, 2004).

The presented approaches are a brief summary of a multitude of advances in traffic flow modeling. Figure 2.2, published in (Wageningen-Kessels et al., 2015), gives a more detailed overview of the variety of models developed since Greenshields (a complete overview is published in (Wageningen-Kessels, 2013)).

2.1.2 Three-Phase Traffic Theory

Parallel to the development of more sophisticated analytical models, the increasing amount of available traffic data was studied and lead to advances in empirical traffic research. Phenomena such as the probabilistic nature of traffic breakdown (Elefteriadou et al., 1995; Kühne et al., 2002; Brilon et al., 2005; Mahnke and Kühne, 2007), induced traffic breakdowns, the wide scattering of traffic flow in congested traffic and the pinch effect (Kerner, 2004) were observed and studied thoroughly by various researchers. A qualitative traffic theory that presents an explanation for all of these traffic phenomena is the Three-Phase traffic theory (Kerner et al., 2004; Kerner, 2009; Kerner, 2017). Its development and evaluation is based on the analysis of extensive datasets of congestion patterns on German and international freeways (Rehborn et al., 2011). In contrast to previous traffic theories which usually distinguish between free traffic and congested traffic, in this theory the existence of three traffic phases is postulated: The congested state is further divided into a synchronized flow phase and a Wide Moving Jam (WMJ) phase. In addition to that, unlike many other theories which are based on a FD describing a one-dimensional relation between flow and density, the Three-Phase traffic theory allows a wide scattering of flow-density pairs in congested traffic. In the theory this is referred to as the two-dimensional state of traffic flow. In Figure 2.3 the three phases are depicted in flow-density and speed-density plane.

They can be distinguished by its characteristics (Kerner et al., 2004):

- In **free traffic**, vehicles move with high speed and are mainly restricted by the free driving speed of the vehicle on that road. Vehicles can change lanes freely and overtake slower ones, such that most of the time a vehicle's motion is not bound by

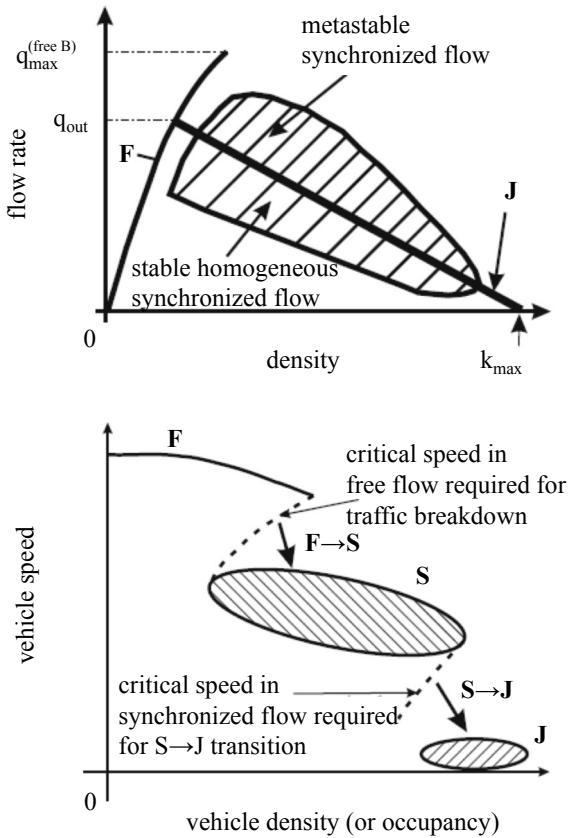


Figure 2.3. Schematic visualization of the Three Phases in flow-density and speed-density plane (Kerner et al., 2004)

a preceding vehicle. Flow and density are in nearly linear relation. In this state, the maximal capacity of the road q_{max} can be achieved.

- In **synchronized flow** and in comparison to free flow, traffic density is significantly higher and vehicles' speeds are significantly lower. Vehicles adapt the gap to a preceding vehicle depending on their current speed (see Figure 2.4). In contrast to many other microscopic models which assume a fixed time headway, the Three-Phase traffic theory allows vehicles to maintain various speed-gap relations. This in turn explains the potential wide scattering of traffic flow in this phase. Usually traffic speeds among different lanes are similar, motivating the name synchronized flow. In case traffic speeds among lanes tend to diverge and the speed is higher on one of the lanes, vehicles switch to the faster lane. As a consequence, the system equilibrates and recovers synchronized conditions. A commonly observable characteristic of the phase region in space-time is that the downstream front sticks to a bottleneck such as an on-ramp (Figure 2.4 right). Even if the synchronized flow phase emerges in absence of a fixed bottleneck and the downstream front propagates upstream, typically the next upstream bottleneck *catches* the downstream

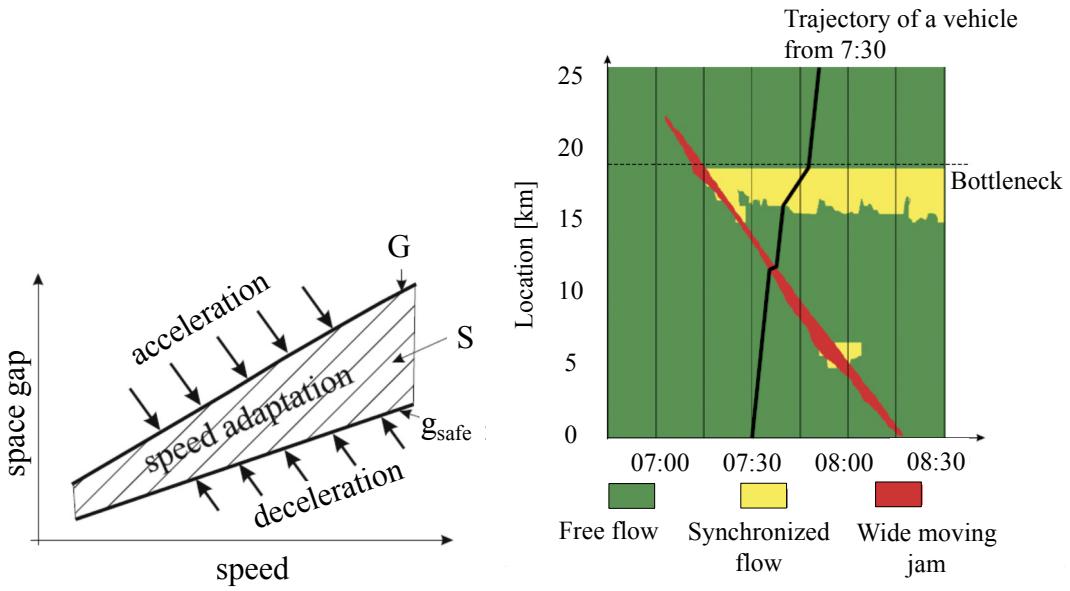


Figure 2.4. The speed adaptation effect according to the Three-Phase theory (left); Propagation characteristics of phases in synchronized flow and WMJ (right) (Kerner et al., 2004)

front (Kerner et al., 2004). Due to the wide scattering of flow-density the synchronized flow phase can be stable and homogeneous, or metastable. In metastable flow there exists a non-zero probability that a WMJ emerges.

- The **WMJ phase** is a phase with low vehicle velocities, which can decrease down to 0 km/h. The main characteristic of this phase is that the downstream phase front propagates upstream with a constant velocity (see Figure 2.4 right). This shock wave propagates through free or congested phases as long as the upstream front of the WMJ phase does not meet the downstream front. In that case, the moving jam dissolves.

The transitions between the phases occur as follows:

- **F→S:** The F→S transition represents a classical traffic breakdown which has been described in many publications (Daganzo, 1996; Hall and Agyemang-Duah K., 1991; Schoenhof and Helbing, 2007; Laval, 2007). Researchers mainly agree that this breakdown is of stochastic nature occurring due to local microscopic disturbances (Elefteriadou et al., 1995; Kühne et al., 2002; Brilon et al., 2005; Mahnke and Kühne, 2007). The probability of such a breakdown is zero as long as traffic flow is below the minimal capacity of the road (Kerner, 2017). With increasing traffic flow, limited by the maximal capacity of the road, the probability of a traffic breakdown increases.

- **S→J:** The metastable flow is defined as a state of traffic flow in which, given a certain traffic density, the corresponding flow exceeds the 'J' line (see Figure 2.3). This line with a gradient of $v_{cong} \approx -15 \text{ km/h}$ (Treiber and Helbing, 2003; Treiber et al., 2010b; Schoenhof and Helbing, 2007) represents the relations between flow and density which suffice to cause a WMJ. Consequently, a transition from metastable synchronized flow to a WMJ phase causes the growth of a WMJ phase over time. If the traffic state upstream of a WMJ phase is in homogeneous conditions, an existing WMJ will dissolve over time. This is important in order to understand the emergence of WMJs. In a microscopic view on metastable traffic flow, time headways between vehicles are relatively low. In case of a local disturbance, such as a vehicle changing lanes or braking abruptly, the following vehicle tends to over-decelerate (Kerner, 2004). Similarly, following vehicles over-decelerate as well. This results in a shock-wave that propagates upstream and each affected vehicle tends to slow down slightly stronger. If traffic density is high, vehicle velocities can decrease down to total stoppage. These microscopic disturbances can emerge in stable synchronized flow as well, though, traffic density in this state does not suffice to develop a full WMJ.

Although the Three-Phase theory is not fully accepted in the scientific community (yet) (Schönhof and Helbing, 2009; Treiber et al., 2010a), it achieves to explain many empirically observed traffic phenomena and combines them into one theory. E.g. it explains the probabilistic nature of traffic breakdown, induced traffic breakdowns by WMJs propagating through a bottleneck, the similar catch-effect, the wide-scattering of traffic flow in congested traffic and the pinch effect, which describes the nucleation of WMJs in congested traffic (Kerner et al., 2004). Furthermore, several microscopic simulation models have been developed that reproduce phenomena described by Kerner et. al. For instance, (Kerner and Klenov, 2010) build a stochastic microscopic model based on the assumptions of the Three-Phase theory, reproducing the phases and effects described in the theory; (Knospe et al., 2002) develop a microscopic model focusing on smooth and comfortable driving that reproduces three phases; (Laval, 2007) proposes a simple microscopic model that reproduces the catch-effect described by Kerner; (Hoogendoorn et al., 2008) construct a macroscopic model based on the LWR with a stochastic component that generates similar congestion patterns as described in the Three-Phase traffic theory. Additionally, (Wagner and Lubashevsky, 2003; Wagner, 2012) show that the net time headway (and speed difference) of individual drivers to a preceding vehicle varies over time. This finding supports the assumed microscopic model of the Three-Phase theory (Figure 2.4) and is a possible explanation for the wide scattering of macroscopic flow and density.

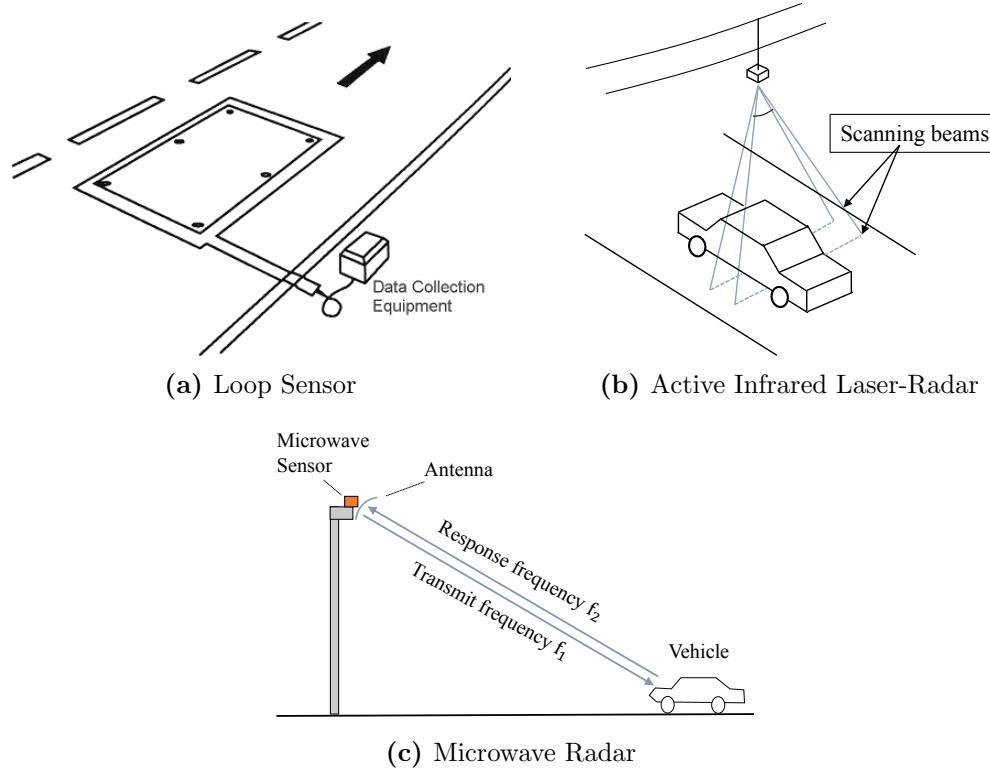


Figure 2.5. Sketches of an inductive loop detection system, a microwave radar and an infrared laser radar¹

2.2 Sensing Traffic Conditions

A key component of traffic systems are measurements of current traffic conditions in a road network. Not only for real-time applications up-to-date measurements of traffic conditions are indispensable, but also for long-term optimizations of networks rich traffic datasets are required. In order to collect data, different sensor technologies are utilized. Most of currently used sensors can be categorized into (stationary) spot sensors and (stationary or mobile) section sensors. Spot sensors refer to devices installed at a fixed position along the road measuring a traffic quantity such as vehicle speed, vehicle count etc. on a pre-defined local (and short) road interval. Section sensors provide information about the traffic state of a road section (or interval). Sensors may be stationary or mobile. Table 2.1 and 2.2 provide a brief summary of current sensors, their capabilities to measure traffic flow (q), density (k) and speed (v) as well as the main advantages and disadvantages of each technology.

¹<https://www.fhwa.dot.gov>

| (Stationary) Spot Sensors | Type | Principle | q | k | v | Strengths / Weaknesses |
|-----------------------------|--|-----------|---|---|---|--|
| | | | | | | |
| Inductive Loops | Electric coils integrated into the road pavement perceive passing vehicles. A connected control unit counts the number of passing vehicles (i.e. the flow). The commonly installed double loops are furthermore able to deduce vehicle speeds per lane (see Figure 2.5). | | ✓ | | ✓ | + Macroscopic traffic quantities per lane sampled at constant time intervals + Relatively robust against adverse weather conditions - Costly and invasive installation as well as maintenance |
| Microwave Radar | A device transmits electromagnetic waves with a constant frequency. It compares the frequency of reflected signals with the original frequency. The Doppler effect allows to determine the speed of the reflecting object. | | | ✓ | | + Direct measurement of vehicle speeds + Relatively cheap, robust against adverse weather conditions - Does not detect standing vehicles |
| Active Infrared Laser-Radar | Two infrared laser positioned above the roadway continuously sample the distance to the ground. Passing vehicles are detected and their speeds are estimated using the distance and time gap between the two laser positions. | | ✓ | ✓ | ✓ | + Accurate measurements of position, speed, count and class - Installation and maintenance is costly |
| Video System | A camera installed with good view on the road takes sequences of images and extracts traffic features such as vehicle speeds, time headways, density and flow | | ✓ | ✓ | ✓ | + Accurate traffic information for the observed road segment - Costly installation and maintenance - Unfavorable weather conditions such as rain, snow or fog reduce the reliability of the video system |

Table 2.1. Overview of (stationary) spot sensor technology

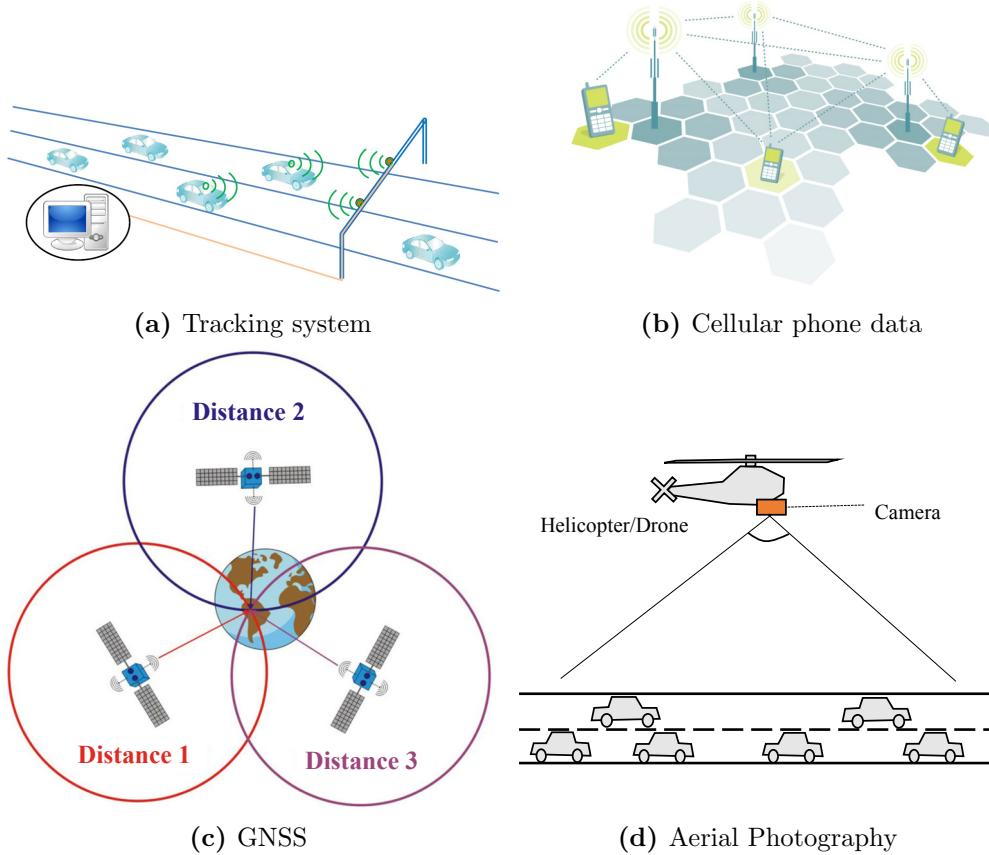


Figure 2.6. Section sensor technologies. Tracking systems (Ni, 2016); Sketch of the cells of a cellular network²; The principle of trilateration in order to determine a GNSS position on earth given satellite signals³; and aerial photography for traffic sensing

| Type | Principle | q | k | v | Strengths / Weaknesses |
|---|---------------------|---|---|---|--|
| (Stationary and Mobile) Section Sensors | Tracking System | | | ✓ | + Cheap variants of this technology are available + Installation does not disturb traffic - Information is limited to travel times |
| | Cellular Phone Data | | ✓ | | + Cheap technology with high availability - Low accuracy due to large cell sizes |

²<https://www.ifrahlaw.com/wp-content/>

³<http://www.dlg.org/>

| | | | | |
|-------------------|--|--|-----|---|
| GNSS Data/ FCD | A vehicle equipped with a GNSS device samples its geolocation on a regular basis. Transmitted positions and timestamps allow to reconstruct the trajectory and speed profile (see Figure 2.6). | | ✓ | + Cheap technology with high space-time accuracy – Observation of one vehicle is only a sample of the macroscopic speed – No lane accuracy |
| Aerial Imaging | Helicopters or drones take sequences of images of the traffic on the road network. Image processing techniques return positions and speeds of all vehicles (see Figure 2.6). | | ✓ ✓ | + Possibility to obtain comprehensive traffic data for a road – Limited continuous surveillance due to high costs – Challenging at night or at unfavorable weather conditions |

Table 2.2. Overview of section sensor technology

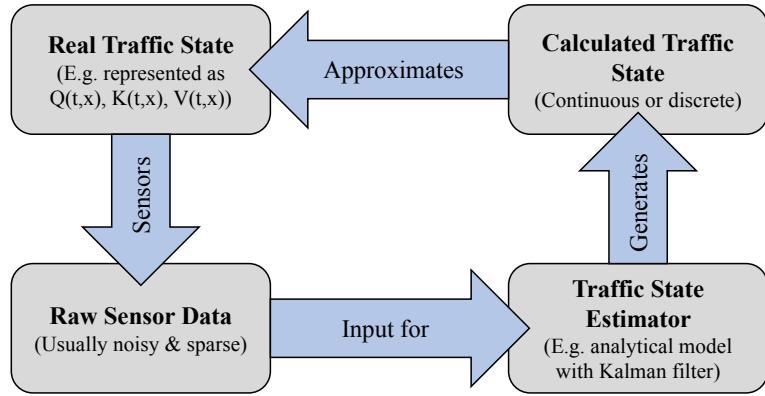


Figure 2.7. The flow of information in a traffic state estimation system

2.3 Traffic State Estimation and Prediction

In this section, first the motivation for traffic state estimation and prediction algorithms is given. Subsequently, some requirements of a practically relevant algorithm are listed. Finally, the state of the art is summarized and the research objectives are put into the overall context.

Assume there is a freeway equipped with inductive loop sensors that report mean traffic flows and speeds each minute. These sensors are distributed along the freeway with a spacing of one kilometer. Thus, a traffic operator or traveler who has access to this data could obtain accurate information about the traffic state at these locations. Further assume, one sensor reports congested traffic and an adjacent, upstream one reports free flow conditions. In this case, it is clear that there is (at least one) transition from free to congested flow in-between the sensors. However, it is unclear which parts of the road segment are truly congested. As a consequence, travelers and traffic operators lack information for decisions on routing and traffic control. Since sensors are prone to noise and outages, the reliability and resulting benefit of raw sensor data suffer even more.

Algorithms that process sensor data in order to provide more accurate and more reliable traffic state information are called traffic state estimation methods (see Figure 2.7). The assumption of most traffic state estimation methods is that traffic on a road can be described as a dynamic (deterministic) system that follows certain rules. Hence, if the rules of the system were perfectly understood, the initial state and the input variables would suffice to deduce the current and future state of the system. In this case, given only few sensor data, accurate traffic state information would be available for each position along the road and for each point in time.

During the last decades great advances in traffic flow theory and modeling (section 2.1) have been accomplished. Though, the perfect traffic flow model does not yet exist. In addition, it is commonly acknowledged that the system embodies stochastic factors. Therefore, some time after model initialization, a model tends to deviate from the real traffic state. Current sensor data is required to update its state.

The preceding introduction focused the estimation of Real-Time Traffic Information (RTTI) given sensor data. There are also many applications that require predictive quantities such as traffic speeds (e.g. in order to determine travel times). Others require traffic state estimations of historical situations. These three types of problems are closely related. The significant difference pertains the target time for which a traffic estimate is determined: In traffic engineering, one usually refers to traffic state *reconstruction* or *off-line* estimation if the target time is in the past. It is called *real-time* or *on-line* traffic state estimation if the target time is the current time and traffic state *prediction* or *forecast* if the target time is in the future⁴. In all of these cases, the same sensor setup might be available. However, for reconstruction problems all measurement data is available at once, while on-line estimation and prediction approaches can access only data that has been collected so far.

For practical large-scale application traffic state estimation algorithms are expected to fulfill the following requirements:

- **Accurate:** Given noisy and sparse data of traffic conditions, the resulting estimate should match the real traffic conditions for all locations on the road and all points in time.
- **Efficient:** The computational resources required to perform the algorithm should be low in order to enable a real-time application in large networks at decent costs.
- **Robust:** According to the IEEE, robustness is "The degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions"⁵. Noisy or erroneous traffic data or in-cautious parametrization should not cause a breakdown of a traffic state estimation method.
- **General:** An algorithm should be able to process heterogeneous types of traffic data, e.g. flow, density or speed measurements with varying accuracy, resolution and spatio-temporal coverage and generate more accurate estimates if more information is available.

⁴There exist several definitions of the terms forecast and prediction that seek to distinguish between their exact meaning. In traffic engineering these terms are often used as synonyms.

⁵1990. IEEE Standard Glossary of Software Engineering Terminology, IEEE Std 610.12-1990 defines robustness

Traffic state estimation and prediction are broad fields of research. In order to give an overview of state-of-the-art approaches, the following four classifications are introduced:

- **Target time:** As mentioned earlier, an algorithm can be designed to target different points in time for which accurate traffic states are determined.
- **Road type:** Traffic dynamics on intersection-free roads (e.g. freeways) and urban roads differ substantially. While the first are modeled as road corridors with high speed-limits, the latter usually have low speed-limits and intersections are controlled by traffic signals. Therefore, in many publications also the applied traffic models for estimation and prediction differ.
- **Data:** Section 2.2 gave an overview of traffic sensors. Since the characteristics of available data impacts the approach to estimate traffic conditions (which is also a motivation for this thesis), published approaches are classified according to the main data type. Most models utilize either inductive loop data and FCD. The remaining ones are summarized as 'other' data sources.
- **Model:** The published approaches are classified with respect to the model type. An analytical model is understood as a set of differential equations describing the dynamics of traffic in space and time. Analytical traffic models usually relate variables such as flow, density and speed. They are discretized in time and space. To find a solution, they are usually integrated numerically over time and assimilated with observed data. The most common type is a CTM based on the LWR equation. Data-driven models omit differential equations. Based on available data, they apply a variety of algorithms in order to estimate and predict traffic states from a combination of historical and current data. Applied algorithms range from simple statistical equations over general Machine Learning (ML) models to dedicated algorithms specialized in traffic state estimation.

Table A.1 provides an overview of current approaches with a focus on traffic speed and/or travel time estimation and prediction. Figures 2.8 and 2.9 depict the number of publications and give one exemplary publication for each category (For a comprehensive overview of methods dedicated to the prediction of other quantities, especially traffic flow, the interested reader is referred to (Vlahogianni et al., 2004; Vlahogianni et al., 2014)). Summarizing the listed and classified publications, some observations can be made:

1. There are significantly more publications dealing with freeway networks than with urban networks.

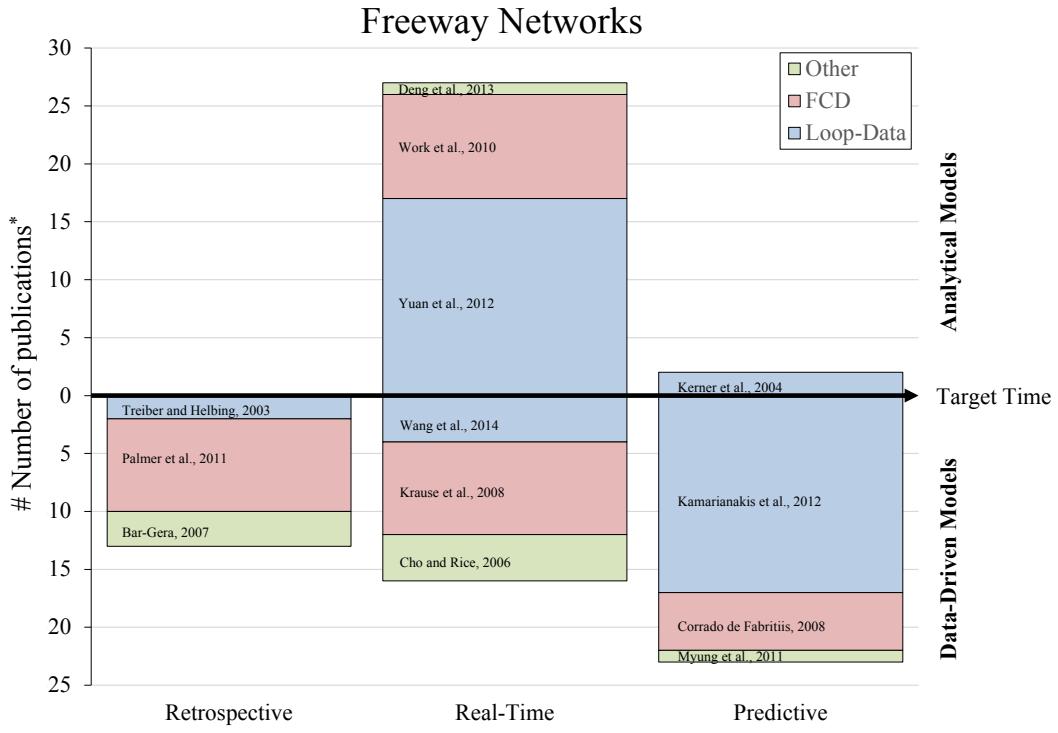


Figure 2.8. Number of publications on traffic speed estimation and prediction on freeway networks. For each category one specific publication is given. (* The number is based on the list in Table A.1)

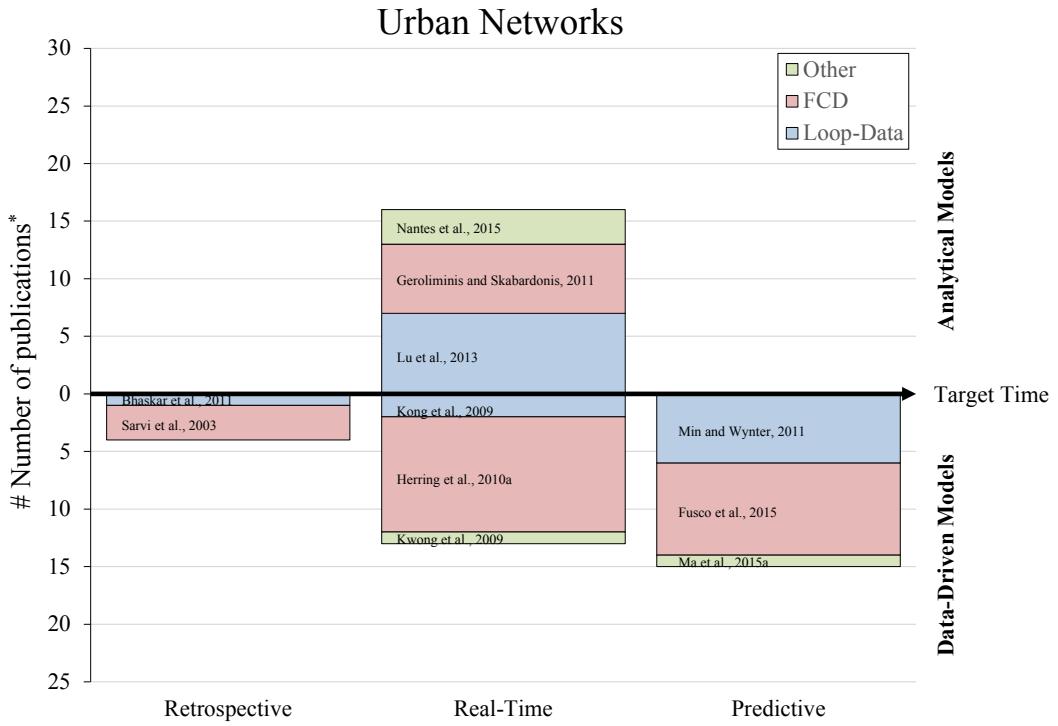


Figure 2.9. Number of publications on traffic speed estimation and prediction on urban networks. For each category one specific publication is given. (* The number is based on the list in Table A.1)

| Research Objective | Target time | Roads | Data | Model |
|-----------------------------|---------------|---------|-----------------|-------------|
| Traffic speed estimation | Retrospective | Freeway | FCD (+ Other) | Data-driven |
| Congestion front forecast | Predictive | Freeway | FCD + Loop-data | Analytical |
| Network congestion analysis | Predictive | Urban | FCD | Data-driven |

Table 2.3. Categorized approaches presented in this thesis

2. For freeway approaches the loop detector is the most common source of data, while for urban networks FCD-based approaches dominate. Other data sources such as Bluetooth trackers, cell-phone data or camera data are relatively rarely utilized.
3. Freeway real-time traffic speed estimators usually develop first or second-order macroscopic traffic models and data assimilation techniques such as Kalman filters. In the past years also some FCD based analytical approaches have been studied (e.g. (Work et al., 2010)). These make use of FDs in order to translate vehicle speeds into densities and, subsequently, apply similar techniques.
4. In urban scenarios the existing approaches can be divided further into two categories: One deals with the accurate estimation of queue lengths at intersections. The typical approach is to couple loop data with FCD. The other category describes approaches that estimate and analyze traffic conditions in entire networks. The main data source for these approaches is FCD.
5. For predictive traffic state estimation most researchers apply data-driven models. Many of the earlier publications study parametric approaches, such as variants of the Autoregressive Integrated Moving Average (ARIMA) model. Later and up to now, more and more non-parametric approaches are applied. These stem from advances in the field of ML, for instance Artificial Neural Networks (ANNs). Among them, a slight trend from univariate to multivariate approaches can be noticed. Unlike univariate models, multivariate ones also consider the influence of neighboring road segments for a traffic state prediction of a certain segment.

The presented results and solutions for traffic speed estimation and prediction described in this thesis can also be categorized according to this scheme (see Table 2.3): The approach described in chapter 4 is designed for the retrospective estimation of freeway traffic speeds using FCD. It integrates findings of the Three-Phase traffic theory into a data-driven approach. Due to its efficiency, it can potentially also be applied in real-time. The evaluation is done using FCD, however, it can be applied to any type of speed data. Based on this approach, chapter 5 describes a model that fuses FCD and loop data in order to provide short-term congestion front forecasts on freeways. An analytical LWR forecast model is applied. Apart from these methods which provide

solutions for freeway traffic, the approach in chapter 6 focuses on urban networks. On top of a general congestion pattern study, a data-driven forecast model for network-wide congestion forecast is developed.

In the following chapters, these approaches are motivated based on a detailed analysis of related works in the respective category. For the evaluation, real FCD collected by a large fleet of vehicles is used. The characteristics of the data and the preprocessing steps are described briefly in the subsequent chapter.

Chapter 3

Data

In this chapter the FCD used throughout this work in order to evaluate estimation and prediction methods is presented. First, the technical implementation is summarized. Second, the so-called map-matching procedure is described, which is necessary in order to link raw GNSS positions with a digital map. Finally, a brief statistical exploration of the data is performed in order to provide an overview of the amount of available data.

3.1 Sampling and Collecting Data

The setup of the applied technology includes a fleet of equipped vehicles, which report GNSS data, and a central server, which collects all data in a database. Each vehicle that is equipped with a RTTI provider has a software system that samples the current GNSS position in intervals of 10 s to 30 s. These positions and according timestamps are stored in the local memory of the vehicle. After sampling a few positions, a filter mechanism decides whether sampled GNSS positions are transmitted to the central server. This filter continuously compares the vehicle's velocity with the velocity given by a traffic provider. If the velocity at one of the sampled data points deviates more than 10% to 30% (depending on the software version of the module) from the provided velocity, the recently sampled positions and according timestamps are transmitted to the central server. In case only small deviations are detected, the recently sampled positions are retained and removed from the local memory. Each transmitted position is linked to an alias that is generated by the vehicle. The alias is random and changes over time. At the server, single transmitted positions of the same alias can be connected in order to reconstruct vehicle trajectories.

The motivation for the implemented filter mechanism and the alias is the protection of drivers' privacy: Since vehicles do not transmit continuously, and hide their vehicle ID, it

is not possible to reconstruct complete trips or infer the driver's identity. Furthermore, since traffic on most roads is usually in free flow conditions, a significant amount of bandwidth is saved. Still, if traffic conditions deviate from the expected conditions, valuable information are gathered and reported to the server.

3.2 Preprocessing

Collected raw data comprises of GNSS positions and respective timestamps. In order to link this data to a road, a process called map-matching needs is applied. A digital map can be represented as a directed graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) \quad (3.1)$$

which is an ordered pair of a (finite) set of vertices \mathcal{V} and a set of edges \mathcal{E} . An edge $e \in \mathcal{E}$ comprises a pair of two vertices $v_1, v_2 \in \mathcal{V}$. In a directed graph, this pair is ordered such that there is a connection from vertex v_1 to v_2 but not necessarily from v_2 to v_1 . In a digital map the vertices usually represent geolocations. Edges represent road segments connecting these locations. The properties of a road segment are represented as attributes of an edge. In this work, the length and regulatory speed-limit of a road segment and the corresponding edge are denominated as $l(e) \in \mathbb{R}_+$ and $V_{Lim}(e) \in \mathbb{R}_+$, respectively.

Given some position and time data, a map-matching algorithm returns a sorted list of edges which are supposed to match the roads that the vehicle generating the data originally passed. In addition to the list of edges, the functions $x_c(t)$ and $v_c(t)$ are calculated, which represent the time-dependent position and velocity of a vehicle on the reconstructed edges.

Map-matching can be challenging: First, the accuracy of GNSS is limited (see section 2.2). Hence, GNSS positions do not match a road exactly, but scatter around its the vehicle's real location. Second, a low sampling rate results in ambiguities of roads that a vehicle might have taken. Third, the reconstruction of a trajectory for which a sequence of positions is available requires extensive computational resources. Several algorithms have been proposed that accomplish map-matching which are summarized in (Quddus et al., 2007). The data used in this work is map-matched based on enhancements of the Multiple Hypothesis Technique (Schuessler and Axhausen, 2009), described in detail in (Heidrich, 2011). It constructs a set of hypotheses of possible trajectories, performs a mutual comparison of all hypotheses and finally selects the most probable one.

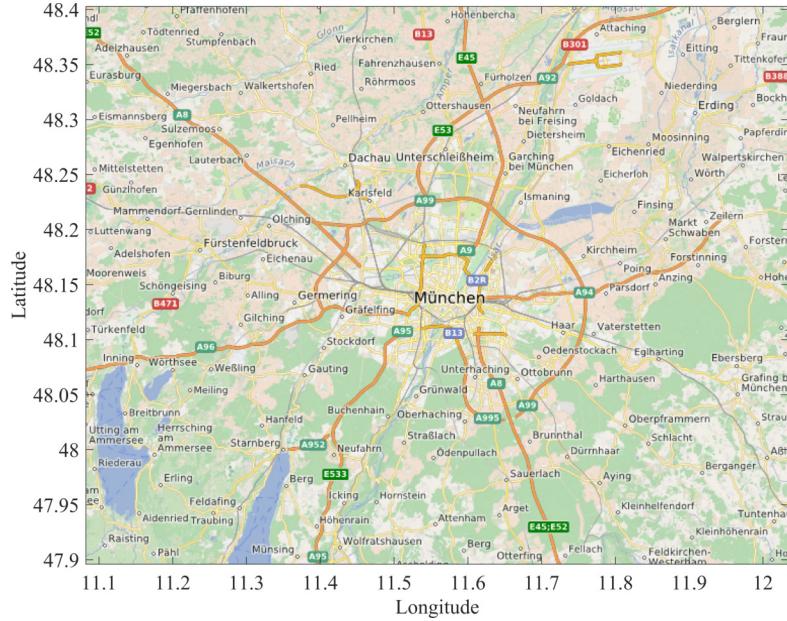


Figure 3.1. Map of Munich and its surrounding¹

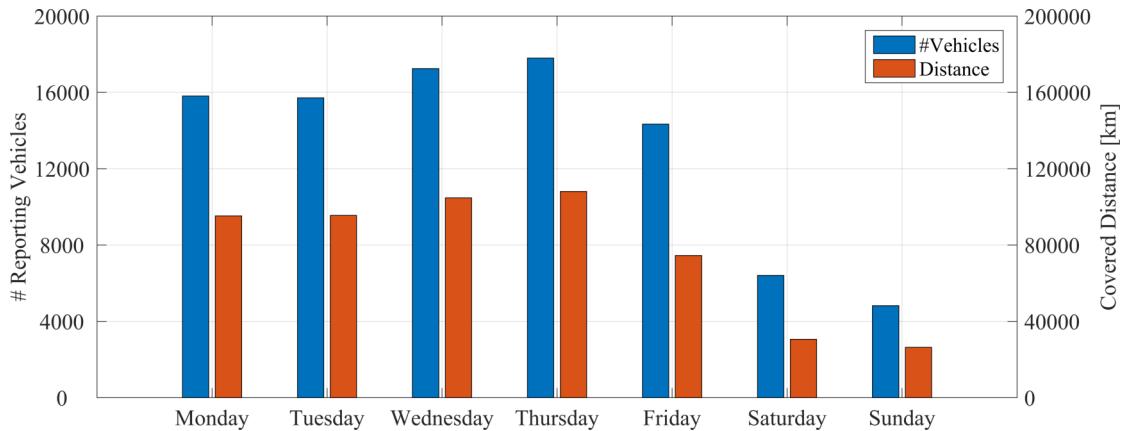


Figure 3.2. Average number of reporting vehicles and average distance covered by the fleet with respect to the day of the week (in 2015)

3.3 Statistical Exploration of Available Data

According to the traffic provider TomTom, Munich was one of the most congested cities in Germany in 2016². Therefore, this region (compare Figure 3.1) suits well for the development and evaluation of algorithms that target traffic congestion. The according digital map consists of 147,108 uni-directional edges spanning a network of 17,219 km. Out of these, 3,186 km comprise major roads comparable to the types freeway and arterials.

¹Map data provided by Open Street Map (OSM)

²http://www.tomtom.com/en_gb/trafficindex/city/munich

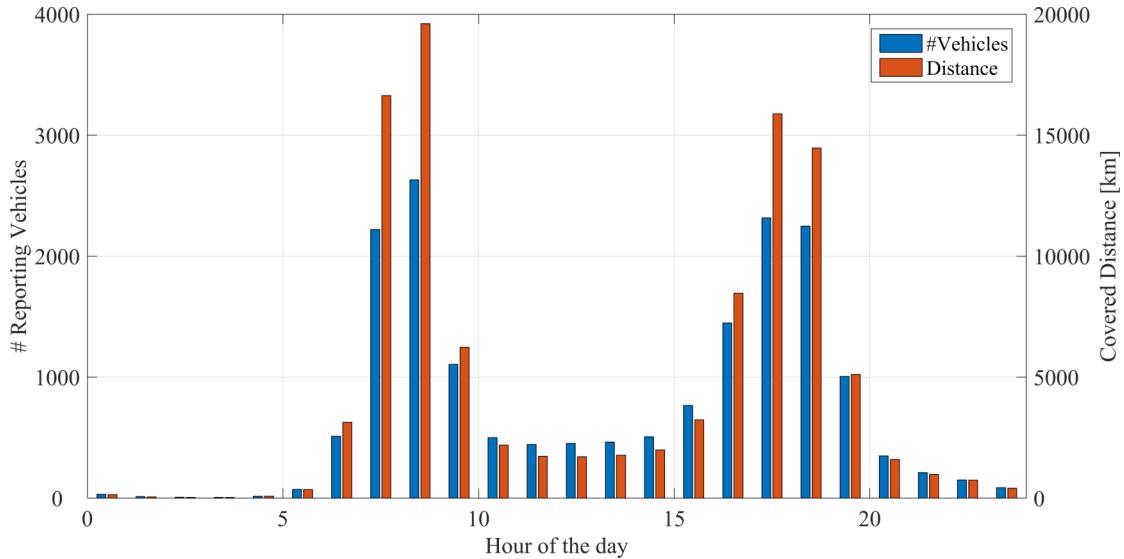


Figure 3.3. Average number of reporting vehicles and average distance covered by the fleet with respect to the hour of the day (Tuesdays-Thursdays in 2015)

Figure 3.2 depicts the average number of vehicles reporting GNSS data to a server with respect to the weekday for the year 2015. In addition, the covered distance of the reporting fleet per day is given. With about 16,000 reporting vehicles, traversing altogether 90,000 km on an average weekday, a great amount of data is available. In comparison to datasets that include vast amount of traces in free flow conditions, the present dataset constitutes a filtered set that describes mostly congested traffic situations. As expected, due to commonly fewer vehicles and less congestion, the amount of data on weekends decreases significantly.

Figure 3.3 illustrates the number of vehicles and total traversed distance with respect to the daytime of a usual working day (Tuesday - Thursday). The bars clearly show the morning and evening peak during which significant parts of the road network are congested. At night barely any data is received. One interesting observation is that the ratio of distance and number of vehicles is higher during peak times than during off-peak times. It means that during peak hours an average vehicle travels larger distances while reporting data than during off-peak times.

Finally, Figure 3.4 illustrates exemplary the average traffic velocity at different times of the day in a three-dimensional plot for the working days in 2015. The average traffic velocity on the major roads in Munich are depicted on the z-axis of the plot. This representation allows to explore the regions and times which are mostly affected by congestion.

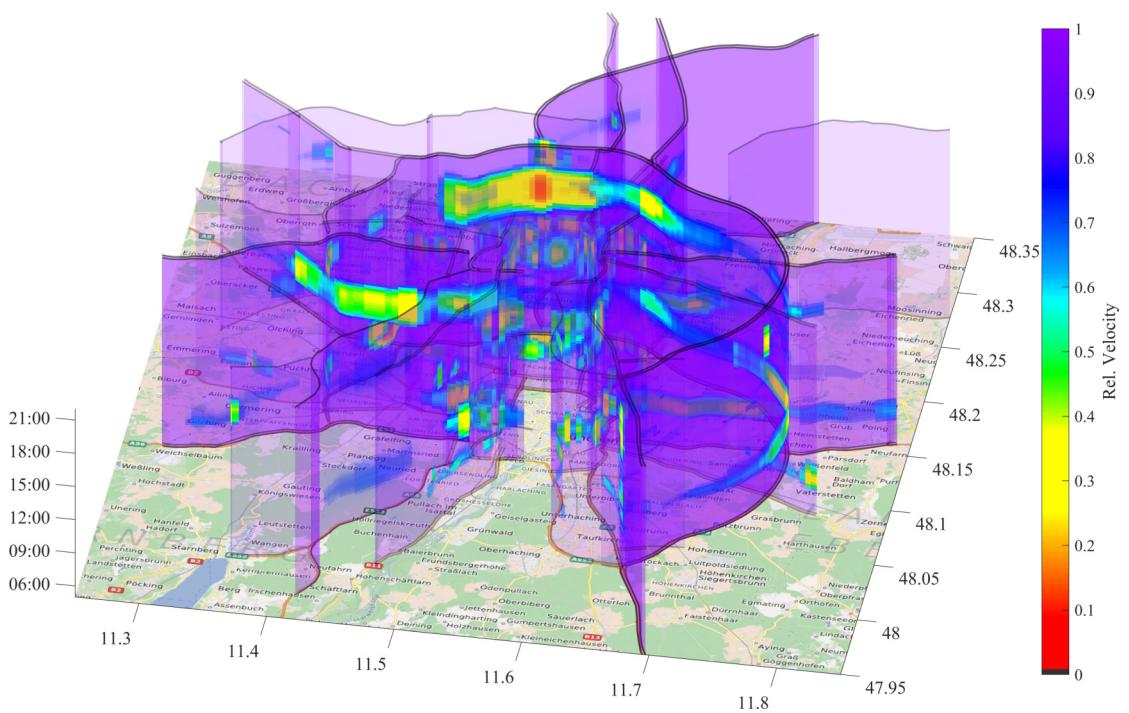


Figure 3.4. Average traffic velocity as 3D plot on major roads of the Munich road network

Chapter 4

Phase-based Traffic Speed Estimation with FCD

In this chapter a novel approach to traffic speed estimation utilizing FCD is presented. The chapter is structured in the following way: First, the motivation and challenges using FCD in traffic speed estimation are given. Second, related work is described and evaluated with respect to their ability to handle FCD as the main data source. Motivated by the strengths and weaknesses of existing approaches a solution approach based on the Three-Phase theory is developed. The evaluation with respect to its accuracy, efficiency and robustness is presented in the section 4.4. A final discussion and conclusion summarizes the results, discusses open issues and proposes future directions.

4.1 Motivation

To exploit the full potential of FCD algorithms are required that deal with the challenges that accompany this new source of data and its characteristics. Traffic estimation methods furthermore need to fulfill the requirements of an accurate, efficient, robust and general traffic estimator (section 2.3) in order to have practical relevance. Compared to conventional fixed sensors such as loop detectors the use of FCD faces the following challenges:

- **Under-determined traffic state:** In contrast to loop detectors providing at least flow and speed of vehicles, which allow to determine the macroscopic traffic state, current Floating Car (FC) technology collects only speed-related data

such as instant velocities or travel times. Consequently, the traffic state is under-determined. Without further assumptions it is not possible to apply analytical traffic models relating flow, density and speed.

- **Sparsity of information:** While loop detectors provide data at pre-defined positions and in pre-defined time intervals, the availability of FCD depends on several factors. Among them is the number of equipped vehicles passing an observed road during an observed time interval. As a consequence, there are times and places for which high densities of FCD are available, and those, for which barely any data has been reported.
- **Samples instead of averages:** By definition, the macroscopic traffic speed is the average speed of all vehicles in a region in space-time ((Treiber and Kesting, 2013) present several ways how to define the macroscopic speed in space-time). Loop detectors and other spot sensors usually determine average traffic quantities at one location from all passing vehicles. FCD stems from individual vehicles and provides information about the traffic state in space and time. Reported speeds can be interpreted as random samples drawn from the distribution of all vehicle speeds in a space-time region. In congested traffic conditions, where speeds among different lanes are synchronized, this distribution is narrow, such that a sample will deviate only slightly from the mean of the distribution. In contrast, in free flow conditions where vehicles can overtake each other, their velocities may vary greatly. To summarize, FCD needs to be interpreted differently than data collected by conventional sensors such as loop detectors.
- **Inaccuracy of lane positioning:** On roads without merging or diverging lanes the traffic state is usually similar among all lanes. The reason is that drivers expecting a personal benefit will usually change from a congested to a free lane and equilibrate the traffic state. However, if e.g. a great part of drivers desires to leave a freeway on the same off-ramp, the traffic state on the diverging lanes might be congested while the main lanes are free. The accuracy of common positioning systems such as GNSS does not always suffice to determine the exact lane a vehicle is driving on. Therefore, using FCD, it is challenging to distinguish heterogeneous traffic conditions on several lanes.

4.2 Related Work

As outlined in section 2.3 the method to be presented here targets retrospective traffic speed estimation on freeways. Most published approaches for this type of problem can be

classified into two categories. The first category comprises analytical flow models coupled with data assimilation techniques. First-order models are usually based on the LWR model (Lighthill and Whitham, 1955; Richards, 1956) and use Kalman filters in order to match model expectation and observation (van Lint and Djukic, 2014). One of the first approaches is published in (Szeto and Gazis, 1972), where the authors assimilate data with a conservation model on a short freeway stretch to estimate traffic density. More recently, in (Suzuki et al., 2003) a Kalman filter is applied in order to estimate traffic conditions based on a mix of loop detector and FCD. Also, results published in (Yuan et al., 2012) present the benefits of a Lagrangian model compared to Eulerian approaches when applied to detector and FCD. In addition, higher order models have been proposed that account for more sophisticated traffic dynamics in (Aw and Rascle, 2000). First approaches are described in (Cremer and Papageorgiou, 1981) who adapt the Payne's model in order to estimate the traffic state on a freeway more accurately. (Wang and Papageorgiou, 2005) describe a second-order model that estimates traffic conditions on a freeway in real-time. Computational issues are addressed by (van Hinsbergen et al., 2012), where the authors demonstrate the development of a localized filter that performs real-time computations.

Though, all of these models rely strongly on flow and density data collected by loop detectors. In contrast, (Herrera et al., 2010; Work et al., 2009; Work et al., 2010; Work et al., 2008) focus on probe data and develop models using a fundamental diagram that estimate densities from probe velocities. Furthermore, (Bekiaris-Liberis et al., 2016) propose a macroscopic model for traffic density estimation using a linear parameter-varying system that relies mainly on probe velocity measurements. Although in the latter mentioned approaches most of the information is obtained from probe data, the proposed models still require flow or density measurements at the boundaries. In practical applications, the need for additional flow or density information drastically limits the applicability of an approach on a large scale since it adds further effort and complexity to data acquisition.

The second category of algorithms comprises of estimation methods that are based on empirical traffic theory. First, in (Kerner, 2004) a model called ASDA/FOTO is introduced. The model assimilates flow, density and velocity data reported by loop detectors with the findings of the Three-Phase traffic theory and provides current and predictive traffic information (Kerner et al., 2004; Kerner, 2009; Kerner, 2017). Therefore, it reconstructs spatio-temporal regions of free flow, synchronized flow and WMJs and tracks phase fronts. While that approach is completely based on loop detector data, (Palmer, 2011; Palmer et al., 2011) study the reconstruction of phase regions with trajectory data exclusively. The phase transitions in space-time of individual vehicles are identified by

means of velocity and time conditions and aggregated into phase objects using a clustering approach. The advantages of this approach include the exclusive use of FCD in order to estimate velocities and phases. Nonetheless, velocities inside a traffic phase are estimated to a constant over space and time, which in turn limits the accuracy of an estimation. Additionally, the trajectories without phase transitions are discarded, which subsequently results in a loss of valuable information. For example, a trajectory in free flow state that passes through an estimated synchronized flow region will not influence the phase estimate since it does not contain phase transitions. Another well-known method is the Generalized Adaptive Smoothing Method (GASM). It is based on the observation that shock waves in congested traffic propagate upstream and shock waves in free traffic propagate downstream (Treiber and Helbing, 2003; Treiber et al., 2010b; Treiber and Kesting, 2013). Using two characteristic convolution processes, traffic data is smoothed and aggregated adaptively. The advantages of the GASM are that it can be applied to velocity data of different sources (Treiber et al., 2010b; van Lint and Hoogendoorn, 2009), that it allows for an efficient implementation (Schreiter et al., 2010) and that it proved to be significantly more accurate than isotropic smoothing (Treiber and Helbing, 2003; Rempe et al., 2016b; van Lint, 2010). On the other hand, it tends to propagate low velocities up- and downstream unconditionally although they might be part of stationary congestion upstream a bottleneck (Treiber et al., 2010b). Thus, when it is applied to sparse probe data the estimated velocities in stationary congestion patterns lack accuracy.

4.3 Solution Approach

Inspired by the GASM and the Three-Phase traffic theory, the present algorithm called Phase-based Smoothing Method (PSM) is developed (published as (Rempe et al., 2017)). The key idea is to divide the estimation process into two steps: The first step aims at the identification of phases in space and time. This is done in accordance to typical characteristics of phases: synchronized flow phases usually stick to bottlenecks and have a stationary character. WMJs are shock waves in traffic with low average vehicle speeds and an approximately constant downstream front speed (Kerner et al., 2004; Kerner, 2009). Their width in time is limited to several minutes and, when full developed, they have been observed to travel upstream for tens of kilometers (Kerner and Rehborn, 1996). In the second step raw data is assigned to the identified phases and traffic speeds for each position and point in time are computed separately for each phase. By assigning traffic data to identified phase regions, it is assured that velocity measurements of one phase do not influence the velocity estimation in adjacent traffic phases. Furthermore, given the phase regions in time and space in the first step, in the second step the estimation

quality can be refined using a local estimator. Thus, effects such as narrow moving jams that emerge in synchronized flow can be reconstructed more accurately (Kerner, 2009).

In order to identify the three traffic phases in space and time, a related approach described in (Palmer et al., 2011) applies a clustering technique that connects estimated phase transitions from individual vehicles. The results are phase regions with sharp phase fronts. However, there are several aspects of this approach that provide potential for further improvements. One is that, although most of the time traffic can be classified clearly into one of the three phases, there exist transitions which take some time to take effect (see section 2.1). E.g. in the pinch region of a congestion pattern narrow moving jams inside a synchronized flow phase may develop into full WMJs over time while propagating upstream (Kerner, 2009). Accordingly, there is no clear front between the synchronized traffic phase and the WMJ front. Another aspect concerns the characteristics that real traffic data such as FCD accompany (see chapter 4). Since data consists of noisy measurements that may deviate from the macroscopic traffic speed, it is important to consider average vehicle speeds and not individual speeds. Otherwise, a single outlier may influence the outcome of the reconstruction vastly. Moreover, data is sparse in space and time. In regions with few data, the reliability of a traffic speed estimate is decreased. Depending on the application the reliability of an estimate may be crucial. For example, a traffic control measure based on wrong traffic information might lead to a significant loss of traffic efficiency or safety. Published approaches with deterministic phase regions such as (Palmer et al., 2011; Kerner et al., 2004) do not provide information about the reliability of the estimated phase regions.

Due to these reasons, the presented approach models phase probabilities instead of deterministic phase regions. Let $\Omega = \{F, S, J\}$ be a set of the phases free flow (F), synchronized flow (S) and WMJ (J). Then, $P_p(t, x) \in [0, 1]$ with $p \in \Omega$ denotes the probability that the traffic state at time $t \in [T_0, T_1]$ and at position $x \in [0, L]$ is in phase p . This probabilistic model allows to consider the aforementioned issues: A narrow moving jam that develops into a WMJ can be handled as a smooth transitions of phase probabilities between the S and the J phase. Furthermore, if only few (potentially noisy) measurements are available, the belief for a specific phase estimate can be expressed with a phase probability below one. This belief can also be interpreted as the quality of an estimate which is a valuable information for practitioners.

Figure 4.1 summarizes the work flow for processing raw trajectory data into a continuous velocity estimate $V_E(t, x)$. As first step, raw data is convolved with different phase-characteristic smoothing kernels. Several fuzzy phase criteria are defined and applied to resulting values. Respective criteria probabilities P_p^i denote their degree of fulfillment. Next, for each phase, the criteria probabilities are aggregated into preliminary phase

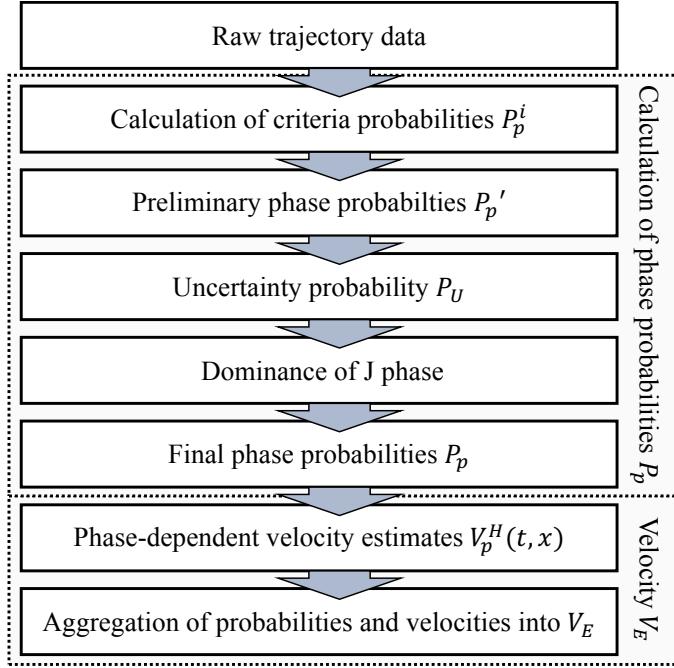


Figure 4.1. Flow diagram of steps taken in order to process raw FCD into a continuous velocity estimate V_E

probabilities P'_p . The probability $P_U \in [0, 1]$ is estimated to establish the level of uncertainty in assigning (t, x) to any of the phases. The relations between phases are modeled in the following step, which results in the final phase probabilities P_p . Based on these and raw trajectory data, for each phase and each point in space-time, a velocity estimate $V_p^H(t, x)$ is computed. Additionally, a fall-back velocity $V_U(t, x)$ is assumed that serves as a best-guess velocity in case the uncertainty $P_U(t, x)$ is high. Finally, the resulting estimate $V_E(t, x)$ is determined by aggregating the probabilities $P_p(t, x)$, $P_U(t, x)$ and their respective velocity estimates V_p^H and V_U .

