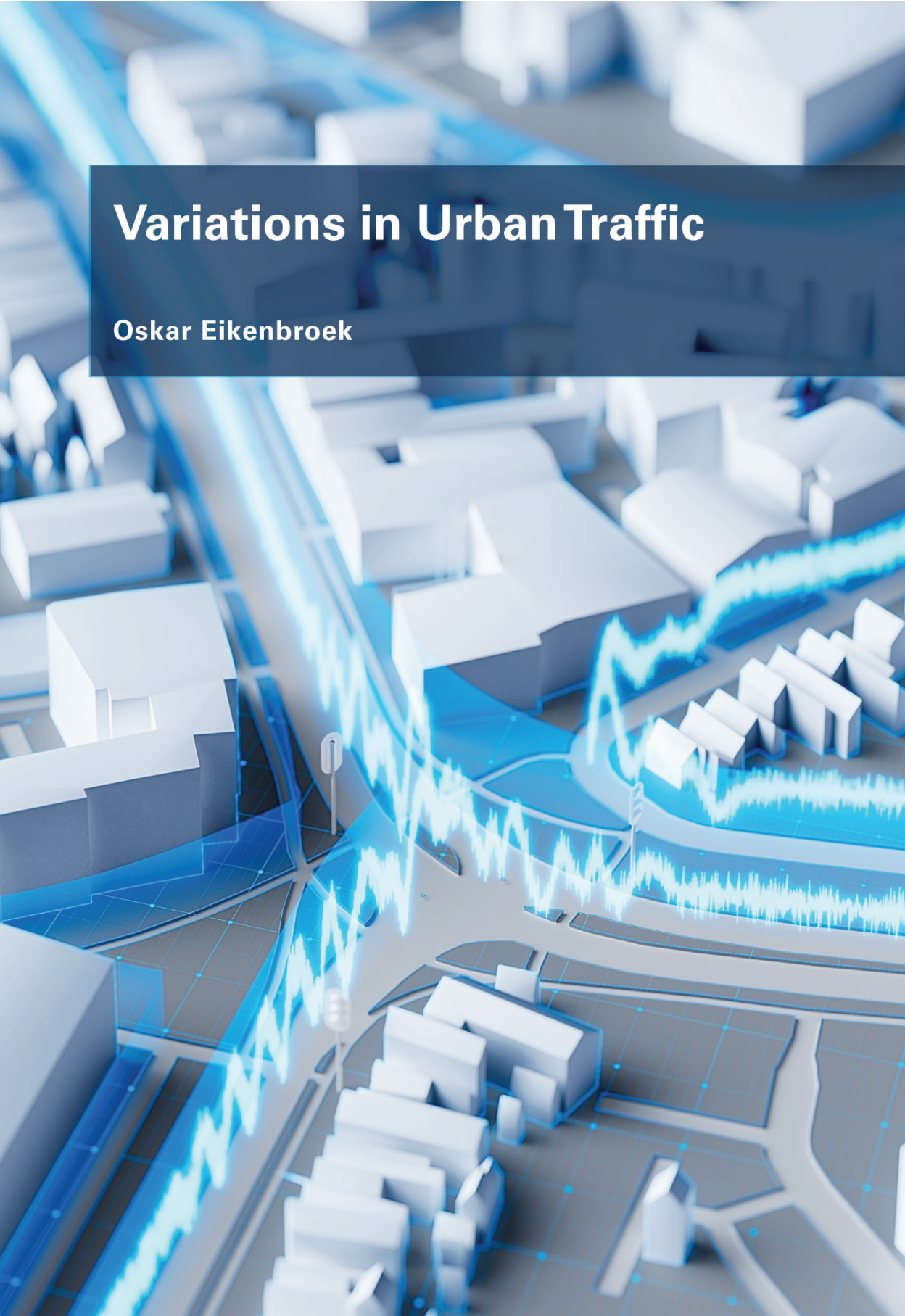


Variations in Urban Traffic

Oskar Eikenbroek

VARIATIONS IN URBAN TRAFFIC

Oskar Adriaan Louis Eikenbroek

VARIATIONS IN URBAN TRAFFIC

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr.ir. A. Veldkamp,
on account of the decision of the Doctorate Board,
to be publicly defended
on Friday 17 February 2023 at 16.45 hours

by

Oskar Adriaan Louis Eikenbroek
born on the 4th of December, 1991
in Emmen, The Netherlands

This dissertation has been approved by:

Supervisor:
prof.dr.ir. E.C. van Berkum

Co-supervisor:
prof.dr.ir. M.R.K. Mes

TRAIL Thesis Series no. T2023/2, the Netherlands Research School TRAIL
Research School TRAIL
P.O. Box 5017
2600 GA Delft
The Netherlands
E-mail: info@rsTRAIL.nl

DSI Ph.D. Thesis Series No. 23-001
Digital Society Institute
P.O. Box 217
7500 AE Enschede
The Netherlands

ISBN: 978-90-5584-321-3
DOI: 10.3990/1.9789055843213
ISSN: 2589-7721
Cover art work: Tiny Giants, tinygiants.nl, @tinygiants3d
URL: https://doi.org/10.3990/1.9789055843213

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

This dissertation is the result of a PhD research carried out from 2017 to 2022 at the University of Twente, Faculty of Engineering Technology, Department of Civil Engineering. This research was funded by the Dutch Research Council (NWO), project number 439.16.103.

Printed in the Netherlands

Graduation committee:

prof.dr.ir. H.F.J.M. Koopman	University of Twente, chair/secretary
prof.dr.ir. E.C. van Berkum	University of Twente, supervisor
prof.dr.ir. M.R.K. Mes	University of Twente, co-supervisor
prof.dr. R. Liu	University of Leeds
prof.dr.ir. C.M.J. Tampère	Katholieke Universiteit Leuven
prof.dr.ing. B. Rosic	University of Twente
prof.dr.ir. G.J. Heijenk	University of Twente

Preface

I would like to use these pages to acknowledge those who have contributed to this thesis and supported me during my time as a PhD researcher.

First of all, my supervisors, Eric and Martijn. Eric, you gave me the opportunity to start as a PhD researcher at the Transport Engineering and Management group, and continue afterwards, for which I am very grateful. I think we are typically on the same page, not only with respect to research but also share a sense of humor. Martijn, thank you for always taking the time to discuss progress and read my work: your feedback has significantly improved this thesis. I would like to thank you both for your support and guidance throughout the project.

Tom, you are probably the person that contributed the most to my research despite the fact that, officially, you were never part of the supervision team. It is difficult to express in a few lines how much I value your support and commitment. I really enjoyed discussing our research, oftentimes without a clear focus or article in mind, but just figuring out how something could or should work. Apart from talking about research, during the nice lunch walks we also discussed many other topics. I miss those walks and the frequent discussions and, all in all, it is a real pity that you left the university.

Since the start of my employment at the UT, many people have come and gone. Constant factor throughout the years was, and still is, Dorette. In the earlier years, we were basically the only ones drinking coffee, and used these morning breaks to talk about the weekend. Furthermore, I highly appreciate your help with all the administrative tasks.

Working at TEM has always been pleasant. Since my PhD trajectory took (a bit) longer than anticipated, I shared the office with many different colleagues. It is virtually impossible to list them all here, but thanks for the nice breaks, cycling trips and outings. In particular, I would like to mention Mariska and Kostas for the collaboration on some research papers and proposals. More recently, Zakir and Georgios have joined our team, and I cherish that every now and then we share a coffee (or Tsipouro) and a laugh. Apart from my (ex-)colleagues at TEM, I am thankful to those with whom I have collaborated on some of the research, most notably Georg, Francesco, Alessio and Xiaojie.

Obviously, this thesis would not exist without the ADAPTATION project. Therefore, I would like to acknowledge the partners of the project, EUR, Albert Heijn Online, DPD and Simacan, and NWO for the funding. Further, Gemeente Enschede and Saxion University of Applied Sciences are acknowledged for providing the data, and DAT.Mobility for their help with the raw data processing.

Although finding a proper balance between work and life has not always been easy for me, Twan, Timm, Stefan, Rigt, GJ, Veld 5, BankCiTters, Huize DrT, H&I, B2012 and The Enschede Group - I have probably also forgotten a few - have always provided me with the right amount of distraction. Many thanks for that.

Finally, the patient Eikenbroek and Hohmann families: Pap, Mam, Felix, Pari, Arian, Rene, Mike and, of course, Mani. I am grateful for having you in my life: thank you for everything.

Contents

Chapter 1

Introduction

In many urban areas, the road network is operating close to capacity. In such networks, small fluctuations in traffic flows can result in large disruptions in the level of service (LOS), e.g., travel speeds and travel times. With local disturbances rapidly influencing a larger part of the network, a full understanding of the spatio-temporal dynamics in the LOS is necessary to prevent or quickly identify disruptions, to predict its network-wide effects, and to support management decisions mitigating negative impacts in the short and long run. In this thesis, we investigate urban traffic variations on different scales and explore the impact of information regarding these variations for the decision-making processes of a selection of actors operating in the urban traffic domain. These actors use traffic information, e.g., for route planning and advice after incidents.

The remainder of this chapter is organized as follows. We introduce the thesis' research on including information regarding urban traffic variations in decision-making processes in Section 1.1. In Section 1.2, we discuss the relevance of information systems for decision making. Section 1.3 provides an overview of the variations in the LOS of the traffic system, and Section 1.4 discusses approaches from literature to model traffic variations on different scales. In Section 1.5, we examine the relevance of these models for decision-making processes of actors in the context of urban traffic. Section 1.6 identifies the research gap, which we consequently use to formulate a research objective and the main research questions. Further, we discuss the scientific and practical relevance of the research in Section 1.7.

1.1 Introduction

The relatively recent increase in the real-time traffic data availability in urban networks allows for constructing a picture of the prevailing traffic conditions. However, using the current situation as a predictor for future conditions during decision-making processes is naive since variations occur on different timescales, and current information may become obsolete in limited time. When making robust decisions, it is inevitable for decision makers operating in the urban traffic domain (e.g., public road authorities, route planners, dispatchers) to anticipate the future state of the system. That is, decisions should be based on forecasts (or: predictions) that not only account for periodic variations in demand and sup-

ply, but additionally incorporate the emergent behavior of travelers in reaction to decisions, and the effects of measures on traffic and society in general. In fact, the uncertain feedback of, e.g., travelers, operators and authorities largely determine future conditions, and predictions therefore need to account for the inter-dependencies between forecasts, management actions and responses.

Constructing such anticipatory forecasts in a timely manner is a difficult task, in particular in the context of urban traffic. In fact, the urban traffic system is highly complex and irregular, with a sophisticated physical structure, a vast variety of trip types and interacting modes, and a range of control measures in place. Moreover, erroneous data and the limited coverage of measurement devices makes that only a small share of the complete network is monitored on a continuous basis. Therefore, not only future conditions are uncertain but also information regarding the historical and current LOS extracted from the data is characterized by a high degree of uncertainty.

Decision makers operating in an urban traffic context employ information regarding the LOS, including its potential future development, using a traffic information system that is part of a larger management system additionally reporting, e.g., information on the status of the actor's operations. In this thesis, we distinguish three actors that potentially benefit from using a traffic information system during different stages of their decision-making processes: (i) logistics service providers (LSPs), (ii) urban traffic managers, and (iii) individual car users.

LSPs in home delivery are faced with a *vehicle routing problem* (VRP): the assignment of parcels to trucks, determining the order of the deliveries, the physical paths between the stops and the departure times (e.g., Laporte, 1992; Toth & Vigo, 2002). To be able to timely communicate an estimated time of arrival (ETA) to their customers, LSPs determine initial plans a long time (hours, days) in advance (e.g., Agatz et al., 2008) after evaluating a possibly large set of alternative route plans. Provided an assignment of parcels to a vehicle, determining the optimal route is known as a variant of the *traveling salesman problem* (TSP). Accounting for variations in driving times is necessary so that accurate ETAs can be communicated to customers while keeping operation costs low. Here, accurate travel time estimates over different timescales are necessary to reduce the uncertainty regarding driving times and to evaluate and construct route plans that possibly require adaptation while being en route.

Urban traffic managers from public road authorities monitor the historical and current network performance using a traffic information or network management center. Historical and real-time information regarding traffic light statuses, queues, delays and/or volumes are typically presented and visualized to support decision making. Further, on major arterials, collected data is used for automatic incident detection. Communication, management and policy measures and scenarios are partly or fully based on this information.

Individual car users employ *advanced traveler information systems* (ATIS), for example route advice in navigation devices, to support or automate a part of their decision making. Often, the origin and destination of the user is known and the information is presented in a descriptive or prescriptive way (Van Essen et al., 2016). Nowadays, a large share of these systems is also used for en-route updates and current and future estimates of the LOS can then be used to improve advice during different stages of the trip.

Independent of the actor, the information from the system can be characterized by two components: dynamism and uncertainty (Gmira et al., 2021; Pillac et al., 2013; Soeffker et al., 2022). The dynamic component reflects the capability of the information system

to incorporate the relevant spatio-temporal variations in the LOS, while uncertainty is due to the inherent capacity limitations of a system to capture all the variability. Indeed, the uncertainty stems from different sources including the unpredictable variability of the traffic system and the user responses from an aggregated perspective, model errors, and the limited access to data.

Decision makers benefit from accounting for both the dynamics in the LOS and the uncertainty in the information system regarding these dynamics. We briefly illustrate how information with respect to traffic variations can support the actors under consideration. Route plans of LSPs can account for the spatio-temporal dynamics of LOS, e.g., by avoiding congested areas during rush hours. At the same time, uncertainty in driving times is important, particularly when considering a sequence of stops where uncertainty on different parts of the traffic network accumulates or cancels out. Urban traffic managers infer an estimate of the performance of the network using a management system. Here, noisy fluctuations in measurements should be distinguished from abrupt changes to prevent the unnecessary deployment and counterproductive effects of measures. ATIS typically account for changing travel times over the course of a day, but could also consider communicating uncertainty in arrival times - or at least provide a robust ETA based on the uncertainty.

Summarizing, LSPs, urban traffic managers and individual car users are faced with decisions over different scales that are influenced by the variability in the urban traffic conditions. Although the actors use independent information systems, all with an individual and fragmented picture regarding the (future) conditions of the network, information regarding the relevant spatio-temporal variations in the LOS as well as the uncertainties therein support these actors in making well-informed and pro-active decisions.

1.2 Information systems, variations and decision making

Information systems extract information from the available data sources and provide an estimate of historical, current, and future developments (Soeffker et al., 2022). In a traffic context, such systems support decision makers by providing descriptive information on the historical and current LOS or predictive information on future events, potentially even for data-poor parts of the network, but can also be used in a prescriptive way when a part of the decision-making process is automated based on the information (Lepenioti et al., 2020; Wang et al., 2016). Often, there is a mismatch between the time of decision and information availability in the sense that not all necessary information is fully known at the time of decision making. Some information (e.g., travel times) is only revealed over time, e.g., during execution, making that the required input for decision or optimization problems is far from static and deterministic (see, e.g., Ferrucci & Bock, 2015; Pillac et al., 2013). Possibly irreversible decisions should be made *here-and-now* using only an a priori characterization of the system's evolution and the corresponding uncertainty. Robust solutions anticipate the dynamics of information (e.g., Pérez Rivera & Mes, 2017) and optimize over multiple possible realizations of the future, which might be infinitely many in contrast to the static and deterministic setting that only considers one scenario. Assuming a fixed horizon and a relatively stable environment, uncertainty decreases over time and continuous refinements of initial decisions are theoretically possible. In fact, newly arriving information can be used to react to changes in the system while anticipating future developments based on an updated characterization of the dynamics and the uncertainty (*recourse*). Such adaptations

could even be necessary when serious unexpected disruptions occur. Although practicalities induce that some decisions should be made a long time before execution, a priori decisions benefit from anticipating that solutions can be adjusted based on the realization of that what was initially considered uncertain.

Models are used to realistically capture the evolution of information but make a trade-off between tractability and simplicity so that they remain applicable in the limited time available (Koot et al., 2021; Shone et al., 2021). Measurements only give an indication of the current status of the system and directly using measurements is challenging in the context of *big data*, e.g., due to the high granularity and the noisy fluctuations. Information extraction is therefore highly related to pattern recognition where regularities are discovered (Bishop, 2006). Domain knowledge, generally expressed in a theoretical model, is used to relate measurements in that the desired hidden information is revealed (Hansen, 2010). A model is even necessary if historical decisions and control actions are inter-related by feedback effects, i.e., if only the realization corresponding to a previous control action is revealed in the data (Bertsimas & Kallus, 2020). Assuming that there are systematic variations and seemingly random fluctuations in the variable of interest as a function of an independent variable (in our context, typically time), past trends are used and extrapolated to forecast the future (Thomas et al., 2010). However, the trend is usually not explicitly observed in the past, and many other trends could have been extracted as well. In any case, even the most detailed model cannot incorporate all real-world mechanisms, and any information system therefore contains some degree of uncertainty (Gneiting et al., 2007; Shone et al., 2021). Even worse, with future conditions possibly being substantially different from the past, no model is able to provide reliable estimates under all circumstances.

With the future being inherently uncertain, the associated limited predictability should be addressed before adopting estimates in a decision-making framework (Lepenioti et al., 2020; Liu & Gupta, 2007; Walker et al., 2003). There have been many attempts to classify uncertainties (e.g., Walker et al., 2003), typically focusing on policy-related uncertainties. Here, we focus on shorter-term decisions on an operational and tactical level and follow the taxonomy of Makridakis et al. (2009), who stated that there are two types of (modeling) uncertainties that make that future outcomes cannot be predicted exactly: *subway uncertainty* and *coconut uncertainty*.

We define subway uncertainty as the accumulated uncertainty in a resultant variable due to the natural variability of the system from the perspective of a modeler under the hypothesis that the underlying environment remains stable (Makridakis et al., 2009). A well-known example of subway uncertainty is the variability accompanying a fair coin toss (assuming one does not model the physics of the toss): the exact outcome of the toss is unpredictable, and the inherent variability induces the uncertainty in the result. Hence, the best possible point prediction is frequently wrong, yet the associated probability distribution of the error can be derived accurately. In this sense, the system is uncertain yet perfectly predictable (Wright & Goodwin, 2009).

Coconut uncertainty refers to events that are unexpected and rare but have far-reaching consequences (Makridakis et al., 2009). One knows that this class of events occurs (with a significant probability), however the time and place of occurrence of such events is rather difficult if not impossible to forecast. As such, individual coconut events occur infrequently and are therefore only sparsely included in historical data, if at all (Goodwin & Wright, 2010).

According to Makridakis et al. (2009), subway uncertainty together with a share of

the coconut uncertainty define the known unknowns, while the other share of the coconut events for which the uncertainty cannot be accurately quantified comprises the unknown unknowns. With the probabilities of occurrences highly dispersed, coconut uncertainties are very difficult to incorporate in forecasts (Goodwin & Wright, 2010). Occurrence of events in the class unknown unknowns cannot be predicted at all - otherwise these events would not be in this class.

Decisions should be considered in the light of the future uncertainties, and one should therefore account for the limited predictability of the system under consideration. Apart from the inherent variability of the system, there is also epistemic uncertainty due to our imperfect knowledge of the system (Walker et al., 2003). Again, considering the fair coin toss example, epistemic uncertainty occurs if the predicted variability in the outcome is different from the 50% chance for the side showing head or tails. In modeling, epistemic uncertainty can be considered the prediction error (when considering forecasts) or the reconstruction error (when considering historical systematic variability). In theory, this uncertainty could be reduced by improving the model. However, it is not always obvious if one is dealing with inherent or epistemic uncertainty - in particular when studying a highly dynamic real-world system. As mentioned, any model introduces uncertainties due to design choices, e.g., assumptions and boundaries, which further complicates distinguishing the different uncertainties. In any case, it is virtually impossible to eliminate all uncertainties using improved models - and one should therefore accept and incorporate uncertainty as an integral part of decision making (Gneiting et al., 2007; Liu & Gupta, 2007; Makridakis et al., 2009; Walker et al., 2003).

Information systems including an estimate regarding the level of the uncertainty allow for robust decisions that anticipate subway uncertainties by accounting for the range of possible yet natural outcomes. Assuming that subway uncertainties can be accurately and continuously captured in a probabilistic framework, the input of the decision problems can be characterized as dynamic and stochastic in the sense that part of the input is revealed over time, but probabilistic information is available regarding the system's evolution (Pillac et al., 2013). Although theoretically this allows for informed decision making, formulating and solving such sequential decision problems is highly complex not only because of the difficulties capturing the dynamic character into a single model but also due to the wilderness of solution methods (Powell, 2014; Shone et al., 2021). Time and place of the occurrence of coconut events cannot be predicted, and therefore asks for reactive contingency plans (Goodwin & Wright, 2010; Makridakis et al., 2009). Decision makers can only respond to such events, and the challenge is to recognize them in a timely manner and dynamically adapt decisions to it. Even though uncertainty-aware decision making introduces additional challenges compared to the conventional static and deterministic setting (Pillac et al., 2013; Shone et al., 2021), it provides better-informed and robust decisions in an inherently uncertain environment.

1.3 Variations in traffic

A large share of the fluctuations in the LOS of the traffic system shows periodic behavior (Turochy & Smith, 2002) and is related to the time of day, e.g., the rush hours repeat themselves from day to day. These fluctuations are inevitably related to the variability in demand and supply. Yet, demand and supply variations do not directly cause a changing LOS, e.g.,

only when demand approaches or exceeds capacity, travel speeds drop and then driving times are likely to increase. We provide a brief overview of factors influencing demand and supply.

Variations in the demand occur on different scales. Longer-term temporal variations and trends in the demand exist from year to year (KiM, 2020; Rijkswaterstaat, 2021), but are also related to the seasons (Thomas et al., 2008), and different types of days (Viti, 2006). Intra-day variations are shown to exist when comparing 24h patterns for different weekdays (Thomas et al., 2008; Viti, 2006; Weijermars & Van Berkum, 2005), e.g., revealing the additional demand because of the extended opening hours of shops (Crawford, 2017). Shorter-term variations also occur due to events (Olabarrieta & Laña, 2020; Polson & Sokolov, 2017; Thomas & Van Berkum, 2009). Exogenous factors known to influence the demand include the weather (Maze et al., 2006) and pandemics (Van der Drift et al., 2022). Spatio-temporal variations in the demand, frequently expressed in terms of network usage, provide an aggregated picture of the variability in travel behavior across travelers (Crawford, 2017), e.g., in time of departure (Crawford, 2020), route choice (Zhu & Levinson, 2015) and mode choice (Heinen & Chatterjee, 2015).

Variability in the supply is caused, among other things, by geometry and the intra-personal variability in car-following behavior, e.g., in headways on major roads (Luttinen, 1996) and departure headways at intersections (Jin et al., 2009). On freeways, capacity is shown to be influenced by the weather (Maze et al., 2006). In an urban setting, the variation in traffic light signal phases is a well-known cause of variation in capacity possibly in interaction with the arrival dynamics (i.e., vehicle-actuated signals). Capacity is also influenced by accidents and incidents, currently causing more than 20% of all congestion on freeways (Rijkswaterstaat, 2021). Spatio-temporal variations in the LOS in an urban setting on a more local and short-term scale occur, e.g., due to the dynamics at intersections, on-street parking, and heterogeneous choice behavior of individual travelers (e.g., Carrion & Levinson, 2012).

1.4 Modeling traffic variations

The LOS of the traffic system shows variability in time and space. Variations are typically expressed on an aggregated and accumulated level in the measurements of variables that provide an indicator of the traffic conditions, e.g., using traffic volume time series. Apart from systematic variations that in theory are predictable, these measurements also show random fluctuations that seemingly show no pattern and cannot be predicted (Banks, 1999; Bates et al., 2001; Thomas et al., 2010). In this section, we discuss literature on modeling systematic and random variations in aggregated or accumulated variables partly reflecting the LOS on a part of the traffic system: volumes, speed, delays, and travel times. For variation in longer-term conditions and predictions, we refer to De Jong et al. (2007).

Fluctuations in traffic conditions-related variables occur on different timescales. In this thesis, we are mainly concerned with changes on the following two timescales. Changes are often measured in intervals in the order of 5-15 minutes, and assumed to reflect the variability in the ‘regime’ or actual performance or usage in the network (Breiman & Lawrence, 1973; Thomas et al., 2008). Very short-term variations occur in the order of seconds. Incorporating these variations is oftentimes not deemed desirable for decision-making processes, and some form of aggregation is applied to obtain ‘stable’ increments (Coogan et al., 2017;

Guo et al., 2007; Oh et al., 2005; Smith & Ulmer, 2003; Vlahogianni & Karlaftis, 2011). Even though very short-term fluctuations seem noisy, both timescales are of practical and theoretical interest (Breiman & Lawrence, 1973; Miller, 1970; Mirchandani & Head, 2001; Vlahogianni et al., 2004). First, variations on an aggregated and/or accumulated level - typically expressed using time series - are directly related to the very short-term variations. For example, the 10min volumes measured at a fixed location can be considered to be the result of a series of events of arrivals or departures with headway variations in the order of (tenths of) seconds. With aggregation applied in the context of decision making oftentimes without consideration of the impact of the resolution (Guo et al., 2007), very short-term variations introduce seemingly random variation when considering scales of several minutes. Indeed, the variability in a time series can correspond to the volatility of an underlying process (Brillinger, 2008). Even periodic variations on short timescales, e.g., due to the variation in supply at signalized intersections, introduce noise on a 5-15min level (Gerlough & Huber, 1976; Koen & Lombard, 1993; Thomas et al., 2010). An understanding of the underlying processes that cause random changes in aggregated time series helps one to separate noise from systematic differences in the conditions and to assess the quality of a model or prediction scheme a priori. Second, the LOS seems to be rapidly changing over time when considering measurements in the order of minutes (e.g., due to the capacity drop and the onset of congestion). In practice, obviously, many of these changes are not so abrupt and can be anticipated or quickly identified when considering shorter timescales (Son et al., 2014; Vlahogianni & Karlaftis, 2011). In particular these events are highly relevant to predict since control actions can then be deployed to mitigate network-wide impact, yet require short-timescale estimates in a very noisy environment.

Systematic variability in the order of minutes can be used to separate different regimes over space and time, which then, in turn, can be used to study changes over longer timescales. Traffic volumes, typically examined in both research and practice using 24h time series, are well known to show systematic volume variations during a day and between days. Time of day is an important predictor for network usage with 24h time series collected at a single point showing a regular M-shaped curve (Crawford et al., 2017; Laña et al., 2019; Weijermars & Van Berkum, 2005). Considering the day-to-day variability, traffic volumes deviate systematically between days (Ma et al., 2021; Stathopoulos & Karlaftis, 2001), and show considerable seasonable variation (Coogan et al., 2017; Thomas et al., 2008). However, as Crawford (2017) points out, there has been less attention for systematic differences in both the shape and the height of 24h traffic volumes on a day-to-day basis. This simultaneous consideration is necessary to accurately consider the impact of short-term systematic variations such as events on a single day and to assess the robustness of time-(in)dependent services and policies such as road pricing (Crawford et al., 2017; Thomas & Van Berkum, 2009). Further, the changing height, width and time of the peak over time are important for estimating the remaining road capacity, and for longer-term monitoring purposes and policy evaluation (Guardiola et al., 2014; Weijermars, 2007).

Since the systematic variation in volumes, and for other measurement series as well, is typically not known in advance nor revealed over time, it needs to be inferred from historical data in a supervised or unsupervised manner. Using a supervised approach (e.g., Crawford et al., 2017), one investigates the influence of explanatory variables next to time as a predictor for systematic differences. Unsupervised approaches (e.g., Guardiola et al., 2014; Muralidharan et al., 2016) aim to infer patterns from the data, which are consequently - if possible - interpreted using auxiliary variables. Hence, from such an approach we cannot

guarantee to find the source of the variation - and therefore it is not easily incorporated in a forecasting scheme. On the other hand, many variations cannot be predicted using typically available exogenous variables only, but can be forecasted in a statistical sense (Ma et al., 2021; Thomas et al., 2010; Wagner-Muns et al., 2018).

Clustering approaches (see Li et al., 2015) are probably the most popular way to recognize and identify groups of 24h time series, typically extracting a group-specific 24h pattern that represents the within-day variability assuming less variation within the group than between groups (Caceres et al., 2012; Chung, 2003; Crawford et al., 2017; Zhong et al., 2020). For example, the approach by Weijermars and Van Berkum (2005) found three regular intra-day working-day patterns when considering 24h traffic volumes collected at a freeway, distinguishing Mondays, core weekdays, and Fridays. Crawford et al. (2017) explicitly accounted for gradual changes in the time series using normalization, and identified day of the week-dependent shapes in 24h flow series on an urban road in the Greater Manchester area. Guardiola et al. (2014) used unsupervised clustering to identify underlying principal components, which turn out to be easily interpreted. For example, here, the first component distinguishes working days and holidays. Wavelet analysis considers one long time series and looks for regularities with various time lags, e.g., Jiang and Adeli (2004) found a daily pattern with two peak periods every day and a periodicity over a period of one week. Longer-term periodic variations are related to weeks and seasons (Jiang & Adeli, 2004), with the winter profile including a smaller proportion of flows in the morning peak, and a larger proportion in the middle of the day (Crawford et al., 2017). Other techniques, e.g., neural networks (Polson & Sokolov, 2017), seasonal ARIMA (Guo et al., 2014) and nearest neighbor (Habtemichael & Cetin, 2016), also infer systematic patterns from the data, but they can be considered black-box approaches in the sense that the inferred systematic variations cannot easily be extracted or are considered in the context of short-term traffic forecasting (Lv et al., 2015; Vlahogianni et al., 2014).

Compared to longer-term variations during regular conditions, short-term systematic differences in time series on timescales shorter than 24h but longer than 5-15min have been studied only to a limited extent. From a modeling perspective, these variations occur in the residual time series, i.e., the deviations compared to the underlying yet unknown 24h pattern which would have been realized without the short-term event (Chen et al., 2012; Li et al., 2015). These variations express systematic differences, e.g., due to events and weather conditions, and are highly relevant to predict since they express non-typical situations. Nonetheless, the time of occurrence of such events is highly variable making that standard 24h clustering methods do not appropriately account for such deviations. Thomas and Van Berkum (2009) show that in case of incidents and events, sinusoidal fluctuations occur around the intra-day pattern. Polson and Sokolov (2017) predict traffic flows related to football matches and snowstorms using a neural network implicitly learning the patterns. Olabarrieta and Laña (2020) show that volume patterns occur in time and space before and after a football match. These studies indicate that events influence volumes in a systematic manner, and in theory can be predicted provided that the time of occurrence and location of the underlying event is known. Periodicities can also be considered in a more abstract form. The occurrence of outliers - independent of the underlying reason - were shown to follow approximately a negative exponential distribution, and can therefore considered to be unpredictable (Chen et al., 2012). Yet, regular fluctuations around the intra-day pattern were shown to have periodicities on a 30min timescale (Thomas et al., 2008).

Regarding the variability in travel times, systematic variations are shown to be related

to the regimes in traffic. Van Lint and Van Zuylen (2005) identified four different regimes with different travel time distributions: free-flow conditions, congestion onset, congestion and congestion dissolve. Typically, a few explanatory variables are used to cover the variations. For example, Li and Rose (2011) consider different independent variables influencing travel times, including time of day, day of week and rainfall. At the same time, recent travel times were shown to influence current and near-future travel times as well (Kwon et al., 2000). Zhong et al. (2020) considered the day-to-day and within-day travel time variability, indicating that under rare events travel times have a longer tail compared to the travel time distribution under recurrent conditions. When considering speeds, variations in speeds over time were shown to follow a quasi-sinusoidal pattern with a daily period (Kamarianakis et al., 2005). Extending the spatial dimension, only a few spatio-temporal patterns were shown to exist on major arterials (Lopez et al., 2017). A difficulty compared to measurements resulting from roadside devices is the additional uncertainty due to the intra-personal variability in travel times.

If the modeler, according to his opinion, has accounted for all systematic variations, or, at least, that the remaining patterns occur on timescales that are shorter than the aggregation level, uncertainty is an inherent part of the variation. Throughout this thesis, following Thomas et al. (2008), we refer to these subway uncertainties that are uncorrelated over successive time intervals as random variation or noise, although what is considered to be uncorrelated for a fixed location might show patterns over other spatial and temporal scales (Ermagun et al., 2017; Koen & Lombard, 1993). Not only the natural variability in the system contributes to this variation, but also temporary measurement errors induce random variations. Quantification of the random variation should be an integral part of the modeling exercise, since it indicates the extent to which a measurement can be 'trusted' (Kaas et al., 2008), and provides an a priori quality assessment of a prediction scheme, including a lower bound on the predictability of the variable under consideration. However, it is currently often only implicitly considered as a part of interpolation exercises possibly including a regularization parameter (Crawford et al., 2017), outlier filtering (Chen et al., 2012), and state-space filtering (e.g., Kalman filter - Haykin, 2004; Kalman, 1960).

Using a time series approach, random variation is also referred to as dispersion or volatility, i.e., the variability over time, for which the time-dependent variance or standard deviation, possibly in relation to the mean value, is a standard measure. In signal processing, this is expressed in the *signal-to-noise ratio* (Kay, 1993). An a priori quantification of the random variation is difficult in the context of urban traffic, and is partly a result from the modeling exercise and the modeler's trade-off between fidelity and simplicity (Shone et al., 2021), and therefore includes a share of uncertainty that is only treated as such for the benefit of the model. Vice versa, without careful consideration of the nature and level of uncertainty one could infer patterns from true noise (*overfitting*), providing a potential reason that simple prediction schemes still provide accurate forecasts (Makridakis et al., 2020). Moreover there is evidence (Chen et al., 2012; Guo & Williams, 2012; Thomas et al., 2008) that the random variation depends on the underlying systematic pattern or conditions and exogenous variables (Li, Chai, et al., 2022) and, hence, random variations should be studied in relation the underlying processes. In the remainder of this section, we provide an overview of capturing and modeling random variation in the context of traffic, typically using a probabilistic framework in contrast to the often-used deterministic setting to capture systematic variations.

Basically, there are two possibilities to study the probability distribution of the random

variation. First, as a function of time and previous squared residuals (compared to a pattern capturing the systematic variability) and variances. Second, as a function of the underlying systematic pattern or exogenous variables. The first approach is used in generalized autoregressive conditionally heteroscedastic (GARCH) models to capture the variance of the residual time series, and to predict the volatility. GARCH has been used by Kamarianakis et al. (2005) to study the volatility in speed, and, among others, by Guo et al. (2015), Huang et al. (2018) and Vlahogianni and Karlaftis (2011) to study the natural variability of traffic flows on arterials and freeways. Stochastic volatility models, where the variance follows a stochastic process, are used, e.g., by Zhang et al. (2014). The second approach, where the distribution is conditioned on the underlying trend or regime, allows for a more intuitive explanation of the random variation. A well-known example is the (discrete) Poisson distribution, for which the variance of the random variable equals the mean. Adopting the latter property for continuously-varying data, Thomas et al. (2008, 2010) showed that this dispersion coefficient is a lower bound on the true dispersion for volume data. Although additional challenges occur when adopting a count time-series approach (Manolakis & Bosowski, 2019), variants of queuing theory can be used to explain the underlying dynamics (e.g., arrival processes) leading to this volatility (Breiman & Lawrence, 1973). Transformations on the data can be used to detect the volatility (Guo et al., 2015; Guo & Williams, 2012).

Regarding travel times, Van Lint and Van Zuylen (2005) observe different types of travel time distributions under different regimes. Where free-flow travel times show little spread around the mean, other regimes show a highly skewed distribution. It is argued that using variance or standard deviation in relation to the mean value is naive when modeling the reliability of travel times (Ramezani & Geroliminis, 2012; Van Lint et al., 2008). Li and Rose (2011) indicated that the vehicle-to-vehicle travel time variability is a sigmoid-like function of the mean travel time meaning that the random variation can be well-modeled as a function of the underlying patterns.

The random variation in traffic conditions-related variables makes that a volatility model can be used for outlier detection, point prediction error estimation and for probabilistic predictions - and thereby for decision assessment. Several standard (point forecasts) error metrics are shown to be influenced by volatility (Kaas et al., 2008; Karlin & Taylor, 2012), but this is not well-acknowledged in prediction studies. Where prediction intervals have been introduced in different contexts (Khosravi et al., 2011; Lin et al., 2018), these are often calibrated based on a single confidence level. However, singe-level optimized intervals may show undesirable performance for other confidence levels. Although various metrics take the width of the prediction interval into account (Khosravi et al., 2011; Makridakis et al., 2020), these measures are more appropriate if a full density forecast is provided and evaluated as a whole (Hong et al., 2016).

The random variation on longer timescales partly results from the very short-term fluctuations in the dynamics. These short-term variations are difficult to study since (i) fluctuations strongly depend on the regime (Breiman & Lawrence, 1973), and (ii) the underlying real-life processes are highly complex particularly in an urban network. The underlying processes are often modeled assuming a relatively slowly-changing or constant regime (stationary conditions), which rarely occurs - if at all - throughout a single day, e.g., volume time series show a highly nonlinear and quickly changing intra-day pattern. In a stationary setting, renewal theory allows one, under certain assumptions, to express the variation in aggregated volumes as the result of the variation in inter-arrival times. In particular for freeways, this approach relates the variations across different timescales (Breiman & Lawrence,

1973; Miller, 1970) but for interrupted flows, e.g., in urban traffic networks, such relations are more difficult to capture directly. Indeed, in the lower urban network a large share of the variations in the LOS is due to the induced fluctuations by interruptions, e.g., signalized intersections and junctions, and traffic flows might show *bursts*, i.e., shorter periods with numerous arrivals alternate with longer periods with no arrivals (Goh & Barabási, 2008; Paxson & Floyd, 1995). Therefore, queuing theory is used to explain the very short-term variations in volumes and delays, the latter as a major travel time-influencing factor. In any case, mirroring real-world traffic dynamics in this setting is very challenging compared to freeways. For example, the arrival events on freeways were shown to approximately follow a renewal process under low-to-moderate flows (Breiman et al., 1977; Luttinen, 1996), and can thus be modeled using solely an inter-arrival or headway distribution (Ha et al., 2012; Hoogendoorn, 2005). In an urban context, under all conditions, one should study events in conjunction with other processes mainly at intersections, e.g., the inter-relations between arrival and departure processes and signal dynamics (Luttinen, 1996; Zheng et al., 2017). Explicitly modeling these interactions is not an easy task: arrivals at an approach are influenced by interruptions upstream and signals are, at least in our Dutch context, typically actuated by vehicle arrivals. Literature focused on the impact of these short-term variations on the mean and variance in the delay (for a single regime). As such, several encompassing formulas approximating the relation between the degree of saturation and delays have been proposed. The most prominent study is the one of Webster (1958). Over time, models have been extended to account for various delay-contributing processes such as different arrival processes (Boon & Van Leeuwaarden, 2018; McNeil, 1968; Viti, 2006), time-dependent processes (Akcelik, 1980), networks of intersections (Boon & Van Leeuwaarden, 2018) and vehicle-actuated signals (Viti & Van Zuylen, 2010b). Here, both the mean as well as the variance in delays increase sharply when the degree of saturation increases.

The presence of systematic variability in the traffic conditions-related variables makes that information systems should explicitly account for the dynamics in the LOS. At the same time, point predictions are not sufficient when evaluating measures in a decision-making process since random variation causes that a realization can substantially deviate from the average. Hence, actors operating in the urban traffic domain should consider both systematic and random traffic variations when considering robust measures, thereby anticipating both the evolution as well as the uncertainty of the LOS.

1.5 Traffic variations and decision-making processes

We distinguish three main actors using information regarding the urban traffic network in the decision making process: LSPs, traffic managers and individual car users. In this section, we illustrate the relevance of using information regarding traffic variations during decision making processes and examine how traffic variations are addressed in the accompanying decision problems.

1.5.1 Traffic variations and decision making for LSPs

LSPs in home delivery construct route plans to visit a set of customers thereby minimizing the costs of transport while satisfying service requirements. Route plans are constructed days or hours before execution (e.g., Agatz et al., 2008), and the corresponding decision

problem involves the assignment of goods to trucks, the sequence of customer visits, the physical paths between the stops as well as the departure times of the trucks. Just before execution, or while being en route, such plans might be adapted since the information on which the initial route plans was based potentially becomes obsolete. In fact, dynamically revealed information makes that adapting route plans in a near real-time fashion is desired when ETAs are exceeded, or to assure that route plans are less vulnerable for future disturbances.

Although route plans are constructed in advance, offline route plans can relatively easily account for variations in the LOS of the traffic system that occur on longer timescales. For example, recurrent congested areas can be avoided by changing the sequence of customer visits or by adapting the departure time as long as the service guarantees are met, i.e., longer-term periodic variations can be accounted for during the offline decision-making processes. However, a large share of the variability in traffic conditions cannot be predicted a long time in advance, making that route planning is characterized by a high degree of unpredictability. Not only the inherent uncertainties of the traffic system are part of this unpredictability, but also systematic variations that occur on relatively short timescales are difficult to predict. Robust offline route plans anticipate these uncertainties and target stable arrival times under a variety of conditions. Here, it is not sufficient to consider local variations in isolation. In fact, variations should be considered on a path rather than a link level since customers and dispatchers are not so much concerned with variations in traffic per se, but are merely focused on the timeliness of the provider (Heim & Sinha, 2001).

Assuming that the planning horizon remains stable over time, a priori information regarding travel times and information over time about realized network delays make that plans can react to changing conditions while anticipating future developments. In fact, prediction errors for recurrent variations decrease over time resulting in shorter-term travel time predictions that can be used to dynamically update routes. However, shorter-term predictions still include subway uncertainties and prediction errors. The online adaptation might even be necessary when coconut events take place but should be avoided when very short-term but natural variations occur that only seem to but not truly impact operations. Making robust offline route plans stable under both subway and coconut uncertainties is virtually impossible. Therefore, anticipatory route plans regarding traffic variations are characterized by relatively stable arrival times under prediction and subway uncertainties and allow for dynamic adaptation when disruptions such as incidents occur.

The well-known routing problems, i.e., shortest-path problem, TSP, and VRP in their basic form assume constant travel times between customers. The periodic variations regarding travel times are included in time-dependent variants (Gendreau et al., 2015; Kok et al., 2012), with speeds or travel times modeled as function of time (of day). Random variations in the LOS of traffic are incorporated using a stochastic version of the optimization problems - implicitly assuming that the uncertainties can be captured by a random variable - thereby theoretically accounting for both the inherent as well as the prediction uncertainties. Periodic variations and the accompanying uncertainty covering timescales longer than the planning horizon can then be included in the offline setting. Possibilities for dynamic refinements are highly determined by the a priori solution. In home delivery, when the trucks depart, parcels are already assigned to vehicles and there is only limited calculation time available to determine improved plans since the LOS might be quickly changing. At the same time, anticipating future adaptations can significantly affect the offline route plans (Powell et al., 1995).

Pillac et al. (2013) provide a taxonomy of the relation between information evolution and the information quality, with all necessary information either known or not known beforehand and with or without uncertainty. The variations in travel times make that the uncertainty-aware decision problem is of dynamic and stochastic nature. In current practice, human planners are often dedicated with re-planning after the occurrence of an incident or if a communicated ETA is exceeded. In theory, a part of re-planning could be automated since the fleet is monitored in real time.

The complexity of the dynamic and stochastic variant of the shortest-path problem, TSP and VRP is a potential reason that anticipatory and adaptive (re-)planning methods have not been widely implemented. First, the time-dependent shortest path problem needs already substantially more calculation time compared to the time-independent variant (Gendreau et al., 2015), which is particularly concerning during the time-critical re-planning process. Second, these problems require a transport management center not only estimating current travel times but also providing travel time forecasts on a continuously changing collection of paths over many different timescales. Here, incidents and accidents should be recognized in a timely manner and forecasts should also account for future consequences (including traffic management) of an event that not necessarily occurred at the intended path. These problems become even more challenging with (time-dependent) uncertainties involved.

The size of the fleet, the uniqueness (in time and space) of incidents, the absence of reliable traffic predictions, and the time-critical nature of (re-)planning makes that human planners decide on sub-optimal route plans. Therefore, there have been several attempts in literature to address these issues, translating the route planning process into a stochastic and time-dependent TSP or VRP. We note that a majority of the VRP-literature, however, considers uncertainties in the demand rather than travel conditions (see, e.g., Soeffker et al., 2022). Static problems with time-dependent stochastic travel times or speeds include the expected shortest path problem (Miller-Hooks & Mahmassani, 2000) and the stochastic time-dependent VRP (Lecluyse et al., 2009; Taş et al., 2014). Re-planning after, e.g., incidents, is studied by Fleischmann et al. (2004) where the shortest path to the next destination is recalculated based on current conditions. Studies that allow the adaptation of route plans based on changing conditions include Ehmke et al. (2015), Ferrucci and Bock (2014) and Köster et al. (2018) - typically making naive assumptions on the propagation of congestion. Gmira et al. (2021) consider a dynamic VRP incorporating dynamic perturbations to the travel speeds after which the current solution is reconsidered. These studies underline the practical and theoretical difficulties to realistically capture and anticipate the spatio-temporal dynamics and the uncertainties in the LOS in the context of routing.

1.5.2 Traffic variations and urban traffic management

Public road authorities and urban traffic managers use historical and near real-time information to design and improve the urban traffic network. Decision making occurs on different levels but is in an urban context often related to intersection-related choices. Strategic and long-term decisions include the physical design of the network, i.e., construction plans regarding a part of the network. On a lower level, junction design decisions are made, involving the choice for the type of junction (e.g., roundabout or signalized intersection) as well as longer-term decisions regarding the geometric design, slow-traffic handling, etc. (Bezembinder, 2021). Medium-term decisions for intersections involve the traffic signal cycle design. Short-term decisions are management and control measures and include, e.g.,

cycle adaptations and communication measures in case of events and road works as well as local management decisions to prevent blockages or the onset of congestion. We refer in the remainder of this thesis to the medium and short-term decisions as urban traffic management, comprising local or network-wide measures on different timescales to improve the utilization of the network (Taale et al., 2018).

The decisions cover various timescales, using diverse types of traffic information. For example, network and junction design are decisions for the long term with infrequent redesigns (Bezembinder, 2021). On the other hand, cycle adaptation occurs more frequently and is potentially deployed on a corridor level with a very short timescale to allow a 'green wave' (Coogan et al., 2017; Mirchandani & Head, 2001). Where longer-term choices are mostly based on estimated trends in yearly or monthly demand, management choices require short-term estimates of the LOS on a local, link or junction-level, scale (Mirchandani & Head, 2001). Decisions on longer temporal scales involve expert judgment with the additional use of design manuals, historical data or rough projections of future developments. As an example, junction and traffic signal cycle design manuals (CROW, 2006; Transportation Research Board, 2010) prescribe estimates of the normative volumes as a major determinant for design choices. Also, rough forecasts in combination with expert judgment and manuals are used for management choices in case of, e.g., events (CROW, 2008). Short-term management decisions are captured in management scenarios, using near real-time data to trigger these scenarios, for instance in case of automated incident detection.

Traffic variations play a pivotal role here. Junction and traffic signal cycle design are based on high-volume occasions, so that the projected day-to-day rush hours can typically be handled. Hence, the trends in the 24h volume time series over days are important to estimate - including structural changes in the height and width of the peak of conflicting directions. For shorter-term decisions, the periodicities of the events influence the extent to which data is used to support decision making. Traffic variations due to non-recurrent events are difficult to predict and short-term management decisions are then reactive and involve the use of near real-time data or recent measurements. This is partly due to the fact that a majority of incidents and accidents cannot be predicted (coconut uncertainty). Urban traffic information centers, at least in the Netherlands, support traffic managers and provide information extracted from the data collected throughout the network, which is oftentimes limited to information on the statuses of loops and traffic light signals, volumes, estimated queues and delays near signalized intersections aggregated on a 1 to 15min scale.

Uncertainty should be an integral part of urban network traffic management. Current management scenarios can be considered to respond to coconut events by comparing measurements to threshold values, so that a minimum LOS can be guaranteed under different conditions. Subway uncertainties, on the other hand, have not been well-integrated in urban traffic management. Yet, it has been suggested to base management measures not only to the systematic variations but also to the accompanying volatility (Tsekeris & Stathopoulos, 2006).

Current data sources, including Bluetooth sensors, floating car data, loop detectors and video cameras, only provide an indication of a portion of the spatio-temporal dynamics that occur in an urban network. An underlying model can also be used translate the available data to a full yet interpolated picture of the traffic state throughout the network (Mahmassani, 2001; Weijermars, 2007). Nonetheless, data is noisy - making that actual measurements not necessarily reflect the actual conditions throughout the network. Moreover, the highly dynamic nature of urban traffic in combination with aggregated and lagged mea-

surements make it difficult to distinguish natural variations from systematic changes that require management intervention. These challenges become even more pressing for parts of the network where no data is collected and therefore, instead, a model is employed. As a matter of fact, it is very challenging to capture all urban dynamics in a single model - let alone make well-informed decisions based on such a model.