In the following sections each taken step will be explained in detail. Before doing so, two preliminary concepts are introduced that are fundamental for the PSM. The first is the representation of trajectory data in time and space. The second is the concept of continuous convolution that is applied multiple times for smoothing purposes.

4.3.1 Representation of FCD in Space-Time

Goal of this section is to develop a general concept for the representation of velocity data collected by individual vehicles that matches best the characteristics of the data while serving as an input for further processing steps, i.e. especially the continuous convolution process. Further benefits of a general concept are discussed in section 4.4.2.

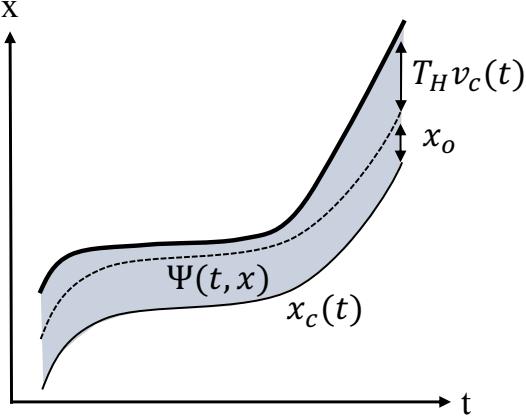


Figure 4.2. Illustration of the space-time region $\Psi(t, x)$ of a vehicle with velocity $v_c(t)$ and length x_0

The trajectory of a vehicle $c = 1, \dots, N_c$ can be described as a function $x_c(t) \in [0, L]$ denoting the position of that vehicle along a road segment with length $L \in \mathbb{R}_+$. Accordingly, vehicle velocity $v_c(t)$ denotes the derivative of $x_c(t)$. Each vehicle that passes through space-time domain $[T_0, T_1] \times [0, L]$ of a road with length L , observed for time period $T_1 - T_0$, provides partial information about the domain. In order to model the space-times for which information is available, a simple car-following model with parameters vehicle length and time headway is assumed. x_0 denotes the length of the vehicle c and the minimal distance to the preceding vehicle in queuing traffic. T_H denotes the time headway to a preceding vehicle. Therefore, at time t , vehicle c occupies the space interval $[x_c(t), x_c(t) + x_0 + T_H v_c(t)]$ (see Fig. 4.2). Occupying a region in space-time means that in the interval only one vehicle can exist and, furthermore, it is assumed that in this interval, the vehicle's velocity is a representation of the traffic velocity. Note that this holds only for single-lane roads. Let $\Psi(t, x)$ be a function that indicates whether time t and position x is occupied by any observed vehicle:

$$\Psi(t, x) = \begin{cases} 1 & \text{if } \exists c : x_c(t) < x < x_c(t) + x_0 + T_H v_c(t) \\ 0 & \text{otherwise.} \end{cases} \quad (4.1)$$

Ψ describes a binary system which has the following properties: If $\Psi(t, x) = 1 \forall t \in [T_0, T_1], x \in [0, L]$ then traffic density is high, all vehicles maintain a time headway of T_H to their preceding vehicles and all vehicle positions as well as velocities are known. Otherwise, there are space-times for which no velocity information is available. The reason can be that traffic density is low such that there are gaps between vehicles, or for a part of the vehicles no position and velocity data is given.

Accordingly, $V_{FCD} \in \mathbb{R}_+$ denotes the velocity information reported by all vehicles,

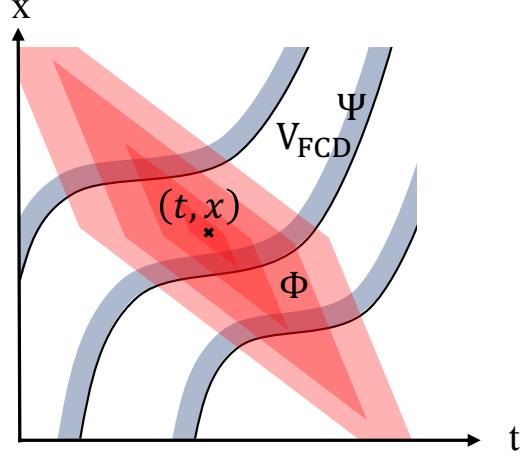


Figure 4.3. Illustration of the convolution process of V_{FCD} and Ψ with kernel Φ

combined into a single two-dimensional function. This velocity is only valid in space-time (t,x) which is occupied by any vehicle c^* , i.e. $\Psi(t,x) = 1$, and is set to the velocity of the closest vehicle upstream of (t,x) :

$$V_{FCD}(t,x) = \begin{cases} v_{c^*} & \text{if } \Psi(t,x) = 1, c^* = \arg \min_{\forall c: x - x_c(t) \geq 0} (x - x_c(t)) \\ 0 & \text{otherwise.} \end{cases} \quad (4.2)$$

4.3.2 Traffic-Motivated Data Smoothing

Due to the sparsity and noise in real data smoothing and interpolation is a necessary step in order to filter outliers and fill gaps between samples. Filtering data in two dimensions with a filter function constitutes a convolution process. The function

$$\Gamma_V(w, \Phi, t, x) = \int_{T_0}^{T_1} \int_0^L \Phi(t - \hat{t}, x - \hat{x}) \cdot w(\hat{t}, \hat{x}) \cdot V(\hat{t}, \hat{x}) \cdot \Psi(\hat{t}, \hat{x}) d\hat{x} d\hat{t} \quad (4.3)$$

represents a weighted continuous convolution. $\Phi(t,x) \in \mathbb{R}$ denotes the kernel function that is applied to the input data $V(t,x)$. $w(t,x) \in \mathbb{R}$ is a space-time dependent weight of this input data. Definitions of traffic-motivated kernel functions are given in eq. (4.7). In order to use the convolution equation for smoothing operations, the results of Γ need to be normalized. The normalization term $D(t,x)$ is similar to the aforementioned convolution but omits velocity input V :

$$D(w, \Phi, t, x) = \int_{T_0}^{T_1} \int_0^L \Phi(t - \hat{t}, x - \hat{x}) \cdot w(\hat{t}, \hat{x}) \cdot \Psi(\hat{t}, \hat{x}) d\hat{x} d\hat{t}. \quad (4.4)$$

Furthermore, the kernel Φ and weighting w are chosen in such a way that D represents an estimate of the local data density. The local density quantifies the amount of information

available for the velocity estimation in (t, x) . The normalized convolution of weighted velocity function $V(t, x)$ and according occupation Ψ with kernel Φ is given by:

$$\Lambda_V(w, \Phi, t, x) = \frac{\Gamma_V(w, \Phi, t, x)}{D(w, \Phi, t, x)}. \quad (4.5)$$

Given for example a kernel function that returns values greater or equal to zero with its maximum in $(0, 0)$ and whose values are monotonically decreasing with increasing distance to the origin in space-time, eq. (4.5) describes a common smoothing process (see Figure 4.3). Then, for space-time (t, x) a weighted average velocity of all nearby velocities valid in their respective occupied regions $\Psi(t, x)$ is computed. The weights depend on the kernel function $\Phi(t, x)$, distance to (t, x) , sampled data as well as the weighting function $w(t, x)$.

Generally, modifying the applied kernel Φ enables a great variety of operations that are often used in the fields of computational image processing (Shapiro and Stockman, 2001). For traffic speed estimation the process of convolution is mainly used for smoothing operations. The basic desired properties of a smoothing kernel are the following:

1. Measurements that are closer in time and space are supposed to influence the final result stronger than more distant measurements.
2. The influence of the distance in time and space shall be controllable using parameters.

(Treiber and Helbing, 2003) propose the following simple and effective smoothing kernel:

$$\Phi_{v^{dir}, \tau, \sigma}(t, x) = \exp\left(-\left|\frac{t - \frac{x}{v^{dir}}}{\tau}\right| - \left|\frac{x}{\sigma}\right|\right). \quad (4.6)$$

Its maximal value is located at $(0, 0)$ and values decrease exponentially with increasing distance to the origin, controlled by parameters τ and σ . Another feature is the subtraction of $\frac{x}{v^{dir}}$ in the nominator of the first term. This relation between space and time results in an anisotropy which grants that one of the main kernel axis (compare Figure 4.3) is rotated into the direction of v^{dir} . Consequently, measurements that are located in direction of v^{dir} relative to (t, x) are weighted stronger. Applied to smoothing of traffic data this preference direction accounts for the propagation of shock waves in traffic dynamics (see (Treiber and Helbing, 2003; van Lint and Hoogendoorn, 2009) for more details).

During the following operations this kernel is applied multiple times. For completeness, the definition is extended in case the shock wave velocity is zero:

$$\Phi_{v^{dir}, \tau, \sigma}(t, x) = \begin{cases} \exp\left(-\left|\frac{t}{\tau}\right| - \left|\frac{x}{\sigma}\right|\right) & \text{if } v^{dir} = 0 \\ \exp\left(-\left|\frac{t - \frac{x}{v^{dir}}}{\tau}\right| - \left|\frac{x}{\sigma}\right|\right) & \text{otherwise.} \end{cases} \quad (4.7)$$

Notice that eq. (4.5) implies that the speeds of those vehicles which occupy more space-time (either because they are longer, or they drive at higher speeds) are weighted stronger than the speeds of those which occupy less space-time. This pertains the interpretation of the resulting macroscopic traffic speed. As explained thoroughly in (Treiber and Kesting, 2013), there exist several ways to aggregate individual vehicle speeds over time and space such as the (arithmetic or harmonic) time mean space, the space mean speed or other methods such as Edie's (Edie, 1963). The average that the PSM resembles has similarities with all of these definitions, but differs in one important aspect: Individual vehicles are converted into homogeneous regions in space-time for which a velocity is available. During that process the information whether a region was occupied by one or several vehicles may get lost: the region may emerge from several short vehicles following each other with time headways T_H , or one long truck. Consequently, the smoothed resulting velocity does not correspond to the mean vehicle speed averaged over all reported vehicles but to the mean speed averaged over all occupied regions in space-time. In that sense, the presented way to average is just another (slightly different) way to define a macroscopic traffic speed. It is motivated by the representation of vehicle speed's with validities Ψ , which can be seen as an approach to advance from a macroscopic to a mesoscopic data representation in time and space.

4.3.3 Modeling Phase Probabilities

This section describes the assignment of space-time (t, x) to the three traffic phases free flow, synchronized flow and WMJ. If an assignment is vague, (t, x) can also be in uncertain state U .

As described in section 2.1 each traffic phase has different characteristics. These empirical characteristics can be used to identify the most likely traffic phase p in space-time (t, x) . Let $P_p^1(t, x), P_p^2(t, x), \dots, P_p^{N_k}(t, x), P_p^i(t, x) \in [0, 1]$ be a number of $N_k \in \mathbb{N} \setminus \{0\}$ criteria that (t, x) needs to fulfill in order to belong to phase p . Each criterion is modeled as a fuzzy decider P_p^i . The combination of several fuzzy decider can be done by applying fuzzy logic. E.g. the 'AND' relation of two fuzzy variables $a, b \in [0, 1]$ is the product of both values: $a \text{ AND } b = ab$. The logical 'OR' is defined as: $a \text{ OR } b = 1 - (1 - a)(1 - b) =$

$a + b - ab$ (Zadeh, 1965). In this case, all criteria are expected to be fulfilled in order to assign (t, x) to one of the phases, which corresponds to an 'AND' relation. Effectively, the independent and preliminary phase probability P'_p is determined as the product of all fuzzy decider:

$$P'_p(t, x) = \prod_i^{N_k} P_p^i(t, x). \quad (4.8)$$

Consequently, P'_p is always lower or equal to the lowest criteria probability P_p^i . In applications where several phase criteria are proposed it might be desired to accumulate criteria of the less rigid probabilities such that the failure of one or more criteria will be tolerated. This would require eq. (4.8) to be adapted. An extension with further logical expressions, e.g. 'OR', 'XOR', 'NOT' would allow to define more complex rules. Other possibilities constitute methods that are applied frequently in related problems which require the fusion of different classifiers into a final decision. Potential candidates are 'opinion pools' (Jacobs, 1995), where each classifier contributes to the final decision depending on its trustworthiness or the Dempster-Shafer theory (Shafer, 1976), which is commonly applied in sensor fusion where estimations of heterogeneous sensors are combined.

For now, two classes of criteria are presented and later applied. The first is the velocity criterion P_p^1 that uses velocity information of smoothed data for determining the phase probability. The second, density criterion P_p^2 , is applied ensuring that a phase hypothesis is supported by nearby data. Table 4.1 gives an overview of all probabilities and will be explained thoroughly in the following sections.

Note that the previous description does not adhere to classical probability axioms since, in this context, it causes a contradiction. E.g. if one considers a hypothesis H with a probability of $P(H)$ that it is true, then the hypothesis being false will have probability $1 - P(H)$. Now, assume for each phase p there is at least one fuzzy decider P_p^i that is zero. This in turn causes all preliminary phase probabilities P'_p to be also zero. In a classical interpretation of probabilities it would follow that traffic does not appear in one of the phases. This naturally contradicts the theory model since traffic must always be in one of the three phases. Therefore, with respect to the evidence theory (Shafer, 1976), the modeled phase probabilities P'_p need to be interpreted as independent beliefs that a phase hypothesis is correct. Thus, each probability represents an estimate of a probability. On the other hand, the probability $1 - P'_p$ needs to be interpreted as the probability that traffic is in any one of the three phases. The higher $1 - P'_p$, the more uncertain is a classification as phase p . That principle of uncertainty is important to fuse data and provide a quality estimate of reconstructed velocities. As a consequence, probabilities in Table 4.1 do not sum up to one.

| | Free Flow | Sync. Flow | WMJ |
|-----------------------------------|--|----------------------------|----------------------------|
| Velocity criterion P_p^1 | P_F^1 | P_S^1 | P_J^1 |
| Density criterion P_p^2 | P_F^2 | P_S^2 | P_J^2 |
| Prelim. Phase Prob. P'_p | $P'_F = P_F^1 \cdot P_F^2$ | $P'_S = P_S^1 \cdot P_S^2$ | $P'_J = P_J^1 \cdot P_J^2$ |
| Final Phase Prob. P_p | $P_F = P'_F(1 - P'_J)$ | $P_S = P'_S(1 - P'_J)$ | $P_J = P'_J$ |
| Uncertainty P_U | $P_U = (1 - P'_F) \cdot (1 - P'_S) \cdot (1 - P'_J)$ | | |
| Quality Q | $Q = 1 - P_U$ | | |

Table 4.1. Overview of probabilities computed during phase identification using a velocity and a density criterion

The probabilistic approach allows a simple extension with further criteria that improve the accuracy in distinguishing between phases. Section 4.3.3.3 proposes further criteria that can be integrated into the PSM given other potentially heterogeneous data sources.

4.3.3.1 Velocity Criterion

Traffic velocity is an essential information in order to determine the traffic phase in (t, x) (Kerner et al., 2004; Kerner, 2009; Kerner et al., 2013; Palmer, 2011). Modern traffic theories state that traffic breakdown, which is the transition from free to congested flow, is a probabilistic event triggered by perturbations (Schoenhofer and Helbing, 2007; Kerner, 2009). A traffic breakdown is usually connected to a capacity drop (Schoenhofer and Helbing, 2007; Laval, 2007; Treiber and Kesting, 2013) and a significant drop in average vehicle velocities (Kerner et al., 2004; Schoenhofer and Helbing, 2007) (see section 2.1). Due to this drop, traffic velocity is a good feature to distinguish between free flow and congested flow. Here, the fuzzy thresholds v_F^{thres} and v_S^{thres} are applied differentiating between these two flow regimes. The distinction between WMJs and synchronized flow, which are both congested states, is less obvious. In fact, the upper velocity of the WMJ is significantly lower than the velocity threshold that separates synchronized and free flow (Kerner et al., 2004). v_J^{thres} denotes this threshold. Also, velocities in this particular phase can decrease down to 0 km/h, therefore in such regions, no lower velocity bound is required. Lastly, the lower velocity bound of the synchronized flow phase is discussed. Kerner suggests that low velocities should be assigned to the WMJ phase.

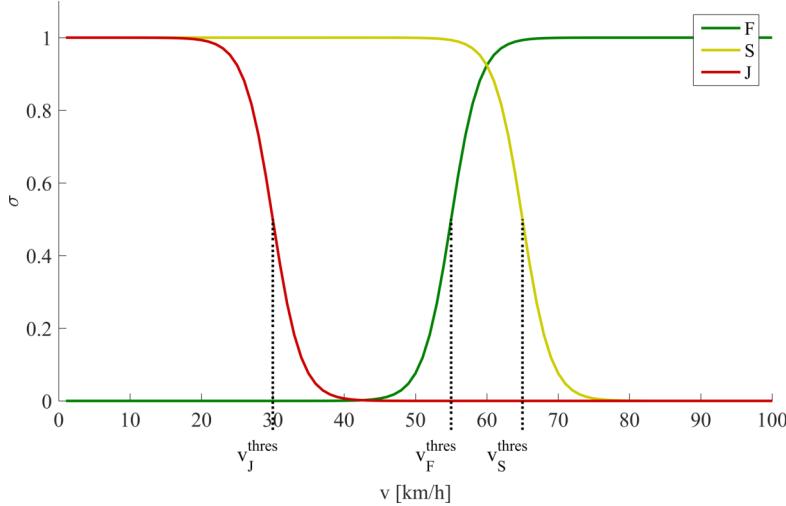


Figure 4.4. Velocity criterion for each phase with respect to traffic velocity

He argues that, with stronger bottlenecks, the type of congestion pattern changes from synchronized pattern, over general pattern to mega-jams (Kerner et al., 2004; Kerner, 2009). According to his definition, mega-jams are WMJs with a great width, which have a very low average speed. In two-phase theories these patterns are usually called Homogeneous Congested Traffic (HCT) (Schoenhofer and Helbing, 2007). In contrast to the ASDA/FOTO model (Kerner et al., 2004) which applies a lower velocity threshold for the synchronized flow, in this approach no threshold is set. Instead, a dominance of the WMJ phase is modeled which assures that in the presence of low velocities the probability of a WMJ is increased (see section 4.3.3.5). This turned out to be more robust than the definition of a lower threshold.

As fuzzy decider function $\sigma_v(v, v^{thres}, \lambda)$ (standard sigmoid function) is applied that translates a velocity v into probability $P_p^1(t, x) \in [0, 1]$. The parameter v^{thres} constitutes the inflection point of the curve, parameter λ the strictness of the threshold, i.e. it controls the gradient of the transition region (the higher λ the higher the gradient):

$$\sigma_v(v, v^{thres}, \lambda) = 1 - \frac{1}{1 + \exp(-\lambda(v - v^{thres}))}. \quad (4.9)$$

Function 4.9 is close to one for low velocities and converges towards zero for velocities that exceed the applied threshold. This is the desired behavior for the probability estimation in congested flow. For free flow, the complementary function is applied:

$$\sigma_v(v, v^{thres}, \lambda) = \frac{1}{1 + \exp(-\lambda(v - v^{thres}))}. \quad (4.10)$$

Figure 4.4 illustrates exemplary the sigmoid functions for the three phases, using the thresholds $v_F^{thres} = 55$ km/h, $v_S^{thres} = 65$ km/h and $v_J^{thres} = 30$ km/h and a common

strictness of $\lambda = 0.5 \text{ h/km}$.

Phases Free and Synchronized Flow Due to the sparsity of data in space and time it is not meaningful to apply the velocity criterion to raw data directly. Instead, smoothed velocities are considered. In order to do so, the previously described two-dimensional smoothing process (section 4.3.2) is applied.

Probabilities $P_F^1(t, x)$ and $P_S^1(t, x)$ are computed as:

$$P_F^1(t, x) = 1 - \sigma_v(V_F(t, x), v_F^{thres}, \lambda_F) \quad (4.11)$$

$$P_S^1(t, x) = 1 - \sigma_v(V_S(t, x), v_S^{thres}, \lambda_S) \quad (4.12)$$

where v_F^{thres} and v_S^{thres} denote the parameters of the decider function with respect to the characteristic velocity of phase p . The velocity estimates V_F and V_S are computed according to the normalized convolution process (eq. (4.5)) as:

$$V_F(t, x) = \Lambda_{V_{FC}D}(w^0, \Phi_F, t, x) \quad (4.13)$$

$$V_S(t, x) = \Lambda_{V_{FC}D}(w^0, \Phi_S, t, x) \quad (4.14)$$

with $\Phi_F(t, x)$ and $\Phi_S(t, x)$ denoting phase specific smoothing kernels with threshold velocities v_p^{thres} and parameters τ_p and σ_p . For the construction of these kernels the parameters v_p^{dir} , τ_p and σ_p are applied (eq. (4.7)):

$$\Phi_F := \Phi_{v_F^{dir}, \tau_F, \sigma_F} \quad (4.15)$$

$$\Phi_S := \Phi_{v_S^{dir}, \tau_S, \sigma_S}. \quad (4.16)$$

At this stage, the data is weighted equally using the weight function $w^0(t, x) = 1 \forall (t, x) \in [T_0, T_1] \times [0, L]$. The standard weighting implies that all smoothed raw data has an equal significance for the determination of the phases.

Phase WMJ The velocity criterion for the J is more distinct. It is postulated that (t, x) can only be assigned to the J phase if, both, up- and downstream of (t, x) low velocities are observed. This ensures that WMJs are not extrapolated far beyond measurements in order to be able to reduce wrongly estimated congested regions. The trick in order to compute this condition efficiently, is to check for measurements up- and downstream of (t, x) smoothing data with differing kernels and to calculate the probabilities independently. Then, the product of both probabilities represents the need to

fulfill both requirements. $P_J^1(t, x)$ is computed as as:

$$P_J^1(t, x) = \sigma_v \left(V_J^d(t, x), v_J^{thres}, \lambda_J \right) \cdot \sigma_v \left(V_J^u(t, x), v_S^{thres}, \lambda_J \right) \quad (4.17)$$

where $V_J^d(t, x)$ and $V_J^u(t, x)$ denote the velocity estimates computed by smoothing data with characteristic kernels $\Phi_J^d(t, x)$ and $\Phi_J^u(t, x)$:

$$V_J^d(t, x) = \Lambda_{V_{FCD}}(w^0, \Phi_J^d, t, x) \quad (4.18)$$

$$V_J^u(t, x) = \Lambda_{V_{FCD}}(w^0, \Phi_J^u, t, x). \quad (4.19)$$

Respective kernels functions are defined as :

$$\Phi_J^d(t, x) = \begin{cases} \Phi_J(t, x) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (4.20)$$

$$\Phi_J^u(t, x) = \begin{cases} \Phi_J(t, x) & \text{if } x \leq 0 \\ 0 & \text{otherwise} \end{cases} \quad (4.21)$$

$$\Phi_J := \Phi_{v_J^{dir}, \tau_J, \sigma_J}. \quad (4.22)$$

Φ_J^u is shaped in such a way that it considers upstream data of (t, x) , Φ_J^d only considers downstream data. In that way, the data is smoothed in different directions. The combination of the sigmoid functions of the velocities V_J^d and V_J^u ensures that space-time (t, x) is only estimated as a WMJ if upstream and downstream data supports the phase hypothesis. Note that eq. (4.17) uses different velocity thresholds v_J^{thres} and v_S^{thres} as parameters for the sigmoid function. The difference stems from the expectation that a WMJ, which is detected by downstream velocities below v_J^{thres} , will propagate upstream as long as the upstream traffic is in a state of critical flow-density (Kerner, 2004). Since no density or flow data is available, this state is assumed to be the congested region with the velocity threshold v_S^{thres} .

By requiring both criteria to be fulfilled it is assured that the J is only reconstructed for low velocity measurements but never extrapolated. It is quite possible that the moving jam emerged earlier than the time from when the first equipped vehicle perceived it and propagated further upstream than the last equipped vehicle passing through the WMJ. Unfortunately, sparse data does not allow one to recognize exactly when a WMJ emerged and when it dissolved. Extrapolating a shock wave upstream or downstream means to risk overestimating. Thus, this approach can be described as the cautious way aimed at minimizing wrongly estimated low velocities.

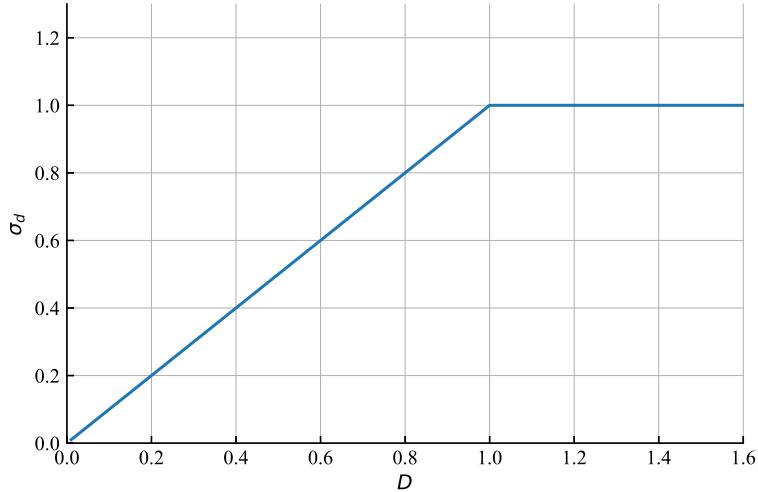


Figure 4.5. Density criterion to translate data density D into phase probability P_p^2

4.3.3.2 Density Criterion

The second criterion, called density criterion $P_p^2(t, x)$ uses data density D (eq. (4.4)) in order to quantify how well a phase hypothesis is supported by nearby data. The necessity of this criterion stems from the varying density the comes along with FCD. Data density $D(w^0, \Phi_p, t, x)$ is computed for each phase p using the respective kernel Φ_p . Since weighting w^0 and the applied kernels are greater than zero, $D(w^0, \Phi_p, t, x)$ is always greater or equal to zero. Note that D is not normalized such that its values often exceed the value one. In order to translate density into the probabilities P_p^2 , its values are converted into phase probabilities by applying conversion function $\sigma_d \in [0, 1]$:

$$P_p^2(t, x) = \sigma_d(D(w^0, \Phi_p, t, x)). \quad (4.23)$$

As a simple variant, the minimum function is chosen here:

$$\sigma_d(x) = \min(1, x). \quad (4.24)$$

This forces the density criteria P_p^2 to equal one if data is nearby, and otherwise converge to zero when the distance between (t, x) and the measurements grows larger (see Figure 4.5). The validity of a measurement in time and space can be parametrized by adapting the kernel function or by modifying the weighting w^0 . Note that the applied minimum operator is one way to do the translation, which is chosen due to its simplicity. Other operators with smoother properties, e.g. sigmoid functions, may also be applied.

4.3.3.3 Further Applicable Criteria

This section provides exemplary two further criteria that may be adopted in specific applications where other types of data are available. This list is not complete and may be extended.

Traffic Density Criterion There are several types of sensors that are able to measure the macroscopic traffic density (section 2.2). Traffic density is a valuable information in order to distinguish between free flow and congested flow. It is clear that low traffic densities indicate free flow states while high traffic densities indicate congested flow. A criterion that could be added to the PSM may model such information using a sigmoid function:

$$P_F^*(t, x) = \sigma_v(k(t, x), k^{crit} + \Delta k, \lambda_k) \quad (4.25)$$

$$P_{S,J}^*(t, x) = 1 - \sigma_v(k(t, x), k^{crit} - \Delta k, \lambda_k) \quad (4.26)$$

where $k(t, x)$ is the prevailing traffic density and k^{crit} the threshold between free and congested flow, as well as a shift of Δk . Classical traffic flow theories assume a triangular fundamental diagram where threshold k^{crit} is the density with maximal flow (see section 2.1). Recent flow theories recognize that there exists a range of traffic densities where traffic can be both, in free and in congested flow (Kerner et al., 2004). If in free flow, this traffic condition is usually unstable and may lead to a spontaneous traffic breakdown. The resulting congested traffic maintains a similar traffic density, though, a lower velocity. Modeling this criterion as a fuzzy decider with overlapping probabilities accounts well for this characteristic.

While currently mostly stationary sensors provide density data, which limits the large-scale application of a traffic estimator, the equipment of vehicles with powerful sensors is advancing. As a result, future FCs will provide not only speed information, but also distance measurements to nearby vehicles as well as their speeds. First promising results are described in (Seo and Kusakabe, 2015), where the spacing information of an Adaptive Cruise Control (ACC) system is compared to loop detector data. Therefore, this type data will play an increasing role in the future.

Car-following Criterion Current vehicle technology includes a great variety of Advanced Driver Assistance Systems (ADASs). One of them is the ACC which measures position and speed differences to a preceding vehicle and adapts the vehicle's speed according to these differences in order to provide comfort and safety to the driver. In traffic flow theory, many microscopic driver models differentiate between the state in

which the driver adapts his speed to one or more preceding vehicles and the state in which they can drive freely (explicitly, or by evaluating the distance to a preceding vehicle) (Kerner and Klenov, 2010; Treiber et al., 2000; Gipps, 1981; Kendziorra et al., 2016). These microscopic models are usually tuned to reconstruct empirically observed macroscopic congestion patterns (Kesting and Treiber, 2008; Brockfeld et al., 2004; Kerner and Klenov, 2010). Consequently, evaluating the state of the ACC system provides information about prevailing traffic states. In accordance to the microscopic model of the Three-Phase theory (see section 2.1), a simple criterion could be that traffic can only be in congested state if a vehicle's time gap is bounded by a preceding vehicle.

4.3.3.4 Probability of Uncertain State

After computing the aforementioned probabilities P_p^i as the product of all phase criteria, uncertainty $P_U(t, x)$ is determined. P_U models the degree of uncertainty in assigning (t, x) to any of the phases:

$$P_U(t, x) = \prod_{p \in \Omega} (1 - P'_p(t, x)). \quad (4.27)$$

In case given data does not allow a classification into one of the phases P_U is high. Consequently also the reliability or trustworthiness of a phase or subsequent velocity estimation is lower than for space-times where the reliability is high. That allows to inherently define a quality estimate enhancing the interpretation of resulting velocities. Probability Q is defined as the complementary probability to P_U :

$$Q(t, x) = 1 - P_U(t, x). \quad (4.28)$$

4.3.3.5 Phase Interactions

Up until now, the independent phase probabilities P'_p have been determined. Since P'_p are computed independently with differing smoothing kernels, particular region characteristics could occur, especially in the presence of shock waves, where more than one of the probabilities P'_F , P'_S or P'_J is estimated to a high value. This is due to the different shapes of the convolution kernels and according velocities that are considered for the determination of the phase.

According to the Three-Phase theory WMJs can propagate through other phases without interruption (Kerner et al., 2004). Therefore, it is reasonable to assume that space-time regions with high probabilities of a J phase and another phase rather belong to the J. This dominance of the J phase over the other phases is modeled (comparable to the

GASM where smoothed low velocities are given a higher weight than high velocities (Treiber and Helbing, 2003)). Final phase probabilities P_p are set as:

$$P_J(t, x) = P'_J(t, x) \quad (4.29)$$

$$P_S(t, x) = P'_S(t, x) \cdot (1 - P_J(t, x)) \quad (4.30)$$

$$P_F(t, x) = P'_F(t, x) \cdot (1 - P_J(t, x)). \quad (4.31)$$

4.3.4 Estimating Phase Velocities

Up to this step, raw data has been used in order to reconstruct the traffic phases in time and space using characteristic smoothing kernels and several criteria that allow one to distinguish between them. Phase information are helpful in order to understand traffic conditions, but most applications require velocity estimates. Though, for estimating traffic speeds it is advantageous to know about the location of these phases. They provide the information about the space and time for which velocity measurements are valid. For example, assume a low velocity measurements that is estimated as part of a WMJ. In this case, the velocity measurement can be used to estimate the average traffic velocity of the WMJ. However, it does not provide information about the traffic velocity in an adjacent free or synchronized flow phase.

In order to estimate traffic speeds, raw data is smoothed a second time for each phase. This time, the estimated phase probabilities P_p are applied as weights in order to assign each measurement a validity:

$$V_p(t, x) = \Lambda_{V_{FCD}}(P_p, \Phi_p, t, x). \quad (4.32)$$

Before aggregating the phase velocities $V_p(t, x)$ into a final traffic speed, two ideas are discussed that enhance the accuracy of the estimate.

First, in eq. (4.32) the same smoothing kernels Φ_p as in the phase estimation process are applied. Those kernels are designed to model the characteristic propagation of phase fronts. For speed estimation, the characteristic propagation velocities are different. In free flow, shock waves propagate downstream with approximately v^{free} (Treiber and Kesting, 2013; Kerner et al., 2004) as can be derived intuitively from a triangular FD (van Lint and Hoogendoorn, 2009). Therefore, it is reasonable to construct the smoothing kernel Φ'_F section 4.3.2 using that velocity. The shock wave propagation in congested flow is v^{cong} (Treiber and Kesting, 2013; Kerner et al., 2004). Note that in Three-Phase traffic theory the FD such that the reasoning for a characteristic shock wave velocity in congested flow is different: Inside the synchronized flow phase so-called narrow moving jams can emerge spontaneously that can develop into WMJs (section 2.1). Similar to

WMJ fronts, these moving jams propagate upstream with a speed of approximately v^{cong} . For the reconstruction of traffic dynamics inside the congested phases respective kernels Φ'_S and Φ'_J with a shock wave velocity of v^{cong} are applied.

The second concept addresses the smoothing process in eq. (4.32) itself. It resembles an arithmetic mean of raw data weighted with phase probabilities and a kernel function. Applying the arithmetic mean means to interpret all values equally important for the description of the center of a data set (Triola, 2014). In most cases that is the desired result. Nonetheless, traffic speed estimates are mainly required for determining accurate travel times. For instance, assume a road with two intervals of lengths $s_1 = 500\text{ m}$ and $s_2 = 500\text{ m}$ and speed-limits $v_1 = 20\text{ km/h}$ and $v_2 = 50\text{ km/h}$. Then, the total travel time would be $t = s_1/v_1 + s_2/v_2 = 90\text{ s} + 36\text{ s} = 126\text{ s}$. The desired average velocity v_{avg} for the entire road would be the one that fulfills $t = (s_1 + s_2)/v_{avg}$. The arithmetic mean of v_1 and v_2 underestimates the required travel time. The harmonic mean

$$v_{avg}^H = \frac{n}{\sum_i^n \frac{1}{v_i}} \quad (4.33)$$

accounts for this bias and preserves the travel times. In general, the harmonic mean "is often used as a measure of center for data sets consisting of rates of change, such as speeds" (Triola, 2014). The original proposal of the GASM also faced this bias. Therefore, (van Lint, 2010) proposed the application of a harmonic smoothing process in order to reproduce traffic speed estimations that allow more accurate reconstructions of travel times. Since the PSM is based on similar smoothing processes the previous arithmetic mean is replaced by a harmonic mean, indicated by the superscript 'H':

$$\frac{1}{V_p^H(t, x)} = \Lambda_{V_{FCD}^{-1}}(P_p, \Phi_p^H, t, x) \quad (4.34)$$

where V_{FCD}^{-1} denote the reciprocal velocities:

$$V_{FCD}^{-1}(t, x) = \frac{1}{V_{FCD}(t, x)} \quad (4.35)$$

Note that real data may contain velocity measurements of 0 km/h which would result in a division by zero. Therefore, a lower velocity limit of 3 km/h is set.

4.3.5 Aggregating Probabilities and Velocities

At this stage all phase probabilities P_p and respective velocities V_p^H are determined. Additionally, P_U is computed and a best-guess velocity V_U is assumed. In order to aggregate all information into a final traffic speed estimate V_E a weighted average is

applied that fuses the phase velocity estimates weighted by the degree of belief into the respective phase:

$$V_E = \frac{P_F V_F^H + P_S V_S^H + P_J V_J^H + P_U V_U}{P_F + P_S + P_J + P_U}. \quad (4.36)$$

The weighted arithmetic average is adopted due to its simplicity. Other variants of aggregation such as maximal values or harmonic average etc. may also be applied here. Though, since most of the time the phase assignment is unambiguous and one phase probabilities exceeds the other ones significantly, a specialized aggregation method is not expected to influence the estimation accuracy noticeably.

4.4 Evaluation

In this chapter the results of an extensive evaluation of the PSM with real data are presented. Specifically, it is analyzed whether the PSM fulfills the four most important requirements of a traffic estimation algorithm: accuracy, efficiency, robustness and generality (see section 2.3).

The evaluation is structured in the following way: First, the parameters of the PSM method are motivated and set with respect to related methods. Second, the discretization of the method in space and time is explained. Third, using the data of an exemplary congestion on German freeway A99 the PSM and two state-of-the-art algorithms are applied and compared qualitatively. In a subsequent quantitative comparison the accuracy of the PSM with respect to the other two methods is assessed. Therefore, all congestion patterns that occurred in the freeway network around Munich during July 7th, 2014 and August 8th, 2014 are considered. One special congestion pattern called 'mega-jam' is afterwards analyzed in detail. A sensitivity analysis in section 4.4.6 identifies the most relevant parameters with respect to the estimation accuracy. In a subsequent run-time analysis the efficiency of the PSM is analyzed. The other two requirements, robustness and generality, constitute qualitative requirements and are included in the conclusion and outlook given in section 4.5.

4.4.1 Setting Parameters

The PSM defines several parameters that have to be set. This section gives an overview of involved parameters and motivates a reasonable setup with respect to traffic theory and related approaches.

Table 4.2 lists all parameters, which are discussed one by one in the following paragraphs.

| | F | S | J |
|---------------|----------|----------|----------|
| v^{thres} | 55 km/h | 65 km/h | 30 km/h |
| λ | | 0.5 | |
| v_p^{dir} | 0 km/h | | -18 km/h |
| τ_p | 250 s | | 30 s |
| σ_p | 150 m | | 500 m |
| $v_p^{dir,H}$ | 70 km/h | | -18 km/h |
| τ_p^H | 100 s | | 30 s |
| σ_p^H | 100 m | | 200 m |

Table 4.2. Parameters of the PSM

The setting of velocity thresholds v_p^{thres} between phases is done according to (Kerner, 2004; Kerner et al., 2013; Palmer et al., 2011). Strictnesses λ_p are set to 0.5 h/km such that the resulting sigmoid function σ_v drops from a value of 92.5% to 7.5% in an interval of ± 10 km/h, where v^{thres} denotes the turning point of the function (compare Figure 4.4).

The kernels Φ_p applied for phase estimation involve τ_p, σ_p and v_p^{dir} . v_p^{dir} can be approximated well based on empirical traffic characteristics. As described in section 2.1, the downstream phase fronts of WMJs propagate upstream with an almost constant velocity of -18 km/h. The downstream phase fronts of synchronized flow phases typically stick to bottlenecks, which corresponds to a velocity of 0 km/h (Kerner et al., 2004; Palmer et al., 2011). Since the downstream front of a synchronized flow phase is the upstream front of a free flow phase, that propagation speed is approximated similarly. Note that not all phase fronts follow propagate with the assumed velocities. Rather, the assignment of a fixed propagation velocity is a heuristic which works well for common congestion patterns (compare (Treiber and Helbing, 2003; van Lint and Hoogendoorn, 2009)).

Parameters τ and σ influence the decay of the kernel function in time and space. The greater the value the lower the absolute gradient of the kernel. In effect, more distant measurements influence the estimation of a velocity estimate at space-time (t, x) . In related work from which the kernel definition has been adopted, different parameter sets have been applied. (Treiber and Helbing, 2003) set $(\sigma, \tau) := (1.1 \text{ min}, 600 \text{ m})$ and in a later publication propose to use half of the distance of the detectors in time and space

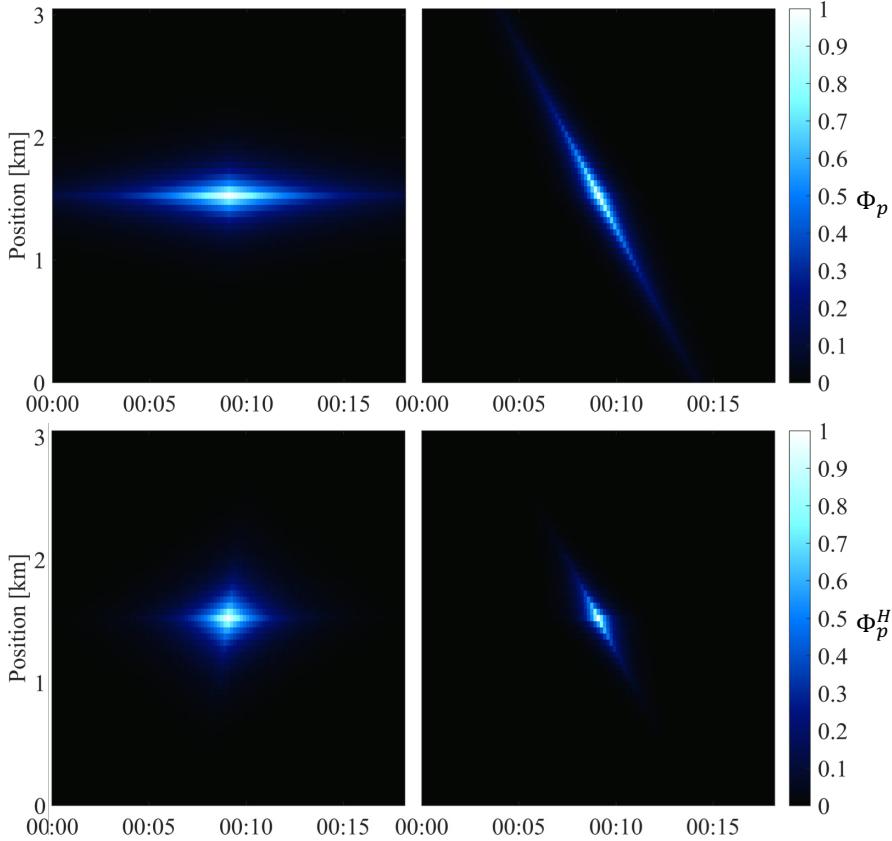


Figure 4.6. Visualization of resulting kernel functions $\Phi_{S,F}$ and Φ_J (up) as well as Φ_F^H and $\Phi_{S,J}^H$ (bottom)

(Treiber et al., 2010b). (Schoenhofer and Helbing, 2007) apply similar values: $(\sigma, \tau) := (1.2 \text{ min}, 600 \text{ m})$. (van Lint and Hoogendoorn, 2009) use $(\sigma, \tau) := (0.5 \text{ min}, 300 \text{ m})$ and acknowledge the need of further work on parametrization of the GASM. (Rempe et al., 2016b) study a range of parameter sets in order to estimate traffic speeds from sparse FCD in an online system. In this case, the kernels Φ_p are parametrized in similar ranges of values but with respect to typically observed properties of traffic phases: WMJs often have a high spatial extent but their width in time is limited (Kerner and Rehborn, 1996). Therefore, σ_J is chosen relatively high and τ_J relatively low. Due to the stationary character of the synchronized flow phase and the adjacent free flow phase, $\tau_{S,F}$ is set significantly greater than τ_J and $\sigma_{S,F}$ significantly lower than σ_J .

For traffic speed estimation, three different kernels Φ_p^H need to be parametrized. Since WMJs and synchronized flow phases belong to the congested traffic phases its parameter sets are merged. The first parameter, $v_p^{dir,H}$, corresponds to the propagation speed of shock waves in traffic. These are well-understood phenomena (Richards, 1956; Newell, 1993; Mika et al., 1969). In free flow shock waves propagate downstream (Treiber and Helbing, 2003; Kerner et al., 2004). Approximations range from 70 km/h (Treiber et al.,

2010b) to 80 km/h (van Lint and Hoogendoorn, 2009; Schoenhofer and Helbing, 2007; Treiber and Helbing, 2003). In congested flow (synchronized flow or WMJ phases) shock waves propagate upstream (Kerner et al., 2004). For the propagation velocity of congested velocities GASM applications apply values that range from -15 km/h (Treiber and Helbing, 2003; Treiber et al., 2010b; Schoenhofer and Helbing, 2007) to -25 km/h (van Lint and Hoogendoorn, 2009). Other empirical studies find propagation values of disturbances in congested traffic of about -15 km/h (Newell, 1993; Kerner, 2009). To conclude, velocities of disturbances in free flow and congested traffic are similar in many publications, which allows to set $v_p^{dir,H}$ on a strong background..

Shock waves in congested traffic phases have relatively short temporal widths and possibly shorter spatial extent than fully developed WMJs. Therefore, $\tau_{S,J}^H$ is set similar to τ_J and $\sigma_{S,J}^H$ is set smaller than σ_J . The kernel parameters τ_F^H and σ_F^H are set to average values compared to all other parameters. Varying these two parameters appeared to influence the estimation accuracy only slightly (see section section 4.4.6). Figure 4.6 illustrates the resulting four kernels with the parameters summarized in Table 4.2.

Two other parameters that stem from the assumed car-following model in section 4.3.1 are time headway T_H and average vehicle length x_0 . They are assumed to be constant for all vehicles. With respect to (Brackstone et al., 2002; Krbalek et al., 2001) the time headway is set to one second and vehicle length to 6 m.

4.4.2 Implementation

In order to apply the PSM to real data, the most simple and efficient way is to discretize time and space into homogeneous segments. The time interval $[T_0, T_1]$ is discretized into N_t intervals of duration $\Delta T = 10$ s and the road of length L into N_x segments of length $\Delta X = 50$ m. The resulting domain can be represented as a matrix of quantities such as velocities $v^{i,j}$ where $i = 1, \dots, N_t$ and $j = 1, \dots, N_x$. Due to the regular structure the discretized domain is denominated as *grid*; one element of the grid as a *grid cell*.

In order to inscribe the trajectory data of vehicle c into the grid, reported GNSS positions and respective timestamps of one vehicle (see chapter 3) are interpolated linearly (see Figure 4.7). Effectively, $x_c(t)$ is a piecewise linear function and $v_c(t)$ a piecewise constant function. Next, the average velocities of the vehicle while in grid cell i, j are computed as the quotient of driven distance inside the cell (max. ΔX) and time inside the cell (max. ΔT). These average velocities are used to compute the occupation Ψ_c of the vehicle (see section 4.3.1). For each valid point in time of the trajectory, the vehicle occupation overlaps with one or more cells. For instance, due to the time headway T_H a fast vehicle occupies more space than a queuing vehicle. Thus, due to larger gaps between fast

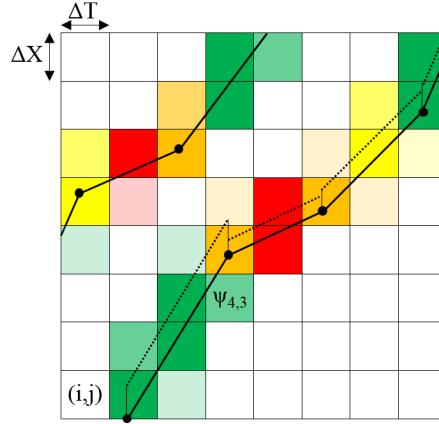


Figure 4.7. Discretization of trajectories and assignment of velocities to grid cells; dotted parallelograms describe the occupation of a trajectory resulting in the cell-wise occupation ψ . Transparent colors indicate low occupation values.

vehicles a vehicle may occupy several cells at the same. In dense and slow traffic, a vehicle may occupy only part of a cell. It follows that each trajectory tr is represented as a set of tuples $tr = \{(i, j, v, \psi)_k\}$ where i, j refers to a grid cell in which the vehicle had velocity v . $\psi \in [0, 1]$ is the cell-wise occupation defined as the ratio between the occupied part of a cell by a vehicle and the cell size. ψ can be interpreted as an estimate of the measurement quality: If a vehicle barely occupies a cell then its velocity is less representative for traffic velocity in the respective cell. If a vehicle occupies an entire cell, then it is probably the only vehicle in this cell its reported velocity equals to the traffic speed in the cell.

The definition of the cell-wise occupation provides two benefits. It allows one to decouple the data from the chosen grid size. If the grid cells are larger in time and/or space, passing vehicles occupy smaller parts of a cell and the weighting ψ decreases. Similarly, if for instance ΔX is smaller than the occupied space of a vehicle, the vehicle occupies more than one cell at the same time. Thus, convolution operations (see section 4.3.2) return the same results irrespective of the chosen grid resolution. Second, this concept allows to place other data sources into the grid, to assign these measurements a level of reliability and to fuse them with FCD. For instance, velocity data obtained from a camera observing a road segment can be mapped to a spatial interval instead of just to one point in space as it is usually done. Since all vehicles and their speeds are observed continuously, the cell-wise occupation would resemble the density of traffic and implicitly quantify the high reliability of the velocity data. In another scenario where GNSS data with large sampling times are given, the cell-wise occupation can be used to model the uncertainty of the speed data: One way to do so is to determine all physically possible trajectories that a real vehicle could have taken in order to travel between two GNSS positions. From the superposition of all trajectories a position density and speed

density function can be derived. The average speed values can be inscribed into the grid cells, while the position density is used as the cell-wise occupation quantifying the reliability of a speed measurement. In this way, the uncertainty that results from large sampling times is considered for fusing heterogeneous data. In order to account for the characteristic of each type of data, further concepts need to be developed in future work.

Besides the trajectories, the continuous 2D convolution (eq. (4.3)) needs to be discretized. A straight-forward implementation using loop-structures is rather in-efficient as shown in (Schreiter et al., 2010). Instead, an implementation using the Fourier-transform is proposed. The idea is to apply the convolution theorem, which states that the convolution of two functions f and g is equivalent to the point-wise product of the Fourier transforms $\mathcal{F}\{f\}$ and $\mathcal{F}\{g\}$:

$$\mathcal{F}\{f \circledast g\} = \mathcal{F}\{f\} \cdot \mathcal{F}\{g\} \quad (4.37)$$

This requires to apply the forward and backwards transform of the involved matrices. The Fast Fourier Transform (FFT) (Cooley and Tukey, 1965) is an algorithm that achieves to perform the transforms very efficiently. (Schreiter et al., 2010) show that an implementation using the FFT allows for significant speed-ups. A run-time analysis of the PSM utilizing the FFT is conducted in section 4.4.7. Note that

- the FFT requires the number of elements in each dimension equals 2^n with $n \in \mathbb{N}$ and
- due to a limited floating point accuracy that current computer systems use, this approach is prone to numerical errors. Especially if a matrix contains both, very large and very small (non-zero) numbers (absolute values), the forward and backward transform likely results in large absolute errors. A simple and effective strategy is to pre-process the matrices and set very small (absolute) values to zero.

Figures 4.8 and 4.9 visualize the flow of data to process discretized raw data into phase probabilities, and subsequently into a velocity estimate and an according quality matrix. For reasons of clarity the Fourier-transforms are not depicted in the chart.

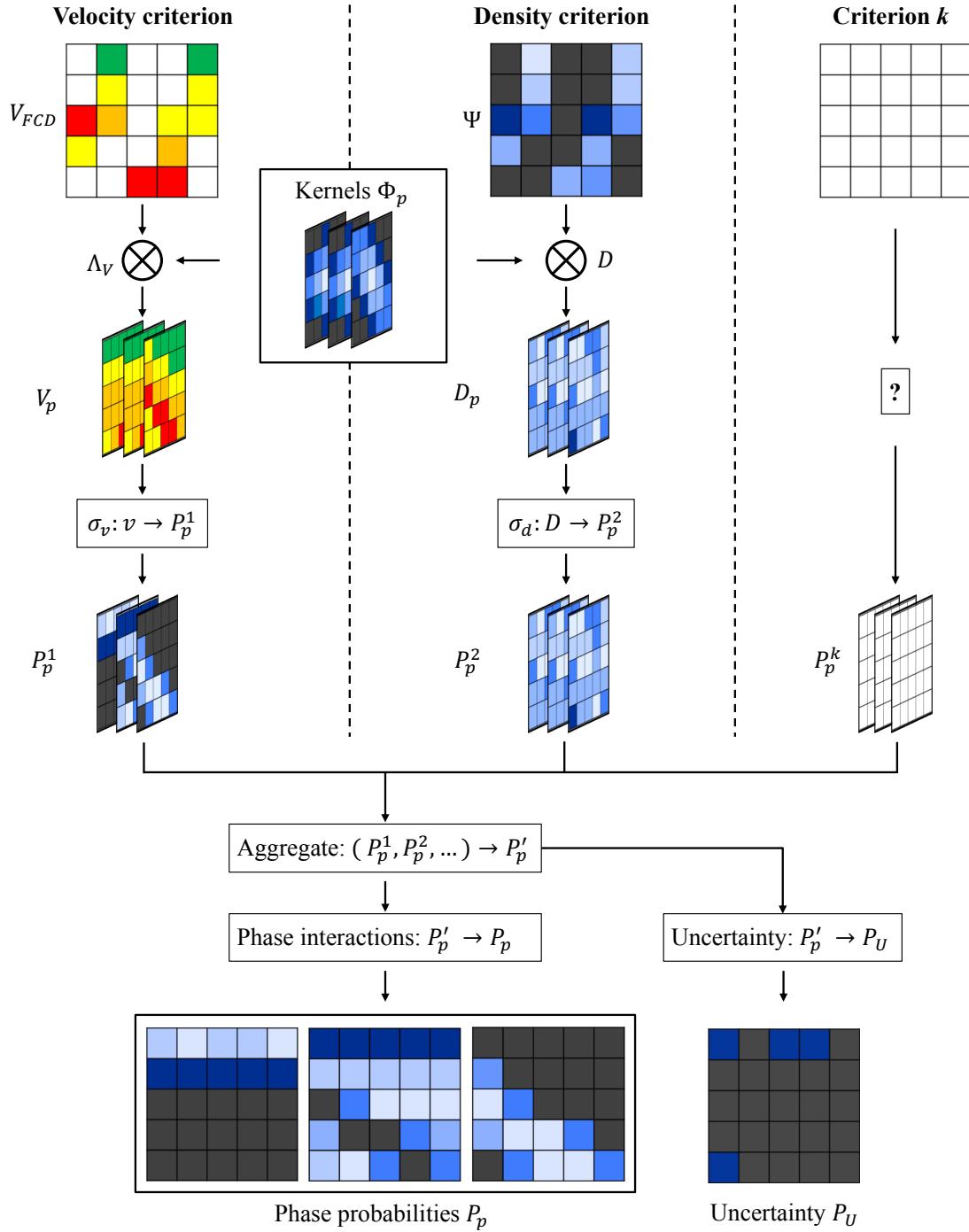


Figure 4.8. Flow of data from raw (discretized) data into phase probabilities.

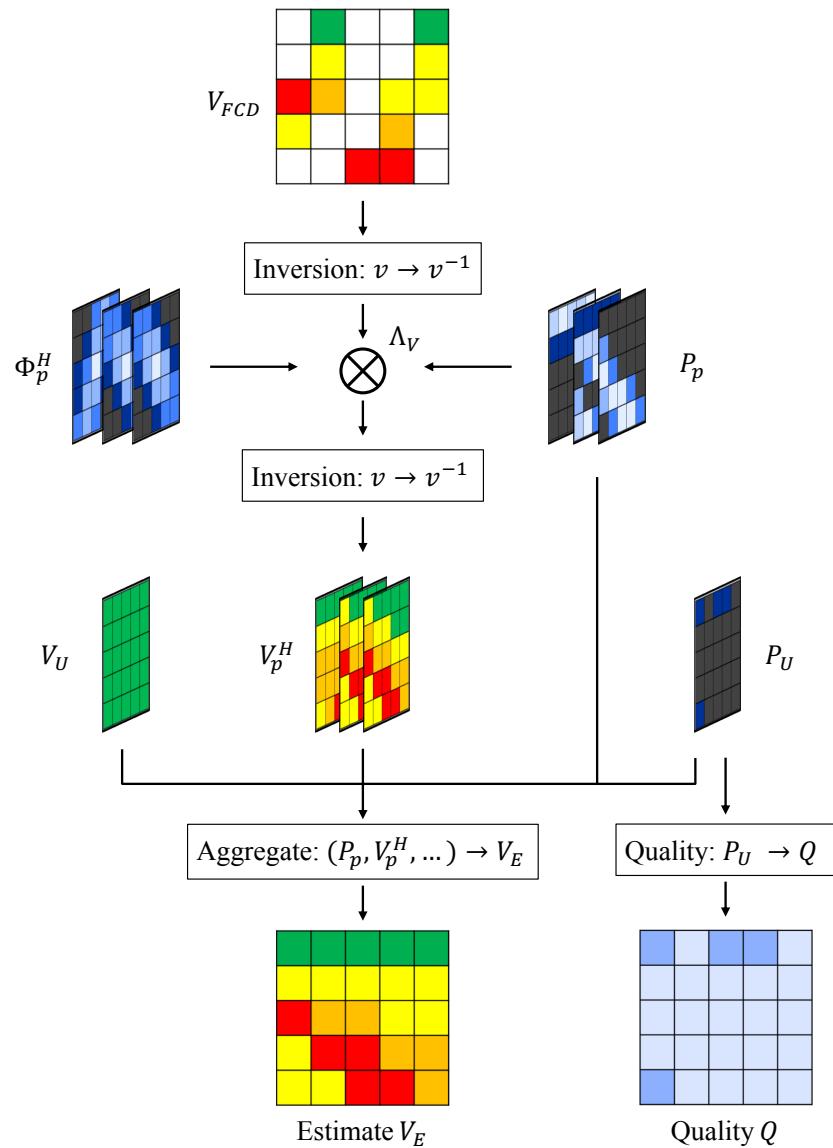


Figure 4.9. Flow of data from phase probabilities to the final velocity estimate.

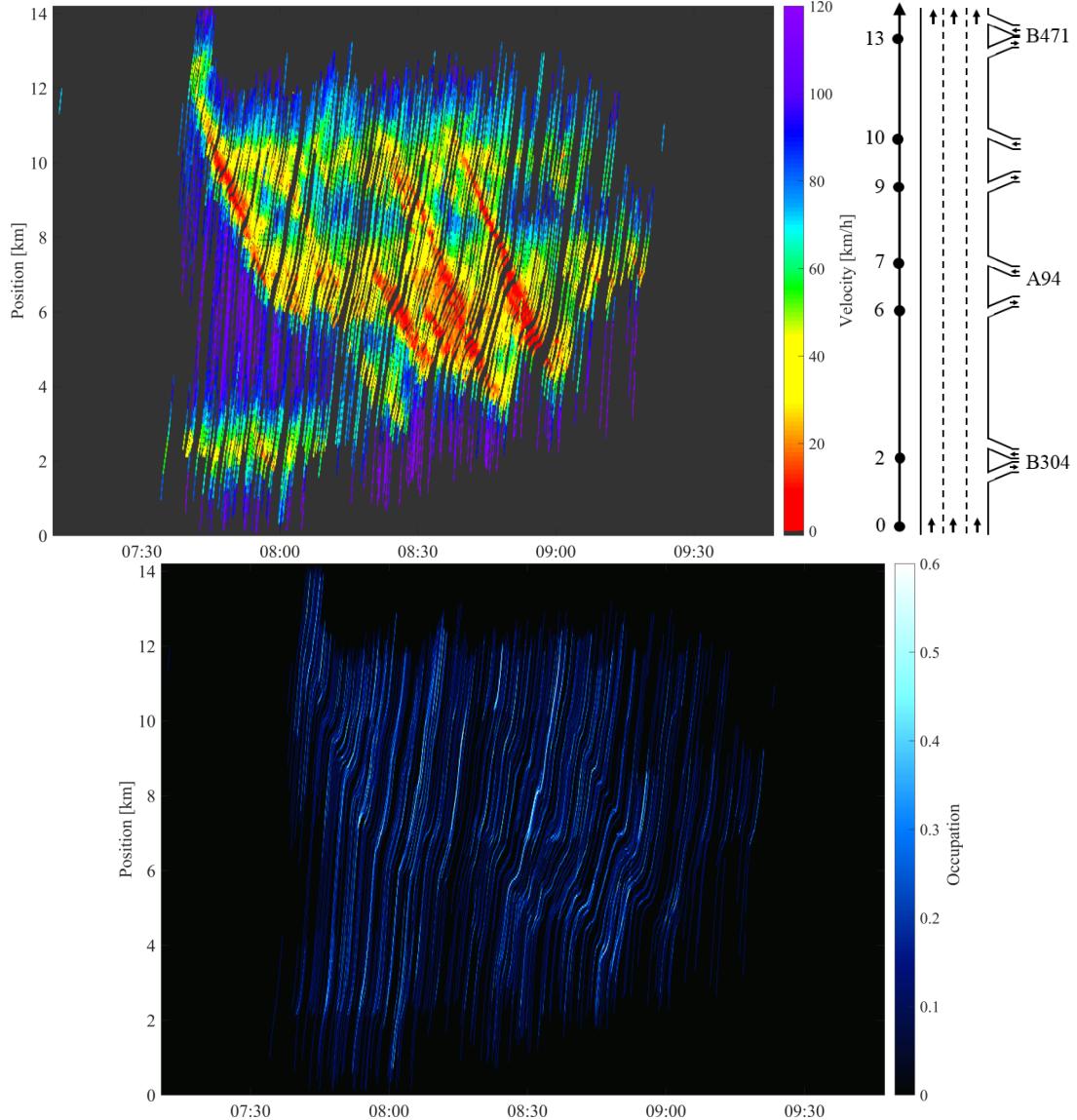


Figure 4.10. Raw trajectory data of a congestion on A99 in eastbound direction (up). Occupation of trajectories in time and space (bottom).

4.4.3 Qualitative Evaluation

Figure 4.10 depicts the trajectories and occupations collected by vehicles during a congestion on A99 in north-bound direction on July 15th 2014. The pattern shows different characteristics often occurring in congested motorway traffic (compare to traffic patterns described in (Kerner, 2009; Helbing et al., 2009; Schoenhofer and Helbing, 2007)). At around 7:45am a moving jam emerged that evolved into a WMJ and induced a traffic breakdown at the on-ramp at kilometer 10. The WMJ propagated further upstream and induced another traffic breakdown at the neighboring bottleneck. The pattern evolves into a General Pattern (GP) expanding over two bottlenecks. The downstream fronts of synchronized flow phases are fixed slightly downstream the on-ramp positions. In

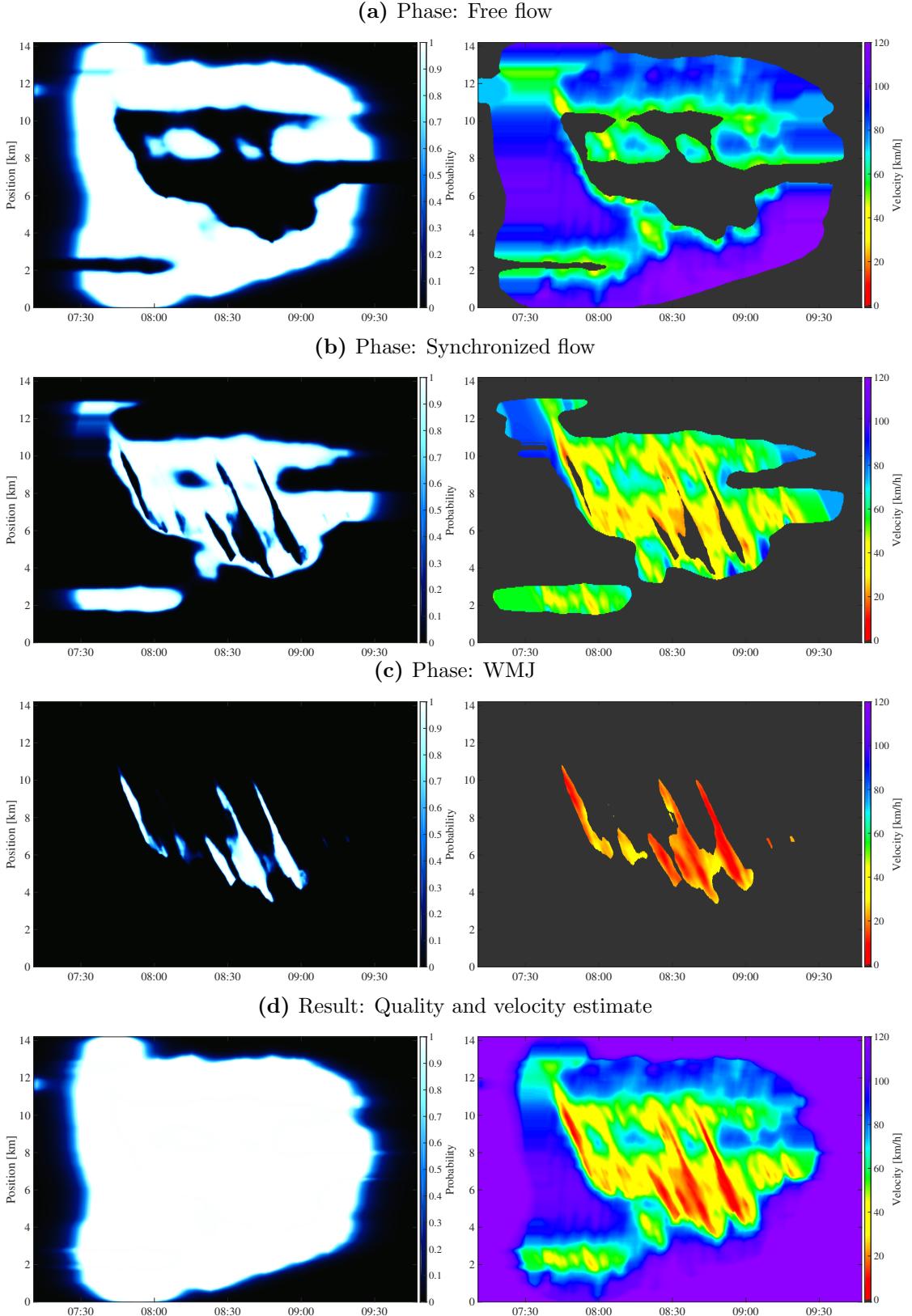


Figure 4.11. Phase probabilities P_p and phase velocities V_p^H for free flow (a), synchronized flow (b) and WMJ (c). The final quality Q and velocity estimate V_E in (d). Note that velocities with phase probabilities below 5 % are colored gray.

the pinch region of the downstream synchronized flow phase a few WMJs originate and propagate upstream.

Applying the PSM results in the phase probabilities P_p and phase-dependent velocities V_p^H illustrated in Figure 4.11. The aggregation of phase probabilities and velocities results in the quality estimate Q and the final velocity estimate V_E depicted contour plots depicted in the last row.

Varying number of available trajectories The amount of data available for reconstruction is an essential influence on the estimation accuracy. In order to give an intuition how the PSM reconstructs this scenario if fewer trajectories are collected, Figure 4.12 illustrates the estimates, assuming that only 10 – 70 % of the original traces are available. The first three contour plots look similar, with small artifacts at the boundaries and less accurate WMJ reconstructions. With 10 % more estimation errors become visible. For instance, the WMJs are not reconstructed completely and there are several space-time regions where a congested part is over- or underestimated.

Comparison with other algorithms In the following the PSM is compared to other state-of-the-art approaches. In order to highlight potential estimation errors, the algorithms are applied to a reduced set of trajectories. A subset of all trajectories is extracted and each algorithm computes a V_E . These estimates are subsequently compared to an estimated Ground Truth (GT) using all trajectory data.

Figure 4.13 illustrates the considered GT. It is the mean of all reported trajectories in a grid of 60 s × 100 m. Note that this is an estimate of the GT since not all vehicle trajectories are known. However, due to the high data density the deviation to the real GT is expected to be low. In comparison to usual GT data that stems from loop detectors, the spatio-temporal resolution of this estimated GT is significantly higher. Figure 4.13 right illustrates the set of trajectories that are used as input for different traffic speed estimation algorithms.

In Fig. 4.13 the velocity estimate $V_E(t, x)$ computed with the PSM and the absolute error $|V_E(t, x) - V_{GT}(t, x)|$ to the GT are depicted. The error plot reveals that most regions of the estimation match well with the GT. Both, the stationary congestion at the bottlenecks at kilometers 8 and 11 as well as the WMJs are accurately reconstructed. Significant differences are marked in the plot: For instances, at (a) the moving jam is reconstructed inaccurately. At (b) the stationary congestion at kilometer 2 starts earlier than estimated and is therefore underestimated. At (c), the PSM overestimates the extent of the congestion at the upstream front for a short range in time.

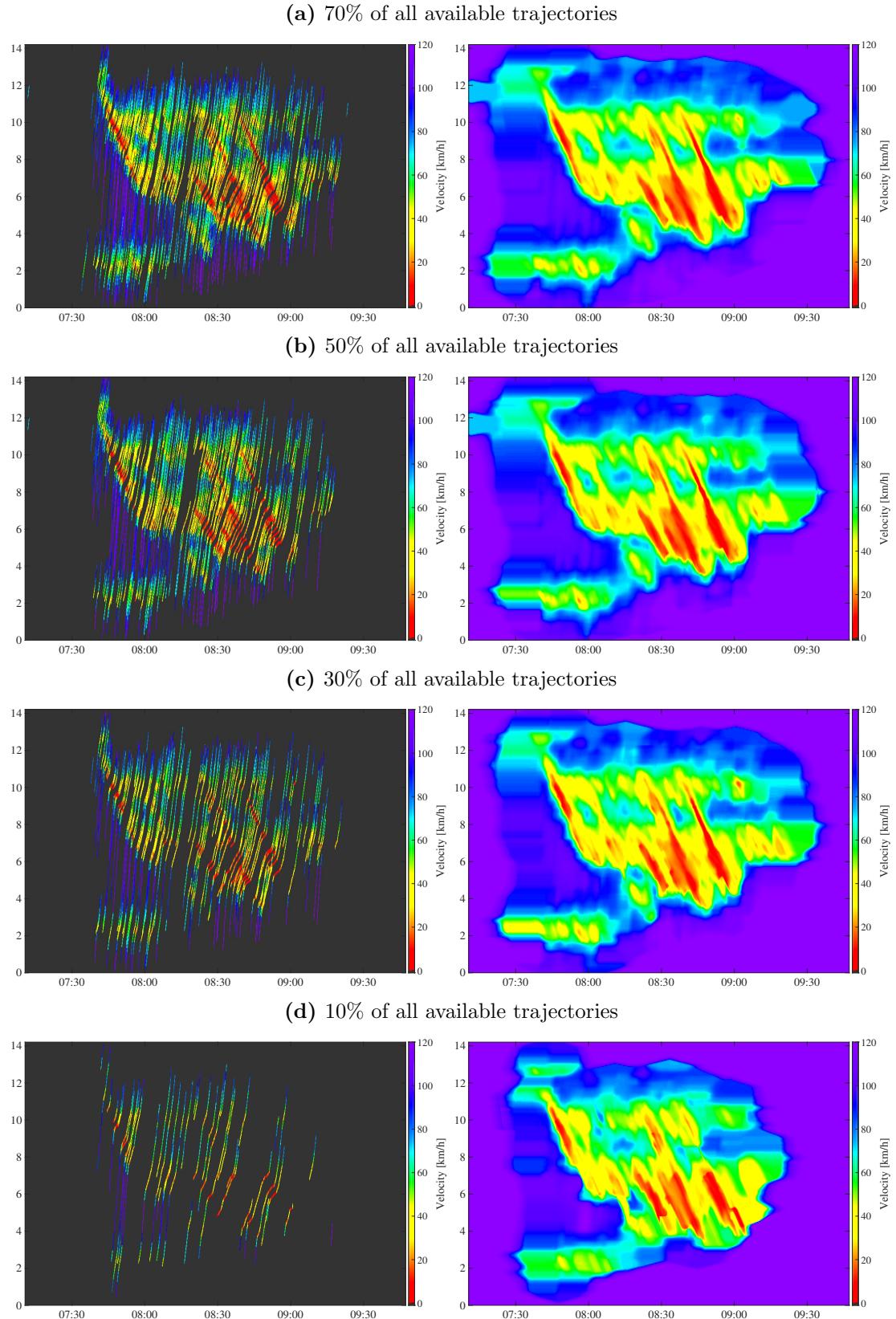


Figure 4.12. Velocity estimates generated by the PSM depending on a varying amount of trajectories used for reconstruction.

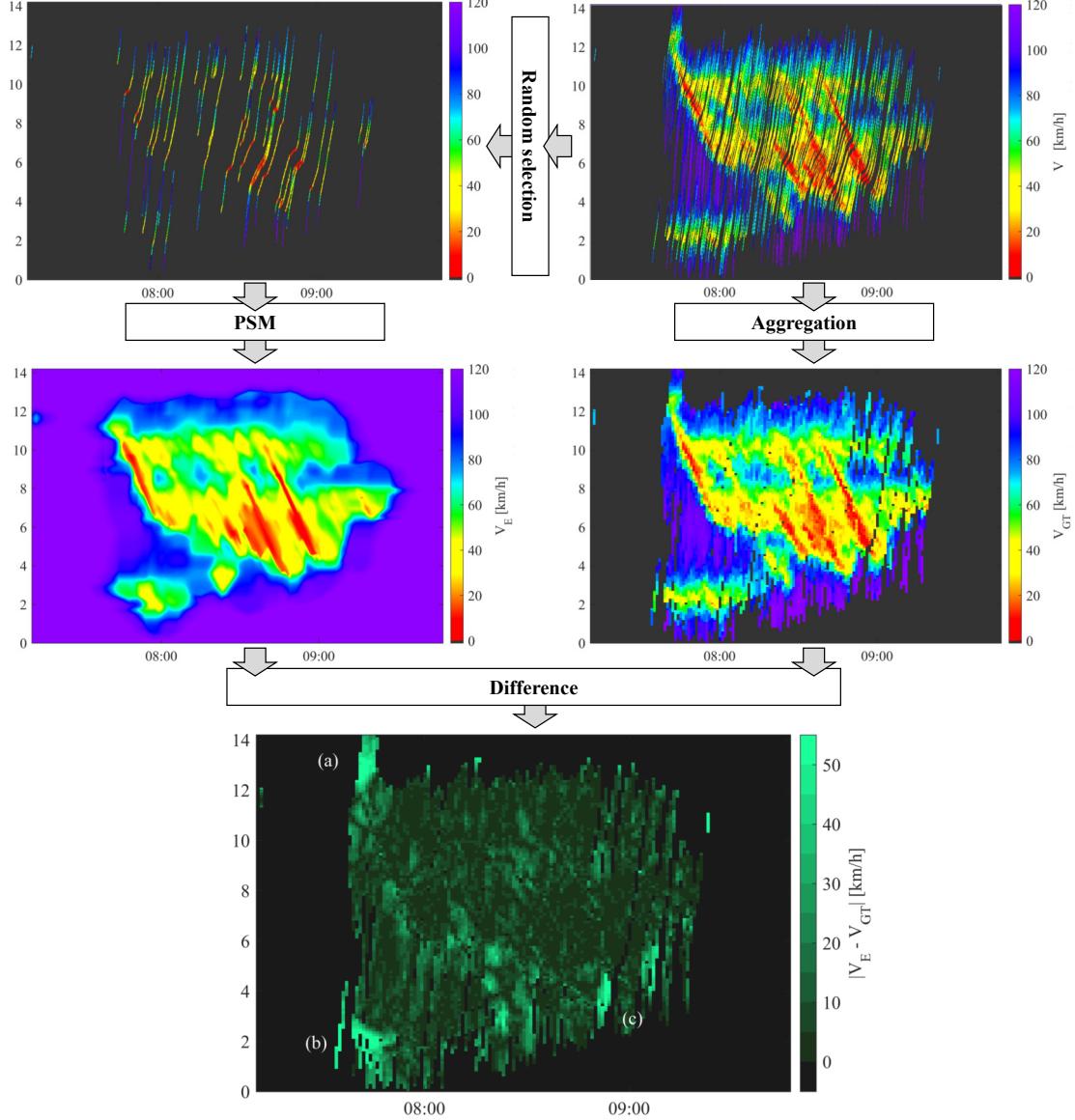


Figure 4.13. Estimated GT constructed using all trajectory data (left) and sparse set of trajectories used as data input for traffic speed estimation (right)

For comparison, the GASM as state-of-the-art traffic speed estimator with sparse data and an isotropic smoothing method are chosen. The GASM is applied as described in (Treiber and Helbing, 2003) and (van Lint and Hoogendoorn, 2009) with an adaption to sparse FCD as presented in (Rempe et al., 2016a). The adaption describes how sparse FCD and a fall-back speed can be fused in order to provide a continuous speed estimate if no data is nearby. The weighting ratio of FCD to the speed fall-back is set as 1000:1. In order to account for travel time accuracy the smoothing processes are applied to the inverted velocities as described in (van Lint, 2010). The parametrization is chosen according to (Treiber et al., 2010a). The isotropic smoothing method is a generalized

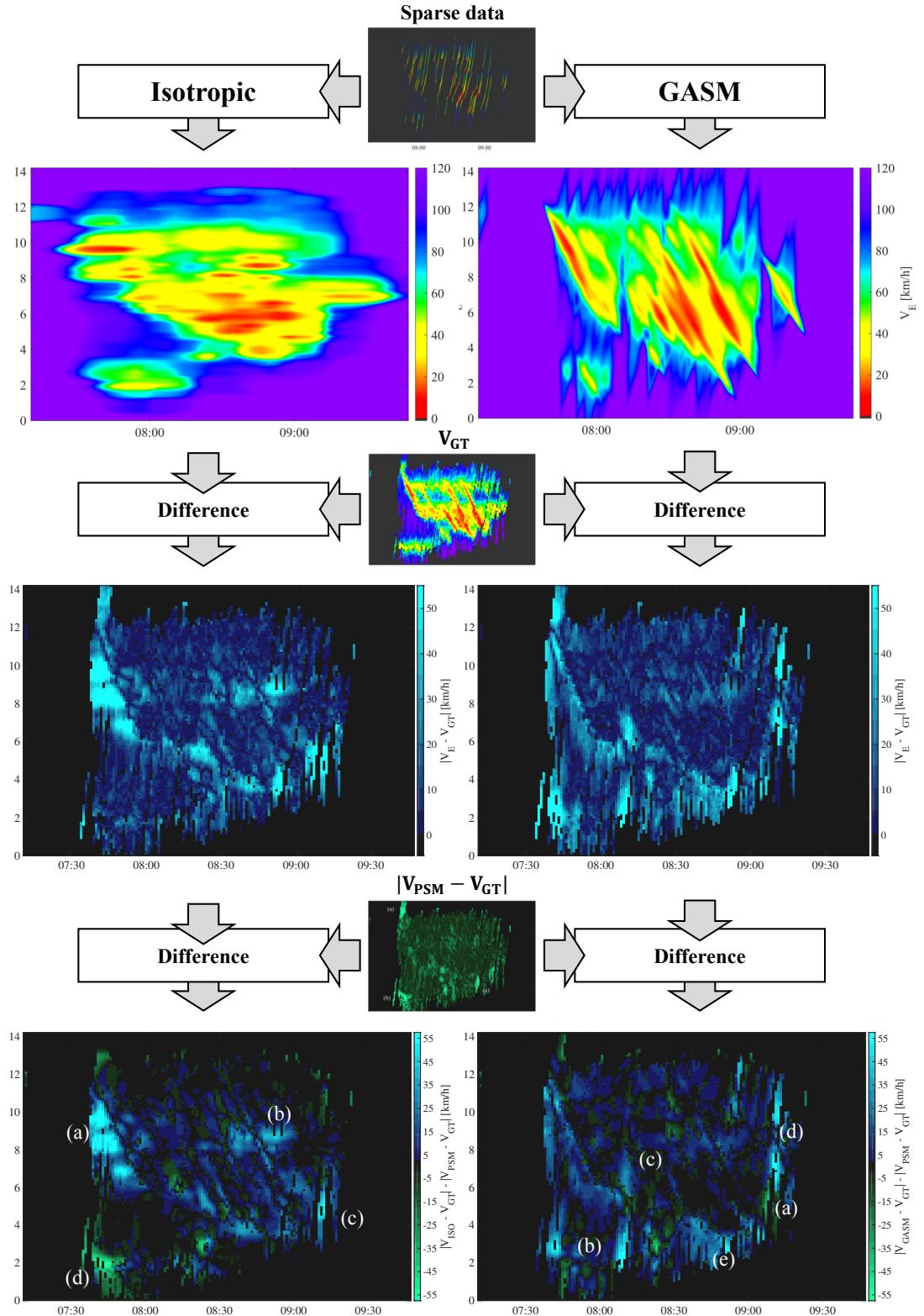


Figure 4.14. Velocity estimate using an isotropic smoothing method and the GASM. Difference of both results to the Ground Truth (mid) and difference of error plots of both methods compared to the PSM

approach that can resemble several ways to average data in space and time. Here, in order to model a standard temporal smoothing of velocity data as it is done in many systems an isotropic smoothing kernel with a small spatial and high temporal width is applied (see eq. (4.7)).

In Fig. 4.14 the velocity estimates computed with the isotropic smoothing method (upper left) and the GASM (upper right) are illustrated. These estimates are compared to the GT and its difference plots are visualized in the second row. In order to compare the reconstruction accuracy of the PSM and these algorithms directly, the differences of the estimation errors are considered:

$$V_{\Delta} = |V_{E,1} - V_{GT}| - |V_{E,2} - V_{GT}|. \quad (4.38)$$

This error term returns positive values if the error of the first velocity estimate $V_{E,1}$ dominates, and negative ones, if the second velocity estimate is less accurate. In the following comparison, blue colored regions indicate that the PSM reconstructed the velocity more accurately than the comparison method.

In Figure 4.14 (left) the isotropic smoothing method and the PSM are compared. At (a), (b), (c) the isotropic smoothing results in large errors as it does not account for moving traffic patterns. Though, at (d) this method estimates the stationary traffic slightly better than the PSM. The results of the GASM compared to the PSM as depicted in Figure 4.14 (right) show that the GASM manages to estimate the dissolving jam at (a) better than the PSM. On the other hand, it does not reconstruct the stationary traffic patterns at (b), (c), (d), but estimates it as free flow instead. Furthermore, the moving jam at (e) is extrapolated too far in upstream direction. Overall, the difference plots reveal that both algorithms result in larger erroneous regions than the PSM, which combines the strengths of both comparative algorithms: To reconstruct stationary as well as moving traffic patterns.

4.4.4 Quantitative Assessment of Estimation Accuracy

This section focuses the overall estimation accuracy of the PSM compared to the other two aforementioned methods. In order to assess the estimation accuracy of an algorithm and compare it to another method, several aspects need to be considered. The first aspect concerns the procedure how to split given data into a training and a test set (section 4.4.4.1). Second is the set of scenarios that are used for evaluation. Section 4.4.4.2 describes an excerpt of the freeway network on which a multitude of congestion patterns are observed and used for comparison. Another aspect is that the amount of available data used for estimation influences the achievable accuracy significantly.

In section 4.4.4.3 this aspect is elaborated and a novel concept called data coverage is developed in order to quantify this factor. Fourth, there are several metrics that can be applied in order to compare estimated speeds and the GT. Section 4.4.4.4 gives an overview of potential quality metrics and selects the one that meets best the requirements for the assessment of a traffic speed estimator

Subsequently, the proposed methodology is applied exemplary to the congestion pattern described in section 4.4.3 and results are presented in section 4.4.4.5. Finally, the estimation results of the three algorithms with different parametrizations applied to all 101 congestion patterns are presented in section 4.4.4.6.

4.4.4.1 Split of Data

In order to assess the quality, i.e. the similarity between an estimated traffic state and the GT the data used for estimation and for evaluation must not be the same. Otherwise, the best algorithm would be the one, that simply reproduces the data. Therefore, the set of all N_c trajectories reported for the space-time domain $[T_0, T_1] \times [0, L]$ is divided into a training and a test set. The training set is used to estimate velocities V_E , and the test set is used to evaluate the accuracy of that estimate. The size of test set $\mathcal{S}_T = \{tr_1, tr_2, \dots, tr_{N_T}\}$ relates to the total number of trajectories as:

$$N_T = \alpha N_c \quad (4.39)$$

with $0 < \alpha < 1$. The training (estimation) set \mathcal{S}_E of size N_E is the $(1 - \alpha)$ part of all trajectories. Additionally, that part is varied with another factor $0 < \beta < 1$ that is used to simulate different data coverages:

$$N_E = \beta(1 - \alpha)N_c. \quad (4.40)$$

4.4.4.2 Scenarios

A great part of publications that propose models for traffic estimation and prediction consider only one scenario or one road for model evaluation. This allows to optimize the model and its parameters to this example. Naturally, this pre-selection and optimization increases the chance that a proposed model performs well; especially better than others developed and optimized for other scenarios. In order to overcome this issue, a model needs to be tested on scenarios that did not influence the development and optimization. In the best and most general case, congestion patterns from all roads over the world collected over many years are taken into consideration. Unfortunately, data availability

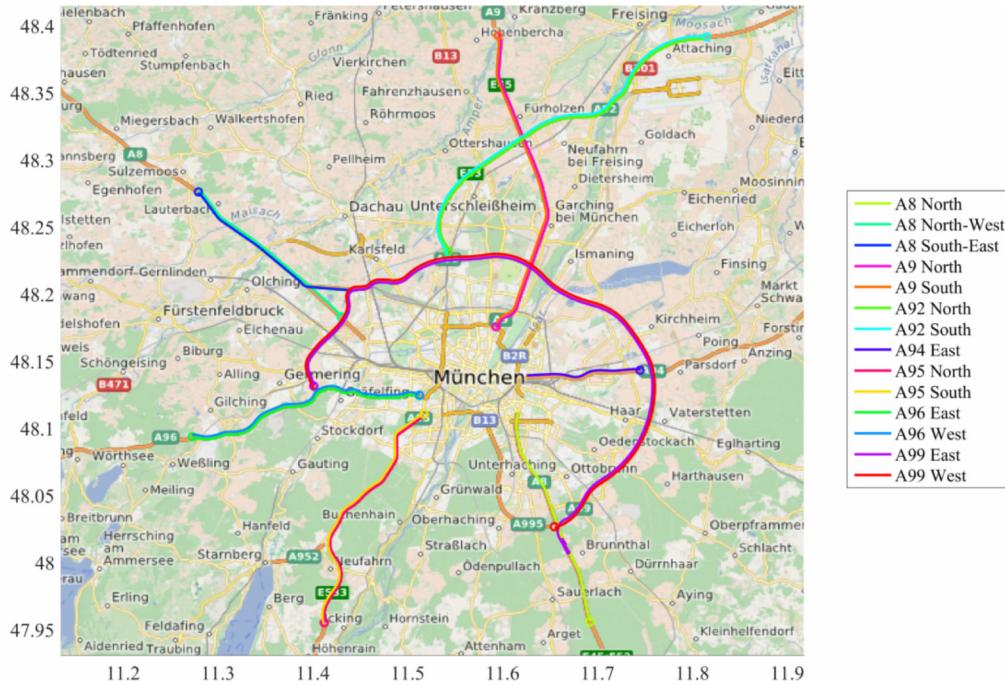


Figure 4.15. Overview of all freeways around Munich where relevant congestion occurred between July 7th, 2014 and August 7th, 2014 ¹

and computational resources are limited. Therefore, a subset of all congestion patterns with a low bias needs to be chosen for evaluation.

In this case, the set of evaluation scenarios comprises all traffic congestion patterns that occurred during July 7th, 2014 and August 7th, 2014 in the freeway network surrounding Munich (see Figure 4.15 and A.1). A congestion pattern is defined as the occurrence of a region in space-time with congested traffic speeds below 60 km/h, a temporal width of at least one hour and a spatial width of at least 2 km. This set consists of in total 101 congestion patterns. Hence, congestion patterns occurring on different road infrastructures at different times and due to several reasons are considered. Although the set of patterns is obviously biased due to the preselected location (Munich road network), the time interval (July and August 2014) and the minimal size of congestion, one important aspect is that patterns itself are not pre-selected. Thus, it is assumed that this set of patterns constitutes a reasonably representative set of congestion patterns for traffic patterns in South Bavaria. Since the Three-Phase traffic theory has been validated on traffic patterns observed on international freeways (Rehborn et al., 2011), it is expected that results can be transferred to roads apart from the presented ones as well.

¹Map data provided by OSM

4.4.4.3 Definition of Data Coverage

Since the amount of available data is an important influence on the estimation accuracy, it is crucial to consider this factor in order to obtain representative results (Palmer, 2011; Rempe et al., 2016b; Bekiaris-Liberis et al., 2016). The term 'data coverage' refers to the overall *amount* of data that is available for estimation (In comparison: data density is defined as a phase criterion in the context of this thesis). An accurate estimator outperforms other approaches for low as well as high data coverages.

Given loop detector data a common approach is to consider the average detector spacing as data coverage (Treiber and Helbing, 2003; Treiber et al., 2010a). With simulated FCD often the penetration rate of reporting vehicles is considered (Palmer et al., 2011). In the context of real FCD and mixes of several data sources both concepts are unfeasible: First, since the total number of vehicles is usually unknown the equipment rate is difficult to approximate. Second, during the observation of a certain road segment the number of reported trajectories usually varies over time. A resulting average equipment rate is a rough simplification. Third, in applications where FCD and other data sources are fused, both concepts fail to account for the level of information that mixes of heterogeneous sensors provide.

Due to these reasons a novel definition of data coverage is proposed that considers incomplete data, local variations of data densities and mixes of different data sources. As described in section 4.3.1 the velocities provided by probes (or by other sensors) occupy space-times $\Psi(t, x)$. For the estimation of the traffic state at (t, x) a traffic state estimator usually considers data in the surrounding of this point. For instance, the PSM combines the results of local smoothing operations. Therefore, in order to approximate the amount of data that is available for the estimation of $V_E(t, x)$, the occupation in the proximity of (t, x) is most relevant. With respect to the definition of the data density at (t, x) (eq. (4.4)), a standard weighting $w_0(t, x) = 1$ and kernel function $\Phi(t, x)$, the normalized local data coverage D_E is defined as the result of smoothing the $\Psi(t, x)$ of all occupation data:

$$D_E(t, x) = \frac{D(w_0, \Phi_D, t, x)}{\int_{T_0}^{T_1} \int_0^L \Phi_D(t - \hat{t}, x - \hat{x}) d\hat{x} d\hat{t}}. \quad (4.41)$$

By definition of the occupation Ψ it holds that $0 < D_E < 1$. For the following studies, a medium-sized kernel Φ_D with $v^{dir} = 0$ km/h, $\tau = 200$ s and $\sigma = 300$ m using kernel definition (eq. (4.7)) is applied. Figure 4.16 depicts exemplary the resulting local data coverage $D_E(t, x)$ of the scenario in section 4.4.3.

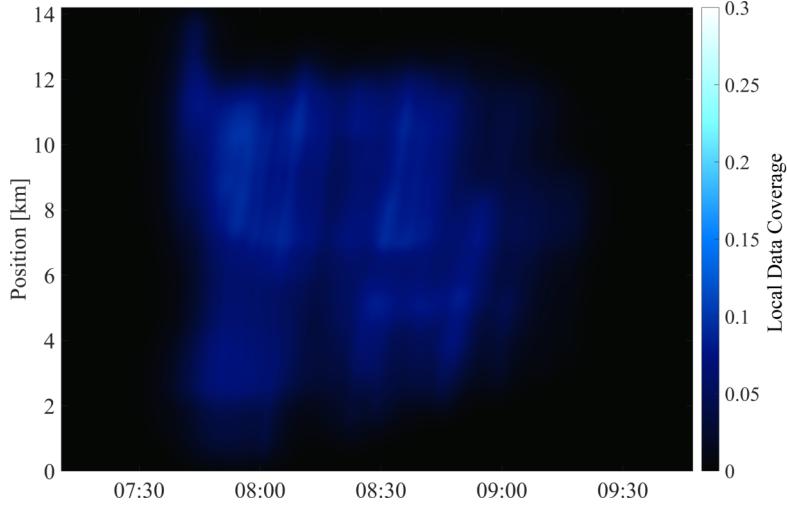


Figure 4.16. Local data coverage D_E based on the occupation of the vehicle data shown in Fig. 4.10

As visible, there are spatio-temporal regions with increased data coverage, and regions where no information is available at all. In the hypothetical case where all trajectories of all vehicles are known, and traffic density is high, it would hold that $D_E(t, x) = 1 \forall (t, x)$. In comparison to the occupation $\Psi(t, x)$ the data coverage is a smoothed quantity in space and time. This allows for a fusion of several sources and models the exponential decay of measurement information in space and time.

4.4.4.4 Quality Metrics

A quantitative assessment of the estimation accuracy requires to apply an error metric that compares a traffic speed estimate V_E and a GT V_{GT} . In order to select the most appropriate metric, it is necessary to understand what exactly an error metric is supposed to penalize. In this section first an overview of possible quality metrics for the evaluation of traffic speed estimates is given and, subsequently, one of them is selected.

In literature a multitude of quality metrics have been proposed. These metrics can be classified according to the quantity that is evaluated: The traffic speed at space-time (t, x) , the slowness or the Travel Time (TT) that a vehicle needs to pass a certain road segment. Moreover, the formulations can be 'absolute' or 'relative'. The latter relate an error term to the GT. Table 4.3 provides an overview of quality metrics that have been proposed for the assessment of traffic speed information (compare to (Bogenberger and Weikl, 2012; Huber et al., 2014)).

The quality metrics Q-FCD and QKZ are designed in order to evaluate the quality of traffic messages. They compare traffic information gathered by probes or detectors

| | Absolute | Relative |
|-------------|--|---|
| Velocity | $MAE = \frac{1}{N} \sum V_{GT} - V_E $ $RMSE = \sqrt{\frac{1}{N} \sum (V_{GT} - V_E)^2}$ Q-FCD, QKZ | $MAPE = \frac{1}{N} \sum \frac{ V_{GT} - V_E }{V_{GT}}$ |
| Slowness | $IMAE = \frac{1}{N} \sum \left \frac{1}{V_{GT}} - \frac{1}{V_E} \right $ | $IMPE = \frac{1}{N} \sum \left \frac{\frac{1}{V_{GT}} - \frac{1}{V_E}}{\frac{1}{V_{GT}}} \right $ $SIMPE = \frac{1}{N} \sum \left \frac{\frac{1}{V_{GT}} - \frac{1}{V_E}}{\frac{1}{V_{GT}}} \right ^2$ |
| Travel Time | $MAETT = \frac{1}{N} \sum TT_{GT} - TT_E $ | $MAPETT = \frac{1}{N} \sum \left \frac{TT_{GT} - TT_E}{TT_{GT}} \right $ Q-BENCH |

Table 4.3. Overview of quality metrics for traffic information assessment

with traffic messages published by e.g. road authorities. A speed threshold is defined that determines whether traffic at (t, x) is congested or not. Two errors are calculated: The first describes the correctly identified congested space-time regions. The second measures the region of 'false positives', which is falsely assumed congested traffic. These metrics allow to evaluate the spatio-temporal description of congestion. Though, since measured speeds are turned into a binary signal, real travel times are neglected. Since the PSM is meant to provide accurate traffic speed estimates for all types of traffic applications, wrongly estimated congested speeds are supposed to be penalized. As a consequence, Q-FCD and QKZ do not meet the requirements of a quality metric.

In order to consider real travel times, there are metrics that evaluate the TT a vehicle requires to pass pre-defined road segments. Besides simple ones such as the MAETT or MAPETT, another procedure is the Q-BENCH metric (Bogenberger and Weikl, 2012; Lotz and Luks, 2011). These metrics compare estimated TTs with real measurements obtained from probe data. In practical application these error metrics are often applied

on road road segments that exceed several hundreds of meters in order to optimize routing systems. However, the goal of the PSM is to accurately reconstruct the spatio-temporal traffic speed of congestion patterns. Therefore, TT-based metrics are not the appropriate choice for this evaluation.

Velocity-based metrics such as the MAE or RMSE constitute continuous metrics that penalize velocity differences. Since accurate spatio-temporal traffic speed estimates automatically result in accurate travel time estimates, these metrics consider important aspects of an appropriate quality metric. However, the error in MAE and RMSE does not meet all requirements: These metrics consider absolute speed errors, although the travel time of a vehicle is inversely proportional to traffic speed. Hence, at low traffic speeds errors are small, although even small differences impact real travel time significantly. Vice versa, velocity differences at high velocities result in large errors, but actually influence the travel time only slightly. In addition, in free flow conditions vehicles' driving speeds tend to diverge (Kerner et al., 2004) which contributes to the conclusion that these error metrics are not sufficiently sensitive to congested conditions in order to be applied for traffic speed evaluation.

The MAPE is an approach to overcome this issue. It relates absolute speed errors to the GT. In this way errors at low GT speeds are weighted stronger, which is consistent with travel time errors. Though, using the GT in order to normalize the deviation causes an antisymmetry: While 'true negatives' (congested conditions which are estimated as free flow) are penalized strongly, 'false positives' are barely penalized since the denominator is large. This also applies to the metrics IMPE and SIMPE, but conversely, since reciprocal speeds are considered. Effectively, these metrics only penalize congestion that is under-estimated (or over-estimated respectively). This antisymmetry is a major drawback.

The IMAE is symmetric, sensitive to spatio-temporal traffic speed dynamics and continuous. Furthermore, the consideration of reciprocal velocities neglects velocity errors in free flow conditions and is in accordance with travel time calculations. Therefore, it is selected as the most appropriate quality metric in order to assess the estimation accuracy of the PSM.

Accordingly, using test set \mathcal{S}_T in order to evaluate estimate $V_E(t, x)$ yields the error:

$$IMAE = \frac{1}{N_{tup}} \sum_{tr \in \mathcal{S}_T} \sum_{(t,x,v) \in \mathcal{S}_T} \left| \frac{1}{v} - \frac{1}{V_E(t, x)} \right| \quad (4.42)$$

with N_{tup} the total number of (t, x, v) - tuples of all trajectories in the test set \mathcal{S}_T . With respect to the data coverage defined in section 4.4.4.3, the mean data coverage MD of

all evaluated tuples is:

$$MD = \frac{1}{N_{tup}} \sum_{tr \in \mathcal{S}_T} \sum_{(t,x) \in \mathcal{S}_T} D_E(t, x). \quad (4.43)$$

4.4.4.5 Accuracy Assessment of Sample Scenario

This section presents the accuracy of the isotropic smoothing method, the GASM and the PSM in estimating traffic velocities of the test scenario presented in section 4.4.3. Furthermore, a generalized error estimator is motivated, which is applied in the subsequent extensive study.

Since the parametrization of an algorithm is the basis for a fair comparison, two parameters sets are considered for each algorithm in order to understand better the influence of its parameters on the estimation accuracy. The isotropic smoothing uses $\tau = \{150\text{s}, 300\text{s}\}$ and $\sigma = \{100\text{m}, 200\text{m}\}$. The most influential parameters of the GASM are set according to the publications by (van Lint and Hoogendoorn, 2009; Treiber and Helbing, 2003) as $\tau = \{30\text{s}, 70\text{s}\}$ and $\sigma = \{300\text{m}, 700\text{m}\}$. The parameters of the PSM are chosen as listed in Table 4.2. According to the sensitivity analysis presented in section 4.4.6 the two most influential parameters, $\tau_{F,S}$ and $\tau_{S,J}^H$, are set to $\tau_{F,S} = \{300\text{s}, 400\text{s}\}$ and $\tau_{S,J}^H = \{20\text{s}, 30\text{s}\}$.

The ratio between train and test set is chosen as 60:40 (i.e. $\alpha = 0.4$). β is varied between 5 – 100 %. All trajectories of the scenario depicted in Figure 4.10 are divided randomly into test and training set according to eq. (4.40). Subsequently, a traffic speed estimate is computed and the IMAE is determined. In order to achieve a robustness of the result, this procedure is iterated 50 times for each considered value of β and mean IMAE values are calculated.

Figure 4.17 depicts the mean *IMAE* and 80 %-quantile with respect to the *MD* in a logarithmic scale. Several observations can be made:

1. As expected, all estimators reconstruct traffic speeds more accurately with an increasing data coverage. Though, the traffic-motivated smoothing algorithms GASM and PSM achieve higher gains and lower absolute errors than the isotropic smoothing.
2. Depending on the parametrization, the GASM achieves either accurate estimations at low data coverages or at high data coverage. This compromise is also described in (Rempe et al., 2016a).

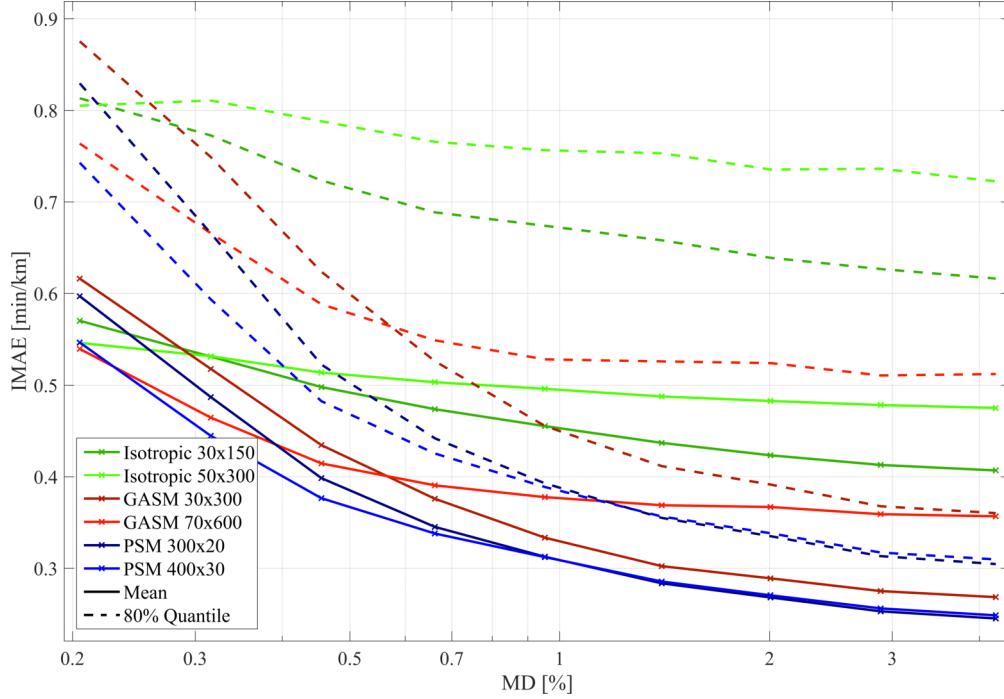


Figure 4.17. Mean IMAE and 80% quantile of two variants of the isotropic smoothing, the GASM and the PSM with respect to the mean data coverage

3. One variant of the PSM achieves the most accurate results for all coverages. The mean errors difference between that PSM variant and one GASM variant is low for some coverages, though the 80 %-quantile is significantly more accurate at all coverages.
4. Also for the PSM the parametrization is relevant. In this case the estimation accuracy for low data coverages is higher using a greater value of $\tau_{F,S}$. Though, both variants produce similar results with greater data coverages.

Note that, alike the mean, a quantile is another characteristic of the error distribution of an algorithm at a certain data coverage. In this case it is displayed in addition to the mean in order to provide more insights into the distribution of errors. A lower 80 %-quantile does not necessarily mean that an algorithm is more accurate than another one. Rather, its distribution of errors is different. The mean error is the quantity that indicates more intuitively if, 'in average, one algorithm is more accurate than another'.

The $IMAЕ(MD)$ allows to analyze one scenario and determine which algorithm outperforms others at a certain data coverage. For instance, if the data coverage is low, the isotropic smoothing yields more accurate results than some variants of the GASM and PSM. However, it lacks accuracy at higher data coverages. In order to identify the most accurate algorithm in the general case, one needs to describe which errors at which data coverages need to be compared. This motivates the definition of an error term

| | Isotropic | | GASM | | PSM | |
|----------------------------|-----------|--------|--------|--------|--------|--------|
| | 30x150 | 50x300 | 30x300 | 70x600 | 300x20 | 400x30 |
| \overline{IMAE} [min/km] | 0.4342 | 0.4871 | 0.3109 | 0.3733 | 0.2881 | 0.2868 |
| Q_{80} [min/km] | 0.6508 | 0.7452 | 0.4244 | 0.5310 | 0.3638 | 0.3599 |

Table 4.4. Mean aggregated estimation errors and error quantiles of the isotropic method, the GASM and the PSM for the scenario depicted in Figure 4.10

that balances the errors for low and high data coverages, and aggregates the $IMAEs$ into one quantity. Assume the data coverage for a given scenario varies depending on several factors such as the penetration rate of equipped vehicles etc. Then, MD can be modeled as a random variable where $MD \in [0, 1]$ and $P(MD)$ as the according probability density function. As a result, the expected error of a situation with random data coverage is:

$$\mathbb{E}(IMAE) = \int_0^1 P(MD) \cdot IMAE(MD) \cdot dMD. \quad (4.44)$$

The goal is to identify the algorithm that performs best, averaged over all data coverages. Therefore, one needs to determine a probability density $P(MD)$ of data coverages in order to aggregate the error function $IMAE(MD)$. The estimation of this probability density for a given fleet is out of scope of this work but may be elaborated in future works. For simplicity, in the following it is assumed that $P(MD)$ has a uniform distribution (i.e. $P(MD) = const$). The aggregated error term \overline{IMAE} simplifies into:

$$\overline{IMAE} = \frac{1}{MD_{max} - MD_{min}} \int_{MD_{min}}^{MD_{max}} IMAE(MD) dMD \quad (4.45)$$

with MD_{max} and MD_{min} the maximal/minimal data coverage that is available for reconstruction. In the following, MD_{min} is set to 5 % of MD_{max} in order to eliminate exceptionally high errors due to extremely low data coverages. Function $IMAE(MD)$ is approximated with a piecewise-linear function between sampled pairs of $IMAE(MD)$ resulting from different values of β and eq. (4.43). For the presented scenario the resulting error values are listed in Table 4.4. These errors reflect well the previous observations: The PSM has the lowest overall error and lowest quantile, followed by the GASM and finally the isotropic method.

4.4.4.6 Comparative Results

This section presents the results of the accuracy assessment of the three algorithms applied to all 101 scenarios. Therefore, first the definition of relative error bounds of

two algorithms which may have several variants are introduced. Next, the isotropic smoothing method and the GASM are compared to the PSM. Therefore, for each pair first the resulting accuracy values are presented and, second, some examples of congestion patterns are analyzed for which one of the algorithms outperformed the other one.

Since congestion patterns may vary significantly, naturally also the absolute values of the resulting errors vary. Furthermore, each algorithm may have different variants, e.g. different parametrizations. Therefore, in order to compare to classes of algorithms it is reasonable to consider relative error bounds. Let \mathcal{A} be a class of algorithms. Each algorithm $a \in \mathcal{A}$ represents a variant of this class of algorithm. Applying a to an estimation problem results in the error $\overline{IMAE}(a)$. The relative error of two variants a and b of possibly different classes of algorithms \mathcal{A}_k and \mathcal{A}_l with $k, l = 1, \dots, N_A$ is defined as:

$$\epsilon(a, b) := \frac{\overline{IMAE}(a)}{\overline{IMAE}(b)}, \quad a \in \mathcal{A}_k, b \in \mathcal{A}_l. \quad (4.46)$$

The lower/upper relative error bound ϵ^L/ϵ^U is defined as the minimal/maximal relative error of all pairs in \mathcal{A}_k and \mathcal{A}_l :

$$\epsilon^L(\mathcal{A}_k, \mathcal{A}_l) := \min_{a \in \mathcal{A}_k, b \in \mathcal{A}_l} (\epsilon(a, b)) \quad (4.47)$$

$$\epsilon^U(\mathcal{A}_k, \mathcal{A}_l) := \max_{a \in \mathcal{A}_k, b \in \mathcal{A}_l} (\epsilon(a, b)). \quad (4.48)$$

The same definition is applied for the quantiles.

PSM vs. Isotropic Smoothing Figure 4.18 visualizes the relative error bounds and quantiles of the PSM compared to the isotropic smoothing method for all 101 congestion patterns. Several observations can be made:

1. The lower and upper error values vary significantly with values between 0.4 and 1.2.
2. For the majority of the scenarios the PSM is more accurate than the isotropic smoothing method.
3. There are a few scenarios where both classes of algorithms produce similar results, or the isotropic smoothing method is slightly more accurate.

Figure 4.19 illustrates the velocity estimates of a variant of the PSM and the isotropic smoothing for data of a congestion where the PSM yields a significantly lower estimation error. The comparison of the contour plots reveals that the PSM achieves to reconstruct

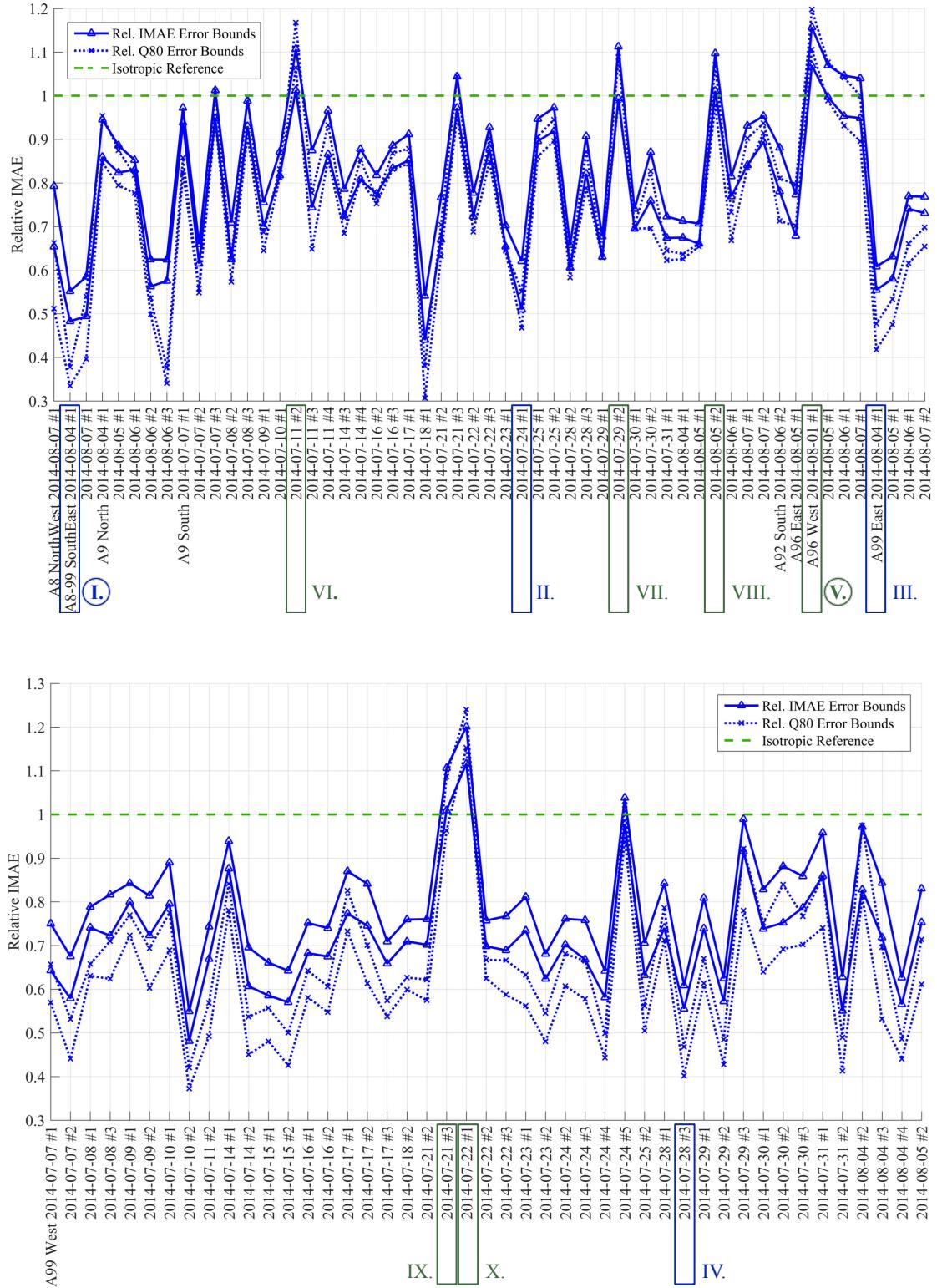


Figure 4.18. Relative \overline{IMAE} and its quantiles of the PSM compared to the isotropic smoothing method. Error bounds include the minimal and maximal relative error with respect to different parametrizations

Scenario I.