Apart from error analysis, subway uncertainties have only been partly covered in decision making. It is recognized in intersection design manuals (CROW, 2006) that traffic shows random behavior and that therefore delays differ from person to person, and that the performance of an intersection changes over time. Although almost all major cities adopted urban traffic management systems in various forms (see, e.g., Hamilton et al., 2013; Nellore & Hancke, 2016), these systems rarely take the accompanying uncertainties explicitly into account (Tettamanti et al., 2011) although, for example, a quantification of the (prediction) uncertainty helps managers to value forecasts (Laña et al., 2019). In the relative comfort of model predictive control, where a model is assumed to mimic the real-world environment, it was shown that taking uncertainties into consideration improves control decisions (Hu & Hellendoorn, 2013; Tettamanti et al., 2011). Current traffic management practice can be considered to be highly reactive, with infrequently re-evaluated management scenarios that are triggered based on (near) real-time data. For example, short(er)-term predictions play a less important role in current practice, but are necessary for anticipatory decision making (Coogan et al., 2017; Li, Yang, et al., 2022; Vlahogianni et al., 2004). In contrast to local anticipatory decisions, network-wide anticipatory decision making should not only focus on local problems but, ideally, also anticipate feedback effects that occur on a larger scale: the emergent behavior of travelers in response to decisions should then be accounted for.

1.5.3 Traffic variations and advanced traveler information systems

ATIS support travelers with their pre-trip and en-route travel decisions (Adler & Blue, 2002) and travel-related information is presented to users in either a descriptive or prescriptive way (Van Essen et al., 2016). Nowadays, car users often use a navigation device supporting them in path choice decisions and to obtain en estimate of their travel or arrival time. The navigation tool is used days to minutes or seconds before departure, and during different parts of the journey. Such estimates mostly contain a (robust) point prediction regarding the travel time or the time of arrival.

ATIS, particularly navigation devices, use predictions regarding the LOS of traffic on different timescales. When suggesting routes days or hours in advance, only longer-term periodic variations can be accounted for - making that the best-possible route can change based on the more actual conditions. Hence, suggested paths are potentially time-dependent. While being en-route, real-time data and short-term predictions are used by the navigation device to propose or prescribe improved routes.

The uncertainty in predictions is usually not communicated to users, although the natural variability in travel times and speeds make that arrival times are inherently variable. The quality of the point prediction is however of high importance, since users oftentimes have a desired or required arrival time. Yet, the subway uncertainties in the traffic network may be accounted for in the point prediction by providing conservative estimates that are typically not exceeded, e.g., when natural variations due to possibly random delays at intersections add up. If in addition to the mean travel time, also predictions intervals would be constructed and presented, users can select their preferred option based on their attitude towards risk

(Tsekeris & Stathopoulos, 2006). For example, Shiftan et al. (2011) showed that experience-seeking users tend to prefer routes characterized with a lower average but larger variance in travel time. However, predictions should be accurate in the sense that an increase in the prediction error causes a decrease in compliance to advice (Ben-Elia et al., 2013).

Where inherent uncertainties and longer-term systematic variations can be (implicitly) included when advising routes, coconut events are generally not anticipated in the forecasts - otherwise point predictions would be inaccurate in the 'business-as-usual' case. This makes that en-route adaptations are potentially necessary to minimize the deviation compared to the initial ETA. Currently, such adaptations are based on near real-time estimates of the LOS over the network. In that sense, these suggestions are only partly anticipatory since these adaptations do not account for feedback effects that particularly occur when penetration rates of information services are high. Indeed, the impact of guidance information on the (future) traffic conditions is rarely accounted for (Ben-Akiva et al., 1991). In addition, whereas route adaptations are often predominantly based on the data collected via users of the navigation system, suggestions are usually based on a limited picture of the LOS and, for example, fail to account for supply variations.

Since ATIS are used during different stages of the journey, travel time estimates are required on various timescales on many different paths. Employing solely the systematic variations in travel times or speeds on links, a time-dependent variant of the shortest path algorithm (Dijkstra, 1959) could be used to calculate such an estimate. Stochastic information on the uncertain LOS can be included in the expected shortest path (Miller-Hooks & Mahmassani, 2000), and - if we look for a policy rather than a path - in a shortest path that is dynamically adapted based on revealed traffic conditions (Levering et al., 2022). In practice, such paths are only recalculated after unexpected disturbances occur and do not anticipate feedback effects.

1.6 Research gap

The urban traffic system is a complex system with its dynamics typically monitored and expressed on an accumulated and aggregated level using spatio-temporal patterns in the measurements. Recurrent patterns express a part of the systematic differences that in theory could be predicted using a model. A share of the occurring fluctuations, however, show seemingly no pattern, are uncorrelated and are considered random and thus unpredictable. Random variation together with unexpected events such as incidents and accidents, and the limited availability of different types of traffic conditions-related measurements on various scales, cause that decision makers in the urban traffic domain are uncertain about the evolution of the LOS. Informed decisions accept the uncertainties and are therefore uncertainty aware. To support anticipatory decision making and to identify the limitations thereof in the context of urban traffic, the inter-relations between systematic variation and uncertainty on different spatio-temporal levels should be understood and quantified.

LSPs, urban traffic managers and individual road users desire during their decision-making processes different types of information regarding the historical and future developments of the traffic network (see Section 1.5). LSPs use travel time predictions on different timescales to make robust route plans that can be dynamically adapted over time. Urban traffic managers employ patterns from historical data for the design of traffic signal control systems and use near real-time data to trigger management scenarios. Individual road users

employ information services to support travel decisions, e.g., estimates regarding the arrival time. In all these decision-making processes, an estimate regarding the spatio-temporal evolution of the LOS is incorporated. The accompanying uncertainty plays typically a less prominent role but might be equally important in the context of robust decision making.

In this thesis, we use historical data to get a grip on the systematic and random variations that occur on different spatio-temporal levels in the urban traffic network and thereby we facilitate a shift from reactive to anticipatory and uncertainty-aware decision making. Since the operations in an urban network are for a large share determined by the dynamics near signalized intersections, we particularly focus on the variations there.

When considering the interrupted flows at signalized intersections, delays are mainly determined by the arrival rate relative to the maximum discharge capacity during the effective green lag (*degree of saturation*). Although the dynamics here are complex, in particular when inter-relations occur between the arrivals at different arms and the capacity in the case of (semi-)actuated traffic control, the academic community (e.g., Akcelik, 1980; Webster, 1958) proposed several encompassing formulas approximating the relation between the degree of saturation and delays. Notwithstanding the accompanying simplification of the actual dynamics, these functions show, under minor assumptions, that there is very little variation in the average delay compared to the volume variability when the degree of saturation is low. With delays showing little variation under low to medium arrival rates relative to capacity, the underlying condition of the network is difficult to infer from delays only. Volumes then particularly support predictions regarding delays, and thus travel times, and thereby allow preemptive measures to be taken. In addition, delays are almost impossible to measure directly and traffic speed measurements are collected by external companies, which can make it expensive to use these data on a continuous basis. In our Dutch context, volume data are collected by induction loop detectors at the arms of intersection and typically easier accessible.

Although estimates regarding future travel times are of interest for road users, traffic volumes throughout the network are an important source for explaining and predicting the variability in driving times. For urban networks, alternative approaches that directly express or forecast the variation in network-wide travel times are less appealing. Forecasting driving times under various conditions based on historical measurements is challenging since many underlying variability-inducing factors are changing over time. For example, not only the network conditions but also the routes may change over time (see Simroth & Zähle, 2010). Hence, in this thesis we mainly focus on the variations in urban traffic volumes near signalized intersections.

Volume variations are mainly studied on two temporal levels. Variations in the conditions are measured on an aggregated level, typically in 5-15min increments. Very short-term fluctuations occur in the order of seconds (Breiman & Lawrence, 1973), and introduce random variation on an aggregated scale. Where systematic and random variations are inter-related and typically difficult to disentangle, current literature mainly study them in isolation. A simultaneous focus, however, is necessary to recognize systematic differences in noisy measurements and to assess decisions in the light of the evolution of the traffic system including the accompanying uncertainty. Reviewing the literature on traffic volume variations, we make the following observations.

The longer-term variability in traffic conditions can be examined by considering the changes in the 24h traffic volume time series. Although a variety of methods have been developed to identify patterns in aggregated and accumulated traffic measurements for re-

construction and prediction purposes, these methods are generally designed for short-term traffic predictions (e.g., Habtemichael & Cetin, 2016; Lv et al., 2015). Many decisions, however, cover a longer timescale and an assessment of the systematic differences in the 24h volume time series support such decisions. Only a few studies (Crawford et al., 2017; Guardiola et al., 2014), however, consider these time series over longer periods, and take both the (slowly) changing shape and the height of the 24h pattern into account when assessing systematic variations. In addition, there has been limited attention for the systematic variability relative to the intra-day pattern, e.g., due to events, but particularly these differences contain valuable information to decision makers. In any case, extracting systematic patterns from the time series is a difficult task, since noise with unknown characteristics corrupt the measurements and the patterns are not observed directly.

Random variations are implicitly addressed by researchers when identifying outliers or developing and applying a state-space model. Volatility models from econometrics (e.g., GARCH) do allow expressing the time-dependent higher order moments in time series yet provide very little intuitive understanding regarding the relation of the volatility with the underlying processes. In addition, when using such time series methods designed for measurements collected at major arterials or freeways, the variability related to the spatio-temporal variation in the demand and supply at the intersections throughout the network is not incorporated. A comprehensive understanding of the relations between arrivals, departure dynamics and signal phases on the one hand and variations in aggregated measurements on the other hand is still missing but is required, e.g., to understand and quantify the information loss due to aggregation (Breiman & Lawrence, 1973; Paxson & Floyd, 1995; Son et al., 2014; Vlahogianni & Karlaftis, 2011). A majority of the studies on local and very short-term dynamics consider a theoretical setting (e.g., Boon & Van Leeuwaarden, 2018; Van Leeuwaarden, 2006; Viti & Van Zuylen, 2010a) rarely based on extensive real-world data, or consider only one factor such as the headway distribution (Hoogendoorn, 2005; Jin et al., 2009; Luttinen, 1996). Yet, the underlying dynamics can only be revealed when simultaneously considering different timescales with the use of empirical data, for example with supply-based information on traffic signal cycles and induction loop data regarding the occupancy.

Anticipatory decision making requires accurate predictions over various timescales, explicitly addressing the uncertainty in such predictions relative to the system's unpredictability including the random variation. Although more and more studies provide probabilistic forecasts (e.g., Guo et al., 2014; Huang et al., 2018; Khosravi et al., 2011; Li & Rose, 2011; Shi et al., 2014), they mostly generate and evaluate a prediction interval for a single or a few confidence levels. The quantification of the individual sources of uncertainty accompanying a prediction supports one to identify potential ways to improve it. In addition, many of the volume prediction methods focus on freeways or major arterials, oftentimes limiting themselves to short-term point predictions for recurrent conditions. With dynamics being substantially different in an urban context, there is a demand for uncertainty-quantifying forecasts, i.e., predictive densities, in this setting for longer prediction horizons under a variety of conditions including events.

Compared to the rerouting in the context of LSPs, anticipatory traffic management and rerouting under high penetration levels suffer from feedback effects in the sense that current decisions influence traffic conditions, making that initial forecasts potentially become obsolete. In fact, intended outcomes might not be achieved when failing to account for the behavior of traffic users in response to management decisions (Ben-Akiva et al., 1991).

Incorporating the dynamic feedback effects in limited time is difficult and an underlying behavioral model should be used to predict the outcome. In this thesis, we explore the potential of anticipatory traffic management on a strategic level and assess the complexity of the corresponding optimization problem as follows. It can be assumed that managers implement management measures to achieve a system optimum: the traffic state with minimum (total or average) travel time (Wardrop, 1952). Without intervention, however, the real-world state is likely to be closer to the user equilibrium than to the system optimum (Klein et al., 2018). We consider a social routing strategy, steering or nudging travelers towards socially-desired routes. However, many approaches in literature (Angelelli et al., 2016; Jahn et al., 2005; Van Essen et al., 2020) relax user constraints and, as a consequence, realized travel time differences can be substantial. In practice, a routing strategy should anticipate user responses to maximize compliance. As a matter of fact, behavioral responses influence travel times, and need to be predicted in order to advise routes that are acceptable to the users. Angelelli et al. (2021), Angelelli et al. (2020) studied such a setting by formulating an integer program using piecewise linearization and developed heuristic solution methods - providing evidence that social rerouting strategies potentially improve network performance but are complex to solve to global optimality in limited time.

Research aim

Initially local and minor disturbances in the urban road network can rapidly impact the conditions on a larger part of the traffic network, and thereby affect the quality of the decisions made by actors operating in the urban traffic domain. This is particularly true for the time-critical delivery operations of LSPs. The research of this thesis is part of the ADAPTATION (ADAptive Planning wiTh Advance Traffic InformatiON) project. The ADAPTATION project has as aim to minimize the impact of traffic disruptions on the operations of LSPs by quickly detecting disturbances, predicting their network-wide impact and adapting the route plan in real time. In the context where small disruptions can have a significant impact on timeliness of LSPs, a better understanding on the spatio-temporal variations in traffic is required - particularly for the highly irregular urban traffic system.

Anticipatory decision-making processes of different actors operating in the urban traffic domain (LSPs, urban traffic managers and individual road users) anticipate the future evolution of the traffic system's LOS as well as the uncertainty therein. Traffic volume data included in traffic information models support such processes. To allow a shift from reactive towards pro-active decision-making processes that account for the development of uncertainties over time, variations in urban traffic volumes near signalized intersections need to be investigated and predicted over multiple timescales. Consequently, the research aim is as follows:

Quantifying and understanding variations that occur in urban traffic volumes at different spatio-temporal levels.

We formulate the following research questions to quantify and understand traffic volume variations:

1. To what degree do 24h urban traffic volume time series show systematic variations, and how to characterize the random variation in volume measurements?
2. What is the influence of the arrival processes near signalized intersections on the varia-

tions in urban volumes and delays?
3. To what degree can the systematic variations be predicted, and how can the characterization of the random variation be used to provide probabilistic volume forecasts over various timescales?
4. What is the potential of anticipatory urban traffic management, in particular a social rerouting strategy, while accounting for different user requirements?

Research approach

The individual research questions will be treated in the subsequent chapters. For each research question, we outline the approach.

Research question 1 is addressed in Chapter 2. In this chapter, we study the variations in the 24h volume time series using 15min increments as collected throughout the Enschede traffic network for two years. Since a share of the systematic variations is recurrent, we infer underlying yet recurrent patterns from the volume data. In this complex setting where patterns are not known in advance, and noise influences our estimates regarding the systematic variability, we develop a 'gray-box' neural network architecture that is not only able to infer non-linear relations, but also extracts the regularities in volumes expressed by of recurrent (long and short-term) profiles. In fact, our method infers physically-meaningful profiles to support application for traffic management purposes and policy making. By allowing small adaptations of the profiles over the days, we reconstruct the systematic variability in the 24h flows over the days while accounting for the noise. To do so, we introduce and estimate a noise-level function, characterizing the natural stochastic fluctuations in the flow as a function of the underlying volume, and adapt the neural network loss function as well as the overall procedure to jointly estimate both the systematic variability and the volume noise characteristics.

Chapter 3 discusses research question 2. We use raw data from induction loop detectors regarding arrival events to statistically characterize arrival processes at signalized intersections. To examine the arrival events while including the inter-dependencies with upstream signals and the dynamics over time and space, we study the arrivals as a point process over different timescales. Using a reconstruction method, we determine the time increments in which demand can be assumed stationary and thereby account for the systematic variability in the arrival rate. For a given interval, we study the arrival process in two ways. We examine the series of events as a counting process, measuring the number of arrivals in an interval, and as a sequence of inter-arrival times. Although these perspectives are fundamentally related, they have very different second-order properties (Daley & Vere-Jones, 2003). We use both the time domain as well as a frequency domain approach to reveal the regularities in the arrivals. In addition, we use simulation to assess the impacts of the arrival structure on the variations in delays.

Chapter 4 addresses research question 3. In this chapter, we study the predictability of systematic variations in urban traffic volumes by developing and consequently evaluating a prediction mechanism. The prediction method uses historical latent profiles to provide volume estimates over multiple timescales, ranging from 15min to 24h in advance. Since profiles yield predictable fluctuations over various horizons, the forecasting task reduces to estimating the magnitude of the profiles over the day. By using the profiles to build up 24h volume time series, we introduce sufficient degrees of freedom to be flexible to adapt the

forecast to a range of scenarios, including situations where recurrent events occur. We start with an initial prediction based on explanatory variables, and construct the 24h, remaining-day and short-term predictions by comparing previous estimates with measurements that are revealed over time. We estimate a systematic error using the noise level function not only to update the forecast over time using state-space smoothing, but also to develop a framework for constructing full density forecasts over the considered timescales. Evaluation is based on an error estimate in the density forecasts relative to the system's predictability as expressed by the statistical characterization of the random variation.

Chapter 5 discusses research question 4. Anticipatory decision making in the context of urban traffic management requires that feedback effects are incorporated in the predictions. To explore the potential of anticipatory traffic management, we propose a social routing strategy that steers the network towards a system optimum while explicitly accounting for the behavioral response of travelers in terms of route choice. We adopt a game-theoretic approach, where a leader (traffic manager) and a follower (travelers) interact according to a Stackelberg game (Josefsson & Patriksson, 2007). In our setting, the response to route advice is anticipated by the leader of the game to propose the best possible advice in terms of total travel time. The corresponding (continuous) optimization problem is formulated as a bilevel program, and we use an implicit reformulation together with parametric analysis of the lower-level solution set to develop and apply a descent method. We explore the potential of the social routing strategy by numerical experiments in test networks.

1.7 Research relevance

The research in this thesis makes several contributions to the literature and to practice. We discuss the scientific relevance in Section 1.7.1, and the relevance for practice in Section 1.7.2.

1.7.1 Scientific relevance

For each chapter in this thesis, we outline the contributions to the literature. A proper theoretical embedding follows in the respective chapters.

- Chapter 2: Despite the attention for the statistical properties of noise in image and signal processing literature, only a few studies statistically characterized the random variation in traffic volume measurements. Although there is evidence that the amount of random variation depends on the underlying systematic patterns, many estimation and prediction methods make restricting assumptions regarding the noise. We introduce a generic noise level model that describes the distribution of the random variation in traffic volume measurements as being signal dependent. Whereas the estimate of the variance of the noise depends on the underlying systematic variation and vice versa, we introduce a method that captures the systematic variability in the shape and the height of the 24h volume time series over the days, including days where events occur. In fact, we reconstruct the systematic variation in volumes using meaningful long and short-term temporal patterns while simultaneously estimating the noise level from the flow measurements;

- Chapter 3: A major share of the variation in volume measurements is the result of the dynamics that occur near signalized intersections. Where many delay estimation methods use naive assumptions regarding the arrival events, we use high-resolution loop detector data to statistically characterize arrival processes. Our approach allows for a comprehensive characterization of an arrival process by considering it as a sequence of inter-arrival times as well as a counting process - in both the time domain and the frequency domain. Correlations over time as well as variations due to upstream interruptions are incorporated, since arrival processes are determined for a high degree by the short-term periodicities related to upstream signals and the platoon dispersion.

- Chapter 4: We introduce a method for predicting urban traffic volumes on various timescales, including a quantification of the uncertainty in the form of full density forecasts. Where long-term patterns are used for forecasts up to 24h ahead, short-term patterns provide predictable fluctuations that cover less time and should not necessarily influence long-term forecasts. Each density prediction accounts for (i) the random variation in volume measurements, (ii) the uncertainty in the current state estimate using a generalized state-space model, and (iii) the systematic prediction error that decreases over time. The method provides accurate and narrow prediction intervals for the various timescales on both minor and major urban roads during recurrent conditions as well as during events.

- Chapter 5: Traffic management measures route drivers towards socially-desired paths in order to achieve a system optimum. In previous attempts, the behavioral response to advice is oftentimes not accounted for since some drivers need to take significantly longer paths in favor of the system. We propose a novel fully anticipatory traffic management strategy called social routing that steers the traffic network towards an efficient but also fair, and therefore achievable and maintainable traffic state. We show that the best possible paths to be proposed by a social routing strategy, while explicitly accounting for behavioral responses to the advice, can be found by solving a bilevel program having a non-unique lower-level solution. A generalized derivative of the of the lower-level link flow solution however exists, and is used in a descent method to find locally-optimal solutions. The strategy can be implemented in a rerouting service to steer traffic towards a fair state with improved network performance.

1.7.2 Practical relevance

The results from this research are relevant for decision makers operating in the urban traffic domain, such as LSPs, traffic managers and individual travelers, as follows. We list a selection of the contributions.

- The data used in this thesis were collected throughout the traffic network of Enschede. The Enschede traffic system is becoming increasingly congested (De Jong, 2020, 2021), and the proposed state-space model together with the inferred temporal patterns can be used, e.g., by urban traffic managers, for monitoring and policy purposes to relieve or mitigate congestion and to meet societal goals in general. In fact, the state-space model provides an estimate of the traffic conditions while accounting for the uncertainty in the aggregated measurements. Although the data were collected in

the Enschede network, the methodological applications go well beyond this network and also apply to traffic conditions-related time series as collected in other urban environments.

- The prediction mechanism developed can be used by LSPs, traffic managers, and traveler information systems to forecast the network-wide traffic volumes on different timescales. In combination with other data sources or models, these predictions can be used to provide probabilistic forecasts regarding delays, travel times or travel speeds. Such forecasts could improve the efficiency of logistics operations in urban areas, mentioned as one the challenges by the Alliance for Logistics Innovation through Collaboration in Europe since urban freight is responsible for 25% of the CO_2 emissions in urban areas (ALICE, 2014).

- The Dutch project 'Talking Traffic' (Partnership Talking Traffic, 2022) focuses, among other things, on reducing fuel and CO_2 emissions for carriers. Trucks communicate with traffic lights and will be provided priority when approaching a signalized intersection in order to prevent the vehicles coming to a standstill. Our developed simulation model based on actual arrival processes under a variety of conditions support authorities in optimizing these dynamic signal plans as well as evaluating the network-wide effects, e.g., by using very short-term predictions regarding the demand, delays and queues at the different approaches.

Chapter 2

Patterns and noise in urban traffic volumes

2.1 Introduction

Traffic networks are increasingly utilized and become less reliable at the same time. Numerous management measures have therefore been designed to improve the utilization of networks. These measures typically react to the prevailing conditions but may fail to achieve their intended outcome since conditions can rapidly change over time. Ideally, management measures are deployed in a pro-active rather than reactive manner in prospect of future conditions resulting from both periodic variations in demand and supply as well as the feedback, e.g., of drivers, in response to the measures and the changing conditions. Making such predictions in limited time is challenging, particularly in an urban setting, which is highly irregular compared to freeways. In any case, to support the development of fast and reliable predictions methods, there is an increasing need to understand the complex urban traffic dynamics.

Fluctuations that occur in traffic networks are typically examined on an accumulated and aggregated level. Network usage is oftentimes expressed by traffic flow (or: volume) time series with regular intervals in the order of 5-15min. These traffic flow time series show clear patterns in time and space, and these patterns can therefore improve predictions and consequently support management decisions since future fluctuations can be anticipated. Apart from the systematic variability, a large share of the fluctuations in time series can be considered random and unpredictable, i.e., noise (Bates et al., 2001; Breiman & Lawrence, 1973; Thomas et al., 2010). Particularly for real-time monitoring and prediction purposes it is important to separate the systematic from the random volume differences so that changing conditions are rapidly recognized in a situation where stochastic high-frequency fluctuations occur in parallel.

Traffic volumes and demand are well known to show systematic variability over various timescales, e.g., within a day and from day to day (Crawford et al., 2017; Rakha & Van Aerde, 1995), and therefore typically examined using 24h time series. Although the intra-

This chapter is based on the following paper: Eikenbroek O.A.L., Thomas, T., Mes, M.R.K., & van Berkum, E.C. Patterns and Noise in Urban Traffic Flows.

day and day-to-day systematic differences are often considered in isolation (e.g., Rakha & Van Aerde, 1995; Stathopoulos & Karlaftis, 2001; Thomas et al., 2008), there is less known regarding the systematic changes in the 24h pattern over the days (Crawford, 2017; Guardiola et al., 2014; Weijermars & Van Berkum, 2005). A simultaneous consideration of both the intra-day and day-to-day variability is however necessary, e.g., to estimate robustness of measures in light of a gradually changing height, width and time of the morning peak (Crawford, 2017). This double timescale examination of traffic volume time series also reveals the systematic variations on timescales longer than 5-15min but shorter than 24h, e.g., due to events and incidents. In fact, these short-term differences are less structured but can still be recurrent. Nevertheless, their frequency of occurrence and magnitude are highly variable, and short-term systematic variations need to be assessed relative to an unknown pattern that would have been realized without the event (Chen et al., 2012; Li et al., 2015). Even though these short-term patterns yield highly valuable information for traffic managers, relatively little is known about such systematic variations in the volumes (Olabarrieta & Laña, 2020; Polson & Sokolov, 2017; Thomas & Van Berkum, 2009) but yield highly valuable information for traffic managers since control actions can deployed to mitigate negative impacts.

Apart from the systematic variability that in theory is predictable, we define the random variation in traffic volume time series as the portion of the fluctuations that shows no pattern and is uncorrelated (Thomas et al., 2008, 2010). With the urban traffic system being inherently variable, an accurate quantification of the distribution of the noise allows for the assessment of measurements and estimates relative to the unpredictability of the system (e.g., by using probabilistic forecasts), but also identifies to what degree control actions can influence the future. In fact, without careful consideration of the random variation during the modeling exercise, one could infer patterns from true noise (*overfitting*) – putting the reliability of forecasts at risk (Makridakis et al., 2018).

Although the systematic variation is the only variability that can be predicted, the true systematic variability in the volumes is rarely revealed. With measurements only providing a single realization of the stochastic and natural variation around the trend, a model is required to separate the signal from the noise. Such a decomposition is highly challenging, since the underlying variability-generating processes are typically far from stationary in an urban environment. In addition, aggregated volume measurements can show patterns over many different timescales which are difficult to model using basic exogenous variables only. An a priori quantification of the noise is nonetheless difficult since systematic volumes can show sudden shocks and there are inter-dependencies between the estimates of the noise characteristics and the systematic variability (Ghosh et al., 2010; Guo et al., 2015; Guo & Williams, 2012; Thomas et al., 2008). Hence, although it is important to disentangle the systematic from the random variation to identify what can(not) be predicted, doing so is challenging in an inherently variable urban traffic environment.

The focus in this chapter is to provide an a posteriori estimate of the amount of repetition and inherent uncertainty in 15min urban traffic volume measurements. In fact, 24h urban traffic volume time series show recurrent patterns that express a major share of the systematic variations. Indeed, time of day and day of the week are important predictors for volumes (Crawford, 2017; Weijermars, 2007), yet traffic flow time series are also impacted by short-term events and show gradual changes over other timescales. Therefore, the measured flows may not necessarily reveal the recurrent patterns *as such*, since the shape, width, and height of the patterns changes over time while noise corrupts the measurements at the

same time. Although these systematic variations show natural variability, they are in fact recurrent and can therefore be incorporated in a model to separate the systematic from the random variation.

Since the characteristics of the noise may depend on the underlying systematic volumes, and vice versa, we propose and apply a data-driven method to infer both jointly. Therefore, we develop a novel uncertainty-aware neural network architecture that retrieves the recurrent patterns in 24h traffic volume time series while accounting for the natural yet systematic variations in these patterns. In addition, we account for the inherent uncertainties in urban traffic volume measurements using a generic noise model that describes the random variation in traffic flow measurements as signal or volume-dependent noise. We apply our method to two years of traffic flow measurements in the city of Enschede, the Netherlands. We discuss throughout the chapter the relevance of the results in light of estimation and prediction methods. The noise model and the results can be used to extract and analyze patterns to be used in predictions with a corresponding confidence interval (Chen et al., 2012; Guardiola et al., 2014; Habtemichael & Cetin, 2016), but also to understand why travel speeds and thus travel times deviate (Li & Rose, 2011).

The remainder of the chapter is organized as follows. In Section 2.2, we discuss a generic volume-dependent noise model. Additionally, we discuss the relevance of estimating the systematic patterns and the amount of random variation. Section 2.3 reviews methods to extract temporal patterns and the noise level. In Section 2.4, we introduce our method to jointly infer patterns and the noise level in urban networks from historical data. We apply our method to study traffic flows in the city of Enschede and show that noise is volume dependent (Section 2.5). Section 2.6 draws the conclusions.

2.2 Systematic and random variations

Separating the systematic from the random variations based on historical measurements requires assumptions on both the dynamics in traffic volumes as well as the source and nature of the noise. In this section, we therefore introduce a generic noise model (Section 2.2.1), and discuss the challenges in distinguishing and estimating systematic and random variations in order to quantify the (un)predictability of volumes (Section 2.2.2 and 2.2.3).

2.2.1 Traffic flow measurements and volume-dependent noise

In urban networks, traffic flow is typically monitored using induction loop detectors near (signalized) intersections. For a fixed loop detector, the signal or measurement x is a mapping $x : D \times T \rightarrow \mathbb{N}^0$, with D the set of days, and T the *time domain* for a single day. Here, a measured 24h traffic flow time series x_d, at day $d \in D$, is a vector $x_d = \big(x_{d,1}, x_{d,2}, \ldots, x_{d,|T|}\big)$, with $|T| = 96$ in our case with 15min measurement intervals. The observed signal $x_{d,t}$ is a realization of a random variable described by the sum of the underlying deterministic *systematic flow* $s_{d,t}$ and random variable $\varepsilon_{d,t}$ (noise), i.e.,

$$s_{d,t} + \varepsilon_{d,t}. \tag{2.1}$$

Hence, volume measurements can be considered outcomes of an experiment due to random and unpredictable behavior of users from an observer's perspective (see Section 2.2.3).

Assuming that a large share of the random variation can be attributed to the random arrival processes, the theory on renewal processes (e.g., Cox & Lewis, 1966) hints on a noise variance in aggregated volume measurements as a function of the underlying arrival rate. Hence, we adopt a general volume-dependent noise model (see Foi et al., 2008; Liu et al., 2014) in the real space with

$$\varepsilon_{d,t} = \sigma(s_{d,t})\eta_{d,t},$$

where $\sigma : \mathbb{R}^+ \to \mathbb{R}^+$ is the *noise level* (as a function of s) and $\eta_{d,t}$ is a zero-mean independent random variable with $\eta_{d,t} \sim \mathcal{N}(0,1)$. Then, the noise is distributed according to a *heteroscedastic* Gaussian with volume-dependent variance, i.e.,

$$\varepsilon_{d,t}|s_{d,t} \sim \mathcal{N}(0, \sigma^2(s_{d,t})),$$

for each (d,t). In this chapter, we estimate a noise function $\sigma(s)$ of the following form (Foi et al., 2008; Liu et al., 2014):

$$\mathrm{Var}(\varepsilon|s) = \sigma^2(s) = \alpha s^{2\gamma} + \beta, \tag{2.2}$$

with parameter vector $\theta = (\alpha, \beta, \gamma)$, where $\alpha, \beta, \gamma \geq 0$. This generic noise-level function (2.2) can in particular take the form of additive white Gaussian noise ($\alpha = 0$), and 'Poisson' noise ($\alpha = 1$, $\beta = 0$, $\gamma = \frac{1}{2}$). The latter noise model would be the result of a renewal process with the inter-arrival times being exponentially distributed (with an interval-dependent arrival rate).

A direct and accurate estimation of the parameters of the noise model $\sigma^2(s)$ requires a reconstruction of s from x, see (2.2). The *true* s, however, is not available. Where 'local' information can be used to obtain a smoothed estimate of s in case of slowly-varying or stationary conditions, traffic volume time series may show rapid changes over the 15min intervals. Moreover, without prior knowledge on s, extracting patterns from historical data is a difficult task: the measurements are corrupted by noise that depends on the systematic flow (Chen et al., 2012; Chen et al., 2008; Thomas et al., 2008). Consequently, noise quantification can only occur by making prior assumptions using domain knowledge about the noise (process) and/or the underlying traffic flow patterns.

2.2.2 Systematic variations

Traffic flow measurements show systematic variations over various timescales. Many of these patterns are recurrent, and can therefore improve prediction and estimation methods (Vlahogianni et al., 2014), by matching recent measurements to historical ones. Such patterns are known to exist in time and space, and often assumed to be the result of the variability in the demand (Thomas et al., 2008). Typical 24h time series show an M-shaped curve, with the exact intra-day shape being dependent on the day of the week (or groups thereof) (Li et al., 2015; Rakha & Van Aerde, 1995; Weijermars & Van Berkum, 2005). Apart from the changing shape of different days of the week, variations also occur on timescales longer than 24h, e.g., due to seasons that can be described using a slowly-changing yet variable magnitude (or: height) (Crawford et al., 2017; Weijermars & Van Berkum, 2005; Zhong et al., 2020).

Not only long-term volume variations are recurrent, but also repeating short(er)-term variations exist that cannot be explained by a daily pattern, e.g., due to events. Such fluctu-

ations are more variable in their frequency of occurrence and accompanying magnitude and shape. The short-term patterns are difficult to recognize and predict, since occurrence might be unpredictable (e.g., accidents). Nonetheless, in particular these patterns can improve estimation and prediction but occur, from a modeling perspective, in the residual time series and therefore requires an estimate of the latent trend. The regularity in the resulting volume differences (relative to the trend) is challenging to capture since impacts highly depend on time and location-specific factors such as the residual capacity.

24h time series are not necessarily the same as the recurrent temporal patterns or *profiles* as we will call them. On the contrary, seemingly different 24h time series may actually consist of the same latent profiles. The 24h time series then look different because the underlying basic profiles are subject to small transformations (e.g., height) that change from day to day. Identifying profiles from historical data is challenging, since additional variables are possibly be needed to explain the occurrence and shape of the pattern (Crawford et al., 2017; Guardiola et al., 2014). Yet, it is virtually impossible to identify and collect all variability-inducing factors. This is particularly true for recurrent patterns related to events and incidents and these fluctuations should not be captured by the intra-day trend. Therefore, we adopt an unsupervised approach in this chapter. In any case, the profiles could be incorporated in prediction methods to quickly anticipate future conditions since the prediction task then occurs in a space with fewer dimensions than the measurement space. For example, provided a recurrent 24h volume shape, a remaining-day forecast can be reduced to predicting a single scaling magnitude thereby providing an estimate for the flows over time (Wagner-Muns et al., 2018).

2.2.3 Random variation

A noise term such as ε in (2.1) is inevitable for any measurement time series. Random variation in traffic flow time series is due to random variation in physical dynamics (*process noise*) (e.g., ad-hoc decisions, stochastic road capacity, queuing dynamics, unknown traffic management measures in place - see, e.g., Breiman and Lawrence (1973) and Chen et al. (2008)). The noise characteristics are not fixed but depend on the information of the observer and the aggregation level (in time and space) (e.g., Oh et al., 2005; Son et al., 2014; Vlahogianni & Karlaftis, 2011). For example, the impact of traffic signal cycles might be systematic and apparent when using short measurement intervals, but is part of the (quasi-)random variation using 15min increments (Thomas et al., 2008), i.e., within 15min, individual red and green times cannot be identified and therefore contribute to the noise. Also, incidental errors in signal processing (*measurement errors*) are included to the noise. Although many researchers are aware of noise due to measurement errors (e.g., Briedis & Samuels, 2010; Yang, Wu, et al., 2019) and noise estimation is intrinsically attached to any traffic measurements-related exercise, less has been published about the amount of variation in traffic time series due to the inherent randomness of the processes (see Section 2.3.2).

Knowledge about the amount of noise supports inference of the current state or trend based on recent measurements, including an accompanying estimate of the level of uncertainty. For example, an abrupt change in measurement volume $x_{d,t}$ compared to $x_{d,t-1}$ can indicate a change in the system, while in practice (a part) the of the fluctuations can also occur due to the inherent variability of the system. A *(3-)sigma clipping approach* (Chen

et al., 2012; Guo et al., 2015; Li et al., 2015) identifies an outlier if

$$|x_{d,t} - F_{d,t}(y)| > 3\sigma(s_{d,t}) \tag{2.3}$$

holds for some (d, t), with $F_{d,t}(y)$ an estimate for $s_{d,t}$ using parameters y (see Section 2.3), e.g., naive estimate $F_{d,t}(y) = x_{d,t-1}$. Obviously, various estimates of $\sigma^2(s)$ lead to the identification different sets outliers. Since outliers as in (2.3) potentially contain information about the prevailing state and the underlying physical dynamics (Koen & Lombard, 1993), an accurate estimate of $\sigma^2(s)$ is required to distinguish random and systematic variations. Moreover, noise levels are explicitly used in the estimates of the systematic variation and accompanying confidence bounds. For example, the widely-applied particle filter techniques (e.g., Kalman filter - see Kalman (1960)) require an estimate of the noise (co)variance to relate systematic flows to measurements. Guo et al. (2015) developed therefore a filter that incorporates the conditional variance of the noise in traffic flows. Hence, even if parameters of the statistical noise model are not of primary interest, they also influence the estimate of other parameters particularly with small data sets (Aravkin & Van Leeuwen, 2012; Kamarianakis et al., 2005).

The stochastic setting makes the noise level function $\sigma^2(s)$ provides a lower bound with respect to the best-possible accuracy of prediction methods. We show that the difference between the total squared error and variance of the random variation in a prediction model indicates the size and nature of the error in the estimate of the systematic variation, i.e., the *systematic error* (Hunt et al., 2007; Thomas et al., 2010). A prediction scheme continuously estimates $\bar{x}_{d,t}$ for some (d, t) in the future. It is typical to evaluate an estimate by comparing it to measurements. The error $(\bar{x}_{d,t} - (s_{d,t} + \varepsilon_{d,t}))$ is a random quantity, and the expected squared error $\mathbb{E}[(\bar{x}_{d,t} - (s_{d,t} + \varepsilon_{d,t}))^2]$ should thus be small on average (Karlin & Taylor, 2012). The squared error becomes (see Karlin and Taylor (2012) for a derivation)

$$\mathbb{E}[(\bar{x}_{d,t} - (s_{d,t} + \varepsilon_{d,t}))^2] = \sigma^2(s_{d,t}) + (\bar{x}_{d,t} - s_{d,t})^2.$$

Hence, the best predictor $\bar{x}_{d,t}$ for measurement $x_{d,t}$ is $\bar{x}_{d,t} = s_{d,t}$, but still has an expected squared error that equals the conditional variance of the noise. In any case, a perfect prediction scheme can only achieve (on average) a squared error that equals the variance of the noise. Interestingly, the previous exercise also provides a network-invariant measure to compare different prediction schemes $\bar{x}$. Since prediction methods are difficult to compare (Vlahogianni et al., 2014), we propose metric $(\bar{x}_{d,t} - s_{d,t})^2$, that requires however an estimate on the amount of random variation (which might be dependent on the network or other factors).

2.3 Estimation in urban traffic networks

In the previous section, we indicated the relevance of labeling variations as either systematic or random. In this section, we show that extracting systematic variations is inter-related with estimation of the noise level function $\sigma^2(s)$.

Provided an identification of similar traffic conditions, one could estimate the systematic value and noise variance using standard statistical techniques (e.g., sample mean and variance). However, identification of more or less uniform conditions is rather difficult in our setting due to the variations with different resolutions. Therefore, a model is used to capture

the systematic variability. Reconstruction problems aim to infer the systematic variations s from noisy historical observations x. Typically, a *forward model* $F(y)$ is defined, mapping parameter $y \in \mathbb{R}^m$ to s so that (see, e.g., Scherzer et al., 2009)

$$F(y) = s.$$

Ideally, either the underlying parameters y and/or model $F(y)$ are known in advance. Given a forward model $F(y)$ (known parameters y), *inverse problems* compute parameters y (model $F(y)$) using measurements x (Hansen, 2010). Reconstruction problems usually solve a variation of the optimization problem

$$(P): \quad \min_{y \in Y, \theta} f(x, y, \theta)$$

to choose the best reconstruction among all available ones (Aravkin & Van Leeuwen, 2012). Objective function $f(x, y, \theta)$ in (P) measures the *reconstruction error* between $F(y)$ and x, using noise model $\sigma^2(s)$ parameterized by θ. $Y \subseteq \mathbb{R}^m$ is the *feasible set* that restricts the choice of y, which appears as variable in (P). $F(y)$ can be non-linear or even 'black box' (e.g., a neural network). In general, (P) is a complex optimization problem that is hard to solve to global optimality. Where an estimation method as (P) makes a trade-off between domain knowledge and fit, the estimated noise level should capture random variations due to the underlying processes rather than being solely a result of the modeling exercise.

$F(y)$ contains beliefs about the patterns in traffic dynamics and the relation with measurements. With respect to a freeway setting, forward model $F(y)$ is often based on partial differential equations (Lighthill & Whitham, 1955) and a fundamental diagram (e.g., Nantes et al., 2016), and dynamics are then assumed to be 'smooth' among neighboring measurement loops. For the urban case, there is less consensus about the forward model $F(y)$. In fact, the underlying network structure is irregular compared to freeways, due to a vast variety of trip types, activities, modes, etc. At the same time, the amount of noise is usually not known apriori, and noise parameters should then be inferred from historical data. In the remainder of this section, we discuss methods to extract patterns that compose s and the statistical properties of ε from x, respectively.

2.3.1 Pattern extraction methods

Optimization problem (P) is not always solved explicitly, but assumptions about the traffic dynamics are used to find systematic variations. Due to the high repetitive nature of activities in time and space, a typical daily pattern is often constructed as the intra-day trend. In line with this assumption, clustering, principal component analysis, and non-negative matrix factorization group similar time series (Chrobok et al., 2004; Jiang et al., 2015; Xing et al., 2015; Yang, Wu, et al., 2019). These approaches retrieve m features, and each feature $y^i \in \mathbb{R}^{|T|}$, $i = 1, 2, \ldots, m$, typically describes a 24h recurrent pattern for a measurement location.

Although the above-mentioned methods are intuitively appealing, and can relatively easily be used to infer spatio-temporal patterns for stretches of freeways, they are not well applicable in the context of an irregular urban system. In general, temporal and spatial patterns show gradual changes in time and space. Clustering does not capture these changes well. In addition, different measurement locations might show different dynamics on dif-

ferent time scales, which need to be considered simultaneously.

In principle, there is a set of underlying yet unobservable profiles that together describe the variations s. These variations might be highly dependent on the location and, from a computational perspective, on the noise characteristics. Because of the location-dependent characteristics, it is natural to analyze network-wide variations by extracting temporal patterns for each location, and then cluster these patterns based on their variations (Ji & Geroliminis, 2012). These profiles should capture the temporal systematic variations over various timescales, without making any assumptions about spatial variations (Koen, 2003). In fact, recent evidence shows that distant locations may still possess strong cross-correlations (Ermagun & Levinson, 2019). It is therefore that also recent big-data methods that scan the state of the network as images (e.g., Ma et al., 2017) are less-suitable for our case since they implicitly define a neighborhood for each measurement location. In an urban setting where oftentimes only a very small share of the network is covered with roadside sensors, such a neighborhood is not trivially defined and can even be dynamic.

2.3.2 Noise level estimation

The noise level is usually not known in advance, and noise parameters are then to be inferred from historical data. A traditional time series approach uses a Fourier Transform and labels all high-frequency fluctuations over time as noise. Disaggregated approaches identify correlations among variations in demand, supply, and measurement devices under different conditions. This approach is frequently applied to estimate random variation for uninterrupted flow in unsaturated conditions (e.g., Breiman & Lawrence, 1973). For aggregated measurements in an urban setting, random variation with respect to arrivals and intersection capacity should then be quantified. Such an approach requires high resolution event data, which is typically less easily accessible. Here, we therefore discuss data-driven approaches using aggregated measurements.

Data-driven approaches often involve some form of *detrending*, i.e., find $F(y)$ so that residuals e statistically follow ε. It is difficult to define termination criteria with respect to detrending since the statistical properties of ε are not known in advance. Therefore, one can either compare $F(y)$ with other fits, or test e for randomness (Koen, 2003). When there is no autocorrelation in e, the noise level can then be estimated by relating the variance of the noise with the underlying estimated systematic flow. In an abstract form, one solves

$$(Q): \qquad \min_{\theta} \tilde{f}(x, y, \theta),$$

where x, y are fixed, and $\tilde{f}(x, y, \theta)$ measures the distance between noise level function $\sigma^2(s)$ and residuals e (as function of s). Nonetheless, assuming no autocorrelation in e, the variance in the residuals only provides an upper bound with respect to the variance of the inherent variability of the urban traffic system: the variance in e consists of (i) measurement errors, (ii) random variation in the underlying dynamics, and (iii) an inaccurate estimation of the systematic variation (Hunt et al., 2007).

Despite the strong interest for the statistical properties of noise in image and signal processing (e.g., Foi et al., 2008; Liu et al., 2014), relatively few studies (Chen et al., 2008; Ghosh et al., 2010; Guo et al., 2015; Guo & Williams, 2012; Huang et al., 2018; Thomas et al., 2008; Tsekeris & Stathopoulos, 2006) investigate the noise properties and volatility in traffic measurements. Although there is evidence that the amount of random variation

depends on the underlying (systematic) flow (e.g., Breiman et al., 1977; Chen et al., 2012; Chen et al., 2008; Luttinen, 1996; Thomas et al., 2010), many estimation and prediction methods (e.g., Wang & Papageorgiou, 2005) assume white Gaussian noise. We cite some exceptions: Li and Rose (2011), Thomas et al. (2010), Wagner-Muns et al. (2018), and Yang, Yang, et al. (2019) on traffic volumes, and Nantes et al. (2015), Tang et al. (2018), and Yang et al. (2010) on travel times.

The (conditional) heteroscedastic nature of the noise induced time series models such as (G)ARCH to model the variance of the noise over time (Guo et al., 2015; Huang et al., 2018; Tsekeris & Stathopoulos, 2006). Only a few (Breiman & Lawrence, 1973; Chen et al., 2008; Ghosh et al., 2010; Thomas et al., 2008), however, tried to estimate the random variation as a function of the systematic flows. In contrast to our research, the mentioned studies were limited to either only a few measurement locations, or made a relatively simple estimate for the systematic flows - typically independent of the (conditional) variance in the noise. Some other studies use characteristics of random variation in their approach. For instance, Ermagun and Levinson (2019) assume that systematic variations are captured if the autocorrelation in the residuals is null. Different measures that quantify the amount of random variation are discussed in Tang et al. (2014), Wang et al. (2013), and Yin and Shang (2016).

2.3.3 Joint estimation

Based on previous subsections, an estimate of the conditional noise variance is required to infer temporal patterns, and systematic variations should be captured to estimate the noise variance. Standard clustering estimates with a dominant day-of-the-week dependent volume pattern do not suffice in this case, since only a share of the systematic variations is then captured - which is likely to lead to an overestimation of the noise variance.

We estimate the predictability of the volumes by quantifying the amount of systematic and random variation in the data set, i.e., we solve (P) and thereby find y and θ jointly. Specifically, we use an estimate on θ to initialize an iterative procedure that finds both y and θ. Here, it is natural to explicitly incorporate the estimate with respect to noise parameters in the optimization problem (P). Therefore, we consider the maximum likelihood estimation (MLE) problem for y given by

$$(P_{\text{ML}}): \qquad \min_{y \in Y} \sum_{d,t} \frac{1}{2} \left(\log\left(2\pi\sigma_{d,t}^2\right) + \frac{(x_{d,t} - F_{d,t}(y))^2}{\sigma_{d,t}^2} \right),$$

with $\sigma_{d,t} = \sigma(F_{d,t}(y))$. In the next section, we describe our method in detail.

2.4 Method

With different daily traffic patterns from day to day (e.g., Zhang et al., 2022), we aim to reconstruct these systematic flows using long and short-term recurrent temporal patterns as building blocks. Due to the dependencies between estimating systematic variations and the statistical properties of random variation, we require a method that is able to infer both from historical data. Therefore, we introduce a general unsupervised learning procedure to

extract the temporal patterns in presence of volume-dependent noise with unknown characteristics in Section 2.4.1, and discuss termination criteria in Section 2.4.2. We use a neural-network architecture (Section 2.4.3 and 2.4.4) as a data-driven method to extract the temporal patterns.

Before we discuss our procedure in detail, we provide a high-level overview in Algorithm 1. This method extracts recurrent profiles y and noise-level parameters θ from measurements x.

1: Initialize noise-model parameters θ;
2: **for** each measurement location in the network **do**
3: Initialize the number of long-term profiles $m^l = 0$, and number of short-term profiles $m^s = 0$;
4: Find the long-term profiles as follows:
5: Let $m^l = m^l + 1$, and use the neural network (**Section 2.4.3**) to retrieve long-term profiles y^i, $i = 1, \ldots, m^l$, by solving (P_{ML}) using forward model $F(y)$ (**Section 2.4.1**) and noise model parameters θ;
6: If termination criterion of the long-term profiles (**Section 2.4.2**) is met, goto line 7 (possibly, $m^l = m^l - 1$), else goto line 5;
7: Find the short-term profiles as follows:
8: Let $m^s = m^s + 1$, and use the neural network (**Section 2.4.3**) to retrieve short-term profiles y^i, $i = m^l + 1, \ldots, m^l + m^s$, (P_{ML}) using forward model $F(y)$ (**Section 2.4.1**) and noise model θ;
9: If termination criterion of the short-term profiles (**Section 2.4.2**) is met, goto line 2 (possibly, $m^s = m^s - 1$), else goto line 8;
10: **end for**
11: Update noise model-parameters θ by solving (Q) (**Section 2.3.2**) based on all measurement locations;
if θ is unchanged, terminate, else goto line 2;

Algorithm 1: High-level overview of the procedure to find profiles and noise level parameters.

2.4.1 Profiles

Although 24h traffic time series show a high degree of regularity, they also show gradual changes in the shape and height over the days (Coogan et al., 2017; Crawford, 2017; Weijermars & Van Berkum, 2005) including shorter-term deviations relative to the intra-day trend, e.g., due to events. A large share of these patterns are recurrent yet show natural variations in their time of occurrence but also in their shape and magnitude. We aim to capture these recurrent variations by means of profiles.