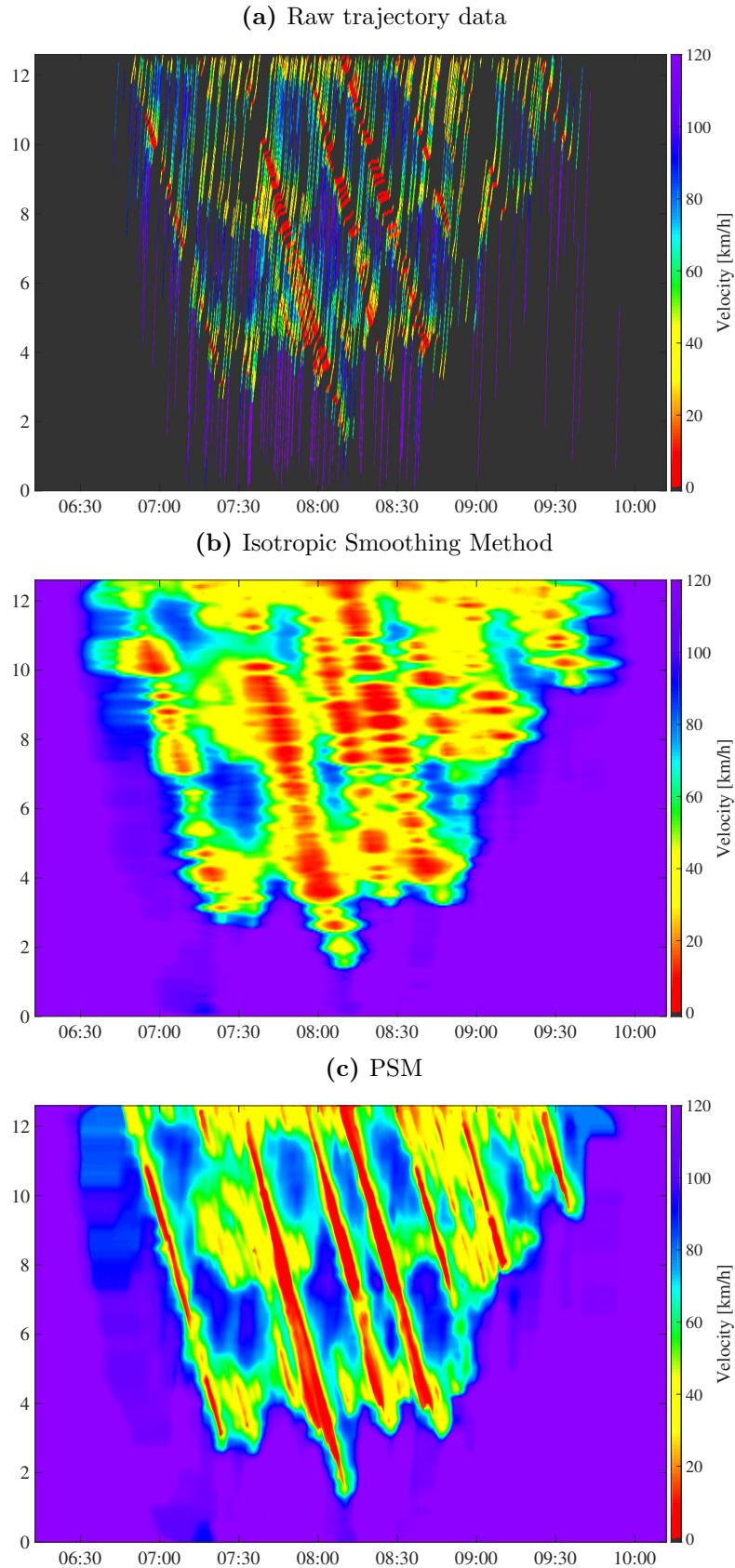


Figure 4.19. Comparison of velocity estimates of an isotropic smoothing and the PSM for a congestion that occurred on A8-99 in south-east direction on August 4th 2014

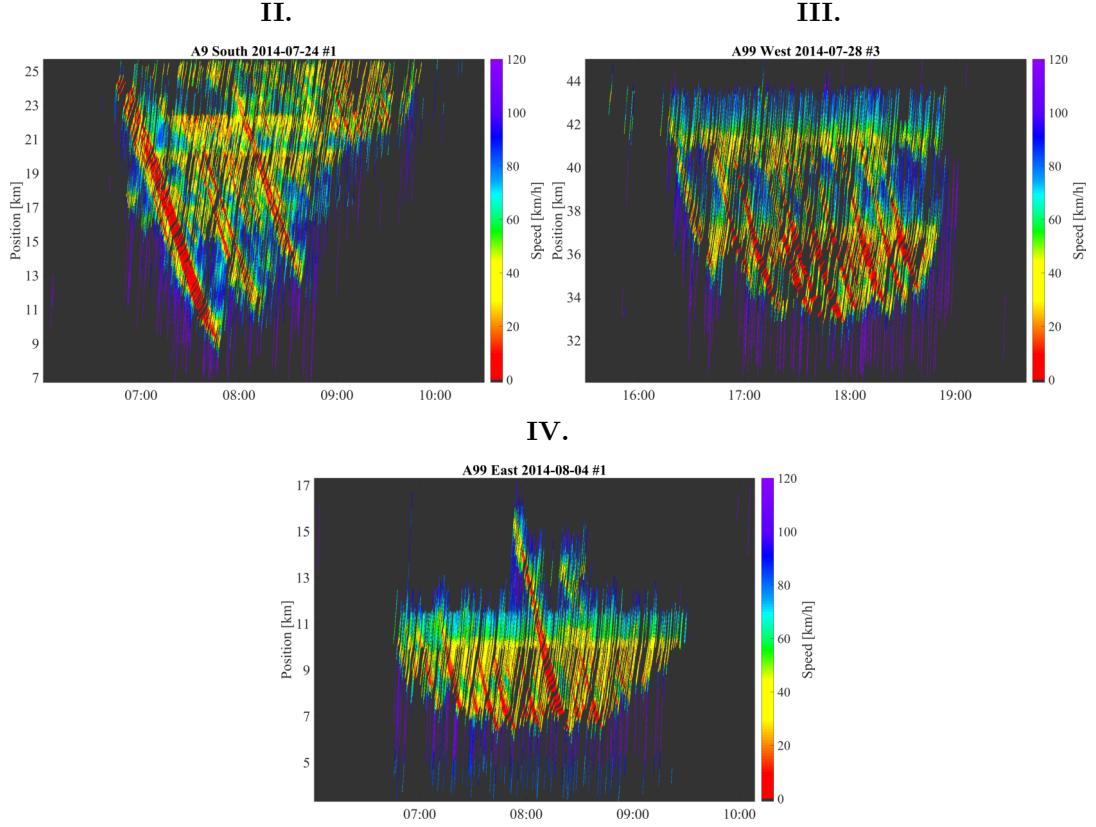


Figure 4.20. Congestion patterns that are reconstructed significantly more accurate with the PSM compared to an isotropic smoothing method

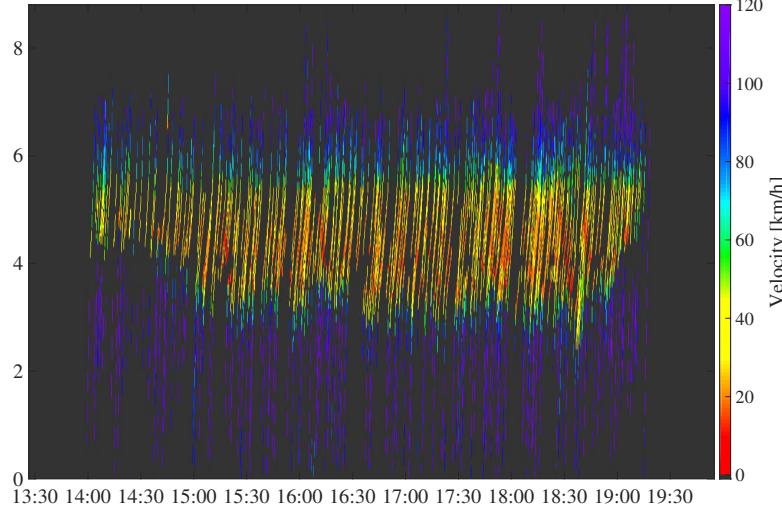
the WMJs while the isotropic smoothing fails to do so. This explains the higher accuracy of the PSM.

Figure 4.20 illustrates the raw data of a few other scenarios where the PSM produces significantly more accurate results. These patterns comprises space-time regions of synchronized flow and WMJs. Since the isotropic smoothing method is not able to reconstruct moving jams, variants of this class of algorithm fail to determine accurate velocity estimates.

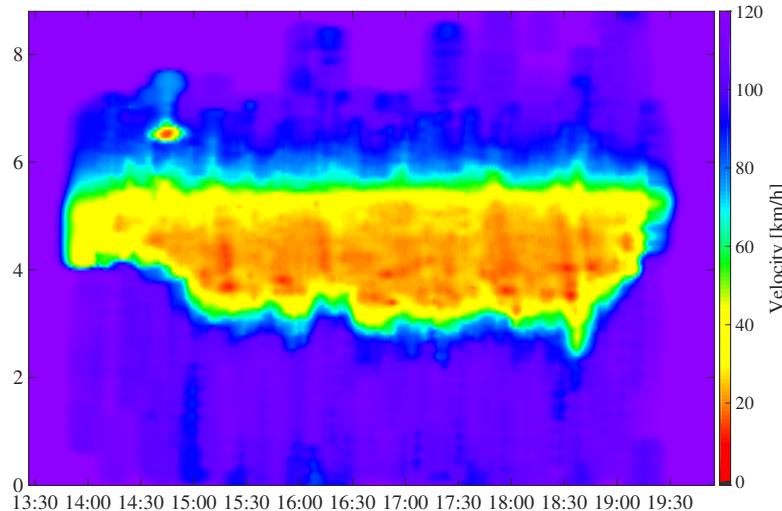
Figure 4.21 shows a scenario where the isotropic smoothing method outperforms the PSM. It is a synchronized flow congestion pattern with a relatively short length of 2-3 km and a temporal width of several hours. The velocity varies only slightly and no noticeably moving jams are visible. Both algorithms reconstruct well the stationary shape of the pattern. The reason for the isotropic smoothing method to be approximately 10% more accurate seems to be the better averaging of the trajectory data in space-time. Due to the relatively low velocities, the PSM identifies emerging moving jams which are not clearly visible in the original data. Thus, it is slightly more inaccurate. Figure 4.22 depicts more congestion scenarios where the isotropic smoothing

Scenario V.

(a) Raw trajectory data



(b) Isotropic Smoothing Method



(c) PSM

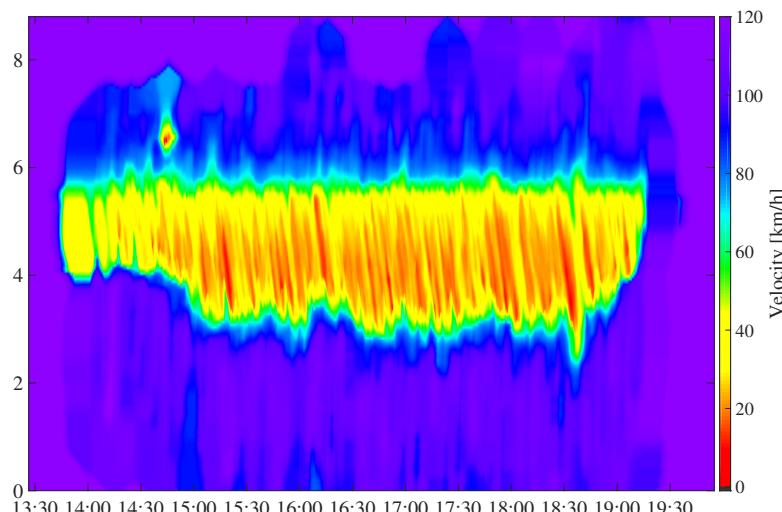


Figure 4.21. Comparison of velocity estimates of an isotropic smoothing and the PSM for a congestion that occurred on A96 in west-bound direction on August 1st 2014

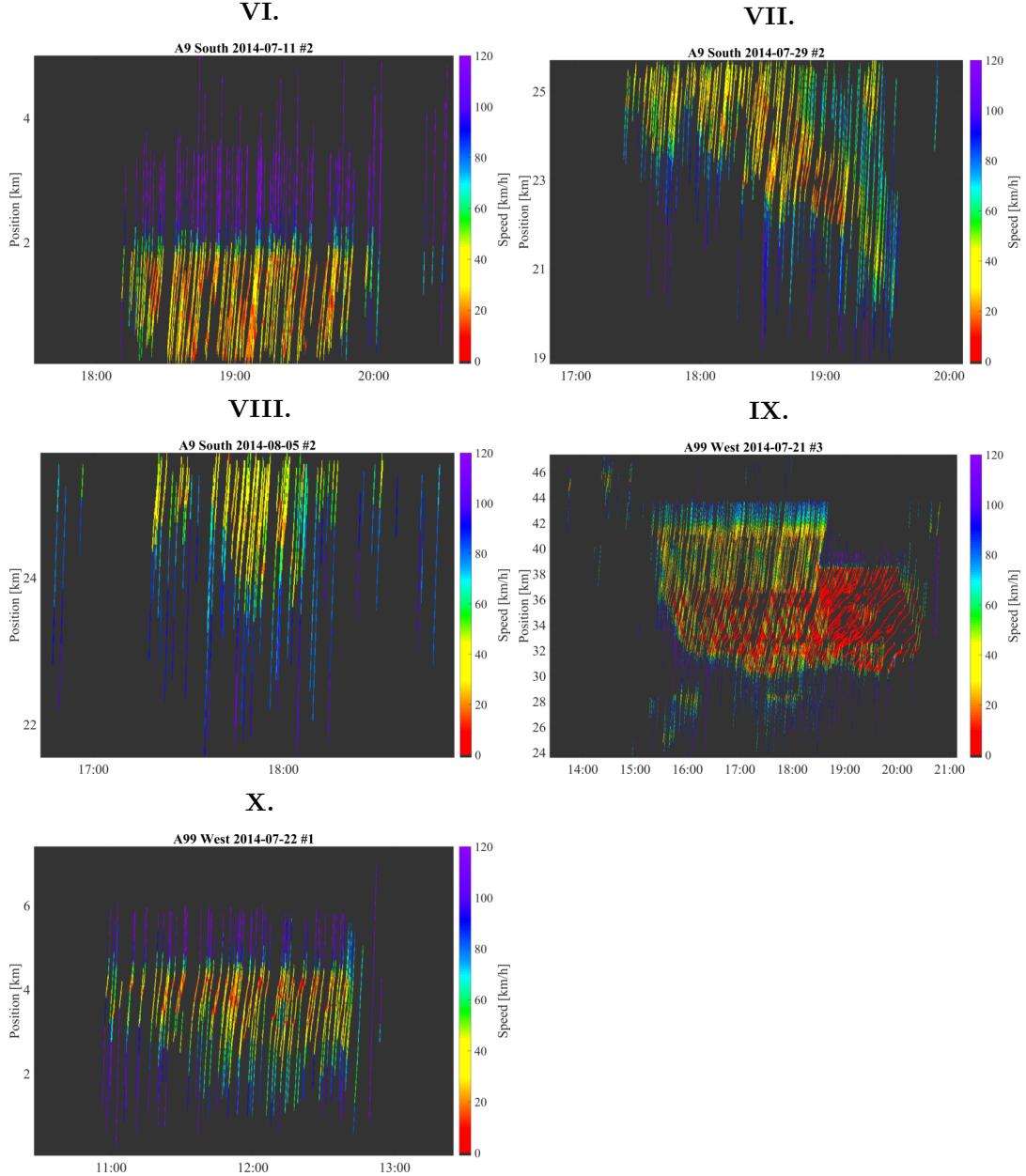


Figure 4.22. Congestion patterns that are reconstructed with similar or slightly better accuracy using an isotropic smoothing compared to the PSM

method and the PSM yield similar results, or the isotropic one is slightly more accurate. Four of five of these patterns represent minor congestion patterns that are rather stationary. Traffic velocities are rather homogeneous. Since both classes of algorithms estimate homogeneous synchronized flow phases similarly, the apparent similar accuracy results can be explained well. Not mentioned yet is the congestion on August 21st on A99 in westbound direction. Until 6pm traffic was in oscillating state (Oscillating Congested Traffic (OCT) (Helbing et al., 2009)), when an accident at kilometer 38 occurred. Due to closed lanes and significantly reduced road capacity, the pattern changed into homogeneous congested traffic (HCT), also called mega-jam (Kerner, 2009). The isotropic

smoothing method reconstructs this pattern more accurately than the PSM. A detailed discussion of this pattern is given in section 4.4.5.

The comparison of the PSM with the GASM results in the relative errors depicted in Figure 4.23. It can be noticed that:

1. The error function is less volatile than the one in Figure 4.18. This can be explained with the higher degree of similarity between PSM and GASM. This higher similarity also explains that
2. the difference of relative errors is generally smaller. Values range between 0.7 and 1.15. Still, most scenarios are reconstructed more accurately with the PSM than with the GASM.
3. There exist a few situations where the GASM catches up with the PSM or is more accurate.

Figure 4.24 depicts one example pattern where the PSM results in significantly lower estimation errors. Several smaller synchronized flow patterns stick to bottlenecks at kilometers 14, 18 and 22. In addition, at 8am an accident occurred at kilometer 12 which resulted in a relatively severe congestion with rather homogeneous traffic speeds. In this pattern a few narrow moving jams emerged that propagated upstream, but which did not develop into full WMJs. All of these patterns consist partly or completely of stationary congestion patterns. As discussed in previous sections and illustrated in Figure 4.14 the GASM is not able to reconstruct stationary patterns accurately using FCD such that the PSM produces more accurate estimates. Figure 4.25 illustrates four other scenarios. The first pattern is a stationary synchronized flow pattern. Three others can be classified as GPs (Kerner, 2009) which comprise synchronized flow and WMJ phases.

Figure 4.26 shows one example where the GASM produces similarly accurate results as the PSM. The pattern is dominated by moving jams. Since the GASM reconstructs moving jams well, the PSM is not able to outperform the GASM in this scenario. Also two other congestion patterns as depicted in Figure 4.27 are less accurately, or similarly as accurately reconstructed using the PSM. The right one comprises several WMJs. The left scenario is known from the preceding comparison between PSM and the isotropic smoothing approach. Apparently, the PSM fails to reconstruct this pattern. The discussion in section 4.4.5 highlights this issue.

Concluding the quantitative comparison, from a total of 101 congestion patterns that occurred on the freeway network surrounding Munich, the PSM reconstructed 89 ones better (i.e. both error bounds were below 1.0), 7 similarly accurate (the value of 1.0 is

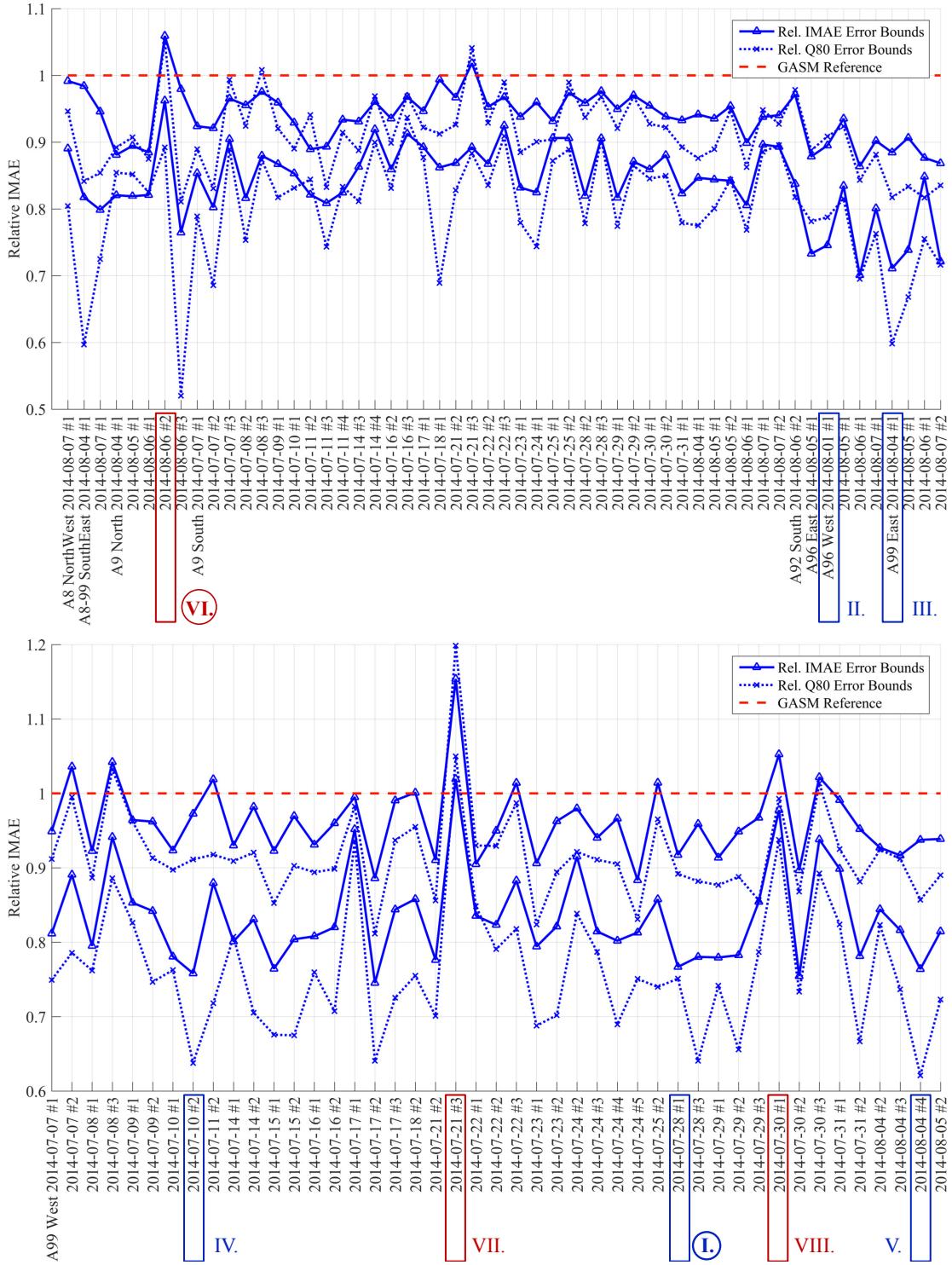
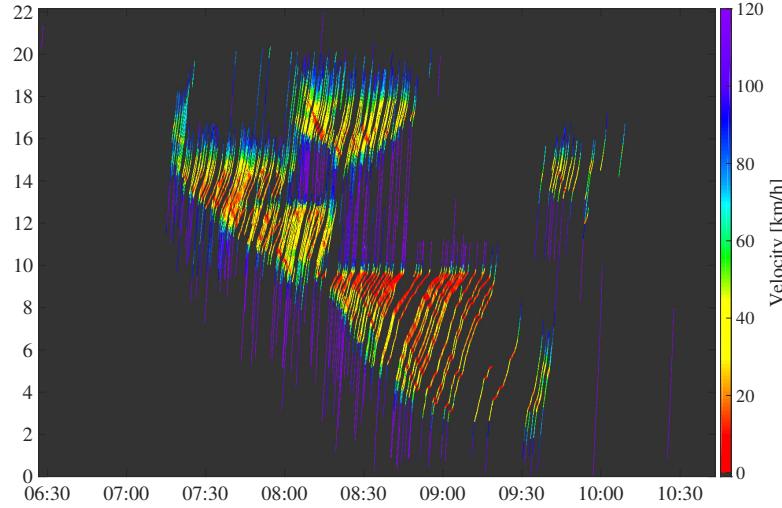


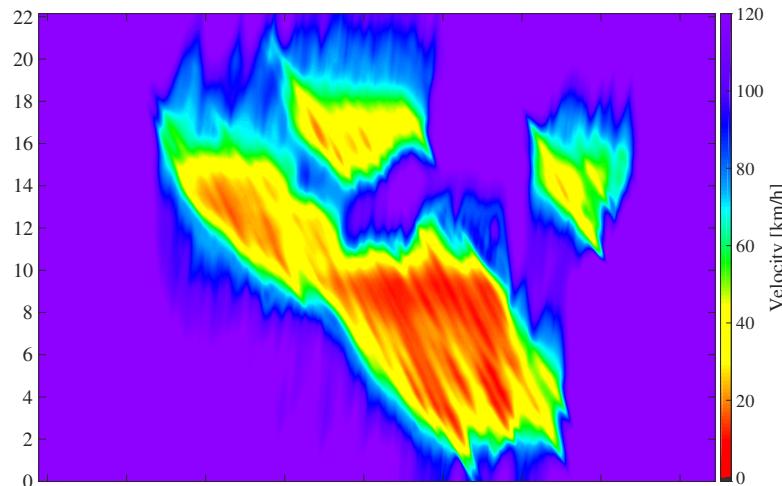
Figure 4.23. Relative \overline{IMAE} and its quantiles of the PSM compared to the GASM. Error bounds include the minimal and maximal relative error with respect to different parametrizations

Scenario I.

(a) Raw trajectory data



(b) GASM



(c) PSM

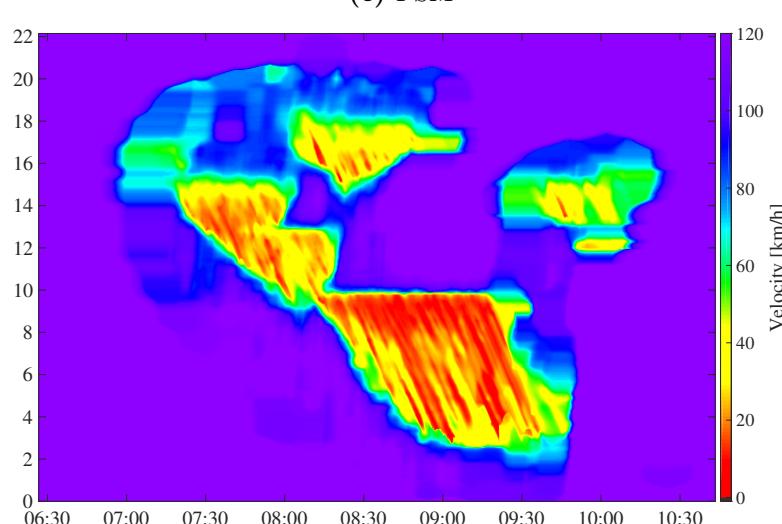


Figure 4.24. Comparison of velocity estimates of the GASM and the PSM for a congestion that occurred on A99 in west-bound direction on July 28th 2014

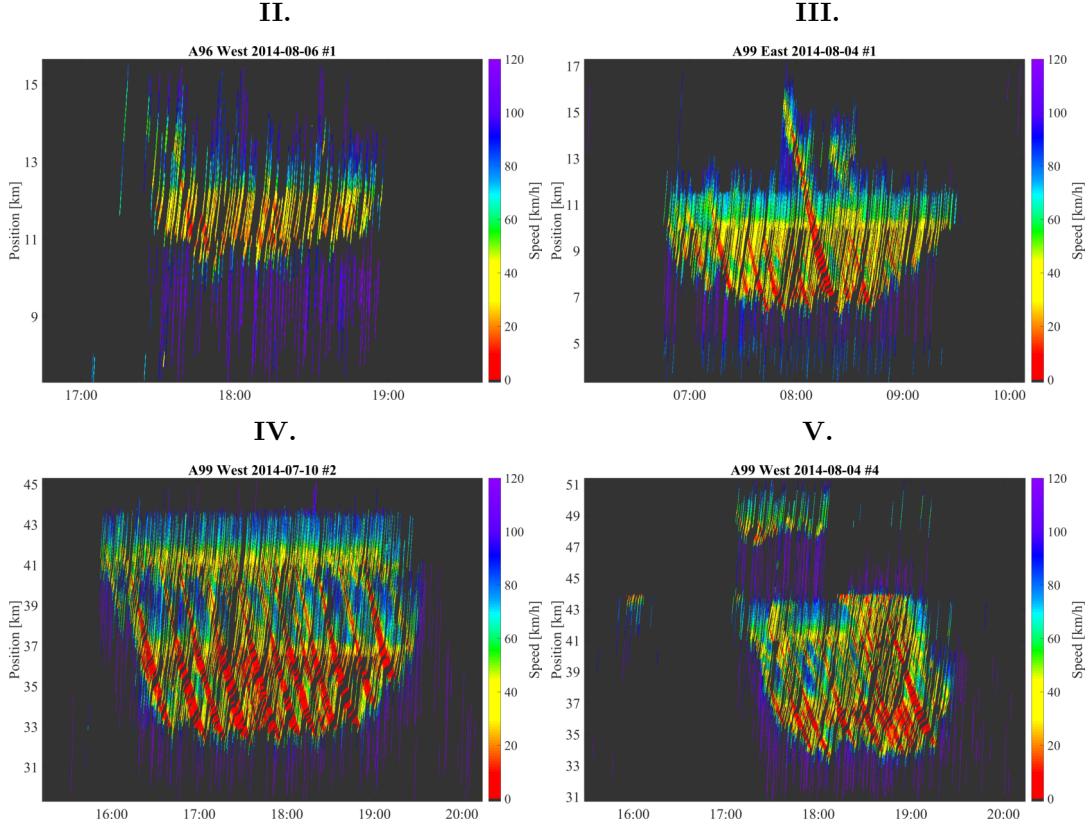


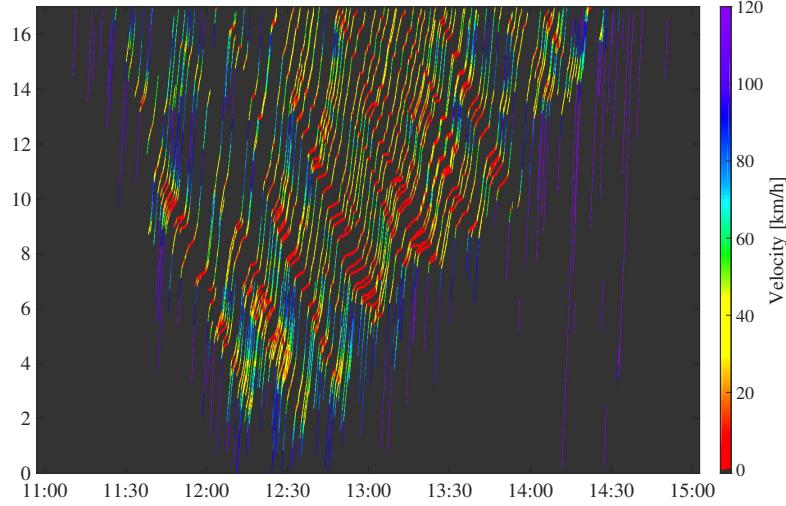
Figure 4.25. Congestion patterns that are reconstructed significantly better with the PSM compared to the GASM

included in the bounds) and 5 worse than the isotropic smoothing method. In average, the improvement varies between 25.7 % and 18.4 %. The 80% quantile is between 33.0 % and 26.2 %. Compared to the GASM, 89 scenarios were reconstructed more accurately, 11 similarly and 1 worse. Average improvements using the PSM are 16.3 % to 5.0 %, the 80% quantile is between 22.8 % and 8.5 %.

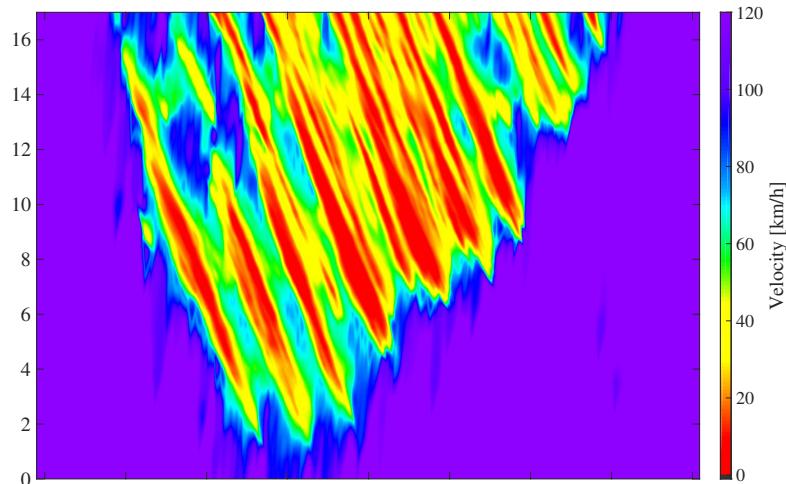
These results, based on a multitude of un-biased congestion patterns, show that the PSM manages to reconstruct most patterns more accurately than an isotropic smoothing method as well as the GASM. Even under consideration of different sets of parameters, taken into consideration using error bounds, the PSM outperforms the other algorithms in 89 of 101 scenarios.

Scenario VI.

(a) Raw trajectory data



(b) GASM



(c) PSM

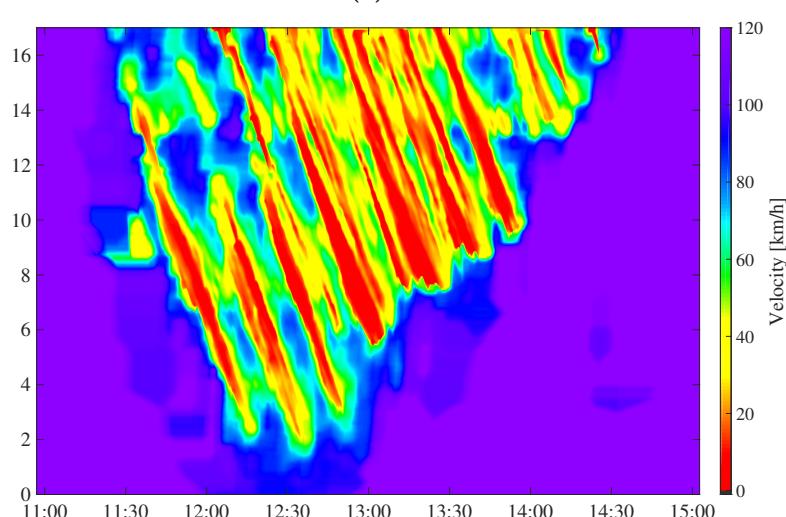


Figure 4.26. Comparison of velocity estimates of the GASM and the PSM for a congestion that occurred on A9 in north-bound direction on August 6th 2014

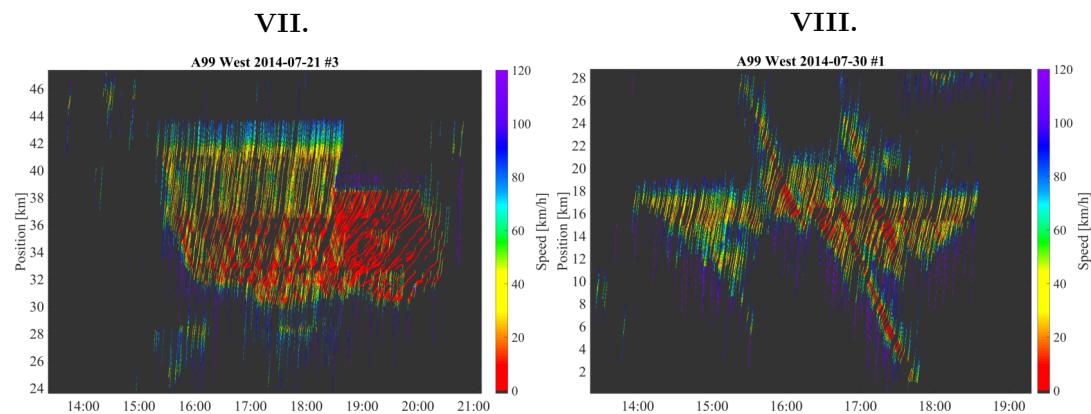


Figure 4.27. Congestion patterns that are reconstructed with similar or slightly better accuracy using the GASM smoothing compared to the PSM

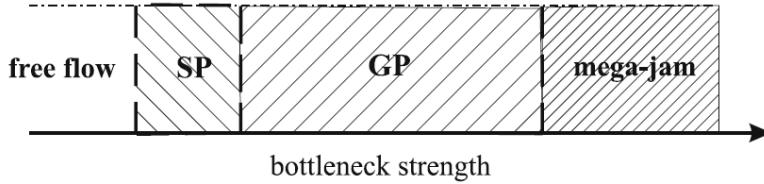


Figure 4.28. Congestion pattern emerging at isolated bottlenecks with respect to the bottleneck strength according to the Three-Phase traffic theory (Kerner, 2009)

4.4.5 Reconstruction of Mega-jams

The reconstruction of a mega-jam pattern with the PSM turned out to lack accuracy. This section provides a deeper analysis of this congestion pattern.

According to the Three-Phase traffic theory, there are three types of congestion patterns (Kerner, 2009): Synchronized Flow Patterns (SPs), GPs and mega-jams (also called HCT in other traffic theories (Helbing et al., 2009)). The emergence of a certain type of pattern depends on the bottleneck strength, i.e. the drop of the road capacity along a road segment (see Figure 4.28). If the bottleneck strength is low, SPs occur. With stronger bottlenecks, patterns develop into GPs which comprise synchronized flow as well as WMJ phases. If a bottleneck is very strong, e.g. due to a lane closure, a mega-jam pattern emerges. According to the Three-Phase traffic theory, a mega-jam is "a wide moving jam with an extremely great width growing continuously over time". It is further characterized as "a non-regular dynamic behavior of wide moving jams as well as random disappearance and appearance of the pinch region of synchronized flow within an GP upstream of the bottleneck", and "the merger of wide moving jams of the GP into a mega-jam" (Kerner, 2009). Thus, although this pattern is basically classified as a WMJ, it shows some irregularities: Instead of individual, distinct waves with constant downstream front velocity, a mega-jam consists of multiple WMJs which merge into one phase region. In contrast to a typical WMJ which emerges as a narrow moving jam in a pinch region (i.e. a synchronized flow phase) and develops slowly into a WMJ, there is no pinch region in a mega-jam. Rather, all moving jams originate at the bottleneck position. As a result, this congestion pattern shows a stationary downstream phase front as long as the bottleneck is active.²

Reconstructing the previously mentioned mega-jam pattern with the PSM results in a $IMAE(MD)$ depicted in Figure 4.29. Apparently, the error of the PSM exceeds the errors obtained by other methods for data coverages below 1 %. Only then, at a coverage greater than 2 %, the PSM catches up and reconstructs slightly more accurate. Figure

²Due to the described irregularities of the mega-jam pattern, this type of pattern is part of the discussion whether two or three states of traffic flow exist (Treiber et al., 2010b; Schönhof and Helbing, 2009; Kerner, 2009).

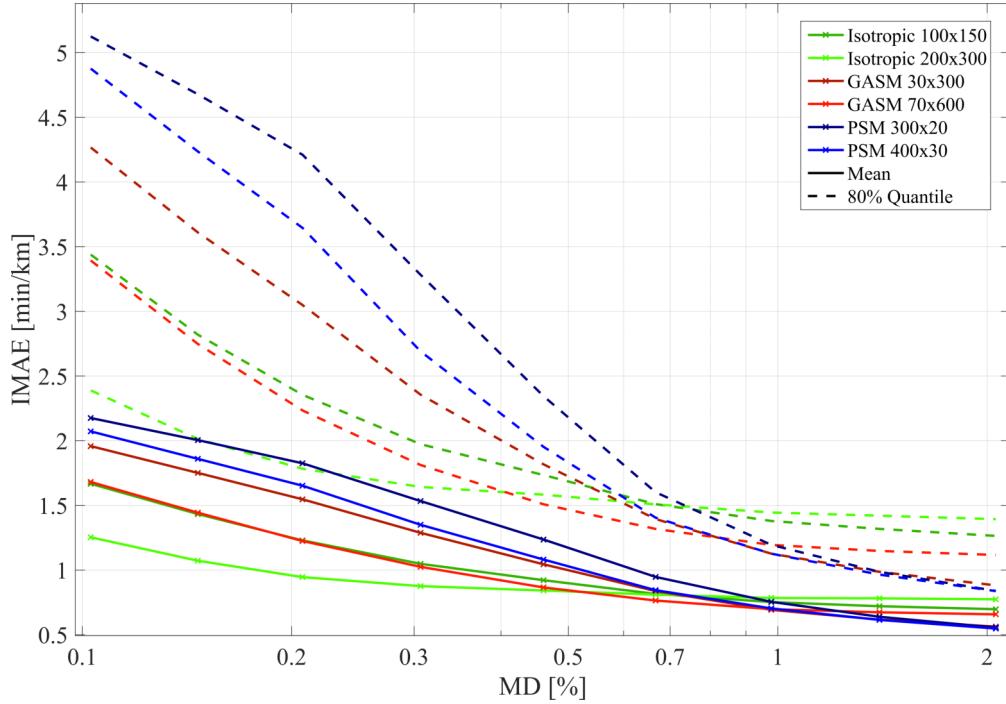


Figure 4.29. Mean IMAE and error quantile of an isotropic smoothing method, the GASM and the PSM when applied to a mega-jam pattern

4.30 illustrates the results of the estimation process with a zoom on the mega-jam. It shows all collected data, a subset of trajectories that correspond to a mean data coverage of 0.5 % and the estimated velocities using the isotropic smoothing method, the GASM and the PSM. The isotropic smoothing method reconstructs this stationary pattern quite well. It smooths all velocity data and produces a homogeneous stationary pattern. The GASM, using the shock-wave characteristic kernels, manages to smooth data widely and reconstructs most of the congested region. Though, part of the congestion is propagated beyond the bottleneck in downstream direction. The PSM reconstructs the downstream front well, however, larger regions of the mega-jam are estimated as free-flow. These wrongly estimated regions result in a large estimation errors.

The reason for the PSM's inability to reconstruct this mega-jam as accurately as other methods can be accounted to its strict accordance to the typical properties of phases as described by the Three-Phase traffic theory. As described earlier, typical properties of synchronized flow phases are its stationary character, while WMJs are shock waves with constant downstream front velocity. The congestion patterns evaluated in section 4.4.4 showed that in most cases this approach results in a gain in accuracy. However, the properties of mega-jams diverge from typical WMJ characteristics ("a non-regular dynamic behavior of wide moving jams"). The PSM is constructed and parametrized for typical phase characteristics, i.e. that temporal widths of WMJs are low (Kerner and

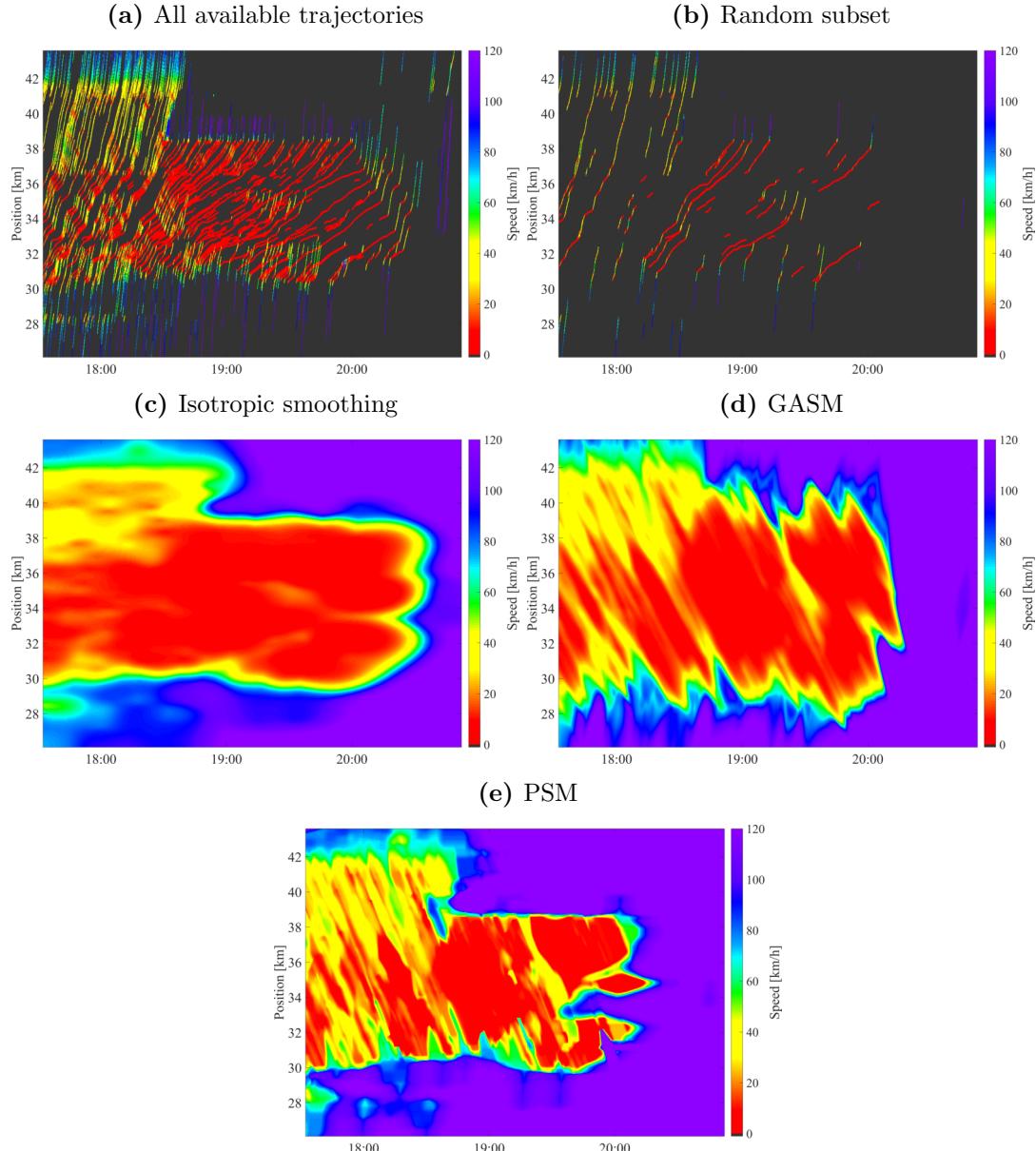


Figure 4.30. Complete dataset (upper left), training data (upper right) and velocity estimates produced by the isotropic smoothing, the GASM and the PSM for a mean data density of 0.5 %

Rehborn, 1996). Therefore, it fails to reconstruct the extensive space-time region of the mega-jam.

In this study the mega-jam pattern occurred only once in 101 patterns. Also studies by other researchers support the hypothesis of a rare observation of this pattern (Schoenhofer and Helbing, 2007). Though, if the accurate reconstruction of this pattern is especially important, it is possible to adapt the PSM in order to improve the estimation accuracy. One way is to optimize its parameter for this pattern. While that would increase

the estimation accuracy for mega-jam patterns, it is likely that the accuracy in reconstructing other more typical patterns would decrease. Since most analyzed patterns are reconstructed well with the proposed set of parameters, this is a less recommended way. Another possibility is to integrate empirical features of mega-jams and the position of the bottleneck into the phase calculation: If its existence and position is known (e.g. because a lane closure has been reported by public authorities or a data-driven approach detected a strong bottleneck) and a mega-jam pattern starts to develop, a queuing congestion upstream the bottleneck is likely. This queuing pattern is in phase J and will not dissolve until upstream and downstream front meet. Additionally, it is very unlikely that any free flow or synchronized flow phase will occur in-between these phase fronts. Such assumptions could be integrated into the PSM using a queue model which tracks the upstream and downstream congestion fronts and assigns enclosed space-times a high phase probability. This approach would be a minor extension and could be integrated easily into the PSM framework in future works.

4.4.6 Sensitivity Analysis

The PSM involves several parameters (see Table 4.2) that need to be set. In order to understand better the influence of certain parameters on the estimation accuracy, in this section the sensitivity of the estimation accuracy with respect to different parameters is presented. This facilitates the adoption and parametrization of the PSM to other scenarios and allows for efficient optimizations of parameters.

Some of the parameters, such as the smoothing directions v_p^{dir} and $v_p^{dir,H}$ as well as the velocity thresholds between the phases are motivated by empirical traffic characteristics. Therefore, in the following, they are seen as constants. The parameters time headway T_H and minimal length x_0 have been subject of several studies publishing distributions of observed values in real traffic (Krbalek et al., 2001; Brackstone et al., 2002). Though, the assumed microscopic car-following model in order to determine $\Psi(t, x)$ is quite simple. If it turns out that parameters T_H and x_0 influence the estimation result significantly, more sophisticated models might be necessary. Finally, the kernel parameters τ_p , τ_p^H , σ_p and σ_p^H do not correspond to any traffic constant. Up to now, they are set in a trial-and-error procedure. Especially for these parameters it is important to know, which one influences the estimation accuracy most, and which ones play a minor role when maximizing the estimation accuracy.

A method that is commonly used to quantify the sensitivity of an input parameter on the cost function is the Variance-Based Sensitivity Analysis (VBSA) (Saltelli et al., 2007). Consider a function $Y(X)$ with $X \in \mathbb{R}^d$ as input vector. The first-order sensitivity S_i

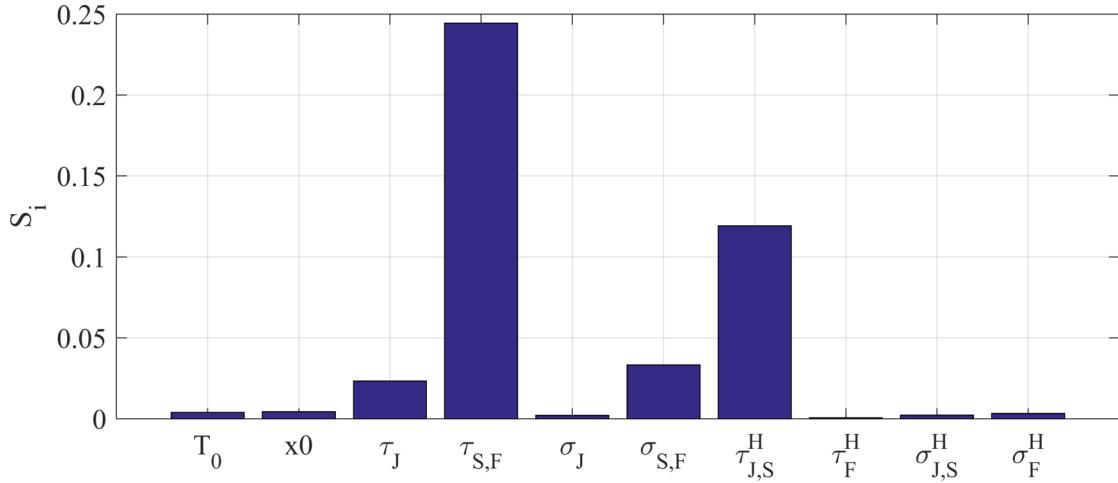


Figure 4.31. First-order sensitivity indices of the parameters of the PSM

measures the effect of varying one input dimension X_i on the cost function Y :

$$S_i = \frac{\text{Var}(\mathbb{E}(Y|X_i))}{\text{Var}(Y)} \quad (4.49)$$

where $\text{Var}(\cdot)$ denotes the variance and $\mathbb{E}(\cdot)$ the expectation of a random variable. Note that $\sum_i S_i < 1$ in case there are interaction terms among the input variables (Saltelli, 2007). The computation of these indices usually requires a vast number of samples of Y (Cukier et al., 1978). In order to reduce this number for time-consuming models a technique called Fourier-Amplitude Sensitivity Test (FAST) is proposed in (Saltelli et al., 1999; Cannavó, 2012).

Applying the FAST to the PSM requires to define a range of valid values for each parameter. They are set as follows: T_H is varied in the range of [0.5, 3], x_0 in [2, 20], all τ values in [20, 500] and all σ in [100, 1000]. The method is applied to the scenario described in section 4.4.3 using the depicted training and test set. All in all, 20,000 model evaluations with varying parameters are done, the *IMAE* is computed and the sensitivities are determined.

Fig. 4.31 shows the resulting first-order sensitivities. Many parameters appear to have minor influence on the estimation accuracy, for instance, the parameters of the assumed car-following model applied for the computation of space-times Ψ . This low sensitivity supports the hypothesis that a simple model is a sufficiently accurate approach for the needs of the PSM. Furthermore, several kernel parameters such as σ_J , τ_F , $\sigma_{S,J}^H$ and σ_F^H are less influential. The highest impact have parameters $\tau_{S,F}$ and $\tau_{J,S}^H$. When applying the PSM to other scenarios, an optimization of these two parameters is likely to produce the greatest gain in accuracy.

Overall, none of the 20,000 model evaluation with strongly varying input parameters resulted in a failure of the model. Thus, since the PSM returns velocity estimates even with parameter sets that deviate strongly from the optimum this method can be described as robust with respect to its parametrization. Contrary, simulation models based on the LWR model diverge if for example the Courant-Friedrich-Levy condition is not satisfied, which results in non-physical velocities (Knoop and Daamen, 2016).

Note that this sensitivity analysis has been performed with one scenario with fixed training and test set. It is likely that other setups will result in slightly different sensitivities. Furthermore, it appears that $\sum_i S_i \approx 0.5 < 1$. As described earlier, it means that there are interaction terms between several parameters. The accurate calculation of interaction terms in a 10-dimensional parameter space requires vast numbers of samples (Saltelli et al., 2007). In order to generate that many function evaluations, a significant amount of computational resources is needed, and a prior optimization of the code is recommended. This could be subject of future work.

4.4.7 Run-time Analysis

Besides accuracy, efficiency is an important requirement of a traffic state estimation method. Especially real-time applications require an algorithm that processes sensor information of a large network as quickly as possible in order to broadcast up-to-date traffic information. But also for processing greater amounts of historical sensor data, efficient estimation algorithms are necessary. In order to provide insights into the efficiency of the PSM, this section presents its computational complexity and average run-times required for scenarios with respect to varying network sizes.

When applied to a space-time domain discretized into N_t cells in time and N_x cells in space, an implementation of the PSM comprises element-wise matrix operations and 2D convolution processes. With respect to the run-time, it is irrelevant whether grid cells contain measurements or not. Element-wise matrix operations have a complexity of $\mathcal{O}(N_t N_x)$. As shown in detail in (Schreiter et al., 2010; Schreiter, 2012), the convolution process can be implemented efficiently using the FFT (Cooley and Tukey, 1965). The resulting complexity is $\mathcal{O}(N_t N_x \log(N_t N_x))$. In complexity considerations, the term with the highest order dominates. With $N_{tot} = N_t N_x$, denoting the total number of grid cells in the domain, the overall complexity of the PSM is:

$$\mathcal{O}(N_{tot} \log N_{tot}). \quad (4.50)$$

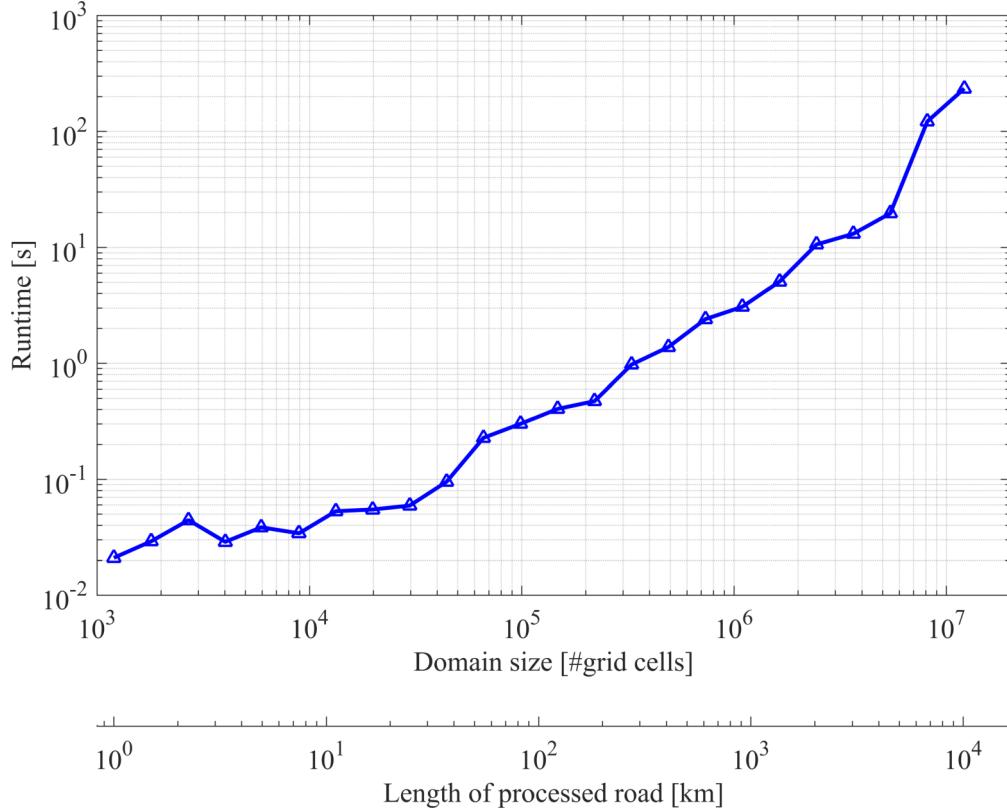


Figure 4.32. Run-time of the PSM with respect to an increasing domain size

This class of complexity is close to linear complexity. Furthermore, when it is applied to increasing problem sizes (number of grid cells), the required number of computations increases only slightly stronger than linearly. If hardware-specific effects such as memory-management are neglected, then also processing times are expected to grow approximately linearly.

The computational complexity is important in order to estimate whether an algorithm is eligible to be applied to huge problems, but does not give insight into the absolute computation times. Therefore, in the following, the computation times for a real-time scenario with increasing domain size are presented. We assume a domain with $\Delta T = 30$ s and $N_t = 60$ which describes a time domain of 30 min duration. ΔX is set to 50 m, and the number of cells in space dimension N_x is varied. We assume that the PSM is called iteratively (every time step) in the real-time case, such that the kernel functions can be preprocessed. The time that it takes to process a matrix with velocity data V_{FCD} into the velocity estimate V_E is measured. Each measurement is repeated for 50 times to ensure robustness of the results. The computations are run on a notebook with i7-4800 processor with 2.7 GHz and 8 GB memory.

Fig. 4.32 illustrates the processing time with respect to an increasing problem size in a

logarithmic scale. As expected from the complexity considerations, run-time increases approximately linearly, with slight deviations at low run-times. Noticeable is that at a cell count of approximately $6 \cdot 10^6$ (5000 km) a significant jump of computation times occurs. The reason is that the required memory exceeds the available memory of the notebook, which causes time-consuming memory management measures of the operating system. Moreover, although the mean run-time of 50 iterations is considered, the curve is still noisy. This stems from the FFT implementation, which expands matrices to the next higher value of 2^{N_t} and 2^{N_x} cells. Therefore, given a continuously growing number of input cells, the resulting processing time is a non-smooth function.

To conclude, the PSM allows to process traffic data in network sizes of up to 1000 km in about 3 s on a standard notebook using an un-optimized implementation. Further optimizations of the code as well as distributed computing techniques will enable to handle far larger network sizes in shorter times. For most real-time applications this processing speed is expected to be satisfactory.

4.5 Conclusion and Outlook

In this chapter, a novel freeway traffic speed estimation algorithm called PSM was developed and its accuracy, sensitivity to parameters and its efficiency were evaluated.

First, a new concept to represent FCD in time and space was introduced that allows to describe the level of information a measurement holds. Afterwards, the two-step approach of the PSM was motivated based on the Three-Phase theory. In the first step, phase regions of free flow, synchronized flow and WMJ are reconstructed. In the second, phase-dependent velocity estimates are computed and aggregated into a final velocity estimate. In a subsequent qualitative analysis of the estimation results of one typical congestion pattern, the strengths and weaknesses of the PSM in comparison with two other state-of-the-art approaches were analyzed. Overall, the PSM outperformed the other approaches.

In order to generalize the results and assess the estimation accuracy of the PSM in comparison with the other algorithms, an extensive evaluation using 101 congestion patterns, which occurred in the freeway network surrounding Munich, was conducted. First a novel methodology was developed that seeks to eliminate the influence of the available amount of data on the overall accuracy of an algorithm. The so-called mean data coverage was introduced which allows to describe the level of information that was available for speed estimation. This allowed to compare the estimation accuracy of several variants of an isotropic smoothing method, the GASM and the PSM. Results

showed that the PSM was in average 18.4 % to 25.7 % more accurate than the isotropic smoothing method and 5.0 % to 16.3 % more accurate than the GASM. Only the reconstruction of a mega-jam patterns lacked substantial accuracy. This pattern was analyzed thoroughly in a subsequent discussion. The results of a first-order sensitivity analysis of the parameters of the PSM on the estimation accuracy were investigated. As a result, two parameters turned out to have the greatest impact. This finding allows to adapt the PSM quickly to new scenarios and achieve accurate results by optimizing only two parameters. Finally, the theoretical complexity and the actual run-times of the PSM were analyzed. Its efficiency is shown in a real-time scenario where the mean processing time is measured depending on the considered road length. For instance, on a standard notebook, network sizes of more than 1,000 km can be computed in about 3 s.

To sum it up, the PSM fulfills the requirements of a practice-ready traffic speed estimation algorithm (section 2.3): It is more accurate than other methods, it is sufficiently efficient to be applied in large-scale networks, it is relatively simply to parametrize to a new setup and it proved to be robust with respect to different parameter sets. Other features are, that it allows for a fusion of different data sources and that it provides a level of trust for each computed velocity value.

There are several aspects that could and should be addressed in future works. One concerns a more sophisticated treatment of mega-jams in order to reconstruct this type of congestion pattern more accurately. Another one is that the PSM is limited to the estimation of traffic speeds on freeways. Since traffic dynamics in urban networks have different characteristics, the PSM needs to be adapted/extended in order to be able to reconstruct traffic speeds accurately on urban roads. Finally, for now, the PSM is able to process velocity data. Due to increasing digitalization and decreasing communication costs in the future more types of data, such as extended vehicle sensor information, accurate density and flow from road side units, lane-information etc., will be available. A comprehensive description and integration of various data types into the PSM framework provides great potential to further increase the estimation accuracy.

Chapter 5

Forecasting Congestion Fronts using FCD and Flow Data

The PSM as described in chapter 4 is a general method that can be used to estimate traffic speeds on freeways given various types of speed data as well as other measurements that support the identification of traffic phases. Its strength are the accuracy, robustness and the efficiency which allows real-time applications (section 4.4.7). Though, it does not provide traffic state forecasts. Short-term traffic state forecasts constitute valuable information for various traffic-related applications (Vlahogianni et al., 2014). Systems such as travel time predictions as well as in-vehicle tail of congestion warnings increase the calculability and safety of individual transportation. The overall efficiency and safety benefit from accurate short-term traffic speed forecasts using effective control strategies such as VSLs (Ackaah et al., 2015) or ramp-metering (Bogenberger and May, 1999).

Especially WMJs are a significant hazard for travelers: Vehicles entering the congestion often need to decelerate strongly. Severe accidents, caused by inattentive drivers, occur frequently. Therefore, effort is done in order to alert the driver and make him slow down decently before arriving at the congestion front. However, time is needed in order to generate and apply a congestion front warning and a driver needs to be warned ahead of time. Therefore, short-term congestion forecasts are required. Nevertheless, the facts that WMJs originate stochastically, the velocity of their upstream fronts varies and they dissolve as soon as upstream and downstream front meet, make it challenging to provide reliable and accurate short-term front forecasts.

The Three-Phase theory provides a few heuristics on congestion front propagation speeds: The downstream front of WMJs propagates upstream with nearly constant velocity and the downstream front of a synchronized flow phase typically sticks to a bottleneck and is therefore stationary. The speed of the danger-prone upstream fronts depend on the flow

and density difference of adjacent phases. Speed data alone does not allow for feeding a physical model, which limits the accuracy of a forecast. However, given heterogeneous types of data such as macroscopic flows and densities or aggregated origin-destination information of individual travelers, more sophisticated forecasts can be determined. A fusion of different types of data constitutes an important aspect of future ITS (see (Faouzi and Klein, 2016)).

This chapter proposes a new method to forecast congestion fronts of WMJs based on FCD fused with detector data for a short time horizon of up to 10 min. In contrast to other methods which usually require data to be collected at fixed positions in fixed time intervals, the proposed method allows data to be sparse in time and space. The fusion of FCD with detector data seeks to combine the strengths of both data sources: The high spatio-temporal coverage of FCD that provides accurate traffic speed estimates and the flow data collected by loop (or other stationary) detectors that is used for congestion front prediction.

The following section briefly summarizes related work. Next, in section 5.2 a simple, yet robust and flexible approach is described that extends the PSM with a short-term congestion front forecast using sparse flow data. Three variants of the approach are described as candidates for further analysis. In section 5.3 the results of an evaluation using the data of a real freeway congestion are presented. Therefore, first a quality metric that measures the prediction accuracy under consideration of a varying prediction horizon is proposed. Second, the accuracy of each variant and a naive predictor is assessed and, finally, the results are compared and discussed. Section 5.4 concludes this chapter and proposes further directions.

5.1 Related Work

Many methods have been developed that predict traffic conditions on freeways for a short time horizon (see section 2.3). Most approaches that are based on analytical models apply a first or second order CTM model using loop detector data and apply data assimilation techniques such as Kalman filters (Lighthill and Whitham, 1955; Richards, 1956; van Lint and Djukic, 2014; Wang and Papageorgiou, 2005; Yuan et al., 2012). A different approach called ASDA/FOTO reconstructs space-time regions of free flow, synchronized flow and WMJ and applies the shock wave equation of traffic flow in order to forecast phase fronts (Kerner, 2004). All of these approaches develop dedicated traffic flow models using detector data. However, they do not allow to incorporate FCD which, due to its high spatio-temporal resolution, is a valuable source of traffic information. Other approaches published in (Bekiaris-Liberis et al., 2016; Work et al., 2010; Work

et al., 2008; Work et al., 2009) integrate probe velocity measurements into an LWR based model. These models convert speed measurements into density or flow estimates using a FD in order to estimate and forecast traffic conditions on a road segment. Disadvantageous is that these methods require complete information at all boundaries of the space-time domain.

To summarize, there are sophisticated CTMs using loop detector data which possibly enrich these measurements with probe data. Provided FCD is rather used as a supplement than as a source of high quality data. These methods rely strongly on flow and density data provided at regular time intervals as well as assumed boundary conditions. This limits the applicability of many methods since boundary conditions are often unknown in practice. Additionally, the accuracy may be low if detector spacings are large.

The present approach seeks to overcome these issues. The idea is to first utilize high resolution FCD in order to identify current traffic phases. Next, these phase regions are used to estimate phase flows and densities from sparse sensor data using phase-characteristic smoothing operations. Subsequently, resulting traffic conditions feed a physical phase front propagation model. In comparison to existing methods, it exploits the high resolution of FCD and does not require boundary conditions at all. Furthermore, it seeks to be efficient, which enables a real-time application, and flexible due to its applicability to heterogeneous sensor data.

5.2 Prediction Model

The idea of the present prediction model applies ideas of different methods such as the ASDA/FOTO model, the GASM and the LWR shock wave equations. Four steps are performed: First, traffic speeds up to the current point in time are estimated using the PSM applied to all available FCD. Second, current phase fronts are identified and phase-dependent estimates of the phase flows and densities are computed using mixes of detector and FC data and phase-characteristic smoothing operations. Third, upstream flows are predicted in time and space. Fourth, phase front propagations are simulated over time depending on predicted flows and densities of adjacent phases.

The following three parts introduce the model: The first describes the definition of a congestion front as well as the propagation of phase fronts according to the shock wave formula. The subsequent two sections describe the estimation procedure of flows and densities of adjacent phases.

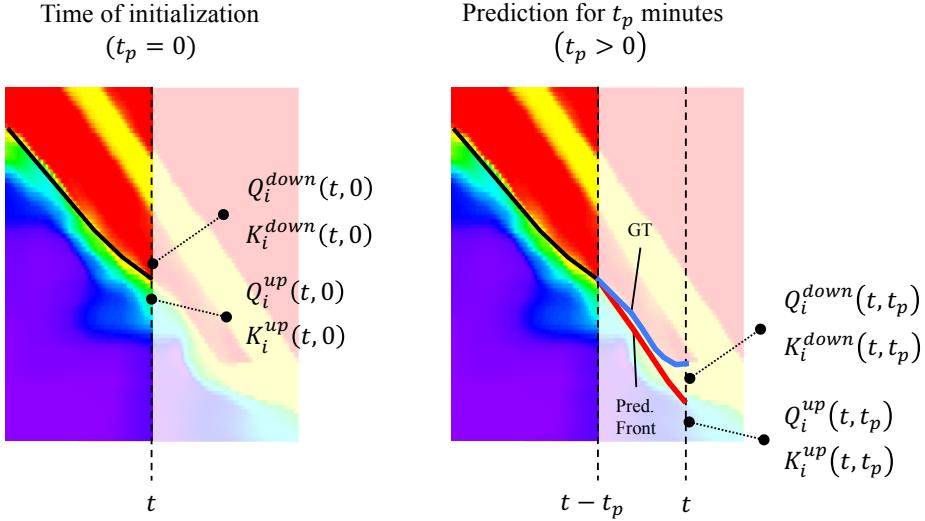


Figure 5.1. Schematic illustration of the front prediction

5.2.1 Phase Front Propagation

Let $V(t, x)$ be the macroscopic traffic speed at time t and at position x on a road segment of length L observed during time interval $[0, T]$, such that $t \in [0, T]$ and $x \in [0, L]$. Given a high data coverage of FCD $V(t, x)$ can be estimated accurately using a state-of-the-art traffic speed estimator such as the PSM described in chapter 4. The positions of the GT upstream jam fronts $X_{GT}^{up}(t)$ are defined as the positions where the traffic speed undergoes a critical velocity of v^{thres} :

$$X_{GT}^{up}(t) := \{x : V(t, x) = v^{thres}, \frac{dV(t, x)}{dx} < 0\}. \quad (5.1)$$

Accordingly, the downstream fronts are defined as:

$$X_{GT}^{down}(t) := \{x : V(t, x) = v^{thres}, \frac{dV(t, x)}{dx} > 0\}. \quad (5.2)$$

Let t_p be the predicted time, i.e. the time that has passed since a front has been initialized with the GT. Then, the goal of the proposed forecast method is to process all given data that is available up to the time $t - t_p$ and provide an estimate of all upstream jam fronts $X_{E,i}^{up}(t, t_p)$, where index i denotes the i -th front in ascending order of x :

$$X_{E,i}^{up}(t, t_p) = X_{GT,i}^{up}(t - t_p) + \int_{t-t_p}^t \dot{X}_{E,i}^{up}(\hat{t}, \hat{t} - (t - t_p)) d\hat{t}. \quad (5.3)$$

This propagation equation is valid as long as the upstream and downstream front do not meet:

$$X_{E,i}^{up}(t, t_p) < X_{E,i}^{down}(t, t_p). \quad (5.4)$$

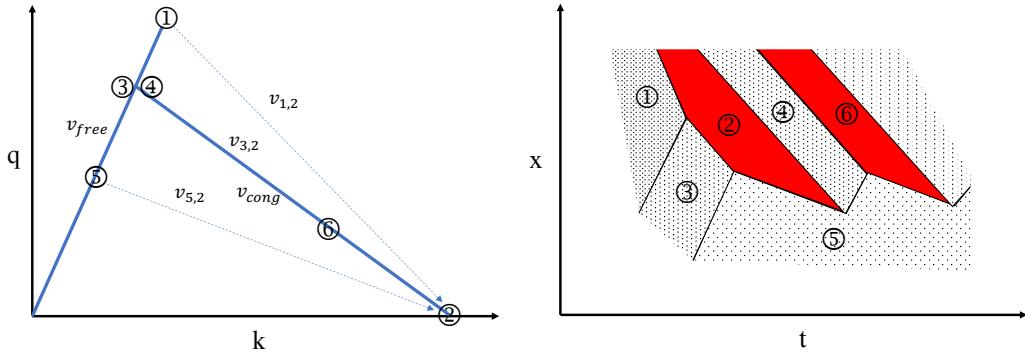


Figure 5.2. Fundamental diagram and corresponding space-time regions with phase fronts and front propagation speeds (compare to (Treiber and Kesting, 2013))

If for any i the condition is violated, both, $X_{E,i}^{up}$ and $X_{E,i}^{down}$ are removed from the sets of currently active fronts. The propagation speed of a front is computed with the well-known shock wave formula (Treiber and Kesting, 2013; Kerner, 2009; Richards, 1956; Lighthill and Whitham, 1955):

$$\dot{X}_{E,i}^{up}(t, t_p) = \frac{Q_i^{down}(t, t_p) - Q_i^{up}(t, t_p)}{K_i^{down}(t, t_p) - K_i^{up}(t, t_p)} \quad (5.5)$$

where Q_i^{down} and Q_i^{up} denote the outflow and inflow into a congestion front, and K_i^{down} and K_i^{up} the traffic density slightly downstream and upstream of the front i (see Figure 5.1).

Figure 5.2 visualizes the front propagation speeds for several pairs of flow and density of adjacent traffic phases computed with the LWR shock wave formula. Since the prediction model is focused on the more dangerous transitions from free to WMJ state, in the visualization mainly examples of free flow and WMJ are given. Still, the potentially wide scattered states of the synchronized flow phase follow the same rules. As described in section 2.1, two distinct velocities have been observed extensively in real traffic patterns: $v^{free} \approx 80 \text{ km/h}$ (Kerner et al., 2004; Kerner, 2009; van Lint and Hoogendoorn, 2009; Treiber and Kesting, 2013) and $v^{cong} \approx -15 \text{ km/h}$ (Newell, 1993; Kerner, 2009; Treiber et al., 2010b). v^{free} has been found to be the average speed of shock waves in free flow, and v^{cong} in congested flow. Transitions between free and WMJs state may have different propagation speeds. For instance, if traffic flow is in free and unstable state (close to q_{max} , (Kerner, 2009)) and streams into a WMJ phase, the (absolute) front speed is higher than v^{cong} . Similarly, if the in-flow into a highly congested state is relatively low, the (absolute) front speed is lower. The other way around, the speed of the downstream front of a WMJ adjacent to free flow varies merely. It has been observed to be quite constant close to v^{cong} (Kerner, 2004).

In order to determine these quantities in practice, measurements provided by e.g. loop detectors are available that are reported sparsely in time and space. Flow data is represented as a set of tuples $\mathcal{Q} = \{(t, x, q)_1, \dots, (t, x, q)_{N_q} : t_j \leq t - t_p, j \in 1, \dots, N_q\}$. According density values (usually determined as q/v) follow the same notation: $\mathcal{K} = \{(t, x, k)_1, \dots, (t, x, k)_{N_k} : t_j \leq t - t_p, j \in 1, \dots, N_k\}$. However, raw data does not correspond to the desired quantities. First, these measurements are available only up to the time of initialization $t - t_p$, but the forecast model requires predictive flows and densities in order to simulate a congestion front. Second, measurements are sparse in time and space. Though, flows and densities up- and downstream and in proximity of a congestion front are required. The following two sections describe the estimation process of flows and densities given sparse flow data.

5.2.2 Estimating Phase Flows

Continuous and predictive flows and densities are determined using traffic-characteristic spatio-temporal smoothing operations as described in section 4.3.2. In free flow, the kernel Φ_F^H is applied which models the propagation of information in free flow conditions. For the WMJ phase the kernel $\Phi_{J,S}^H$ is used. In addition, the probability $P_{\neg J} := 1 - P_J$ is defined which assigns each space-time (t, x) a weight according to the phase it belongs to. This results in the definition of the flow inside ($Q_J(t, x)$) and outside ($Q_{\neg J}(t, x)$) a WMJ phase:

$$Q_J(t, x) = \frac{\sum_{(t^*, x^*, q^*) \in \mathcal{Q}} \Phi_{J,S}^H(t - t^*, x - x^*) \cdot P_J(t^*, x^*) \cdot q^*}{\sum_{(t^*, x^*, q^*) \in \mathcal{Q}} \Phi_{J,S}^H(t - t^*, x - x^*) \cdot P_J(t^*, x^*)}. \quad (5.6)$$

$$Q_{\neg J}(t, x) = \frac{\sum_{(t^*, x^*, q^*) \in \mathcal{Q}} \Phi_F(t - t^*, x - x^*) \cdot P_{\neg J}(t^*, x^*) \cdot q^*}{\sum_{(t^*, x^*, q^*) \in \mathcal{Q}} \Phi_F(t - t^*, x - x^*) \cdot P_{\neg J}(t^*, x^*)}. \quad (5.7)$$

Thus, given sparse flow data up to time $t - t_p$, $Q_J(t, x)$ and $Q_{\neg J}(t, x)$ describe continuous flow estimates using the phase regions determined with the potentially high spatio-temporal resolution of FCD. Moreover, the smoothing process constitutes an extrapolation of data into the future. These estimates allows to set the upstream and downstream predictive flows.

Using this model, two phase transitions are possible: A transition from a non-WMJ phase to a WMJ, and from a WMJ to a non-WMJ phase. The in-flow $Q_i^{up}(t, t_p)$ for the first case is set as:

$$Q_i^{up}(t, t_p) := Q_{\neg J}\left(t, X_{E,i}^{up}(t, t_p)\right). \quad (5.8)$$

$Q_i^{down}(t, t_p)$ could be set analogously with $Q_J(t, x)$. Though, since the flow in the WMJ phase is very low, an extrapolation (i.e. forecast) of the current phase flow does not add

information. Therefore, Q_i^{down} is set to the flow at time of initialization:

$$Q_i^{down}(t, t_p) := Q_J \left(t - t_p, X_{E,i}^{up}(t - t_p, 0) \right). \quad (5.9)$$

It is possible to measure and set input and output flows for the second case as well. Though, it has been observed that the downstream front of a WMJ has a relatively constant speed. This constitutes already an accurate forecast, which a dynamic model can hardly outperform. Therefore, with respect to (Kerner, 2009), $\dot{X}_{E,i}^{down}(t, t_p)$ is set to:

$$\dot{X}_{E,i}^{down}(t, t_p) = v^{cong}. \quad (5.10)$$

5.2.3 Estimating Phase Densities

The estimation of phase densities is similar, but differs in one important aspect. Since loop detectors do not measure density, but it is determined as the quotient of flow and speed, traffic density can be biased. In free flow conditions macroscopic speeds and flows can be measured with high precision, such that also traffic densities are estimated accurately. In turn, in congested traffic conditions vehicle speeds and vehicle counts are low. The macroscopic flow and speeds, which constitute averages, are determined using only a few samples. Additionally, technical measurement errors influence speed values. Therefore, in congested traffic the accuracy of density data deduced from loops is limited.

Due to low speeds in WMJs phase regions this impacts strongly the forecast of severe congestion fronts. In order to analyze the effects of this issue and identify a possibly more accurate density estimator, in the following three variants to estimate the densities are contrasted.

The first variation, denominated as K-DET, smooths density quantities \mathcal{K} determined from detector data in the same way as flow data. The resulting smoothed and continuous functions $K_{\neg J}(t, x)$ and $K_J(t, x)$ are used to set the respective density values:

$$K_i^{down}(t, t_p) := K_J \left(t - t_p, X_{E,i}^{up}(t - t_p, 0) \right) \quad (5.11)$$

$$K_i^{up}(t, t_p) := K_{\neg J} \left(t, X_{E,i}^{up}(t, t_p) \right). \quad (5.12)$$

The second variant, denominated as K-MAX, is based on the ASDA/FOTO model (Kerner et al., 2004). In that model the authors precompute a density which represents the maximal density in congested traffic where vehicle velocities are very low. This

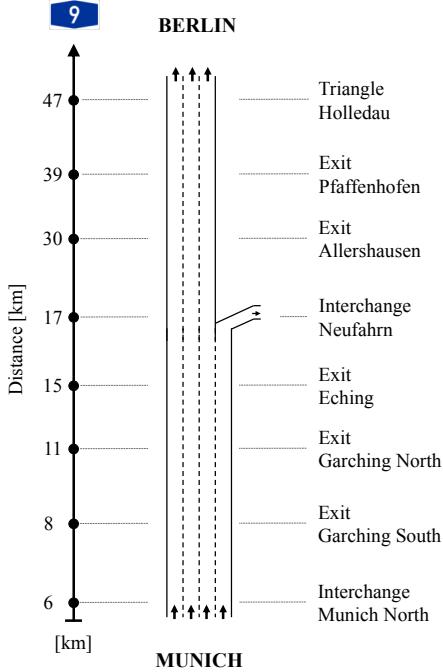


Figure 5.3. Schematic representation of the German freeway A9 in north-bound direction. The distance refers to the origin of the freeway.

approach is integrated into this framework by setting:

$$K_i^{down}(t, t_p) := k_{max} \quad (5.13)$$

and $K_i^{up}(t, t_p)$ similar to K-DET. In this case, k_{max} is set to 90 veh./km based on (Kerner et al., 2004).

The idea of the third variant, denominated as K-FCD, is that great part of the estimation error of the densities in the congested flow regime possibly stems from the inaccuracy of the traffic speed measurements. Since the velocity estimate $V_E(t, x)$, obtained from FCDs is expected to have a greater accuracy than a detector-based speed estimate, the overall accuracy of densities could increase if the FCD-based velocity estimate is utilized. $K_i^{down}(t, t_p)$ and $K_i^{up}(t, t_p)$ are computed using the velocity estimate at the time of initialization and the flow forecast described in the previous section:

$$K_i^{down}(t, t_p) := \frac{Q_J(t-t_p, X_{E,i}^{up}(t-t_p, 0))}{V_E(t-t_p, X_{E,i}^{up}(t-t_p, 0))} \quad (5.14)$$

$$K_i^{up}(t, t_p) := \frac{Q_{\neg J}(t, X_{E,i}^{up}(t, t_p))}{V_E(t-t_p, X_{E,i}^{up}(t-t_p, 0))}. \quad (5.15)$$

5.3 Evaluation

The test site for evaluation is the German Autobahn A9 close to Munich in northbound direction. On April 30th, 2015 a severe congestion occurred due to a lane closure after an accident (Fig. 5.3). One-minute flow measurements on several lanes are collected using detectors, averaged over all lanes and normalized by the number of lanes (Figure 5.4 up)¹. Figure 5.4 mid visualizes the raw trajectory data that was reported by the fleet of vehicles (see chapter 3) during that day on this road segment. The velocity estimate $V_E(t, x)$ (Figure 5.4 down) is computed using the PSM.

The downstream congestion front of the pattern is fixed upstream at the accident location at kilometer 43. There, WMJs emerge and propagate upstream. These WMJs separate as they travel upstream., forming distinct phase regions. At the time when the congestion occurs (approx. 5pm), the upstream congestion front propagates upstream with about 15km/h until 6pm. At 6pm a significant drop of upstream flow is measured. The upstream front propagates much slower and some WMJs dissolve. At around 7pm, the maximum length of the congestion pattern is reached.

5.3.1 Accuracy Assessment

In order to evaluate the accuracy of a front forecast an error metric is required. A first approach would be to consider the RMSE of a GT position and a forecast position. Though, this metric has one significant drawback. In the case a jam dissolves and its phase fronts vanish but a forecast front still exists, there is no pair of fronts to calculate an error. Hence, the RMSE is not able to represent the over- or underestimated existence of a phase front.

Therefore, a metric is applied that penalizes the following errors:

1. If the simulated front deviates more than x_{tol} from the GT front.
2. If the forecasted front dissolved, but the GT front is still active (true negative).
3. If the GT front already dissolved, but the simulated front is still active (false positive).

The fulfillment of these conditions is summarized in the binary value $Hit_i(t, t_p)$. It indicates whether for predicted front i there is a corresponding real front in the proximity

¹Thanks to *Autobahndirektion Südbayern* for providing the detector data.

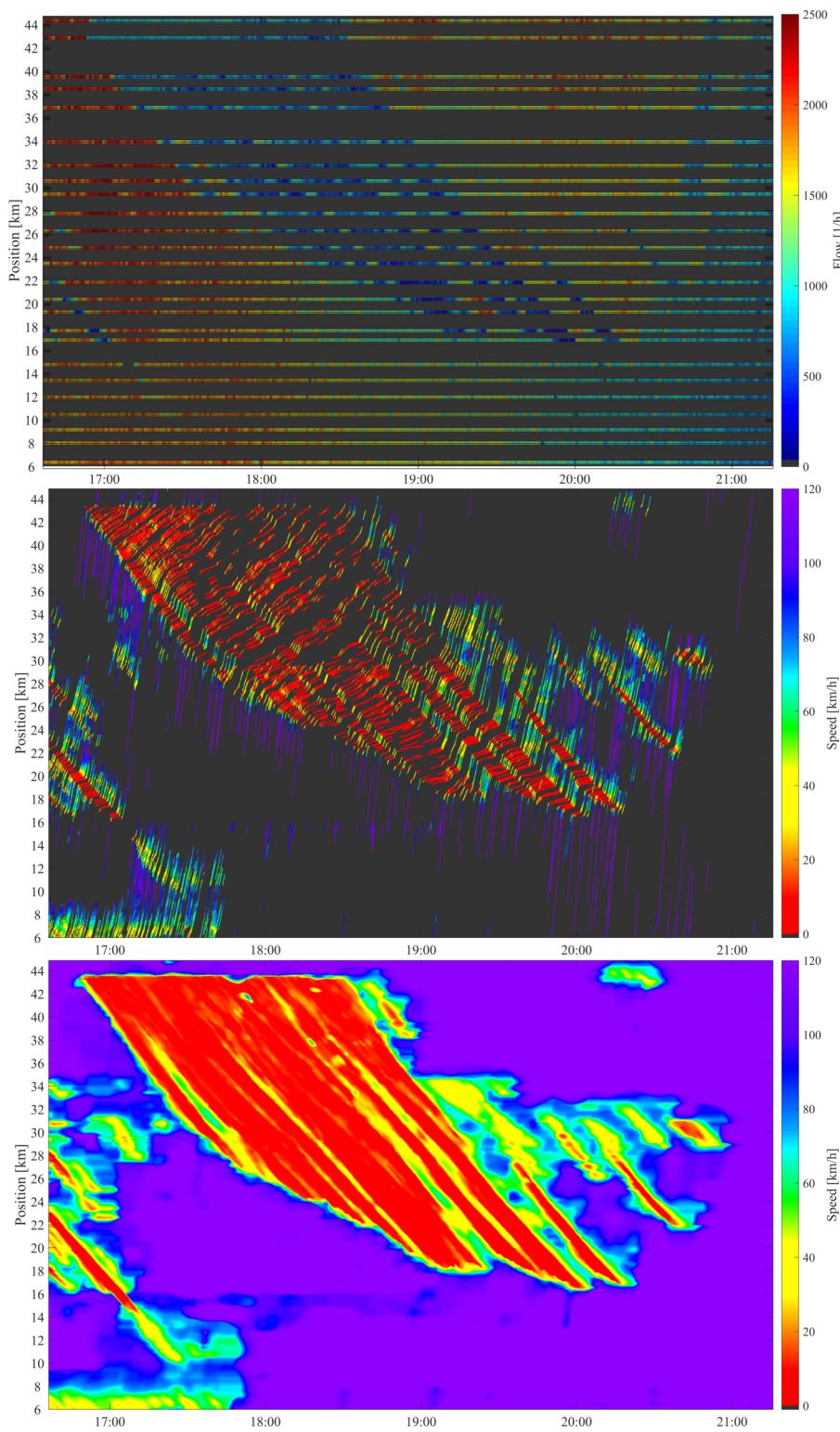


Figure 5.4. Congestion scenario used for evaluation. (Up) Normalized flow values collected by loop detectors. (Center) Collected FCD. (Bottom) Estimated traffic speed using the PSM

of x_{tol} depending on the current time t and the forecast time t_p :

$$Hit_i(t, t_p) = \begin{cases} 1 & \text{if } \exists j : |X_{GT,j}^{up}(t) - X_{E,i}^{up}(t, t_p)| < x_{tol}, \ j = 1, \dots, |X_{GT}| \\ 0 & \text{otherwise .} \end{cases} \quad (5.16)$$

Additionally, $Tot_i(t, t_p)$ indicates whether there exists either a GT or simulated front of index i at time t :

$$Tot_i(t, t_p) = \begin{cases} 1 & \text{if } |X_{GT}^{up}(t)| \geq i \vee |X_E^{up}(t, t_p)| \geq i \\ 0 & \text{otherwise .} \end{cases} \quad (5.17)$$

For instance, if at prediction time t_p there exist two GT fronts and one simulated front (since the algorithm has predicted the dissolution of a jam), $Tot_1(t, t_p)$ would equal one, and $Tot_2(t, t_p)$ would equal one. $Hit_1(t, t_p)$ equals one if the still existing forecasted front is in proximity of x_{tol} to any GT front, and $Hit_2(t, t_p)$ equals zero. Thus, for this point in time and this prediction horizon the algorithm has an accuracy of 50 %.

The accuracy $A(t_p, i)$ aggregates all $Hit_i(t, t_p)$ and $Tot_i(t, t_p)$ of time interval $[T_1, T_2]$ and calculates the ratio. Its parameters are the prediction time t_p and front index i :

$$A(t_p, i) = \int_{T_1}^{T_2} \frac{Hit_i(t, t_p)}{Tot_i(t, t_p)} dt. \quad (5.18)$$

5.3.2 Results

In this comparison four algorithms are implemented and its accuracies depending are compared. In addition to the aforementioned three variants K-DET, K-MAX and K-FCD, a naive predictor is applied. It propagates any front with a constant velocity of v^{cong} upstream.

Two influences on the accuracy are evaluated: The prediction horizon t_p and the front index i . Therefore, time is discretized into intervals of $\Delta T = 1\text{s}$ and space into $\Delta X = 50\text{m}$. For each point in time for which at least one GT front exists, the identified (upstream and downstream) fronts are forecasted for horizons of $t_p \in [0, 10\text{ min}]$. For accuracy estimations x_{tol} is set to 500 m.

Figure (5.5) depicts the positions of the simulated upstream fronts compared to the GT fronts for $t_p = 5\text{ min}$ and $t_p = 10\text{ min}$ respectively. For a concise description of the observations, in the following upstream fronts with index i equal to one are denominated as 'first order' fronts, and the remaining upstream fronts (with $i > 1$) as 'higher order'

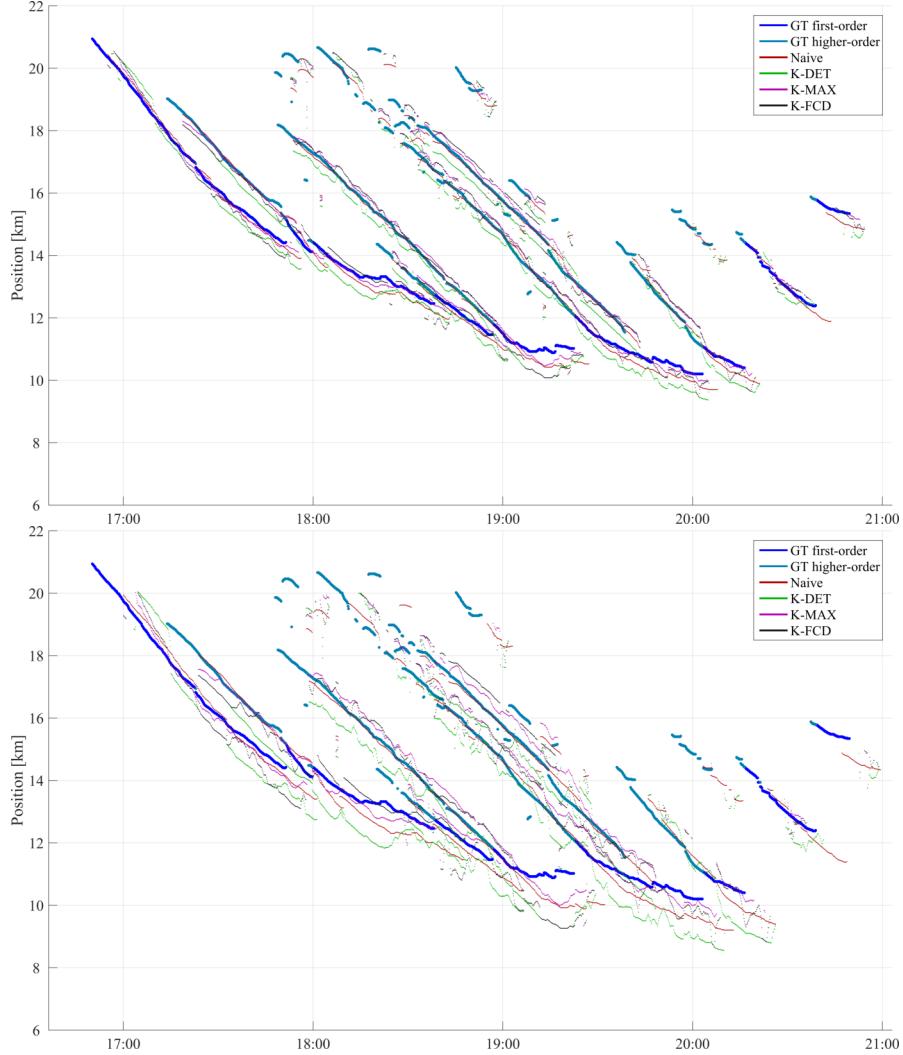


Figure 5.5. Comparison of Ground Truth fronts and predicted upstream fronts for several variations of the proposed algorithm and a prediction horizon of 5min (up) and 10min (bottom)

fronts. Note that in this context it does not refer to any mathematical concept but is simply used as an abbreviation.

Comparing the GT fronts and the predictors several observations can be made:

- During 5pm and 6pm all predictors produce reasonably accurate results for a 5min horizon.
- In this time, the second order front is forecasted most accurately with the naive predictor.
- During 6pm and 8pm the first order fronts have lower propagation speeds and several WMJs dissolve. For both time horizons the K-DET and naive approach

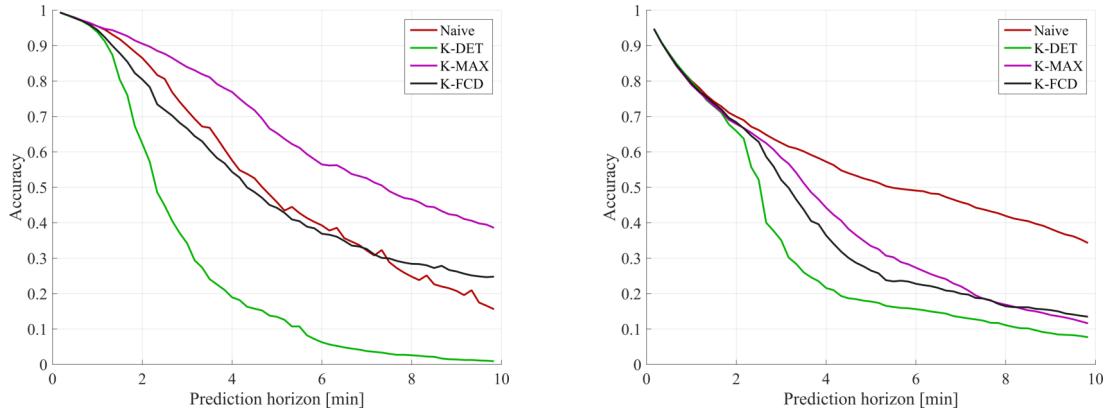


Figure 5.6. Accuracy of several variations of the proposed algorithm with respect to the prediction horizon. On the left, the accuracy for the prediction of the most upstream congestion front; on the right the accuracy for all other fronts

frequently overestimates the front propagation, significantly. K-MAX seems to be the most accurate one.

- Higher order fronts seem to have quite constant propagation speeds. All approaches except the naive forecast strongly varying fronts.

Fig. 5.6 visualizes the accuracies of the three variants and the naive algorithm with respect to the prediction horizon. A distinction between the prediction of the first order and higher order fronts is made since in both cases different effects influences the front propagation: the first order front is mostly influenced by the prediction of the upstream flow, while the inflow of higher order fronts is given by the outflow of the neighboring fronts (compare Fig. 5.2).

The accuracy for the first order front forecast are visualized in the left diagram. As expected all algorithms achieve worse accuracies with increasing prediction horizon. The K-DET is the most inaccurate. After a relatively short time of 5 min its performance drops below 15 %. The naive approach and K-FCD perform similarly: After 5 min they still forecast correctly in about 45 % of the time. With an accuracy of 65 % K-MAX yields most accurate results for all horizons.

The accuracy for higher order fronts shows different results. Here, the naive approach outperforms the other ones: After 5 min its accuracy is at about 50 % whilst K-MAX, K-FCD and K-DET forecast with an accuracy of 34 %, 27 % and 18 %, respectively. Noticeable is the equal accuracy of all predictors for a horizon between 0 min and 2 min. This is due to the discontinuity of higher order fronts in complex congestion patterns. As a pattern evolves over time, WMJs diverge and partly merge, the exact location of higher order fronts is difficult to define (and estimate from sparse data). Therefore, as visible in Fig. 5.5, there are several identified front positions of very short duration. These

fronts are forecasted for several minutes, but since the GT front already vanished, all predictions are wrong. At the same time, a simulated front requires some time to deviate more than x_{tol} from the initialized GT. Therefore, the systematic error dominates over the algorithm-specific prediction accuracy during the first minutes, which causes similar accuracies for all approaches. This effect is also visible in the first order prediction accuracy, but it is less striking.

To conclude the results, K-DET is the least accurate predictor. Apparently, the usage of density data deduced from loop detectors results in the most inaccurate phase front forecast. A detailed analysis of the calculated densities reveals that in the congested regime the traffic density is often underestimated. The consequence are overestimated (absolute) front propagation speeds. The K-FCD approach is able to correct the estimate to a certain degree, such that its performance is significantly higher for first order and higher order fronts. Though, it does not manage to eliminate the bias completely and underestimates traffic densities similarly. In average, the naive predictor is comparably as accurate for first order fronts. Setting the traffic density based on empirical values as done in the K-MAX variant seems to be the most accurate way to calculate realistic first order front speeds. For higher order front speeds it appears to be more difficult to estimate realistic traffic conditions up- and downstream of the phase fronts. Hence, the analytical models lack accuracy and are outperformed by a naive approach. The good performance of the naive approach for higher order fronts can be explained well with the FD (Fig. 5.2): The in-flow into the congested state is the outflow of the upstream free flow (or synchronized flow) state whose in-flow, in turn, equals the outflow of another congested regime. The transition from a WMJ state with nearly constant downstream front speed of v^{cong} implies that flow and density of the downstream phase lie on the according line of the FD (compare to states (2,3,4,6) in the diagram). Another transition from any state on this line to a WMJ state results consequently in phase front propagation speeds of v^{cong} . Hence, the obtained results of the naive predictor match the expectations that stem from the assumed FD.

5.3.3 Discussion

The proposed method is able to handle sparse FCD and sparse flow measurements in order to provide short-term front forecasts of WMJs fronts. One variant of the method showed to be more accurate than a naive predictor for first order fronts. Up to now, the model focuses fronts of WMJs which are especially dangerous for road users due to significant drops of velocity. Nevertheless, for some applications also the fronts of synchronized flow phases may be relevant. In order to provide short-term forecasts also for synchronized flow states the proposed model needs to be extended. Therefore,

the phase probabilities computed during the PSM can be utilized. Though, since the synchronized flow phase is characterized by a wide-scattering of traffic states (Kerner, 2009), the exact determination of traffic flow and density is more challenging. This increases the requirements of high quality and high resolution of data. Additionally, for synchronized flow phases also a distinct downstream phase front forecast need to be calculated. This is especially challenging since it depends on a forecast of bottleneck capacities and ramp in-flows (e.g. at merging regions).

Another limitation is that the flow forecast is basically a linear extrapolation (including a smoothing) of current measurements into the future. However, if there are off- or on-ramps the real flow may change over time. A more sophisticated model could include current (and predictive) in- and outflows at ramps as well as infrastructural properties of the road (e.g. lane reductions).

5.4 Conclusion and Outlook

In this chapter a novel method is proposed that combines the strengths of flow as well as FC data in order to provide short-term congestion front forecasts. Using the high spatio-temporal resolution of FCD, congested regimes and according congestion fronts are identified. Subsequently, sparse flow data is utilized to forecast the congestion front positions. The developed model is based on a basic FD and the shock wave formula, which are widely accepted among different schools of traffic theory (Richards, 1956; Newell, 1993; Treiber and Kesting, 2013; Kerner et al., 2004; Laval, 2007; Nagatani, 2002). It combines different data sources in a flexible, robust and efficient way using smoothing operations. This allows to apply the method to various types of data and in real-time.

The evaluation of the method is done using the FCD and loop data reported during a severe congestion on a German Autobahn. The accuracies of three variants of the proposed method and a naive predictor are compared. The results show that for the first upstream congestion front one variant of the proposed method outperforms the other approaches significantly. For cascades of WMJs, further upstream fronts are forecasted more accurately using a naive predictor, which complies with the expectations that stem from the assumed traffic model. Hence, in a future traffic system a combination of both approaches is likely to achieve the most accurate forecasts.

For future work, further studies should be conducted that focus the fusion of various types of data, e.g. density and flow measurements collected via vehicle sensors (see (Seo and Kusakabe, 2015) who study the estimation of traffic density using in-vehicle distance

sensors). Furthermore, the proposed model could be extended with more sophisticated flow forecasts which consider in- and outflows and infrastructural properties.

Chapter 6

Congestion Analysis and Prediction in Urban Road Networks

In the preceding two chapters methods were described that estimate traffic speeds and forecast congestion fronts on freeways. In this chapter, urban road networks are focused. Compared to traffic congestion on freeways, congestion in urban road networks differs in the following ways:

- Freeways are usually designed as long multi-laned roads. Ramps with dedicated acceleration and deceleration lanes seek to harmonize vehicle speeds on the main lanes in order to increase traffic flow and safety. The free driving speed often exceeds 100 km/h. Congestion patterns usually origin due to interactions between vehicles at on- and off-ramps in dense traffic conditions (see (Kerner, 2004)). A congestion pattern is usually represented as a congested interval on a road corridor (see chapter 4). Urban networks usually have low speed-limits. They comprise many short road segments connected at signalized intersections. Traffic congestion usually spreads over several connected segments and branches intersections. Therefore, the common way to represent a congestion pattern on a road corridor is not applicable. Rather, branched subgraphs of the network need to be considered.
- Stationary sensors providing data are costly and freeways are covered only sparsely. Urban road networks are usually equipped even less dense. Therefore, the estimation of traffic conditions is additionally challenging.
- Many freeway traffic forecast approaches develop analytical models of flow, density and speed based on the LWR equation and assimilate data with the model (see

section 2.3). A forecast of traffic conditions in urban road networks is particularly challenging. It requires further predictive information such as turn ratios of vehicles and signal timings. Both are information that are difficult to obtain in practice.

As source of traffic data with potentially high coverage FCD allows to estimate urban traffic congestion on a wide scale. Though, the estimation and prediction of urban traffic states is still challenging. Vehicles frequently accelerate and brake, change lanes etc. even if traffic density is low. This complicates the detection and description of congestion. For instance, if a vehicle in a dense network reports low velocities, it is unclear whether the vehicle is in congested traffic, queuing at a signal, parking at the road side, not able to overtake a slow bike etc. Thus, data from individual vehicles may be non-representative for the traffic conditions. Moreover, a network comprises a huge number of road segments with different properties such as lengths, speed-limits and number of lanes. Processing data on a multitude of segments can be computationally expensive if algorithmically complex methods are applied. Finally, sparse FCD is barely able to feed a data-hungry analytical forecast model.

This chapter presents a novel way to cope with the challenges of traffic prediction in urban networks. The idea is to reduce the traffic network to the most vulnerable parts that are frequently congested and analyze network-wide traffic congestion based on a few variables representing the level of congestion in said parts. This simplifies the manual analysis of traffic in large networks, and furthermore facilitates the training of data-driven prediction models.

Section section 6.1 presents related works in urban traffic analysis and prediction and motivates a novel approach. Section 6.2 provides a formal definition of the developed clustering algorithm and travel time prediction method. The evaluation is done using one year of FCD reported in the Munich traffic network (section section 6.4). First, the sensitivity of the method with respect to its parameters is investigated. Subsequently, traffic conditions inside the clusters are examined for spatio-temporal patterns. Finally, the key results of the pattern analysis are integrated into the proposed forecast methodology and its accuracy is assessed. Section 6.5 summarizes the chapter with a critical discussion and an outlook.

6.1 Related Work and Solution Approach

In (Vlahogianni et al., 2014) ten major challenges of traffic forecasting are pointed out which mark promising directions to increase forecast accuracy. Among them is the need

to consider temporal as well as spatial dependencies between traffic conditions on different edges in the network. Several recent works published results aligned with that direction. To mention a few, (Min and Wynter, 2011) apply a multivariate spatial-temporal autoregressive model on a sample network with different road categories. (Kamarianakis and Prastacos, 2005) model the traffic flow in space and time using a Space-Time Autoregressive Integrated Moving Average (STARIMA) model. (Cheng et al., 2012a) compute correlations between edges in the London traffic network in order to analyze required model complexities for models such as STARIMA. (Ma et al., 2015a) study congestion propagation in networks using probe data. They apply a Restricted Boltzmann machine in order to predict the congestion evolution in the road network. Most of the literature is based on small to medium sized networks that model dependencies between road segments. For bigger networks, the computational expense to compute the correlations between all edges increases dramatically. In order to eliminate mutual dependencies, neighborhood selection techniques, such as a Graphical Lasso are proposed (Gao et al., 2011). (Haworth and Cheng, 2014) give a comparison about different methods. Still, these approaches are on an edge-level, plus, the neighborhoods are still local which does not allow to consider network-wide relations. In order to aggregate similarly behaving edges of a large network, (Asif et al., 2014) apply several clustering techniques. Edges do not need to be connected and, therefore, resulting clusters are disseminated. (Anwar et al., 2014) propose algorithms in order to create so-called supernodes that represent connected subgraphs of large road networks. Edges of a subgraph are expected to have high similarities. Though, the approach is validated with simulated data only and the number of resulting supernodes is still large. A different way is to analyze congestion from a on network-level. (Ji and Geroliminis, 2012; Ji et al., 2014) partition a road network dynamically into connected and congested subgraphs with similar properties. They observe the evolution of such congested regions over time and seek to determine a macroscopic FD for urban networks.

The presented method is based on the work by (Ji and Geroliminis, 2012) and extended with further concepts that allow for a more sophisticated congestion pattern analysis and forecast. One observation when applying the dynamic partitioning of the network is that congestion often emerges and resides in the same parts of the network. A possible explanation is that similar commuting patterns of travelers cause high traffic demands at the same bottlenecks every day, which ultimately leads to recurrent spatio-temporal congestion patterns. As a side-effect, there are also many edges of the network which are rarely or never congested. For traffic monitoring and prediction, these edges are less relevant and may be neglected in favor of a reduced model complexity and decreased computational times. Thus, the idea is to identify these regions in a network in which congestion occurs on a regular basis and focus monitoring and forecasting on said parts.

Besides a reduced computational effort, the identification of such regions has further benefits. One advantage is that the level of congestion at one bottleneck can be quantified with one variable. Compared to edge-based approaches where each edge has a different length and may be in a different traffic condition, an aggregation of all involved edges results in a more robust and representative variable for a subgraph. These aggregates computed for several congestion-prone parts of a network finally enable to analyze the spatio-temporal relations between traffic conditions at distant bottlenecks in the network using just a small number of variables. A small number of variables is advantageous for analysis and visualization as well as for the training of data-driven forecasting methods with limited amount of data.

In the following, these congestion-prone regions are called 'congestion clusters'. They have the following desired properties:

1. They span the regions that are frequently congested.
2. They are static over time.
3. All edges of a congestion cluster are connected.

6.2 Definition of Congestion Clusters

Figures 6.1 gives an overview of the steps taken in order to determine static congestion clusters. These steps will be explained in detail in the following sections. First, for each time step an edge is determined as free or congested depending on reported velocity data and the edge's speed limit. Subsequently, subgraphs of congested and connected edges are identified which are denominated as congestion pockets. A spatial smoothing is performed that closes gaps in-between these subgraphs. These congestion pockets are determined for each point in time. In a next step, a so-called connectivity matrix is computed that counts the number of time steps two edges are assigned to the same congestion pocket. Finally, using this matrix an iterative algorithm constructs static congestion clusters comprising edges which are frequently congested simultaneously.

6.2.1 Dynamic Congestion Pockets

Assume that for each edge $e \in \mathcal{E}$ of the graph \mathcal{G} (see section 3.2) the average traffic velocity at time t is determined from reported GNSS data and denominated as $V_{Rec}(e, t)$. Further assume that the length of an edge e is $l(e)$ and the speed limit $V_{Lim}(e)$. The relative driving speed is defined as:

$$V_{Rel}(e, t) = \frac{V_{Rec}(e, t)}{V_{Lim}(e)}. \quad (6.1)$$

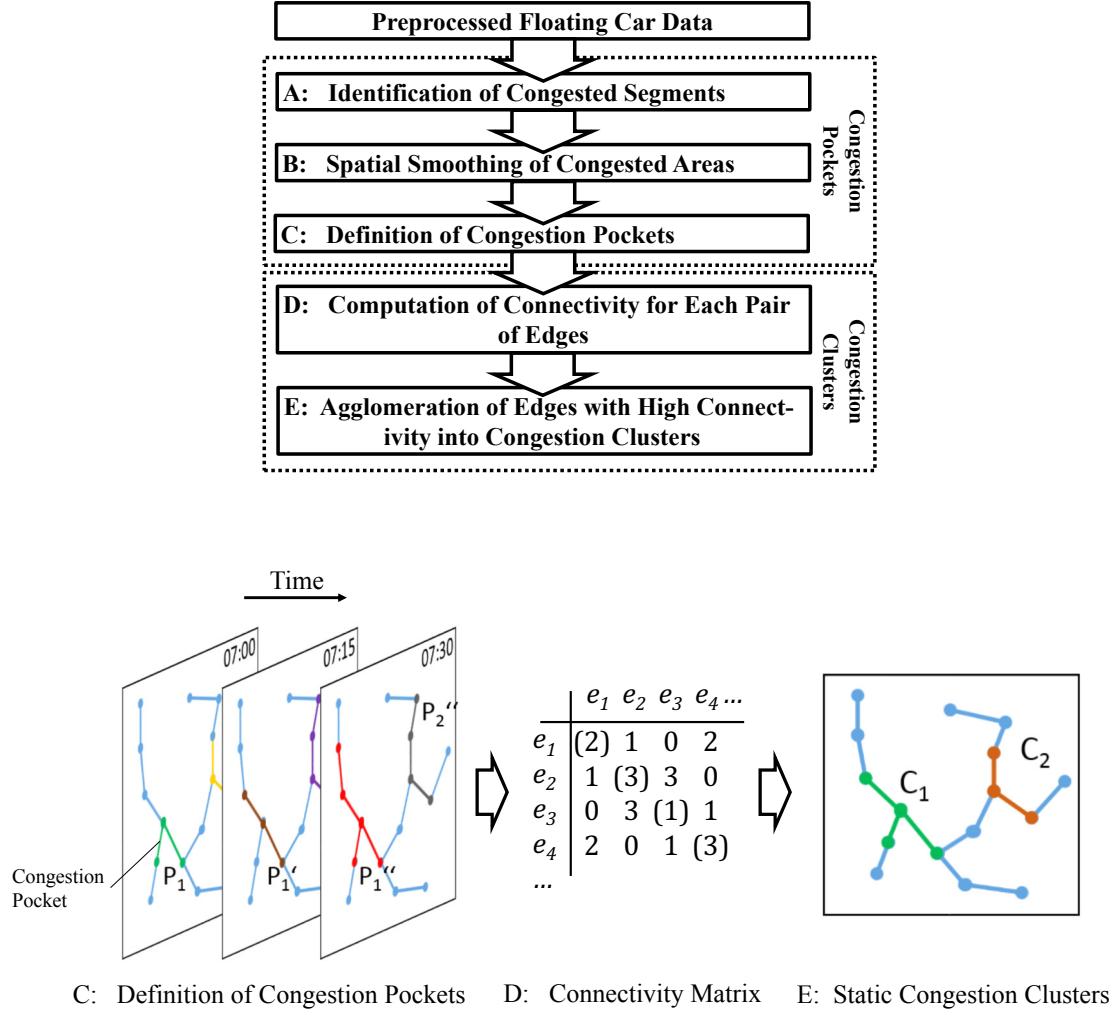


Figure 6.1. Overview of the steps taken to process FCD into static congestion clusters

The relative driving speed is compared to a threshold $V_{Rel}^{thres} \in [0, 1]$ that distinguishes between free and congested traffic. Accordingly, the function $J(e, t)$ ('J' for 'jam') is defined that indicates whether edge e is congested at time t :

$$J(e, t) := \begin{cases} 1 & \text{if } V_{Rel}(e, t) \leq V_{Rel}^{thres} \\ 0 & \text{otherwise.} \end{cases} \quad (6.2)$$

Similar to (Ji et al., 2014) a dynamic congestion pocket \mathcal{P} is defined. A congestion pocket denotes the spatial extent of an occurring traffic jam at a certain time t . Each congestion pocket is a time-dependent subgraph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ of $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, i.e. $\mathcal{V}' \subseteq \mathcal{V}, \mathcal{E}' \subseteq \mathcal{E}$. For each point in time the number and size of the congestion pockets may change. Formally, for some time t and a congested edge e^* , a congestion pocket $\mathcal{P}(e^*, t)$ is defined as the set of all edges $e \in \mathcal{E}$ that have the subsequent properties:

1. $J(e, t) = 1 \quad \forall e \in \mathcal{P}(e^*, t)$.

2. Either there exists a path from e^* to e on \mathcal{G} or a path from e to e^* on \mathcal{G} that consists solely of edges to which $J(e, t)$ assigns a value of one.

Thus, $\mathcal{P}(e^*, t)$ describes a set of edges that are associated with edge e^* which is part of the congestion pocket. For non-congested edges this set is consequently empty. This relation from edge to congestion pocket is surjective, which means that each edge at time t is mapped to one congestion pocket at maximum. On the other hand, a congestion pocket may refer to several edges.

When considering real traffic data, the number of congestion pockets can become large. An often occurring effect is that regions, which seem to belong to the same congestion pocket are separated by single, non-congested edges. In many cases, this is rather a lack of data, an error of the map-matcher or a short change of traffic conditions than a real separation of two congestion pockets. Hence, a spatial smoothing is carried out as suggested in (Ji et al., 2014). There, after determining the set of congested edges for some time t , all edges which have more congested neighbors than non-congested neighbors are defined as congested as well. This means that function J is redefined as stated below:

$$J(e, t) := \begin{cases} 1 & \text{if } V_{Rel} \leq V_{Rel}^{thres} \\ 1 & \text{if } |\{e' \in \mathcal{N}(e) : V_{Rel}(e', t) \leq V_{Rel}^{thres}\}| > |\mathcal{N}(e)| / 2 \\ 0 & \text{otherwise.} \end{cases} \quad (6.3)$$

Thereby, $|\cdot|$ denotes the cardinality of a set and $\mathcal{N}(e)$ denotes the neighborhood of e , i.e. the set of all edges in \mathcal{E} that share at least one node with e (except for edge e itself). Having function J adjusted, the definition of congestion pockets remains basically the same.

6.2.2 Static Congestion Clusters

If two edges are frequently congested at the same time and they have a high proximity it is likely that those edges belong to a congestion-prone region at the same bottleneck. A static congestion cluster is supposed to agglomerate these edges. Dynamic congestion pockets model the proximity and congested state of edges such for one point in time. In the following, the temporal clustering of congestion pockets is described.

Assume that for all discrete time intervals $\mathcal{T} = \{T_0, T_0 + \Delta T, T_0 + 2 \cdot \Delta T, \dots, T_1\}$ congestion pockets are computed. The function $Y(e_1, e_2)$ is defined counting the number of

time intervals in which two edges $e_1, e_2 \in \mathcal{E}$ are part of the same congestion pocket:

$$Y : \mathcal{E} \times \mathcal{E} \rightarrow \{0, 1, \dots, |\mathcal{T}|\} \quad (6.4)$$

$$Y(e_1, e_2) := |\{t \in \mathcal{T} : e_1 \in \mathcal{P}(e_2, t)\}|. \quad (6.5)$$

The resulting quantities can be represented as a matrix (compare Figure 6.1 step D). It is symmetric with the total number of time steps for which an edge is congested on its diagonal. For the clustering, the duplicate entries (due to the symmetry) and the diagonal elements are not relevant. Therefore, in the following matrix Y^* as the strictly lower triangular matrix of Y is considered. Two edges with a relatively high corresponding entry in Y^* are called edges with high connectivity.

All edges with a high connectivity are supposed to be clustered into a finite number of static clusters $\mathcal{C}_i \in \mathcal{E}, i = 1, \dots, n_c$. An iterative algorithm is applied which assigns edges to different clusters \mathcal{C}_i based on Y^* . In short, the algorithm finds the pair of edges with the highest connectivity in the matrix. It checks, whether one of the edges is already assigned to any cluster. If not, a new cluster is defined that comprises these two edges. If one of the edges is already assigned to a cluster, the other edge is assigned to the same cluster. If both edges are already assigned to differing clusters both clusters are merged. Finally, the connectivity of these two edges is set to zero and the algorithm evaluates the next pair of edges. This procedure is done as long as:

$$\max(Y^*(e_1^*, e_2^*)) > Y_{min} \quad (6.6)$$

with

$$Y_{min} = \alpha \cdot \max(\{Y(e_1, e_2) : e_1 \neq e_2\}). \quad (6.7)$$

Parameter $\alpha \in [0, 1]$ is introduced to decouple the clustering from the number of analyzed time intervals \mathcal{T} .

Algorithm 1 describes the iterative pseudo code for the cluster generation. Auxiliary function $c(e)$ returns the index of the cluster \mathcal{C}_i to which e is assigned. If it is not assigned yet it returns the value zero.

After the clustering algorithm has stopped it holds that the union of all edges in all clusters represents a set of edges where any two edges are at least Y_{min} time slices part of the same cluster:

$$\mathcal{C} = \bigcup_{i=1,2,\dots,n_c} \mathcal{C}_i = \{e \in \mathcal{E} : \exists e' \in \mathcal{E} \text{ where } Y^*(e, e') \geq Y_{min}\}. \quad (6.8)$$

Algorithm 1: Static clustering

Data: Connectivity matrix Y^* , connectivity threshold Y_{min}
Result: Static clusters \mathcal{C}

```

 $n_c := 0;$ 
while  $\max(Y^*(e_1^*, e_2^*)) > Y_{min}$  do
    | find  $(e_1^*, e_2^*) := \text{argmax}(Y^*(e_1, e_2) : e_1 \neq e_2)$  ;
    | if  $c(e_1^*) = 0 \wedge c(e_2^*) = 0$  then
    |   |  $n_c := n_c + 1$  ;
    |   |  $\mathcal{C}_{n_c} := \{e_1^*, e_2^*\}$  ;
    | else if  $c(e_1^*) > 0 \wedge c(e_2^*) = 0$  then
    |   |  $\mathcal{C}_{c(e_1^*)} := \mathcal{C}_{c(e_1^*)} \cup e_2^*$  ;
    | else if  $c(e_1^*) = 0 \wedge c(e_2^*) > 0$  then
    |   |  $\mathcal{C}_{c(e_2^*)} := \mathcal{C}_{c(e_2^*)} \cup e_1^*$  ;
    | else
    |   |  $\mathcal{C}_{c(e_1^*)} := \mathcal{C}_{c(e_1^*)} \cup \mathcal{C}_{c(e_2^*)}$  ;
    |   |  $\mathcal{C}_{c(e_2^*)} := \{\}$  ;
    | end
    |  $Y^*(e_1^*, e_2^*) := 0$  ;
end

```

Remove empty clusters from list and update n_c ;

This approach is designed to identify connected edges of a network that are frequently congested. As such, they are expected to be most relevant for traffic management and for individual travelers. Therefore, in the following section a method is developed that utilizes the clusters for traffic forecasts.