We define a normalized (basic) profile as a recurrent sub time series (temporal pattern) that starts at midnight (00:00h) and has a total flow of 1. Basically, we describe the systematic intra-day flow using a linear combination of these (transformed) profiles. Distinct from earlier approaches (see Section 2.3.1), we explicitly consider profiles of different temporal scales to distinguish long- (24h) and short-term (less than 24h) variations, and use the transformations to capture the natural yet systematic differences over the days.

We introduce the long and short-term profiles. A long-term profile is a 24h time series, while a short-term profile is (a part of a) 7.5h time series. We hypothesize that a 7.5h time series is sufficient to capture short(er)-term fluctuations while excluding long-term variations. The profiles correspond to the temporal recurrent patterns in the time series. Although time series for a measurement location show a high degree of correlation from day to day, there are small and gradual but natural variations in, e.g., period between morning peaks, which should not be included in the random variation and are in fact systematic. To appropriately capture these natural fluctuations, we introduce transformations on the profiles. In our case, measurements occur with 15min increments but we define the profiles on time domain $\mathcal{T}$, which is finer than the measurement domain T (i.e., $T \subset \mathcal{T}$). That is, each profile i has a corresponding shape, described by a 24h time series y^i with 3min increments.

We use domain knowledge to assure that our method extracts profiles that are *physically meaningful*. To capture natural variations in systematic flows over the days, we introduce transformations with respect to magnitude (scale in magnitude), period (shift in time), and shape (scale in time). Since the (shape of the) underlying profiles are fixed over days (i.e., day-invariant), we allow transformations to change from day to day (and from location to location) so that the variations are captured on different temporal scales. First, the magnitude (height) of each profile is variable. Second, we introduce shift $\delta \in \mathbb{N}^0$, with δ the shift in time (in $3 \times \delta$ minutes) that the start time of the profile deviates from midnight. Short-term profiles are allowed to start throughout the day, long-term profiles are assumed to be highly related to the daily demand and can therefore only shift for a maximum of 15 minutes (e.g., to capture a small shift in rush hours). Finally, we allow short-term profiles to be *stretched* (scaled in time).

We formally discuss these possible transformations in our forward model $F(y)$. Each measurement location has a corresponding set of profiles $\mathcal{I}$. Each profile $i \in \mathcal{I}$ is transformed and then contributes to the final reconstruction $F_d(y) \in \mathbb{R}^{|T|}$ for a day d by means of 24h time series $z_d^i = z_d^i(y) \in \mathbb{R}^{|\mathcal{T}|}$. Note that the resulting 15min reconstruction $F_{d,t}(y)$, for some (d,t)-combination, is the sum of 5 consecutive 3min volumes:

$$F_{d,t}(y) = \sum_{\tau=5(t-1)+1}^{5t} \sum_{i \in \mathcal{I}} z_{d,\tau}^i. \tag{2.4}$$

We discuss how the resulting flow vector z_d^i for profile i at day d is the result of a set of transformations, expressed by a, C and $\mathcal{F}$:

$$z_d^i = a_d^i \left(C^\delta \mathcal{F}(y^i) \right).$$

We formally introduce these transformation mappings:

- *Scale in magnitude*; $a_d^i \geq 0$ describes the magnitude of profile i at day $d \in D$. Hence, the magnitude a_d^i is equal to the net flow that profile i contributes to $F_{d,t}(y)$;
- *Shift in time*; C^δ denotes a shift operation of $\delta = \delta_d^i \in \mathbb{N}^0$ of profile i at day d. For long-term profiles we assume that this shift is cyclic, i.e.,

$$C^\delta = \begin{bmatrix} 0^{(|\mathcal{T}|-\delta)\times\delta} & I^{(|\mathcal{T}|-\delta)} \\ I^\delta & 0^{\delta\times(|\mathcal{T}|-\delta)} \end{bmatrix},$$

with I and 0 the identity and null-matrix, respectively. For short-term profiles, we introduce a conventional shift, i.e., for $\delta \in \mathbb{N}^0$,

$$C^{\delta} = \begin{bmatrix} 0^{\delta \times (|\mathcal{T}|-\delta)} & 0^{\delta} \\ I^{|\mathcal{T}|-\delta} & 0^{(|\mathcal{T}|-\delta)\times\delta} \end{bmatrix};$$

- *Stretch in time*; for a short-term profile, $\mathcal{F}(y)$ denotes the discrete-time convolution of profile y with kernel g^q ($q \geq 0$), i.e.,

$$\mathcal{F}(y) = (y * g^q)_\tau := \sum_{l=-\infty}^{\infty} y_l g_{\tau-l}, \tag{2.5}$$

with g^q being a discretized Gaussian distribution $\mathcal{N}(0, q)$. The convolution operator approximates a 'stretch in time'.

We underline that the above approach allows multiple short-term profiles at the same day by considering *copies* of, e.g., one short-term profile. Note that it is not necessary that all profiles are *active* at day d, i.e., possibly $a_d^i = 0$ for some i.

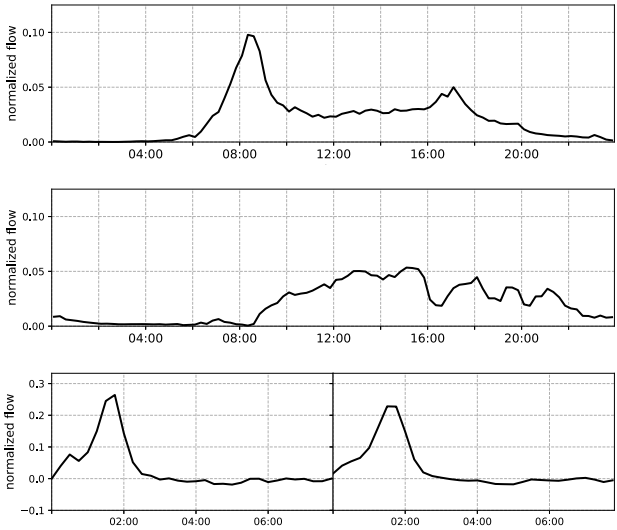


Figure 2.1: Extracted profiles for measurement location A from Figure 2.4. The upper and middle figure show the long-term profiles. The lower figure shows the extracted and the stretched (single) short-term profile.

Figure 2.1 displays the normalized profiles y for a measurement location in Enschede (A in Figure 2.4, see Section 2.5.1). We show the extracted long-term profiles (upper and middle figure) and the short-term profile (lower figure, with an example of the short-term profile being stretched) on time domain T. After analyzing the patterns and corresponding activations, we interpret the shape of the profiles as follows: the upper long-term is the 'peak profile', the middle pattern is the 'base profile', and the short-term profile mostly corresponds to additional traffic due to soccer matches. Note that the short-term profiles are

only limitedly stretched, i.e., many of the short-term systematic variations show a similar pattern in time. In Section 2.4.3 we describe the computational method used to extract profiles.

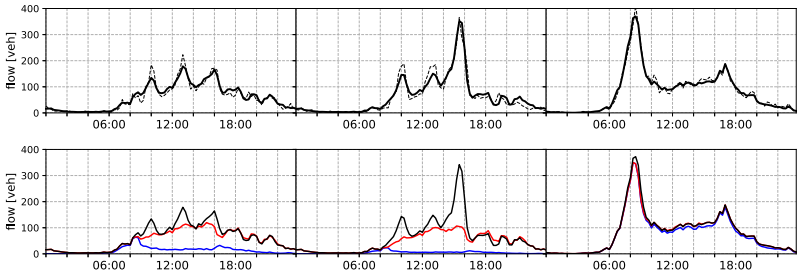


Figure 2.2: Upper figure: measurements (dashed line) and reconstruction (solid line) for three consecutive days (January 30, 2016 (Sat) - February 1 (Mon)) at measurement location A. Lower figure: each reconstruction is a linear combination of the same - but transformed - underlying profiles of Figure 2.1. The different colors indicate the different profiles.

Together with the possible transformations, the profiles capture systematic deviations in the measurements over various timescales. Figure 2.2 depicts a set of reconstructions (upper figure) of our model. In the lower figure, we show that the two long-term profiles (peak profile in blue, base profile in red) together with the short-term profile (black) are cumulatively used to reconstruct measured time series.

We remark that the peak profile might not be apparent in resulting measurements (e.g., in the left and middle time series in Figure 2.2), but is actually part of the underlying flow. The short-term profile is used in the middle of the reconstructed 24h time series to capture short-term variations. Although these short-term deviations look different, they are captured by transforming a single profile. The profiles of Figure 2.1 are similarly used to reconstruct every 24h time series of this measurement location in the data set. As such, we accurately capture the changing shape and height of the volume pattern over the days.

2.4.2 Termination criteria

In the previous subsection, we explained the reconstruction method for a given number of long and short-term profiles, which we denote here by m^l and m^s, respectively. Theoretically, we could add as many profiles as days, and obtain a perfect reconstruction of the measurements (including the noise), which would lead to poor performance of a foreseen prediction method (see Section 2.2.3). Hence, a crucial question is thus the number of profiles that is needed to capture systematic variations without fitting noise. Here, we use an iterative approach, i.e., we begin with a single (unknown) profile and iteratively add profiles (and solve corresponding (P_{ML})) until a termination criterion is met.

Our goodness-of-fit procedure compares a reconstruction (for an estimated noise model $\bar{\theta}$) assuming m (either m^l or m^s) profiles with the reconstruction having $m+1$ profiles. In fact, if the reconstruction using $m+1$ profiles is not substantially improved compared to the previous one, we reject profile $m+1$, and conclude that we found all m profiles.

The likelihood function in (P_{ML}) allows us to adopt a set of statistical termination criteria, i.e., to test the relative improvement among different reconstructions. We mention the theoretically-appealing likelihood-ratio, Wald, and Lagrange multiplier test (Greene, 2003), and the intuitive 'elbow'-test (finds the number of profiles for which an added profile is not substantially improving: the 'elbow' in the profiles-error graph (Yang, Wu, et al., 2019)). Both type of methods are not well-suited for our setting. The first set of tests requires a regularity condition to hold, which is not necessarily true in our case. Current 'elbow'-tests might be highly dependent on the estimate $\bar{\theta}$ of θ. We desire a generic yet practical method that is relatively independent of this estimate.

We propose a goodness-of-fit method that uses a characteristic of the noise: the noise is uncorrelated in time and space. The statistical test uses the empirical distribution of the autocorrelation among successive residuals. With respect to long-term profiles m^l, we basically add long-term profiles until *on average* no correlation in successive residuals is left. We iteratively increase the number of short-term profiles m^s until the outliers do not follow a systematic pattern.

In the remainder of this section, let ρ_d be the remaining autocorrelation with lag 1 in residuals e_d of day d (for a given number of profiles m^l). We are mainly interested in the correlation in an aggregated sense (i.e., over days), since large absolute values of ρ_d are highly influenced by disruptions that cover a substantial amount of time, but should not be covered by a long-term profile. For a given number of profiles m^l, let $G_{m^l}(\rho)$ be the corresponding cumulative distribution function of the remaining autocorrelation among the residuals over the days. We use the two-sample Kolmogorov-Smirnov (K-S) test

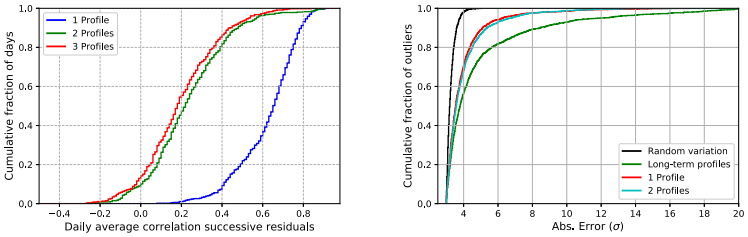
$$D = \sup_{\rho} |G_{m^l+1}(\rho) - G_{m^l}(\rho)| \tag{2.6}$$

to test whether G_{m^l} and G_{m^l+1} have the same underlying distribution. The K-S test is mainly sensitive with respect to changes in the median value and the shape (Babu & Feigelson, 2006). Other tests (Stephens, 1974) that use a distance between empirical distribution functions are typically more sensitive to differences in the tails of the distributions. Whereas these tails are mainly caused by short-term systematic variations that should not be covered by long-term profiles, we use the two-sample K-S test (2.6) to identify whether a new profile should be added. If the null-hypothesis that distribution functions G_{m^l} and G_{m^l+1} have the same underlying distribution is rejected, we accept the new profile $m^l + 1$. Otherwise, we conclude that there are m^l underlying long-term temporal profiles. Figure 2.3a shows an example of a measurement location in which we decide on two long-term profiles. Indeed, the third profile is rejected based on (2.6).

To decide on the number of short-term profiles m^s, we proceed as follows. We determine the cumulative distribution function of the outliers for which

$$|e_{d,t}| > 3\sigma(s_{d,t}) \tag{2.7}$$

holds. Given that there is only random variation left, we still expect some outliers that occur by chance (approximately 0.3% of the measurements). Therefore, we compare the empirical distribution of the outliers (2.7) with the distribution of outliers under $\varepsilon_{d,t} \sim \mathcal{N}(0, \sigma^2(s_{d,t}))$. We use the one-sample K-S test to test whether the cumulative distribution of outliers is different from outliers arising from noise. If, according to the K-S test, these distributions are not different, we conclude that there are m^s underlying short-term profiles



(a) Distribution function of the remaining autocorrelation over the days. *(b) Empirical distribution function for outliers as in* (2.3).

Figure 2.3: Illustration of termination criteria for long-term profiles (left figure), and short-term profiles (right figure) at location A in Figure 2.4.

(which could be 0). Similar to the procedure to determine the long-term profiles, we adopt a two-sample K-S test to test whether cumulative distribution function of outliers in $m^s + 1$ is different from m^s. Figure 2.3b shows an example in which we have one underlying short-term profile, because adding a second one does not improve the distribution of the outliers. We underline that the different empirical distributions in Figure 2.3b are not necessarily based on the same set of outliers. After termination, there are still outliers left which cannot be explained by the noise. These outliers are probably caused by disruptions or events with no recurrent pattern.

Our procedure is quite powerful, because the cumulative distribution function of the residuals turns out the be sensitive to large outliers. These are exactly the outliers we are interested in, and thus should be captured by the profiles.

2.4.3 Neural network architecture

Neural networks are able to derive complex and non-linear relations that appear in large-scale data sets. Typically, neural network methods require very few assumptions in advance and show impressive performance in prediction tasks (e.g., Lv et al. (2015), Ma et al. (2017), and Polson and Sokolov (2017)). On the other hand, neural networks are often criticized for being 'black box' approaches, i.e., the derived relations can only be limitedly extracted after the learning process. Concurrently, there is increasing need to extract patterns that are physically meaningful (Bascol et al., 2016). In this chapter, we develop a *convolutional neural network* (Bascol et al., 2016; Krizhevsky et al., 2012) that reproduces the flow s from x following the conditions in Section 2.3: the method is designed such that it is capable to recognize and extract temporal patterns (profiles) from measurements.

We use a convolutional autoencoder (see Appendix 2.7). We divide this architecture into two levels; the upper level is designed to extract long-term profiles (Line 4-6 in Algorithm 1), the lower level is such that it finds the short-term profiles (Line 7-9 in Algorithm 1). To simplify the presentation in this section, we assume here that there is only a single long and short-term profile (V^{long} and V^{short}, respectively) to be learned.

We define a noisy *softmax* function to replace some of our scaling functions to improve

convergence. The softmax function for vector $z \in \mathbb{R}^m$ is defined as follows:

$$\text{softmax}(z)_j = z_j \frac{e^{Lz_j+\eta_j}}{\sum_i e^{Lz_i+\eta_i}}, \quad j = 1, 2, \ldots, m,$$

with η_j randomly sampled from $\mathcal{N}(0, 1)$ during training, and $\eta_j = 0$ during reconstruction. Here $L > 0$ is a *sufficiently large* number.

Consider time series vector $x = x_d \in \mathbb{R}^n$ with $n = 96$ as input of the network. Let $W^{\text{magn},l} \in \mathbb{R}^n$ be a filter to extract the magnitude from input x with $l = 1, 2, \ldots, 10$. Then,

$$a^{\text{magn},l} = g(W^{\text{magn},l}x + b^{\text{magn},l})$$

and

$$a^{\text{magn}} = \max_l(a^{\text{magn},l})$$

is the final magnitude corresponding long-term profile V^{long}. In parallel, we learn the shift of the long-term profile. Therefore, let $W^{\text{shift},l} \in \mathbb{R}^m$ be a filter to find the shift in time,

$$a_t^{\text{shift},l} = g\left(\sum_{s=1}^{m} W_s^{\text{shift},l} x_{t+s-1}\right), \quad t = 1, 2, \ldots, \delta + 1$$

and

$$a_t^{\text{shift}} = \max_l(a_t^{\text{shift},l}),$$

with $m = 96 - \delta$, and δ is the shift in time (see Section 2.4.1). Scaling function g is the *leaky relu* function (Maas et al., 2013).

We construct a single activation vector that indicates the magnitude of the profile at time t, i.e., for each t,

$$a_t = (\text{softmax}(a_t^{\text{shift}}))a^{\text{magn}}.$$

Hence, the magnitude and shift of the profile for a day d are learned independently, and activation vector a indicates the magnitude of the profile at each activation time. To obtain reconstruction z, we apply the transposed convolution operation with cyclic shifts, see (2.13). Formally,

$$z^{\text{long}} = \sum_t \max\{a_t, 0\} C^{t-1}(V^{\text{long}}). \tag{2.8}$$

Note that $V^{\text{long}} \in \mathbb{R}^n$ is the (in this case) long-term profile (i.e., temporal pattern) of interest. Then, $z \in \mathbb{R}^n$ is the contribution of this profile to the final reconstruction of x, see (2.4). Intuitively, activation vector a is 0 for all but one entry that corresponds to the non-negative magnitude and the cyclic shift.

We use a similar approach to learn short-term profile V^{short}. Now, we feed the lower level with residuals $e^{\text{long}} = x - z^{\text{long}}$ interpolated, so that it is a vector in $\mathbb{R}^{|\mathcal{T}|}$. We discuss the changes compared to the upper-level architecture. For the lower level we use (2.13) rather than (2.8), since we consider a conventional shift rather than a cyclic shift. Second, we allow short-term profiles to be stretched in time. Therefore, we add separate filters to choose among 4 different stretch operations (see (2.5)). One stretch operation is the identity mapping (i.e., no stretching). We already showed examples of learned stretched profiles in Figure 2.1 (lower figure). In addition, we allow short-term profiles to be activated before

the day starts, in order to capture short-term variations at night.

We make some comments regarding the case with multiple profiles. The upper-level architecture is naturally extended; there is however explicit interaction among short-term profiles. Based on Bascol et al. (2016), to assure that short-term profiles independently capture systematic variations, an activation of a single short-term profile implies no activation of any other profile 2.5 hours before and after.

2.4.4 Initialization and learning process

The procedure to find the temporal patterns and the noise-level parameters (Algorithm 1) has outer and inner iterations. In the outer iterations, the statistical parameters θ are learned. Therefore, we initialize our procedure with an estimate θ^i, learn the profiles for each measurement location until the termination criteria are met (Section 2.4.2), and then obtain a new estimate θ^{i+1} of θ, by relating the variance in the residuals with the mean flow (i.e., solving (Q), see Section 2.3.2). This procedure is repeated, until θ^i, $i = 0, 1, \ldots$, converges. During the inner iterations, we use the neural network to solve (P_{ML}), using estimate θ^i, and thereby extract the profiles for each measurement location.

Our neural network procedure of the previous section basically re-trains the neural network after we added a new profile (see Algorithm 1). However, what is learned by the neural network does not need to be learned again. Therefore, we feed the neural network with information obtained during the process of previous profile(s). That is, each time we learn a new profile, we provide an initial estimate of the shape of the profile to the neural network. This estimate is for the first long-term profile the mean flow over all days. For the other long-term profiles it is the standard deviation of the residuals since a large standard deviation is a possible indicator for a poor fit. The estimate for the shape of a short-term profile is the 7.5h time series of residuals with maximum error with respect to $|e_{d,t}| \, / \sigma(s_{d,t})$.

Note that in our above-mentioned setting, there might be multiple long and short-term profiles active at a day. To speed up the learning process, we do not allow multiple activations of the same short-term profile for a single day. During the final reconstruction (i.e., after training for a fixed number of profiles), we allow multiple activations of the same short-term profile, as long as the magnitude is at least 10% of the maximum magnitude among all days (for the same profile) as in Bascol et al. (2016).

We implemented our neural network in Tensorflow, and trained the network using the stochastic gradient descent-method ADAM (Kingma & Ba, 2015) (parameters: $\eta = 1e^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$), with 2400 iterations for each long-term profile, and 600 iterations for each added short-term profile. We use a batch size of 25. We note that parameters after adding a new profile can still be updated. This does not apply to the shape of the long-term profile while learning short-term profiles, since it might highly impact the number of short-term profiles to be learned (e.g., when short-term profiles cover long-term variations). Notice that our objective is to infer the systematic variation from the measurements. Therefore, we do not have a separate training and test set. The procedure in Algorithm 1 is so that an overfit is prevented (i.e., our iterative procedure finds the minimum number of profiles to explain the variations).

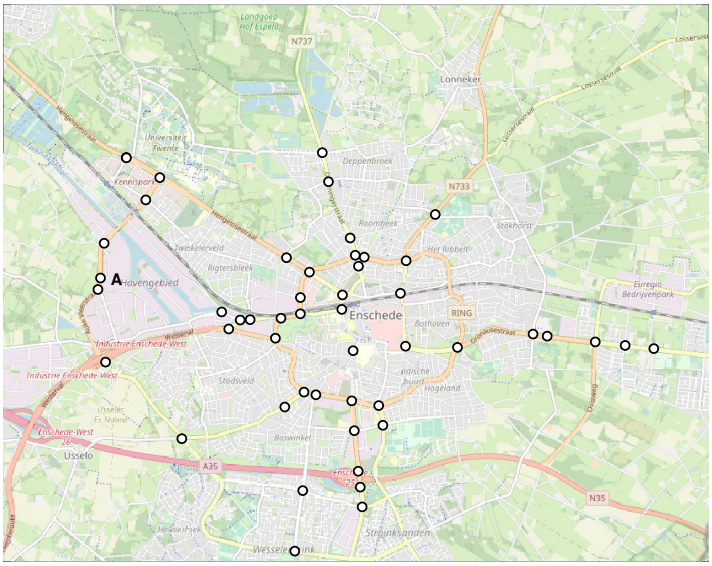


Figure 2.4: City of Enschede, the Netherlands. Dots indicate the signalized intersections with (multiple) measurement locations (source map: OpenStreetMap, 2019).

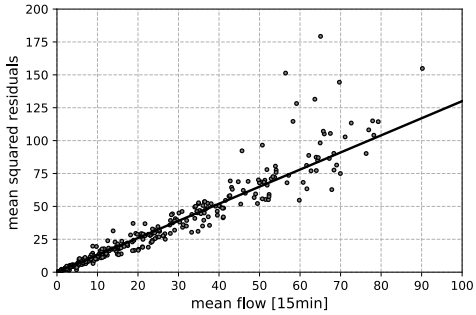
2.5 Results

We apply our procedure to estimate systematic variations and statistical properties of the noise in traffic volumes measured by loop detectors in the city of Enschede.

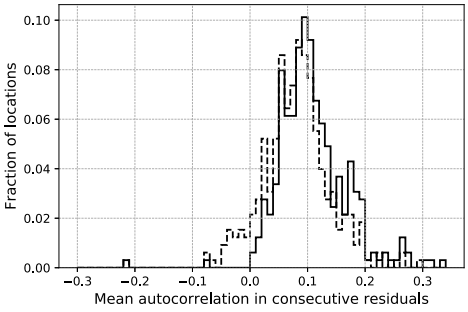
2.5.1 Data

We study traffic volumes in the city of Enschede (+/- 160.000 inhabitants). Data were collected at 49 signalized intersections (see Figure 2.4), from January 2016 until December 2017. Vehicles were detected at each approaching lane by inductive loop detectors a few meters from the stop line, and were aggregated to measurements with a 15min interval. At the junctions under consideration different modes of transport interact, including, depending on the intersections, buses (with possible priority), cyclists, and pedestrians. The intersections are located at both major and minor urban roads and include intersections near freeway off-ramps and on-ramps.

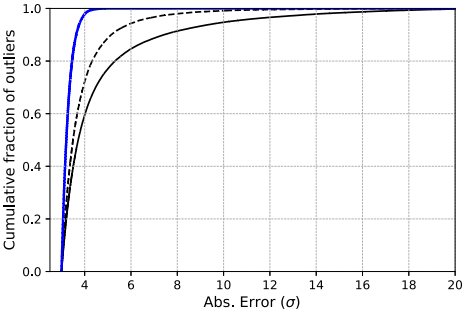
We inspected the data, and rejected volume measurements based on straightforward criteria. We rejected a complete day of loop measurements if either the time series has missing data, or contains a consecutive 6 hours of measurements with no counts between 5AM and midnight. The measurements for lanes that share directions (i.e., origin and destination are locally shared) were summed. 326 loops are in our data set after disregarding loops with less than 365 (of the possible 731) complete 24h time series.



(a) Noise level estimation using a fit (line) with the mean flow (x-axis) and mean squared of the residuals (y-axis). Each dot corresponds to a single measurement location.



(b) Distribution of mean autocorrelation in residuals with lag 1 over the measurement locations. The solid line corresponds to the reconstruction with only long-term profiles, the dashed line corresponds to the final reconstruction.



(c) Distribution of outliers for locations with at least one short-term profile (black lines), and the cumulative distribution function for outliers of random variation (blue line). The solid line corresponds to the reconstruction with only long-term profiles, the dashed line corresponds to the final reconstruction.

Figure 2.5: Results of noise level estimation in the Enschede traffic network.

2.5.2 Noise level estimation

We applied Algorithm 1 and reconstructed, for each location under consideration, the systematic volumes in the historical data by means of long and short-term profiles. We initialized the procedure with noise parameters θ^0 so that $\varepsilon_{d,t}$ is assumed to be distributed according to $\mathcal{N}(0, s_{d,t})$ for each (d,t)-combination (i.e., Poisson-like noise). After we solved (P_{ML}) using the neural network (Section 2.4.3), we estimated the noise parameters θ for the next (outer) iteration as follows.

Per location, we estimate the variance in the residuals using the mean square of the residuals, i.e.,

$$\frac{1}{|D|\,|T|}\sum_{d,t} e_{d,t}^2, \tag{2.9}$$

and relate it with the estimated mean of the reconstructed systematic flows. We applied robust regression for $\gamma \in \{0, 0.5, 1\}$ to estimate the parameters of (2.2). We used robust regression to prevent overestimates of the conditional variance, since time series can still include outliers that have no recurrent character but dominate the estimate (2.9). After re-estimation of θ, we repeated the procedure.

The algorithm terminated after the third iteration. Figure 2.5a shows the estimate of the noise quantity. The random variation in this network is roughly distributed according to a heteroscedastic Gaussian distribution:

$$\varepsilon \sim \mathcal{N}(0, 1.3s), \tag{2.10}$$

i.e., $\alpha = 1.3, \beta = 0, \gamma = 0.5$. A noise quantity of $\varepsilon \sim \mathcal{N}(0, s)$ was found when we applied our framework with an initial estimate of white Gaussian noise $\mathcal{N}(0, 1)$, which will eventually lead to the same random variation as we found. Estimate (2.10) indicates that variance in the noise is proportional to the volume, and there is typically no additive noise. Thomas et al. (2008) found a lower bound $\varepsilon_{d,t} \sim \mathcal{N}(0, 1s_{d,t})$ in a comparable urban setting.

The estimate (2.10) is based on the assumption that there is no correlation among successive residuals, i.e., all systematic variation is captured. Figure 2.5b shows the distribution (over the measurement locations) of the mean remaining autocorrelation (with lag 1). This distribution indicates that on average there is still some serial correlation left. However, for each location, the addition of another profile does not substantially improve reconstruction and one can conclude that a majority but not all systematic variations have a recurrent character.

Regarding the short-term profiles, Figure 2.5c shows that the distribution of outliers (3σ) is different from $\mathcal{N}(0, 1.3s)$. Further research should investigate whether these outliers can be explained by spatial and/or spatio-temporal correlations. On the other hand, the noise distribution may also have a different shape (e.g., with a heavy tail) (Buckley, 1967), or depend on other factors such as the weather conditions. In any case, a large share of the systematic variations occur on timescales longer than 15min.

In Figure 2.6, we show the number of extracted long and short-term profiles per location in the network. Interestingly, traffic volume time series show much more regularity than can be expected from the volumes as such. That is, not only the typical day-of-the-week dependent 24h pattern can be expressed by only a few underlying recurrent profiles, also the day-to-day variations in the daily traffic pattern can be captured by these profiles. Most locations under consideration require only two or three 24h profiles and at most one

short-term profile, adapted over various timescales, to express a majority of the systematic variations. For prediction purposes, however, it can be necessary to explain the time of occurrence, the magnitude and the shape of these profiles. This may require significant efforts, since additional variables may need to be collected.

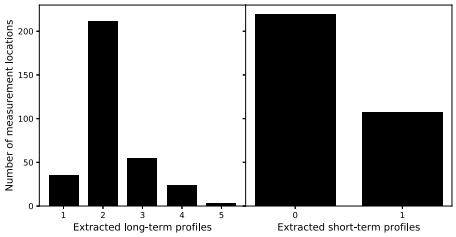


Figure 2.6: Number of extracted profiles over the network.

2.6 Conclusion

Patterns in urban traffic flow time series can be used for accurate predictions of urban traffic flows. At the same time, a portion of the variations in measurements is unpredictable. In this chapter, we analyzed urban traffic dynamics by means of systematic (predictable) and random (unpredictable) variations in 15min traffic flow time series. Many of the systematic variations are recurrent, and can therefore be incorporated in a prediction method. Capturing these systematic variations, however, is a challenging task since they may occur on various timescales, they are not revealed by the noisy measurements, and the patterns may show small yet natural adaptations over time. In this chapter, we proposed a framework to reconstruct systematic volumes from historical data by means of recurrent long- and short-term patterns. In fact, we use a neural network to extract profiles and show that for many locations the systematic variation can be captured using a few temporal profiles. This description of the volume variations allows for a lower-dimensional estimation and prediction method, since a profile provides the predictable fluctuations over multiple time horizons.

Random variation makes that a part of the flow measurements is unpredictable. In this chapter, we estimate the statistical parameters of the volume-dependent noise. We found that on a network-wide level the random variation has roughly a conditional variance of 1.3 times the underlying systematic volume. This result can directly be applied for incident and outlier detection, probabilistic density forecasts (see Chapter 4) or measurement interval choice (Smith & Ulmer, 2003).

Our approach is generic in the sense that it can be applied to any (traffic) time series. In a future study, we intend to expand our architecture to study traffic speeds and extract urban network-wide, thus spatio-temporal, profiles. Such long and short-term profiles express network processes under different conditions. In particular, short-term profiles may reveal the impact of disruptions over the network, which can be used to efficiently predict traffic speeds on a network-wide scale.

2.7 Appendix: Convolutional autoencoder

We briefly discuss the concept of a simple autoencoder that aims to reproduce the input by means of a parse (hidden) representation (*cf.* Goodfellow et al., 2016). Consider input vector $x \in \mathbb{R}^n$, the hidden state $a^l \in \mathbb{R}$ with respect to filter $l = 1, 2, \ldots, N_l$, (with N_l the number of filters) that corresponds to x is $a^l = a^l(x)$ with

$$a^l(x) = g(W^l x + b^l). \tag{2.11}$$

Here, $W^l \in \mathbb{R}^n$, and $b^l \in \mathbb{R}$ are the weights and bias, respectively, of filter l. $g(x)$ is a function that scales a (Goodfellow et al., 2016). To construct output $z = F(y) \in \mathbb{R}^n$, an autoencoder uses *profiles* $V^l \in \mathbb{R}^n$ so that

$$z^l = a^l V^l, \quad \text{and} \quad z = \sum_{l=1}^{N_l} z^l. \tag{2.12}$$

The neural network estimates the parameters $y = (V, W, b)$ using a training set of x.

In a convolutional autoencoder (2.11) is replaced by convolution operation

$$a_t^l = g\left(\sum_{s=1}^{m} W_s^l x_{t+s-1}\right), \quad t = 1, 2, \ldots, n - m + 1$$

with $W^l \in \mathbb{R}^m$, and $m \leq n$, in which we assumed $b^l = 0$. A *transposed convolution* (Zeiler & Fergus, 2014) operation decodes the hidden state a to the input space. Then, the first equation of (2.12) becomes

$$z_t^l = \sum_{s=1}^{m} V_s^l a_{t-s+1}^l, \quad t = 1, 2, \ldots, n, \tag{2.13}$$

with $V^l \in \mathbb{R}^m$. Remark that not all these equations are well-defined, and therefore *padding* (Goodfellow et al., 2016) is applied.

Chapter 3

A statistical characterization of arrival processes at urban signalized intersections

3.1 Introduction

Delays at signalized intersections determine the travel times in urban traffic networks to a high degree. Numerous studies (e.g., Akcelik, 1980; Boon & Van Leeuwaarden, 2018; Viti & Van Zuylen, 2010a) have therefore been conducted to model the operations that occur at such junctions. These models are consequently used for strategic, tactical, and operational decision making by, for example, road authorities to optimize the level of service, both on a local level as well as on a network-wide scale.

A share of the fluctuations in delays is recurrent and is related to the patterns in demand over time (Viti, 2006). Many of the models are therefore concerned with the demand-dependent average level of service (see Cheng et al. (2016) for an overview). Even though the average performance for various demand scenarios is relevant on a strategic level, it is increasingly recognized that also the distribution of the delays is important (Chen et al., 2017; Fu & Hellinga, 2000; Zheng & Van Zuylen, 2010). Whereas decision makers are particularly concerned with the reliability of their services, delay fluctuations, e.g., faced by logistics service providers, may cause that under comparable conditions an intended time of arrival is exceeded or not. Indeed, accumulated local and short-term variations significantly impact operations, and thereby impose serious costs since substantial slack must be introduced in route plans. Despite the fact that many of the longer-term variations in the level of service are periodic and can therefore be anticipated well-before departure, road users suffer from variations that occur on shorter timescales. In an urban network, these uncertainties mainly originate from the dynamics near (signalized) intersections.

Where delays and travel time information are of interest for road users, volume information improve predictions regarding these variables, e.g., in near-saturated conditions when

This chapter is based on the following paper: Eikenbroek, O.A.L., Thomas, T., Mes, M.R.K., & van Berkum, E.C. Statistical characterization of arrivals at urban signalized intersections.

the onset of congestion needs to be predicted. In fact, traffic management decision making, at least in the Netherlands, is typically based on stop-line departure volume measurements (counts). With future travel conditions being inherently variable, even if one accounts for the predictable variations in demand and supply (Viti, 2006; Zheng et al., 2017), these unpredictable fluctuations impose an inevitable uncertainty to decision makers. As such, point predictions for traffic conditions-related variables are not sufficient when evaluating route plans and designing robust management measures. Anticipatory decision-making processes allow stochastic and robust optimization techniques to be applied, but at the same time requires that the underlying variability-inducing processes are accurately captured in a model.

Anticipatory decision making requires an understanding and quantification of the uncertainties that occur. In the lower urban network, travel time uncertainty is largely caused by interrupting processes at signalized intersections. To account for the variability in travel times, delays, and volumes, several studies focused on describing such fluctuations on an aggregated scale (Chen et al., 2017; Luo et al., 2019). Anticipating uncertainties, however, requires not only that the resultant variability is described but additionally asks for an understanding of the contributing factors (Zheng et al., 2017). This is particularly relevant in the context of anticipatory decision making in which predictions have a feedback loop with control measures and therefore require a model to incorporate the dynamics at signalized intersections due to the emergent behavior of travelers.

A large share of the fluctuations in delays at signalized intersections can be traced back to the arrivals of the vehicles at the different approaches, and the inter-dependencies with the signal timings particularly in case of vehicle-actuated signals. Although the importance of these dynamics for decision makers operating in the urban traffic domain, and the wide availability of models and simulation tools to mirror the interactions, these tools are rarely validated based on empirical data. Fortunately, the increasing availability of a variety of measurement devices and data-processing tools allows us to assess the empirical consistency of, and possibly improve, existing models.

In this chapter, we use empirical data to study arrival processes at signalized intersections on different spatial and temporal scales, and quantify how these processes contribute to the shape of the delay and volume distributions under a range of conditions. In fact, we provide a statistical description of arrival processes and include the inter-dependencies with (upstream) signals, and the dynamics over time and space. Therefore, arrival patterns are studied on various scales simultaneously to not only account for the change in the demand between days and within a day for different parts of the network, but also to incorporate the dynamics on a much smaller level, e.g., due to traffic signal cycles upstream. We illustrate that failing to accurately capture the structure of arrivals in a simulation setting might underestimate the variations in volumes and overestimate delays and thereby have serious implications particularly for tactical and operational decisions, e.g., when designing the cycle settings.

We collected millions of arrival events in an urban network, and study real-world arrival processes using the data collected at various intersections on different parts of an urban network. We introduce a comprehensive statistical framework to examine the structure of the arrival processes accounting for both correlations over time as well as for the variations introduced by upstream interruptions, thereby incorporating short-term periodicities related to upstream signals, the formation of platoons, and the changing structure as traffic proceeds. By studying the arrivals on such a detailed resolution, we are able to assess the information loss by using counts on aggregated scales as is typical for many road authorities. In fact, the

local fluctuations introduce noise on an aggregated scale, making it difficult to anticipate or respond quickly to changing situations. However, very short-term variations and predictions are highly relevant so that preemptive control or coordination actions can be deployed to mitigate or optimize network-wide performance (Li, Yang, et al., 2022; Vlahogianni et al., 2004), e.g., by recognizing changes in the arrival process or for using coordinated traffic signal systems (Robertson & Bretherton, 1991).

The remainder of this chapter is organized as follows. In Section 3.2, we formally introduce the problem and explore the relevance of studying arrival dynamics as both a counting process as well as a sequence of inter-arrival times [1]. Section 3.3 discusses the data collection and filtering process. In Section 3.4, we introduce our statistical framework which we use consequently in Section 3.5 and Section 3.6 to assess the structure in the inter-arrival times and the counting process, respectively. Based on this characterization, we discuss the impact on the variations in delays in Section 3.7 and draw conclusions in Section 3.8.

3.2 Problem formulation

A substantial share of the variability in delays and volumes at signalized intersections can indirectly be traced back to the arrivals of vehicles (or cyclists, pedestrians) at the approaches. Arrivals are typically modeled using a stochastic process, i.e., the exact inter-arrival times cannot be predicted. The changes in arrivals occur on different scales, typically studied on scales exceeding 5-10 minutes, but fluctuations also occur in the time gaps in the order of tenths of seconds (Banks, 1999; Breiman & Lawrence, 1973).

Variations that occur in the order of several minutes indicate the varying conditions or regimes in network usage, oftentimes related to the demand that is known to show systematic variability in time and space (Crawford et al., 2017). Changes in the arrival rate regime separate, for example, the rush hour from a quiet period. Considering a fixed time of day, the demand also shows systematic variation over space with some parts of the network heavily utilized while other parts perform well below capacity. The variability in the regime for a fixed point in the network can be captured using a time-varying mean arrival or demand rate, ideally considering both the day-to-day as well as the within-day variations in the demand. This rate is slowly changing compared to fluctuations in the inter-arrival times of the individual vehicles under regular conditions. Typically, a location-dependent 24h pattern is used to express the non-stationarity in the demand - thereby incorporating that 'normal' or random fluctuations naturally exist even under stationary demand (Breiman et al., 1977; Breiman & Lawrence, 1973; Sparks, 1976). Very short-term fluctuations, in the order of tenths of seconds, roughly describe the non-predictable (random) fluctuations in the arrival events (Dion et al., 2004; Gwiggner & Nagaoka, 2014). Considering a fixed location in an urban setting, short-term fluctuations show *bursts*: short periods of many arrivals alternate with longer periods in which there are no arrivals (Goh & Barabási, 2008; Vázquez et al., 2006). Indeed, arrivals typically occur in clusters or platoons, separated by relatively long periods without any arrival.

The two timescales are fundamentally related, perfectly illustrated by renewal theory (e.g., Karlin & Taylor, 2012), and the point process describing the random occurrence of events over time is a building block of the aggregated time series of volumes (Brillinger,

[1] Throughout this chapter, we use inter-arrival times interchangeably with (time) headways

2008; Cox & Lewis, 1966). Even though there has been considerable attention for describing the demand over time (Crawford et al., 2017; Weijermars, 2007), empirically, very little is known about the stochastic fluctuations in the arrivals in the complicated context of urban traffic - and typically some form of aggregation is applied (Guo et al., 2007; Li, Yang, et al., 2022; Oh et al., 2005; Vlahogianni & Karlaftis, 2011). At the same time, it is widely recognized that the arrival pattern impacts delays at signalized intersections (Akcelik, 1980; Chen et al., 2017; Olszewski, 1990; Transportation Research Board, 2000; Zheng & Van Zuylen, 2010). For a given demand rate, the impact of the arrival pattern on the delay can be considerable (Van Leeuwaarden, 2006), although its effect is expected to be less compared to the time-changing demand (Olszewski, 1990).

In this chapter, we provide a statistical characterization of arrivals in urban networks based on a large set of recorded arrivals. Compared to the mathematically-tractable processes often assumed in models and simulations settings (Luo et al., 2019; Mohajerpoor et al., 2019; Zheng et al., 2017), real-world arrival processes show a different structure which is only revealed when studied on different scales. We use data collected throughout an urban network to provide evidence that it is important to account for the actual arrival process for estimates regarding the variability in aggregated volumes and delays.

3.2.1 Exploratory analysis

In this subsection, we provide exploratory evidence regarding the characteristics of real-world arrival processes.

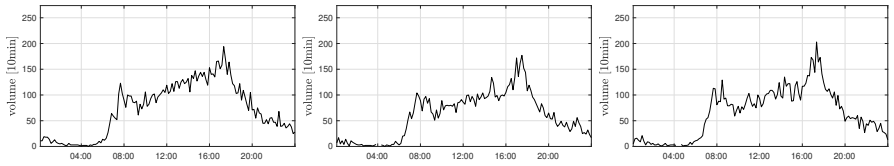


Figure 3.1: Examples of measured 10min volume time series at an approach of an intersection.

The non-stationarity in the demand under regular conditions is mainly related to the variability in within-day and day-to-day travel behavior. Figure 3.1 provides examples of 24h volume time series collected at an approach of an intersection, measured using 10min increments (see Section 3.3). The M-shaped pattern provides a first indication of the within-day variability, and, as illustrated, day-to-day variations are oftentimes considered to be less compared to the within-day variations. Hence, a typical daily pattern, deterministic in nature, can be constructed where stochastic fluctuations occur around the time-varying mean demand rate (Sparks, 1976). Even if the day-to-day variability is explicitly addressed, days can either be divided into (mutually exclusive) sub-intervals with each interval assumed to have a constant arrival rate, or there is a continuously-varying latent process describing the time-varying rate. Using the first approach, the arrival process is then said to be piecewise stationary or piecewise constant (Law et al., 2007; Paxson & Floyd, 1995) - and shows sudden shocks or 'piecewise discontinuities' over time when implemented on volume time series such as the ones in Figure 3.1 (Shone et al., 2021). Determining intervals with a constant arrival rate is not trivial since also systematic variations may occur on

timescales shorter than the interval. In addition, short (e.g., events) and longer-term (e.g., seasonal) variations exist, which makes that spatio-temporal changes in the demand are natural. Continuously-changing demand patterns have only recently been introduced in a traffic context (e.g., Guardiola et al., 2014), but a non-homogeneous Poisson process where a stochastic process describes the time-varying arrival rate has found its way in literature in different contexts and was shown to capture real-world dynamics accurately (e.g., Kim & Whitt, 2014). The majority of the studies in our context (e.g., Breiman & Lawrence, 1973; Webster, 1958), however, assume that a day can be divided into different subintervals with fixed demand, and model the interactions between arrivals and intersection performance for a given interval. A mathematically-appealing arrival process for an interval is the (fixed-rate or homogeneous) Poisson process, in which inter-arrival times of vehicles are assumed to be independently distributed and follow an exponential distribution. Such a process is not only tractable, but also shows agreement with stochastic 5-15min volume fluctuations (Thomas et al., 2008).

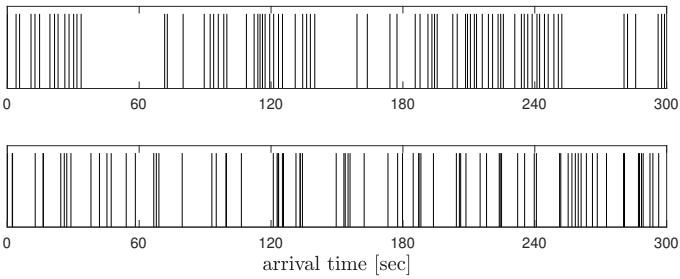


Figure 3.2: Comparison of a sample empirical inter-arrival times (top row) and inter-arrival times from a homogeneous Poisson process (bottom row).

Although arrival counts resemble a Poisson process on a resolution in the order of 5-15min, the empirical arrival times show a very different structure. Figure 3.2 shows a sample of recorded inter-arrival times (top row, measured with 0.1s increments) and a sample of exponentially-distributed inter-arrival times (with the same arrival rate). For a 5min interval, we plotted a vertical line for each arrival, under more or less stationary conditions, assuming in both cases that the first arrival occurs at $t = 0$. When comparing both figures, the exponential distribution allows very small inter-arrival times that are physically impossible due to a minimum spacing between two vehicles. Second, the top row indicates that the true arrivals are more clustered compared to a homogeneous Poisson process. Third, visually, the relatively long gap times seem to show more regularity compared to the bottom figure, e.g., the top row hints on a periodicity in the absence of arrivals and approximately every 100 seconds there is a longer period in which no arrival occurs. To further illustrate this, Figure 3.3 provides a real-world sample for the number of arrivals per 15s for a 30min interval, and indicates a higher degree of regularity in arrival events over time than can be expected from a homogeneous Poisson process. Moreover, the regularity in the arrivals is already 'lost' when considering 2min increments (blue line in Figure 3.3).

Based on this exploratory analysis, we cannot reasonably expect to model arrivals using a simple (non-)homogeneous Poisson process or by solely using a headway density function (see Section 3.2.2). In fact, a stochastic arrival model that aims to realistically mimic real-world urban arrivals should, at least, consider the following aspects:

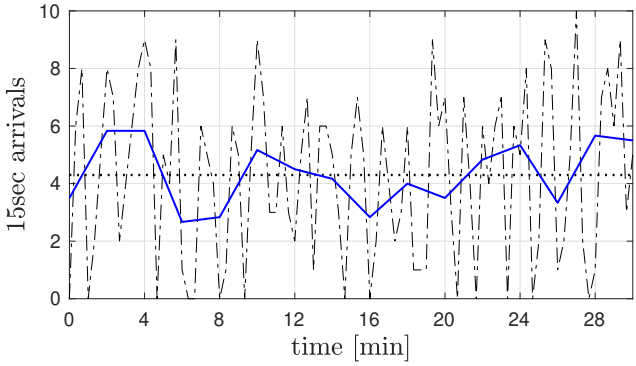


Figure 3.3: 30min sample of arrival counts, measured with 15sec increments (dashed line). The blue line provides the counts on a 2min scale while the dotted line indicates the 30min average.

- the non-stationarity in the demand over time and space due to the location-specific arrival rate, including its intra-day and day-to-day variability;
- the marginal distribution of inter-arrival times, where the left-hand side should reflect the minimum physical and the desired time gap between vehicles, while the right-hand tail of the marginal distribution should reflect the burst structure related to both non-interacting vehicles as well upstream disturbances such as traffic lights;
- the periodicities in the arrival events, indicating that stationary demand and a headway distribution are not sufficient to capture the regular patterns in the arrivals, e.g., due to platoon forming and upstream traffic light influences. Indeed, trends and periodicites might be particularly observed as a function of time rather than as a trend in the inter-arrival times only (Lewis, 1970).

We refer to Goh and Barabási (2008), Kim and Whitt (2014), Lancia and Lulli (2020), and Paxson and Floyd (1995) for similar observations in different contexts. In the following subsection, we discuss earlier approaches to capture the above-mentioned characteristics.

3.2.2 Literature review

We discuss studies from literature on modeling arrival processes regarding the non-stationarity in the demand, marginal distribution of inter-arrival times, and the periodicities in arrival events.

Considering the arrival rate (and possibly, departure rate) as a piecewise constant function of time and space makes that even the early models (e.g., McNeil, 1968; Miller, 1963; Webster, 1958) can implicitly account for the changing demand scenarios. This is a possible reason why many of these models are still used in design manuals such as in CROW (2006).