6.3 Data-Driven Congestion Prediction in a Clustered Network

In this section a data-driven prediction model is developed. Alike the clustering, it is based on the assumption that traffic congestion follows certain patterns. For instance, that congestion occurs recurrently in similar regions of the network and at similar times. Furthermore, it is assumed that there are spatio-temporal dependencies between the level of congestion in different clusters of the network. These dependencies might be of various nature. For instance, on a day there may be especially low or high traffic demand, which impacts the level of congestion in all clusters. Or a higher demand occurs only on some origin-destination relations affecting only a few clusters. To model these dependencies explicitly and collect the necessary data in order to apply the model to forecast problems is challenging. The advantage of a data-driven approach is that no explicit modeling is required but that the model is deduced from collected data. Though, there exist

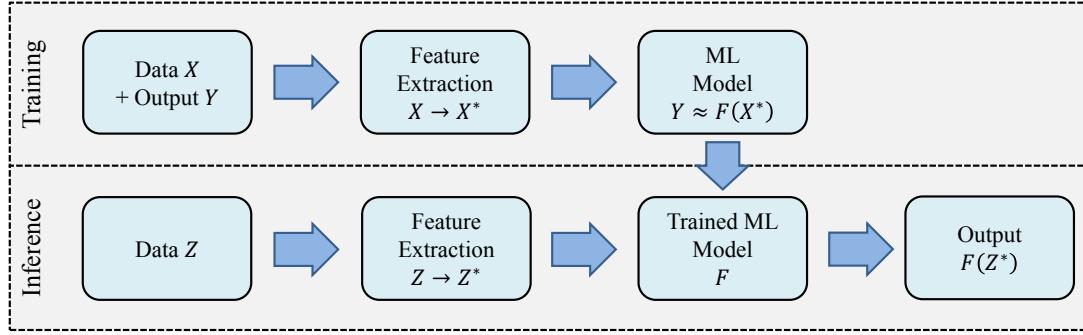


Figure 6.2. Flowchart of common supervised learning algorithms

numerous ways to formulate a data-driven method. This concerns the considered input and output variables, the model itself and the set of provided training data.

This approach focuses two aspects of a data-driven forecast method and thus distinguishes from other published works (see section 2.3): The first is the usage of the congestion clusters in order to develop a small set of expressive features as model input. The second is that not only the local neighborhood of a bottleneck but also network-wide traffic conditions are a valuable input for a traffic forecast in order to increase the accuracy.

6.3.1 Overview of the Machine Learning Pipeline

Many data-driven methods can be classified as ML approaches. ML is a subcategory of Artificial Intelligence (AI). The idea is that a computer program or mathematical model is derived from data solely, in contrast to traditional programs and models that are specified by humans. The task of an ML algorithm is to process possibly huge amounts of data and learn from these such that it is able to provide a response to unseen data.

One distinguishes between supervised, unsupervised and reinforcement learning (Russell et al., 2003). In supervised learning the algorithms process pairs of input and respective output during learning phase and use a trained model for inference. These types of algorithms are frequently utilized for the prediction of outputs given unseen data. Unsupervised learning seeks to identify hidden structures and patterns in data without a label (for classification) or function output. In reinforcement learning an algorithm seeks to learn strategies in order to optimize the reward that a dynamic environment returns. These algorithms require that the environment provides flexible amounts of data.

For congestion prediction based on observed data supervised learning suits. Reported data can be used to generate a model of a system that returns predictions for a future time. Figure 6.2 depicts the common process of supervised learning. First, during training phase, input and corresponding output data are used to train a function that approximates outputs from given inputs. Often, input data is transformed into another space. This process is called feature extraction. After training a model, the model can be used for inference. This denominates the process of predicting an output with unseen data input.

Compared to many other ML approaches in traffic forecasting this approach presented here is characterized by the definition of congestion clusters. In the context of ML this constitutes a feature extraction: data of tens of thousands of edges is transformed into a few variables describing the level of congestion inside the clusters. Feature extraction is a fundamental process in ML applications, summarized by one of the leading ML scientists Andrew Ng¹: 'Coming up with features is difficult, time-consuming, requires expert knowledge. Applied machine learning is basically feature engineering.' It increases the accuracy of a model, enables a deeper problem understanding and reduces the computational complexity of an algorithm (Krupka et al., 2008; Domingos, 2012; Guyon and Elisseeff, 2003). Furthermore, as shown in (Levi and Weiss, 2004) for an object detection problem, good features reduce the number of required training samples. The reduced number of training samples is a significant advantage for traffic forecasts. Since traffic congestion is a temporal phenomenon, there exists only a strongly limited number of days on which traffic data can be collected. Due to changes of the infrastructure and commuting patterns of travelers, data that has been collected several years ago might not be descriptive for the current system. This makes data-efficient approaches highly relevant.

6.3.2 KNN Travel Time Predictor

Many variables can be subject of a prediction in a traffic network. Usually, traffic speed, flow, density as fundamental variables are predicted. For travelers the most relevant information is the travel time, or, likewise, the Travel Time Loss (TTL) due to congestion. Traffic managers are interested in the reduction of TTLs of all vehicles in order to reduce the economical impacts of congestion and ensure smooth traffic flow. The instantaneous TTL on a set of edges $\mathcal{E}^* \subseteq \mathcal{E}$ at time t is approximated with the

¹Former professor for machine learning at Stanford, in leading positions for machine learning and AI at Google, Coursera and Baidu in the past years

recorded velocities $V_{Rec}(e, t)$ as:

$$TTL_{\mathcal{E}^*}(t) = \sum_{e \in \mathcal{E}^*} l(e) \left(\frac{1}{V_{Rec}(e, t)} - \frac{1}{V_{Lim}(e)} \right). \quad (6.9)$$

Note that the instantaneous TTL considers current traffic conditions. A real vehicle that passes through congestion may perceive a different travel time due to dynamically changing traffic conditions.

As described in the preceding section, the selection of expressive features is fundamental for all machine-learning tasks. For the choice of features in this application the following considerations are done: The assumption of the clustering and this predictor is that there are similar commuting patterns. At times with low traffic overall traffic demands, the infrastructure's capacity suffices to meet the demand. Traffic conditions are free in all parts of the network, including the congestion clusters. With increasing demand more vehicles intend to pass certain bottlenecks of the network. When the bottleneck's capacity is exceeded, traffic gets congested and the TTL in these regions rises. With more vehicles entering the bottleneck region, more edges get congested and the TTL in the congestion clusters increases. Thus, the level of congestion of a cluster relates to the current traffic demand on the corresponding bottleneck. Observing not only one but all clusters at the same time and their respective levels of congestion gives a picture of the traffic demand of the entire network. Note that this is a rough approximation of the real demand of the network. However, given incomplete FCD, more accurate estimates of the network demand are difficult to obtain. These current demands at bottlenecks are promising features for traffic forecast methods: they relate to the physical reason for congestion and each congestion cluster can be represented with just one quantity, which allows to represent the current state of the network with a small number of expressive variables.

Hence, one possibility is to use the TTL of an observed time interval of the current day as a feature for a data-driven traffic forecast. The TTL is a time series and each of the n_c clusters provides one time series discretized into intervals of ΔT . The consequence is a feature vector of size $n_c \times |\mathcal{T}^*|$ where \mathcal{T}^* represents the observed time interval considered for model training and inference. This feature vector is further simplified due to the following reasons: Because of a changing level of congestion and measurement inaccuracies the perceived TTL is volatile. This noise reduces the expressiveness of each individual measurement. Furthermore, for short-term predictions of a few minutes the current traffic conditions are highly relevant since congestion is relatively inert, i.e. the system requires time to change significantly. Thus, for short-term forecasts the most up-to-date measurements give important information about the traffic conditions that will prevail likely in the next minutes. Considering a longer-term prediction there is no

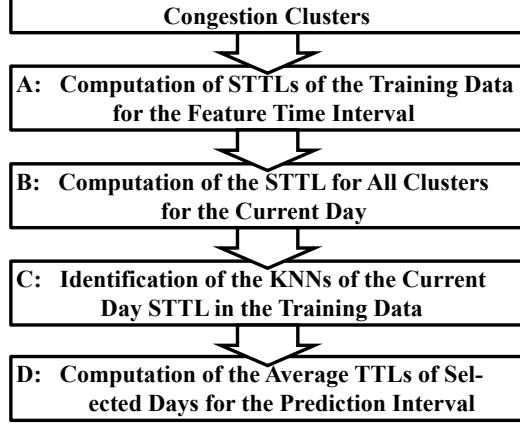


Figure 6.3. Data-driven TTL prediction using a KNN model

distinct time step of the TTL which has more predictive power than any other time step. Rather, it is assumed that the overall development of the function up to the current time provides the most valuable information. Due to these reasons the TTL over time interval \mathcal{T}^* is aggregated. The Summed Travel Time Loss (STTL) is defined as:

$$STTL_{\mathcal{E}^*}^{\mathcal{T}^*} = \sum_{t \in \mathcal{T}^*} TTL_{\mathcal{E}^*}(t). \quad (6.10)$$

Thus, the STTL represents the approximated traffic demand of edge set \mathcal{E}^* aggregated over interval \mathcal{T}^* . This step further reduces the number of features while trying to keep the essential information.

Since congestion occurs recurrently data is structured day-wise. Given $d = 1, \dots, n_d$ days for which time-discrete data $V_{Rel}(e, t)$ is available, the TTL and STTL for cluster $i = 1, \dots, n_c$ are defined as:

$$TTL_{d,t^*,i} := TTL_{C_i}(T_{d,t^*}) \quad (6.11)$$

$$STTL_{d,i}^{\mathcal{T}_F} := STTL_{C_i}^{T_{d,\mathcal{T}_F}} \quad (6.12)$$

where $T_{d,t}$ denotes a point in time that is described by day d and daytime t and \mathcal{T}_F the feature time interval represented as a set of discrete times. Thus, for a specific feature time the STTLs can be represented as a matrix of size $n_d \times n_c$. Figure 6.3 summarizes the process to provide TTL predictions for a cluster. First, for all training days \mathcal{D}_P the corresponding STTLs for a specific feature time are determined. Next, for the current day d^* the STTL for the same feature time is computed, denominated as $STTL_{d^*,i}$. A KNN classifier is applied which returns the K most similar historical days, i.e. the days with the lowest distance between all rows in $STTL_{\mathcal{D}_P} \in \mathbb{R}^{|\mathcal{D}_P| \times n_c}$ and $STTL_{d^*} \in \mathbb{R}^{1 \times n_c}$. The set of similar days is denominated as \mathcal{K} . In order to compare

multi-dimensional vectors, a distance metric such as the (normalized) euclidean or the manhattan distance etc. needs to be selected. The according TTLs of the similar historical days are aggregated into a TTL prediction TTL^{Pred} for the current day d^* :

$$TTL_{d^*,t,i}^{Pred} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} TTL_{k,t,i}. \quad (6.13)$$

In the last decades hundreds of machine learning algorithms and uncountable variants of each algorithm have been proposed. The KNN classifier is a rather simple method with only one parameter (i.e. the number of neighbors K), which is frequently applied in machine learning tasks (Bishop, 2013; Kuhn and Johnson, 2016). The focus of this work is to develop a framework that benefits from the definition of congestion clusters for the advantageous definition of prediction features. The KNN is selected as an exemplary algorithm which may be replaced with other more sophisticated methods. The identification of the most accurate algorithm requires extensive studies and could be part of future work.

To summarize, the proposed framework for traffic prediction constitutes a novel way to predict network-wide traffic congestion. It is based on the definition of congestion clusters whose traffic conditions are forecasted using a simple KNN approach. In section 6.3 this method is evaluated using one year of FCD.

6.4 Evaluation

In the following, the proposed methods are evaluated with the FCD described in chapter 3. Therefore, first the preparation of the data for this approach is presented. Next, the results of the static clustering are visualized and compared using several metrics and varying clustering parametrization. Subsequently, traffic conditions inside the clusters are analyzed statistically and spatio-temporal congestion patterns are extracted. Based on the statistical results, the travel time predictor is applied and its prediction accuracy is assessed. Section 6.5 summarizes this chapter and gives an outlook.

6.4.1 Data Preparation

The FCD collected in the Munich region in year 2015 is map-matched as described in section 3.2. In total 318 days of data are available, resulting in approximately 400,000 velocity measurements per day on a network comprising 17413 major road edges (corresponds to 1826 km). In order to process the measurements of individual vehicles into a

space-time continuous traffic estimate and fill gaps for which no data is reported, the following considerations are done: First, time is discretized into intervals of $\Delta T = 1$ min, such that $\mathcal{T}_d = 1, \dots, 1440$. Then, all velocity measurements V_{Rel} are interpreted as macroscopic traffic velocities for the respective time interval. If there are several measurements for the same time interval and the same edge e , their arithmetic average is computed. Since usually not for all time slices a velocity measurement is available, there are gaps in time and space. A simple estimation algorithm is applied: Each measurement is assumed to be valid for either 15 min or until a new measurement is reported. This allows to have a mostly continuous representation of traffic conditions for peak hours of the day. For all remaining gaps, which appear mostly at night or on minor roads, free flow conditions are assumed.

Figure 6.4 and Figure 6.5 depict the resulting traffic conditions during the morning and evening peak in the Munich road network for one exemplary day (March 11th, 2015). It is a regular Wednesday, in the sense, that no relevant events occurred that influenced traffic significantly. As visible most parts of the network remain in free conditions throughout the day. Some parts of the A99 get congested in morning traffic and others get congested during evening hours. Also certain parts of the 'Mittlerer Ring' get congested significantly. Inside this ring road several roads with low velocities are reported.

6.4.2 Cluster Metrics

In order to assess the quality of a clustering quantitatively and evaluate if they meet certain requirements, in the following four metrics are developed.

The first is the intra-cluster dissimilarity or homogeneity, originally proposed by (Ji and Geroliminis, 2012). One desired property of the clustering is that all edges of a cluster behave similarly, i.e. these edges are clustered optimally if all edges of that cluster are either congested or free at the same time. In this case a cluster is most homogeneous, or respectively, its dissimilarity is lowest. In accordance to (Ji and Geroliminis, 2012) the following metric measuring the homogeneity of a cluster (and extended with a temporal component) is applied:

$$\delta(\mathcal{E}^*) := 1 - \frac{\sum_{t \in \mathcal{T}} \sum_{e_1 \in \mathcal{E}^*} \sum_{e_2 \in \mathcal{E}^*} (V_{Rel}(e_1, t) - V_{Rel}(e_2, t))^2}{|\mathcal{T}| \cdot |\mathcal{E}^*|^2}. \quad (6.14)$$

Homogeneity is one important property, but not the only one. For example, diminutive clusters can have a high homogeneity, but do not provide information about the network since they are not able to cover a significant part of the network. Therefore, a second

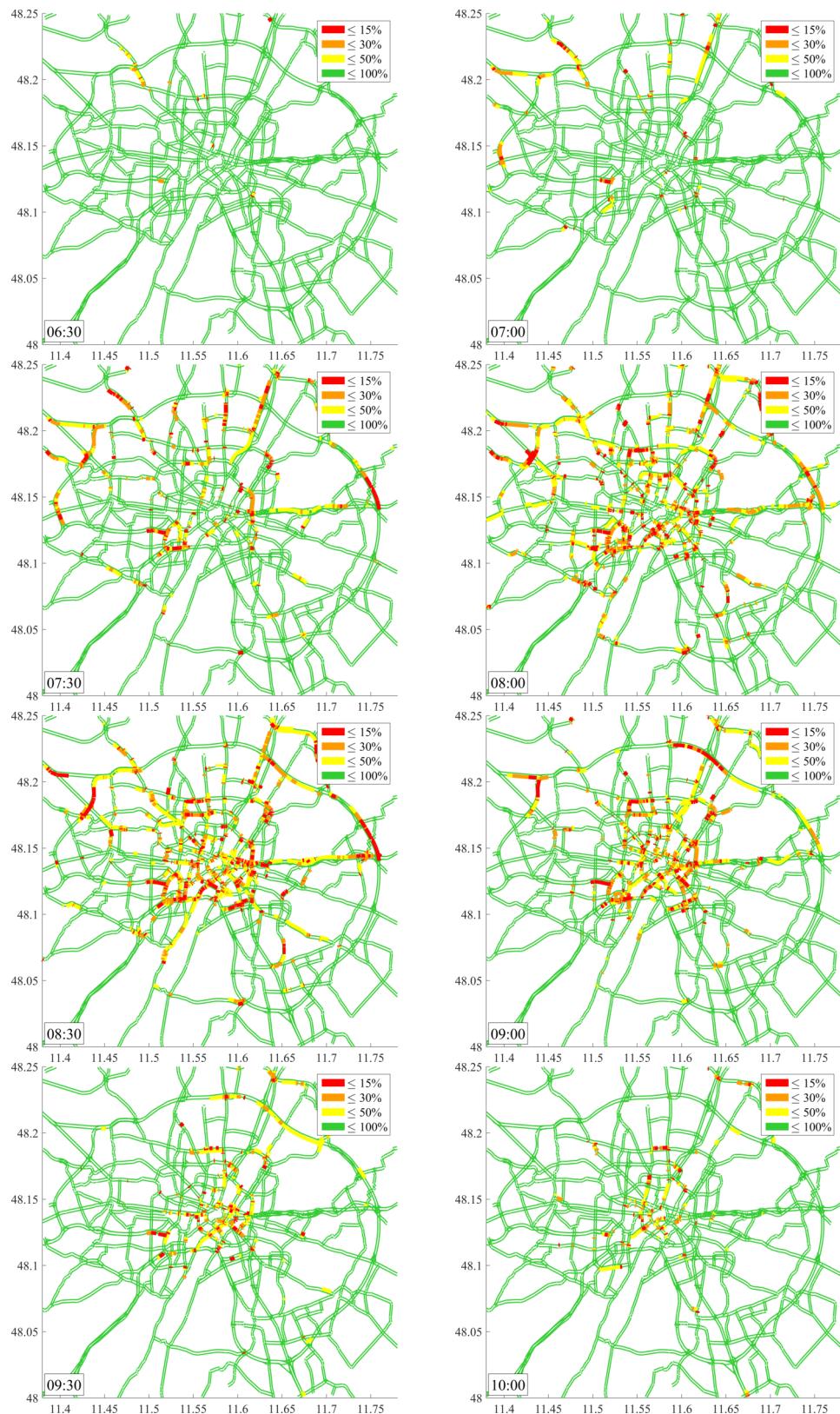


Figure 6.4. 30 min-Snapshots of the traffic conditions during morning peak on March 11th, 2015

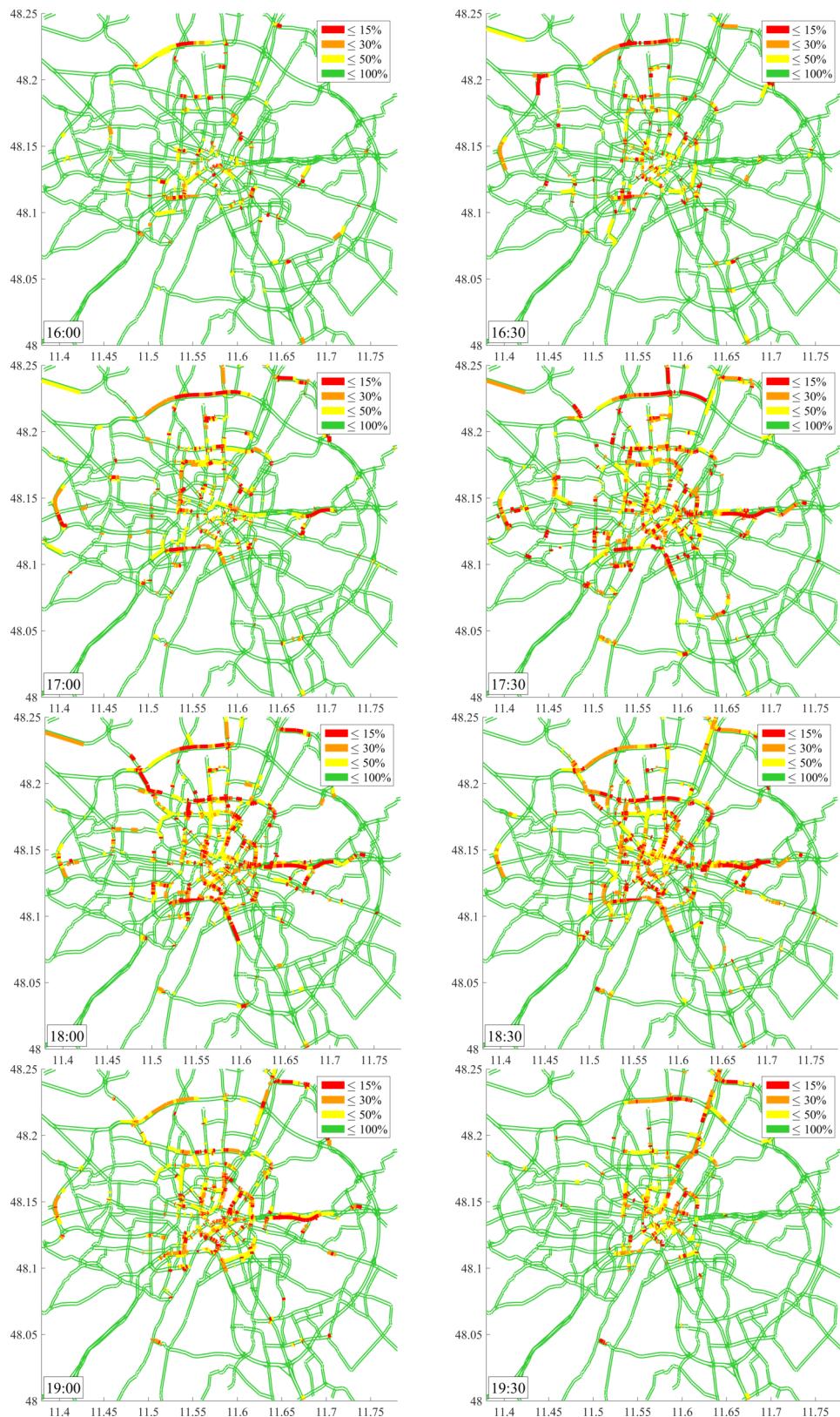


Figure 6.5. 30 min-Snapshots of the traffic conditions during evening peak on March 11th, 2015

metric is introduced: the network representativeness:

$$\rho(\mathcal{E}^*, \mathcal{E}) := \frac{\text{cov}(\kappa_{\mathcal{E}^*}, \kappa_{\mathcal{E}})}{\sigma(\kappa_{\mathcal{E}^*}) \cdot \sigma(\kappa_{\mathcal{E}})} \text{ with } \kappa_{\mathcal{E}^*, i} = \kappa(\mathcal{E}^*, \mathcal{T}_i) \quad (6.15)$$

with cov the covariance, σ the standard deviation and κ the level of congestion of a set of edges (corresponds to the congested part of a cluster):

$$\kappa(\mathcal{E}^*, t) := \frac{\sum_{e \in \mathcal{E}^*} J(e, t) l(e)}{\sum_{e \in \mathcal{E}^*} l(e)}, \quad \mathcal{E}^* \subseteq \mathcal{E}. \quad (6.16)$$

The network representativeness describes the correlation of the traffic conditions in a set of edges with the traffic conditions in all edges of the network. A high correlation means that an overall increase of congestion is related to an increase of congestion in a cluster. Consequently, the observation of cluster(s) allows to infer traffic conditions in the entire network. The best value of representativeness can naturally be reached if all edges of the network are clustered, while only few clustered edges correspond to a low representativeness.

A third metric is denominated as the specificity. It is the average level of congestion that occurs in all clusters over time period \mathcal{T}^* :

$$\xi(\mathcal{E}^*) := \frac{\sum_{t \in \mathcal{T}^*} \kappa(\mathcal{E}^*, t)}{|\mathcal{T}^*|}. \quad (6.17)$$

It expresses the requirement that a congestion cluster is supposed to span only mostly congested parts of the network during a relevant time interval. Connected free flow edges expand the size of the cluster but do not provide additional information. A high specificity describes that a cluster comprises edges that are mostly congested (during a peak interval \mathcal{T}^*).

Finally, the coverage described the ratio of congestion that occurs inside the clusters compared to the overall congestion in the network:

$$\zeta(\mathcal{E}^*, \mathcal{E}) := \frac{\sum_{t \in \mathcal{T}} \kappa(\mathcal{E}^*, t)}{\sum_{t \in \mathcal{T}} \kappa(\mathcal{E}, t)}. \quad (6.18)$$

It is similar to the network representativeness with the difference that it does not determine correlations but compares absolute levels of congestion. A resulting high value expresses that most part of the network congestion is covered by the congestion clusters.

In the following congestion clusters with a varying parametrization are determined and the proposed metrics are applied and compared.

6.4.3 Static Clusters in the Munich Road Network

A great part of road users that drive during peak hours commute between their homes and workplaces. The high number of trips at similar times ultimately causes congestion during peak hours. As visible in Figures 6.4 and 6.5, different parts of the network get congested during morning compared to evening peak. The congestion clusters are designed to be static over time. Hence, if applied to the data of a complete day the clustering algorithm would merge several areas of morning and evening peak into one cluster. As a result the specificity of the clusters would decrease. In order to differentiate between morning and evening traffic, each day is split at 12.00 noon and for each half-day a cluster set is determined.

All in all, 318 days of traffic data are used for cluster generation. Figure 6.6 illustrates the ten largest clusters resulting from the clustering with varying parameter α for the morning and evening period. Several observations can be done:

1. The higher is parameter α , the smaller the clusters.
2. There are substantial differences between the clusters of the morning and evening period. That justifies the approach to distinguish between these peaks.
3. During morning peak there seems to be frequent congestion on the freeway roads that lead into the urban region of the city.
4. During evening congested regions seem to be concentrated on more central parts of the network
5. Even with a low α there are many roads in the network that are not covered by any of the ten largest clusters during morning or evening peak. Hence, congestion here seems to be less severe in general.
6. In the center of the city there are clusters during morning and evening hours, which quickly merge into larger clusters with decreasing α .

Figure 6.7 depicts the homogeneity, representativeness, specificity and coverage of the evening clusters with respect to a varying parameter α . As expected the coverage and representativeness decrease as the clusters become smaller since fewer regions are covered. The specificity increases since the clustered edges tend to be congested more often if the clustering algorithm quits earlier. The homogeneity seems to have a minimal value at $\alpha \approx 20\%$. Apparently, small clusters as well as extensive clusters behave more homogeneously than mid-sized clusters. The reason is that small clusters (i.e. high α), which comprise only a few connected edges, are usually congested or free during the same time intervals and thus behave homogeneously. With a decreasing α more edges are added.

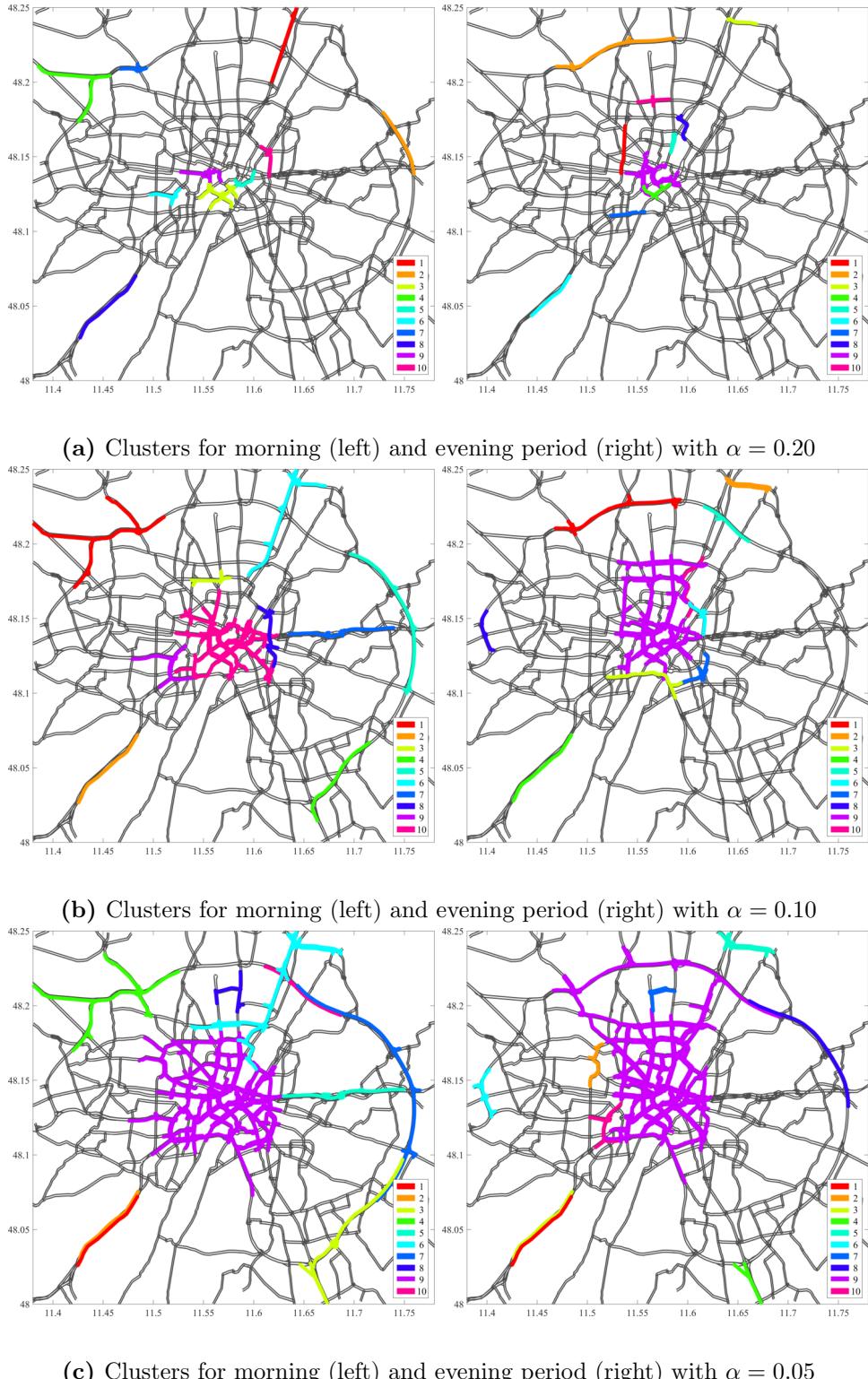


Figure 6.6. Results of the static clustering with respect to parameter α

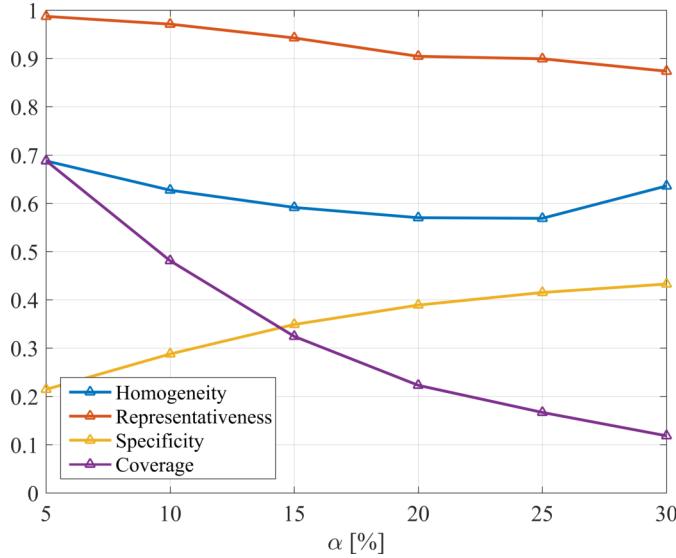


Figure 6.7. Quality metrics of the evening clusters with respect to parameter α

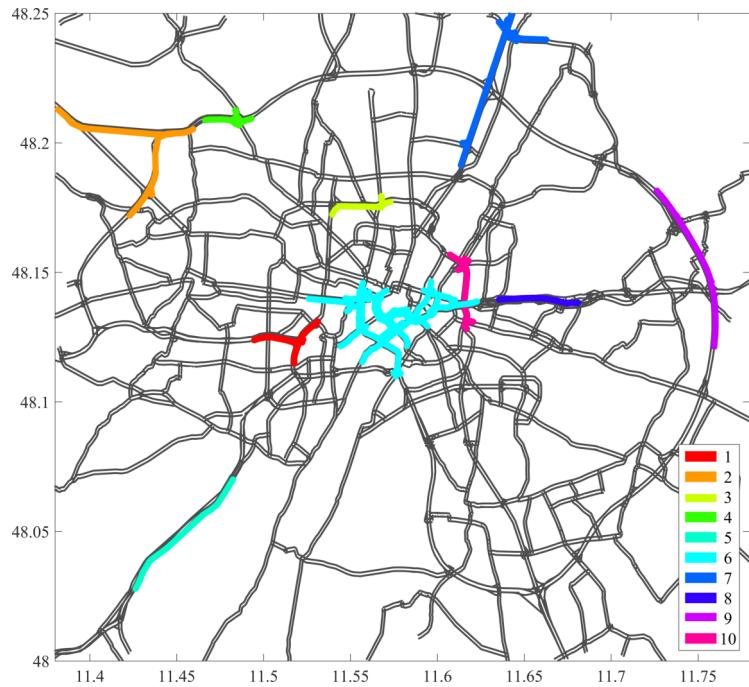
Due to the clustering algorithm these edges have a lower chance to be congested than the ones assigned before. As a result the homogeneity decreases. With even further growing clusters (decreasing α) the number of edges that are frequently in free flow state further increases. These edges are mutually highly homogeneous. As a consequence, the overall homogeneity of big clusters increases due to the high number of free flow edges. To summarize, the parameter α can be varied in order to fit the cluster properties to the application's needs.

6.4.4 Congestion Pattern Analysis

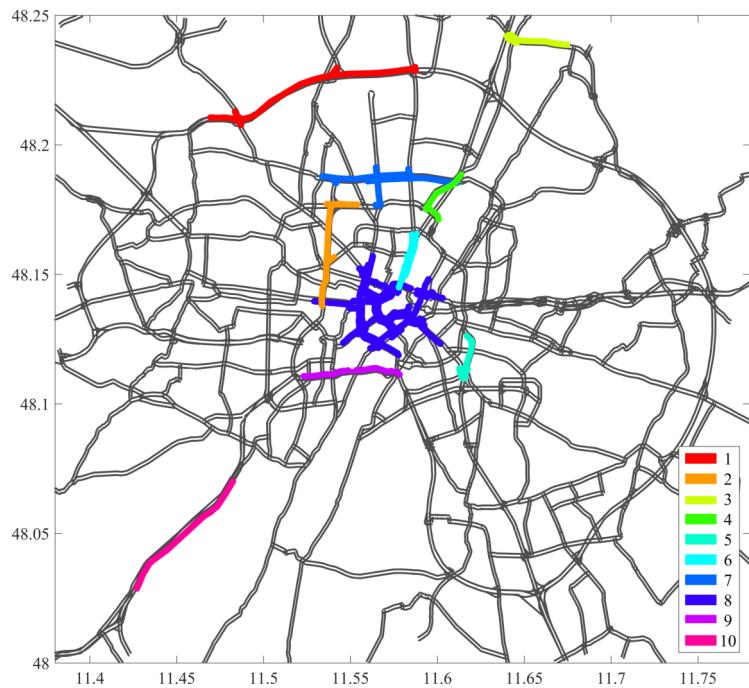
In this section a set of statistical tools is presented that support a practitioner analyzing network-wide traffic congestion using the clustering approach. The focus of this analysis is to identify spatio-temporal congestion patterns and use these for enhanced traffic prediction in networks. For the following analyses, a congestion cluster set generated with $\alpha = 0.15$ is selected (Figure 6.8). The ten largest clusters are considered which cover a length of 96.2 km (52 % of all clustered distance) during morning and 95.7 km (55 % of all clustered distance) during evening hours.

6.4.4.1 Average Levels of Congestion

Figure 6.9 depicts the total median cluster congestion κ_C of all clusters with respect to different weekdays. Monday to Thursday are grouped as they represent usual working days. On these days the level of congestion starts to increase at around 6:30am and



(a) Clusters for morning traffic with $\alpha = 0.15$



(b) Clusters for evening traffic with $\alpha = 0.15$

Figure 6.8. Static clusters chosen for the following congestion pattern analyses.

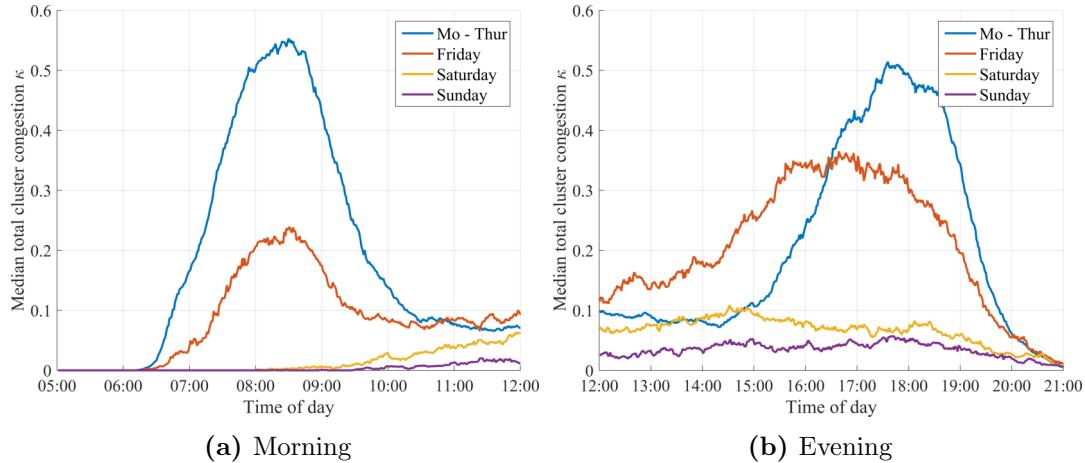


Figure 6.9. Median total cluster congestion grouped by days of the week

reaches the peak between 8am and 9am. Afterwards, the level of congestion decreases to a low level again. In the afternoon the congestion starts to increase again at around 4pm and reaches its maximum at around 6pm. All congestion dissolves at around 8pm. On Fridays the congestion during morning peak is significantly less severe. During afternoon the level of congestion starts to increase significantly earlier compared to the other working days, reaches a lower maximum and decreases earlier. The less severe congestion on Friday morning might be related to relatively less people who commute at all. The earlier, broader and less severe congestion during afternoon is probably related to the earlier time that people tend to leave from offices, which relaxes the peak time. During Saturdays and Sundays the level of congestion is low in general.

Figure 6.10 illustrates the median cluster congestion $\kappa(\mathcal{C}_i, t)$ of each cluster during working days Monday-Thursday. Under consideration of the cluster locations during morning the following observations can be done:

1. Some of the clusters get congested earlier than others. For instance, congestion in clusters 1, 2, 4, 7 begins earlier than in the clusters 3, 6, 8, 9, 10. The time offset can be approximated with the value of about half an hour. These early congested clusters are located on the freeways 'A99', 'A9' and 'A96' which conduct traffic streams from the North and West into the city. Thus, apparently there is a high traffic demand early in the morning that causes a traffic breakdown in these regions. Clusters 3, 6 and 9 on the other hand are located closer to the city center. An explanation for the delayed congestion in these clusters may be that the major freeways that guide vehicles into the more central road network work like valves with limited throughput: They get congested early, but it takes time until the limited throughput suffices to exceed the capacity of the central road network.

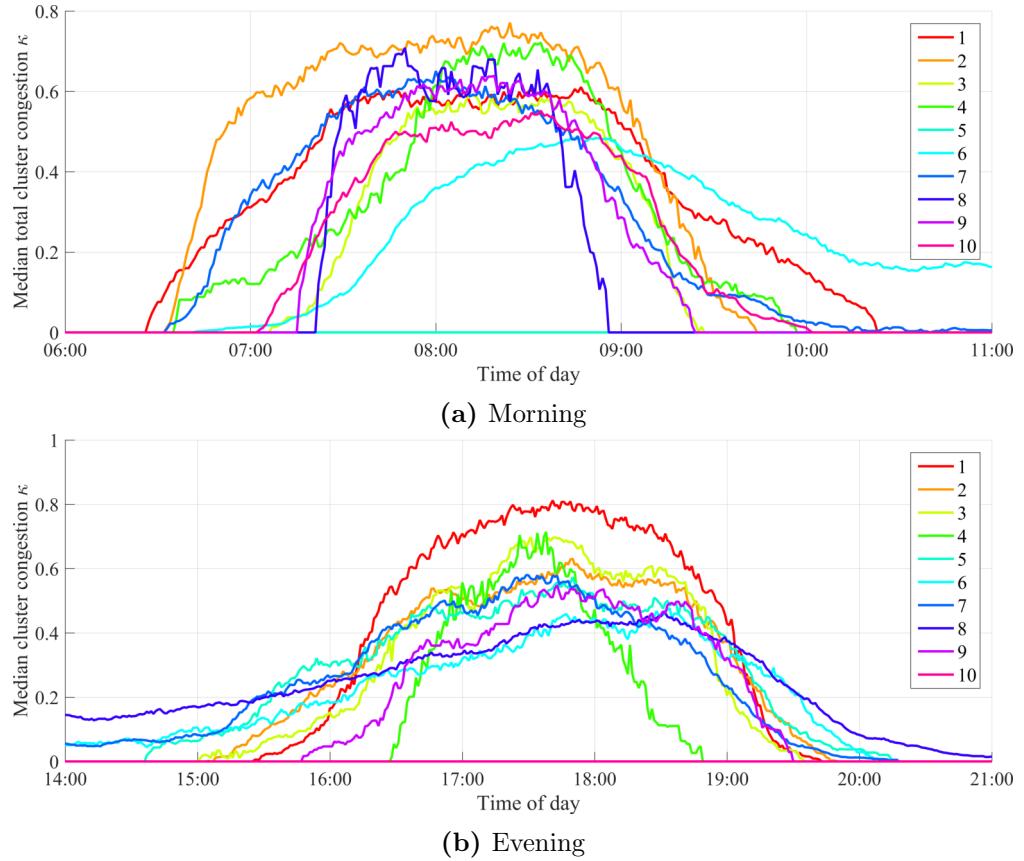


Figure 6.10. Median cluster congestion of each cluster for morning and evening peak for all Mondays-Thursdays in 2015

Though, since there are also vehicles starting their trips inside the central part this is not an exclusive explanation.

2. Cluster 6 gets congested relatively late and, during morning hours, does not get entirely free again. This cluster covers the edges that are most central. These are affected by a more balanced traffic demand over the day.
3. The median cluster congestion of cluster 5 is zero. Apparently this cluster is not a common bottleneck during working days (Mo-Thur). Though, since it has been identified as a congestion cluster there must have been frequent congestion at this location.

During evening the situation changes:

1. Cluster 1 and 10 are the only freeway clusters, whereas cluster 10 corresponds to cluster 5 during morning. Similarly, it is not congested on Mondays to Thursdays during evening hours. Cluster 1 on the other hand is severely congested with a high median κ and long peak time. An explanation for congestion occurring on more central parts of the network during evening could be that commuters leaving

the city center cause high demand on the central road infrastructure. Their limited throughput causes congestion. The outflow of these bottlenecks is not sufficient in order to induce a traffic breakdown in outer parts of the network.

2. Cluster 1 and 3 get congested at similar times. Cluster 3 is located on a parallel road to 'A99' (cluster 1). Possibly, the demand is distributed among these two roads as they constitute alternative routes for similar destinations.
3. Clusters 4 and 9 get congested slightly later than the other ones. Though, time delays between start of congestion are less obvious when compared to the ones during morning peak.

In the following, a more thorough analysis of the spatio-temporal relations between clusters is conducted in order to assess its potential to be used for network-wide traffic predictions.

6.4.4.2 Cluster Correlations

In this section correlations between the level of congestion in different clusters are examined. One way to determine the similarity of two signals is the cross-correlation. For two continuous functions f and g it is defined as (see (Stoica and Moses, 2005) for more details):

$$(f * g)(\chi) = \int_{-\infty}^{+\infty} f(t) g(t + \chi) dt \quad (6.19)$$

where χ is called the displacement or lag. Note its similarity to the convolution (eq. (4.3)), which differs only with respect to the sign of the displacement. In order to compare two discrete processes X and Y , often a normalized cross-correlation \hat{R} is considered that scales the output to values between zero and one:

$$R_{X,Y}(\chi) = \sum_{-\infty}^{+\infty} Y[i] X[i + \chi] \quad (6.20)$$

$$\hat{R}_{X,Y} = \frac{R_{X,Y}(\chi)}{\|X\| \|Y\|} \quad (6.21)$$

where $\|\cdot\|$ denotes the Euclidean norm.

The consideration of the cross-correlations is a simple way to compare two signals that may have a temporal displacement and determine whether these signals follow a similar behavior. For congestion pattern analyses this allows to compare the congestion in different clusters even if there is a temporal shift between starts and ends of congestion as it appears to be the case. Therefore, it can be applied as a flexible tool that allows

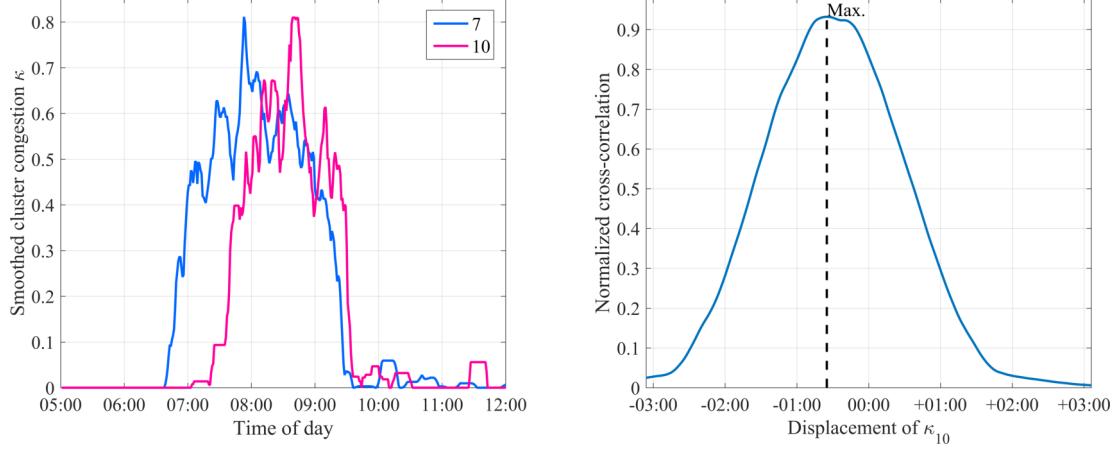


Figure 6.11. Normalized cross-correlation coefficients between congestion in cluster 7 and cluster 10 on July 14th, 2015

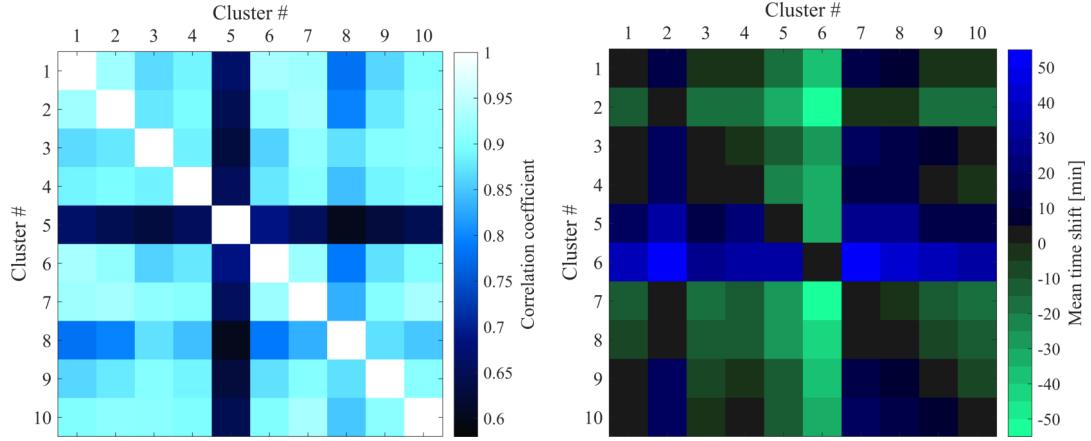


Figure 6.12. Mean normalized cross-correlation and mean time shift between morning congestion inside congestion clusters over all workdays in 2015. (Notice: A positive lag for a cluster combination (i, j) means that congestion in cluster i starts later than congestion in cluster j)

to obtain a first general overview of potential dependencies. Figure 6.11 visualizes the cross-correlation between the $\kappa(t)$ of morning cluster 7 and 10. As visible, the congestion in cluster 10 starts later. Furthermore, the peak of congestion is reached approximately 40 minutes later. On the right, the result of the cross-correlation is illustrated. It reaches its maximum at a displacement of about $\chi \approx -40$ min. Due to the definition of cross-correlation the position of the maximum can be interpreted as the amount of time that the second signal needs to be shifted in order to match best with the first signal. Formally, the lag between two signals is defined as:

$$\chi^{max} := \arg \max_{\chi} \hat{R}_{X,Y}(\chi). \quad (6.22)$$

Figure 6.12 illustrates the cross-correlation coefficient and the average lag between all

clusters averaged over 260 weekdays in 2015 during morning peak. Several observations can be made:

1. Having coefficients greater than 0.8 it can be stated that many clusters correlate strongly. Due to the overall peak hour and the median cluster congestions this result meets the expectations.
2. Cluster 5 correlates less with all other clusters. This matches with the previous analysis that cluster 5 behaves differently than the others.
3. Partially, there are strong lags between the congestion patterns: In average, congestion in cluster 6 is delayed about 40 min to 50 min compared to other cluster congestions. Clusters 2 and 7 tend to get congested earlier than other clusters. This also matches the previous findings. For comprehensiveness, the evening cluster correlations and lags are visualized in A.2.

To summarize, the cross-correlation analysis is a tool that provides an overview of the similarities between the level of congestion in the clusters. As such, it points out which clusters are more similar and which tend to get congested earlier/later than others. Though, its results may be misleading: If e.g. the duration of one congestion is significantly longer than the other one, the resulting cross-correlation is low since the maximal overlap between the two signals is reduced. Furthermore, the lag may be less obvious when the signals are complex. Therefore, cross-correlation coefficients and lags are only indicators of dependencies. Further analyses are required in order to verify hypotheses based on the cross-correlations.

6.4.4.3 Relating Starts and Ends of Daily Congestion

The previous studies revealed that, in average, there are correlations and lags between the time series of congestion among several clusters. In this section, the dependencies between congestion starts and ends on a daily basis are analyzed. The question that is investigated can be summarized as follows: 'Does the observation of congestion in one cluster allow to predict the level of congestion in another cluster for the past, current and future time?'. Especially the future time is most interesting since it enables traffic forecasts which are relevant for travelers and traffic managers.

In order to analyze starts and ends of congestion a definition of these events is required. Qualitatively, a congestion start is defined as the point in time when the level of congestion increases from a mean low level to a mean high level. Using the cross-correlation

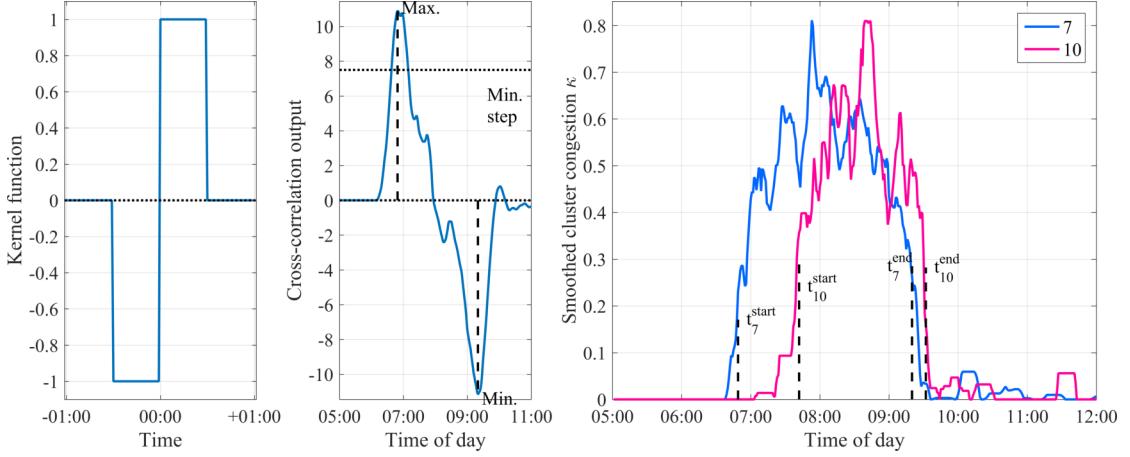


Figure 6.13. Definition of congestion start and end times. Left: The applied kernel function, mid: the result of the cross-correlation with κ of cluster 10 observed on July 14th, 2015, right: the resulting start and end times of congestion for clusters 7 and 10

function again, this definition is applied as:

$$t_i^{start} := \begin{cases} \arg \max_{\chi} R_{\kappa_i, S}(\chi) & \text{if } \max(R_{\kappa_i, S}(\chi)) > R^{thres} \\ 0 & \text{otherwise} \end{cases} \quad (6.23)$$

where

$$S(t) := \begin{cases} -1 & \text{if } 0 > t > -T_A \\ 1 & \text{if } 0 < t < T_A \\ 0 & \text{otherwise} \end{cases} \quad (6.24)$$

and $T_A, R^{thres} \in \mathbb{R}_+$ denote the time window used for averaging and a threshold that filters minor congestions, respectively. Congestion ends are defined similarly. A mandatory prerequisite is that a congestion start has been detected for the same day and cluster:

$$t_i^{end} := \begin{cases} \arg \min_{\chi} R_{\kappa_i, S}(\chi) & \text{if } t_i^{start} > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (6.25)$$

Figure 6.13 illustrates the definition with $T_A = 30$ min and $R^{thres} = 7.5$ and the resulting start and end times for congestion levels of cluster 7 and 10 on one exemplary day. Note that the definition utilizes the unnormalized cross-correlation. A threshold value of $R^{thres} = 7.5$ in combination with $T_A = 30$ min means that the average increase of congestion at the time of congestion start needs to exceed $7.5/30 = 0.25$ in order to be classified as a congestion start.

In the following the starts and ends of congestion of several pairs of clusters are discussed. Examples of clusters-pairs with significant lag in congestion, pairs with high correlation

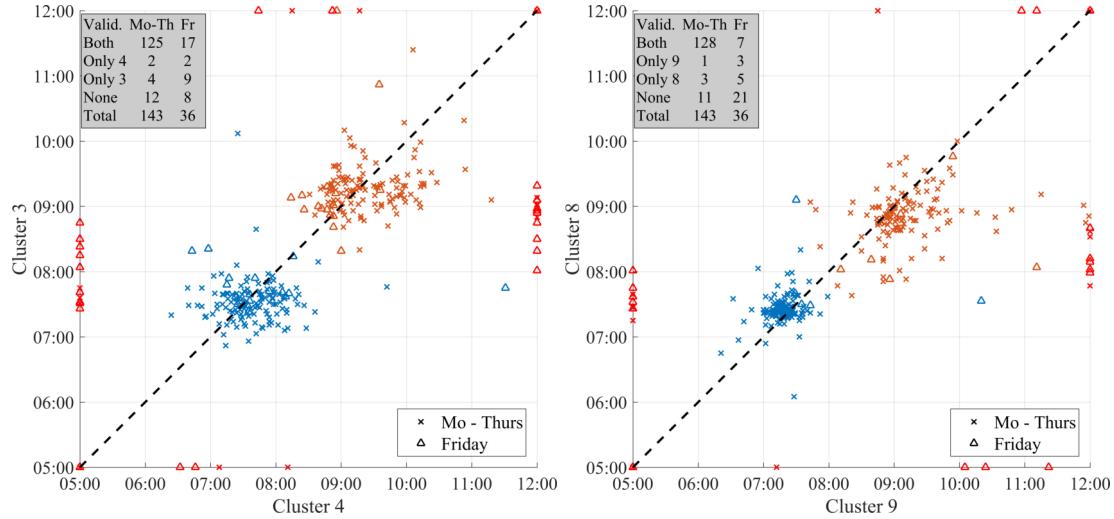


Figure 6.14. Congestion starts and ends of several cluster combinations with high correlation and minor lag. Blue markers indicate the start times and orange markers the end times of congestion. Missing congestion are visualized as symbols on the axis of the diagram.

and pairs with low correlation are presented exemplary.

Figure 6.14 illustrates two examples of cluster pairs with high correlation and minor lag for a total of 179 workdays in 2015 (due to technical reasons not for all working days data is available). For each day one (blue) marker indicates start time t^{start} of congestion and one (orange) end time t^{end} . Two types of symbols distinguish between Monday-Thursday and Friday. If no congestion start and end for a cluster is identified for a certain day, these markers are drawn on the axis of the diagram. The dotted, isochronal line is a visual aid. A table in the upper left counts the numbers of starting/ending times with respect to different combinations of congestion status. Several interesting observations can be made:

1. In most days, both, cluster 3 and 4 get congested between 7am and 8am and congestion ends around 9am.
2. Outliers occur relatively often on Fridays; during Monday-Thursday outliers are relatively rare.
3. The distribution of congestions starts in cluster 8 and 9 is much narrower. It spreads around 7.30am.
4. Both cluster pairs are located in similar parts of the network, though not on the same road. The narrow distributions of starts and ends and the low number of outliers indicate a strong dependency between these bottlenecks.