Frequent studies (e.g., Fu & Hellinga, 2000; Webster, 1958)) assume a homogeneous Poisson arrival process where inter-arrival times are *independent and identically distributed* (iid) and follow an exponential distribution. That the distribution of inter-arrival times is not exponential was already recognized decades ago (Branston, 1976; Buckley, 1968;

Wasielewski, 1974, 1979). Nonetheless, the (semi-)parametric estimation of the distribution is typically limited to headways measured at freeways (see, Ha et al., 2012; Hoogendoorn, 2005; Luttinen, 1996). In any case, these studies point out that one should distinguish *leaders* and *followers* when studying inter-arrival times. Leaders have, by definition, no interaction with the preceding vehicle and typically have a relatively long inter-arrival time compared to its predecessor - reflected in the right tail of the distribution (see, e.g., Vogel, 2002). Followers interact with the preceding vehicle leading to a relatively short inter-arrival time reflected in the left part of the distribution. Where the (right) tail of the distribution is due to the lack of interaction between two vehicles, in interrupted networks it is highly influenced by the arrivals, clearance time and red times upstream. Traffic signal control in our setting is vehicle-actuated but becomes more or less static when saturation levels are approached, meaning that, in general, the tail of the distribution statistically reflects a combination of the variable red times and inter-arrival times upstream. The speed patterns of individual vehicles and traffic lights upstream lead to a different probability of medium and high inter-arrival times compared to an exponentially-tailed inter-arrival distribution. When the arrivals are measured close to but downstream from another intersection, the inter-arrival times are expected to be similar to the inter-departure times, reflected in the left part of the headway distribution (Li & Chen, 2017).

Clustered arrivals occur at an approach mainly due to interacting vehicles and upstream interruptions, which makes that inter-arrival times are likely to be correlated (Boon & Van Leeuwaarden, 2018), in contrast to the iid assumption often assumed in literature (Viti & Van Zuylen, 2010b; Zheng & Van Zuylen, 2010) and in manuals (CROW, 2006). Furthermore, this means that the variance-to-mean or standard deviation-to-mean ratio in the headways is not sufficient to characterize arrival processes in delay models (Hutchinson, 1972; McNeil, 1968; Miller, 1963; Olszewski, 1990; Webster, 1958), and that formulating a mathematically-tractable stochastic arrival process that mirrors the true structure is difficult since standard renewal theory assumptions do not hold (Wang et al., 2018). For departures from the Poisson process assumption in traffic queuing models we refer to, e.g., Boon and Van Leeuwaarden (2018), Chen et al. (2016), Li (2017), Van As (1991), Wang et al. (2018), and Yang and Shi (2018) - typically considering the variance-to-mean ratio for discrete time increments. The arrival process is however continuous in nature, and the structure in the urban traffic arrivals - particularly its periodicities - can potentially be revealed in the frequency domain. Yet, we are only aware of the empirical evidence provided by Miller in Miller (1970) and in the discussion following the paper of Bartlett (1963). In the latter, a point-process periodogram shows a dominating frequency corresponding to the cycle periodicity at the upstream signal. Such periodicites were shown to exist in high-resolution but aggregated volume measurements (Touhbi et al., 2018).

In different contexts, it was shown that the structure of arrival processes is only revealed when studying the arrival patterns over multiple timescales (Crovella & Bestavros, 1997; Gwiggner & Nagaoka, 2014; Paxson & Floyd, 1995), and that bursts and regularity in the events highly influence the second-order characteristics of the delays and counts. These phenomena are, to the best our knowledge, not yet studied in a comprehensive manner for the urban traffic network. In this chapter, we use high-resolution loop detector data collected throughout an urban traffic network to characterize arrivals on different temporal and aggregation scales - accounting for the travel distance relative to other interruptions. In fact, we explore volume and location-(in)variant properties of the bursts in arrivals. Furthermore, we use a simulation approach to explore the implications of using real-world rather than naive

arrival processes in the light of delay variations.

3.3 Data

To empirically study arrival patterns, we use high-resolution event data collected at approaches of signalized intersections. In this section, we discuss the data collection process in Section 3.3.1. In Section 3.3.2, we discuss our data filtering approach. Based on 10min volume counts, we identify volume noise properties (Section 3.3.3), that we consequently use to identify 10min volume regimes in Section 3.3.4.

3.3.1 Data collection

We study arrival patterns in the city of Enschede, the Netherlands (+/- 160,000 inhabitants). Data were collected at 14 signalized intersections (see Figure 3.4), from August 2019 until March 2020 for 11 intersections, and from April 2020 until December 2020 for three intersections. The intersections are located throughout the network, include off-ramps, and the corresponding operations are thus location-dependent and complex in the sense that different modes of transport might interact, including cyclists, buses (possibly, with priority), and pedestrians.

Figure 3.4: Map of Enschede (OpenStreetMap, 2021), where dots indicate the signalized intersections in Enschede where data are collected.

Intersections are equipped with a set of induction loop detectors at each arm to detect approaching or waiting vehicles. Figure 3.5 shows a sketch of a possible configuration for a single signal group. Here, each lane is equipped with a *stop loop detector*, typically 1 to 2 meters from the stop line, to request green. A *long loop*, 10 to 15m from the stop line, detects queues. An *upstream* or *advance* loop detector, often 1m long, is installed 60 to 100m from the stop line to detect approaching vehicles. In our case, we focus on arms with a single-lane approach in which there is an, often additional, upstream loop detector that is able to

detect spillback or upstream blocking. These distant loops are located up to 300m from the stop line. Such distant upstream loop detectors are better able to measure arrivals compared to other loops. We note that Figure 3.5 is only an example of a possible configuration with here left, straight and right-only lanes, although at other arms and intersections, direction of travel from a lane can be shared.

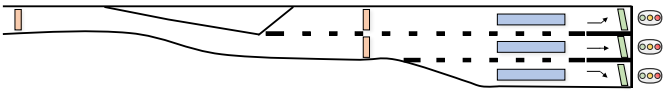


Figure 3.5: Sketch of a typical loop detector configuration, with the stop loop detectors in green, long loops in blue, and upstream loop detectors in pink.

In total, we have 24 different arms under consideration, to which we refer as *measurement locations*. Two of these measurement locations are at off-ramps, while the other are at urban road segments. To support our results throughout the chapter, we often provide the type of road segment as well as the distance to the midpoint of the closest major upstream intersection (ranging from 20m to 2km). Note that the arrivals at measurement locations with a very short distance to an upstream intersection can be assumed to mainly reflect a combination of the various discharge processes at the corresponding signals. However, in this case, these discharge processes characterize the arrival process at the next intersection. In any case, the variety of locations in combination with the fact that volumes are collected over long periods allow us to study arrival processes under different conditions.

3.3.2 Data filtering

We use high-resolution event data from the induction loop detectors. For loop detectors, these events are the time instants that the loop status is switched 'on' (i.e., becomes occupied) or 'off'. The data are collected in a continuous fashion at 0.1s increments for each individual loop detector, typically stored in batches with the events that occurred during a 5min interval. Errors that occurred during data collection and processing makes that for each intersection full days and individual batches might be missing from the data set. Further, we removed public holidays and parts of days with recurrent events such as football matches from the data set.

We filtered data as follows. First, in order to extract vehicle *detections* from the loop statuses, we assumed that every 'on' status of a loop detector should be followed by an event that the loop status is 'off', and we filtered those that did not follow this assumption or had an unrealistically long inter-event time between them. In the remainder, we define a vehicle detection as a pair of 'on' and 'off' loop statuses. Second, we checked for flow conservation, i.e., a full batch is deemed unreliable if the volumes measured at the stop loop detectors showed significantly more or less counts compared to the distant loops. Despite the fact that it is difficult to assess the quality of this procedure directly, the induction loop detectors generally show reliable performance - and are in practice used for the vehicle-actuated control.

3.3.3 Non-stationary demand

Time of day and day of the week are important predictors for the arrival rate. However, variations occur over different timescales (see Chapter 2) and 24h patterns with a changing shape and height over the days explain almost all non-stationarity of the arrivals (Crawford et al., 2017; Weijermars & Van Berkum, 2005). We need to explicitly account for non-stationarity, since (i) we are particularly interested in the natural variation in arrival patterns, delays and counts under similar conditions, and (ii) the character of the arrivals is heavily related to the underlying demand. Here, we discuss our method to identify similar conditions. Therefore, we infer the systematic variations from the historical data, which we then use to find subintervals where the arrival rate is assumed to be constant.

We now introduce the notation. For a single upstream loop detector, we construct aggregated 10min volumes. In fact, let $x_{d,t} \in \mathbb{Z}_{\geq 0}$ denote the 10min volume count at day $d \in D$ at time $t \in T$. Here, D is the set of days, and T is the time domain with 10min increments for a single day. The 24h volume time series x_d consists of a latent yet deterministic systematic flow s_d and a residual vector e_d so that $x_d = s_d + e_d$. Assuming that residuals $e_{d,t}$ are a realization of an independent random variable $\varepsilon_{d,t}$, we infer an underlying estimate of $s_{d,t}$ based on patterns that exist in the 24h volumes over the days.

We refer to ε as volume noise, and identify the random arrival process as the major source of this random variation. Indeed, variation in inter-arrival times is one the causes that $s_{d,t} + \varepsilon_{d,t}$ is a random variable (Banks, 1999; Thomas et al., 2010). We assume that ε is uncorrelated over time, i.e., the actual (realized) number of arrivals relative to the systematic demand is unpredictable when considering 10min timescales. As mentioned, correlations exist when considering very short timescales (e.g., when vehicles arrive in platoons) and also spatial correlations exist over different timescales, e.g., an increase in volume at the upstream loop detector is a predictor for an increase in volume at the stop loop or downstream locations. In this case, we are both interested in s_d as well as in the noise-level function $\sigma^2(s)$ that describes the volume noise variance as a function of the underlying systematic flow, i.e.,

$$\mathrm{var}(\varepsilon_{d,t}|s_{d,t}) = \sigma^2(s_{d,t}), \quad d \in D, t \in T.$$

In fact, we use s to identify conditions with similar arrival processes (see Section 3.3.4), while the noise-level function gives the volume dispersion resulting from the randomness in arrivals when considering 10min timescales. Based on literature (Chen et al., 2012; Guo & Williams, 2012; Thomas et al., 2008), we assume that the noise variance $\sigma^2(s_{d,t})$ scales linearly with volume $s_{d,t}$, i.e., the (conditional) noise variance is a linear function of the underlying systematic volume. Although different mechanisms exist to estimate s from x (Habtemichael & Cetin, 2016; Lancia & Lulli, 2020; Thomas et al., 2008), in the next section we use an approach that explicitly accounts for volume-dependent noise as further elaborated upon in Chapter 4.

3.3.4 Demand patterns and interval selection

We identify similar traffic conditions based on the 10min systematic variation s as measured at the upstream loop detector. In fact, we partition the raw detection data occurring within batches of 10min increments into mutually exclusive bins so that arrivals are piecewise stationary: within each bin the mean arrival rate is assumed to be constant. This partition is

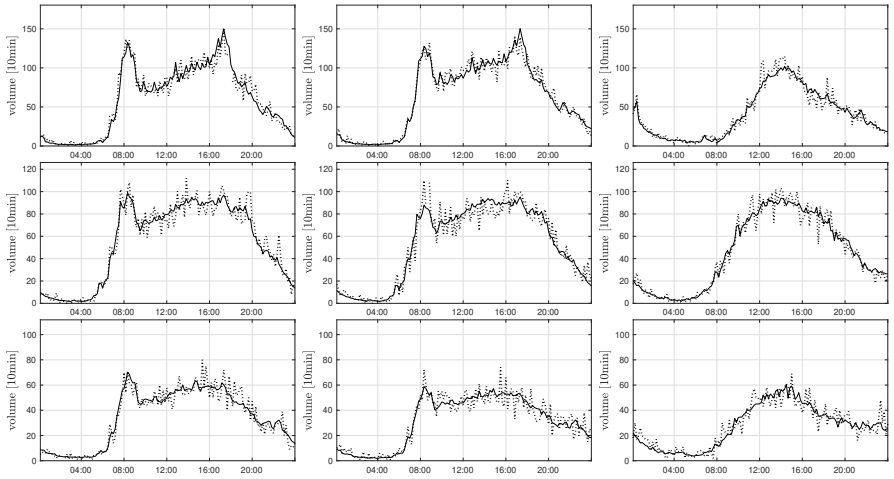


Figure 3.6: 10min aggregated arrival measurement time series (dotted line) and systematic variation estimate (solid line) for three days at three different approaches (upper, middle, and lower row of figures).

based on the 10min systematic volume estimate s, and, in the remainder of this chapter, we study the arrival processes within each bin.

Figure 3.6 shows examples of 10min volume time series (dotted line), including the estimate of the systematic variation (solid line). Here, we clearly observe different volume patterns for the different measurement locations and days, typically M-shaped for weekdays. Visually, we quite accurately cover the patterns in the 24h volume time series within a day and between days - and we conclude that a major share of the variations in the arrival volumes is accounted for. However, changes in the arrivals on a shorter timescale in the range of minutes might still exist, e.g., due to temporary blockages or slowly-moving heavy vehicles. It is difficult to account for such variations explicitly since they occur on timescales shorter than 10min. At the same time, simulation processes might need to explicitly incorporate such conditions as well whereas they naturally arise in a real-world setting. Short-term deviations compared to the pattern occurring on relatively short timescales that take longer than 10min, e.g., additional demand due to events, also occur but such situations are excluded from further consideration by using a sigma-clipping approach. In fact, we remove 10min instances from further consideration if (i) the residual $|e_{d,t}|$ is more than $4\sigma(s_{d,t})$ away from the pattern-based estimate $s_{d,t}$ or (ii) if two consecutive 10min intervals each have a deviation larger than $2\sigma(s_{d,t})$ compared to the estimate. In any case, the estimate of the systematic variation might be such that a majority but not all arrival rate variations are accounted for. However, longer-term variations (on timescales of at least 20min) are captured using our approach.

When the queue length exceeds the distance from the stop line to the used upstream loop, a departure process rather than a series of arrival events are measured by the distant detector (see Figure 3.5). We refer to such a situation as (over)saturation, which should be excluded from our analysis. We therefore remove a full 10min batch of events if either (i) a single long occupancy time is measured at the upstream loop, (ii) the mean occupancy time at this loop suggests that traffic is slowly-moving, or (iii) the 95th percentile in occupancy times

indicates that traffic was in saturated condition for a shorter period. We further remove a full day from further consideration if the serial correlation (at lag 1) of the 10min residuals e_d has an absolute value larger than 0.25. After accounting for these situations, we are left with a large number of reliable recorded arrival times ranging from 35,000 to 1,250,000 events, dependent on the location. Using this filtering approach, our aim is to obtain intervals with uniform conditions (Son et al., 2014; Tan et al., 2016; Vlahogianni & Karlaftis, 2011). We note, however, that we remove some 10min intervals with many arrivals if the queue length exceeded the distance to the upstream loop detector, and therefore underestimate the variation in the actual arrivals. Therefore, we only consider those demand bins in which less than 30% of the days are excluded from analysis. Combined with the relatively long distance from the stop loop to the used loop detector, we can be confident that we characterize actual arrival processes.

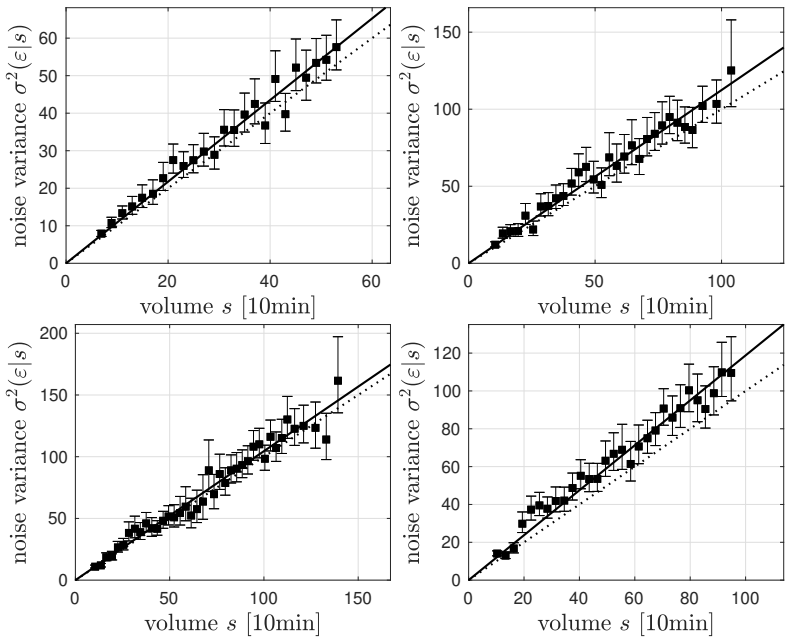


Figure 3.7: Noise-level estimates, with each square indicating the mean-variance estimate for a bin. The solid line gives the least squares estimate, the dashed line shows the estimate in case of Poisson noise.

The unpredictability of the actual number of arrivals is quantified by the noise variance compared the mean volume. We illustrate estimates of 10min volume linear noise-level functions in Figure 3.7, including the 90% confidence interval of the estimates assuming that volumes have a Gaussian distribution on a 10min level. On average, the random volume variation shows slight overdispersion relative to Poisson noise, i.e., $\mathrm{var}(\varepsilon|s) > \mathbb{E}[s]$, and the linear noise-level function gives quite a good fit. In fact, there is little variation in the noise level between the different locations. We underline that by this binning approach, there is systematic flow variance left within each bin. However, this variation is very small compared to the noise variance and therefore we neglect this remaining variation in the underlying arrival rate in the remainder of this chapter. In any case, the absolute noise

variance increases approximately in a linear fashion with growing volumes independent of the location under consideration. Hence, a heteroscedastic Gaussian distribution or even a Poisson distribution are appropriate to capture natural stochastic fluctuations on timescales longer than 10min. Based on this observation, we cannot invalidate the assumption of a renewal arrival process yet since the asymptotic distribution of a renewal process, under minor assumptions, also scales linearly with the demand. In the next sections, we therefore study the arrivals within each volume bin on shorter timescales.

3.4 Framework

In previous section, we identified increments with similar arrival rates on a 10min level. Consider a fixed roadside measurement location with a stationary demand in that all regularities in the arrival process occur on shorter timescales. At this location, we measure arrival events over time (see Figure 3.2), which we consider as a point process with binary events in (continuous) time t, with the probability of more than one event in an interval $\Delta t > 0$ being $o(\Delta t)$, $\Delta t \to 0$. The arrival process can be characterized in two ways. Either by a counting process $N(t)$, measuring the cumulative number of events in an interval $(0, t]$ assumed to start from but not including an arbitrary event (as is typical in traffic engineering), or by the sequence of inter-arrival times $\{X_i\}$ between successive subsequent events. These two aspects of the process are directly related since (see, e.g., Karlin and Taylor (2012))

$$S_n := \sum_{j=1}^{n} X_j > t \quad \text{if and only if} \quad N(t) < n, \quad n = 1, 2, 3, \ldots$$

For all $\Delta t > 0$, we consider the difference process corresponding to $N(t)$ as follows:

$$\Delta N(t) = \frac{N(t + \Delta t) - N(t)}{\Delta t}, \qquad t \geq 0, \Delta t > 0,$$

or the differential process $\{dN(t)\}$ being $\Delta N(t)$ with $\Delta t \to 0$.

Under minor assumptions, the first-order properties of the intervals are directly related to the first-order properties of the counts. Indeed, let $M(t) = \mathbb{E}[N(t)]$ be the expected number of arrivals in an interval, and assume that $\mathbb{E}[dN(t)]/dt$ exists for all t. Then for constant arrival rates we have $\lambda(t) = dM(t)/dt = \frac{1}{\mathbb{E}[X]}$, that is, the mean inter-arrival time can be inferred from the flow rate. The second-order properties, on the other hand, are not equivalent (Daley & Vere-Jones, 2003) but reveal much more and different information about the process. We discuss in the following subsections the second-order properties of both the inter-arrival times as well as the counts.

3.4.1 Statistical characterization of inter-arrival times

The first-order properties do not reveal the structure in the arrival events, e.g., the fluctuations in the inter-arrival times. The set of intervals or inter-arrival times $\{X_i\}$ is a sequence of random variables, assumed to have a common marginal distribution function denoted by $F_X(x)$. A relatively simple and often-used metric to assess the dispersion in the intervals

under different conditions is by means of the ratio

$$C^2(X) = \frac{\mathrm{var}(X)}{\mathbb{E}[X]^2},$$

i.e., the squared *coefficient of variation* $C(X)$. $C^2(X)$ measures departures from the Poisson process driven by exponentially distributed inter-arrival times for which $C^2(X) = 1$, and thereby provides a first indication for invalidating the Poisson process in the setting under consideration. We note that in some cases, however, this coefficient is not meaningful if the tail of the distribution is such that the variance is infinite. Here, however, we do not cover this case as we have no evidence of *long-range dependency* in the process.

The memory structure over the intervals $\{X_i\}$, i.e., the dependency between consecutive inter-arrival times, however, cannot be revealed using the coefficient of variation. For example, it could be that a short inter-arrival time is followed by another short one. Such dependencies can be examined using the serial or autocorrelation sequence

$$\rho_k := \frac{\mathrm{cov}(X_i, X_{i+k})}{\mathrm{var}(X)}, \quad k = \ldots, -1, 0, 1, \ldots,$$

with $\rho_{-k} = \rho_k$ in our case. Alternatively, the Fourier transform can be employed to characterize the autocorrelational properties using

$$\rho_k = \int_{-\pi}^{\pi} g_X(\omega) cos(k\omega) d\omega, \quad k = \ldots, -1, 0, 1, \ldots,$$

where

$$g_X(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \rho_k e^{-ikw},$$

with $i^2 = -1$, is the corresponding power spectral density for discrete time series. These techniques are well known for time series collected at regular (time) increments (Stathopoulos & Karlaftis, 2001; Sun et al., 2018). In our case, however, the increments are the consecutive vehicles, and the time series reflects then the inter-arrival times over the vehicles.

Looking directly at the autocorrelational sequence between consecutive inter-arrival times may not provide one with a clear picture regarding the impact of the memory structure over a multitude of vehicles. When looking at k consecutive stationary inter-arrival times, the corresponding variance function (of k)

$$\begin{aligned} V_k &:= \mathrm{var}(\sum_{j=1}^{k} X_j) \\ &= k\,\mathrm{var}(X) + 2\sum_{j=1}^{k-1}\sum_{l=1}^{j} \mathrm{cov}(X_j, X_{j+l}), \end{aligned} \tag{3.1}$$

reveals the correlation structure in that the variance of a sum of random variables possibly deviates from the sum of variances (Fendick et al., 1989; Gusella, 1991). Using V_k, and allowing direct comparison with a Poisson process by a demand-independent metric, we

define the *dispersion index* (for inter-arrival times X) at lag k as

$$J_k = \frac{V_k}{k\mathbb{E}[X]^2} = C^2(X)(1 + 2\sum_{j=1}^{k-1}(1 - \frac{j}{k})\rho_j).$$

For renewal processes, $\rho_k = 0$ for all $k \neq 0$, hence J_k is a constant independent of k, and has therefore a flat power spectral density (so-called 'white noise' spectrum in the typical time series setting). Even stronger, $J_k = 1$ for a Poisson process. For long-range dependent processes, i.e., with $\sum_{j=1}^{k} \rho_k \to \infty$ as $k \to \infty$, it follows that $J_k \to \infty$. In our setting, as we will see, J_k converges rather quickly. In any case, the dispersion index J_k reveals the persistence of the correlation in consecutive inter-arrival times, even if the individual serial correlation coefficients are weak. Disadvantage, however, is that the true structure of the arrival events in an urban setting is likely to be a function of time. This structure is difficult to reveal using the characterization of this subsection, but can be revealed using the second-order properties of the counts.

3.4.2 Statistical characterization of the counts

We consider the second-order properties of the counts following the conventions in Bartlett (1963), Cox and Lewis (1966), Hawkes (1971), and Lewis (1970). Consider an arrival process with a constant arrival rate over time $\lambda = \lambda(t) = \frac{dM(t)}{dt}$ - also known as the *intensity function*. We consider the covariance properties of the counting process $N(t)$. Therefore, we introduce the *cross-intensity function* $\gamma(t, t+\tau)$ so that

$$\mathbb{E}[dN(t+\tau)dN(t)] = \gamma(t, t+\tau)(dt)^2,$$

or, equivalently,

$$\gamma(t, t+\tau) = \text{Prob}\{\text{event at } (t+\tau, t+\tau+dt] \text{ and event at } (t, t+dt]\},$$

roughly, the probability that an arrival occurs at t and $t+\tau$ (see also Miller, 1970). Here, dt is the differential of t. Under stationarity assumptions, the arrival rate is independent of t, and $\gamma(t, t+\tau)$ is a function of the time difference τ only. Then, the *covariance density* $\mu(\tau)$ of the differential process $\{dN(t)\}$ becomes (Bartlett, 1963)

$$\mu(\tau) = \frac{\mathbb{E}[dN(t+\tau)dN(t)]}{(dt)^2} - \lambda^2, \qquad 0 \neq \tau \in \mathbb{R}.$$

The case $\tau = 0$ needs to be considered as well, and we extend the definition of $\mu(\tau)$ to $\tilde{\mu}(\tau)$ so that $\tilde{\mu}(\tau) = \lambda$ for $\tau = 0$, and $\tilde{\mu}(\tau) = \mu(\tau)$ otherwise. Indeed, $\mathbb{E}[dN(t)] = \text{var}(dN(t)) = \lambda$. The Bartlett power spectrum $g_N(\omega)$ - corresponding to signal $N(t)$ - is the Fourier transform of $\tilde{\mu}(\tau) = \delta(\tau)\lambda + \mu(\tau)$, with $\delta(\tau)$ the Dirac-delta function. Bartlett's spectrum is thus defined as

$$g_N(w) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \lambda\delta(\tau)e^{-iw\tau}d\tau + \frac{1}{2\pi}\int_{-\infty}^{\infty} \mu(\tau)e^{-iw\tau}d\tau, \qquad \omega \in \mathbb{R}$$

For comparative purposes, we normalize $g_N(\omega)$ and only consider frequencies $0 \leq \omega \leq \pi$ by defining

$$g_N^+(\omega) := \frac{\pi 2 g_N(\omega)}{\lambda}, \qquad \omega \geq 0$$

We estimate Bartlett's spectral density $g_N^+(w)$ of a point process $N(t)$ using a periodogram. For a recorded series of n ordered arrivals $t_1, t_2, \ldots, t_n$, in an interval of length t_0, the estimated spectrum $\hat{g}_N^+(\omega)$ corresponding to $g_N^+(\omega)$, $\omega \geq 0$, is estimated using (Cox & Lewis, 1966)

$$\hat{g}_N^+(\omega) = \frac{1}{\lambda t_0} \sum_{j=1}^{n} \sum_{l=1}^{n} e^{i\omega(t_j - t_l)},$$

which, in our case, is smoothed by averaging over the periodograms corresponding to the individual 10min batches each containing a sequence of recorded arrivals. As illustrated, e.g., in Bartlett (1963), Cox and Lewis (1966), and Hawkes (1971), Bartlett's spectrum can be used - similar to the traditional power spectral density - to reveal initially-hidden periodicities. However, in this case we aim to extract periodicities in the arrival events. Note that for the memoryless Poisson process $\mu(\tau) = 0$, so that the corresponding spectrum $g_N^+(\omega) = 1$, i.e., it plays a similar role as white noise in the standard time series approach (Lewis, 1970).

For volumes collected on longer timescales in the order of seconds, the dispersion in the counts can be measured as the variance-over-mean ratio using the increment of interest. The variance-time curve $V(t) := \mathrm{var}(N(t))$ - a continuous function of aggregation level t - can be expressed as

$$V(t) = \lambda t + 2 \int_0^t \int_0^v \mu(u) du dv, \qquad t \geq 0, \tag{3.2}$$

and the corresponding dispersion index (of counts or volumes) $I(t)$ is defined as (Gusella, 1991)

$$I(t) = \frac{V(t)}{M(t)}, \qquad t \geq 0.$$

Directly calculating $I(t)$ is difficult, and therefore we apply discretization as follows based on Cox and Lewis (1966) and Paxson and Floyd (1995). We calculate the number of arrivals during a 1s interval, and smooth over consecutive $1s$ intervals similarly as in (3.1). In this way, we construct a variance-time curve over different timescales.

In general, the second-order properties of the inter-arrival times and the counts provide complementary information. For example, dominant frequencies in the arrival events due to upstream interruptions are difficult to recognize using the correlational structure the inter-arrival times only. Yet, the dispersion indices are related in that (Gusella, 1991)

$$\lim_{t \to \infty} I(t) = \lim_{k \to \infty} J_k,$$

assuming both limits exist. Before we consider the real-world arrival data, we consider second-order properties under theoretical arrival processes in the upcoming sections.

3.4.3 Headway distribution and platooning

First, we consider the influence of the headways and the platoon distribution on the statistical characterizations of the arrival process. In fact, under minor assumptions, we can find a theoretical Bartlett spectrum which we can use to examine the periodogram for the measured arrival events. For example, for stationary renewal processes with each inter-arrival time X having distribution function $F_X(x)$ (Daley & Vere-Jones, 2003),

$$g_N^+(\omega) = 1 + \frac{\mathcal{L}\{F_X\}(i\omega)}{1 - \mathcal{L}\{F_X\}(i\omega)} + \frac{\mathcal{L}\{F_X\}(-i\omega)}{1 - \mathcal{L}\{F_X\}(-i\omega)}.$$

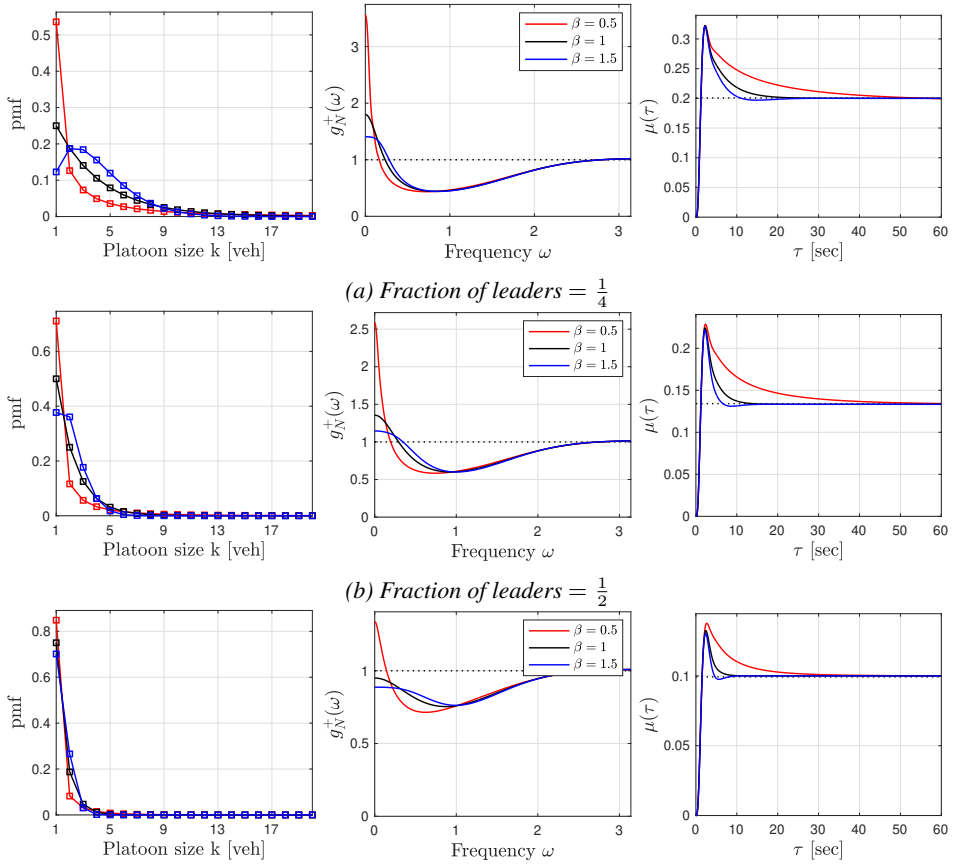
Here, $\mathcal{L}\{F_X\}(s) = \int_0^\infty e^{sx} dF_X(x)$ is the Laplace-Stieltjes transform of X. As mentioned, if inter-arrival times are exponentially distributed, the normalized Bartlett spectrum is flat with $g_N^+(\omega) = 1$ for all $\omega \geq 0$. Hence, departures from a Poisson process can directly be observed in this manner. In literature, more realistic yet complicated headway distributions were proposed, for example Branston's Generalized Queuing (BGQ) model (Branston, 1976; Hoogendoorn & Botma, 1997; Luttinen, 1996), in practice estimated solely based on the recorded inter-arrival times (and not the order therein). While this headway distribution model accounts for the fraction of leaders and followers, it cannot be directly used in a simulation model since the platoon structure does not follow from the distribution function alone.

Assume that vehicles are classified as either a leader or a follower (see Luttinen, 1996). A platoon P is then assumed to consist of one leader and zero or more followers. Let us model the platoon size distribution according to a discretized Weibull distribution (Khan et al., 1989; Nakagawa & Osaki, 1975), i.e.,

$$\text{Prob}\{P = k; q, \beta) = (q)^{(k-1)^\beta} - (q)^{(k)^\beta}, \quad k = 1, 2, 3, \ldots, \tag{3.3}$$

with $\beta > 0$, and $0 < q < 1$. The distribution as defined by its probability mass function (3.3) has two parameters, with β determining the failure rate. $0 < \beta \leq 1$ indicates a decreasing failure rate, while $\beta \geq 1$ expresses an increasing failure rate relative to a Geometric distribution (where $\beta = 1$). In our context, $0 < \beta < 1$ means that the probability of another vehicle (follower) in the platoon increases when platoons grow in number (so-called overdispersion). For a given 10min volume estimate s, we can assess the platoon dispersion by solely considering β, since the other parameter q can be considered to follow directly to match the moments of the headway distribution. Indeed, the fraction of leaders p_l and followers $p_f = 1 - p_l$, respectively, are usually part of the parametric headway distribution. In any case, one can use, e.g., the correlation in consecutive headways ρ_k (or V_k), the variance-time curve $V(t)$, or the estimated spectrum $\hat{g}_N^+(\omega)$ to infer both the platoon structure simultaneously with the headway distribution.

Figure 3.8 shows the platoon distribution, the Bartlett spectrum and the covariance density function under a variety of scenarios regarding the fraction of leaders and the platoon dispersion index. Here, we model the distribution in inter-arrival times for followers and leaders according to the BGQ model. We note that for a given fraction of leaders - independent of β - the distribution in inter-arrival times is fixed. Hence, a changing platoon structure cannot be revealed based on the headway distribution only but can be clearly distinguished based on the spectrum. In the short-term, with overdispersion, the probability of another arrival increases and thereby the covariance density function is more slowly de-



(a) Fraction of leaders $= \frac{1}{4}$

(b) Fraction of leaders $= \frac{1}{2}$

(c) Fraction of leaders $= \frac{2}{3}$

Figure 3.8: Impact of platoon dispersion on platoon distribution (left), spectral density (middle), and covariance density function (right) under different settings.

creasing compared to a geometric distribution. According to (3.2), $\beta < 1$ would introduce variance in the counts on longer timescales. However, the noise-level function indicates that the (count) dispersion index is volume and location-invariant and thus almost constant. Hence, increasing overdispersion in the platoons is likely to introduce positive yet minor serial correlations in consecutive inter-arrival times (compare the limit of J_k with the one of $I(t)$).

3.4.4 Traffic light

We consider the influence of a traffic signal directly upstream of the location of interest on the second-order properties of the counts and the inter-arrival times. For the sake of illustration, we assume that one traffic signal with static, fixed-time, control filters an upstream Poisson arrival process having arrival rate λ^0. Therefore, we introduce a stochastic continuous-time signal $Y(t)$ roughly mirroring the traffic light cycle phases, with $Y(t) = 1$ if the light is green at t, and $Y(t) = 0$ otherwise. For a single realization $y(t)$ of $Y(t)$, the arrival rate at the measurement location is no longer constant, i.e., $\lambda(t) = \lambda^0 y(t)$, and it follows that

$$\mathbb{E}[dN(t+\tau)dN(t)] = (\lambda^0)^2 y(t)y(t+\tau)(dt)^2. \tag{3.4}$$

By taking expectations over $Y(t)$, $\mu(\tau) = (\lambda^0)^2(\mathrm{Cov}(Y(t), Y(t+\tau))$, and the covariance density function is proportional to the autocovariance of the signal $Y(t)$. Indeed, straightforward calculus shows that in this case

$$g_N^+(\omega) = 1 + \lambda^0 \sigma_Y^2 \int_{-\infty}^{\infty} \rho_Y(\tau) e^{-iw\tau} d\tau, \quad \omega \geq 0.$$

Here, σ_Y^2 is the variance of $Y(t)$, and $\rho_Y(h)$ is the corresponding continuous-time autocorrelation at time lag h. It follows from the definition of the Fourier transform of a continuous-time signal that the power spectral density of the traffic signal's phase times is revealed in the Bartlett spectrum of the point process. Slightly more complicated but similar calculus indicates that periodicities in the (dynamic) traffic light cycles are also visible in $g_N^+(\omega)$ under more complex and merging arrival processes upstream. In Figure 3.9, we show the covariance density function and the theoretical Bartlett spectrum assuming that the upstream signal is static and both the green and red time are 30sec. Both the spectrum and the density function clearly reveal the impact of the traffic signal, i.e., we can expect sine wave-like behavior in the covariance density function and clear spikes in the spectrum when considering interrupted arrival processes. If no information regarding the traffic light signal control is available, its corresponding autocovariance structure can be revealed by comparing the ratio between the modeled and the observed density function, see (3.4).

The downstream headways are also impacted in the sense that sine wave-like behavior is introduced in its distribution function and is likely to introduce variance in the inter-arrival times. In any case, the spikes in the frequencies cause that the variance time-curve $V(t)$ shows a more complicated and sinusoidal-like structure over time and highly impacts the serial correlation coefficients ρ_k of the inter-arrival times.

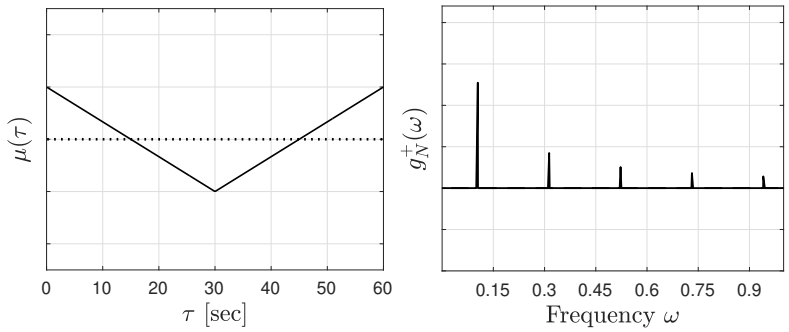


Figure 3.9: Theoretical covariance density function and Bartlett spectrum for an arrival process interrupted by a static traffic light upstream.

3.4.5 Spatial dynamics

Arrival processes at locations can be considered in isolation, but, e.g., for traffic control purposes, the changes in the process over space are of interest. These dynamics are much more difficult to capture since so-called displacements in time are far from independent with vehicles clearly interacting due to the limited take-over opportunities in an urban setting, i.e., typically first-come first-serve properties should be guaranteed. At the same time, complex urban dynamics make that even for small stretches conservation of flow is not guaranteed (e.g., due to on-street parking), and the arrival rate is thus not necessarily constant over space - not even for short stretches.

For the sake of the example, consider the case of random displacements with flow conservation, that is, an artificial setting in which vehicles depart from one lane and then each vehicle travels in its own lane at their desired speed. We label the upstream location 1, and the downstream (cross-section) location as 2. We can model the displacement according to the density function $h(x)$ - likely to be similar to a Gaussian or Gamma density function with the mean and standard deviation a function of distance and (average) speed. Following Brillinger (1975) and Cox and Miller (1977), the resulting Bartlett spectrum $g_N^{+,2}(\omega)$ of the superposed process downstream can be expressed as a function of the upstream spectrum $g_N^{+,1}(\omega)$ by

$$(g_N^{+,2}(\omega) - 1) = (g_N^{+,1}(\omega) - 1)\mathcal{F}[h(x) * h(-x)], \quad \omega \geq 0,$$

with $\mathcal{F}[h(x) * h(-x)]$ the Fourier transform of the difference of two independent random displacements. Hence, only the non-Poissonian part of the spectrum is influenced. Considering the setting in which a displacement is modeled by a Gamma distribution, it is interesting to see that the low-frequency periodicities are preserved over space, while high frequency-parts of the spectrum rapidly become flat. In other words, the variations in the short time gaps will become similar to an exponential distribution, while the longer-term gaps introduced by the upstream interruptions will be preserved for a longer period of time.

Obviously, the dynamics in practice are much more complex and difficult to model directly. Yet, one can estimate relations by comparing the spectra or by using the cross-

intensity function $\gamma^{1,2}(t, t+\tau)$ between upstream location 1 and 2:

$$\gamma^{1,2}(t, t+\tau) = \text{Prob}\{\text{event at 2 at}(t+\tau, t+\tau+dt] \text{ and event at 1 at } (t, t+dt]\}, \quad (3.5)$$

or the corresponding cross-covariance density function $\mu^{1,2}(t, t+\tau)$ and its Fourier transform. We note that $\mu^{1,2}(t, t+\tau)$ makes no assumptions regarding the order among the vehicles nor on the conservation of flow and thus can be used in a variety of settings.

3.5 Statistical characterization of inter-arrival times

We statistically characterize inter-arrival times based on the marginal distribution and the correlation in consecutive inter-event times. The dispersion is assesses relative to the mathematically appealing Poisson process, where the marginal distribution is exponential, and there is no serial correlation in successive inter-arrival times. In this section, we investigate the arrivals under different conditions for the approaches under consideration using a set of metrics describing the empirical distribution and the autocorrelation structure relative to Poisson.

3.5.1 Marginal distribution

The characterization of the noise in the 10min volumes (Section 3.3.4) indicates that, neglecting the slight overdispersion, it shows agreement with the Poisson distribution. In fact, earlier studies (e.g., Mahalel & Hakkert, 1983) mention that the exponential distribution provides an adequate description of the marginal distribution in inter-arrival times beyond a certain cut-off point. We further study the distribution of the empirical inter-arrival times in detail.

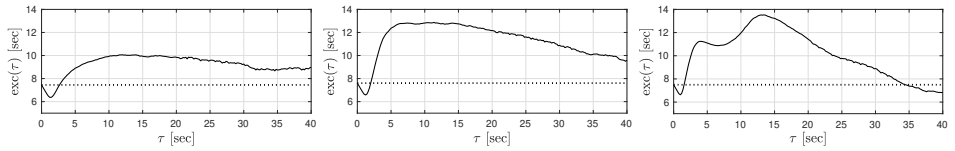


Figure 3.10: Three examples of the mean excess function of the inter-arrival times, as a function of threshold τ. The left figure shows the mean excess function for an approach with an upstream intersection at 1.9km, the middle figure shows the mean excess function for an approach with an upstream intersection at 0.2km and the right figure corresponds to an approach with an upstream intersection at 30m. The dotted line provides the mean excess function in case of a Poisson process.

The distribution function can be considered to be a mixture of simpler distributions. We use the mean excess function,

$$\text{exc}(\tau) = \mathbb{E}[X - \tau | X > \tau],$$

as a function of $\tau \geq 0$, as a tool to recognize different regimes in the distribution by indicat-

ing the information change provided by the 'age' τ. Figure 3.10 shows a sample estimate

$$\mathrm{exc}_n(\tau) = \frac{\sum_{i=1}^{n}(X_i - \tau)\mathbb{1}(X_i > \tau)}{\sum_{i=1}^{n}\mathbb{1}(X_i > \tau)},$$

with $\mathbb{1}(A)$ the indicator function of event A (see, Markovich & Krieger, 2010), for two approaches under similar demand conditions. With the excess function being a constant for all $\tau \geq 0$ in case of a Poisson process (i.e., *memoryless*), in our case, we roughly observe a mixture of three distributions. In the left figure, we can identify leaders beyond 10-12 seconds with an exponential or Weibull-like tail. In the right figure, we have a clear follower distribution for $\tau \in [0, 5]$, an intermediate regime showing similarities with an exponential distribution ($\tau \in (5, 15]$), and a light-tailed regime for $\tau > 15$. This light tail indicates an increasing probability of an arrival in case no arrival occurred for a longer period, and is in that sense less bursty compared to a Poisson process. With the right figure mirroring the inter-arrival times closer to an intersection, the follower distribution shows less variation possibly since in this case it is mirroring the inter-departure headways under near-saturated conditions. The mean excess function for a location directly downstream of another intersection is substantially more complex, and it is difficult to distinguish leaders and followers due to the highly variable curve additionally reflecting the all-red time for the approaches towards the location under consideration. When comparing the different figures as a function of distance, we observe that an increasing distance filters the inter-arrival times between 5-15 seconds, and the light-tailed regime becomes more like an exponential one. In any case, a one-size-fits all headway distribution is unlikely particularly due to the complicated structure introduced by traffic lights upstream. In fact, we were only able accurately capture the distribution in inter-arrival times using the BGQ model provided that the traffic light (visually) played no prominent role.

3.5.2 Correlational structure and platoon dispersion

Without any further information on the memory structure in the intervals, we can try to model arrivals according to a renewal process where inter-arrival times are iid. This assumption can relatively easily be tested if both the headways and aggregated volume noise are available, since the asymptotic distribution of the counting process is known in is such a case. One can assume that this limiting distribution is achieved at a 10min level. Hence, we compare the renewal process-assumed variance at a 10min level with the volume noise variance of Section 3.3.4, and any significant difference indicates the persistence of serial correlations ρ_k on this resolution.

We calculate for each approach and each bin under consideration the sample $\hat{C}^2(X)$ of the squared coefficient of variation, and under the renewal assumption the (noise) variance in the counts on a 10min level is approximately $C^2(x) \times s$, with s the bin-dependent 10min demand. Figure 3.11 shows the 10min volume variance as a function of the demand for four different approaches - representative for all the approaches in our data set. In fact, independent of the location, $C^2(X)$ can be assumed to scale in a linear fashion with respect to the demand, meaning that in the renewal case the 10min volume variance is a quadratic function of the underlying volume.

Figure 3.11 illustrates that $C^2(X)$ is not volume-invariant in general, that the renewal assumption does not hold, and that the dispersion index highly differs from location to

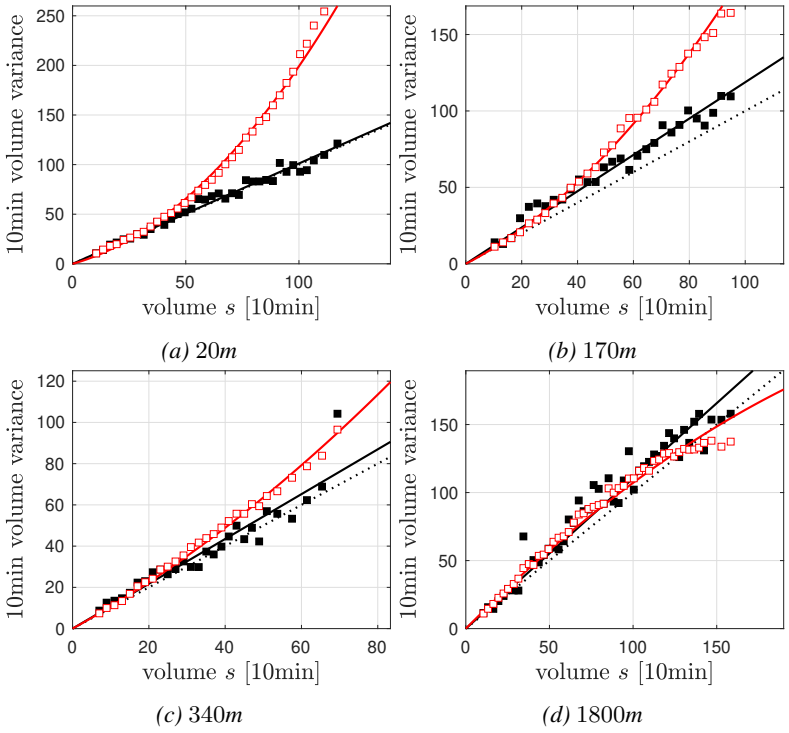


Figure 3.11: Coefficient of variation multiplied by mean 10min volume (red squares) - including quadratic fit (red line) as a function of 10min volume for different approaches. Black dots and black line provide the 10min noise variance and linear fit, respectively. The dashed line indicates the volume-dependent variance corresponding to a Poisson process.

location. In fact, where we see slight underdispersion when the distance to the upstream interruption increases (similar to freeway observations in Luttinen (1996)), obvious overdispersion occurs directly downstream of another intersection. The relative dispersion becomes more prominent when volumes increase - indicated by the quadratic nature of the red line in Figure 3.11. The difference between the black and red line in the figure reflects the persistence of the serial correlation coefficients on a multitude of inter-arrival times (see (3.1)). Hence, when looking on a 10min scale, negative serial correlations dominate near-intersection headways while positive serial correlations dominate the locations further downstream.

Although the distribution of inter-arrival times alone is not sufficient to describe the volume noise variance, the individual correlation coefficients tend to be very small (typically, absolute values less than 0.05). Hence, they persist over many intervals as indicated by Figure 3.11. We briefly consider the dispersion index J_k as a function of k, i.e., the variance in inter-arrival times over a multitude of vehicles (Figure 3.12). The volume dispersion index is an exponential function of the lag. Where absolute serial correlation coefficients damp beyond 3-5 lags, their relative impact on a multitude of vehicles is substantial. Comparing the dispersion indices for the different locations, these serial correlation coefficients grow when

volumes increase and are particularly prominent for near-traffic light locations. For such locations, the correlation structure might only be revealed when looking at the headways of dozens of vehicles in a row.

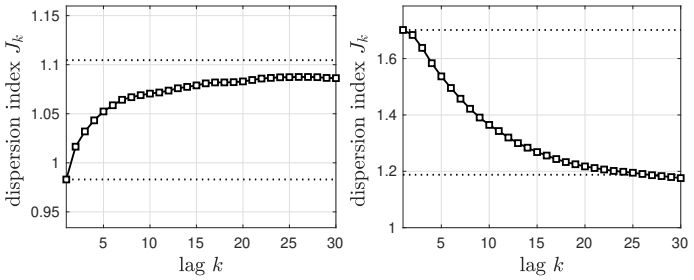
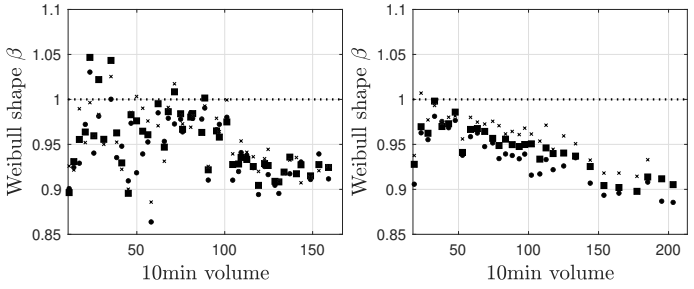


Figure 3.12: Examples of estimates of the volume dispersion index at two approaches under high volumes.

It is difficult to assess the origin of the serial correlation coefficients if the headway distribution is impacted by the traffic light upstream. To assess the impact of platooning, we consider the arrival structure (see Section 3.4.3) for those locations and volume bins where the traffic light has visually little impact on the headway distribution (*uninterrupted processes*). We estimated the platoon dispersion at these locations as follows. We use a cutoff point, and identify the followers as those having an inter-arrival time (compared to the preceding vehicle) of less than this cutoff value. The other vehicles are identified as leaders. Provided this characterization we can identify the distribution of platoons using the discretized Weibull distribution (3.3), where for $\beta < 1$ the distribution has a stronger peak compared to a Geometric distribution to which we refer as overdispersion (i.e., an increasing chance of very long platoons).



(a) 1.8*km* *(b) Off-ramp*

Figure 3.13: Platoon dispersion index β as a function of the volume for uninterrupted processes. Black squares provide the dispersion index estimates using 5 *seconds, black dots provide the dispersion index using* 4 *seconds and black crosses is the dispersion index provided a cutoff point of* 6 *seconds.*

Figure 3.13 shows the estimate of β under a variety of conditions, using cutoff points of 4, 5 and 6 seconds. For all cutoff points, we observe a similar trend, with increased overdispersion as volumes increase, i.e., the probability of another vehicle in a platoon

increases as the platoon size grows. In fact, this is confirmed by small yet significant serial correlation coefficients when considering the inter-arrival times, as is in line with the theory of Section 3.4.3. In this case, the probability of single-vehicle platoons increases as well meaning that arrival processes are characterized by alternating short and long platoons.

3.6 Statistical characterization of the counts

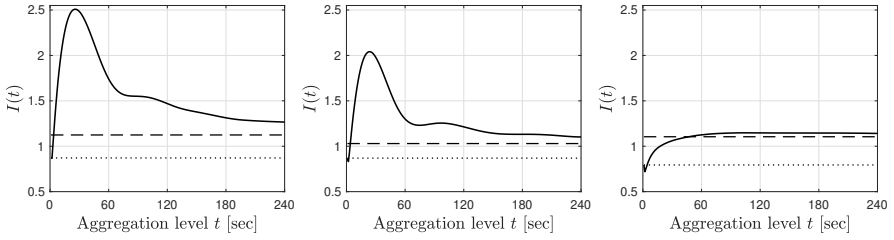
In the previous section, we considered the arrivals based on the inter-event times, showing a very different burst and memory structure compared to a Poisson or renewal process. This structure persists over a multitude of vehicles. In this section, we consider the variations in the counts at different scales.

The dispersion in counts is typically measured using variance-to-mean ratio $I(t)$ as a function of the aggregation interval $t \geq 0$. Thereby, one assesses the burstiness of the process by 'smoothing' over time (Paxson & Floyd, 1995). The corresponding variance-time curve $V(t)$ converges rapidly and scales in a linear fashion with the inverse of the mean inter-arrival time if the autocorrelation function is rapidly decaying. Yet, this curve shows more complex behavior under interrupted settings. The covariance density function $\mu(\tau)$ - and thereby $V(t)$ according to (3.2) - is then expected to show damped wave-like behavior. Hence, $V(t)$ is expected to converge only slowly, if at all.

Figure 3.14 shows typical examples of the (count) dispersion index over different accumulation intervals, estimated using 1sec increments. Considering the behavior of $I(t)$ over t, one is mostly interested in obtaining 'stable' counts for time series analyses in the sense that the random fluctuations in consecutive volume measurements are uncorrelated. On the other hand, the existance of serial correlation in volume measurements can benefit predictions. When the increment-levels approach 0 (from above), the dispersion index approaches 1 since $\mathbb{E}[dN(t)] = \mathrm{var}(dN(t))$. In the order of a few seconds, counts follow a Bernoulli distribution, and the dispersion index is less than 1 by definition. The curves in Figure 3.14 indicate that correlations in consecutive small increments are substantial, i.e., variations on shorter timescales are predictable when considering short time intervals but are unpredictable when aggregation levels increase. In addition, this substantiates that discrete-time arrival processes are solely useful if the correlations are incorporated (Boon & Van Leeuwaarden, 2018). Further, as illustrated by Figure 3.3, the volume measurements using 15-45sec increments may show substantial overdispersion particularly when traffic control influence arrival processes (see also Ritchie, 1983). Then, stable volume increments are only reached on 3 to 5min level, compared to the 1min level in the uninterrupted case.

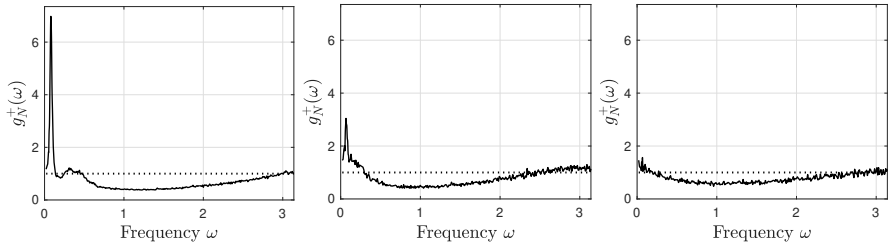
Rather than looking at the density or intensity function directly, we consider the point process periodogram to reveal the underlying reason for the behavior of the count dispersion index over time (Figure 3.15). As already illustrated in Section 3.4.4, interrupted processes can be characterized by the dominant frequencies in the power spectrum, while uninterrupted processes show a much smoother spectral density under similar demand conditions. When volumes grow, dominant periods appear even for locations that are relatively far from the traffic light upstream. Yet, the limiting value $g_N^+(0^+)$ is relatively stable over the locations and the different volume scenarios. Indeed, the influence of signal control fades if aggregation levels increase.

For a few measurement locations, we have additional information of the realized traffic light cycles upstream, and we confirm that the dominating frequencies in the estimate $\hat{g}_N^+(\omega)$



(a) 160*m* *(b)* 200*m* *(c)* 1800*m*

Figure 3.14: Estimates of the count dispersion index $I(t)$ at three approaches. The dashed line indicates the 10min noise-level estimate, while the dotted line is the dispersion index at a 1*sec level*

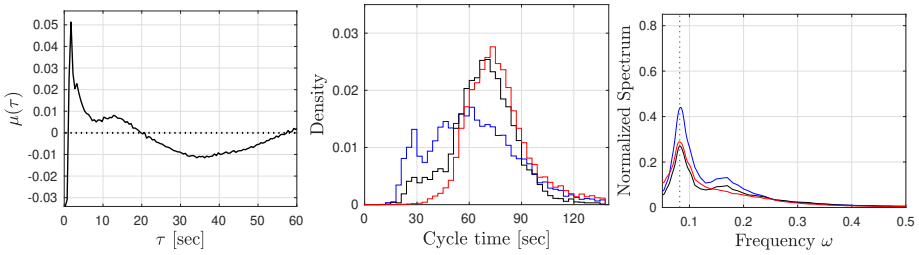


(a) 20*m* *(b)* 200*m* *(c)* 1800*m*

Figure 3.15: Estimated spectral density for the arrivals measured at three different locations under comparable demand levels.

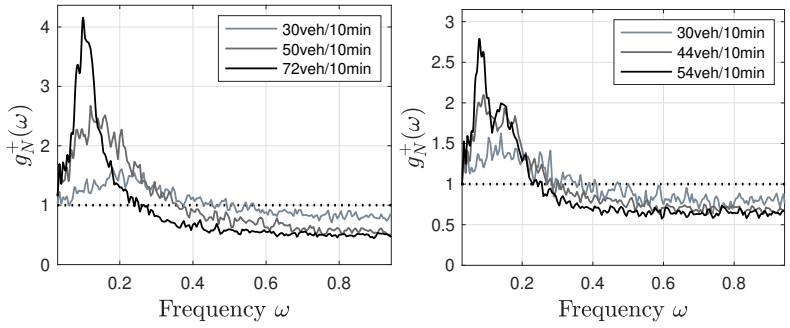
correspond to the periodicities in the cycle times. For the scenario considered in Figure 3.15a, Figure 3.16 shows the covariance density function, and the cycle-time densities and the spectral density of the traffic light cycles upstream - solely considering those upstream legs with the appropriate direction of travel. Here, we normalized the spectrum so that it reflects the relative volume contribution of the legs upstream and depict a virtual line at the frequency where the point process spectrum has its peak. The periodogram in Figure 3.16c is in agreement with the one in Figure 3.15a. In any case, the covariance density function shows clear sinusoidal behavior, indicating that there is a higher probability of no arrival approximately 35 seconds after an arrival event. This is also reflected in the second smaller peak in Figure 3.16c.

When considering the spectrum as a function of the volume, we observe as expected that the spectrum shows a smaller spike with a higher peak at lower frequencies when volumes increase - see Figure 3.17 for illustrative examples. Indeed, vehicle-actuated signals tend to behave similar to a traffic light with fixed-time control in near-saturated and saturated conditions and thus introduce clear periodicities in the arrival events. For very low volumes, there are no clear periodicities. Such periodicity-corresponding peaks in the spectrum appear to be less prominent if the distance to the closest upstream intersection increases. In fact, we tried to relate the spikes in the periodogram with the (log10) distance to the closest upstream intersection. Although with increasing distance, less-strong periodicities occur, the inverse relation is not true in general. Even during high-volume occasions, some locations show



(a) Covariance density function *(b) Probability density of the cycle times* *(c) Periodogram of the cycle times*

Figure 3.16: Covariance density function (a) for the location-regime combination depicted in Figure 3.15. For the traffic light signals directly upstream, we indicate the cycle-time distribution (b), including its periodogram (c).



(a) 200*m* *(b)* 350*m*

Figure 3.17: Bartlett's periodogram as a function of volume.

no major spike. Hence, complex location-dependent characteristics such as the capacity are necessary to account for.

We further study the space-time dynamics by comparing the discharge processes upstream with the arrival process at the location of interest as follows. For two measurement locations of interest, we have accurate upstream stop loop detector data available from which we construct an artificial superposed process including all departure event times at the stop loops of interest. Under conservation of flow, this process shows much similarities with the arrival process downstream provided that the corresponding distance is short. One might expect that spikes decrease in height over space - but the dynamics over space, again, are highly link dependent. Considering two close points in the network, we can estimate a cross conditional-intensity function expressing the probability of an arrival at $t + \tau$ at a downstream location provided a departure occurs at t. As expected, the cross covariance density functions for points that can be assumed to be on the same link (in both cases, approximately 250m apart) show Gaussian-like behavior (top row in Figure 3.18), with its mean roughly reflecting the distance in relation to the average speed which can be used to predict arrival events downstream. Note, here, that conservation of flow does typically not occur and is also not assumed. The corresponding estimated spectra at the individual

locations (bottom row in Figure 3.18) show different characteristics. The left periodogram indicates that the dominant periodicities fade out, while the right one indicates that they persist over space. For all locations, however, we observe a change in the spectrum at the higher frequencies. Indeed, discharge headways typically have limited variability - but variation in these short headways is rapidly introduced just after departure.

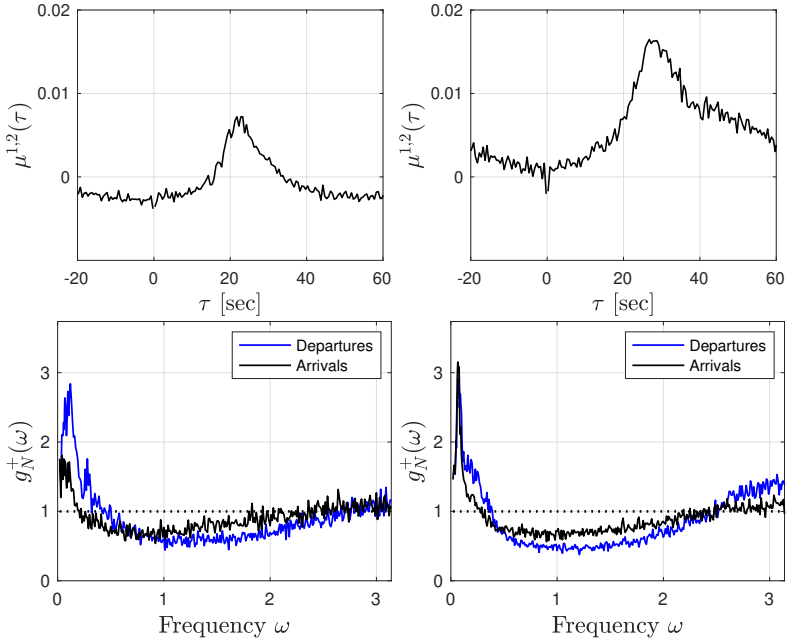


Figure 3.18: Cross-covariance density function between departures (upstream) and arrivals (downstream). The bottom row provides estimates of the power spectral density of the point processes at the individual measurement locations.

3.7 Variations in delays

We statistically characterized arrival processes in an urban setting over different temporal and spatial scales. In this section, we discuss the implications of our results for the variations in the delays experienced at intersections.

Although the stochastic fluctuations in 10min volume measurements are appropriately modeled using a Poissonian framework, such a process fails to reflect the structure in the arrival process on short timescales. Yet, the most popular and still widely used delay approximation function of Webster (1958) is estimated with exponentially distributed inter-arrival times. As Van Leeuwaarden (2006) already pointed out, differences in delay estimates under various stochastic arrival processes can be considerable – even under static traffic light control. We further study this statement in the light of our results in a simulation environment with vehicle-actuated traffic signals.

We simulate a setting with an artificial signalized intersection having three single-lane approaches. Traffic signal control is vehicle actuated, i.e., phase timings are variable, there-

fore using two loop detectors. A stop loop is used for green requests, while an upstream loop is employed for green time extension (see Figure 3.5). Provided a signal turns green, it remains green for at least 6 seconds and is extended as long as a vertical queue is present, or arrivals occur at the upstream loop. If the queue is dissolved, and there has been no arrival (with constant speed) at the upstream loop detector for 3 seconds, the light turns amber and then red for at least 6 seconds. Now, a different service point is provided with green – assuming there is a pending green request since a queue is present there or a vehicle arrived at the corresponding stop loop. Alternatively, an all-red period occurs until such a request is made. During the green period, vehicles can depart instantly if they arrive during a green time, the queue is empty, and the 'server' is available. After every departure, the server remains unavailable for 2 seconds.

We mirror the real-world arrival process at a measurement location using data regarding the realized phase timings as follows. In (near-)saturated conditions, the realized phase timings can be assumed to be roughly independent from the arrival process. For the measurement point under consideration, we confirm this assumption since a BGQ point process - expressed as a signal - 'multiplied' with the continuous-time binary signal corresponding to the realized phase timings (see Section 3.4.4) mirrors the measured intensity, covariance density function and the Bartlett spectrum at the stop loop quite accurately. We simulate an arrival process by superposing the artificial departure processes each consisting of a simulated BGQ point process filtering out those arrivals that occur during red times, whereby we randomly sample from 10min phase timings as a whole to capture within-cycle and between-cycle variations. When comparing the simulated arrival process with the measured one, we see that we quite accurately mirror the dynamics – including the periodicities and thus the dominant frequencies - particularly under high volume occasions.

In the simulation setting, we consider one major approach and two minor approaches assuming a stationary demand for 35min (including the 5min warm-up period). For the major approach, we consider varying arrival rates expressed in 10min volumes, while the minor approaches have independent arrival processes with the same characteristics (compared to the major approach) but with a relative arrival rate of 40 or 60% by independently filtering arrival events. We use 500 replications for each considered scenario and compare the delay estimates using the constructed so-called periodic arrival process with the Poisson process. We assume a maximum green time of 60s for the major approach, and 30s for the minor approach if the relative arrival rate is 40%, and 40s otherwise.

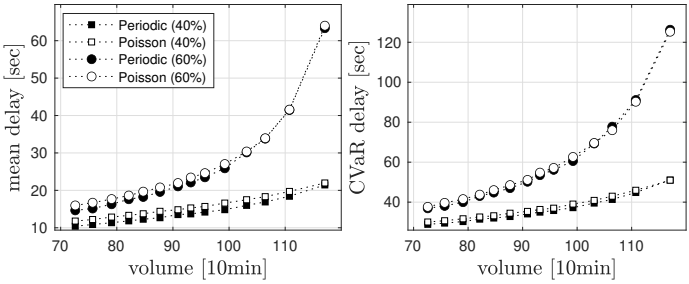


Figure 3.19: Impact of arrival process on the mean delay and 10% *CVaR for different volume levels.*

Figure 3.19 expresses the impacts of the arrival process on two indicators regarding the

delay distribution, the mean delay and the 10% conditional value at risk (CVaR). The CVaR provides the mean delay of the worst 10% of the travelers, i.e., an indicator the tail of the delay distribution. We observe that the arrival process can have considerable impact on the delay distribution, in particular in unsaturated conditions. In fact, the mean and CVaR of the delay is overestimated when using a Poisson process by approximately 1.5-2s. Even though these impacts seem limited, they can be considerable when designing strategies for coordinated control. However, when volumes grow these delay differences become smaller, indicating that in saturated condition the actual arrival process has little impact on the delay estimates - although other performance indicators may be considered as well (Viti & Van Zuylen, 2010a). The difference in performance can be explained by the predictive behavior of real-world arrival processes. Indeed, in practice, an arrival event increases the probability of another arrival within limited time because of the platoon structure and the periodicities in the arrival events - see, e.g., the covariance density function in Figure 3.16a. Conditioning on an event in a Poisson process provides no additional information regarding the next event, one can substantiate that traffic signals with extended green times are likely to perform better in practice than can be expected from a simulation environment assuming exponentially distributed inter-arrival times.

3.8 Conclusion

A major share of the variability in travel times in urban networks is determined by delays imposed at signalized intersections. A full understanding of the delay-contributing factors is required for decision makers explicitly anticipating the evolution of uncertainty in delays under various conditions. In this chapter, we provided a statistical characterization of arrival processes at signalized intersections using real-world data regarding the arrival events collected throughout an urban traffic network.

The structure of arrival processes can only be truly uncovered if it is studied on various scales. Indeed, arrivals show a different memory and burst structure when considering it as a sequence of inter-arrival times compared to looking at it as a counting process. Therefore, we studied the arrival processes using both perspectives, and applied both a time and frequency-domain approach. Arrival processes, in general, can be characterized as bursty: longer periods with no arrivals alternate with short periods with relatively many arrivals (platoons).

Non-stationarity occurs in the arrival rate, indicating that the demand - and thereby the phase timings - substantially deviates in time and space. The arrival rate is typically captured using a 24h pattern, and provided that the systematic variability in the network usage is captured, the corresponding fluctuations around this pattern provide a rough estimate of the arrival process on a larger temporal scale. These natural fluctuations were shown to have a slight overdispersion compared to Poisson noise.

When considering the marginal distribution of the inter-arrival times, the headway distribution functions as suggested in literature do not provide an adequate description under all conditions. In an urban setting, there is a higher probability of medium and high inter-arrival times compared to an exponential tail. This excess probability is introduced by traffic lights, and statistically reflects a combination of variable red times and the (interaction with) inter-arrivals upstream. The left-hand side of the distribution shows minor variation during the discharge processes under high-volume occasions, but the variability is rapidly introduced

after departure. In addition, inter-arrival times turn out to show (weak) serial correlation - but this effect accumulates to a significant level when looking at a multitude of vehicles. For uninterrupted processes, serial correlation is introduced due to the platoon formation.

By using the power spectral density corresponding to the sequence of arrival events, we revealed the periodicities in the arrivals that were not visible when looking at a sequence of inter-arrival times. These periodicites are shown to correspond to the cycle times of the traffic light upstream and can be significantly influencing the dispersion indices even under lower arrival rates. As a matter of fact, in particular on shorter temporal scales of less than 30sec, the dispersion in counts can be very high when looking at interrupted processes. The power spectrum density is smooth for uninterrupted processes and dominant frequencies do not appear. Hence, for such processes, the dispersion index is a much smoother function of the aggregation level. Nonetheless, almost all effects smooth out when aggregation levels increase beyond 4-5min.

In general, theoretical delay and queuing models and simulation tools typically make naive assumptions regarding the arrival processes in urban networks. It is complex to accurately capture the structure of the arrivals, since regularities appear on many different scales. In any case, failing to capture this structure for the benefit of tractable arrival processes can underestimate the variations in volumes and overestimate the delays. In case of vehicle-actuated control, in particular for lower volume occasions, a Poisson arrival process overestimates both the mean as well as the CVaR of the delays. Hence, delay estimates - and thereby traffic management control - need to account for location-dependent characteristics although the true impact should be assessed further using real-world delay measurements.

In this research, we limited ourselves to signalized intersections with single-lane approaches. In future research, one could further investigate the arrival dynamics on a corridor and a multi-lane setting. The burst and memory structure of the arrival process become more complex in such a case and dominant frequencies can even be amplified over space. In any case, arrival processes in an urban setting remain an important topic to study.

Chapter 4

Pattern-based prediction of urban traffic volumes

4.1 Introduction

Logistics service providers (LSPs) construct route plans days or hours before execution (e.g., Agatz et al., 2008) but can only base these plans on approximate travel times. Unexpected variations in traffic conditions (e.g., volumes, speeds, delays) make that the actual travel times may strongly deviate from the estimates with initially small deviations potentially accumulating and propagating throughout the plan. Updating route plans while being en route, alongside robust planning, could mitigate a share of the negative impact of the uncertainty on performance. With the initial route plan accounting for the uncertainty in the longer-term prediction of travel times, the route is then adapted during execution in response to changing short-term forecasts so that service requirements are met.

Nowadays, the en-route adaptation of route plans is possible since LSPs continuously monitor the position of their fleet, and can communicate with drivers directly. However, dynamically updating route plans requires a *control center* monitoring and forecasting not only the travel times on current routes, but also throughout the network (Fleischmann et al., 2004). Network-wide traffic volumes (or: flow rates) support the prediction of travel times as follows. First, delays occurring at intersections mainly determine travel times in urban areas and strongly depend on traffic volumes, in particular when these volumes approach capacity. Second, traffic speed measurements are often collected by external companies making it costly to use this data source on a continuous basis. Volume measurements, on the other hand, are, at least in the Netherlands, more easily accessible and provided by induction loop detectors near intersections. Finally, directly predicting driving times under different conditions based on historical travel time measurements is less appealing with many variability-inducing factors changing over time. For example, routes might change from day to day or even within a day, with conditions on different parts of the network continuously changing.

This chapter is based on the following paper: Eikenbroek, O. A. L., Thomas, T., Mes, M.R.K. & Van Berkum, E. C. (2023). Pattern-based probabilistic prediction of urban traffic volumes. *Under review.*

For logistics operators it is important to have accurate forecasts for different timescales including a 24h forecast, an updated forecast for the remainder of the day, as well as a short-term prediction. Indeed, offline route plans are typically constructed overnight which makes that long-term predictions are needed. Dynamic (re-)planning calls for updates of these predictions (remaining-day predictions), as well as short-term forecasts. In this chapter, we therefore include these types of forecasts for volume predictions. Robust (re-)planning demands that also the uncertainty in predictions should be accounted for since solely using point forecasts is treacherous: they give the impression being accurate (Makridakis et al., 2020). In this chapter, we construct predictive densities for both the long and short term explicitly accounting for the uncertainties that occur on the different timescales.

The remainder of this chapter is organized as follows. In Section 4.2, we provide a literature overview on traffic volume prediction methods and discuss the research contributions of the chapter. In Section 4.3, we discuss the data, the uncertainty in predictions, and the relation with patterns and random variation in volume time series. Section 4.4 introduces a novel method to jointly extract the noise-level characteristics as well as recurrent patterns from the data. These patterns are used in Section 4.5 to construct a 24h prediction. We use smoothing in order to update the remaining-day prediction as well as the short-term forecast based on the measurements (Section 4.6). In Section 4.7, we evaluate the prediction mechanism and compare our prediction with the predictions using a method from literature. In Section 4.8, we draw conclusions and discuss topics for further research.

4.2 Literature overview and research contribution

Although the traffic system shows repeating behavior, there is a typical discrepancy between the time of decision of route planners and the available information in that some relevant data is only revaled during execution. Hence, LSPs use predictions with various spatial and temporal scales to make or update route plans.

Many traffic volume prediction methods have been proposed in literature, with a vast majority focusing on short-term point predictions for recurrent freeways conditions (see Vlahogianni et al., 2014). Taking into account the considered application for route planning, volume forecasting methods should (i) offer reliable forecasts for different network settings (urban and freeway) and conditions, (ii) provide predictions for both the long and short term and (iii) quantify the uncertainty in predictions by producing probabilistic forecasts.

We roughly distinguish between data-driven and model-based methods when classifying volume prediction mechanisms. Data-driven methods include time series methods, clustering methods, neural networks, etc. Model-based methods use a traffic model (e.g., CTM (Tampère & Immers, 2007), METANET (Wang & Papageorgiou, 2005)) to describe the underlying state. Obviously, both types of methods use data to update future forecasts, possibly with a filtering technique to relate the measurement space to the state. Model-based methods use a theoretical description of the traffic dynamics in time and space (Van Lint & Van Hinsbergen, 2012), while data-driven models (implicitly) relate recent measurements to (a part of a) historical pattern with the remainder of the pattern providing the predictable fluctuations. Although a merit of using model-based methods is that they are able to provide reasonable forecasts under different conditions, even for situations for which the model has neither been validated nor calibrated, it is difficult to capture the complex dynamics of lower urban roads, including on-street parking and slowly-moving vehicles with limited take-over

opportunities, in a single model. In addition, they are often not designed for long(er)-term forecasts. Taking into account the LSP-imposed requirements that forecasts should cover various timescales and network settings, we limit therefore our discussion in the remainder of this section to data-driven methods.

Conventional time series approaches, most notably autoregressive methods (e.g., ARIMA), identify trends in measurements by decomposing the time series into different underlying series and use extrapolation for forecasts (Ahmed & Cook, 1979; Kumar & Vanajakshi, 2015; Lippi et al., 2013). Such methods have been widely applied for predictions, typically with a focus on freeways, yielding impressive performance for short-term predictions (Lippi et al., 2013). With the more recent development of generalized autoregressive conditional heteroscedasticity (GARCH) models (Kamarianakis et al., 2005; Shi et al., 2014; Vlahogianni & Karlaftis, 2011), time series methods are in addition able to provide reasonable prediction intervals (Guo et al., 2014). Yet, conventional approaches assume that recent measurements strongly correlate with the predicted value(s), i.e., performance is expected to decrease with sudden changes or when prediction horizons increase. Therefore, time series methods are to be used in conjunction with clustering or pattern recognition, e.g., an autoregressive model then captures the behavior of the residuals compared to the intra-day pattern (Chen et al., 2012). Machine learning methods, e.g., neural networks, often use unsupervised learning to infer systematic variations and thereby reconstruct measurements. Learned relations then implicitly (neural networks (Van Lint & Van Hinsbergen, 2012)) or explicitly (clustering methods (Weijermars & Van Berkum, 2005)) contain the patterns in the time series, and how they relate to recent measurements. Such methods benefit from including domain-specific knowledge, e.g., dynamics are 'smooth' among adjacent segments when considering freeways (Polson & Sokolov, 2017). Machine learning methods are shown to provide good predictions in the short term (Lv et al., 2015). Disadvantage is the 'black-box' nature and the often time-intensive learning procedure attached. In addition, machine learning methods do not generalize well beyond their training data meaning that conditions that have not been seen during training cannot be predicted accurately. Indeed, only a minor share of the training set consists of measurements corresponding to non-recurrent conditions, and the training set should then be enriched with additional yet difficult-to-collect variables to particularly focus on such situations. In comparison to time series methods, uncertainty has not been well-accounted for when applying machine learning methods (see also Chapter 2 and Makridakis et al., 2018), and there is evidence in literature that combining statistical methods with machine learning can improve performance (Makridakis et al., 2020).

Probabilistic forecasts provide estimates beyond the point predictions, typically using prediction intervals or density functions. The prediction uncertainty depends on (i) the uncertainty in our knowledge about the current and previous states of the network, (ii) the limits of predictability, i.e., the inherent random behavior of the system, and (iii) the (stochastic) prediction error. Uncertainty in the state is often accounted for by using state-space filtering or smoothing (e.g., Wang & Papageorgiou, 2005), while the uncertainty in a prediction can be accounted for by fitting a distribution to previous prediction errors or by sampling from the residuals. For example, quantile regression is applied by Dutreix and Coogan (2017) to provide density forecasts. Concerning the unpredictability of the system, there has been relatively limited attention for the flow or time-dependent random variation (volatility) in traffic measurements. Regarding volume-dependent random variations, Guo and Williams (2012) apply a transformation to the data (see also Guo et al., 2014), while Thomas et al.

(2010) account for the random variation when constructing and evaluating point predictions. Shi et al. (2014) and Huang et al. (2018) estimate a time-dependent (i.e., time-of-day and day-of-week) variance of the random variation based on an underlying SARIMA model. More authors (e.g., Guo et al., 2014; Guo & Williams, 2010; Yang et al., 2010) construct probabilistic forecasts for the short term by applying a (variant of a) GARCH model. Although probabilistic predictions are provided, evaluation is difficult and often limited to a measure for a single or a few prediction intervals (e.g., Guo et al., 2014; Huang et al., 2018; Khosravi et al., 2011; Li & Rose, 2011; Wagner-Muns et al., 2018; Zhang et al., 2014). Among others, Guo et al. (2014) and Huang et al. (2018) stress the need for a uniform performance measure for constructing and evaluating probabilistic forecasts.

The choice of the prediction method highly depends on the situation under concern, and a direct quantitative comparison of methods - in particular of probabilistic forecasts - is difficult. With a vast majority of literature focusing on short-term predictions for freeways, LSPs require a method that additionally provides long(er)-term forecasts and accompanying uncertainties for urban networks. In an urban context, time series show a high degree of regularity. In fact, many of the systematic patterns are recurrent on timescales longer than 24h and the explicit incorporation of 24h patterns in prediction mechanisms might substantially improve performance (see Chapter 2). At the same time, the intra-day pattern provides the predictable fluctuations for longer prediction horizons (e.g., Ma et al., 2021; Song et al., 2018), particularly useful for constructing offline route plans. Hence, in this chapter we use a pattern-based method for probabilistic urban traffic volume predictions, for 15min to 24h ahead.

We discuss recent approaches using a pattern-based method. Wagner-Muns et al. (2018) applied a functional data approach to freeway data where an underlying set of 24h patterns (so-called components) are fitted to measurements. These patterns provide the intra-day variations in volumes given a corresponding scaling magnitude ('score'). This method predicts component scores using a time series approach, with magnitudes updated based on recent measurements. The uncertainty in forecasts is accounted for by bootstrapping residuals. Although useful for longer-term predictions, shorter-term fluctuations that cover less than 24h, e.g., events with varying starting times, are not incorporated. 24h patterns are also used by Laña et al. (2019), where several clusters of time series are constructed using a neural network. Based on extracted features, a 24h prediction is constructed by selecting the mean time series of the corresponding cluster. Throughout the day, the remaining-day prediction is replaced if the measurements substantially deviate from the initial prediction. Although prediction intervals are not provided, uncertainty is implicitly accounted for when detecting substantial deviations from the predictions. However, small deviations are not used to update forecasts. Habtemichael and Cetin (2016) used a pattern-based approach, to which we refer as the K-Nearest Neighbor (K-NN) method, where recent measurements are compared with volumes from other days. The short-term prediction is a weighted average of succeeding measurements on similar days. This method is straightforward and effective, but neither anticipatory nor adaptive (i.e., there is no feedback between previous errors and current forecasts). In our case, some locations show a Wednesday-noon peak in volumes, which are difficult to predict using the K-NN method since volumes preceding this peak are very similar to the ones on other weekdays (without a peak). This might be a reason that the performance of point predictions is shown to decrease when the forecasting horizon increases. Longer-term point predictions improving a baseline forecast in the order of hours for the benefit of timing plans are constructed using *partial least squares* in Coogan

et al. (2017). There, events with variable starting times are not explicitly incorporated, and the (24h) forecast is not updated based on previous errors. Recently, Ma et al. (2021) predict traffic flows for future days using a neural-network accounting for inter- and intra-day variations. They, however, do not consider short- and mid-term predictions.

In summary, a multi-timescale prediction method benefits from using long- and short-term patterns to capture the systematic variations. Compared to the afore-mentioned studies, we make the following contributions in this chapter. First, we identify long- (24h) and short-term (related to events) recurrent patterns in measurements and use these patterns for predicting volumes over time. Observed patterns in measurements are assumed to be a combination of underlying recurrent temporal patterns (*profiles*) and we use small adaptations (or: transformations) of these profiles to predict volumes. In contrast to the above-mentioned studies, in this chapter we do not only use long-term patterns for forecasts up to 24h ahead, we explicitly incorporate short-term patterns that provide predictable fluctuations that cover less time and should not necessarily influence long-term forecasts, e.g., in case of events. Second, we go beyond point predictions and provide probabilistic forecasts in the form of density functions which we evaluate as a whole. We show that this density function is actually a natural result of the volume-dependent random variation in the measurements and the (stochastic) prediction error. Third, where most studies focus on freeways or major arterials, our data set also covers minor inner-city roads characterized by low speed limits and very low average volumes.

4.3 Patterns and uncertainty in predicting urban traffic volumes

This section describes the traffic volume data we use (Section 4.3.1). In Section 4.3.2, we outline our prediction method. In addition, we discuss the relevance of labeling variations in volume measurements as either being systematic or random and explain how such a decomposition supports probabilistic forecasts (Section 4.3.3).

4.3.1 Data

We predict traffic volumes in the city of Enschede, the Netherlands (+/- 160,000 inhabitants). Volume data were collected at signalized intersections throughout the city (see Figure 4.1) from January 2016 until December 2017. At the arms of the intersections, vehicles were detected through an induction loop at each lane, a few meters from the stop line. These detections are aggregated to 15min-interval count data. The dynamics that occur at the intersections under consideration are complex in the sense that different modes of transport interact, including, depending on the intersection, buses, cyclists, and pedestrians. The intersections are located at both major and minor urban roads, include intersections near off-ramps and on-ramps, and have a speed limit of 30, 50, 70, or 80km/h.

In the remainder of this chapter, we study 36 collections of 24h time series. Each time series consists of elements with 15min volume measurements as measured by a single stop-line loop detector. These time series correspond to multiple stop loop detectors when lanes share direction of travel. We refer to the resulting segments as *measurement locations*, and predict volumes for each of the 36 measurement locations in isolation.

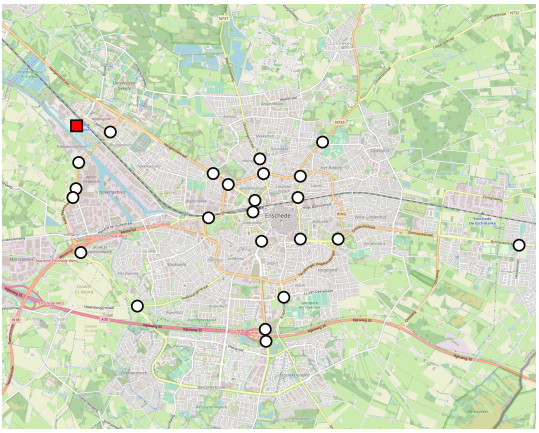


Figure 4.1: City of Enschede (source map: OpenStreetMap (2021)). Dots indicate the signalized intersections under consideration with (multiple) measurement locations. The red square gives the location of the FC Twente stadium.

Throughout the chapter, we particularly focus on the impact of the football matches of FC Twente on the volumes. FC Twente plays its matches in the stadium with a capacity of 30,000 and is located in the north-western part of Enschede (see Figure 4.1). In our data set, we have 32 different matches, including league, cup, and women's matches. These matches were played at different days of the week, with varying kick-off times.

We briefly introduce notations. For a measurement location, let $x_{d,t} \in \mathbb{Z}_{\geq 0}$ denote the 15min volume measurement at day $d \in D$ at time $t \in T$. Here, D is the set of days in the database, and T is the time domain with 15min increments for a single day. Throughout the chapter, depending on the context, we use both $x_{d,t}$ and x_t to refer to a 15min volume measurement.

After inspection of the data, we rejected volume measurements based on the following criteria. First, we excluded a complete day of volume measurements from further consideration if more than 20 elements of a daily time series (of the maximum-possible 96) were either missing or had zero volume. Subsequently, we checked for longer-term shocks in the demand caused by local road works and closures. Since reliable information regarding historical road works and closures was not available, we used the following approach to identify outliers in the weekly demand (see also Thomas et al., 2008). For each location, a simple estimate for the 24h pattern is

$$\bar{x}_t = \frac{1}{|D|} \sum_{d \in D} x_{d,t}, \quad t \in T. \tag{4.1}$$

Based on this 24h pattern $\bar{x}$, the mean squared error

$$\zeta = \frac{1}{|D|} \sum_{d} \tau_d, \quad \text{with} \quad \tau_d = \frac{1}{|T|} \sum_{t \in T} (x_{d,t} - \bar{x}_t)^2, \tag{4.2}$$

gives a sample estimate for the squared deviation of a single day compared to the 24h

pattern. For each week $w \subseteq D$, we calculate the mean distance compared to the 24h pattern, and compare it with $\sqrt{\zeta}$. In fact, we exclude a full week w of measurements if

$$\sqrt{\sum_{d \in w} \tau_d} > 2\sqrt{\zeta} \tag{4.3}$$

holds. Visual inspection confirms that this criterion identifies weeks with unrepresentative high or low demand. After the identification of weeks with unrepresentative demand, we used a modification of the above criterion to diagnose individual days with unrealistic high or low 24h demand. Directly comparing τ_d with ζ turns out to be overly sensitive to events or incidents that cover only a few hours. Therefore, we recalculate ζ and τ_d as in (4.2) but restrict the time domain by excluding the first two hours of each day. We identify days for which $\tau_d > 2\sqrt{\zeta}$ holds. We use an iterative approach by recalculating ζ and τ_d and only excluding the next two hours compared to previous iteration. We remove those days from the data set for which $\tau_d > 2\sqrt{\zeta}$ holds for all possible sub time domains.

We note that data were collected in batches, with each batch covering 3 months of data. For some of the intersections, a single batch is missing from the data set. The considered measurement locations have at least 500 representative days in the database. Here, the above criteria led typically to a rejection of $1-3\%$ of the 24h time series. We underline that we do not artificially replace (missing) data, making that the resulting time series still include days with missing elements, miscounts, temporarily malfunctioning loops, events, incidents, etc. For each measurement location, the final data set is divided into a training and a test set. The training set is used for model development and covers the first two thirds of the data, while the test set consists of the last third of the days in the database and is only used for evaluation in Section 4.7.

4.3.2 Outline of the prediction method

Observed patterns in the 15min volume measurements are considered to be a combination of underlying recurrent temporal patterns, or profiles as we will call them. Not all variations in 24h time series can be explained using exogenous variables, but can be expressed using small yet systematic transformations of long- and short-term profiles that change from day to day. Although extracting the underlying profiles is a difficult task (see Section 4.4), these profiles benefit a prediction mechanism as follows. First, only a few underlying profiles with day-dependent adaptations were shown to be needed to explain almost all intra-day and day-to-day volume variations - even with recurrent events (see Chapter 2). Given such profiles, the prediction task then reduces to forecasting which profiles are *active* at which part of the day. Second, the profiles and corresponding magnitudes typically yield a physically-meaningful interpretation albeit additional variables might be needed to explain or forecast the variations. Hence, the impact of external factors or determinants can possibly be incorporated. Third, by decomposing 24h volume time series, we introduce additional degrees of freedom compared to, e.g., clustering and time series approaches that solely consider variations in the resulting (accumulated) volumes. Since underlying profiles can be transformed independently, we are flexible in adapting the prediction to a variety of situations occurring.

The pattern-based prediction method of this chapter provides three types of point and density forecasts: a 24h prediction, a remaining-day prediction and a short-term prediction. The 24h forecast provides a prediction for a full day before the start of the day. The

remaining-day prediction gives at any time of day a forecast for the volumes during the remainder of the day (i.e., up to 24h ahead). Short-term predictions cover up to 1.5h. For all these predictions, we use the underlying volume profiles (extracted in Section 4.4). Considering the application for LSPs, a 24h prediction is particularly useful for offline planning (at night), while the remaining-day and short-term forecasts are to be used for dynamically updating initial plans. Now, we provide a brief outline of the substeps of our methodology, with each type of prediction being an update of a longer-term prediction.

We start with an initial prediction, to which we refer as the baseline prediction, that can be constructed a long time in advance using day-dependent features (i.e., exogenous variables such as the day of the week). The baseline prediction gives a typical 24h volume pattern, based on the feature-dependent scaling magnitudes of the profiles.

We construct the 24h, remaining-day and short-term predictions by comparing a previous estimate with initially unavailable measurements that now can be used to improve the forecasts. We do this in several steps, with each step having its own modifications that we discuss in detail in the corresponding sections:

- The 24h prediction (Section 4.5.2) for day $d \in D$ is a prediction for the volumes throughout the day. This prediction is an update of the baseline prediction for the same day by scaling the initial forecast based on the relative difference in 24h volumes between measurements and the baseline during previous days. Thereby, the 24h prediction accounts for slowly-changing variations such as seasonal variation. In addition, the 24h prediction incorporates traffic flow rate changes due to recurrent events;

- The remaining-day prediction (Section 4.6.2) accounts for systematic errors that are correlated over longer time periods (but less than 24h), which are assumed to be manifested in a changing demand for (a part of) the day. We begin with the 24h prediction at the start of the day and update the prediction for the remainder of a day by comparing the initially unavailable measurements throughout the day with the previous remaining-day prediction;

- A short-term prediction (Section 4.6.3) is the prediction for the next 15min to 1.5h by comparing recent measurements to the remaining-day prediction.

The collected 24h measurements x_d, $d \in D$, are *noisy* in the sense that typically a slowly-changing pattern and high-frequency fluctuations can be recognized by eye. The systematic part s_d of these measurements need to be predicted, and is in fact the only variability that can be predicted. Random variations, on the other hand, are the fluctuations that show no pattern: they are uncorrelated and these deviations are therefore unpredictable (noise) (see Chapter 1). Although a perfect prediction scheme estimates $s_{d,t}$ over time, the *true* systematic flow is unavailable and random variations make it difficult to distinguish systematic from random errors. In any case, an estimate of what can(not) be predicted identifies potential ways to improve the forecasts. However, the presence of noise complicates this process.

Roughly, we make use of two techniques to reduce the negative effect of the noise when updating the predictions. For the 24h prediction, we use aggregation among previous residuals to reduce the relative impact of the noise (compared to the systematic variations). When constructing remaining-day and short-term predictions, we only use a relatively short

history of measurements for updating. Rather than using the recent measurement directly, we adapt the prediction based on a *smoothed* estimate of the volumes. In fact, in Section 4.6.1 we introduce a state-space smoothing method to relate recent volume measurements to the underlying profile magnitudes (the state), while explicitly accounting for noise as well as for model and prediction errors. In the next subsection, we discuss the relevance of decomposing systematic and random variations for our probabilistic predictions.

4.3.3 Patterns, noise and uncertainty when predicting volumes

Many of the patterns in the time series are recurrent, and can therefore in principle be predicted. In our urban traffic time series, intra-day patterns show a high degree of regularity over the days. With the use of 24h patterns, we can thus describe a major share of the variations, in particular when a variable scaling magnitude is used for variations covering timescales longer than 24h (e.g., seasons). Also short(er)-term systematic variations that cover less than 24h exist, often related to temporarily changing demand and/or supply due to events, incidents, etc. Time-of-occurrence and magnitude of these short-term patterns are much more variable and therefore more difficult to predict (Vlahogianni et al., 2014). Where most prediction methods forecast regular conditions, i.e., days without any events occurring, forecasts can be improved when explicitly accounting for these short-term fluctuations.

Random variation, or noise, in our traffic flow time series is mainly due to the inherent variability in the dynamics that occur near signalized intersections (*process noise*) (e.g., random arrival processes and variable traffic signal cycles), but also measurement errors contribute to the noise. These variations might show patterns on timescales shorter than 15min yet they are uncorrelated on longer timescales and considered to be unpredictable and uncorrelated noise in our case. We assume that residuals $e_{d,t} = x_{d,t} - s_{d,t}$, $d \in D$, $t \in T$, are realizations of independent random variables $\varepsilon_{d,t}$. There is evidence (Chen et al., 2008; Guo et al., 2015; Thomas et al., 2010) that the amount of random variation depends on the underlying pattern. Indeed, count time-series (e.g., of random arrivals) are often modeled using a Poisson distribution (with the variance equal to the mean). Here, we model that the noise is distributed according to a heteroscedastic Gaussian distribution $\varepsilon_{d,t}|s_{d,t} \sim \mathcal{N}(0, \sigma^2_\theta(s_{d,t}))$, with the noise-level function $\sigma^2_\theta(s)$ linear in s so that

$$\sigma^2_\theta(s) = \theta \cdot s, \quad \text{with } \theta \in \mathbb{R}_+. \tag{4.4}$$

Note that for $\theta = 1$, the noise function approximates a Poisson arrival process (for sufficiently large s). In this chapter, we explicitly account for the statistical properties of the noise for predicting volumes over time as follows.

First, we jointly estimate systematic patterns together with the noise level so that we obtain a robust estimate of the temporal profiles while preventing *overfitting*. In addition, for short-term predictions, the random variation substantially influences our updating scheme. Therefore, we infer a *denoised* estimate of the measurements by using a Kalman-like smoothing method that explicitly incorporates the volume-dependent noise.

Second, an accurate estimate of the statistical properties of the noise is important when making volume point forecasts, since the conditional variance of the noise provides a lower bound on the best-possible accuracy of the point prediction method, and therefore indicates the model and prediction error (Hunt et al., 2007). Consider a point predictor $y_{t+h|t}$ for

time $t+h$ made at time t, the corresponding expected squared error becomes

$$\mathbb{E}[(y_{t+h|t} - X_{t+h})^2 \mid s_{t+h}] = \sigma_\theta^2(s_{t+h}) + (y_{t+h|t} - s_{t+h})^2, \tag{4.5}$$

with X_{t+h} the random variable for the 15min volume at $t+h$. Although at time t the best-possible point prediction for $t+h$ is given by $y_{t+h|t} = s_{t+h}$, in general we have a biased prediction. In the remainder, we assume $y_{t+h|t} = s_{t+h} + \tau_{t+h|t}$, with random variable $\tau_{t+h|t}$ so that $\mathbb{E}[\tau_{t+h|t}|s_{t+h}] = 0$ and $\mathbb{E}[\tau^2_{t+h|t}|s_{t+h}] = (c_{t+h|t}s_{t+h|t})^2$, with $c_{t+h|t} \in \mathbb{R}_+$. Hence, when predicting similar conditions, *on average* we have a good forecast but we consistently have a relative error of c. Our prediction of the corresponding (Gaussian) density function is

$$f_{t+h|t}(x) = \phi(x; y_{t+h|t}, \sigma_\theta^2(y_{t+h|t}) + (c_{t+h|t}y_{t+h|t})^2), \tag{4.6}$$

with $\phi(x; \mu, \sigma^2)$ the probability density function of a Gaussian distribution with mean μ and variance σ^2. The $100(1-\alpha)\%$ prediction interval directly follows from (4.6) and is given by

$$[y_{t+h|t} - z_{\alpha/2}\sqrt{\mathbb{E}[e^2_{t+h|t}]}, y_{t+h|t} + z_{\alpha/2}\sqrt{\mathbb{E}[e^2_{t+h|t}]}], \tag{4.7}$$

with $z_{\alpha/2}$ the $\alpha/2$-corresponding z-score of a standard Gaussian distribution (Chatfield, 2001). Here, $\mathbb{E}[e^2_{t+h|t}]$ follows from the error (4.5) and is thus approximated by $\sigma_\theta^2(y_{t+h|t}) + (c_{t+h|t}y_{t+h|t})^2$. Hence, our goal is to minimize $(c_{t+h|t})^2$, which in parallel narrows the quantile forecasts as in (4.7). One should note that the density forecast $f_{t+h|t}(x)$ can be considered to be a direct result of the point prediction $y_{t+h|t}$ by assuming normality and estimating the variance of the Gaussian distribution using $\mathbb{E}[e^2_{t+h|t}]$. However, one can also relax the normality assumption and construct $y_{t+h|t}$ by combining a series of quantile predictions (e.g., using the median) (Hong et al., 2016), or one could consider both prediction tasks independently. In our case, we use an intermediate approach, and estimate $f_{t+h|t}(x)$ from $y_{t+h|t}$ with some additional tuning parameters (see Section 4.5.3 for details).