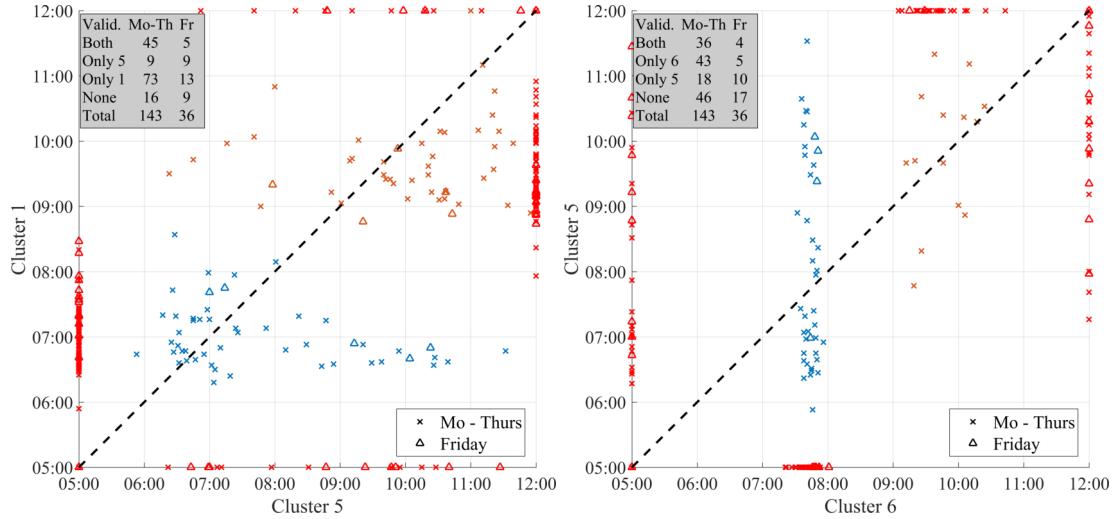


Figure 6.15. Congestion starts and ends of several cluster combinations with low correlation. Blue markers indicate the start times and orange markers the end times of congestion. Missing congestion are visualized as symbols on the edges of the diagram.

Figure 6.15 illustrate two examples of pairs with low correlation. The start and end of congestion in cluster 5 compared to cluster 1 and cluster 6 spread across the entire morning. Furthermore, there are many days in which no start and end times are detected. As a result, the observation of congestion in cluster 5 does not allow to infer information on one of the other clusters.

The most interesting cases are the ones where there is a significant lag between congestion starts and/or ends. Figure 6.16 illustrates four examples of pairs that show these characteristics. Observations are:

1. Cluster 2 gets congested relatively reliably between 6.30 and 7am in the morning. Cluster 8 and 9 do so at around 7.30am. Both distributions are narrow with only few outliers. This allows to formulate a statistically strong prediction rule: If cluster 2 happens to be congested between 6.30am and 7am, it is very likely that cluster 8 and 9 will get congested half an hour later.
2. In most cases cluster 7 gets congested earlier than cluster 3 and 6. Thus, also the observation of congestion in cluster 7 allows to deduce likely states for the other two clusters. However, the distribution of start times is wider and there are relatively many cases in which cluster 6 does not get congested at all.
3. Unfortunately, the ends of congestion distribute wider than the starts, which challenges a statistical forecast. Though, in some of the presented cases there are significant lags between congestion ends as well.

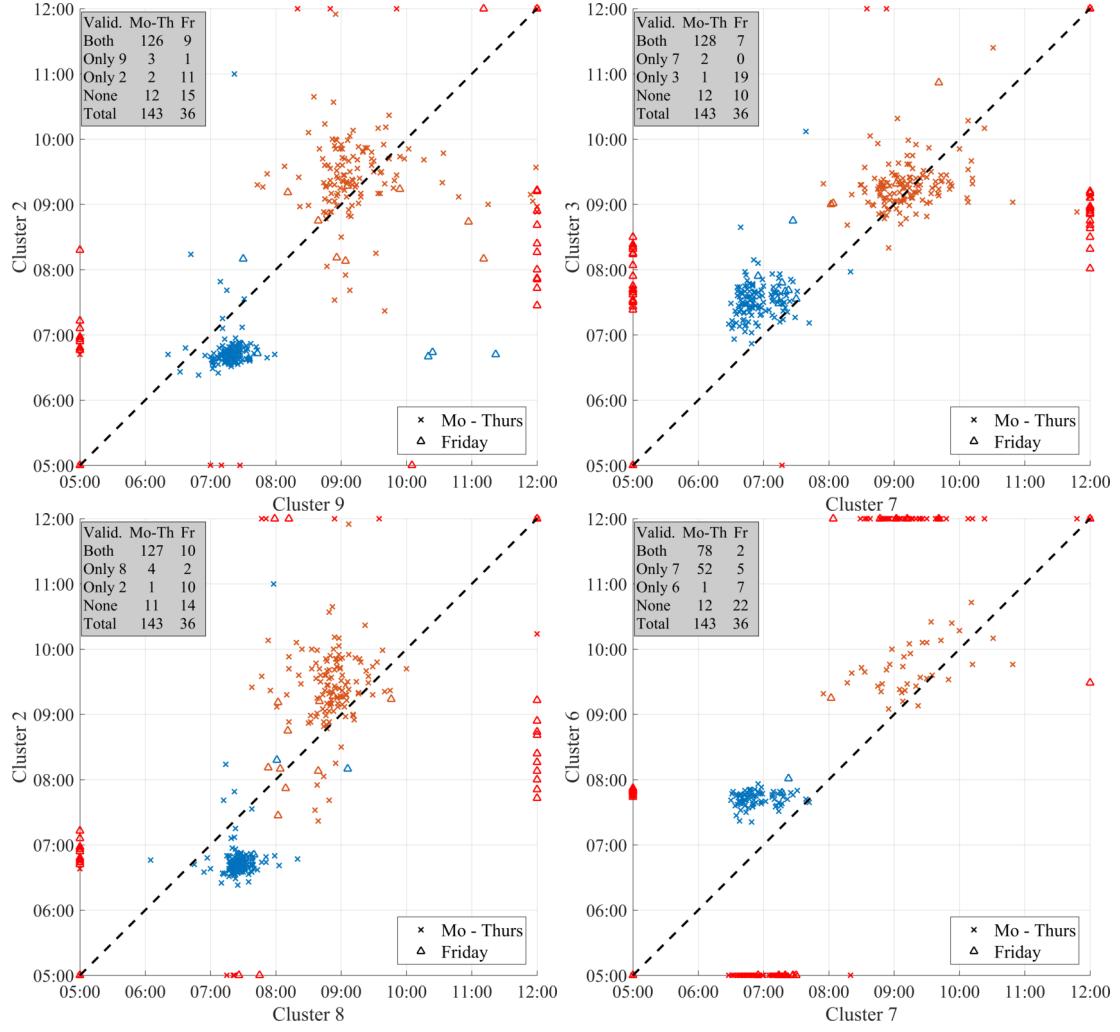


Figure 6.16. Congestion starts and ends of several cluster combinations with significant lag. Blue markers indicate the start times and orange markers the end times of congestion. Missing congestion are visualized as symbols on the edges of the diagram.

Compared to the cross-correlation analysis, this type of analysis of cluster relations provides insights into daily traffic patterns of several dependent clusters. This allows to identify the distribution of starts and ends of congestion. Given narrow distributions with few outliers, this enable accurate traffic forecasts on a network level. Moreover, if lags are low, highly correlated clusters can be used to formulate expected states of the network at different times. These expectations can be fused with sparse data in order to provide more accurate current traffic estimates. Furthermore, irregularities of traffic in the network can be detected using observed traffic conditions in the clusters. However, applying these methods one has to consider that these results are based on the assumption of an invariant traffic system. If significant changes of the infrastructural supply or road usage demand occur, gathered data may become outdated. Thus, the application of such methods requires a continuous update of data. Furthermore, this

analysis is performed only for the Munich road network. Similar results are expected to be found in other metropolitans, though, this proof remains for future work.

All in all the congestion pattern analysis in this section reveals two important aspects that are relevant for traffic forecasts in networks: First, there are recurring congestion patterns that show a high degree of regularity. They can be explained well with the commuting patterns of travelers. Second, there are spatio-temporal relations between congestion at different bottlenecks of the network. For instance, some clusters are usually congested earlier than others and some clusters on alternative roads are congested at similar times. In the following section, these findings are (inherently) applied using the forecast methods as described in section 6.3.

6.4.5 Cluster-Based Congestion Prediction

The goal of the prediction algorithm is to provide accurate predictions of TTLs for each cluster given data of the current day as well as historical data. Specifically, in this evaluation part it is shown that the consideration of network-wide congestion provides relevant information for the prediction of the TTL in one cluster.

The evaluation of the method presented in section 6.3 is structured in the following way: First, some parameters of the KNN based predictor are presented and a few variants are motivated. Next, the evaluation methodology is described. Subsequently, the errors of several variants and comparative algorithms are presented. Finally, the influence of two parameters on the prediction accuracy is analyzed.

Basically, the proposed KNN predictor and its accuracy is influenced by the following four settings:

- **Prediction horizon:** Since the predictor utilizes complete historical time-series, technically it is not limited to any prediction horizon. In this case, the TTL is based on the congestion inside the clusters generated for morning and evening period, respectively. With respect to the results of the congestion pattern analysis which detects congestion ends usually before 10am, the maximal prediction horizon is limited to this time.
- **Feature time interval:** For the computation of the STTLs which serve as features for the prediction model, a time interval has to be specified for which the TTL is aggregated. Since the STTL is supposed to be an indicator for the overall congestion inside a cluster, the time of earliest congestion in a cluster is considered. For morning hours it is set to 5am up until the current daytime.

- **Number of Neighbors:** The KNN-method searches for the K most similar (historical) neighbors and computes an average TTL reported on these days. With higher K , random fluctuations are smoothed. However, also more dissimilar neighbors are considered. In other approaches where a KNN method is applied to traffic forecasting, optimal numbers of neighbors K are found in the range of 5-30 (Zheng and Su, 2014; Myung et al., 2011; Smith et al., 2002; Bustillos and Chiu, 2011). For the following evaluations a value of $K = 10$ neighbors is used. Later, a brief analysis studies the influence of K on the accuracy.
- **Distance measure:** The distance measure is a crucial influence since it decides which features are similar and which not. In this case, two distance metrics are applied. One is the standardized Euclidean norm:

$$\Delta^s(STTL_d, STTL_{d^*}) := \left\| \sigma_T^{-1} (STTL_d - STTL_{d^*}) \right\| \quad (6.26)$$

where σ_T^{-1} denotes a vector with reciprocal standard deviations of each dimension of the training dataset. The standardization is applied in order to convert the features to the same scale, such that each feature dimension is weighted similarly.

The second metric makes use of the cluster correlations. Integrated into a distance metric, it scales the distance for the similarity of cluster i with the powered correlation values \hat{R} of cluster i and all other clusters:

$$\Delta_i^c(STTL_d, STTL_{d^*}) := \left\| \hat{R}_i^\gamma \sigma_T^{-1} (STTL_d - STTL_{d^*}) \right\| \quad (6.27)$$

where $\gamma \in \mathbb{R}$ is a parameter that controls the impact of the correlation coefficients. The idea of this metric is that clusters that are correlated strongly are more relevant for the determination of similarity. With increasing γ the coefficients of weakly correlated clusters vanish such that their impact on the selection of similar patterns is reduced.

For the following comparison three variants of the KNN based predictor are considered: One, denominated as 'KNN Uni', is a univariate formulation where for the selection of similar days for cluster i considers only the training STTLs of the same cluster. This approach can be seen as a representative for a time-series prediction which is not influenced by network dynamics. 'KNN All' uses as distance function the normalized euclidean norm (eq. (6.26)). Finally, 'KNN Cov' applies the cluster-based distance function that integrates the covariance (eq. (6.27)). γ is set to a value of 10.

As comparative algorithms three commonly applied approaches are selected: One ('All days') takes the average of all training TTL for each time for all working days and determines the average. The second builds two clusters: One average time-series for all

Mondays-Thursdays and one for all Fridays. The third further distinguishes between usual days and school holidays and thus computes four time-series in total.

As error the *RMSE* is applied:

$$RMSE(t) = \sqrt{\frac{1}{|\mathcal{D}_T| n_c} \sum_{d^*}^{|\mathcal{D}_T|} \sum_i^{n_c} \left(TTL_{d^*, i, t}^{Pred} - TTL_{d^*, i, t}^{GT} \right)^2} \quad (6.28)$$

where \mathcal{D}_T denotes the test set used for evaluation of the method. The set is chosen as a random subset of all days for which data is available. The training set \mathcal{D}_P is used to determine the *TTL* forecasts for all methods. Size of training and test are divided in a ratio of 80:20. Note, that the KNN has only one parameter such that no over-fitting of the method is possible and no validation set is required. In order to obtain robust results, average *RMSE* values over 50 iterations are considered.

Figure 6.17 depicts the *RMSE* for the six methods applied to different times of the day for predicting the *TTLs* of the morning peak until 10am. At a time of 6.30am the errors for all prediction horizons are high. The most dedicated historical average is the most accurate one. The KNN predictors are inaccurate since no congestion has been detected yet, such that no distinct prediction is possible. This changes at 7am. While the historical average stay unaffected since they do not consider the current traffic situation, the 'KNN All' and 'KNN Cov' result in significantly more accurate predictions. Compared to the situation at 6.30am the 'KNN Uni' also improved, but less than the other KNN approaches. A similar result is illustrated in the figure depicting the errors at 7.30am. Here, the KNN predictors that consider network-wide traffic for the forecast outperform the other approaches significantly. This shows that the current network-wide level of congestion is a valuable feature for accurate traffic forecasts. At later times of prediction the 'KNN Uni' variant improves, outperforms the 'KNN All' slightly and forecasts with similar accuracy as the 'KNN Cov'. This result can be accounted to the fact that all clusters are congested at that time such that the consideration of congestion in other clusters is decreasingly relevant. All in all, the 'KNN Cov' is the most accurate or similarly as accurate as the other KNN-based predictors. It indicates that a weighting of the feature dimensions with the cluster correlations is a way to increase the expressiveness of the features. The errors for the evening period are depicted in A.3. Since their results are similar to the ones described for the morning period a description is omitted here.

Figure 6.18 depicts the prediction error for the best historical average and the best KNN predictor in comparison to a free flow prediction ($TTL = 0$ for all time steps) and an RTTI prediction (the current traffic conditions are kept constant over the entire prediction period). As point in time 7.30am is chosen. As expected, the free flow

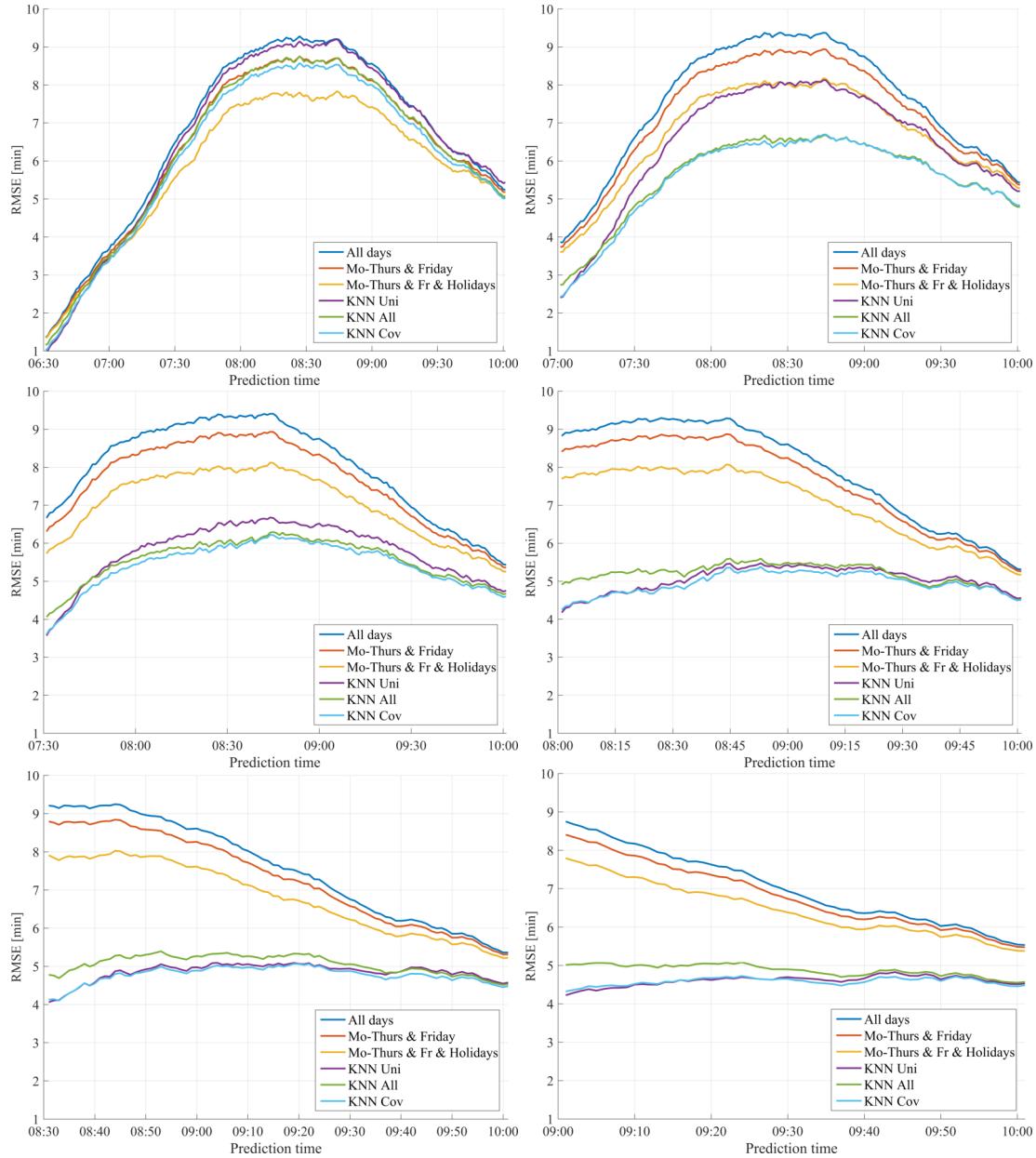


Figure 6.17. Time-dependent prediction error of several variants of a historical average and a KNN based predictor with respect to a varying start of prediction for the morning peak

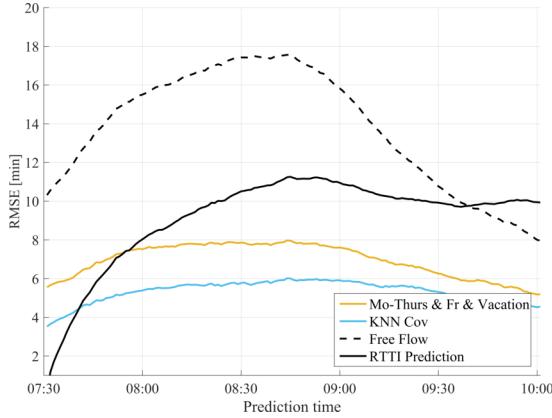


Figure 6.18. Prediction error of a Free Flow predictor and an RTTI predictor compared to the best historical and best KNN predictor

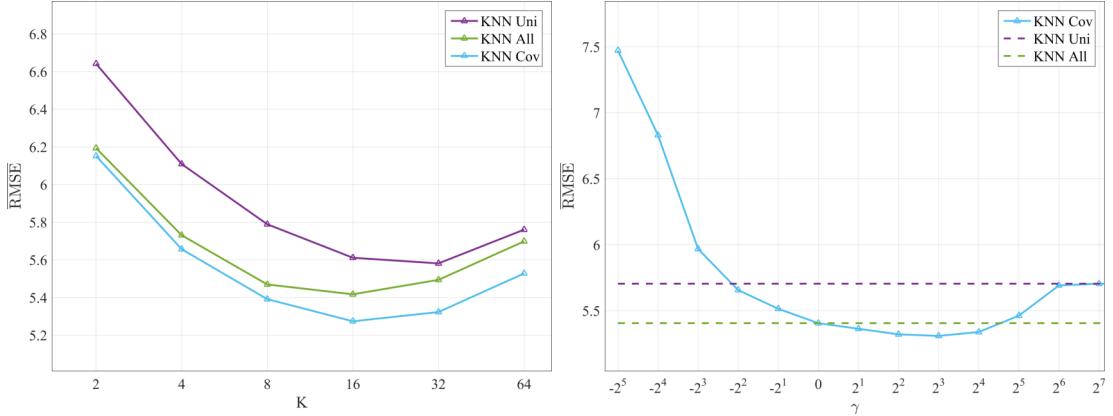


Figure 6.19. Mean RMSE of the prediction accuracy using variants of the KNN predictor with respect to the considered number of neighbors (left) and the parameter γ (right)

prediction is the most inaccurate one with errors that exceed the other approaches over the entire prediction time. The RTTI prediction is more accurate than the other ones for a short horizon of 10-15 minutes. After this time the historical average and especially the KNN predictor are more accurate. This shows that historical averages and KNN predictors are specialized for longer term prediction horizons. Since the historical average does not consider current traffic conditions at all, and the KNN-based approaches do so only indirectly, they are not able to account for accurate short-term traffic evolutions. One simple way to gain accuracy in short-term predictions would be to fuse the current traffic situation with the longer term prediction. The weight between these two methods would favor the current traffic conditions for a short-term horizon, and favor the KNN predictor for long-term prediction horizons.

Figure 6.19 (left) illustrates the mean *RMSE* of the 'KNN Cov' predictor at 7.30am with respect to the number of considered neighbors K . The convex error function concurs

with the hypothesis that too many neighbors as well as too few neighbors decrease the overall accuracy. In this case an optimal number of $K \approx 16$ neighbors is found. Though, the exact number may vary depending on the time of prediction and the number of available training days. Figure 6.19 (right) depicts the influence of the parameter γ on the mean prediction error. Apparently there exists an optimal value for parameter γ which enables a higher prediction accuracy than the 'KNN All' and the 'KNN Uni' approach. The method relates to 'KNN All' and 'KNN Uni' in the following way: With $\gamma = 0$ the correlation matrix is turned into a matrix of ones, resulting in an equal weighting of all features. Hence, it matches the 'KNN All' method. With $\gamma \rightarrow +\infty$ the correlation matrix turns into the identity matrix since all entries in the matrix converge towards zero except the main diagonal (which equals always one). This in turn matches the univariate KNN approach.

To conclude, if the features and coefficients are selected cautiously, the consideration of network-wide congestion for the prediction of TTLs in one congestion cluster yields the most accurate predictions. Even if γ is not optimized, the equal weighting of the features still produces more accurate results than the consideration of only one STTL value. Especially during the hours where there is substantial lag between the starts of congestion in several clusters, the presented approach outperforms other methods. However, the presented studies constitute only results for a limited number of available techniques. For instance, there are several ways to define correlation between cluster congestion, the current variant does not respect the lag between the time-series of cluster congestion and as predictor a simple KNN is applied. The quick advances in the fields of machine learning in the last decade rises the expectation of significant further improvements of the prediction accuracy considering more sophisticated feature selection and prediction methods.

6.5 Conclusion and Discussion

In this chapter a novel approach to analyze and predict network-wide traffic conditions was presented and its strengths and limitations were assessed. The basic idea was to reduce a road network to the regions that are frequently congested and focus the congestion analysis and forecast on these so-called congestion clusters. The benefits of this approach are:

- The result of the clustering is an abstract representation of the road network originally comprising thousands of edges with varying properties as a limited number of homogeneous regions. This simplification improves the computational efficiency and robustness of all depending methods.

- For understanding network-wide traffic congestion (as the first step for long-term traffic improvements), the abstraction enables meaningful and clear statistical analyses such as the presented correlation study. Hereby, the homogeneity of clusters contributes to the reduction of measurement noise.
- The observation of a few number of regions, whose level of congestion can be described with one variable each, simplifies network-wide congestion estimation and allows for systems with 'humans are in the loop', e.g. some kind of traffic control system.
- Compared to other edge-based methods the reduced computation efficiency and reduction of noise are especially valuable in networks where significant lags between the congestion starts and ends can be observed. For data-driven prediction, these traffic conditions inside the congestion clusters constitute expressive features which showed to improve the efficiency and accuracy of an applied machine learning algorithm.

Aside the aforementioned advantages, the clustering approach comes with some disadvantages and open issues that need to be considered in further work:

- A major drawback is that, while the clusters cover frequently congested regions, non-recurrent congestion that may occur due to special events such as weather conditions, accidents, festivals etc. is possibly not covered by the congestion clusters. In this case, traffic estimation and control algorithms based solely on the clusters would lack sufficient information. Therefore, it is necessary to develop and assess ways to identify irregularities. Since traffic congestion in the road network is connected to a certain degree, the observation of the congestion clusters for irregularities might indicate that current traffic conditions are non-recurrent.
- The statistical analysis provides spatio-temporal congestion patterns which can be used to identify dependencies between different regions in the network. However, these correlations do not reflect causal connections. This is a drawback since it does not allow to reason with certainty. For instance, if traffic conditions in one cluster change, it is not secure that a strongly correlated cluster will change as well. Rather, a statistical likelihood can be determined.
- A third issue is that the presented forecast methods are data-driven. They rely on the assumption that data collected from a system will provide information for future states of the system. In the presence of substantial changes of the infrastructure or the commuting patterns of travelers the validity of collected data decreases. In this case, outdated observations need to be discarded and new data

need to be collected. As a result, it may take time until the previous prediction accuracy is restored.

Currently, the definition of the congestion clusters as well as the TTL use only velocity data in order to detect the state of congestion as well as travel time losses. The total number of vehicles that is affected is neglected. For future developments, a fusion of this approach with flow data should be considered. The multiplication of speeds or travel time losses with flow data would allow to measure the loss of all travelers, which is a more meaningful quantity for traffic control than individual losses. Furthermore, in order to improve the interpretation of cluster correlations, trajectory data could be evaluated: For instance, if there are relatively many road users that pass through several bottlenecks during peak hours, there is an apparent strong dependency between these clusters. Finally, it is promising to apply the clustering to more cities world-wide that may have different congestion patterns. Especially metropolitans regions with several millions of inhabitants which suffer strongly from congestion could reveal interesting spatio-temporal patterns and could benefit from a network-wide traffic congestion prediction.

Chapter 7

Conclusion

7.1 Summary

The goal of this thesis was to advance the state of the art in traffic speed estimation and prediction using FCD. Therefore, three research objectives have been identified that focus on the effective use of FCD in order to improve specific traffic systems.

The first objective targets the basis of many traffic-related applications: The estimation of accurate traffic speeds on freeways. In contrast to many approaches that require flow data, a novel approach based on the Three-Phase traffic theory was presented using only FCD. Sparse data in time and space is processed in phase- and shock-wave-characteristic ways in order to reconstruct the continuous macroscopic traffic speed. In an evaluation with 101 congestion patterns, the accuracy of the PSM was compared to state-of-the-art methods. As a result, it outperformed the isotropic smoothing method by 18.4 % to 25.7 % and the GASM by 5.0 % to 16.3 %, depending on the parametrization. A subsequent run-time analysis proved its efficiency. The performed Global Sensitivity Analysis (GSA) indicated its robustness and resulted in recommendations for practical parametrization. An advantage relevant for future applications is that the method can be applied to fuse heterogeneous traffic speed data. Due to these reasons, the PSM is seen as a relevant contribution for the field of traffic speed estimation with sparse sensor data on freeways.

The second objective focused on the short-term forecast of congestion fronts in order to provide hazard warnings and improve traffic control measures. Based on a real-time velocity estimate provided by the PSM and sparse flow data, an analytical forecast model was developed. In an evaluation with real data one variant of the method was able to produce significantly more accurate short-term forecasts than comparable approaches.

This demonstrates the advantages of fusing several data sources for short-term forecasts: the high spatio-temporal accuracy of FCD and flow data gathered by loops.

The third objective focused on traffic congestion in urban road networks. First, a clustering method was developed that reduces complex road networks to those regions which are frequently congested. The consideration of congestion levels of these clusters allowed to study network-wide congestion using only a few abstracted variables. Next, the clustering approach was applied to one year of FCD in Munich city. Using several statistical tools clear spatio-temporal congestion patterns and dependencies between the clustered regions could be identified. These findings motivated the development of a network-wide traffic prediction algorithm. A data-driven approach was proposed in order to predict travel-time losses in congested clusters. In a comparison of several predictors this approach outperformed other data-driven and naive methods. These results demonstrated that these abstracted quantities are valuable features for a network-wide prediction model.

In section 2.3 four requirements which traffic state estimation and prediction algorithms are supposed to fulfill have been pointed out: Accuracy, efficiency, robustness and generality. Figure 7.1 provides a qualitative evaluation of the developed approaches with respect to these requirements. The last row summarizes the most promising open issues of each approach that could be addressed in future work.

7.2 Outlook

These developed methods show that FCD is a valuable source of traffic information which, processed in an appropriate way, allows for traffic speed estimation and prediction systems on freeways as well as urban roads. Still, there are still several limitations of FCD and current models that need to be addressed in order to increase the utility of this technology. One limitation is that the proposed as well as most other traffic models assume the homogeneity of traffic conditions among all lanes. This assumption fails in those cases where one or more lanes diverge from the main lane and measurements of different traffic conditions are fused. Another limitation is that traffic flow and density, which are fundamental quantities for analytical traffic flow models, are still difficult to deduce from FCD. A promising development that may alter these limitations is the steadily increasing number of in-vehicle sensors. Data from cameras as well as distance sensors are processed in order to create a model of traffic in the vehicle's proximity. This model includes all types of traffic-related information: the positions and speeds of surrounding vehicles, the signal status of traffic lights, the lane a vehicle is driving on etc. Thus, in the future, extended data will facilitate to assign vehicle data to certain

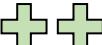
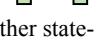
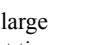
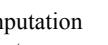
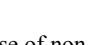
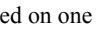
| | | Traffic Speed Estimation Chapter 4 | Congestion Front Forecast Chapter 5 | Network Congestion Analysis Chapter 6 |
|-------------|-------------|---|--|---|
| | | Accuracy | Efficiency | Robustness |
| | Future work | | | |
| Accuracy | |   <ul style="list-style-type: none"> Outperformed other state-of-the-art methods in an extensive study with 101 congestion patterns |  <ul style="list-style-type: none"> Improved accuracy compared to a naive forecast approach Study is limited to one congestion pattern |  <ul style="list-style-type: none"> Study show an accuracy gain for long-term TTL predictions Less accurate for short-term predictions |
| Efficiency | |   <ul style="list-style-type: none"> Able to process large networks in short time |   <ul style="list-style-type: none"> Comparable to the PSM efficiency since similar operations are applied Numerical integration of few fronts is fast |   <ul style="list-style-type: none"> Clusters pre-computation allows for efficient use KNN approach with relatively low sample sizes (100-1000) is fast |
| Robustness | |   <ul style="list-style-type: none"> GSA showed its robustness in case of non-optimal parametrization Applied successfully to different congestion patterns under varying data coverage |  <ul style="list-style-type: none"> Smooths data to reduce noise and eliminate outliers May fail with low data availability or complex road infrastructure |   <ul style="list-style-type: none"> Clustering is based on one parameter only KNN is a robust predictor by design |
| Generality | |  <ul style="list-style-type: none"> Can be applied to any speed data on freeways Integration of other types of data (flow, density) elaborated theoretically |  <ul style="list-style-type: none"> Requires flow and speed data with acceptable data coverage Is limited to the forecast of WMJ fronts |  <ul style="list-style-type: none"> Applicable to all link-based speed data Cluster TTLs cannot be used for routing applications |
| Future work | | <ul style="list-style-type: none"> Study of fusion capabilities with heterogeneous data sources Transfer to urban networks Study accuracy in real-time application | <ul style="list-style-type: none"> Model the in- and outflows at ramps as well as varying road capacities Usage of in-vehicle sensors to estimate flows/densities Sensitivity analysis with respect to data availability and type of congestion | <ul style="list-style-type: none"> Validate results in other cities Combine speed with flow data in clustering to consider road importance Reformulate to link-based prediction and fuse with real-time data to enable application in routing |

Figure 7.1. Overview of research objectives and evaluation with respect to the requirements of traffic state estimation and prediction algorithms (see section 2.3) as well as a summary of future work

lanes, or groups of lanes. Additionally, more sophisticated vehicle sensors will allow to determine traffic densities and flows using vehicle sensors. One advance towards this goal is already published in (Seo and Kusakabe, 2015) where traffic densities are estimated from the data of vehicle's distance sensors.

With increasing accuracies and amounts of FCD, further aspects might be investigated. One concerns the de-facto standard to represent traffic flow with the macroscopic quantities flow, density and speed. A macroscopic representation aggregates all vehicle information into the quantities flow, density and speed. The necessity to aggregate vehicle information over time and space constitutes a lower bound on the maximal accuracy of such a representation. In order to overcome these bounds and describe traffic in a general and accurate way, it is promising to advance to a meso- or microscopic level. In this thesis, the concept of a spatio-temporal vehicle occupation was introduced, which can be seen as a first step. Though, higher penetration rates of equipped vehicles and more detailed sensor data could lead towards a new level of traffic information. The development of such representations is highly encouraged. Related to this aspect are the chances, challenges and risks of augmenting data acquisition and exchange. Whilst the acquisition and exchange of data might enable applications with great effectiveness, the privacy of the travelers must be respected. Therefore, concepts are necessary that make data anonymous but preserve its essential information.

To conclude, FCD turned out to be a powerful source of traffic data that suits well for traffic speed estimation and prediction. The methods developed and described in this thesis hopefully contribute to an optimal use of this source of traffic data in order to reduce traffic congestion in the future.

Appendix A

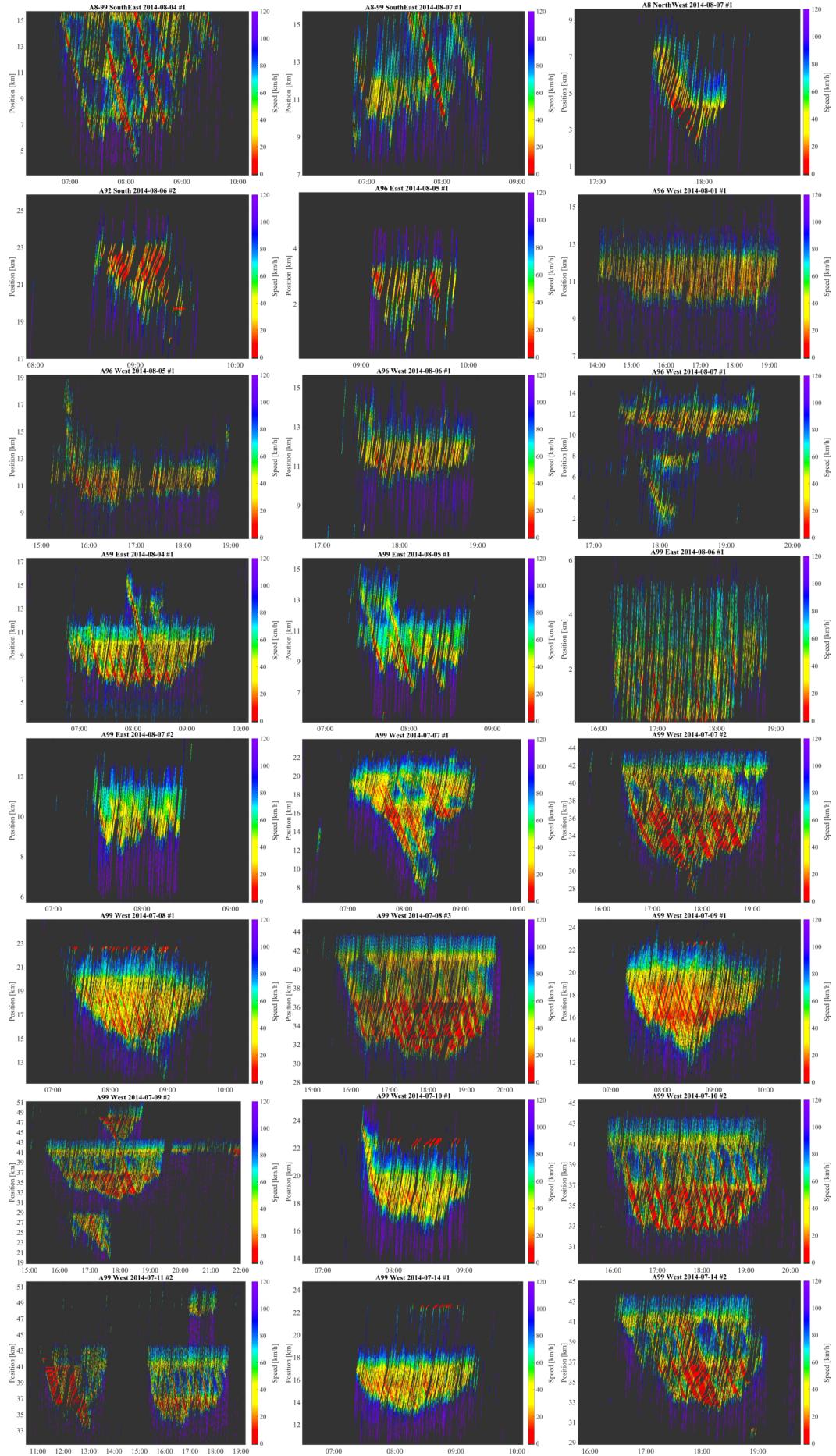
Appendix

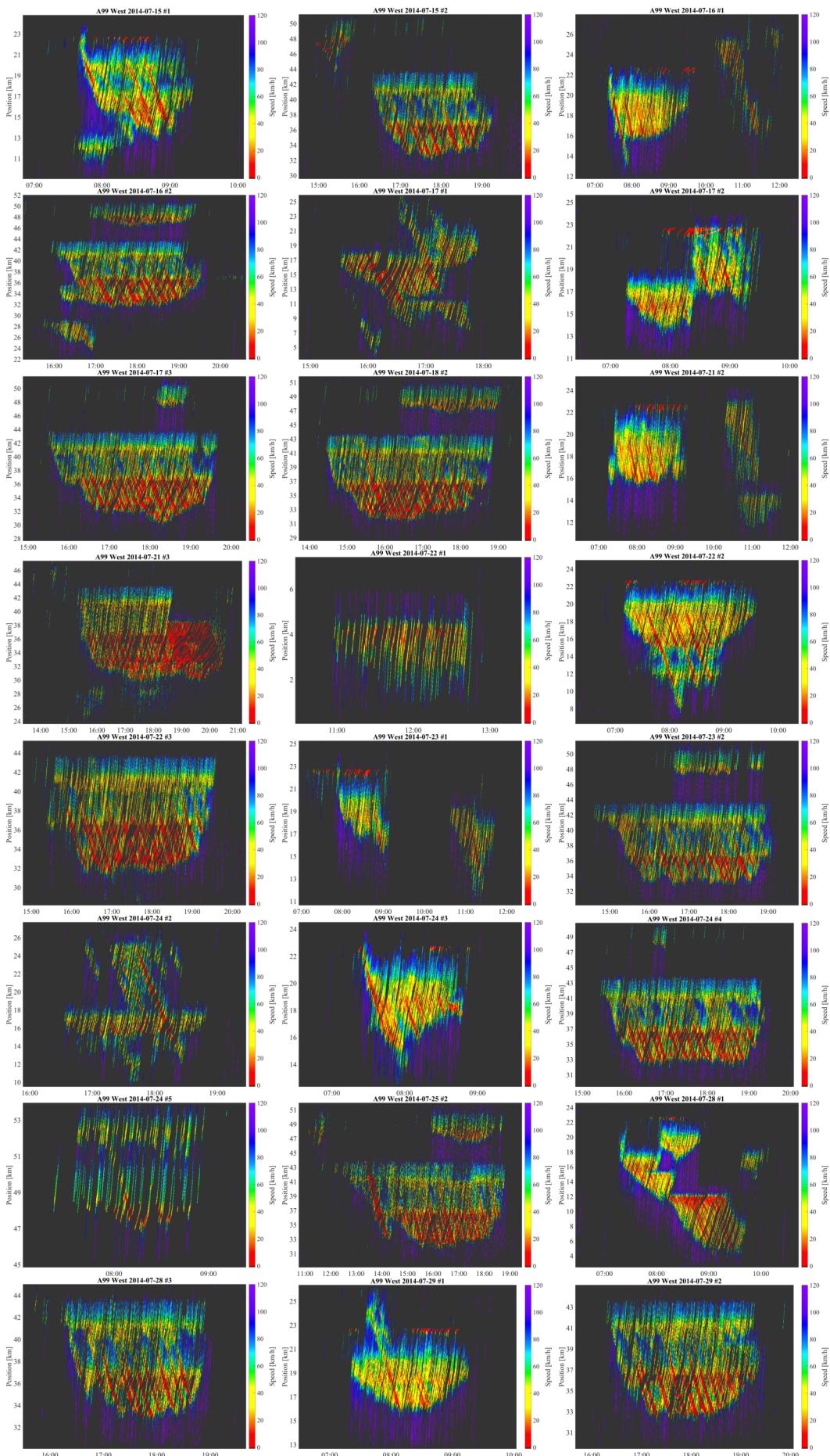
| Authors & Year | Time | Roads | Sensors | Model |
|--------------------------------|---------------|---------|--------------|-------------|
| Treiber and Helbing, 2003 | Retrospective | Freeway | Loops+FCs+.. | Data-driven |
| Treiber et al., 2010a | Retrospective | Freeway | Loops+FCs | Data-driven |
| van Lint and Hoogendoorn, 2009 | Retrospective | Freeway | FCs+Other | Data-driven |
| Kerner et al., 2013 | Retrospective | Freeway | FCs | Data-driven |
| Palmer et al., 2011 | Retrospective | Freeway | FCs | Data-driven |
| Herrera et al., 2010 | Retrospective | Freeway | FCs | Data-driven |
| Kerner et al., 2005 | Retrospective | Freeway | FCs | Data-driven |
| Seo and Kusakabe, 2015 | Retrospective | Freeway | FCs | Other |
| Bar-Gera, 2007 | Retrospective | Freeway | Other | Data-driven |
| Bhaskar et al., 2011 | Retrospective | Urban | Loops+FCs | Data-driven |
| Sarvi et al., 2003 | Retrospective | Urban | FCs | Data-driven |
| Hong et al., 2007 | Retrospective | All | FCs | Data-driven |
| Cho and Rice, 2006 | Real-Time | Freeway | Other | Data-driven |
| Wang et al., 2014 | Real-Time | Freeway | Loops+FCs+.. | Data-driven |
| Deng et al., 2013 | Real-Time | Freeway | Loops+FCs+.. | Analytical |
| Westerman et al., 1996 | Real-Time | Freeway | Loops+FCs | Data-driven |
| Astarita et al., 2006 | Real-Time | Freeway | Loops+FCs | Other |
| Blandin et al., 2013 | Real-Time | Freeway | Loops+FCs | Analytical |
| Piccoli et al., 2015 | Real-Time | Freeway | Loops+FCs | Analytical |
| Suzuki et al., 2003 | Real-Time | Freeway | Loops+FCs | Analytical |
| Qing Ou et al., 2011 | Real-Time | Freeway | Loops+Other | Other |
| Guo et al., 2009 | Real-Time | Freeway | Loops | Analytical |
| Hegyi et al., 2006 | Real-Time | Freeway | Loops | Analytical |
| Leclercq et al., 2007 | Real-Time | Freeway | Loops | Analytical |

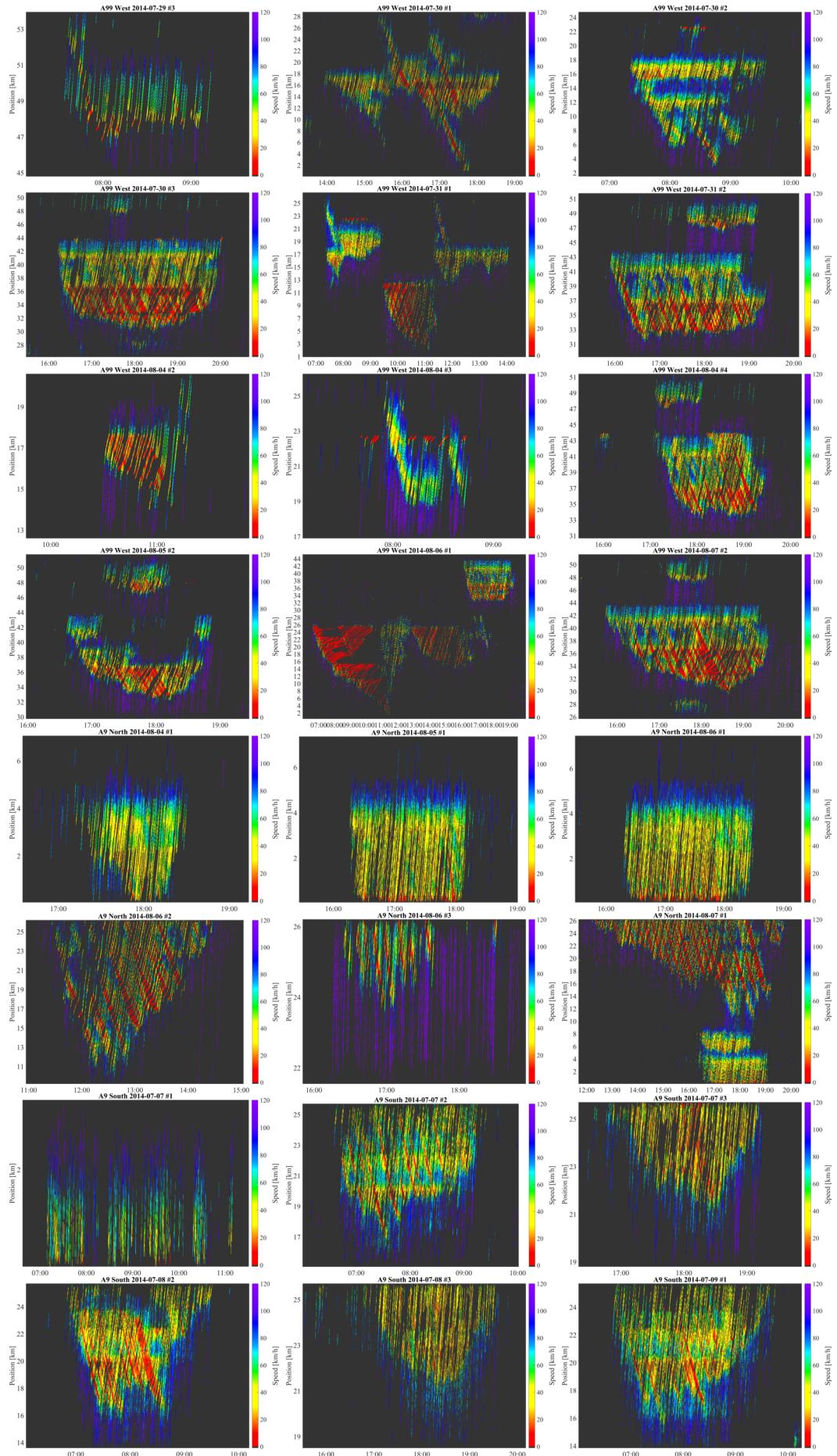
| | | | | |
|-------------------------------------|-----------|---------|--------------|-------------|
| Mihaylova et al., 2007 | Real-Time | Freeway | Loops | Analytical |
| Mihaylova et al., 2012 | Real-Time | Freeway | Loops | Analytical |
| Morarescu and Wit, 2011 | Real-Time | Freeway | Loops | Analytical |
| Morbidi et al., 2014 | Real-Time | Freeway | Loops | Analytical |
| Ngoduy, 2011 | Real-Time | Freeway | Loops | Analytical |
| Tampere and Immers, 2007 | Real-Time | Freeway | Loops | Analytical |
| van Wageningen-Kessels et al., 2010 | Real-Time | Freeway | Loops | Analytical |
| Wang and Papageorgiou, 2005 | Real-Time | Freeway | Loops | Analytical |
| Yuan et al., 2012 | Real-Time | Freeway | Loops | Analytical |
| Rempe et al., 2016a | Real-Time | Freeway | FCs | Data-driven |
| Krause et al., 2008 | Real-Time | Freeway | FCs | Data-driven |
| Sanwal and Walrand, 1995 | Real-Time | Freeway | FCs | Data-driven |
| Ygnace et al., 2000 | Real-Time | Freeway | FCs | Data-driven |
| Bekiaris-Liberis et al., 2016 | Real-Time | Freeway | FCs | Analytical |
| Herrera and Bayen, 2010 | Real-Time | Freeway | FCs | Analytical |
| Work et al., 2010 | Real-Time | Freeway | FCs | Analytical |
| Work et al., 2008 | Real-Time | Freeway | FCs | Analytical |
| Work et al., 2009 | Real-Time | Freeway | FCs | Analytical |
| Cheng et al., 2006 | Real-Time | Freeway | Other | Data-driven |
| Kwong et al., 2009 | Real-Time | Urban | Other | Data-driven |
| Bhaskar et al., 2014 | Real-Time | Urban | Loops+Other | Analytical |
| Geroliminis and Skabardonis, 2011 | Real-Time | Urban | Loops+Other | Analytical |
| Nantes et al., 2013 | Real-Time | Urban | Loops+Other | Analytical |
| Nantes et al., 2015 | Real-Time | Urban | Loops+FCs+.. | Analytical |
| Kong et al., 2009 | Real-Time | Urban | Loops+FCs | Data-driven |
| Mehran et al., 2011 | Real-Time | Urban | Loops+FCs | Analytical |
| Liu et al., 2009 | Real-Time | Urban | Loops | Analytical |
| Lu et al., 2013 | Real-Time | Urban | Loops | Analytical |
| Chen et al., 2007 | Real-Time | Urban | FCs | Data-driven |
| Feng et al., 2014 | Real-Time | Urban | FCs | Data-driven |
| Herring et al., 2010a | Real-Time | Urban | FCs | Data-driven |
| Hofleitner et al., 2012b | Real-Time | Urban | FCs | Data-driven |
| Ramezani and Geroliminis, 2012 | Real-Time | Urban | FCs | Data-driven |
| Uno et al., 2009 | Real-Time | Urban | FCs | Data-driven |
| van Zuylen et al., 2010 | Real-Time | Urban | FCs | Data-driven |
| Ban et al., 2011 | Real-Time | Urban | FCs | Analytical |
| Cheng et al., 2012b | Real-Time | Urban | FCs | Analytical |

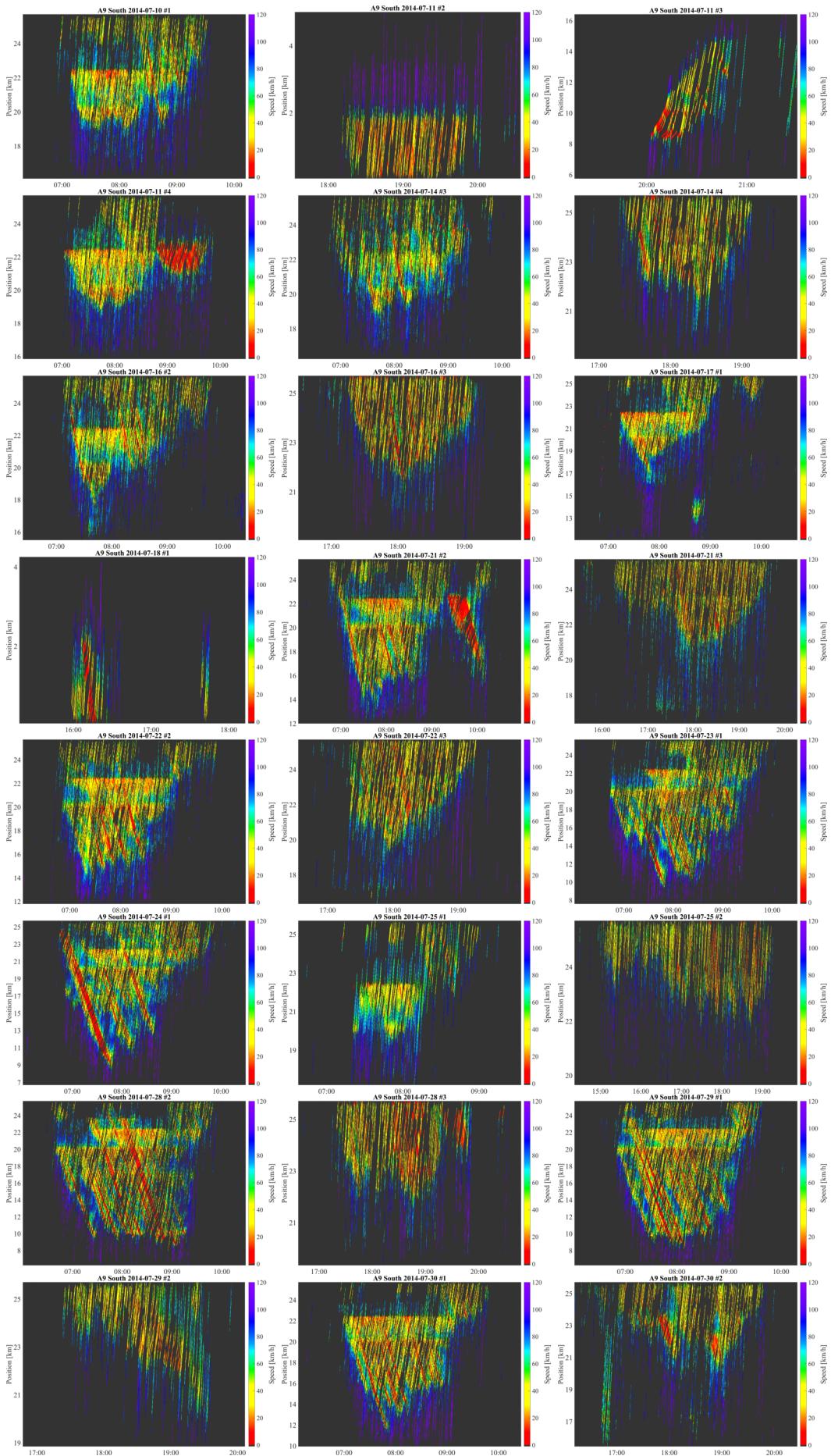
| | | | | |
|------------------------------------|------------|---------|-----------|-------------|
| Ramezani and Geroliminis, 2015 | Real-Time | Urban | FCs | Analytical |
| Hofleitner et al., 2012a | Real-Time | Urban | FCs | Analy.+DD |
| Ladino et al., 2017 | RT+Pred. | Freeway | Loops | Data-driven |
| Kerner et al., 2004 | RT+Pred. | Freeway | Loops | Analytical |
| Corrado de Fabritiis, 2008 | RT+Pred. | Freeway | FCs | Data-driven |
| Rehborn and Klenov, 2009 | Predictive | Freeway | Loops+FCs | Data-driven |
| Zou et al., 2014 | Predictive | Freeway | Loops | Data-driven |
| Dong et al., 2014 | Predictive | Freeway | Loops | Analy.+DD |
| Helbing et al., 2009 | Predictive | Freeway | Loops | Data-driven |
| Barimani et al., 2012 | Predictive | Freeway | Loops | Data-driven |
| Kamarianakis et al., 2012 | Predictive | Freeway | Loops | Data-driven |
| McFadden et al., 2001 | Predictive | Freeway | Loops | Data-driven |
| Park et al., 2011 | Predictive | Freeway | Loops | Data-driven |
| Hashemi et al., 2012 | Predictive | Freeway | Loops | Data-driven |
| van Hinsbergen et al., 2009 | Predictive | Freeway | Loops | Data-driven |
| van Lint, 2006 | Predictive | Freeway | Loops | Data-driven |
| van Lint, 2008 | Predictive | Freeway | Loops | Data-driven |
| van Lint et al., 2005 | Predictive | Freeway | Loops | Data-driven |
| Yildirimoglu and Geroliminis, 2013 | Predictive | Freeway | Loops | Data-driven |
| Zhang et al., 2014 | Predictive | Freeway | Loops | Data-driven |
| Elhenawy et al., 2014 | Predictive | Freeway | FCs | Data-driven |
| Ye et al., 2012 | Predictive | Freeway | FCs | Data-driven |
| Myung et al., 2011 | Predictive | Freeway | Other | Data-driven |
| Ma et al., 2015b | Predictive | Urban | Other | Data-driven |
| Park and Lee, 2004 | Predictive | Urban | Loops+FCs | Data-driven |
| Asif et al., 2014 | Predictive | Urban | Loops | Data-driven |
| Csikos et al., 2015 | Predictive | Urban | Loops | Data-driven |
| Min and Wynter, 2011 | Predictive | Urban | Loops | Data-driven |
| Fusco et al., 2015 | Predictive | Urban | FCs | Data-driven |
| Herring et al., 2010b | Predictive | Urban | FCs | Data-driven |
| Kong et al., 2016 | Predictive | Urban | FCs | Data-driven |
| Ma et al., 2015a | Predictive | Urban | FCs | Data-driven |
| Yao et al., 2017 | Predictive | Urban | FCs | Data-driven |
| Leonhardt, 2009 | Predictive | All | Loops+FCs | Data-driven |

Table A.1. State-of-the-art approaches for traffic speed estimation and prediction









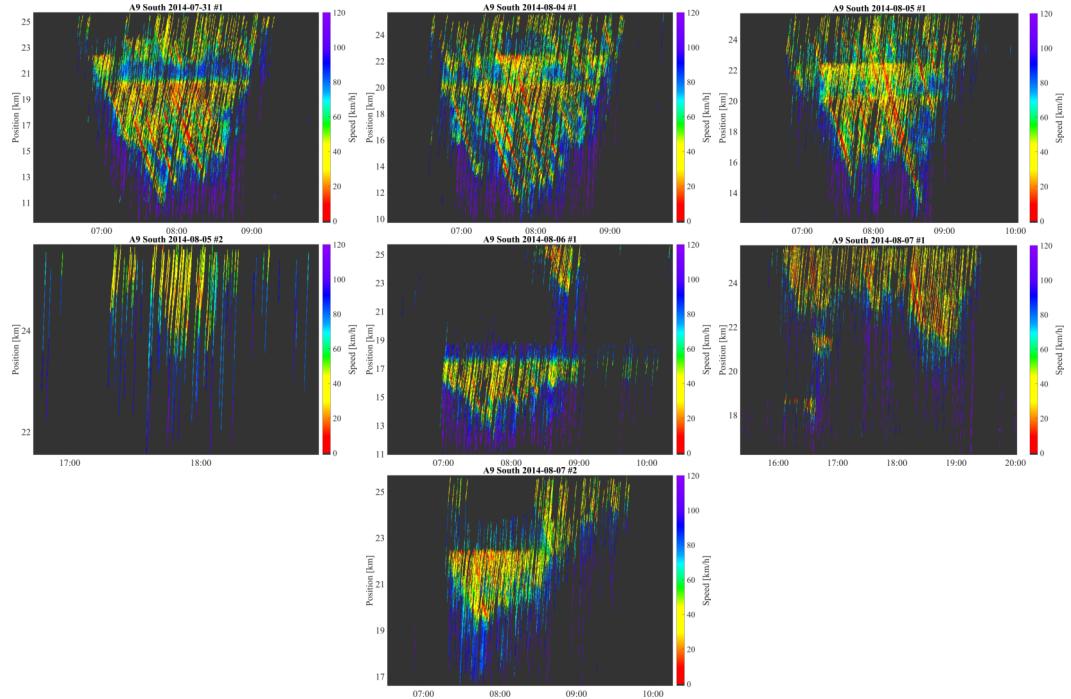


Figure A.1. Raw trajectory data of all congestion patterns collected on the freeway network surrounding Munich between July 7th, 2014 and August 7th, 2014