Even with the best-possible prediction $y_{t+h|t} = s_{t+h}$ and a prediction interval as in (4.7), still $100\alpha\%$ of the volume measurements are expected to fall outside the interval (4.7). The intuitive *absolute coverage difference* (ACD) metric measures the absolute difference between expected and true coverage (Makridakis et al., 2020) and is similar to the kickoff percentage in, e.g., Guo et al. (2014). For example, when 89% of the measurements fall within the prediction interval, and we expect on average that 90% are within the bounds, then the ACD will be $|0.89 - 0.90| = 0.01$. Using the ACD (and other measures, e.g., score interval as in Makridakis et al. (2020)) for a single confidence level allows for carefully-crafted prediction intervals (Diebold et al., 1998; Khosravi et al., 2011). These measures are therefore more appropriate to use when, in addition to the point forecast $y_{t+h|t}$, also a full density forecast $f_{t+h|t}(x)$ is offered. In order to evaluate the density forecast, for each prediction, we additionally extract the 1st, 2nd, ..., 99th quantiles, denoted by $q_1, q_2, \ldots, q_{99}$, respectively. An integrated measure to evaluate these quantiles is the *pinball loss* function (Hong et al., 2016)

$$L(q_\alpha, x) = \begin{cases} (1 - \frac{\alpha}{100})(q_\alpha - x) & \text{if } x < q_\alpha \\ \frac{\alpha}{100}(x - q_\alpha) & \text{if } x \geq q_\alpha \end{cases}, \quad \alpha = 1, 2, \ldots, 99, \tag{4.8}$$

averaged over all target quantiles. In the remainder of this chapter, we employ the pinball

loss function (to be minimized) to evaluate and optimize predictive densities as a whole.

4.4 Temporal volume patterns

In previous section, we discussed that the estimate of the noise level depends on the systematic variation and vice versa. In this section, we show how the noise level can be inferred based on a reconstruction of the systematic variation and the other way around (Section 4.4.1). In Section 4.4.2, we introduce a method to jointly infer both at the same time. In Section 4.4.3, we give an overview of the noise levels and the extracted temporal patterns that we use for our predictions in the remainder of the chapter.

4.4.1 Noise levels and pattern recognition

In Section 4.3.3, we showed that metrics measuring the performance of a prediction method actually capture both the variance of the random variation as well as the true prediction error (see (4.5)). To disentangle these two components, we need an estimate for the noise-level function $\sigma_\theta^2(s)$.

We introduce a segmentation-based approach to estimate the volume-dependent noise-level function. Here, we partition an a posteriori estimate (i.e., reconstruction) of the systematic variation $\bar{s}$ of s into different bins Ω_k, $k \in K$, so that 'similar' data is in the same group. That is, the reconstructed volumes $\bar{s}$ of days in training set can be divided into mutually-exclusive bins, and within each bin the noise level is assumed to be constant. Although the segmentation can be based on e.g., time of day or the week (seasonal noise level) (see also Shi et al., 2014), in our approach we model the noise as being volume-dependent (see Section 4.3.3) and relate therefore the bins to the flow rate. In fact, for a given interval size $\omega > 0$, we define clusters

$$\Omega_k(\bar{s}) = \{(d,t) \in D \times T \mid \bar{s}_{d,t} \in [\omega(k-1), \omega k)\}, \quad k = 1, \dots, |K|\,.$$

Hence, Ω_k includes all time intervals for which the underlying systematic flow estimate $\bar{s}$ is in the interval $[\omega(k-1), \omega k)$. Under the assumption that for each cluster the underlying random noise process is similar, we can relate the sample volume mean

$$\mu_k(\bar{s}) = \sum_{(d,t)\in\Omega_k} \frac{1}{|\Omega_k|} \bar{s}_{d,t},$$

of cluster k to the corresponding variance of the residuals

$$\sigma_k^2(\bar{s}) = \sum_{(d,t)\in\Omega_k} \frac{1}{|\Omega_k|} (\bar{e}_{d,t})^2, \quad \text{with } \bar{e}_{d,t} = x_{d,t} - \bar{s}_{d,t}.$$

We estimate the parameters of the noise-level function by relating the sample mean and the variance, i.e., by solving the least-squares problem (see (4.4))

$$(Q(\bar{s})): \quad \min_{\theta\in\Theta} \sum_{k=1,2,\dots,|K|} \left(\sigma_k^2(\bar{s}) - \theta\mu_k(\bar{s})\right)^2.$$

Figure 4.2 shows two examples (of two different measurement locations) of the noise-level estimate for a given estimate $\bar{s}$ of s, which is estimated in Section 4.4.3. Here, a black square shows a (μ_k, σ_k^2)-combination for a cluster $k \in K$, and the solid line shows the least squares estimate $\bar{\theta}$ of θ (i.e., the optimal solution of $(Q(\bar{s}))$). The dashed line illustrates the estimate with $\bar{\theta} = 1$ (so-called Poisson noise). Our assumption having a linear noise-level function leads to quite a good fit. We noticed that by segmentation we introduced some variation of $\bar{s}$ within a bin, yet this variation is very small compared to the noise variance and is therefore neglected.

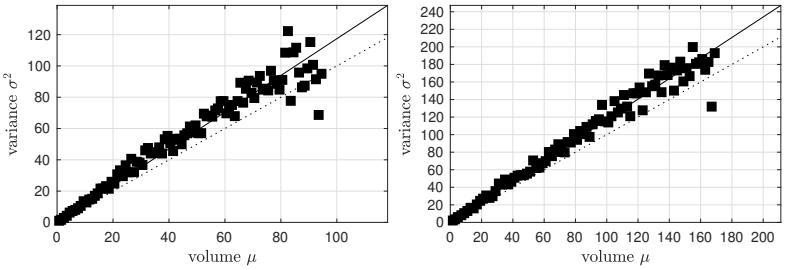


Figure 4.2: Noise-level estimates, with each square indicating the mean-variance estimate for a cluster. The solid line gives the least squares estimate, the dashed line shows the estimate in case of Poisson noise.

The noise-level fits as in Figure 4.2 are solely valuable if a majority of the systematic variation is captured. Therefore, we need a method finding a systematic variation estimate, and a criterion (independent of θ) that determines if the desired systematic flow rate is captured. Here, we discuss how to estimate the systematic volumes, while in Section 4.4.3 we discuss the criterion we employed.

Given a noise-level parameter $\bar{\theta}$, the problem to estimate the systematic volumes s explicitly accounting for the noise can be considered to be optimization problem

$$\min_{s \in \mathcal{S}} f(s, \bar{\theta}) = \frac{1}{2} \sum_{d,t} \left(\log(2\pi\sigma_{\bar{\theta}}^2(s_{d,t})) + \frac{(x_{d,t} - s_{d,t})^2}{\sigma_{\bar{\theta}}^2(s_{d,t})} \right), \tag{4.9}$$

with $\mathcal{S}$ denoting the feasible set regarding s.

As mentioned, we hypothesize that observed 24h patterns in the measurements are a combination of underlying and unobservable profiles. Many of the variations within and between days can be explained by a linear combination of these profiles. In this chapter, we apply non-negative matrix factorization (NMF) as a framework to find 24h profiles and thereby estimate s in (4.9).

Let us assume for now that we know that we need m profiles to capture the 24h systematic variations. Denote the shapes of the (normalized) profiles by matrix $W \in \mathbb{R}_+^{|T| \times m}$, i.e., a profile covers 24h and consists only of non-negative elements. The corresponding day-dependent magnitudes $H \in \mathbb{R}_+^{m \times |D|}$ express the scaling of each profile. Together we have $s = WH \in \mathbb{R}^{|T| \times |D|}$ as estimate for the systematic variation. Having m as estimate of the number of 24h profiles, we can apply NMF within the optimization framework, by solving (4.9) using $s = WH$. In Section 4.4.3, we discuss a criterion to find an estimate for

m.

4.4.2 Joint estimate of noise levels and patterns

To estimate $\bar{s}$ and $\bar{\theta}$, one could follow an iterative approach. First, an initial estimate θ^0 of noise-level parameter θ is required to solve (4.9), which leads in return to an estimate s^1 of the systematic flow rate. By solving $(Q(s^1))$, one obtains a new estimate θ^1 of θ. However, such an iterative approach does not anticipate the changes in the noise level while finding a new estimate of the systematic variation. Here, we propose an alternative optimization-based approach that jointly estimates the temporal volume patterns as well as the parameter of the noise-level function for each location in isolation.

We formulate an optimization problem to find patterns $\bar{s} \in \mathbb{R}^n$ and noise-level parameters $\bar{\theta} \in \mathbb{R}^k_+$. Note that, without loss of generality, we assume s to be in vector form rather than in matrix form. Also notice that $\theta \in \mathbb{R}_+$ (i.e., $k = 1$) is a variable in this framework. In a general form, the optimization problem can be written as (Aravkin & Van Leeuwen, 2012)

$$(P): \quad \min_{s \in \mathcal{S}, \theta \in \Theta} f(s, \theta),$$

where $f : \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}$ is the reconstruction error between measurements x and systematic variation estimate s, depending on the noise-level function parameterized by θ. $\mathcal{S}$ is the feasible set regarding s, and Θ is the set resulting from the box constraints restricting the choice of θ. Typical examples of $f(s, \theta)$ include the (negative) log likelihood of x given s and θ (as in (4.9)), or the sum of squared residuals.

In iterative approaches, one finds noise parameters $\bar{\theta}$ based on an estimate $\bar{s}$ of s. Here, $\bar{\theta}$ is the optimal solution of a parametric optimization problem, i.e., $\bar{\theta}$ results from an estimate of s and in that sense should explicitly depend on it. Formally, $\bar{\theta}$ is the optimal solution corresponding to

$$(\tilde{Q}(s)): \quad \min_{\theta} g(s, \theta), \quad \text{with parameter } s = \bar{s}.$$

Here, $g : \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}$, and is, e.g., a least squares estimate between the volumes s and a variance estimate as in $(Q(s))$. However, in our case the least-squares solution to $(\tilde{Q}(s))$ is unique for all fixed $\bar{s}$, and it follows that the corresponding optimal solution θ is an *implicit function* of s, i.e., $\theta : \mathbb{R}^n \to \mathbb{R}^k_+$ with

$$\theta(s) = \{\theta \mid \theta \text{ is a global minimizer of } \tilde{Q}(s)\}.$$

This allows us to rewrite problem (P) as an optimization problem only in s, since θ directly follows from optimization problem $(\tilde{Q}(s))$. Indeed, we solve

$$(P'): \quad \min_{s \in \mathcal{S}} f(s, \theta(s)),$$

rather than (P). This reformulation (P') is shown to have some advantages compared to (P) (Bell et al., 1996). Under some additional assumptions $\theta(s)$ is continuously differentiable in s (Klatte & Kummer, 2006), and sometimes even a closed-form expression is possible (Aravkin & Van Leeuwen, 2012). However, under more general assumptions, $\theta(s)$ is typically not continuously differentiable everywhere.

For the remainder of this section, we rewrite (P') as a bilevel optimization problem, i.e.,

$$(P'') : \quad \min_{s \in \mathcal{S}} f(s, \theta) \quad \textit{s.t.} \quad \theta \text{ solves } (\tilde{Q}(s)).$$

We have that $(\tilde{Q}(s))$ is a convex optimization problem with linear (box) constraints (see $(Q(s))$). Under these assumptions, the *Karush-Kuhn-Tucker* (KKT) conditions are both necessary and sufficient for optimality and we can replace the difficult constraint in (P'') by its system of KKT equations, abstractly written as

$$H(\theta, s) = 0,$$

with $H : \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}^m$ being a C^1-function (Klatte & Kummer, 2006; Still, 2018).

The *implicit function theorem* (IFT) states under additional conditions the following, for some reference point $(\bar{\theta}, \bar{s})$ with $\bar{\theta}$ solving $(Q(\bar{s}))$. If $H(\bar{\theta}, \bar{s}) = 0$, and $\nabla_\theta H(\bar{\theta}, \bar{s})$ is non-singular, there exists a neighborhood in s around $\bar{s}$, and a C^1-function $\theta(s)$ for which $\theta(\bar{s}) = \bar{\theta}$ so that $H(\theta(s), s) = 0$. Moreover, in this neighborhood in s around $\bar{s}$, $\nabla_s \theta(s)$ exists and is given by

$$\nabla_s \theta(s) = -(\nabla_s H(\theta, s))^{-1} \nabla_\theta H(\theta, s).$$

Intuitively, the IFT says that the noise-level parameter θ is a locally continuous function of the systematic variation s around $\bar{s}$, and, in addition, is continuously differentiable in this neighborhood. Now, we can use this information for solving (P') and thereby jointly find the noise-level parameter as well as the patterns in the measurements.

This framework does not directly apply to finding the patterns and the noise level as in Section 4.4.1. At some points in the domain, the IFT conditions might not hold and $\theta(s)$ is shown to be only *piecewise smooth* at some s (Dempe & Vogel, 2001). In addition, (4.9) is not defined everywhere, and we replace therefore $\sigma^2_\theta(s)$ with $\tilde{\sigma}^2_\theta(s) = \max\{3, \theta \cdot s\}$ to prevent the estimate to be overly sensitive to fluctuations during very low-volume occasions (e.g., at night). We ignore possible nonsmoothness and apply a Gauss-Newton method with inexact line search to solve the problem at hand.

We found that the solution θ corresponding to $Q(s)$ is highly influenced by outliers, and we use therefore a sigma-clipping approach to diagnose and remove outliers while estimating the noise level. Visually, we should mark measurements as outliers if they are either 3 to 4 standard deviations $\sqrt{\sigma^2(s)}$ away from s, or two consecutive measurements having a distance of 2 to 3 standard deviations. Hence, we applied a 3.5 and 2.5-sigma criterion, respectively. Obviously, outliers and clusters depend on estimate $\bar{s}$ of s and are therefore re-identified after updating $\bar{s}$.

4.4.3 Extracted patterns and estimated noise levels

The optimization framework of Section 4.4.2 is applied in conjunction with the noise-level estimation and NMF framework of Section 4.4.1. This resulting optimization problem can be applied to find m profiles assuming to describe the underlying systematic variation for a single measurement location, together with an accompanying noise-level function. A remaining yet essential question to answer is the number of profiles needed to capture the systematic volumes without fitting noise. After all, m is is not known beforehand.

Here, we use an iterative approach and begin with a single (unknown) profile and solve (P'). Then, we add profiles until a termination criterion, diagnosing if on average all systematic variations are captured, is met. Where the noise is assumed to be uncorrelated over successive increments (see Section 4.3.3), the reconstruction does not capture all slowly-changing systematic variations if there is still correlation in successive residuals (serial correlation). In fact, we add 24h profiles until on average no correlation in the successive residuals is left. We iteratively add profiles $1, \dots, m$ (and solve the corresponding (P')) until the two-sample Kolmogorov-Smirnov (KS) test indicates that the daily-average serial correlation in the reconstruction using m profiles is not a significant improvement compared to the correlation with $m-1$ profiles. Then, we decide on $m-1$ profiles. We found that a significance level $\alpha = 0.001$ is suitable in our case.

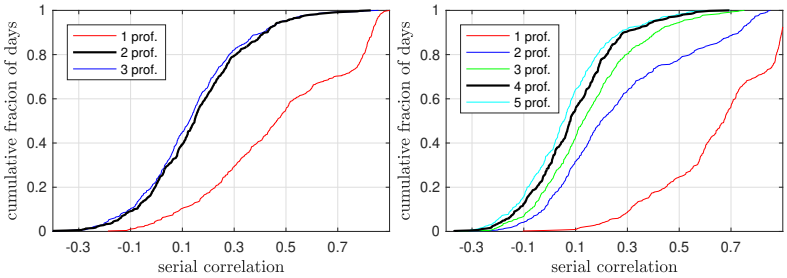


Figure 4.3: Distribution function of the remaining serial correlation of the residuals over the days for reconstructions using different number of profiles. The black-colored line gives the distribution function for the final number of profiles.

Figure 4.3 shows two examples of measurement locations where we decide on 2 and 4 profiles, respectively, since the reconstruction with an additional profile does not substantially improve the distribution of the serial correlation in residuals (according to the KS test). The corresponding noise-level estimates are given in Figure 4.2. Similarly, we applied this method to extract 24h patterns and the noise levels for all measurement locations. On average, the random variation has slight overdispersion relative to Poisson noise (i.e., $\mathrm{Var}(\varepsilon|s) > \mathbb{E}[s]$). The variation in the noise-level parameter between the different locations under consideration is depicted in Figure 4.4.

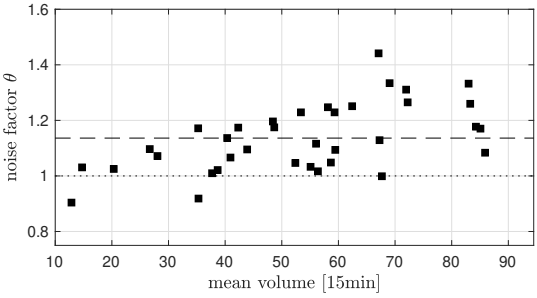


Figure 4.4: Noise-level parameter estimates for the different measurement locations as function of the mean 15min volume.

Figure 4.5 shows examples of reconstructions obtained using the 24h profiles for three different measurement locations. Interestingly, with only limited degrees of freedom (in this case, 2 or 3 profiles with a day-dependent magnitude per profile), we can capture a substantial share of all the systematic variations. In fact, these profiles are able to capture the different intra-day patterns as well as long-term variations by day-to-day adaptations of the underlying patterns (i.e., inter-day variations).

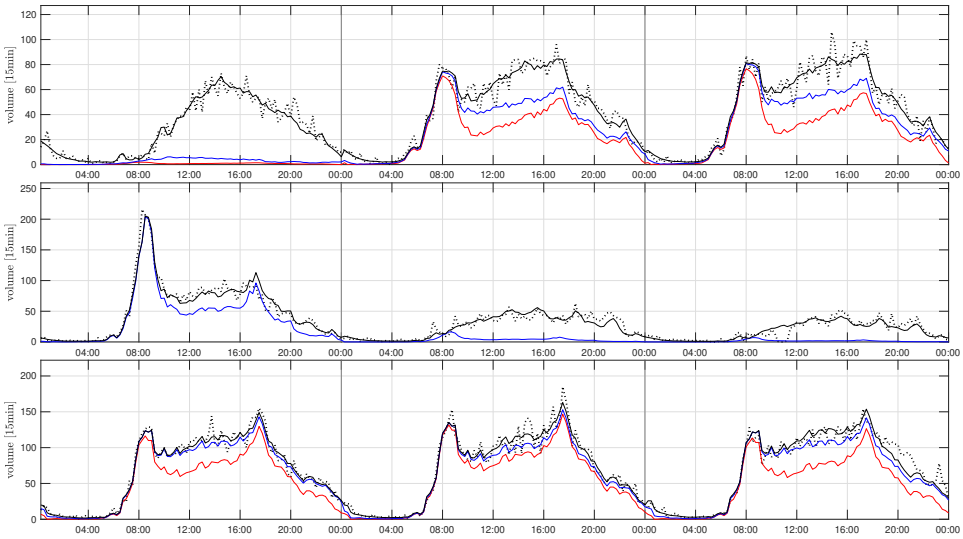


Figure 4.5: For three days, measurements (dashed) and reconstruction of the systematic variation (black). The systematic variability uses the profiles with day-dependent magnitudes as building blocks (cumulative volume illustrated by the different colors).

We found short-term patterns related to football matches and other recurrent events as follows. Based on the days in the training set with an event occurring during that day, we use the long-term reconstruction as obtained by the long-term profiles and fitted a Gaussian curve on the daily residuals. Although the measurements do not necessarily follow a Gaussian curve, it occurs to be quite a good fit. As a result, we have a short-term pattern describing the change in volume (compared to the 24h reconstruction) due to a recurrent event. Figure 4.6 shows examples of time series of two measurement locations where the Gaussian curve expresses the additional volumes due to traffic to and from the stadium. In fact, Figure 4.6 shows the same dates (with a FC Twente football match). Where the time series in the upper row show additional traffic before the match, the series in the lower row indicate more traffic after the match. We find that the occurrence of this curve is highly related to the (varying) kick-off time of the matches. Hence, these short-term profiles are incorporated in the forecasts as well. In the next sections, we use these profiles and the noise level for predictions.

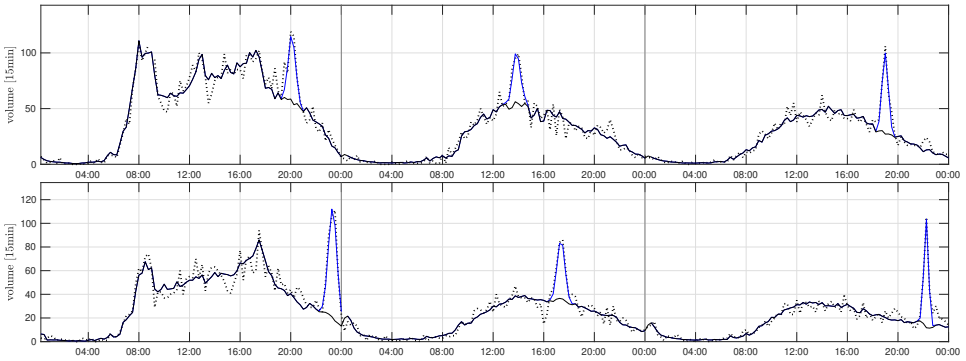


Figure 4.6: Volume measurements (dashed) for three days with football matches, and reconstruction using the long-term profiles (black) and a short-term profile (blue).

4.5 Baseline and 24h prediction

The 24h prediction provides an estimate of the upcoming-day volumes before the start of the day. This prediction is an update of the baseline forecast: a typical 24h pattern based on *predictable features* accompanying a day, i.e., characteristics of the day that can be identified a long time in advance using straightforward exogenous variables.

We construct the final 24h prediction in two steps. First, we construct a baseline forecast using the predictable features. Then, we update this baseline prediction based on the realized difference between the measurements and the baseline predictions of the previous days. Rather than predicting the resulting volumes directly, we predict the flow rate by means of forecasting the scaling magnitude of each profile as found in previous section. In the remainder of this section, we first discuss the baseline (Section 4.5.1) and the 24h point prediction (Section 4.5.2). In Section 4.5.3, we discuss the estimate of the density function forecast.

4.5.1 The baseline prediction

The baseline forecast is constructed based on day-dependent features. Indeed, there is a large share of variation from day to day, and we therefore group days based on assigned labels to the dates. We used the following labels to differentiate between varying 24h patterns from day to day: day of the week (Monday - Sunday), shopping Sunday, and (school) holiday period. Since various labels might be assigned to a single day, we identified 15 mutually exclusive groups of days: Monday - Saturday outside school holiday period, Monday - Saturday during school holiday period, (regular) Sunday outside school holiday period, Sunday during school holiday period, and shopping Sunday. An alternative method uses features extracted from the data, e.g., as in Laña et al. (2019), but here we use straightforward features, independent of the time series and location, that can easily be predicted in advance.

Consider a group of days G in the training set, i.e., $G \subseteq D$. In Section 4.4, we reconstructed the systematic volumes of all training days with the use of m underlying profiles $w_1, w_2, \ldots, w_m$ (the column vectors of W), with corresponding day-dependent magnitude vectors $h_1, \ldots, h_{|G|}$ of $H_{\cdot,G}$ (i.e., the columns of H corresponding to the days in G with

the scaling magnitude for each profile during these days). For a single group, the baseline prediction s^{bl} gives a typical 24h pattern. This prediction is constructed by means of the scaling magnitudes, i.e., $s^{bl} = Wh^{bl}$. Here, h^{bl} is the vector $h \in \mathbb{R}_+^m$ that solves (4.9) given W and noise-level function $\sigma_\theta^2(s)$ - restricting ourselves to the training set and the days in group G. We repeat this process for all identified groups thereby constructing a typical daily pattern for each group of days.

In Figure 4.7, we show examples (in red) of the baseline reconstruction of days in the training set, and find that this baseline forecast incorporates location-dependent characteristics including day-to-day variation in the shape of the 24h volumes. For example, the most right time series in the lower row of figures shows the recurrent Wednesday-noon peak, which does not occur on the preceding weekdays.

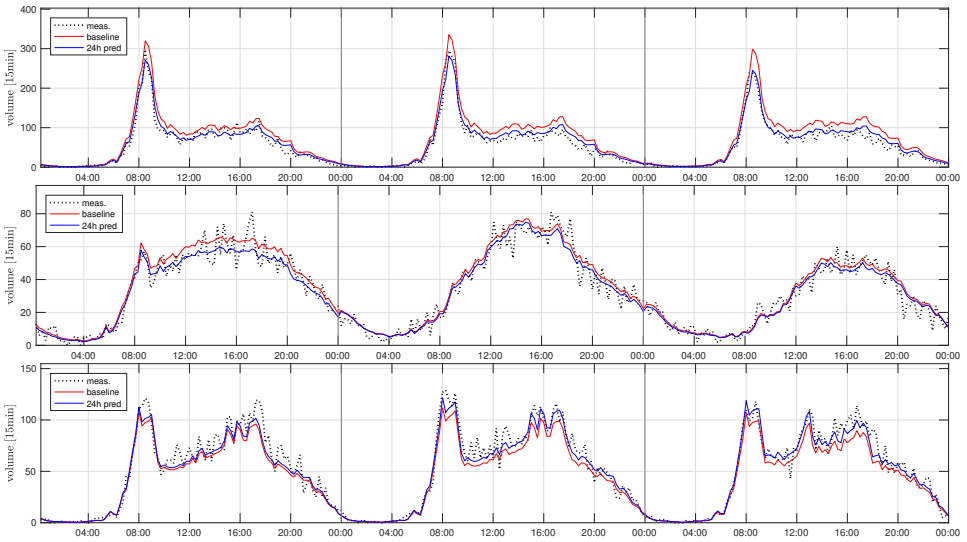


Figure 4.7: Example of baseline (red) and 24h (blue) point predictions for three consecutive days, with measurements (dashed line).

4.5.2 24h prediction

Although the baseline provides quite an accurate forecast regarding the shape of the 24h volumes during regular days, we found that it can be improved in two ways. First, the baseline prediction does not incorporate recurrent events. Second, we found that under regular conditions the measurements show an almost constant relative error over the days compared to the prediction. These deviations are consistent in the sense that these variations occur on timescales longer than 24h (i.e., slowly changing), and thus could be accounted for when making a long-term prediction.

During events, measurements do not follow the baseline for a shorter period of time. We have reliable information regarding the occurrence of football matches and university-related events (e.g., open days). We observe that in particular the football matches influence volumes at some locations, typically leading to a substantial increase in flows before or after the match (see Figure 4.6). We also found other days with comparable short-term variations,

but we could not identify the underlying event that caused this. Hence, in the 24h forecast we only account for short-term variations due to the football matches and university-related events as follows. In the previous section, we showed that a Gaussian curve approximates the short-term deviations compared to the long-term pattern. Although some variability in the short-term patterns is observed, they are relatively consistent over the different events when accounting for varying starting and ending times. Two Gaussian curves are therefore used to predict short-term variation; one for the volume changes due to FC Twente matches, and one for changes due to university-related events. The corresponding parameters (scale, time of occurrence relative to starting time, and width) are the average over the individual fits in the training set. The resulting curve is then the short-term pattern, which is included in the 24h prediction (see Figure 4.8 for an example).

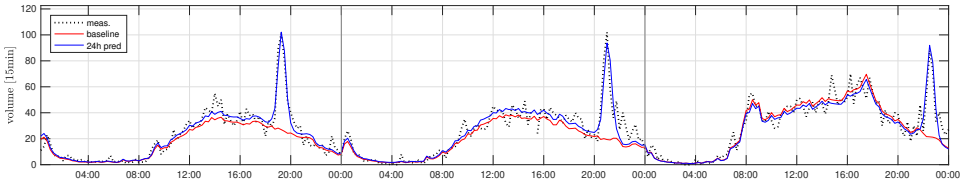


Figure 4.8: Example of baseline (red) and 24h (blue) point predictions during days with FC Twente matches, with measurements (dashed line).

We update our baseline prediction to arrive at a 24h prediction by using the residuals as follows (see also Thomas et al., 2010). We found that the relative daily residual

$$\Delta_d^{bl} = \frac{\sum_{t\in T}(x_{d,t} - s_{d,t}^{bl})}{\sum_{t\in T} s_{d,t}^{bl}},$$

at day d, is highly correlated with the relative daily residual Δ_{d-1}^{bl} of the previous day. Now we construct the 24h prediction of the scaling magnitude h^{24} of profiles $i = 1, 2, \ldots, m$ at day d so that

$$h_{i,d}^{24} = h_{i,d}^{bl}(1 + \beta_1\Delta_{d-1}^{bl} + \beta_2\Delta_{d-2}^{bl}), \tag{4.10}$$

with β_j, $j = 1, 2$, location-dependent parameters that are optimized using the training set with the objective function of (4.9). Here, we note that for a weekday d, $d-1$ and $d-2$ refer to previous most recent weekdays, while for days in the weekend we take into account the daily residuals at the previous weekend days. We note that Δ is not very sensitive to noise, since we aggregate the residuals over a complete day.

After accounting for the slowly-changing demand, we also update our prediction of the individual 24h profiles. Indeed, at some locations, the residuals in the scaling magnitudes show strong serial correlation over the days. Here, for each profile, we use a similar mechanism as in (4.10), but only take into account the absolute residual in scaling magnitude of the previous day. Based on this method, we have a matrix H^{24} denoting the 24h prediction of the scaling magnitude of each profile. Since the profiles cover 24h, the long-term volume point predictions are given by $y^{24} = WH^{24}$. In addition, we have a prediction of the change in volume (compared to the 24h prediction) due to recurrent events. This is also included in y^{24}.

Figure 4.7 shows examples of the 24h prediction (in blue) on the training set, including

a comparison to the baseline prediction. We see that the method is adaptive in the sense that the prediction is improved over time based on previous prediction errors. At the same time, the update of the baseline is slowly changing over the days, i.e., comparable scaling is used for the consecutive days, although these corrections might change for weekdays compared to weekends. Figure 4.8 shows time series illustrating the 24h prediction during days with events. Here, including recurrent events improves the prediction, and the forecast quite accurately covers the substantial deviations compared to the long-term pattern.

4.5.3 From point prediction to density prediction

Following Section 4.3.3, the error of the 24h point prediction on the training set is used to provide a 24h density function forecast. In fact, let us denote by $c^{24}_{d,t}$ the relative prediction error (here: reconstruction error) of the 24h reconstruction based on the training set. We find that this relative error is related to the time of day. Using the law of total expectation, we have an expected squared error

$$\begin{aligned}\mathbb{E}[(y^{24}_{d,t} - X_{d,t})^2] &= \mathbb{E}_s[\mathbb{E}[(\tau^{24}_{d,t} + \varepsilon_{d,t})^2 | s_{d,t}]] \\ &= \mathbb{E}_s[\theta s_{d,t} + (c^{24}_{d,t} s_{d,t})^2].\end{aligned}$$

Hence, we estimate the time-of-day prediction error by comparing the mean squared error of the residuals minus the noise sample variance with the mean quadratic volumes, i.e.,

$$(c^{24}_t)^2 = \frac{\frac{1}{|D|}\sum_{d\in D}(x_{d,t} - y^{24}_{d,t})^2 - \frac{1}{|D|}\theta\sum_{d\in D} y^{24}_{d,t}}{\frac{1}{|D|}\sum_{d\in D}(y^{24}_{d,t})^2}. \tag{4.11}$$

Interestingly, assuming a constant relative error over time, large volumes dominate the mean squared error. Figure 4.9 shows an example of the relation between the mean squared error of the 24h residuals and the corresponding error for days in the training set. Where the mean squared error is large for higher volumes, the relative error as in (4.11) shows that predictions are actually worse during the night. In addition, days of the weekend are typically more difficult to predict and the relative error is 50% larger compared to weekdays. Therefore, we estimate independently a time-of-day prediction error for weekdays and weekends.

Based on the relative error, we construct a first estimate of the density predictions $f^{24}(x)$ with

$$f^{24}_{d,t}(x) = \phi(x; y^{24}_{d,t}, \sigma^2_\theta(y^{24}_{d,t}) + (c^{24}_t y^{24}_{d,t})^2).$$

The corresponding quantile predictions $q_\alpha, \alpha = 1, 2, \ldots, 99$, then follow from the density function additionally taking into account that flows are non-negative, i.e., $q_\alpha \geq 0$. We can improve this density forecast by introducing additional degrees of freedom and allowing θ and c^{24}_t to be optimized. However, (4.11) reveals a direct relation between θ and c. For the sake of improving the density forecast, we add θ as an additional degree of freedom for the sake of the prediction, and use the resulting c from (4.11) so that the relative pinball loss is minimized, i.e., the solution of optimization problem

$$\min_{\theta\geq 0} \frac{1}{|T|}\sum_{t\in T} \frac{\frac{1}{|D|}\frac{1}{99}\sum_{d\in D,\alpha=1,2,\ldots,99} L_{d,t}(q_\alpha, x)}{\frac{1}{|D|}\sum_{d\in D} y^{24}_{d,t}}$$

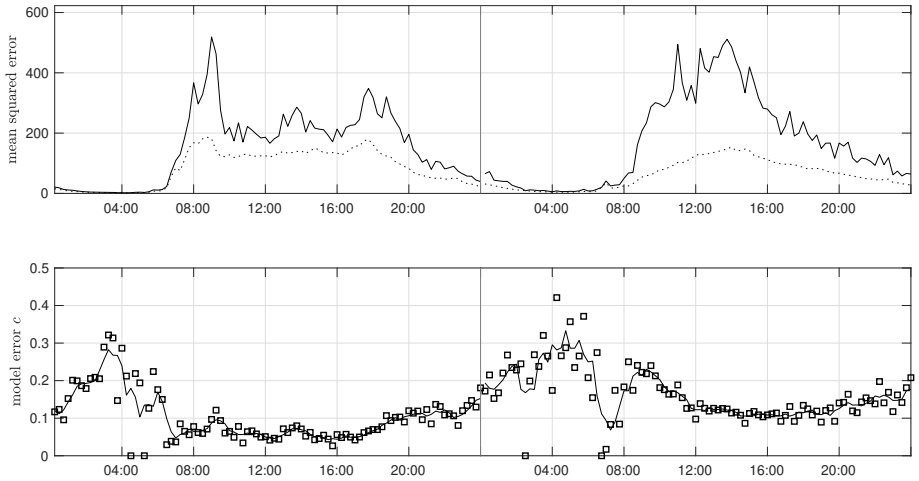


Figure 4.9: Relation between mean squared error (solid line in top row of figures) and relative error (bottom figures) for both weekdays (left figures) and weekends (right figures) at a high-volume location. In the top figures, the dashed line shows the mean 15min volume. In the bottom figures we additionally show the (1h) moving mean of the relative error.

with L as in (4.8). Then, $f^{24}(x)$ accounts for both the model and prediction error as well as the random variation in the volumes. Figure 4.10 provides examples of the 90% prediction interval derived from the 24h density function forecast. We note that the 90% prediction interval is only used for illustration purposes. The relative width of the prediction interval mainly differs due to the difference in prediction error. We assess the quality of the 24h point and density prediction on the test set in Section 4.7.

4.6 Remaining-day and short-term prediction

Over the course of the day, we are able to adapt the prediction since the initial 24h forecast can be compared with the recently realized volumes. Adaptive predictions over shorter time horizons are difficult to construct since (i) measurements are noisy making it difficult to distinguish random from systematic errors particularly when only a few measurements are available, (ii) during the night prediction errors increase and volumes can wrongly indicate that the remainder of the 24h prediction is off, and (iii) the 24h prediction accounts for day(-of-week)-dependent characteristics which might or might not need to be maintained when updating.

In this section, we use state-space smoothing to relate recent volume measurements (space) to the underlying profile magnitudes (state), while explicitly accounting for volume-dependent noise and prediction errors (Section 4.6.1). This smoothing method is used to update our point and density forecast by adapting the scaling magnitudes of the profiles throughout the day (Section 4.6.2). In Section 4.6.3, we use similar techniques for forecasts in the short term.

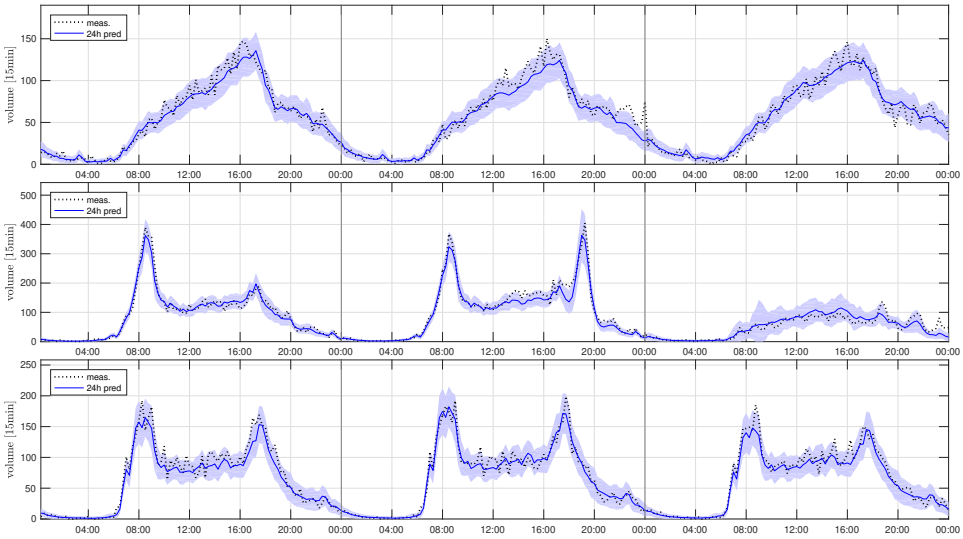


Figure 4.10: Examples of 24h point predictions and 90% prediction intervals for three different measurement locations at three consecutive days, with measurements (dashed line).

4.6.1 State-space smoothing

Recent measurements used to adapt the remaining-day prediction are highly influenced by noise, since aggregation (as used for the 24h forecast) is only partly possible and desirable. We use state-space smoothing to denoise measurements by obtaining an estimate of the systematic flow in a continuous fashion, and at the same time infer the underlying scaling magnitudes of the profiles. Many traffic prediction methods (e.g., Wang & Papageorgiou, 2005) make use of a filtering or smoothing technique, most notably the linear Kalman filter and Rauch-Tung-Striebel smoother (Haykin, 2004; Kalman, 1960). However, this widely-adopted method is rather restrictive in our case since it does neither account for volume-dependent noise nor for constraints. We cite Huang et al. (2018), who incorporated the time-of-day dependent noise variance using a rolling horizon framework combined with a Kalman filter approach, but independent of the underlying systematic flow estimate. Since we assume to have non-negative scaling magnitudes h (demand is non-negative), and the noise variance is modeled to be a linear function of the volume, we use a nonlinear smoother that does account for these two complicating phenomena.

In the previous section, we predicted the counts throughout a day using a single scaling magnitude h^{24} per profile. In fact, the scaling magnitude together with the profiles W provided the 24h prediction. Here, we expand our state and let the profile-scaling magnitudes to be variable throughout the day. With abuse of notation, we denote the magnitude vector at time t by h_t, for which we already have a 24h prediction, denoted by h_t^{24}. Since the profiles are fixed, we have a linear system and can adopt the constrained Kalman-Bucy model

$$\begin{aligned} \eta_t &= \eta_{t-1} + u_t, \\ x_t &= W_{t,\cdot}(h_t^{24} + \eta_t) + \varepsilon_t, \\ h_t &= \eta_t + h_t^{24} \geq 0. \end{aligned} \tag{4.12}$$

Here, $\eta_t \in \mathbb{R}^m$ is the unknown state vector measuring the deviation in profile magnitude compared to the estimate h_t^{24}. $u_t \in \mathbb{R}^m$ is the noise in the underlying state dynamics, $x_t \in \mathbb{R}_+$ is the measurement at time t, and $\varepsilon_t \in \mathbb{R}$ is the random volume-dependent error in the measurements. The noise vectors are assumed to be i.i.d. and Gaussian with

$$u_t \sim \mathcal{N}(0, U_t), \quad \text{and} \quad \varepsilon_t | s_t \sim \mathcal{N}(0, \sigma_\theta^2(s_t)),$$

with $s_t = W_{t,\cdot} h_t$, so that the measurement noise is a function of the systematic variation (see Section 4.3.3). Using the model in (4.12), we assume that the scaling magnitude $h_t = \eta_t + h_t^{24}$ is slowly changing over time. In fact, the majority of the variation in the volumes is expressed by the 24h volume profiles W.

Before adapting our prediction, we are concerned with *state estimation* at time instant N. That is, we infer an estimate of the underlying state $\hat{\eta}_{t|N}$ of η_t (and thereby also flow estimate $\hat{y}_{t|N}$ of x_t) at time N. If $t < N$, we are concerned with *smoothing*, and we can use all information up to time instant N for the estimation. If $t = N$, we have a *filtering* problem (Aravkin et al., 2017; Haykin, 2004).

We follow arguments and notations from Aravkin et al. (2017) and Bell et al. (2009). For a state sequence denoted by $\{\eta_t\}_{t=1}^N$, the negative log likelihood of the (normal) density for the measurements $\{x_t\}_{t=1}^N$ given $\{\eta_t\}_{t=1}^N$ is

$$-\log p(\{x_t\}_{t=1}^N | \{\eta_t\}_{t=1}^N) = \\ \frac{1}{2} \sum_{t=1}^N \log\det\left(2\pi\sigma^2(s_t)\right) + (x_t - W_{t,\cdot}(h_t))^T (\sigma^2(s_t))^{-1} (x_t - W_t(h_t)).$$

Before retrieving the measurements, the negative log of the probability density of the state sequence is given by

$$-\log p(\{\eta_t\}_{t=1}^N) = \frac{1}{2} \sum_{t=1}^N \log\det\left(2\pi U_t\right) + (\eta_t - \eta_{t-1})^T U_t^{-1} (\eta_t - \eta_{t-1}).$$

Since $p(\{x_t\}_{t=1}^N, \{\eta_t\}_{t=1}^N) = p(\{x_t\}_{t=1}^N | \{\eta_t\}_{t=1}^N) p(\{\eta_t\}_{t=1}^N)$. The optimization problem to infer estimate $\{\hat{\eta}_{t|N}\}_{t=1}^N$ at time N based on the measurement sequence becomes

$$(R_N): \min_{\{\eta_t\}_{t=1}^N} -\log p(\{x_t\}_{t=1}^N | \{\eta_t\}_{t=1}^N) - \log p(\{\eta_t\}_{t=1}^N) \quad \textit{s.t.} \quad h_t \geq 0, t = 1, \ldots, N.$$

As shown by Aravkin et al. (2017), under volume-independent noise and without the linear constraint, the problem (R_N) can be solved using the Rauch-Tung-Striebel scheme, with $\hat{\eta}_{N|N}$, part of the optimal solution of (R_N), the (Kalman) filtered state. However, the general formulation (R_N) can be applied to our problem at hand with linear constraints and signal-dependent random variation.

We use a similar framework as in Aravkin et al. (2017), and solve (R_N) by solving a sequence of quadratic sub-problems. For every quadratic problem, we fix $\sigma^2(s_t)$ based on the estimate of s_t resulting from the previous subproblem. Moreover, since (R_N) is an optimization problem essentially growing in time, we only use the most recent 24h measurements when solving the program. We found that this smoothing method is highly sensitive to new measurements supposedly being outliers. Therefore, we use a 4-sigma clipping cri-

terion to exclude outliers in the last hour while solving (R_N). When N increases over time, these initially considered outliers are again incorporated.

Before applying this method, we need an estimate of the error covariance U_t. The estimate for the resulting model error covariance matrix is based on the training set. The (relative) model error is denoted by b. We apply this relative model error independently to the scaling magnitudes $h_1, h_2, \ldots, h_m$, thereby having higher absolute errors when profile volumes increase. Since the relative error is time-of-day dependent (see Section 4.5.3), we scale the model error to arrive at

$$U_t = (bc_t^{24})^2 \cdot \mathrm{diag}(\{(h_{i,t}^{24})^2\}_{i=1,2,\ldots,m}\}),$$

with (location-dependent) c_t^{24} as in (4.11). Visual inspection showed that the smoothing method shows reasonable performance with $b = 0.03$. Here, the model error is both volume and time(-of-day) dependent. Again, we observe that the relative model error is typically larger during the night, which means that the smoothing method should follow the measurements more closely during these intervals compared to, e.g., the between-peak period (see also Figure 4.9).

Figure 4.11 shows time series examples of days where the smoothing method improved the systematic volume estimates by solving (R_N). As illustrated, the state-space smoothing method is able to obtain an visually-accurate estimate of the systematic variation by independently adapting the scaling magnitudes of the underlying 24h volume profiles.

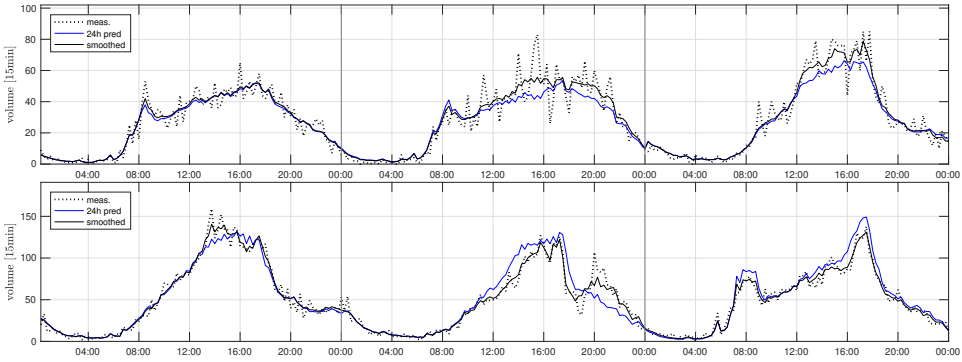


Figure 4.11: Measurements (dashed) for three days and smoothed reconstruction of the systematic variation (black) and the 24h prediction (blue) .

When events occur, performance of the smoothing method deteriorates since the underlying state equation in (4.12) does then not mirror the actual dynamics. In fact, the state-space model in its current form only accounts for long-term profiles. To allow for both long- and short-term systematic variations, in parallel to solving (R_N) we solve a similar problem with an extended space model

$$x_t = W_{t,\cdot}(h_t^{24} + \eta_t) + V_t + \varepsilon_t,$$

where V_t corresponds to the volumes of the short-term pattern (i.e., following the Gaussian function as estimated). We switch to this alternative space model if we expect a short-term pattern to be present (e.g., when football matches occur), or when a substantial improve-

ment compared to the original problem's state estimate can be obtained. In order to check for improvement, we apply a χ^2-test on the likelihood ratios. After accepting the alternative space model, we adapt the parameters of the Gaussian curve so that we minimize the negative log likelihood. Initial tests on the training set with the alternative space model showed that continuously updating V_t often decreases performance if only a very small share of the short-term deviations is revealed. Therefore, we solely adapt the parameters corresponding to V_t after a substantial share of the short-term pattern is revealed. In Figure 4.12, we show three examples of smoothed measurements in which we switched to the alternative space model, with the most-right time series an example of a day with an event that was initially not accounted for.

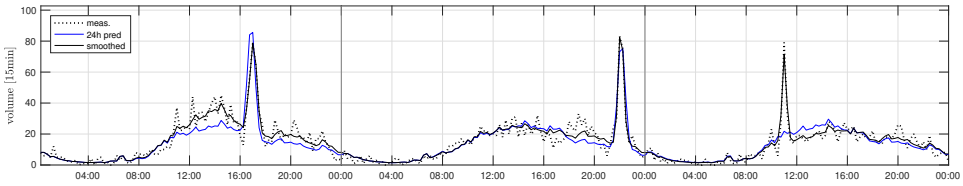


Figure 4.12: Measurements (dashed) for three (non-consecutive) days and smoothed reconstruction of the systematic variation including short-term patterns (black) and the 24h prediction (blue) .

4.6.2 Remaining-day point and density prediction

Systematic prediction errors in the 24h forecast are often correlated over time, and we hypothesize that a major share of the systematic error is due to a short or longer-term yet temporarily changing demand compared to what was expected. In this section, we update the volume forecast for the remaining day by comparing smoothed volume sequence $\{\hat{y}_t\}_{t=1}^{N}$ with the initial 24h prediction $\{y_t^{24}\}_{t=1}^{N}$. The predictions for the short-term are considered in Section 4.6.3.

We adopt a (point-)adaptive updating mechanism, i.e., the prediction is updated based on the errors in the point estimate during previous time intervals. Consider time N at day d. Just before the measurement $x_{d,N}$ is processed, we assume to have a current prediction of the 24h volumes, $\{\bar{y}_{t|N-1}\}_{t\in T}$, i.e., a full-day estimate based on the measurements before N. We compare the filtered volumes with the prediction, and adopt the recursive updating mechanism

$$\bar{y}_{t|N} = \bar{y}_{t|N-1}\left(\frac{\sum_{k=0,\dots,3}\max\{3,\hat{y}_{N-k|N}\}}{\sum_{k=0,\dots,3}\max\{3,\bar{y}_{N-k|N-1}\}}\right)^{\gamma}, \quad t\in T \tag{4.13}$$

with $\gamma > 0$ a location-dependent updating factor. In (4.13), we adapt our complete underlying 24h estimate (even for a few time intervals in the past) based on the ratio between the smoothed and the expected volumes in the last hour. In this way, we quickly adapt the prediction to a changing demand while maintaining day-dependent characteristics. We included the $\max$ operator in (4.13) to prevent being sensitive to changes in low-volume estimates. Preliminary tests show that it is difficult to distinguish noise from slowly-changing yet systematic changes in the demand. To make the prediction more robust against noise,

we combine predictions and therefore use a combination of the initial 24h prediction and the updated prediction (4.13). In fact, the final remaining-day prediction $\{y^{rd}_{t|N}\}^{T}_{t=N+1}$ is a linear combination of both so that

$$y^{rd}_{t|N} = \psi_N \bar{y}_{t|N} + (1-\psi_N) y^{24}_{t|N}, \quad t = N+1, \ldots, |T|\,, \tag{4.14}$$

where $0 \leq \psi_N \leq 1$ is the weighting factor. With more substantial deviations in the demand we aim for $\psi_N \to 1$, while small changes should lead to $\psi_N \to 0$. To continuously balance between the two predictions, we employ a bell-shaped membership function, i.e.,

$$\psi_N = S\left(\frac{\sum \bar{y}_{N-k,\ldots,N|N}}{\sum y^{24}_{N-k,\ldots,N|N}}\right), \quad \text{with } S(x) = 2 \cdot \min\left\{\frac{1}{2}, \frac{1}{1+\left|\frac{x-1}{a}\right|^{2b}}\right\}.$$

We determine the parameters a and b of the membership function $S(x)$ based on the training set. In addition, changes during the night are shown to be substantial yet often unrelated to deviations in demand during the rest of the day. Therefore, we let $\psi = 0$ for the first part of the day - again so that the prediction is optimized using the training set.

To arrive at a sequence of remaining-day density function predictions $\{f^{rd}_{t|N}(x)\}^{|T|}_{t=N+1}$ at time N, we use the method as outlined in Section 4.5.3. We tune the parameters corresponding to the density function independently for each horizon v, thereby accounting for the model (estimation) error and the noise as well as for the horizon-dependent prediction error.

Figure 4.13 shows three examples of 24h time series predictions, and for each day three remaining-day predictions with the accompanying 90% prediction intervals based on the density forecasts. Here, we see the adaptive but also robust behavior of the method. The initial 24h prediction is in the first two cases underestimating volumes for a share of the day. Throughout the day, we adapt the remaining-day forecast. In the top figure, the increased demand covers almost the whole day, and the prediction is updated accordingly. In the middle row, the increase in demand was actually short-term in nature but covered several hours. Hence, the remaining-day forecast is initially predicting an increase in demand for the remainder of the day, but adapted again. The lower set of time series shows that the remaining-day prediction is not very sensitive to sudden but temporary shocks in the flow rate. Note that the width of the 90% prediction interval is rather robust, i.e., even for longer horizons the intervals are relatively small.

4.6.3 Short-term prediction

Thus far, we concentrated on the long-term and the remaining-day forecast. We further update our prediction to also incorporate short-term variations. Where in the remaining-day prediction we aimed at capturing longer-term changes in the demand, here we focus on taking short-term changes into account that not endure in the longer term. The short-term prediction provides forecasts up to 1.5h ahead with 15min increments.

We use the underlying and individual temporal profiles to update the prediction as follows. We compare the magnitudes $\hat{h}_{t|N} = h^{24}_t + \hat{\eta}_{t|N}$ from the smoothing method (see (4.12)) with the remaining-day forecast $h^{rd}_{t|N}$ regarding these scaling magnitudes as inferred from (4.14). The short-term prediction h^s follows from the mean difference between the

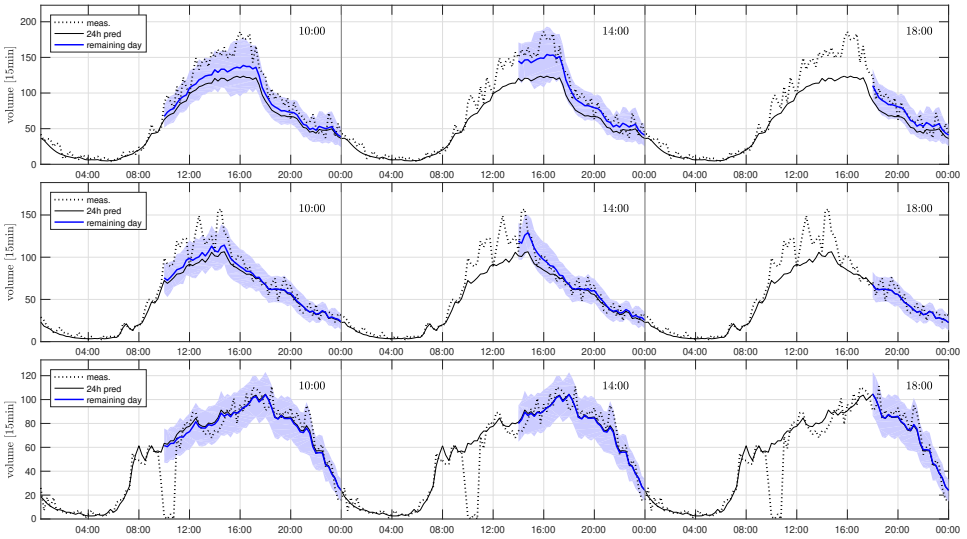


Figure 4.13: Examples of remaining-day predictions at different measurement locations. The solid blue line gives the point predictions for the remaining day, and the shaded blue area is the 90% prediction interval. The solid black line is the initial 24h prediction, and the dashed line shows the measurements.

two estimates, so that the vector-estimate for the profile magnitudes becomes

$$h^s_{N+\upsilon|N} = h^{rd}_{N+\upsilon|N} + \gamma^s \frac{1}{T_\upsilon} \sum_{k=0,1,\ldots,(T_\upsilon-1)} (\hat{h}_{N-k|N} - h^{rd}_{N-k|N}), \quad \upsilon = 1, \ldots, 6.$$

The final volume prediction y^s at time N for forecasting horizon υ is then

$$y^s_{N+\upsilon|N} = W_{N+\upsilon,.} h^s_{N+\upsilon|N}.$$

We determine the parameters $\gamma^s \in [0,1]$ and $T_\upsilon \in \mathbb{Z}_{\geq 0}$ based on the training set. By increasing the history parameter T_υ the prediction becomes less sensitive to the noise in the recent measurements, but we did not observe an improvement in our prediction. The short-term density function predictions $f^s_{N+\upsilon|N}(x)$, $\upsilon = 1, 2, \ldots, 6$, at time N, are constructed following the method of Section 4.5.3. The parameters accompanying the density forecast are independently optimized for each prediction horizon υ.

Figure 4.14 shows examples of the 15, 30, and 45min point predictions, including the 90% prediction intervals derived from the accompanying density forecasts. We observe that where the remaining-day prediction might substantially differ from the 24h prediction, short-term predictions show less-strong improvements over time since the remaining-day forecast turns out to be quite accurate already. We further assess the quality of the point and density forecasts in Section 4.7.

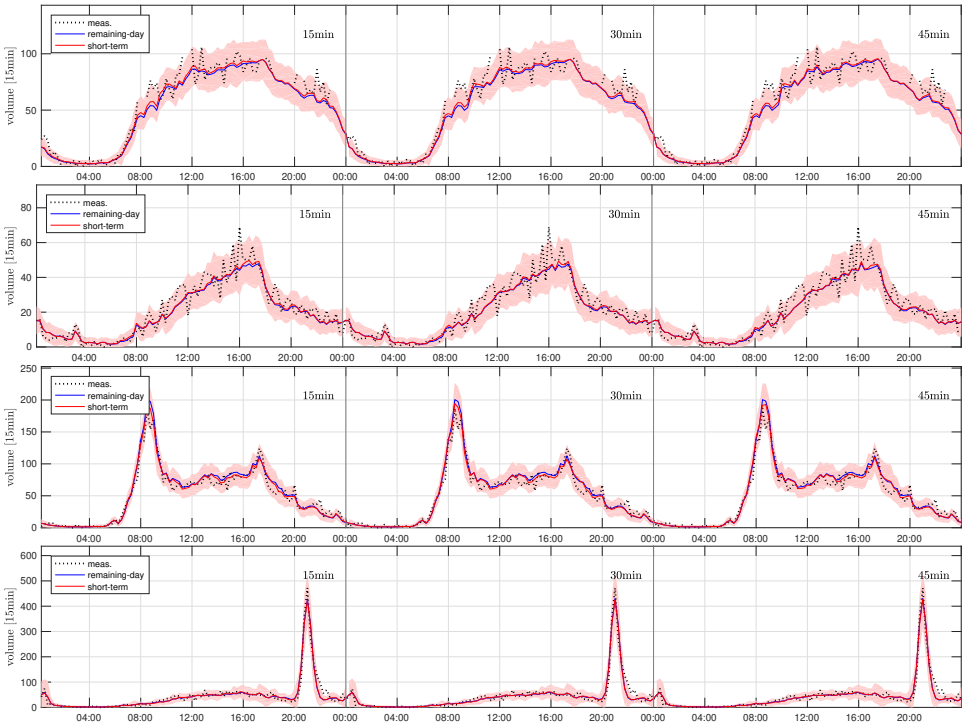


Figure 4.14: For different measurement locations, examples of the short-term point prediction and 90% prediction interval (red) with different forecast horizons (15min, 30min and 45min), including the remaining-day forecast (blue) and the measurements (black).

4.7 Prediction results

In the previous sections, we constructed point and density function forecasts for timescales ranging from 15min up to 24h. In this section, we assess the quality of the predictions using the test set. First, we discuss in Section 4.7.1 the forecast errors for point and density predictions for longer prediction horizons (i.e., the 24h and remaining-day forecast). In Section 4.7.2, we assess the quality of the short-term predictions in the test set by comparing the prediction error with the lower bound as indicated by the noise level. In Section 4.7.3, we compare our volume forecasts with point predictions using a mechanism from literature.

4.7.1 Longer-term predictions

We use an a posteriori comparison between the longer-term predictions and the measurements in the test set, independently for each measurement location under consideration. In this subsection, we compare three different predictions: the baseline (point) prediction, the 24h prediction and the remaining-day prediction. We first consider the point forecasts, and then discuss quality of the density predictions.

Figure 4.15 shows the prediction error as in (4.11) for the remainder of the day at three

different time instants (10:00, 14:00 and 18:00). Here, for each time interval, we calculate the prediction error for each measurement location based on the volume-dependent noise variance (Figure 4.4) and depict the mean over all locations. We note that the baseline as well as the 24h prediction are not adapted throughout a day.

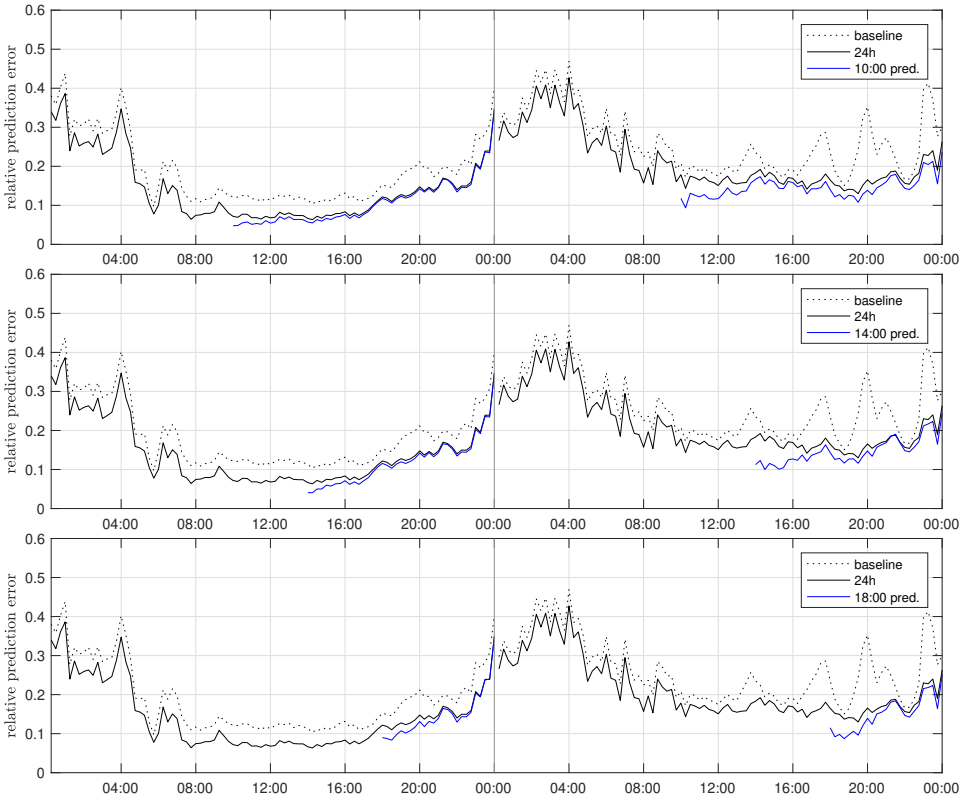


Figure 4.15: Relative prediction error for weekdays (left) and weekends (right) for the baseline (dashed), 24h (black) and remaining-day prediction (blue) at three time instants.

We observe that the 24h prediction substantially reduces the relative prediction error compared to the baseline forecast. This improvement is consistent over time, meaning that a large share of the variations in volumes can be explained by seasonal (i.e., longer than 24h) variations. When comparing weekdays with weekends, we see an increased prediction error during the weekend suggesting that the seasonal variations are more dominant during weekdays. For weekends, a large share of the improvement is obtained by incorporating recurrent events in the 24h prediction. The larger prediction error during the weekend can partly be explained by local (non-recurrent) events that are difficult to account for on timescales longer than 24h. A similar observation is made when comparing the error during the night with the relative error during the day: variations during the night show less repetitive behavior on longer timescales.

For shorter prediction horizons (i.e., the first few hours after the start of the blue line), the improvement of the remaining-day update is substantial. However, for long time horizons

(several hours after the most-recent update of the remaining-day forecast), the remaining-day prediction is not much better than the 24h prediction. Larger improvements are obtained for weekends where volume fluctuations seem to show repetitive behavior over shorter timescales. In any case, Figure 4.15 shows that the prediction is improving over time. However, the adaptation of the remaining-day forecast is insensitive to deviations during the night. Possibly, variations during the night can better be explained by volume fluctuations from the day before even though the prediction error also increases at the end of the day. We also note a small peak in the error between 6:00-7:00, i.e., the start of morning congestion is more difficult to predict than the peak itself.

We constructed 24h and remaining-day point and density function predictions, which can be used for forecasting volume intervals. As mentioned, increased point-prediction errors lead to wider intervals. Now, we consider the optimized density forecasts and assess the quality of the accompanying 90% prediction interval for a remaining-day forecast as a whole. Ideally, the 90% prediction interval for the forecast for the remainder of the day covers - on average - 90% of the measurements. Figure 4.16 shows the average absolute coverage difference (ACD) for the forecast for the remainder of the day, as a function of time of day. Here, we calculated the ACD for the 90% prediction interval independently for each location, which we averaged over all locations. This ACD does not reflect the width of the interval, but the width can directly be derived based on an estimate of the noise variance and error as in Figure 4.15. In Section 4.7.2, we evaluate the density forecasts in more detail.

The constructed density predictions are accurate in the sense that on average 88.5 to 91.5% of the remaining day's measurements are within the 90% prediction intervals. Hence, the uncertainty accompanying a forecast and the random variation in the volumes is well-accounted for in the probabilistic predictions. Although the prediction intervals for the 24h volume predictions are wider compared to the remaining-day forecast (see Figure 4.15), the intervals are accurate in the sense that before the start of the day we expect an ACD of 1.2% for the daily volumes. At the same time, the probabilistic prediction is less accurate during weekends and during the night. Here, the ACD's for this particular uncertainty level can be improved by optimizing the ACD over the training set.

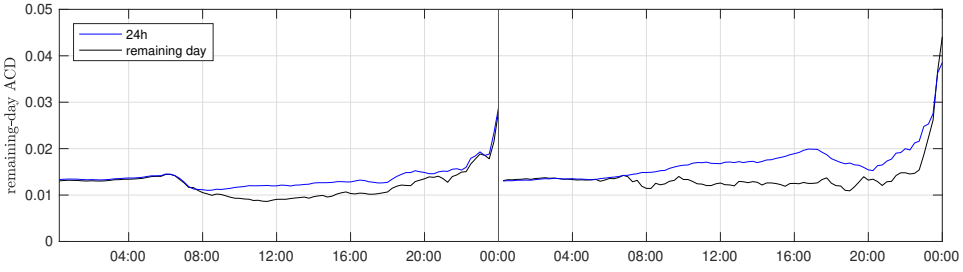


Figure 4.16: Absolute coverage differences for the volumes during the remainder of the day using the 90% *prediction interval for weekdays (left) and weekends (right).*

4.7.2 Short-term predictions

We assess the quality of the short-term point predictions by comparing the forecast error with the lower bound as provided by the noise level as in (4.5). Figure 4.17 shows the 15min relative prediction error as in (4.11) for both weekdays and weekends, with averages over the time-of-day forecast errors for the different locations under consideration. In general, relative prediction errors are small during weekdays (on average around 10%), in particular for the between-peak period, and increase during weekends, during the night and in the early morning. These intervals have typically (very) low volumes, and higher-volume increments are thus shown to be easier to be predicted overall. As already mentioned (Section 4.7.1), longer-term variations do not fully express the fluctuations in the volumes during the night and in the weekend. Although the remaining-day prediction shows comparable performance with the short-term prediction after the early morning, Figure 4.17 shows that the short-term prediction reduces the prediction error substantially during the night.

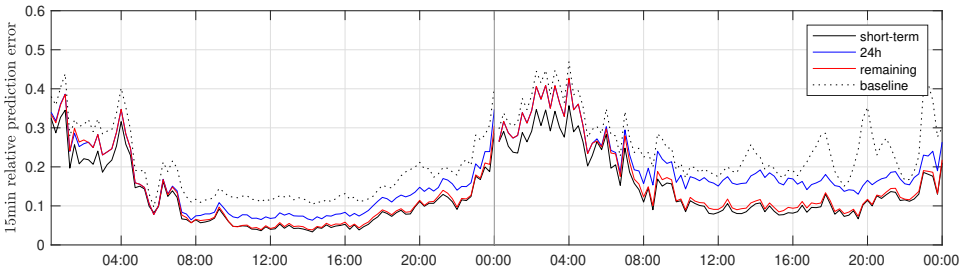


Figure 4.17: Time-of-day average 15min prediction errors for weekdays (left) and weekends (right).

We can draw similar figures as in Figure 4.17 for forecast horizons from 30min to 1.5h. Overall, point predictions can be improved by 10 to 15%, with prediction errors growing with an increasing horizon. Figure 4.18 shows the prediction errors for forecasts covering 15min, 60min and 24h. Indeed, the prediction error of the 60min forecast is worse compared to the 15min one, yet improving the 24h estimate. Interestingly, many measurement locations show similar performance for both the 15min, 60min and 24h forecasts which means that 24h patterns are particularly useful for these locations. The largest prediction errors occur during the weekend at major arterials serving traffic to and from the freeways. 24h patterns are then less beneficial to explain volume variability, and short-term and spatial relations are likely to explain a larger share of the variation.

Regarding the predictive densities, we assess the quality of the constructed target percentile forecasts for the 15min horizon using the ACD in Figure 4.19a. A decreasing ACD for updated forecasts for the different quantile predictions shows that improving point predictions are accompanied with improved density forecasts. Overall, the proposed framework for constructing probabilistic forecasts results in quite accurate coverage, with the ACD for the short-term 15min prediction fluctuating around 2%. Further improvements can be obtained by relaxing the normality assumption and/or constructing independent quantile forecasts. Figure 4.19b shows the ACD for the constructed quantile predictions for different prediction horizons. The quality of the density forecasts for increasing forecasting horizons is very similar, again substantiating that the proposed framework provides accurate density

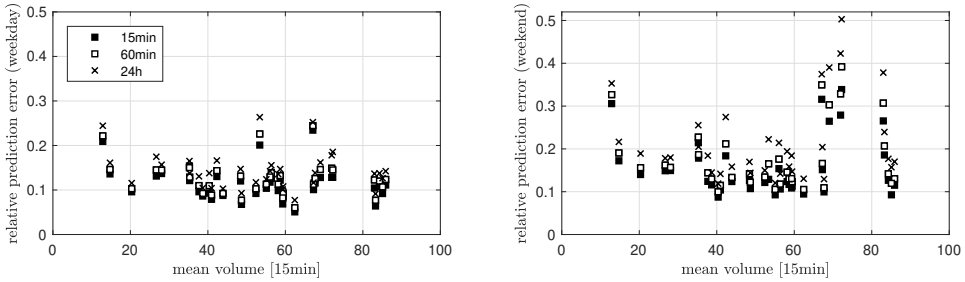
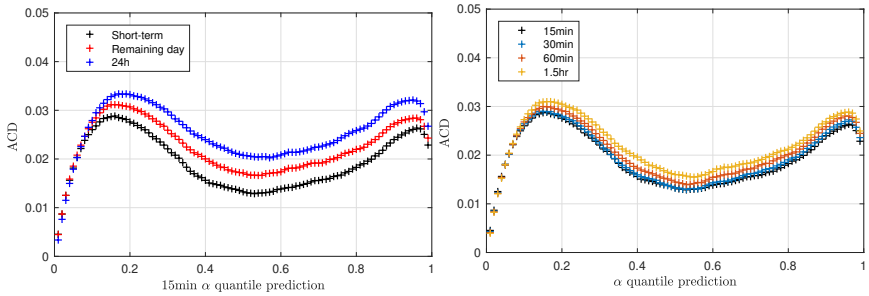


Figure 4.18: Average prediction errors for point predictions with different horizons for the measurement locations under consideration.

forecasts over different timescales.



(a) ACD for 15min forecasts

(b) ACD with different prediction horizons.

Figure 4.19: Target quantiles - absolute coverage differences (ACD) for predictions covering different timescales.

4.7.3 Prediction comparison

In previous subsections, we assessed the quality of the predictions by comparing the forecast error with a lower bound as suggested by the noise level. In this subsection, we compare our short-term point predictions with forecasts by the K-NN method of Habtemichael and Cetin (2016) that has shown to outperform some other point-prediction methods for motorway and freeway traffic. First, we briefly discuss the adopted K-NN method and then we compare the results based on several performance metrics.

K-NN Method

We adopted the K-NN method as in Habtemichael and Cetin (2016) as follows. At time N of day d, we have measured a part of the traffic volume for that day, i.e., sequence $\{x_{t,d}\}_{t=1}^{N}$ is observed. The K-NN method compares a part of this sequence $\{x_{t,d}\}_{t=N-m}^{t}$ (the m-lagging part) with flow rates in the historical data set. In fact, we select the K-best days $d' < d$ with $\{x_{t,d'}\}_{t=N-m}^{t}$ similar to $\{x_{t,d}\}_{t=N-m}^{t}$. Hence, we have a set (of size K) of nearest neighbors which have shown a similar traffic pattern in the past, and can

therefore be used for predicting volumes. In fact, the prediction for time $t + v$ is then a weighted sum of the succeeding measurements in the past, i.e., $y^{knn}_{t+v,d} = \sum_{d'} x_{t+v,d'} w_{d'}$, with $w_{d'}$ a normalization weight so that $\sum_{d'} w_{d'} = 1$.

As in Habtemichael and Cetin (2016), we use the weighted Euclidean distance (see Eq. (4) and (10) in the aforementioned paper) to measure similarity between candidate and subject profiles. We also apply Windsorization, loess-smoothing on the lagging-part of the historical data (span of 0.2), and the Rank-exponent method (with $Z = 2$) of weight assignment as suggested. In contrast, for each location we solely use the data set that corresponds to that location, and when predicting we are only allowed to select neighbors from the historical data set. For each prediction horizon v, we select the location-specific lag m and value of K based on the mean squared error of the residuals on the training set. Compared to the study of Habtemichael and Cetin (2016) with $K \approx 10$, we observe that it is beneficial to increase K so that typically $K > 30$. In addition, we found values of m typically exceeding 20 while Habtemichael and Cetin (2016) found values of $m \approx 4$.

Comparison

We compare our prediction method with the K-NN mechanism. For the different prediction horizons we use the following three metrics to compare forecasts y with measurements x: the relative prediction error as in (4.11); the (low-volume adjusted) mean absolute percentage error (MAPE) $\frac{1}{n}\sum_t \left|\frac{y_t - x_t}{\max\{3, x_t\}}\right| \cdot 100\%$; and the root mean square error (RMSE) $\sqrt{\frac{1}{n}\sum_t (y_t - x_t)^2}$. Here, n is the number of out-of-sample measurements. In Section 4.3.3, we argued that the latter two performance indicators are biased but we use them here since they are widely adopted to assess quality. Moreover, we use the relative prediction error to explicitly account for the random variation when comparing the different predictions.

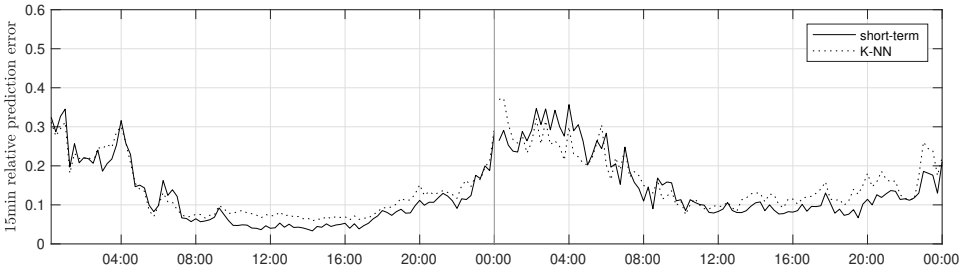


Figure 4.20: Time-of-day average 15min prediction errors for the forecasts of the K-NN method and the proposed short-term prediction method during weekdays (left) and weekends (right).

Figure 4.20 shows the relative prediction error of our 15min prediction and the K-NN method relative to the time of day. Overall, our method shows slightly better performance regarding this metric, with more substantial improvements during weekdays. The K-NN prediction error is smaller during the nights preceding weekends. This difference can be explained by the fact that we use - in this case restricting - 24h profiles, while the K-NN method looks more explicitly at volumes in the recent past, which might occur during the previous day.

When the prediction horizon increases to 60min, our method shows more substantial improvements regarding the performance metrics. Figure 4.21 shows the difference in MAPE for the 15min and the 60min forecasting horizons, and Figure 4.22 compares the RMSE for both mechanisms with a 15min and 60min timescale for each location under consideration. These results suggest that our method is on average better able to predict traffic volumes in the urban setting we consider, with an average reduction of about 3-7% per metric. Furthermore, the underlying patterns benefit predictions when the forecasting horizon increases. Yet, performance of each method might differ from day to day.

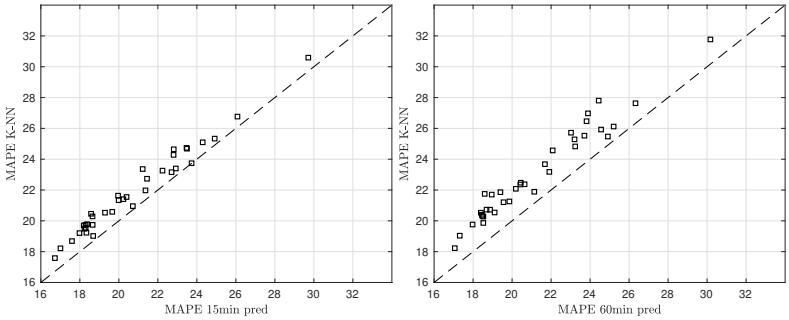


Figure 4.21: Comparison of the MAPE our short-term forecast and K-NN prediction for 15min (left) and 60min (right) predictions for the different measurement locations.

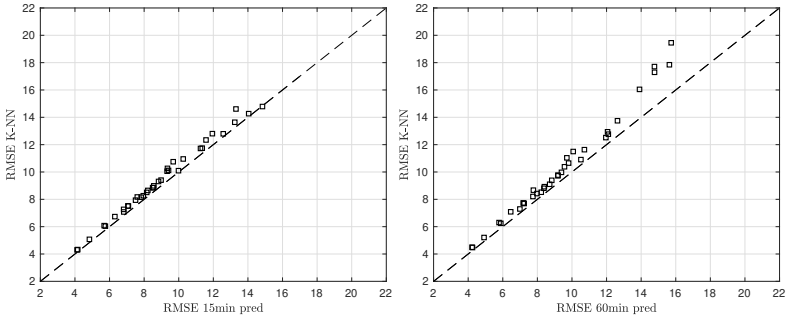


Figure 4.22: Comparison of the RMSE our short-term forecast and K-NN prediction for 15min (left) and 60min (right) predictions for the different measurement locations.

The difference in the prediction error metrics does not fully express the differences in predictions. For example, the K-NN method is in its current form not capable to incorporate short-term patterns with varying starting times. In fact, the time domains between subject and candidate profiles are assumed to be shared. For instance, the upper row in Figure 4.23 shows days in which our method has a smaller prediction error compared to the K-NN method. The most-left figure shows the difference in the 15min prediction on a day where a weather (snow) alarm was issued during the day. Although our method requires some time to adapt, it slowly converges to the measurements. Since the training data did

not include such patterns, the K-NN method shows almost no adaptation. The middle figure shows that event-induced volume changes are not accounted for by the K-NN method. The most-right time series illustrates that the K-NN method does not well-anticipate the regular Wednesday-noon peak at this location. The lower row of 24h time series in Figure 4.23 shows examples of days at the same location with the K-NN reducing the error compared to our method on a 15min timescale. Here, absolute differences are less substantial. In fact, for this measurement location, our 15min short-term prediction leads to an overall improvement of the RMSE and MAPE of about 5%.

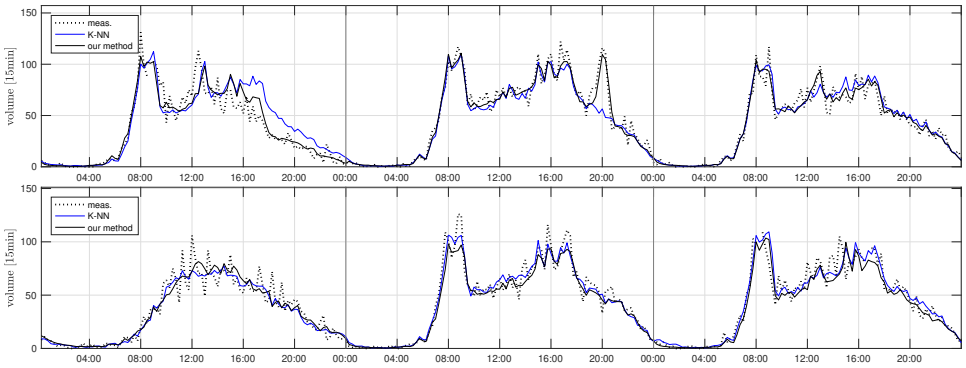


Figure 4.23: Comparison of our 15min predictions (black) with forecasts using the K-NN method (blue).

4.8 Conclusion

Logistics service providers are faced with uncertain driving times when designing route plans. Therefore, they desire network-wide travel time predictions for different timescales to construct robust route plans that can be dynamically adapted over time. Predictions of the traffic volumes support the anticipation of travel time fluctuations in particular when the saturation rate approach capacity, i.e., when the onset of congestion is to be predicted. Uncertainty on different timescales influence predictions, and can only be reduced to a limited extent. Probabilistic forecasts express the uncertainties in conditions and predictions, and are, ideally, offered for both the long and short term for a range of settings and conditions.

In this chapter, we provided a prediction mechanism for urban traffic volumes from 15min to 24h ahead. Where most prediction mechanisms in literature offer point forecasts for regular freeway conditions, our model provides both point and full density function estimates for regular days as well as for days with special events. In fact, the model constructs three types of point and density function predictions: the 24h prediction, a remaining-day prediction and a short-term prediction. The 24h prediction accounts for the day-specific pattern and seasonal variations, and can be predicted before the start of the day. The remaining-day prediction updates the 24h forecast to account for deviations in the demand by comparing the initial 24h forecast with the smoothed measurements. The short-term forecast additionally accounts for fluctuations covering shorter timescales.

We presented a framework for constructing density forecasts by explicitly quantifying

the uncertainties accompanying a prediction. The uncertainty cannot be eliminated (Gneiting et al., 2007), since next to the natural but random variation in the counts, also the prediction error and the measurement errors or limitations are part of the remaining variation. The predictions use a set of 24h and short-term profiles that together describe the 24h volumes. These temporal profiles turn out to be very useful for longer-term predictions. Indeed, the provided predictions are shown to be accurate, with point predictions having an average prediction error of 10-15%, while density forecasts have coverage differences of 1-3%. Forecasts are robust in the sense that the prediction error and the coverage difference only marginally increase when the prediction horizon increases. When comparing our predictions with the forecasts of the K-NN method from literature, we reduce the error metrics with up to 7%.

Further research includes improving the prediction mechanism, particularly for increments during the night and during the weekend for major arterials with higher volumes. Here, 24h patterns express a smaller share of the variation, and spatial correlations can benefit predictions. In fact, shorter-term variations due to events and incidents can then be anticipated. In addition, where we have focused on predicting traffic flow rates, forecasting travel times based on volume estimates is a non-trivial task and topic for further investigation.

Chapter 5

Improving the performance of a traffic system by fair rerouting of travelers

5.1 Introduction

Transport authorities face the daily challenge to reduce congestion. Traditionally, this was solved by increasing road capacity through building new or expanding existing infrastructure. However, the construction of infrastructure is costly, and may also lead to an increase in demand. Nowadays, authorities implement management measures alongside to improve utilization of existing roads.

The need for policy measures in general stems from the observation that individuals typically behave selfishly, i.e., travelers are mainly concerned with their own utility when making decisions. The resulting traffic state (i.e., flow distribution) with respect to route choice, the *user equilibrium*, does mostly not correspond to the *system optimum*: the traffic state with minimum (total or average) travel time (Wardrop, 1952). Without intervention, in particular with the increasing use of real-time routing apps, the real-world traffic state is likely to be closer to the inefficient user equilibrium than to the system optimum (Klein et al., 2018). In the user equilibrium, travelers with the same origin-destination pair have equal travel times. The system optimum, on the other hand, is 'unstable' since it is unfair: some drivers may travel longer than others for the same origin-destination pair. Hence, we can characterize the system optimum as (perfectly) efficient but unfair, while the user equilibrium is inefficient and perfectly fair.

Recently, traffic management measures, e.g., social routing, have been proposed that steer or nudge travelers towards socially-desired routes. The 'pure' system optimum is difficult to achieve (Klein et al., 2018) and maintain over time, because only some travelers use and comply with advice from information systems, and the individual intra- (within the

This chapter is based on: Eikenbroek, O. A. L., Still, G. J., & Van Berkum, E. C. (2022). Improving the performance of a traffic system by fair rerouting of travelers. *European journal of operational research*, 299(1), 195-207. Available at: https://doi.org/10.1016/j.ejor.2021.06.036

system optimum) and inter-state (compared to the user equilibrium) travel time differences might be substantial (Jahn et al., 2005; Van Essen et al., 2020). Hence, any social routing strategy should in essence anticipate user responses and persuade travelers to comply with socially-oriented advice.

Empirical evidence (e.g., Djavadian et al., 2014) shows that some (travelers) are receptive for advice that proposes reasonable routes for the system's benefit. A possible explanation is that individuals have a so-called *indifference band* (Simon, 1997), which means in our context that when a route is only slightly longer than the best one, it is still acceptable to use (Vreeswijk et al., 2015). A social routing strategy can 'exploit' the indifference band and propose acceptable routes (possibly, *sub-optimal* from an individual's perspective) to *receptive drivers* (those that use and comply with advice from the service), and thereby potentially steer the network to a state close to the system optimum. Compared to the system optimum, the resulting distribution is easier to achieve and maintain over time.

In this chapter, we propose and evaluate a centrally coordinated social routing strategy that improves overall efficiency, while we explicitly account for the above-mentioned practical requirements. The routing strategy incorporates user-induced constraints in the sense that travel time differences in the resulting state are explicitly limited, and only a fraction of the travelers is asked to take an acceptable detour to the system's benefit. We note that a routing service adopting the strategy, in practice, offers a single route advice using a personalized information device to its users before departure.

Research contribution

Although empirical research has shown that social routing has great potential in real life, there is not yet a corresponding routing strategy that improves efficiency while explicitly incorporating user responses to advice in terms of route choice behavior. Route choice behavior is crucial for the strategy's performance in practice. Compliance is expected to be much higher when the advised route is only slightly longer than the shortest route. Behavioral responses influence the travel times, and should thus be anticipated in order to advise routes that are acceptable with respect to travel time.

In this chapter, we propose a social routing strategy that explicitly accounts for behavioral responses to a routing service. In fact, changes in route choice may occur from travelers that comply with the advice but also from those that do not comply, but are now confronted with altered travel times on routes as a result of behavioral changes by others. We introduce a bilevel optimization problem that calculates the best possible paths (with respect to efficiency) with a limited (realized) detour to be proposed to the compliant travelers. Although in this chapter we limit ourselves to a static environment, the bilevel problem is already highly challenging to solve. Many of the theoretical difficulties that occur in our case, also apply to a real-world social route guidance service in which limited detours are suggested in a dynamic fashion. Hence, before considering such a guidance system we should address the theoretical challenges and potential impact in a static traffic assignment first. In particular, the service as proposed in this chapter can serve as a proof-of-concept for a dynamic variant.

Related social routing approaches

We discuss related social routing approaches from literature. Jahn et al. (2005) proposed a routing strategy that limits the 'normal length' difference before and after implementation, assuming that the normal length is independent of the traffic flow. This mechanism was numerically evaluated on realistic network instances, and showed performance (with respect to efficiency) close to the system optimum. The intra-state time differences, however, were not explicitly limited. A related approach by Angelelli et al. (2016) considers a mathematical program that tries to achieve an optimal flow with a regularization term to minimize the 'total inconvenience' alongside. Here, the travel time is assumed to be independent of the flow. Both studies assume a full market penetration of the routing service. Bagloee et al. (2017); Van Essen et al. (2020) and Zhang and Nie (2018) proposed systems to route a fraction of the demand onto social routes. We refer to Li et al. (2018) and Zhou et al. (2017), for dynamic (day-to-day) variations on such routing policies.

In contrast to the above-mentioned studies, we propose a routing strategy that steers the network to a system optimum while explicitly limiting the intra-state time differences whereas we argue that travelers evaluate acceptability of routes in terms of realized travel time rather than free-flow travel time or distance. This necessary user-induced constraint makes the accompanying optimization problem substantially harder to solve, which might be a reason that a majority of the studies relax this real-life constraint or introduce heuristic approaches (e.g., Angelelli et al., 2018; Roughgarden, 2005). Recently, Angelelli et al. (2020) studied a similar setting, with a so-called 'constrained system optimum'. They used an integer linear program and matheuristic to formulate and solve the corresponding optimization problem, respectively. In contrast to our study, they do not incorporate the route choices of travelers that do not comply with route advice. Angelelli et al. (2021) proposed a fast heuristic to find the constrained system optimum and use a piecewise linearization of the travel time function. In our chapter, we formulate the problem as a continuous optimization problem and keep the nonlinearity of the travel time function.

We note that our optimization problem is a generalized case of finding the *boundedly rational user equilibrium* (BRUE) with minimum travel time. Although there is a body of literature on BRUE (e.g., Di et al., 2013; Lou et al., 2010), a thorough quantitative analysis of this problem is still lacking. Thus far, analyses have been based on relatively strong assumptions which reduce the complexity of the problem but might not hold in practice.

Bilevel problem

The success of the social routing strategy, as discussed, hinges on the (travel time of the) paths suggested to the drivers. We show that best possible paths can be found by solving a bilevel optimization problem. Our bilevel problem (see Section 5.2.2) can be seen as a game between a leader (authority) and a follower (travelers) (Josefsson & Patriksson, 2007). The leader chooses the paths to be proposed, while the travelers update their route choice based on this advice. The compliant travelers follow the advice if the travel time differences (based on the route choice of the travelers) compared to the fastest paths are limited, while non-compliant travelers find the cheapest paths available. These dynamics should be anticipated to find the best possible advice in terms of total travel time.

Bilevel problems are typically difficult to solve directly, and therefore often reformulated as single-level problems. In this chapter, we use an implicit reformulation, and require

parametric analysis of the lower-level problem to describe the behavior of the corresponding solution set as a function of the upper-level variable. Parametric analysis is either quantitative or qualitative in nature (Fiacco & Ishizuka, 1990b). The qualitative analysis is mainly concerned with the continuity of the optimal solution set. Here, we require a - local - quantitative analysis that focuses on the estimation of generalized derivatives of the optimal solution set, e.g., to be used in numerical procedures. We refer to Eikenbroek et al. (2018) for the qualitative analysis of this problem (the mentioned paper's setting is however different).

Techniques from variational analysis are used to study the quantitative behavior of the lower-level problem. We refer to Luo et al. (1996), Mordukhovich (2018), and Rockafellar and Wets (2009) for an overview of theoretical results. Many of these results, however, are presented in general form and require relatively strong conditions when applying an implicit reformulation. A critical issue in our case is that the lower-level solution is not unique for a given upper-level variable. This leads to practical and theoretical challenges, since the desired lower-level solution might not be realized, and small changes in the parameter might lead to major changes in the solution (Dempe, 2002). Hence, at first sight, many of the algorithms designed for bilevel problems do not apply in our context.

In this chapter, we theoretically assess the lower-level problem and use techniques from variational analysis to show that we can guarantee the existence and calculation of a generalized derivative of the lower-level solution projected onto a subspace. The generalized derivative of the solution of the lower-level problem contributes to the understanding of the optimization problem finding the best possible paths. Indeed, the theoretical analysis in this chapter allows one to formulate the necessary optimality conditions of the bilevel problem. Moreover, the derivative can be used in exact numerical procedures to find descent directions. Hence, not only can the generalized derivative be used in standard algorithms to solve bilevel programs, it can also support the assessment of heuristic procedures (e.g., Angelelli et al., 2020). Although a comparative analysis of algorithms that solve the formulated program is beyond the scope of our research, we provide nonetheless a numerical procedure and refer to related algorithms that could be applied.

In a static traffic assignment context, bilevel problems are well-known, mainly in *Network Design Problems* (NDPs) in which optimal network settings (e.g., link tolls) are determined. Parametric analysis has been topic of a body of literature in this context (Chung et al., 2014; Josefsson & Patriksson, 2007; Lu, 2008; Lu & Nie, 2010; Outrata, 1997; Patriksson, 2004; Patriksson & Rockafellar, 2002, 2003; Qiu & Magnanti, 1989; Robinson, 2006; Tobin & Friesz, 1988; Yin et al., 2009). Mainly, these papers concern perturbations that occur in the (parameters of the) link cost function and/or demand vector. Our research is different from the aforementioned studies since we basically consider perturbations in a path-dependent parameter. It turns out that the analysis and the computational results rely on the choice of a suitable route flow corresponding to a link flow solution. This forces us to study the behavior of the (multi-valued) route flow solution set in dependence of the parameter. In the context of perturbations in the parameter of the demand vector, the analysis of Qiu and Magnanti (1989) also depends on the choice of a specific route flow solution. However, Patriksson and Rockafellar (2002) show that the results of Qiu and Magnanti (1989) are actually independent of a specific choice. In our context, this does not hold (as we will show in Example 1). Some of our findings show similarities to the results for parametric optimization problems with a unique minimizer but non-unique multipliers (Dempe, 1989; Dempe, 1993; Ralph & Dempe, 1995). There, a generalized derivative of the optimal solu-

tion can be calculated by choosing a suitable multiplier (which might be difficult to find). In our case, we consider a setting with a non-unique optimal solution.

Considering the practical application, we assess a possible implementation of the social routing strategy. Specifically, we evaluate the interaction among compliance rate, acceptable travel time differences, and network-wide performance in a static setting. The numerical experiments provide insight which minimum penetration rate and indifference band might be required to substantially lower the total travel time in the network, and, thus, how much some travelers have to sacrifice for the network's benefit. These experiments substantiate the opportunities for a real-life implementation of a social routing service or guidance mechanism.

Summarizing, the main contributions of this chapter are as follows:

- We propose a social routing strategy that steers the traffic network towards an efficient but also fair, and therefore achievable and maintainable, traffic state. We show that the best possible paths to be proposed by a social routing service can be found by solving a bilevel program that explicitly accounts for behavioral responses to the service;
- We use parameteric analysis to prove that the generalized derivative of the lower-level link flow solution problem exists and can be calculated efficiently. The generalized derivative can be used to find descent directions and to formulate optimality conditions of the bilevel problem;
- We use the generalized derivative in a descent algorithm to solve the bilevel problem and numerically evaluate our proposed social routing strategy in test networks. Here, only a small fraction of the travelers need to take a limited detour to substantially improve the traffic system's performance.

The remainder of our chapter is organized as follows. We formally introduce our social routing strategy in Section 5.2. In Section 5.3, we analyze qualitatively the 'behavior' of the optimization problem that relates to our social routing strategy. In Section 5.4, we investigate the existence and calculation of the directional derivative of the link flows, which we use in Section 5.5 in a descent algorithm for solving the bilevel problem. Section 5.6 reports on numerical experiments and management implications. Section 5.7 draws the conclusions.

5.2 Problem formulation

We study the static traffic assignment with fixed demand. Given is a directed traffic network $G = (V, E)$, with V being the set of nodes, and E is the set of directed edges (roads or links) $e = (i, j)$, with $i, j \in V$. The network has a set of *origin-destination pairs* (OD pairs) $\mathcal{K} \subseteq V \times V$, with static demand $d_k > 0$, $k \in \mathcal{K}$. Each OD pair $k \in \mathcal{K}$ is connected by the set $\mathcal{P}_k$ of simple directed paths. The set $\mathcal{P}$ of all paths in the network is the union of the path sets per OD pair, i.e., $\mathcal{P} = \cup_{k \in \mathcal{K}} \mathcal{P}_k$.

A feasible *traffic flow* or *flow* for given demand $d \in \mathbb{R}_+^{|\mathcal{K}|}$ (we denote by $|.|$ the cardinality

of a set) is a pair of vectors $(f, x) \in \mathbb{R}^{|\mathcal{P}|} \times \mathbb{R}^{|E|} = (f_p, p \in \mathcal{P}; x_e, e \in E)$ so that

$$\Lambda f = d, \quad \Delta f - x = 0, \quad f \geq 0. \tag{5.1}$$

The matrix $\Lambda \in \mathbb{R}^{|\mathcal{K}| \times |\mathcal{P}|}$ is the *OD-path incidence matrix* with $\Lambda_{kp} = 1$ if $p \in \mathcal{P}_k$, and $\Lambda_{kp} = 0$ otherwise. $\Delta \in \mathbb{R}^{|E| \times |\mathcal{P}|}$ denotes the *link-path incidence matrix*: $\Delta_{ep} = 1$ if edge e is in route p, and $\Delta_{ep} = 0$ otherwise. For each edge, $e \in E$, $l_e(x_e)$ is the non-negative, continuous, and non-decreasing link cost (or: travel time) function for a given flow x_e on that edge. The cost of a route $c_p(f)$, $p \in \mathcal{P}$, is the sum of travel costs of all edges in that path, $c_p(f) = \sum_{e \in p} l_e(x_e)$.

Throughout our chapter we make the following (natural) assumption regarding the travel time function (we refer to Patriksson and Rockafellar (2002) for a study that relaxes this assumption).

Assumption 1. *We assume throughout the chapter that the travel time functions $l_e(x_e)$ are continuous, convex, and strictly monotone: $l_e(x_e) < l_e(x_e^0)$, for $x_e < x_e^0$, for all $e \in E$.*

5.2.1 A social routing strategy

We consider the setting in which a central authority asks travelers to take a small detour for the system's benefit (see Section 5.1). The *social travelers* comply with such an advice if the alternative route is reasonable, i.e., the route is not perceived to be substantially worse (in terms of travel time) compared to the fastest path. The remaining drivers do not comply with travel advice and behave in a selfish manner, i.e., choose the fastest path available.

The demand vector $d^s \in \mathbb{R}_+^{|\mathcal{K}|}$ ($d_k^s \leq d_k$ for all $k \in \mathcal{K}$) denotes the travelers that receive and comply with a route advice from the authority (superscript s refers to the social travelers). The remaining demand $d^n \in \mathbb{R}^{|\mathcal{K}|}$, so that $d = d^s + d^n$, behaves selfishly. (Superscript n refers to Nash equilibrium - see (5.2b) below: a driver cannot improve travel time by changing strategy (route)).

We define $\mathcal{F}$ as the set of feasible flows. Formally,

$$\mathcal{F} = \left\{ (g, h, x) \in \mathbb{R}^{|\mathcal{P}|} \times \mathbb{R}^{|\mathcal{P}|} \times \mathbb{R}^{|E|} \;\middle|\; \begin{array}{l} \Lambda g = d^s, g \geq 0, \\ \Lambda h = d^n, h \geq 0, \\ \Delta(g + h) - x = 0 \end{array} \right\}.$$

Obviously, any $(g, h, x) \in \mathcal{F}$ is a flow as in (5.1) for $f = g + h$.