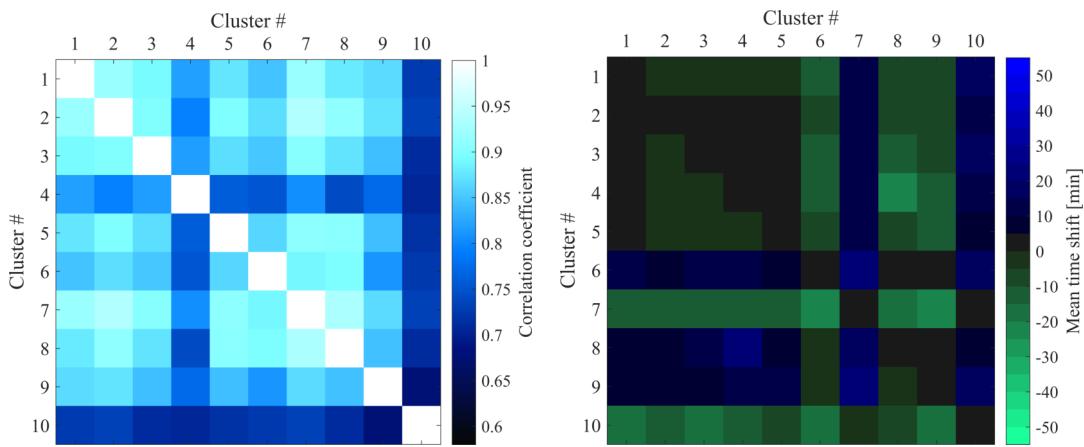


Figure A.2. Mean normalized cross-correlation and mean time shift between evening congestion inside congestion clusters over all workdays in 2015. (Notice: A positive lag for a cluster combination (i, j) means that congestion in cluster i starts later than congestion in cluster j)

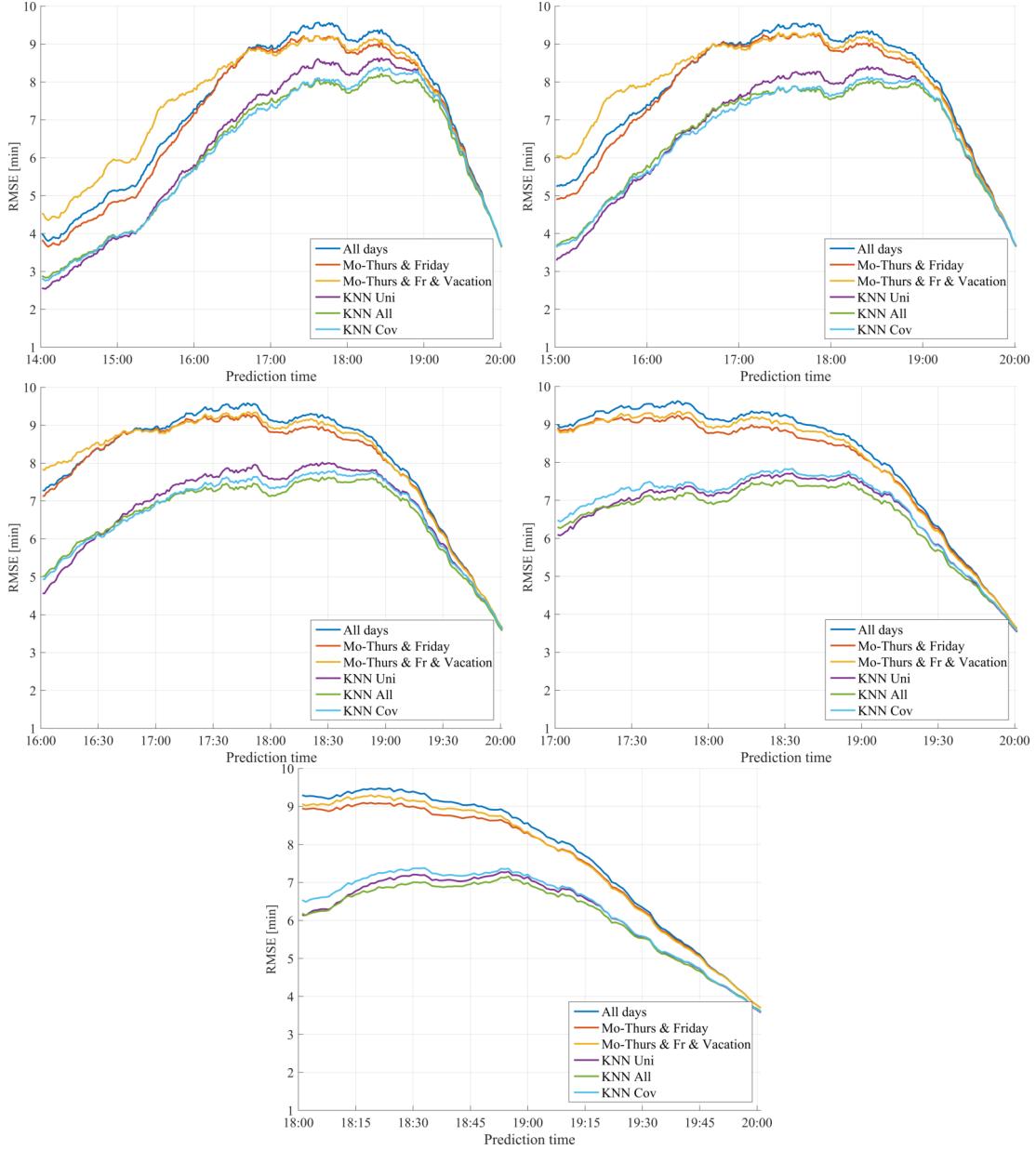


Figure A.3. Time-dependent prediction error of several variants of a historical average and a KNN based predictor with respect to a varying start of prediction for the evening peak

Abbreviations

| | |
|--------------|--|
| ACC | Adaptive Cruise Control |
| ADAS | Advanced Driver Assistance System |
| AI | Artificial Intelligence |
| ANN | Artificial Neural Network |
| ARIMA | Autoregressive Integrated Moving Average |
| CTM | Cell Transmission Model |
| FAST | Fourier-Amplitude Sensitivity Test |
| FCD | Floating Car Data |
| FC | Floating Car |
| FD | Fundamental Diagram |
| FFT | Fast Fourier Transform |
| GASM | Generalized Adaptive Smoothing Method |
| GP | General Pattern |
| GNSS | Global Navigation Satellite System |
| GSA | Global Sensitivity Analysis |
| GT | Ground Truth |
| HCT | Homogeneous Congested Traffic |
| ITS | Intelligent Transportation Systems |
| KNN | K-Nearest Neighbors |
| LWR | Lighthill-Whitham-Richards |
| ML | Machine Learning |
| OCT | Oscillating Congested Traffic |
| OSM | Open Street Map |
| PSM | Phase-based Smoothing Method |
| RTTI | Real-Time Traffic Information |
| SP | Synchronized Flow Pattern |

| | |
|----------------|---|
| STARIMA | Space-Time Autoregressive Integrated Moving Average |
| STTL | Summed Travel Time Loss |
| TT | Travel Time |
| TTL | Travel Time Loss |
| V2X | Vehicle-to-Everything |
| VBSA | Variance-Based Sensitivity Analysis |
| VSL | Variable Speed Limit |
| WMJ | Wide Moving Jam |

List of Symbols

| | |
|-----------|--|
| x | Position on a road |
| t | Time |
| L | Length of a road corridor |
| T | Specific point in time |
| N_c | Total number of vehicles providing trajectory data |
| x_c | Trajectory of vehicle c |
| v_c | Velocity of vehicle c |
| x_0 | Minimal length of a vehicle |
| T_H | Time headway |
| Ψ | Region occupied by any vehicle |
| Ω | Set of traffic phases |
| F | Phase: free flow |
| S | Phase: synchronized flow |
| J | Phase: wide moving jam |
| U | Uncertain state |
| P_p | Probability that traffic is in phase p |
| P_p^i | Criteria probabilities for phase p |
| V_{FCD} | Velocity input generated from FCD |
| V_U | Fall-back velocity |
| V_p^H | Harmonic velocity estimate for phase p |
| V_E | Estimate of macroscopic traffic velocity |
| Q | Quality estimate |
| Φ | Kernel function |
| τ | Parameter for a kernel function |

| | |
|-------------------|--|
| σ | Parameter for a kernel function |
| Γ | Convolution process |
| Λ | Normalized convolution |
| D | Data density |
| w | Continuous weighting function as input for the convolution process |
| v^{free} | Propagation speed of shock waves in Free Flow |
| v^{cong} | Propagation speed of shock waves in congested flow |
| v^{thres} | Velocity threshold of a phase |
| v_p^{dir} | Propagation speed of phase front p |
| σ_v | Function converting velocities into probabilities |
| σ_d | Function converting densities into probabilities |
| λ | Strictness of sigmoid function |
| ΔX | Space discretization |
| ΔT | Time discretization |
| \mathcal{A} | Set of estimation algorithms |
| ϵ | Relative estimation error of two algorithms |
| tr | List of tupels representing a trajectory |
| \mathcal{S} | Set of Trajectories |
| D_E | Normalized local data coverage |
| ψ | Cell occupation |
| S_i | First-order sensitivity index |
| \mathcal{G} | Graph representing a road network |
| \mathcal{V} | Nodes of the graph |
| \mathcal{E} | Edges of the graph |
| l | Length of an edge |
| V_{Lim} | Speed limit of an edge |
| V_{Rec} | Average recorded driving speed on an edge of the network |
| V_{Rel} | Relative driving speed |
| V_{Rel}^{thres} | Relative velocity threshold between free and congested traffic |
| J | Binary function that indicates the congestion status of an edge |
| \mathcal{P} | Congestion pocket as a set of connected and congested edges |
| Y | Function that counts the time two edges are congested simultaneously |
| \mathcal{C} | Static congestion cluster as a set of edges |

| | |
|---------------|---|
| α | Parameter for static clustering |
| \mathcal{K} | Set of most similar datasets resulting from KNN classification |
| δ | Intra-cluster dissimilarity |
| κ | Level of congestion |
| ρ | Cluster representativeness |
| ξ | Specificity of a cluster |
| ζ | Coverage of clusters |
| χ | Temporal lag between two congestion patterns |
| R | Cross-correlation |
| γ | Parameter controlling the influence of the correlation coefficients |
| \mathcal{D} | Set of days |

List of Figures

| | | |
|------|---|----|
| 1.1 | Structure of thesis | 3 |
| 2.1 | Fundamental diagram according to Greenshields' studies | 6 |
| 2.2 | Overview of traffic flow models | 8 |
| 2.3 | Schematic visualization of the Three Phases in flow-density and speed-density plane | 10 |
| 2.4 | The speed adaptation effect according to the Three-Phase theory | 11 |
| 2.5 | Sketches of stationary spot sensors | 13 |
| 2.6 | Section Sensor technologies | 15 |
| 2.7 | Flowchart of traffic state estimation | 17 |
| 2.8 | Publications for freeway traffic speed estimation | 20 |
| 2.9 | Publications for urban traffic speed estimation | 20 |
| 3.1 | Map of Munich and its surrounding | 25 |
| 3.2 | Average number of actively reporting vehicles and average distance covered by the fleet with respect to the day of the week | 25 |
| 3.3 | Average number of actively reporting vehicles and average distance covered by the fleet with respect to the hour of the day | 26 |
| 3.4 | Average traffic velocity as 3D plot on major roads of the Munich road network | 27 |
| 4.1 | Flow diagram: Processing FCD into a velocity estimate | 33 |
| 4.2 | Illustration of the space-time region of a vehicle with respect to its velocity and length | 34 |
| 4.3 | Illustration of the convolution process of FCD and occupation with a kernel | 35 |
| 4.4 | Velocity criterion for each phase with respect to traffic velocity | 40 |
| 4.5 | Density criterion to translate data density into phase probability | 43 |
| 4.6 | Visualization of kernel functions | 50 |
| 4.7 | Discretization of trajectories and assignment of velocities to grid cells . . | 52 |
| 4.8 | PSM data-flow to calculate phase probabilities | 54 |
| 4.9 | PSM data-flow to calculate the traffic estimate | 55 |
| 4.10 | Raw trajectory data of a congestion on A99 in eastbound direction and occupation of trajectories in time and space | 56 |
| 4.11 | Phase probabilities and phase velocities for free flow, synchronized flow and WMJ | 57 |
| 4.12 | Qualitative estimation accuracy depending on number of available trajectories | 59 |
| 4.13 | Estimated GT constructed using all trajectory data and sparse jectories used as data input for traffic speed estimation | 60 |

| | | |
|------|---|-----|
| 4.14 | Velocity estimate using an isotropic smoothing method and the GASM. Difference of estimation errors of both algorithms compared to the estimation error of the PSM | 61 |
| 4.15 | Overview of all freeways around Munich where relevant congestion occurred between July 7th, 2014 and August 7th, 2014 | 64 |
| 4.16 | Local data coverage based on the occupation of the vehicle data | 66 |
| 4.17 | Mean IMAE and error quantile of two variants of the isotropic smoothing, the GASM and the PSM with respect to the mean data coverage | 70 |
| 4.18 | Relative IMAEs and its quantiles of the PSM compared to the isotropic smoothing method | 73 |
| 4.19 | Comparison of the velocity estimates of an isotropic smoothing and the PSM | 74 |
| 4.20 | Congestion patterns that are reconstructed significantly more accurate with the PSM compared to an isotropic smoothing method | 75 |
| 4.21 | Comparison of the velocity estimates of an isotropic smoothing and the PSM | 76 |
| 4.22 | Congestion patterns that are reconstructed with similar or slightly better accuracy using an isotropic smoothing compared to the PSM | 77 |
| 4.23 | Relative IMAEs and its quantiles of the PSM compared to the GASM . . | 79 |
| 4.24 | Comparison of the velocity estimates of GASM and the PSM | 80 |
| 4.25 | Congestion patterns that are reconstructed significantly better with the PSM compared to the GASM | 81 |
| 4.26 | Comparison of the velocity estimates of GASM and the PSM | 82 |
| 4.27 | Congestion patterns that are reconstructed with similar or slightly better accuracy using GASM compared to the PSM | 83 |
| 4.28 | Congestion pattern emerging at isolated bottlenecks with respect to the bottleneck strength | 84 |
| 4.29 | Mean IMAE and error quantile of an isotropic smoothing method, the GASM and the PSM when applied to a mega-jam pattern | 85 |
| 4.30 | Complete dataset, training data and velocity estimates produced by the isotropic smoothing, the GASM and the PSM | 86 |
| 4.31 | First-order sensitivity indices of the parameters of the PSM | 88 |
| 4.32 | Run-time of the PSM with respect to an increasing domain size | 90 |
| 5.1 | Schematic illustration of the front prediction | 96 |
| 5.2 | Fundamental diagram and corresponding space-time regions with phase fronts and front propagation speeds | 97 |
| 5.3 | Schematic representation of the German freeway A9 | 100 |
| 5.4 | Congestion scenario used for evaluation. (Up) Normalized flow values collected by loop detectors. (Center) Collected FCD. (Bottom) Estimated traffic speed using the PSM | 102 |
| 5.5 | Comparison of Ground Truth fronts and predicted upstream fronts for several variations of the proposed algorithm and a prediction horizon of 5min (up) and 10min (bottom) | 104 |
| 5.6 | Accuracy of several variations of the proposed algorithm with respect to the prediction horizon. On the left, the accuracy for the prediction of the most upstream congestion front; on the right the accuracy for all other fronts | 105 |

| | | |
|------|--|-----|
| 6.1 | Overview of the steps taken to process FCD into static congestion clusters | 113 |
| 6.2 | Flowchart of common supervised learning algorithms | 117 |
| 6.3 | Data-driven TTL prediction using a KNN model | 120 |
| 6.4 | Snapshots of the traffic conditions during morning peak | 123 |
| 6.5 | Snapshots of the traffic conditions during evening peak | 124 |
| 6.6 | Results of the static clustering | 127 |
| 6.7 | Quality metrics of the evening clusters | 128 |
| 6.8 | Static clusters chosen for the following congestion pattern analyses | 129 |
| 6.9 | Median cluster congestion during morning and evening peak | 130 |
| 6.10 | Median cluster congestion of each cluster during morning and evening peak | 131 |
| 6.11 | Cross-correlation between the congestion in two clusters | 133 |
| 6.12 | Normalized cross-correlation and mean time shift of morning congestion clusters | 133 |
| 6.13 | Definition of congestion start end end times | 135 |
| 6.14 | Congestion starts and ends of several cluster combinations with high correlation and minor lag | 136 |
| 6.15 | Congestion starts and ends of several cluster combinations with low correlation | 137 |
| 6.16 | Congestion starts and ends of several cluster combinations with significant lag | 138 |
| 6.17 | Prediction errors of several variants of a historical average and a KNN predictor | 142 |
| 6.18 | Errors of an RTTI and a Free Flow Predictor compared to historical methods | 143 |
| 6.19 | Influence of the number of neighbors and correlation coefficients for KNN clustering | 143 |
| 7.1 | Research objective and comparison with requirements | 149 |
| A.1 | Raw trajectory data of all congestion patterns collected on the freeway network surrounding Munich between July 7th, 2014 and August 7th, 2014 | 159 |
| A.2 | Normalized cross-correlation and mean time shift of evening congestion clusters | 159 |
| A.3 | Prediction errors of several variants of a historical average and a KNN predictor during evening period | 160 |

List of Tables

| | | |
|-----|---|-----|
| 2.1 | Overview of (stationary) spot sensor technology | 14 |
| 2.2 | Overview of section sensor technology | 16 |
| 2.3 | Categorized approaches presented in this thesis | 21 |
| 4.1 | Overview of probabilities computed during phase identification | 39 |
| 4.2 | Parameters of the PSM | 49 |
| 4.3 | Overview of quality metrics for traffic information assessment | 67 |
| 4.4 | Mean aggregated estimation errors and error quantiles of the isotropic method, the GASM and the PSM for the scenario for the example scenario | 71 |
| A.1 | State-of-the-art approaches for traffic speed estimation and prediction . . | 154 |

Bibliography

- [1] Williams Ackaah, Gerhard Huber, and Klaus Bogenberger. “Quality evaluation method for variable speed limit systems: incident detection and warning potential”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2484 (2015), pp. 80–89.
- [2] Tarique Anwar et al. “Spatial Partitioning of Large Urban Road Networks”. In: *EDBT*. 2014, pp. 343–354.
- [3] Muhammad Tayyab Asif et al. “Spatiotemporal Patterns in Large-Scale Traffic Speed Prediction”. In: *IEEE Transactions on Intelligent Transportation Systems* 15.2 (2014), pp. 794–804.
- [4] Vittorio Astarita et al. “Motorway traffic parameter estimation from mobile phone counts”. In: *European Journal of Operational Research* 175.3 (2006), pp. 1435–1446.
- [5] A. Aw and M. Rascle. “Resurrection of Second Order Models of Traffic Flow”. In: *SIAM Journal on Applied Mathematics* 60.3 (2000), pp. 916–938.
- [6] Xuegang Ban, Peng Hao, and Zhanbo Sun. “Real time queue length estimation for signalized intersections using travel times from mobile sensors”. In: *Transportation Research Part C: Emerging Technologies* 19.6 (2011), pp. 1133–1156.
- [7] Hillel Bar-Gera. “Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel”. In: *Transportation Research Part C: Emerging Technologies* 15.6 (2007), pp. 380–391.
- [8] Nasim Barimani, Behzad Moshiri, and Mohammad Teshnehlab. “State Space Modeling and Short-Term Traffic Speed Prediction Using Kalman Filter Based on ANFIS”. In: *International Journal of Engineering and Technology* 4.2 (2012), pp. 116–120.
- [9] Nikolaos Bekiaris-Liberis, Claudio Roncoli, and Markos Papageorgiou. “Highway Traffic State Estimation With Mixed Connected and Conventional Vehicles”. In: *IEEE Transactions on Intelligent Transportation Systems* (2016), pp. 1–14.
- [10] Ashish Bhaskar, Edward Chung, and André-Gilles Dumont. “Fusing Loop Detector and Probe Vehicle Data to Estimate Travel Time Statistics on Signalized Urban Networks”. In: *Computer-Aided Civil and Infrastructure Engineering* 26.6 (2011), pp. 433–450.

- [11] Ashish Bhaskar et al. “Urban traffic state estimation: Fusing point and zone based data”. In: *Transportation Research Part C: Emerging Technologies* 48 (2014), pp. 120–142.
- [12] Christopher M. Bishop. *Pattern recognition and machine learning*. 11. (corr. printing). Information science and statistics. New York [u.a.]: Springer, 2013. ISBN: 978-0387310732.
- [13] Sébastien Blandin et al. “Phase transition model of non-stationary traffic flow: Definition, properties and solution method”. In: *Transportation Research Part B: Methodological* 52 (2013), pp. 31–55.
- [14] K. Bogenberger and S. Weikl. “Quality Management Methods for Real-Time Traffic Information”. In: *Procedia - Social and Behavioral Sciences* 54 (2012), pp. 936–945.
- [15] Klaus Bogenberger and Adolf D. May. “Advanced coordinated traffic responsive ramp metering strategies”. In: *California Partners for Advanced Transit and Highways (PATH)* (1999).
- [16] Mark Brackstone, Beshr Sultan, and Mike McDonald. “Motorway driver behaviour: Studies on car following”. In: *Transportation Research Part F: Traffic Psychology and Behaviour* 5.1 (2002), pp. 31–46.
- [17] Werner Brilon, Justin Geistefeldt, and Matthias Regler. “Reliability of freeway traffic flow: a stochastic concept of capacity”. In: *Proceedings of the 16th International symposium on transportation and traffic theory*. Vol. 125143. College Park Maryland. 2005.
- [18] Elmar Brockfeld, Reinhart Kühne, and Peter Wagner. “Calibration and validation of microscopic traffic flow models”. In: *Transportation Research Record: Journal of the Transportation Research Board* 1876 (2004), pp. 62–70.
- [19] DJ Buckley. “A semi-poisson model of traffic flow”. In: *Transportation Science* 2.2 (1968), pp. 107–133.
- [20] Brenda Bustillos and Yi-Chang Chiu. “Real-Time Freeway-Experienced Travel Time Prediction Using N -Curve and k Nearest Neighbor Methods”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2243 (2011), pp. 127–137.
- [21] Flavio Cannavó. “Sensitivity analysis for volcanic source modeling quality assessment and model selection”. In: *Computers & Geosciences* 44 (2012), pp. 52–59.
- [22] Yikai Chen et al. “A New Method For Urban Traffic State Estimation Based On Vehicle Tracking Algorithm”. In: *2007 IEEE Intelligent Transportation Systems Conference*. 2007, pp. 1097–1101.
- [23] Peng Cheng, Zhijun Qiu, and Bin Ran. “Particle filter based traffic state estimation using cell phone network data”. In: *2006 IEEE Intelligent Transportation Systems Conference*. 2006, pp. 1047–1052.
- [24] Tao Cheng, James Haworth, and Jiaqiu Wang. “Spatio-temporal autocorrelation of road network data”. In: *Journal of Geographical Systems* 14.4 (2012), pp. 389–413.

- [25] Yang Cheng et al. “An Exploratory Shockwave Approach to Estimating Queue Length Using Probe Trajectories”. In: *Journal of Intelligent Transportation Systems* 16.1 (2012), pp. 12–23.
- [26] Young Cho and John Rice. “Estimating Velocity Fields on a Freeway From Low-Resolution Videos”. In: *IEEE Transactions on Intelligent Transportation Systems* 7.4 (2006), pp. 463–469.
- [27] James W. Cooley and John W. Tukey. “An algorithm for the machine calculation of complex Fourier series”. In: *Mathematics of Computation* 19.90 (1965), p. 297.
- [28] Roberto Ragona Gaetano Valenti Corrado de Fabritiis. “Traffic Estimation and Prediction Based on Real Time Floating Car Data”. In: *Intelligent Transportation Systems (ITSC)* 11 (2008), pp. 197–203.
- [29] M Cremer and Markos Papageorgiou. “Parameter identification for a traffic flow model”. In: *Automatica* 17.6 (1981), pp. 837–843.
- [30] Alfred Csikos et al. “Traffic speed prediction method for urban networks — an ANN approach”. In: *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. 2015, pp. 102–108.
- [31] R. I. Cukier, H. B. Levine, and K. E. Shuler. “Nonlinear Sensitivity Analysis of Multiparameter Model Systems”. In: *Journal of Computational Physics* 16 (1978), pp. 1–42.
- [32] Carlos F. Daganzo. “The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory”. In: *Transportation Research Part B: Methodological* 28.4 (1994), pp. 269–287.
- [33] Carlos F. Daganzo. “The cell transmission model, part II: Network traffic”. In: *Transportation Research Part B: Methodological* 29.2 (1995), pp. 79–93.
- [34] Carlos F. Daganzo. “The nature of freeway gridlock and how to prevent it”. In: *Proceedings of the 13th International Symposium on Transportation and Traffic Theory* (1996), pp. 629–646.
- [35] Wen Deng, Hao Lei, and Xuesong Zhou. “Traffic state estimation and uncertainty quantification based on heterogeneous data sources: A three detector approach”. In: *Transportation Research Part B: Methodological* 57 (2013), pp. 132–157.
- [36] Pedro Domingos. “A few useful things to know about machine learning”. In: *Communications of the ACM* 55.10 (2012), p. 78.
- [37] Chunjiao Dong et al. “Flow rate and time mean speed predictions for the urban freeway network using state space models”. In: *Transportation Research Part C: Emerging Technologies* 43 (2014), pp. 20–32.
- [38] Leslie C Edie. *Discussion of traffic stream measurements and definitions*. Port of New York Authority, 1963.
- [39] Nils Gustaf Eissfeldt. “Vehicle-based modelling of traffic. Theory and application to environmental impact modelling”. PhD thesis. Universität zu Köln, 2004.
- [40] Lily Elefteriadou, Roger P Roess, and William R McShane. “Probabilistic nature of breakdown at freeway merge junctions”. In: *Transportation Research Record* 1484 (1995).

- [41] Mohammed Elhenawy, Hao Chen, and Hesham A. Rakha. “Dynamic travel time prediction using data clustering and genetic programming”. In: *Transportation Research Part C: Emerging Technologies* 42 (2014), pp. 82–98.
- [42] Nour-Eddin El Faouzi and Lawrence A. Klein. “Data Fusion for ITS: Techniques and Research Needs”. In: *Transportation Research Procedia* 15 (2016), pp. 495–512.
- [43] Yiheng Feng, John Hourdos, and Gary A. Davis. “Probe vehicle based real-time traffic monitoring on urban roadways”. In: *Transportation Research Part C: Emerging Technologies* 40 (2014), pp. 160–178.
- [44] Gaetano Fusco et al. “Short-term traffic predictions on large urban traffic networks: Applications of network-based machine learning models and dynamic traffic assignment models”. In: *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. 2015, pp. 93–101.
- [45] Ya Gao, Shiliang Sun, and Dongyu Shi. “Network-Scale Traffic Modeling and Forecasting with Graphical Lasso”. In: *Advances in Neural Networks – ISNN 2011*. Vol. 6676. Springer Berlin Heidelberg, 2011, pp. 151–158.
- [46] Nikolas Geroliminis and Alexander Skabardonis. “Identification and Analysis of Queue Spillovers in City Street Networks”. In: *IEEE Transactions on Intelligent Transportation Systems* 12.4 (2011), pp. 1107–1115.
- [47] Peter G. Gipps. “A behavioural car-following model for computer simulation”. In: *Transportation Research Part B: Methodological* 15.2 (1981), pp. 105–111.
- [48] Harold Greenberg. “An Analysis of Traffic Flow”. In: *Oper. Res.* 7.1 (1959), pp. 79–85.
- [49] B. D. Greenshields. “A study of traffic capacity”. In: *Proceedings of the Highway Research Board* 14 (1935), p. 448.
- [50] Jianhua Guo, Jingxin Xia, and Brian L. Smith. “Kalman filter approach to speed estimation using single loop detector measurements under congested conditions”. In: *Journal of Transportation Engineering* 135.12 (2009), pp. 927–934.
- [51] Isabelle Guyon and André Elisseeff. “An Introduction to Variable and Feature Selection”. In: *J. Mach. Learn. Res.* 3 (2003), pp. 1157–1182.
- [52] D.L Hall and Agyemang-Duah K. “Freeway capacity drop and the definition of capacity”. In: *Transportation Research Record: Journal of the Transportation Research Board* (1991), pp. 20–28.
- [53] S.Mehdi Hashemi et al. “Predicting the Next State of Traffic by Data Mining Classification Techniques”. In: *International Journal of Smart Electrical Engineering* 01.03 (2012), pp. 181–193.
- [54] J. Haworth and T. Cheng. “A Comparison of Neighbourhood Selection Techniques in Spatio-Temporal Forecasting Models”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 2 (2014), pp. 7–12.

- [55] A. Hegyi et al. “A comparison of filter configurations for freeway traffic state estimation”. In: *Intelligent Transportation Systems Conference, 2006. ITSC’06. IEEE*. 2006, pp. 1029–1034.
- [56] A. H. Heidrich. “Mapmatching von GPS-Tracks zur automatisierten Qualitätsanalyse von Verkehrsinformationen”. In: *Diploma Thesis* (2011).
- [57] Dirk Helbing et al. “Theoretical vs. Empirical Classification and Prediction of Congested Traffic States”. In: *The European Physical Journal* 2009.B 69 (2009), pp. 583–598.
- [58] Juan C. Herrera and Alexandre M. Bayen. “Incorporation of Lagrangian measurements in freeway traffic state estimation”. In: *Transportation Research Part B: Methodological* 44.4 (2010), pp. 460–481.
- [59] Juan C. Herrera et al. “Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment”. In: *Transportation Research Part C: Emerging Technologies* 18.4 (2010), pp. 568–583.
- [60] Ryan Herring et al. “Estimating arterial traffic conditions using sparse probe data”. In: *IEEE Transactions on Intelligent Transportation Systems* 13 (2010), pp. 929–936.
- [61] Ryan Herring et al. “Using Mobile Phones to Forecast Arterial Traffic Through Statistical Learning”. In: *Transport Research Board* 89 (2010).
- [62] Aude Hofleitner, Ryan Herring, and Alexandre Bayen. “Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning”. In: *Transportation Research Part B: Methodological* 46.9 (2012), pp. 1097–1122.
- [63] Aude Hofleitner et al. “Learning the Dynamics of Arterial Traffic From Probe Data Using a Dynamic Bayesian Network”. In: *IEEE Transactions on Intelligent Transportation Systems* 13.4 (2012), pp. 1679–1693.
- [64] Jun Hong et al. “Spatial and temporal analysis of probe vehicle-based sampling for real-time traffic information system”. In: *Intelligent Vehicles Symposium, 2007 IEEE*. 2007, pp. 1234–1239.
- [65] Serge Hoogendoorn, Hans van Lint, and Victor Knoop. “Macroscopic Modeling Framework Unifying Kinematic Wave Modeling and Three-Phase Traffic Theory”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2088 (2008), pp. 102–108.
- [66] Serge P Hoogendoorn and Piet HL Bovy. “Generic gas-kinetic traffic systems modeling with applications to vehicular traffic flow”. In: *Transportation Research Part B: Methodological* 35.4 (2001), pp. 317–336.
- [67] Andreas Horni, Kai Nagel, and Kay W Axhausen. *The multi-agent transport simulation MATSim*. Ubiquity Press London, 2016.
- [68] G. Huber, K. Bogenberger, and R. Bertini. “New Methods for Quality Assessment of Real Time Traffic Information”. In: *Transport Research Board* 93 (2014).
- [69] Robert A. Jacobs. “Methods for combining experts’ probability assessments”. In: *Neural computation* 7.5 (1995), pp. 867–888.

- [70] Yuxuan Ji and Nikolas Geroliminis. “On the spatial partitioning of urban transportation networks”. In: *Transportation Research Part B: Methodological* 46.10 (2012), pp. 1639–1656.
- [71] Yuxuan Ji, Jun Luo, and Nikolas Geroliminis. “Empirical Observations of Congestion Propagation and Dynamic Partitioning with Probe Data for Large-Scale Systems”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2422 (2014), pp. 1–11.
- [72] Yiannis Kamarianakis and Poulicos Prastacos. “Space–time modeling of traffic flow”. In: *Computers & Geosciences* 31.2 (2005), pp. 119–133.
- [73] Yiannis Kamarianakis, Wei Shen, and Laura Wynter. “Real-time road traffic forecasting using regime-switching space-time models and adaptive LASSO”. In: *Applied Stochastic Models in Business and Industry* 28.4 (2012), pp. 297–315.
- [74] Andreas Kendziorra, Peter Wagner, and Tomer Toledo. “A Stochastic Car Following Model”. In: *Transportation Research Procedia* 15 (2016), pp. 198–207.
- [75] B. S. Kerner and H. Rehborn. “Experimental features and characteristics of traffic jams”. In: *Physical Review E* 53.2 (1996), R1297–R1300.
- [76] Boris S. Kerner. *Breakdown in Traffic Networks: Fundamentals of Transportation Science*. Springer Berlin and Springer, 2017.
- [77] Boris S. Kerner. *Introduction to Modern Traffic Flow Theory and Control*. Springer Berlin Heidelberg, 2009. ISBN: 978-3-642-02604-1.
- [78] Boris S. Kerner. *The Physics of Traffic*. New York: Springer, 2004.
- [79] Boris S. Kerner and Sergey L. Klenov. “A theory of traffic congestion at moving bottlenecks”. In: *Journal of Physics A: Mathematical and Theoretical* 43.42 (2010), p. 425101.
- [80] Boris S. Kerner and P. Konhäuser. “Structure and parameters of clusters in traffic flow”. In: *Physical Review E* 50.1 (1994), p. 54.
- [81] Boris S. Kerner et al. “Recognition and tracking of spatial–temporal congested traffic patterns on freeways”. In: *Transportation Research Part C: Emerging Technologies* 12.5 (2004), pp. 369–400.
- [82] Boris S. Kerner et al. “Traffic dynamics in empirical probe vehicle data studied with three-phase theory: Spatiotemporal reconstruction of traffic phases and generation of jam warning messages”. In: *Physica A: Statistical Mechanics and its Applications* 392.1 (2013), pp. 221–251.
- [83] Boris S. Kerner et al. “Traffic State Detection with Floating Car Data in Road Networks”. In: *Intelligent Transportation Systems (ITSC)* (2005), pp. 44–49.
- [84] Arne Kesting and Martin Treiber. “Calibrating car-following models by using trajectory data: Methodological study”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2088 (2008), pp. 148–156.
- [85] Victor L. Knoop and Winnie Daamen. *Traffic and Granular Flow '15*. Cham: Springer International Publishing, 2016.

- [86] Wolfgang Knospe et al. “Human behavior as origin of traffic phases”. In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 65.1 Pt 2 (2002), p. 015101.
- [87] Qing-Jie Kong et al. “An Approach to Urban Traffic State Estimation by Fusing Multisource Information”. In: *IEEE Transactions on Intelligent Transportation Systems* 10.3 (2009), pp. 499–511.
- [88] Xiangjie Kong et al. “Urban traffic congestion estimation and prediction based on floating car trajectory data”. In: *Future Generation Computer Systems* 61 (2016), pp. 97–107.
- [89] Daniel Krajzewicz et al. “Recent Development and Applications of SUMO - Simulation of Urban MObility”. In: *International Journal On Advances in Systems and Measurements* 5.3&4 (2012), pp. 128–138.
- [90] Andreas Krause et al. “Toward community sensing”. In: *Proceedings of the 7th international conference on Information processing in sensor networks*. 2008, pp. 481–492.
- [91] M. Krbalek, P. Seba, and P. Wagner. “Headways in traffic flow: remarks from a physical perspective”. In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 64.6 Pt 2 (2001), p. 066119.
- [92] Eyal Krupka, Amir Navot, and Naftali Tishby. “Learning to select features using their properties”. In: *Journal of Machine Learning Research* 9.Oct (2008), pp. 2349–2376.
- [93] Max Kuhn and Kjell Johnson. *Applied predictive modeling*. Springer, 2016.
- [94] Reinhart Kühne et al. “Probabilistic description of traffic breakdowns”. In: *Physical Review E* 65.6 (2002), p. 066125.
- [95] Karric Kwong et al. “Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors”. In: *Transportation Research Part C: Emerging Technologies* 17.6 (2009), pp. 586–606.
- [96] A. Ladino et al. “A real time forecasting tool for dynamic travel time from clustered time series”. In: *Transportation Research Part C: Emerging Technologies* 80 (2017), pp. 216–238.
- [97] Jorge A. Laval. “Linking Synchronized Flow and Kinematic Waves”. In: *Traffic and Granular Flow'05*. Springer Berlin Heidelberg, 2007, pp. 521–526.
- [98] J-P. Lebacque. “The Godunov scheme and what it means for first order traffic flow models”. In: *Internaional symposium on transportation and traffic theory*. 1996, pp. 647–677.
- [99] Ludovic Leclercq, J. Laval, and E. Chevallier. “The Lagrangian coordinates and what it means for first order traffic flow models”. In: *Proc. 17th International Symposium on Transportation and Traffic Theory* (2007), pp. 735–753.
- [100] Axel Leonhardt. *Ein instanzbasiertes Lernverfahren zur Prognose von Verkehrskenngrössen unter Nutzung räumlich-zeitlicher Verkehrsmuster*. Vol. 9. Lehrstuhl für Verkehrstechnik, Technische Universität München, 2009. ISBN: 978-3-937631-09-7.

- [101] K. Levi and Y. Weiss. “Learning object detection from a small number of examples: the importance of good features”. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*. 2004, pp. 53–60.
- [102] M. Lighthill and G. Whitham. “On kinematic waves II. A theory of traffic flow on long crowded roads.” In: *Proceedings of the Royal Society A* 229 (1955), pp. 317–345.
- [103] Henry X. Liu et al. “Real-time queue length estimation for congested signalized intersections”. In: *Transportation Research Part C: Emerging Technologies* 17.4 (2009), pp. 412–427.
- [104] Christine Lotz and Malte Luks. *Qualität von on-trip Verkehrsinformationen im Straßenverkehr*. Vol. 82. Berichte der Bundesanstalt für Straßenwesen, Fahrzeugtechnik. Wirtschaftsverl. NW Verl. für neue Wiss, 2011.
- [105] Yang Lu, Ali Haghani, and Wenxin Qiao. “Macroscopic Traffic Flow Model for Estimation of Real-Time Traffic State Along Signalized Arterial Corridor”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2391 (2013), pp. 142–153.
- [106] Xiaolei Ma et al. “Large-scale transportation network congestion evolution prediction using deep learning theory”. In: *PloS one* 10.3 (2015), e0119044.
- [107] Xiaolei Ma et al. “Long short-term memory neural network for traffic speed prediction using remote microwave sensor data”. In: *Transportation Research Part C: Emerging Technologies* 54 (2015), pp. 187–197.
- [108] Reinhard Mahnke and Reinhart Kühne. “Probabilistic description of traffic breakdown”. In: *Traffic and Granular Flow'05*. Springer, 2007, pp. 527–536.
- [109] John McFadden, Wen-Tai Yang, and S. Durrans. “Application of Artificial Neural Networks to Predict Speeds on Two-Lane Rural Highways”. In: *Transportation Research Record: Journal of the Transportation Research Board* 1751 (2001), pp. 9–17.
- [110] Babak Mehran, Masao Kuwahara, and Farhana Naznin. “Implementing Kinematic Wave Theory to Reconstruct Vehicle Trajectories from Fixed and Probe Sensor Data”. In: *Procedia - Social and Behavioral Sciences* 17 (2011), pp. 247–268.
- [111] Lyudmila Mihaylova, René Boel, and Andreas Hegyi. “Freeway traffic estimation within particle filtering framework”. In: *Automatica* 43.2 (2007), pp. 290–300.
- [112] Lyudmila Mihaylova et al. “Parallelized particle and Gaussian sum particle filters for large-scale freeway traffic systems”. In: *IEEE Transactions on Intelligent Transportation Systems* 13.1 (2012), pp. 36–48.
- [113] H. S. Mika, J. B. Kreer, and L. S. Yuan. “Dual mode behavior of freeway traffic”. In: *Highway Research Record* 279 (1969), pp. 1–12.
- [114] Wanli Min and Laura Wynter. “Real-time road traffic prediction with spatio-temporal correlations”. In: *Transportation Research Part C: Emerging Technologies* 19.4 (2011), pp. 606–616.

-
- [115] Irinel-Constantin Morarescu and Carlos Canudas-de Wit. "Highway traffic model-based density estimation". In: *American Control Conference (ACC), 2011*. 2011, pp. 2012–2017.
 - [116] Fabio Morbidi et al. "A new robust approach for highway traffic density estimation". In: *Control Conference (ECC), 2014 European*. 2014, pp. 2575–2580.
 - [117] Jiwon Myung et al. "Travel Time Prediction Using k Nearest Neighbor Method with Combined Data from Vehicle Detector System and Automatic Toll Collection System". In: *Transportation Research Record: Journal of the Transportation Research Board* 2256 (2011), pp. 51–59.
 - [118] Takashi Nagatani. "The physics of traffic jams". In: *Reports on Progress in Physics* 65.9 (2002), pp. 1331–1386.
 - [119] Kai Nagel and Michael Schreckenberg. "A cellular automaton model for freeway traffic". In: *Journal de physique I* 2.12 (1992), pp. 2221–2229.
 - [120] Alfredo Nantes et al. "Real-time traffic state estimation in urban corridors from heterogeneous data". In: *Transportation Research Part C: Emerging Technologies* (2015).
 - [121] Alfredo Nantes et al. "Traffic state estimation from partial Bluetooth and volume observations: case study in the Brisbane metropolitan area". In: *Proceedings of the 18th International Conference of Hong Kong Society for Transportation Studies*. 2013.
 - [122] G. F. Newell. "A simplified car-following theory: A lower order model". In: *Transportation Research Part B: Methodological* 36.3 (2002), pp. 195–205.
 - [123] G. F. Newell. "A simplified theory of kinematic waves in highway traffic, part II: Queueing at freeway bottlenecks". In: *Transportation Research Part B: Methodological* 27.4 (1993), pp. 289–303.
 - [124] Dong Ngoduy. "Low-rank unscented Kalman filter for freeway traffic estimation problems". In: *Transportation Research Record: Journal of the Transportation Research Board* 2260 (2011), pp. 113–122.
 - [125] Daiheng Ni. *Traffic flow theory: Characteristics, experimental methods, and numerical techniques*. Amsterdam, Boston, and Heidelberg: Butterworth-Heinemann, 2016.
 - [126] J. Palmer. *Fahrzeugautonomie und verteilte Erkennung räumlich-zeitlicher Verkehrsmuster zur Nutzung in Fahrerassistenzsystemen*. Universität Tübingen, 2011.
 - [127] Jochen Palmer, Hubert Rehborn, and Ivan Gruttaduria. "Reconstruction Quality of Congested Freeway Traffic Patterns Based on Kerner's Three-Phrase Traffic Theory". In: *International Journal on Advances in Systems and Measurements* 4 (2011), pp. 168–181.
 - [128] Jungme Park et al. "Real time vehicle speed prediction using a Neural Network Traffic Model". In: *2011 International Joint Conference on Neural Networks (IJCNN 2011 - San Jose)*. 2011, pp. 2991–2996.

- [129] Taehyung Park and Sangkeon Lee. “A Bayesian Approach for Estimating Link Travel Time on Urban Arterial Road Network”. In: *Computational Science and Its Applications – ICCSA 2004*. Vol. 3043. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2004, pp. 1017–1025.
- [130] Harold J. Payne. “Models of freeway traffic and control”. In: *Mathematical models of public systems* (1971).
- [131] Benedetto Piccoli et al. “Second-order models and traffic data from mobile sensors”. In: *Transportation Research Part C: Emerging Technologies* 52 (2015), pp. 32–56.
- [132] Qing Ou et al. “A Theoretical Framework for Traffic Speed Estimation by Fusing Low-Resolution Probe Vehicle Data”. In: *IEEE Transactions on Intelligent Transportation Systems* 12.3 (2011), pp. 747–756.
- [133] Mohammed A. Quddus, Washington Y. Ochieng, and Robert B. Noland. “Current map-matching algorithms for transport applications: State-of-the art and future research directions”. In: *Transportation Research Part C: Emerging Technologies* 15.5 (2007), pp. 312–328.
- [134] Mohsen Ramezani and Nikolas Geroliminis. “On the estimation of arterial route travel time distribution with Markov chains”. In: *Transportation Research Part B: Methodological* 46.10 (2012), pp. 1576–1590.
- [135] Mohsen Ramezani and Nikolas Geroliminis. “Queue Profile Estimation in Congested Urban Networks with Probe Data”. In: *Computer-Aided Civil and Infrastructure Engineering* 30 (2015), pp. 414–432.
- [136] Hubert Rehborn and Sergey L. Klenov. “Traffic Prediction of Congested Patterns”. In: *Encyclopedia of Complexity and Systems Science*, Springer (2009).
- [137] Hubert Rehborn, Sergey L. Klenov, and Jochen Palmer. “An empirical study of common traffic congestion features based on traffic data measured in the USA, the UK, and Germany”. In: *Physica A: Statistical Mechanics and its Applications* 390.23-24 (2011), pp. 4466–4485.
- [138] Felix Rempe et al. “A phase-based smoothing method for accurate traffic speed estimation with floating car data”. In: *Transportation Research Part C: Emerging Technologies* 85 (2017), pp. 644–663.
- [139] Felix Rempe et al. “Online Freeway Traffic Estimation with Real Floating Car Data”. In: *Intelligent Transportation Systems (ITSC)* (2016), pp. 1838–1843.
- [140] Felix Rempe, Gerhard Huber, and Klaus Bogenberger. “Spatio-Temporal Congestion Patterns in Urban Traffic Networks”. In: *Transportation Research Procedia* 15 (2016), pp. 513–524.
- [141] A Reuschel. “Fahrzeuggbewegungen in der Kolonne bei gleichförmig beschleunigtem oder verzögertem Leitfahrzeug”. In: *Zeitschrift des Oesterreichischen Ingenieurund Architekten-Vereines* 95.9 (1950), pp. 59–62.
- [142] P. Richards. “Shock waves on the highway”. In: *Operations Research* 4 (1956), pp. 42–51.

- [143] Stuart J. Russell, Peter Norvig, and John Francis Canny. *Artificial intelligence: A modern approach*. 2nd ed. Englewood Cliffs, N.J.: Prentice Hall, 2003. ISBN: 978-0137903955.
- [144] A. Saltelli. *Sensitivity analysis of scientific models*. Hoboken, N.J. and Chichester: John Wiley, 2007.
- [145] A. Saltelli, S. Tarantola, and K. P.-S. Chan. “A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output”. In: *Technometrics* 41.1 (1999), p. 39.
- [146] Andrea Saltelli et al. *Global Sensitivity Analysis. The Primer*. Chichester, UK: John Wiley & Sons, Ltd, 2007. ISBN: 9780470725184.
- [147] K. Sanwal and J. Walrand. “Vehicles as Probes: California PATH Working Paper UCB-ITS-PWP-95-11: Institute of Transportation Studies, University of California, Berkeley”. In: *California PATH Working Paper UCB-ITS-PWP-95-11* (1995).
- [148] Majid Sarvi et al. “A Methodology to Identify Traffic Condition Using Intelligent Probe Vehicles”. In: *Proceedings of 10th ITS World Congress* (2003).
- [149] Martin Schoenhofer and Dirk Helbing. “Empirical Features of Congested Traffic States and Their Implications for Traffic Modeling”. In: *TRANSPORTATION SCIENCE* (2007), pp. 135–166.
- [150] Martin Schönhofer and Dirk Helbing. “Criticism of three-phase traffic theory”. In: *Transportation Research Part B: Methodological* 43.7 (2009), pp. 784–797.
- [151] Thomas Schreiter. *Vehicle-class Specific Control of Freeway Traffic*. TRAIL Thesis Series, 2012.
- [152] Thomas Schreiter et al. “Two Fast Implementations of the Adaptive Smoothing Method Used in Highway Traffic State Estimation”. In: *Intelligent Transportation Systems (ITSC)* 13th International IEEE Conference on (2010), pp. 1202–1208.
- [153] Nadine Schuessler and Kay Axhausen. “Processing Raw Data from Global Positioning Systems Without Additional Information”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2105 (2009), pp. 28–36.
- [154] Toru Seo and Takahiko Kusakabe. “Probe vehicle-based traffic state estimation method with spacing information and conservation law”. In: *Transportation Research Part C: Emerging Technologies* 59 (2015), pp. 391–403.
- [155] Glenn Shafer. *A mathematical theory of evidence*. Vol. 42. Limited paperback editions. Princeton and London: Princeton University Press, 1976.
- [156] Linda G. Shapiro and George C. Stockman. *Computer vision*. Upper Saddle River, N.J.: Prentice-Hall, 2001. ISBN: 978-0130307965.
- [157] Brian L. Smith, Billy M. Williams, and R. Keith Oswald. “Comparison of parametric and nonparametric models for traffic flow forecasting”. In: *Transportation Research Part C: Emerging Technologies* 10.4 (2002), pp. 303–321.
- [158] Petre Stoica and Randolph Moses. *Spectral analysis of signals*. Upper Saddle River, NJ: Pearson Education, 2005. ISBN: 978-0131139565.

- [159] Hironori Suzuki, Takashi Nakatsuji, and Chumchoke Nanthawichit. “Application of Probe Vehicle Data for Real-Time Traffic State Estimation and Short-Term Travel Time Prediction on a Freeway”. In: *Transportation Research Record: Journal of the Transportation Research Board* 1855 (2003), pp. 49–59.
- [160] Michael W Szeto and Denos C Gazis. “Application of Kalman filtering to the surveillance and control of traffic systems”. In: *Transportation Science* 6.4 (1972), pp. 419–439.
- [161] Chris M.J Tampere and L. H. Immers. “An Extended Kalman Filter Application for Traffic State Estimation Using CTM with Implicit Mode Switching and Dynamic Parameters”. In: *2007 IEEE Intelligent Transportation Systems Conference*. 2007, pp. 209–216.
- [162] Martin Treiber and Dirk Helbing. “An adaptive smoothing method for traffic state identification from incomplete information”. In: *Interface and Transport Dynamics* 32 (2003), pp. 343–360.
- [163] Martin Treiber and Arne Kesting. *Traffic Flow Dynamics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. ISBN: 978-3-642-32459-8.
- [164] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. “Congested traffic states in empirical observations and microscopic simulations”. In: *Physical Review E* 62.2 (2000), pp. 1805–1824.
- [165] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. “Derivation, properties, and simulation of a gas-kinetic-based, nonlocal traffic model”. In: *Physical Review E* 59.1 (1999), p. 239.
- [166] Martin Treiber, Arne Kesting, and R. Wilson. “Reconstructing the Traffic State by Fusion of Heterogeneous Data”. In: *Computer-Aided Civil and Infrastructure Engineering* 2011.26 (2010), pp. 408–419.
- [167] Martin Treiber, Arne Kesting, and Dirk Helbing. “Three-phase traffic theory and two-phase models with a fundamental diagram in the light of empirical stylized facts”. In: *Transportation Research Part B: Methodological* 44.8-9 (2010), pp. 983–1000.
- [168] Mario F. Triola. *Elementary statistics*. 12th edition. Always learning. Boston: Pearson, 2014. ISBN: 9780321836960.
- [169] Nobuhiro Uno et al. “Using Bus Probe Data for Analysis of Travel Time Variability”. In: *Journal of Intelligent Transportation Systems* 13.1 (2009), pp. 2–15.
- [170] Chris P. I. J. van Hinsbergen et al. “Localized Extended Kalman Filter for Scalable Real-Time Traffic State Estimation”. In: *IEEE Transactions on Intelligent Transportation Systems* 13.1 (2012), pp. 385–394.
- [171] C.P.IJ. van Hinsbergen, J.W.C. van Lint, and H. J. van Zuylen. “Bayesian committee of neural networks to predict travel times with confidence intervals”. In: *Transportation Research Part C: Emerging Technologies* 17.5 (2009), pp. 498–509.

- [172] Hans van Lint and Tamara Djukic. “Applications of Kalman Filtering in Traffic Management and Control”. In: *New Directions in Informatics, Optimization, Logistics, and Production* (2014), pp. 59–91.
- [173] Hans van Lint and Serge Hoogendoorn. “A Robust and Efficient Method for Fusing Heterogeneous Data from Traffic Sensors on Freeways”. In: *Computer-Aided Civil and Infrastructure Engineering* 24 (2009), pp. 1–17.
- [174] J. van Lint. “Empirical Evaluation of New Robust Travel Time Estimation Algorithms”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2160 (2010), pp. 50–59.
- [175] J. W. van Lint. “Reliable Real-Time Framework for Short-Term Freeway Travel Time Prediction”. In: *Journal of Transportation Engineering* 132.12 (2006), pp. 921–932.
- [176] J. W. C. van Lint. “Online Learning Solutions for Freeway Travel Time Prediction”. In: *IEEE Transactions on Intelligent Transportation Systems* 9.1 (2008), pp. 38–47.
- [177] J.W.C. van Lint, S. P. Hoogendoorn, and H. J. van Zuylen. “Accurate freeway travel time prediction with state-space neural networks under missing data”. In: *Transportation Research Part C: Emerging Technologies* 13.5-6 (2005), pp. 347–369.
- [178] Femke van Wageningen-Kessels et al. “Lagrangian formulation of multiclass kinematic wave model”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2188 (2010), pp. 29–36.
- [179] Henk J. van Zuylen, Fangfang Zheng, and Yusen Chen. “Using Probe Vehicle Data for Traffic State Estimation in Signalized Urban Networks”. In: *Traffic Data Collection and its Standardization*. Vol. 144. International Series in Operations Research & Management Science. Springer New York, 2010, pp. 109–127.
- [180] Eleni I. Vlahogianni, John C. Golias, and Matthew G. Karlaftis. “Short-term traffic forecasting: Overview of objectives and methods”. In: *Transport Reviews* 24.5 (2004), pp. 533–557.
- [181] Eleni I. Vlahogianni, Matthew G. Karlaftis, and John C. Golias. “Short-term traffic forecasting: Where we are and where we’re going”. In: *Transportation Research Part C: Emerging Technologies* 43 (2014), pp. 3–19.
- [182] Femke Wageningen-Kessels. *Multi class continuum traffic flow models: Analysis and simulation methods*. Doctoral thesis at the department of Transport and Planning, TU Delft, 2013, p. 268.
- [183] Femke van Wageningen-Kessels et al. “Genealogy of traffic flow models”. In: *EURO Journal on Transportation and Logistics* 4.4 (2015), pp. 445–473.
- [184] Peter Wagner. “Analyzing fluctuations in car-following”. In: *Transportation Research Part B: Methodological* 46.10 (2012), pp. 1384 –1392.
- [185] Peter Wagner and Ihor Lubashevsky. “Empirical basis for car-following theory development”. In: *arXiv preprint cond-mat/0311192* (2003).

- [186] Chunhui Wang et al. “Fusing heterogeneous traffic data by Kalman filters and Gaussian mixture models”. In: *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on.* 2014, pp. 276–281.
- [187] Yibing Wang and Markos Papageorgiou. “Real-time freeway traffic state estimation based on extended Kalman filter: a general approach”. In: *Transportation Research Part B: Methodological* 39.2 (2005), pp. 141–167.
- [188] Marcel Westerman, Remco Litjens, and Jean-Paul Linnartz. “Integration of probe vehicle and induction loop data: Estimation of travel times and automatic incident detection”. In: *California Partners for Advanced Transit and Highways (PATH)* (1996).
- [189] R. Wiedemann. “Simulation des Straßenverkehrsflusses”. In: *Schriftenreihe des IfV* 8 (1974).
- [190] D. B. Work et al. “A Traffic Model for Velocity Data Assimilation”. In: *Applied Mathematics Research eXpress* (2010).
- [191] Daniel B. Work et al. “An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices”. In: *2008 47th IEEE Conference on Decision and Control.* 2008, pp. 5062–5068.
- [192] Daniel B. Work et al. “Lagrangian sensing: traffic estimation with mobile devices”. In: *2009 American Control Conference.* 2009, pp. 1536–1543.
- [193] Baozhen Yao et al. “Short-Term Traffic Speed Prediction for an Urban Corridor”. In: *Computer-Aided Civil and Infrastructure Engineering* 32.2 (2017), pp. 154–169.
- [194] Qing Ye, W. Y. Szeto, and S. C. Wong. “Short-Term Traffic Speed Forecasting Based on Data Recorded at Irregular Intervals”. In: *IEEE Transactions on Intelligent Transportation Systems* 13.4 (2012), pp. 1727–1737.
- [195] Jean-Luc Ygnace et al. “Travel time estimation on the san francisco bay area network using cellular phones as probes”. In: *California Partners for Advanced Transit and Highways (PATH)* (2000).
- [196] Mehmet Yildirimoglu and Nikolas Geroliminis. “Experienced travel time prediction for congested freeways”. In: *Transportation Research Part B: Methodological* 53 (2013), pp. 45–63.
- [197] Yufei Yuan et al. “Real-Time Lagrangian Traffic State Estimator for Freeways”. In: *IEEE Transactions on Intelligent Transportation Systems* 13.1 (2012), pp. 59–70.
- [198] L. A. Zadeh. “Fuzzy sets”. In: *Information and Control* 8.3 (1965), pp. 338–353.
- [199] Xiao Zhang et al. “Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction”. In: *Transportation Research Part C: Emerging Technologies* 43 (2014), pp. 127–142.
- [200] Zuduo Zheng and Dongcai Su. “Short-term traffic volume forecasting: A k-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm”. In: *Transportation Research Part C: Emerging Technologies* 43 (2014), pp. 143–157.

- [201] Yajie Zou et al. “A space–time diurnal method for short-term freeway travel time prediction”. In: *Transportation Research Part C: Emerging Technologies* 43 (2014), pp. 33–49.