The advised routes to compliant travelers d^s have to be fair in the sense that the realized (i.e., traffic flow-dependent) travel time differences are limited. We assume that social travelers accept any travel time difference (compared to the shortest path for the same OD pair) with a maximum of $\varepsilon_k \geq 0, k \in \mathcal{K}$. Hence, the resulting state in the network is so that no social traveler for OD pair $k \in \mathcal{K}$ can improve travel time with more than ε_k by unilaterally changing routes. At the same time, the selfish travelers choose the fastest path. The following definition (Definition 1) formalizes our notion of the resulting state among social (receptive) and selfish travelers. We refer to this state as a *mixed equilibrium*.

Definition 1 (Mixed equilibrium). *Given $\varepsilon \in \mathbb{R}_+^{|\mathcal{K}|}$, a traffic flow $(g, h, x) \in \mathcal{F}$ with corresponding path costs $c(f)$, $f = g + h$, is called a mixed equilibrium among social and selfish travelers if for all $k \in \mathcal{K}$, the following conditions are satisfied for all $p \in \mathcal{P}_k$:*

$$g_p > 0 \Rightarrow c_p(f) \leq \min_{q \in \mathcal{P}_k} c_q(f) + \varepsilon_k \tag{5.2a}$$

$$h_p > 0 \Rightarrow c_p(f) = \min_{q \in \mathcal{P}_k} c_q(f) \tag{5.2b}$$

Assuming only selfish demand, in a traffic state in user equilibrium as in (5.2b), travelers with the same OD pair share travel times. However, it is well-known that this state does not necessarily minimize total travel time $\sum_{e \in E} x_e l_e(x_e)$. The traffic state (f, x) as in (5.1) which minimizes the total travel time, is referred to as the system optimum (Wardrop, 1952). Typically, it may be assumed that in practice, without intervention, a state close to a user equilibrium arises.

Condition (5.2a) gives a range of acceptable travel times for a receptive user. We assume that any social traveler that is routed onto an *acceptable path* (i.e., any route $p \in \mathcal{P}_k, k \in \mathcal{K}$ for which $c_p(f) \leq \min_{q \in \mathcal{P}_k} c_q(f) + \varepsilon_k$) complies with such an advice although the user might be aware that it is not necessarily the fastest path available. The condition as defined in (5.2a) is equivalent to the BRUE condition (see Section 5.1). The mixed equilibrium as in (5.2), i.e., (5.2a) and (5.2b), has the user equilibrium as a special case and does not correspond (even if $\varepsilon \to \infty$) to a mixed user equilibrium and system-optimal flow, e.g., as in Yang et al. (2007).

In (5.2a), we model the band ε as being *additive*. In particular for shorter travel times, an additive indifference band is more appropriate compared to a multiplicative one as in, e.g., Roughgarden, 2005. In combination with ε_k, $k \in \mathcal{K}$, being OD-pair dependent, we allow a range of scenarios regarding the maximum detour to be modeled using the condition in (5.2a).

The mixed-equilibrium conditions (5.2) do not provide a unique state (yet all travelers are satisfied with their route), which is key for the social routing strategy. We exploit this range of allowed distributions to find one which is the best for the system. That is, our routing strategy is designed so that we achieve - among all $(g, h, x) \in \mathcal{F}$ that satisfy (5.2) - the one with the minimum total travel time. Hence, the optimal strategy can be found by solving the following optimization program for a known $\varepsilon \geq 0$:

$$\min_{(g,h,x) \in \mathcal{F}} \varphi(x) \qquad \textit{s.t.} \qquad (g, h, x) \text{ satisfies (5.2)}, \tag{5.3}$$

where $\varphi(x) = \sum_{e \in E} x_e l_e(x_e)$ is the total travel time.

For a routing service, the optimal solution of (5.3) with respect to g is typically the variable of interest, since g represents the distribution of the social travelers over the different acceptable paths. The selfish demand basically responds to the choices of the social demand in the sense of (5.2b). In fact, selfish travelers are confronted with a change in travel times on routes due to the choices of others. When determining the best distribution g (with condition (5.2a)), the authority needs to anticipate the travel times depending on the route choices of both the social and selfish demand. This *Stackelberg* mechanism is implicitly in (5.3). After solving (5.3), the route to be suggested to a social traveler can be extracted from solution g.

One should note that, in principle, while solving (5.3), one is free to choose any (g, h, x)

satisfying (5.2). In practice, for a given g, the distribution h is a result of the route choice behavior of the selfish travelers and cannot be precisely predicted (if there are multiple h satisfying (5.2b)). However, as we will see in Theorem 1, the response to g with respect to the link flows x is uniquely determined. Since x is the only variable appearing in the objective function, it is therefore not necessary to consider a pessimistic variant of (5.3).

5.2.2 Bilevel reformulation

The optimization problem in (5.3) is difficult to solve. Indeed, Eikenbroek et al. (2018) and Lou et al. (2010) show that the feasible set corresponding to (5.3) is in general not convex, does not satisfy a regularity condition, and different local minimizers can coexist. We use the following proposition (Proposition 1) to reformulate our problem. In the remainder of the analysis we drop parameter ε in the notation: we assume it is known and fixed. During the experiments (Section 5.5 and Section 5.6), we numerically investigate the impact of a varying ε.

Proposition 1 (Di et al. (2013); Eikenbroek et al. (2018))**.** *The following are equivalent for* (g, h, x)*:*

1. $(g, h, x) \in \mathcal{F}$ *is a mixed equilibrium as in* (5.2)*;*
2. *There exists*
$$\rho \in \Xi := \left\{ \ \rho \in \mathbb{R}^{|\mathcal{P}|} \ \middle| \ 0 \leq \rho \leq \Lambda^T \varepsilon \ \right\}$$
such that (g, h, x) *solves the convex optimization problem*
$$Q(\rho): \qquad \min_{(g,h,x)} z(\rho, g, x) = z_0(x) + \rho^T g \qquad \text{s.t.} \qquad (g, h, x) \in \mathcal{F}, \tag{5.4}$$
where $z_0(x) = \sum_{e \in E} \int_0^{x_e} l_e(\omega) d\omega$.

We omit the proof, which is a generalization of Proposition 2.2 in Di et al. (2013) or Proposition 1 in Eikenbroek et al. (2018). These references use objective function $z_0(x) - \tilde{\rho}^T g$, but the two problems are equivalent by selecting $\tilde{\rho} = \Lambda^T \varepsilon - \rho$. We prefer our objective function in (5.4) whereas it eases the upcoming analysis. We note that ρ does not necessarily have an intuitive interpretation.

Problem (5.3) is a *mathematical program with equilibrium constraints*. According to Proposition 1, we can rewrite (5.3) as a bilevel problem. We use the following reformulation, which eases the parametric analysis in Section 5.3 and 5.4 (Eikenbroek et al., 2018):

$$(BL): \qquad \min_{(g,h,x,\rho)} \varphi(x) \qquad \textit{s.t.} \qquad \begin{array}{l} \rho \in \Xi \\ (g, h, x) \text{ solves } Q(\rho). \end{array}$$

(BL) is a technical reformulation of the bilevel problem in which the leader finds the best possible paths to be proposed, while anticipating route choices. Basically, $Q(\rho)$ describes the route choice behavior of both the social and selfish travelers for a given ρ.

In the remainder, we refer to *parametric optimization problem* $Q(\rho)$ as the *lower-level problem*. Here, ρ is a parameter in the lower-level problem but a variable in the *upper-level problem*. Note that in (BL) both lower-level variables (g, h, x) as well as upper-level variable ρ appear as variables. Even in case there is no upper bound with respect to ρ, i.e.,

the social travelers can be routed onto any path, the problem (BL) might be difficult to solve. In the upcoming sections we rewrite and (numerically) solve (BL) as a single-level optimization problem.

5.3 Parametric analysis

Based on reformulation (BL) of previous section, one basically needs to find an appropriate $\rho \in \Xi$ so that the corresponding (g, h, x) that solves $Q(\rho)$ minimizes total travel time $\varphi(x)$. In this chapter, we apply parametric analysis with respect to problem $Q(\rho)$, i.e., we investigate the 'behavior' of (g, h, x) that solves $Q(\rho)$ under perturbations in ρ.

The purpose of the analysis is, from a computational perspective, as follows. The parametric analysis provides an estimate for the rate of change in the lower-level solution as the lower-level parameter (which is an upper-level variable) changes (Patriksson, 2004). Then, we use this estimate to move into a direction that decreases the total travel time. In this and next section (Section 5.4), we provide the parametric analysis of the lower-level problem. The results of these sections are used to reformulate and solve (BL) as single-level optimization problem (Section 5.5).

5.3.1 Notation, definitions and preliminary results

We introduce notations that correspond to lower-level problem $Q(\rho)$ (see (5.4)) with parameter ρ:

$$v(\rho) = \min \{z(\rho, g, x) \mid (g, h, x) \in \mathcal{F}\},$$

$$S(\rho) = \{(g, h, x) \mid (g, h, x) \text{ is a global minimizer of } Q(\rho)\}.$$

We refer to $\mathcal{F}$ as the *feasible set*, $v(\rho)$ as the *optimal value function*, and to $S(\rho)$ as the *solution set* at ρ.

To study the parametric problem $Q(\rho)$, we introduce definitions that describe the behavior of functions. In this chapter, we consider both single and multi-valued functions (or: mappings). A multi-valued function F assigns to each $\varepsilon \in X \subseteq \mathbb{R}^n$ a possibly empty subset $F(\varepsilon) \subseteq Y \subseteq \mathbb{R}^m$. We denote by $\text{dom}(F) := \{\varepsilon \in X | F(\varepsilon) \neq \emptyset\}$ the domain of multifunction F. We further define for $\tau > 0$, $\delta > 0$, the neighborhoods $U_\tau(F(\varepsilon^0)) := \left\{x \in \mathbb{R}^m | \|x - x'\| < \tau \text{ for some } x' \in F(\varepsilon^0)\right\}$ and $U_\delta(\varepsilon) := \{x \in \mathbb{R}^n | \|x - \varepsilon\| < \delta\}$.

We use the following definitions (Bank et al., 1983; Robinson, 1982):

Definition 2. *A multifunction $F(\varepsilon)$ is said to be:*

1. *closed at ε^0 if for any sequences ε^l, x^l, $l \in \mathbb{N}$, with $\varepsilon^l \to \varepsilon^0$, $x^l \in F(\varepsilon^l)$, the condition $x^l \to x^0$ implies $x^0 \in F(\varepsilon^0)$;*

2. *upper/outer semicontinuous at ε^0, if for any $\tau > 0$, exists $\delta > 0$ such that*

$$F(\varepsilon) \subseteq U_\tau(F(\varepsilon^0)), \qquad \text{for all } \varepsilon \in U_\delta(\varepsilon^0);$$

3. *lower/inner semicontinuous at ε^0, if for any $\tau > 0$, exists $\delta > 0$ such that*

$$F(\varepsilon^0) \subseteq U_\tau(F(\varepsilon)), \qquad \text{for all } \varepsilon \in U_\delta(\varepsilon^0);$$

4. *(locally) upper Lipschitz continuous at ε^0 if there exists a $\delta > 0$ and Lipschitz constant $L < \infty$ such that*

$$F(\varepsilon) \subseteq F(\varepsilon^0) + L\|\varepsilon - \varepsilon^0\|\mathbb{B}, \qquad \text{for all } \varepsilon \in U_\delta(\varepsilon^0),$$

where $\mathbb{B} := \{x \in \mathbb{R}^m \mid \|x\| \leq 1\}$;

5. *(locally) Lipschitz continuous at ε^0 if there exists a $\delta > 0$ and Lipschitz constant $L < \infty$ such that*

$$F(\varepsilon) \subseteq F(\varepsilon') + L\|\varepsilon - \varepsilon'\|\mathbb{B}, \qquad \text{for all } \varepsilon, \varepsilon' \in U_\delta(\varepsilon^0).$$

The following results are from Eikenbroek et al. (2018). Here, $S^x(\rho), S^g(\rho), S^h(\rho)$ denote the projections of $S(\rho)$ onto the x, g, and h-space, respectively.

Theorem 1 (Eikenbroek et al. (2018))**.**

1. *$S(\rho^0) \neq \emptyset$ for all $\rho^0 \in \Xi$;*
2. *$S(\rho^0)$, $S^g(\rho^0)$, and $S^h(\rho^0)$ are (polyhedral) convex sets for each $\rho^0 \in \Xi$;*
3. *$S^x(\rho^0)$ is a singleton for each $\rho^0 \in \Xi$, i.e., $S^x(\rho^0) = \{x(\rho^0)\}$, and $x(\rho)$ is a continuous function on Ξ, i.e., $x(\rho)$ is upper and lower semicontinuous at each $\rho^0 \in \Xi$. Moreover,*
$$\psi(\rho) := \left\{ \ \rho^T g \ \middle| \ g \in S^g(\rho) \ \right\}$$
is uniquely determined at each $\rho^0 \in \Xi$;
4. *The mappings $S(\rho)$, $S^g(\rho)$, and $S^h(\rho)$, are upper semicontinuous at each $\rho^0 \in \Xi$;*
5. *The mapping $S(\rho)$ is not injective, i.e., different $\rho^0 \neq \rho^1 \in \Xi$ might have a common solution $(g^0, h^0, x^0) \in S(\rho^0) \cap S(\rho^1)$.*

We underline that in our setting we cannot expect Theorem 1 to be stronger in the sense that $S^g(\rho)$ is also lower semicontinuous at each ρ^0. The route flow set

$$S^g(\rho) = \left\{ \ g \in \mathbb{R}^{|\mathcal{P}|} \ \middle| \ \exists h, \ \begin{array}{l} \Lambda g = d^s, \Lambda h = d^n, \Delta(g+h) = x(\rho), \\ g \geq 0, h \geq 0, \rho^T g = \psi(\rho) \end{array} \right\},$$

is a polyhedral convex set at each $\rho^0 \in \Xi$. So, although the x-part of the solution to $Q(\rho)$ is uniquely determined, there might be multiple route flow solutions that correspond to a single link flow solution $x(\rho)$. In the context of perturbations of a parameter in the link-cost and/or demand vector, the route flow set is a continuous mapping relative to its domain (Lu & Nie, 2010), given that the link flow changes continuously. We demonstrate later (Section 5.4) it is in fact the absence of lower semicontinuity of $S^g(\rho)$ at some $\rho^0 \in \Xi$ that causes the practical difficulties for the calculation of the directional derivative $x'(\rho^0; r)$ of $x(\rho)$ at ρ^0 in direction $r \in \mathbb{R}^{|\mathcal{P}|}$.

Remark 1. *To improve readability, we assume for now that $d^s = d$ (i.e., $d^n = 0$). We prove in Section 5.4.4 that we can extend the results to the more general case $d^n \neq 0$.*

5.3.2 Directional derivative of the optimal value function

This subsection covers the parametric analysis of the optimal value function $v(\rho)$. We show that the directional derivative $v'(\rho^0; r)$ of $v(\rho)$ exists for any $\rho^0 \in \Xi$ and direction r, $\|r\| = 1$, and we use - in Section 5.4.3 - the sensitivity of the optimal value function to find a specific route flow.

Definition 3 (Directional derivative)**.** *A function $f(\rho)$ is said to be directionally differentiable at $\rho^0 \in dom(f)$ in direction r, $\|r\| = 1$, if*

$$f'(\rho^0; r) := \lim_{t \to 0^+} \frac{f(\rho^0 + tr) - f(\rho^0)}{t}$$

exists.

The following proposition (Proposition 2) demonstrates that the optimal value function $v(\rho)$ is directionally differentiable at any ρ^0 for any direction r, $\|r\| = 1$. This is a well-known result in parametric optimization (see Fiacco & Ishizuka, 1990a), but the accompanying proof (provided in the Appendix) is easier in our case.

Proposition 2. *The optimal value function $v(\rho)$ is directionally differentiable at each $\rho^0 \in \Xi$ and in each direction $r \in \mathbb{R}^{|\mathcal{P}|}$, $\|r\| = 1$. In fact, $v'(\rho^0; r)$ is the optimal value that corresponds to a solution of the parametric linear program*

$$P(r): \qquad \min r^T g \quad \text{s.t.} \quad g \in S^g(\rho^0).$$

In this section, we proved that the directional derivative $v'(\rho^0; r)$ of $v(\rho)$ exists for any ρ^0 and direction r ($\|r\| = 1$). In the upcoming section, we treat the (existence and calculation of the) directional derivative of the link flows $x(\rho)$. The sensitivity analysis of $v(\rho)$ can also be used to formulate a single-level problem, see, e.g., Dempe and Zemkoho (2012) and Mordukhovich (2018).

5.4 Parametric analysis of the optimal solution

Intuitively, directional derivative $x'(\rho^0; r)$ is the rate of change of the optimal solution $x(\rho)$ at ρ^0 along r. This section investigates the existence and calculation of the directional derivative, which we use in Section 5.5 to formulate a solution method for bilevel program (BL).

In the remainder, we repeatedly use the following assumption (Assumption 2), which states that the Jacobian of the link cost function is a positive definite matrix. This assumption is stronger than necessary for some of the upcoming results, and it does not follow directly from Assumption 1 (e.g., when using the Bureau of Public Roads-function (Bureau of Public Roads, 1964) with $x_e = 0$, for some $e \in E$). See Lu (2008) for conditions that can replace Assumption 2.

Assumption 2. *Assumption 2 is said to hold at x^0 if $\nabla_x^2 z_0(x) (= \nabla_x l(x))$ is a positive definite matrix at x^0.*

Let $\rho^0 \in \Xi$ be in the remainder of this section a *reference value* and we consider *reference point* (ρ^0, x^0), with $x^0 \in S^x(\rho^0)$.

We prove that the *Karush-Kuhn-Tucker* (KKT)-set mapping corresponding to $Q(\rho)$ is an upper Lipschitz continuous multifunction at ρ^0, given that Assumption 2 holds at x^0. Consider therefore the system of KKT optimality conditions for $Q(\rho)$. For each ρ, this system can be written as

$$\begin{array}{ll} l(x) - \beta = 0 & g^T\gamma = 0 \\ \Delta^T\beta - \gamma - \Lambda^T\lambda + \rho = 0 & (g,x) \in \mathcal{F}, \end{array} \tag{5.5}$$

with accompanying Lagrange multiplier vector $\phi := (\beta, \lambda, \gamma), \gamma \geq 0$. The KKT-set mapping $S_{KKT}(\rho)$ is the function that maps ρ onto the set of (g, x, ϕ) that satisfies (5.5), i.e., for $\rho \in \Xi$:

$$S_{KKT}(\rho) = \left\{ \ (g,x,\phi) \ \middle| \ (g,x,\phi) \text{ satisfies (5.5)} \, , \gamma \geq 0 \ \right\}.$$

In our context, the Lagrange multiplier vector ϕ is uniquely determined at ρ^0. Indeed, for each fixed ρ^0, $x^0 = S^x(\rho^0)$ is a singleton, which implies that $l(x^0)$ and thus β^0 are uniquely determined (with $(g^0, x^0, \phi^0) \in S_{KKT}(\rho^0)$). Whereas ρ^0 is fixed, and there exists at least one $p \in \mathcal{P}_k$, for which $\gamma_p^0 = 0$ (which is true by $d_k > 0$) for all $k \in \mathcal{K}$, it follows that also λ^0 (and thus γ^0) are uniquely determined given ρ^0.

We state the main result of this section (Theorem 2): $S_{KKT}(\rho)$ is (locally) upper Lipschitz continuous at ρ^0. We moved the (rather technical) proof to the Appendix.

Theorem 2. *Let Assumption 2 hold at x^0, the multifunction $S_{KKT}(\rho)$ is upper Lipschitz continuous at $\rho^0 \in \Xi$.*

We need the auxiliary result of this section in the upcoming subsections to prove existence of the directional derivative $x'(\rho^0; r)$, under Assumption 2 at x^0.

5.4.1 Directional derivative of the link flow solution

This and upcoming subsections (Section 5.4.2 and 5.4.3) are devoted to treat the existence and calculation of the directional derivative

$$x'(\rho^0; r) = \lim_{t \to 0^+} \frac{x(\rho^0 + tr) - x^0}{t},$$

with $x^0 = x(\rho^0)$, since in particular the link flows are of interest for authorities (i.e., the upper-level objective function $\varphi(x)$ in (BL) is a function of x). Some of our arguments are taken from Dempe (1993) and Pang and Ralph (1996).

Let ρ^0 be the reference value and $r \in \mathbb{R}^{|\mathcal{P}|}$, $\|r\| = 1$, is an arbitrary direction. Let $t^k > 0$, $k \in \mathbb{N}$, so that $t^k \to 0$. From previous analysis (Theorem 2), we know that, if Assumption 2 holds at x^0, for each

$$(g^k, x^k, \phi^k) \in S_{KKT}(\rho^k), \quad \rho^k := \rho^0 + t^k r,$$

exists

$$(\tilde{g}^k, x^0, \phi^0) \in S_{KKT}(\rho^0) \tag{5.6}$$

so that

$$\frac{(g^k, x^k, \phi^k) - (\tilde{g}^k, x^0, \phi^0)}{t^k} \tag{5.7}$$

is a bounded sequence, and thus has (for a certain subsequence) a limit point $w = (w^g, w^x, w^\phi)$. We investigate whether w^x of w is unique and independent of the choices of t^k and $\tilde{g}^k$.

The complexity of the analysis lies in the fact that $S^g(\rho)$ is only upper semicontinuous at ρ^0. Intuitively, for some $\rho^k \to \rho^0$, not all $g \in S^g(\rho^0)$ can be *reached* by some (sub)sequence $g^k \in S^g(\rho^k)$. We follow the strategy of Dempe (1993), and introduce *reachable set* $V(S^g(\rho^0); r)$ of $S^g(\rho)$ at $\rho^0 \in \Xi$ into direction r:

$$V(r) = V(S^g(\rho^0); r) = \left\{ \quad g \in \mathbb{R}^{|\mathcal{P}|} \quad \middle| \quad \begin{array}{l} \text{exists sequence } t^k > 0, k \in \mathbb{N}, t^k \to 0, \\ \text{and } g^k \in S^g(\rho^k) \text{ so that } g^k \to g \end{array} \right\}.$$

We first show that $V(r)$ is nonempty, and that it is a subset of $SP(r)$ (and thus $S^g(\rho^0)$) (*cf.* Dempe, 1993). $SP(r)$ is the solution set corresponding to problem $P(r)$ with parameter r, i.e.,

$$SP(r) = \left\{ \; g \in S^g(\rho^0) \; \middle| \; g \text{ solves } P(r) \; \right\}.$$

Lemma 1. *For arbitrary direction r, $\|r\| = 1$:*

$$\emptyset \neq V(r) \subseteq SP(r) \subseteq S^g(\rho^0). \tag{5.8}$$

Proof. We prove the lemma in two parts. First, we prove that $\emptyset \neq V(r)$, and then we prove that $V(r) \subseteq SP(r)$. It is trivial that $SP(r) \subseteq S^g(\rho^0)$.

($\emptyset \neq V(r)$). Consider ρ^k, $k \in \mathbb{N}$, so that ρ^k converges to ρ^0. Choose $g^k \in S^g(\rho^k)$. Since $\|g^k\|$ is bounded, there exists subsequence g^{k_j} of g^k so that g^{k_j} converges to some g^0. $S^g(\rho)$ is a closed mapping at ρ^0, and thus $g^0 \in S^g(\rho^0)$. So, $V(r) \neq \emptyset$.

($V(r) \subseteq SP(r)$). Choose any $g^0 \in V(r)$. By definition, there exists $g^k \in S^g(\rho^k)$ so that $g^k \to g^0 \in S^g(\rho^0)$. In the proof of Proposition 2, we established that

$$r^T g^0 \geq v'(\rho^0; r) = \lim_{k \to \infty} \frac{v(\rho^k) - v(\rho^0)}{t^k} \geq \lim_{k \to \infty} r^T g^k = r^T g^0$$

So, $r^T g^0 = v'(\rho^0; r) = \min_{g \in S^g(\rho^0)} r^T g$. That is, $g^0 \in SP(r)$. □

In general, it holds that $V(r)$ is a proper subset of $S^g(\rho^0)$ (as we show in Example 1 in Section 5.4.3), and $V(r) = S^g(\rho^0)$ follows if $S^g(\rho)$ is lower semicontinuous at ρ^0. $S^g(\rho)$ is lower semicontinuous relative to its domain if ρ is a parameter in the link cost function (see Lu & Nie, 2010).

Lemma 2. *Let Assumption 2 hold at x^0. For direction r, $\|r\| = 1$, any limit point w of* (5.7) *satisfies the following system:*

$$\begin{aligned} 0 = \Delta^T(\nabla_x l(x^0) w^x) + \rho^0 - \Lambda^T w^\lambda - \sum_{p \in I(g^0)} (w_p^\gamma) \mathbb{1}_p; \\ \Delta w^g - w^x = 0, \Lambda w^g = 0; \\ w_p^g = 0, p \in I(g^0) : \gamma_p^0 > 0; \\ w_p^g \geq 0, p \in I^0; \\ w_p^\gamma \geq 0, p : \gamma_p^0 = 0; \\ w_p^\gamma w_p^g = 0, p \in I^0, \end{aligned} \tag{5.9}$$

for some $I^0 \subseteq I(g^0)$*, with* $g^0 \in V(r)$.

Here, $\mathbb{1}_p \in \{0,1\}^{|\mathcal{P}|}$ is the indicator vector. $I(g) \subseteq \mathcal{P}$ denotes the active index set at $g \in \mathcal{F}^g$:

$$I(g) = \left\{ \; p \in \mathcal{P} \;\middle|\; g_p = 0 \; \right\}.$$

Proof. We prove this lemma in three parts. In the first part of the proof, we prove that for $g^0 \in V(r)$, $\tilde{g}^k$ of (5.6) converges to g^0. In Part 2, we prove that the limit point w satisfies the first equality of (5.9), that w satisfies the (in)equalities of (5.9) is proven in Part 3.

(Part 1). Note that we can assume (by passing to a subsequence) that $g^k \to g^0$, i.e., $g^0 \in V(r)$. We prove that for $g^0 \in V(r)$, $\tilde{g}^k$ as in (5.6) converges to g^0. Let $g^0 \in V(r)$. By definition there exists a sequence $t^k > 0$, with $t^k \to 0$, and $g^k \in S^g(\rho^k)$ so that $g^k \to g^0$. Since

$$\|(g^k, x^k, \phi^k) - (\tilde{g}^k, x^0, \phi^0)\| \to 0,$$

and $g^k \to g^0$ as $k \to \infty$, it follows that $\tilde{g}^k \to g^0$.

(Part 2). Consider the set $S_{KKT}(\rho^k)$ for each $k \in \mathbb{N}$. Recall, for each $k \in \mathbb{N}$, and $g^k \in S^g(\rho^k)$ exists unique (x^k, ϕ^k) so that $(g^k, x^k, \phi^k) \in S_{KKT}(\rho^k)$. That is, for each k, (g^k, x^k, ϕ^k) satisfies the KKT conditions that correspond to $Q(\rho^k)$, i.e., with $\gamma^k \geq 0$,

$$\begin{array}{ll} l(x^k) - \beta^k = 0 & (g^k)^T \gamma^k = 0 \\ \Delta^T \beta^k + \rho^0 + t^k r - \Lambda^T \lambda^k - \gamma^k = 0 & (g^k, x^k) \in \mathcal{F}. \end{array} \tag{5.10}$$

Since $g^0 \in V(r)$, $g^k \to g^0$ with $g^0 \in S^g(\rho^0)$. Hence, $g^0_p > 0$ implies $g^k_p > 0$ for sufficiently large k, and thus $I(g^k) \subseteq I(g^0)$ for these k. Now, we can rewrite the first three KKT conditions in (5.10) as

$$0 = \Delta^T l(x^k) + \rho^0 + t^k r - \Lambda^T \lambda^k - \sum_{p \in I(g^0)} \gamma^k_p \mathbb{1}_p. \tag{5.11}$$

Taylor's expansion of $l(x)$ around x^0 says that

$$l(x^k) = l(x^0) + \nabla_x l(x^0)(x^k - x^0) + o(\|x^k - x^0\|), \tag{5.12}$$

where $o(\|x^k - x^0\|)/t^k$ converges to zero for $k \to \infty$.

We repeat a similar argument for $(\tilde{g}^k, x^0, \phi^0)$. We have that $\tilde{g}^k$ converges to g^0, and thus $I(\tilde{g}^k) \subseteq I(g^0)$ for large k. The KKT conditions of $Q(\rho^0)$ say that $(\tilde{g}^k, x^0, \phi^0) \in S_{KKT}(\rho^0)$ satisfies (at least) the following condition for sufficiently large k (using the uniqueness of γ^0):

$$0 = \Delta^T l(x^0) + \rho^0 - \Lambda^T \lambda^k - \sum_{p \in I(g^0)} \gamma^0_p \mathbb{1}_p. \tag{5.13}$$

Subtracting (5.13) from (5.11), and using the Taylor expansion (5.12), we obtain

$$0 = \Delta^T (\nabla_x l(x^0)(x^k - x^0)) + t^k r - \Lambda^T (\lambda^k - \lambda^0) - \sum_{p \in I(g^0)} (\gamma^k_p - \gamma^0_p) \mathbb{1}_p + o(\|x^k - x^0\|). \tag{5.14}$$

We divide (5.14) by t^k, using that the quotient in (5.7) is bounded by upper Lipschitz continuity of $S_{KKT}(\rho)$ at ρ^0, then the limit point w of (5.7) satisfies (at least) the following

equation:

$$0 = \Delta^T(\nabla_x l(x^0)w^x) + r - \Lambda^T w^\lambda - \sum_{p\in I(g^0)} (w_p^\gamma)\mathbb{1}_p. \tag{5.15}$$

(Part 3). The last KKT condition in (5.5) for ρ^k and ρ^0 says that for each $k \in \mathbb{N}$, $(g^k, x^k) \in \mathcal{F}$ and $(\tilde{g}^k, x^0) \in \mathcal{F}$. Therefore,

$$\Delta w^g - w^x = 0, \quad \text{and} \quad \Lambda w^g = 0.$$

Also, for any $p \in \mathcal{P}$, we find that (for a subsequence)

$$\left(\frac{g^k - \tilde{g}^k}{t^k}\right)_p \to w_p^g \begin{cases} = 0, & \text{if } \gamma_p^0 > 0 \\ \geq 0, & \text{if } \gamma_p^0 = 0 \text{ and s.t. exist infinitely many } k \text{ with } \tilde{g}_p^k = 0, \end{cases}$$

which yields

$$w_p^g \begin{cases} = 0, & p \in I(g^0) : \gamma_p^0 > 0 \\ \geq 0, & p \in I^0, \end{cases} \tag{5.16}$$

for some $I^0 \subseteq I(g^0)$.

By the non-negativity constraint with respect to multiplier γ in (5.5), in combination with the fact that γ is a singleton for each ρ, we have that for $p \in \mathcal{P}$,

$$\left(\frac{\gamma^k - \gamma^0}{t^k}\right)_p \to w_p^\gamma \quad \geq 0, \quad \text{if } \gamma_p^0 = 0.$$

Finally, note that also a complementarity condition arises:

$$w_p^\gamma w_p^g = 0, \quad \text{for all } p \in I^0. \tag{5.17}$$

□

We recall that, in order to determine whether directional derivative $x'(\rho^0; r)$ exists, we have to show that the limit point w^x of $\frac{x^k - x^0}{t^k}$ does not depend on choices of $t^k, g^k, \tilde{g}^k$, and $I^0 = I^0(\tilde{g}^k)$. Based on the result as presented in Lemma 2, even in the case that $V(r)$ is a singleton, different choices of I^0 could possibly lead to different solutions w^x of (5.9). In the following section, we present a method that finds $x'(\rho^0; r)$ without the trouble finding an appropriate I^0.

5.4.2 A quadratic program reformulation

Recall reference point (ρ^0, x^0). As mentioned, even if $V(r)$ is a singleton ($V(r) = \{g^0\}$), different $I^0 \subseteq I(g^0)$ in (5.16), (5.17), might be possible, which makes it difficult to calculate (a) limit point w. In this subsection, we demonstrate that, under the assumption that $V(r) = \{g^0\}$, w^x is actually independent of I^0 and can be found efficiently by solving a convex optimization problem.

Before we continue, we define $T_{\mathcal{F}}(g,x)$ as the *tangent cone* to $\mathcal{F}$ at $(g,x) \in \mathcal{F}$, i.e.,

$$T_{\mathcal{F}}(g,x) = \left\{ \; (w^g, w^x) \in \mathbb{R}^{|\mathcal{P}|} \times \mathbb{R}^{|E|} \;\middle|\; \begin{array}{ll} \Lambda w^g = 0 & \\ \Delta w^g - w^x = 0 & \\ w^g_p \geq 0 & p \in I(g) \end{array} \right\}.$$

We introduce the following parametric (convex) quadratic optimization problem (with parameter (g^0, r), and for now $V(r) = \{g^0\}$):

$$QP(g^0, r): \qquad \min_w \frac{1}{2}(w^x)^T A w^x + r^T w^g \quad \textit{s.t.} \quad (w^g, w^x) \in \mathcal{C}(g^0, x^0, \phi^0),$$

where $A := \nabla_x l(x^0) = \nabla^2_x z_0(x^0)$, and

$$\mathcal{C}(g^0, x^0, \phi^0) := T_{\mathcal{F}}(g^0, x^0) \cap T_{\mathcal{D}(\phi^0)}(g^0)$$

is the *critical cone* to $\mathcal{F}$ at (g^0, x^0, ϕ^0). Here,

$$T_{\mathcal{D}(\phi^0)}(g^0) = \left\{ \; w^g \in \mathbb{R}^{|\mathcal{P}|} \;\middle|\; w^g_p = 0, p \in \mathcal{P} : \gamma^0_p > 0 \; \right\}$$

is the tangent cone to

$$\mathcal{D}(\phi^0) = \left\{ \; g \in \mathbb{R}^{|\mathcal{P}|} \;\middle|\; g_p = 0, p \in \mathcal{P} : \gamma^0_p > 0 \; \right\},$$

at g^0. Under Assumption 2 at x^0, $QP(g^0, r)$ is a convex problem (strictly convex in w^x).

Lemma 3. *Let Assumption 2 hold at x^0. For direction r, $\|r\| = 1$, for which $V(r) = \{g^0\}$, w^x of any limit point w of* (5.7) *is the (global) optimal solution of $QP(g^0, r)$.*

Proof. Consider a limit point w of (5.7). We prove that the w^x-part of w is the optimal solution of $QP(g^0, r)$ with $V(r) = \{g^0\}$. Therefore, we first show that w^x, with accompanying w^g, is a feasible solution of $QP(g^0, r)$, then we prove (w^g, w^x) is a global optimal solution of $QP(g^0, r)$.

(Feasibility). For given direction r and $V(r) = \{g^0\}$, with $g^k \to g^0$, we note that for sufficiently large k, $(g^k, x^k) \in S(\rho^k)$ are also optimal solutions to

$$\tilde{Q}(\rho): \qquad \min z_0(x) + \rho^T g \quad \textit{s.t.} \quad (g,x) \in \tilde{\mathcal{F}} := \mathcal{F} \cap \mathcal{D}(\phi^0),$$

with $\rho = \rho^k$. So, $x^k \in \tilde{\mathcal{F}}^x$ (the projection of $\tilde{\mathcal{F}}$ onto the x-space) for all these k, and therefore w^x of any limit point w of (5.7) satisfies $w^x \in T_{\tilde{\mathcal{F}}^x}(x^0)$. Since $\tilde{\mathcal{F}}^x$ is a polyhedral set (the projection of a polyhedral set is a polyhedral set), $x^1 = x^0 + \alpha w^x \in \tilde{\mathcal{F}}^x$ for some $\alpha > 0$. Hence, exists $g^1 \in \tilde{\mathcal{F}}^g$ so that $\Delta g^1 = x^1$ (see Rockafellar & Wets, 2009, Theorem 6.43). Now, let $\bar{w}^g = \frac{g^1 - g^0}{\alpha}$, then $\bar{w}^g \in T_{\tilde{\mathcal{F}}}(g^0)$, since $g^1 = g^0 + \alpha \bar{w}^g \in \tilde{\mathcal{F}}^g$. In particular, it holds that $\bar{w}^g_p \geq 0$ for all $p \in I(g^0)$. Thus, limit point w^x of (5.7) is in the feasible set $\mathcal{C}^x(g^0, x^0, \phi^0)$. We underline that $\bar{w}^g$ is different from the w^g-part of w in (5.7), i.e., it might hold that $\bar{w}^g \neq w^g$.

(Optimality). We showed that w^x of the limit point of (5.7) with an accompanying $\bar{w}^g$ is a feasible solution of $QP(g^0, r)$. Now, we demonstrate that $(\bar{w}^g, w^x)$ is the optimal solution of $QP(g^0, r)$.

Note from (5.15) that there exists w^ϕ so that (w^x, w^ϕ) satisfies

$$\Delta^T(\nabla_x l(x^0)w^x) + r - \Lambda^T w^\lambda - \sum_{p\in I(g^0)} (w_p^\gamma)\mathbb{1}_p = 0$$

Then let $(u^g, u^x) \in \mathcal{C}(g^0, x^0, \phi^0)$ be arbitrary, we find that

$$\begin{aligned} 0 &= (\Delta^T(\nabla_x l(x^0)w^x) + r)^T u^g - (\Lambda^T w^\lambda)^T u^g - \sum_{p\in I(g^0)} ((w_p^\gamma)\mathbb{1}_p)^T u^g \\ &\leq (\Delta^T(\nabla_x l(x^0)w^x) + r)^T u^g, \end{aligned}$$

which is exactly the first-order optimality condition of convex problem $QP(g^0, r)$. Note that the latter inequality holds whereas $w_p^\gamma < 0$ for some $p \in I(g^0)$ implies that $\gamma_p^0 > 0$, and thus $u_p^g = 0$. Since $\mathcal{C}(g^0, x^0, \phi^0)$ is a polyhedral cone, and by strict convexity of the objective function in $QP(g^0, r)$ with respect to w^x, the limit point w^x is contained in the optimal solution w (unique with respect to w^x) of $QP(g^0, r)$. We show in the proof accompanying Lemma 4 that the optimal solution is bounded. □

For any given direction r ($\|r\| = 1$), in combination with the extra assumptions that $|V(r)| = 1$ and Assumption 2 holds at x^0, we proved that the directional derivative $x'(\rho^0; r)$ exists. This directional derivative is the optimal solution (with respect to w^x) of $QP(g^0, r)$ with $V(r) = \{g^0\}$. An opportunity to force uniqueness of $V(r)$ (and also $S^g(\rho^0)$) is to include a regularization term in the objective function of the lower-level problem.

5.4.3 $V(r)$ not a singleton

The more interesting case occurs when $V(r)$ is not a singleton. Note that only a finite number of different $I(g^0)$, $g^0 \in V(r)$, can occur, and, under Assumption 2 at x^0, finitely many w^x exist.

The previous analysis in Section 5.4.2 relied on the choice of $g^0 \in V(r)$. One might ask the question whether we can choose any $g^0 \in S^g(\rho^0)$, and solve $QP(g^0, r)$ to obtain directional derivative w^x, if it exists. The following example illustrates that an arbitrary $g^0 \in S^g(\rho^0)$ may lead to an unbounded solution of $QP(g^0, r)$.

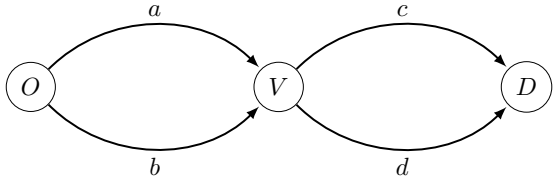


Figure 5.1: Example traffic network

Example 1 (Unbounded Solutions)**.** *In this example, we show that optimization program $QP(g^0, r)$ with $g^0 \in S^g(\rho^0) \setminus V(r)$, may have a corresponding unbounded solution.*

Figure 5.1 shows the single OD pair ($|\mathcal{K}| = 1$) network we consider. The network has 4 *links with travel time function* $l_e(x_e) = x_e$ *for all* $e \in E$. *Demand for the OD pair is* 1. *The paths*

$$p_1 = \{a, c\}, \quad p_2 = \{a, d\}, \quad p_3 = \{b, c\}, \quad \text{and} \quad p_4 = \{b, d\},$$

connect the OD pair (O, D). *Define* $\rho = (\rho_{p_1}, \rho_{p_2}, \rho_{p_3}, \rho_{p_4})$, *and let*

$$\rho(t) = t \cdot r, \quad \text{with } r = (1, 0, 0, 0), \quad \text{and } t \in [0, 1],$$

for the sake of this example. We solve $Q(\rho(0))$*: the traditional user equilibrium problem (Beckmann et al., 1956). We denote this solution with respect to* x *by* x^n *and find*

$$x^n = (x_a^n, x_b^n, x_c^n, x_d^n) = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right).$$

Since the link cost functions are strictly increasing, we find the optimal solution vector $x(\rho(t))$ *as a function of* t*:* $x(\rho(t)) = x^n, t \in [0, 1]$. *Consider* $S^{g_{p_1}}(\rho(t))$, *the route flow solution* g *on path* p_1, *as a multifunction of* t*:*

$$S^{g_{p_1}}(\rho(t)) = \begin{cases} [0, \frac{1}{2}] & \text{if } t = 0; \\ 0 & \text{if } t \in (0, 1]. \end{cases}$$

It is clear that $S^{g_{p_1}}(\rho(t))$ *is not a lower semicontinuous function at* $t = 0$. *Moreover, choose* $g^0 \in S^g(\rho(0))$ *so that* $g^0_{p_1} > 0$. *It is easy to check that* $QP(g^0, r)$ *gives an unbounded solution for* $r = (1, 0, 0, 0)$. *In fact,* $g^0 \notin V(r)$ *and observe that* g^0 *is not a solution of* $P(r)$.

Example 1 illustrates the practical difficulties calculating the directional derivative. In fact, if we choose $g^0 \in S^g(\rho^0)$ arbitrarily, we might not be able calculate $x'(\rho^0; r)$ using $QP(g^0, r)$ (even if it exists - see Theorem 3). We should select therefore $g^0 \in S^g(\rho^0)$ carefully. From a practitioner's perspective, this result is undesirable since some $g^0 \in S^g(\rho^0)$ is often a by-product of the algorithm that solves $Q(\rho^0)$. In the upcoming analysis, we prove that $g^0 \in S^g(\rho^0)$ could be selected so that $g^0 \in SP(r)$.

Lemma 4. *Let Assumption 2 hold at* x^0. *For arbitrary* r, $\|r\| = 1$, $QP(g^0, r)$, *with* $g^0 \in SP(r)$, *has a bounded solution* w *which is unique in* w^x.

Proof. Let $g^0 \in SP(r)$, and $(g^0, x^0, \phi^0) \in S_{KKT}(\rho^0)$. From Corollary 2.1 in Lee et al. (2005) it follows that $QP(g^0, r)$ has a solution if and only if

$$\left.\begin{array}{l} (u^g, u^x), (w^g, w^x) \in \mathcal{C}(g^0, x^0, \phi^0) \\ (u^x)^T A u^x = 0 \end{array}\right\} \Rightarrow (u^x)^T A w^x + r^T u^g \geq 0. \tag{5.18}$$

By Assumption 2, A is a positive definite matrix, and $(u^x)^T A u^x = 0$ implies $u^x = 0$ and it automatically follows that $(u^x)^T A w^x = 0$. Suppose now that the right-hand side of (5.18) is not satisfied, i.e., $r^T u^g < 0$ for some $(u^g, u^x) \in \mathcal{C}(g^0, x^0, \phi^0)$. Note that

$$\nabla_x z_0(x^0)^T u^x + (\rho^0)^T u^g = 0, \tag{5.19}$$

for any $(u^g, u^x) \in \mathcal{C}(g^0, x^0, \phi^0)$ (see Luo et al., 1996, p. 225). Since $u^x = 0$, by (5.19), $(\rho^0)^T u^g = 0$. So, for small $t > 0$, $(g^0 + tu^g) \in S^g(\rho^0)$ and $r^T(g^0 + tu^g) < r^T g^0$, which contradicts that $g^0 \in SP(r)$. The uniqueness of w with respect to w^x can then be concluded from the fact that A is positive definite matrix and that $\mathcal{C}(g^0, x^0, \phi^0)$ is a polyhedral cone. □

Hence, selecting $g^0 \in SP(r)$ makes that the issue as described in Example 1 cannot occur. Now, we prove the main result of the chapter. For direction r, rather than explicitly using $V(r)$, we can choose an arbitrary $g^0 \in SP(r)$ to calculate directional derivative $x'(\rho^0; r)$ of $x(\rho)$ at ρ^0.

Theorem 3. *Let Assumption 2 hold at x^0. For arbitrary direction r, $\|r\| = 1$, $x'(\rho^0; r)$ exists and is the optimal solution (with respect to w^x) of optimization problem $QP(g^0, r)$, $g^0 \in SP(r)$.*

Proof. Based on Lemma 3 and 4, we only need to prove that for any r the solution w^x of w that corresponds to $QP(g^0, r)$ is independent of the choice $g^0 \in SP(r)$. Assume r to be fixed, and let $g^1 \neq g^2 \in SP(r)$. Suppose $(w^{g,1}, w^{x,1})$ solves $QP(g^1, r)$, and $(w^{g,2}, w^{x,2})$ solves $QP(g^2, r)$, but $w^{x,1} \neq w^{x,2}$. Note that both problems have an optimal solution by Lemma 4.

We may assume, without loss of generality, that

$$\frac{1}{2}(w^{x,1})^T A(w^{x,1}) + r^T w^{g,1} \leq \frac{1}{2}(w^{x,2})^T A(w^{x,2}) + r^T w^{g,2}.$$

Since $w^{x,1} \neq w^{x,2}$, and the optimal solution of $QP(g^2, r)$ is unique with respect to $w^{x,2}$, we have

$$\frac{1}{2}(w^{x,1})^T A(w^{x,1}) + r^T w^{g,1} < \frac{1}{2}(w^{x,1})^T A(w^{x,1}) + r^T \bar{w}^{g,2}, \tag{5.20}$$

for all $\bar{w}^{g,2}$ so that $(\bar{w}^{g,2}, w^{x,1}) \in \mathcal{C}(g^2, x^0, \phi^0)$. It directly follows from (5.20) that $r^T w^{g,1} < r^T \bar{w}^{g,2}$ for all such $\bar{w}^{g,2}$, given that there exist such $(\bar{w}^{g,2}, w^{x,1}) \in \mathcal{C}(g^2, x^0, \phi^0)$.

Note that for all sufficiently small $\alpha > 0$, $g^1 + \alpha w^{g,1} \in \tilde{\mathcal{F}}$. Hence, for any such α, let

$$\bar{w}^{g,2} = \frac{g^1 + \alpha w^{g,1} - g^2}{\alpha}.$$

Then, $g^2 + \alpha \bar{w}^{g,2} \in \tilde{\mathcal{F}}$, hence $(\bar{w}^{g,2}, w^{x,1}) \in \mathcal{C}(g^2, x^0, \phi^0)$. Since $r^T(g^1 + \alpha w^{g,1}) = r^T(g^2 + \alpha \bar{w}^{g,2})$, it follows that $r^T g^1 > r^T g^2$, which contradicts that $g^1 \in SP(r)$. □

Theorem 3 proves that $x'(\rho^0; r)$ exists for any ρ^0 in any direction r, $\|r\| = 1$, provided that Assumption 2 holds globally (i.e., for all $x(\rho^0)$ with $\rho^0 \in \Xi$). Now, for $\rho^0 \in \Xi$, we can estimate $x(\rho^1) \approx x^0 + tx'(\rho^0; r)$, with $\rho^1 = \rho^0 + tr$, $t > 0$ small, and $\|r\| = 1$. To do so, we have to choose $g^0 \in SP(r)$, and subsequently solve $QP(g^0, r)$. We use this result to formulate an optimization method for (BL) in Section 5.5.

We compare the result of Theorem 3 with Theorem 2 in Ralph and Dempe (1995). There, the directional derivative of a solution of a parametric nonlinear program (with a locally unique minimizer) can be calculated (under a constraint qualification) by selecting a suitable KKT multiplier as a solution of auxiliary program. In our case, we have a non-unique solution, and need a linear program to find directional derivative $x'(\rho^0; r)$ of the link flows $x(\rho)$ at ρ^0.

5.4.4 General results

In previous sections, we assumed $d^n = 0$. We extend the results to the case $d^n \neq 0$. We omit the corresponding proofs which are straightforward extensions of the proofs in previous sections.

For $\rho^0 \in \Xi$ and r, $\|r\| = 1$, arbitrary, and $(g^0, h^0, x^0) \in S(\rho^0)$, the linear program

$$P(r): \qquad \min_{g,h} r^T g \quad \textit{s.t.} \quad (g,h) \in S^{(g,h)}(\rho^0),$$

finds $(g^0, h^0) \in SP(r)$. The quadratic (convex) optimization problem to find directional derivative $x'(\rho^0; r)$ of $x(\rho)$ at ρ^0 in direction r corresponds to

$$QP(g^0, h^0, r): \qquad \min \frac{1}{2} w^{xT} A w^x + r^T w^g \quad \textit{s.t.} \quad (w^g, w^h, w^x) \in \mathcal{C}(g^0, h^0, x^0),$$

with

$$\mathcal{C}(g^0, h^0, x^0) = \left\{ (w^g, w^h, w^x) \;\middle|\; \begin{array}{ll} w_p^g \geq 0 & p \in \mathcal{P}_{g,1} \\ w_p^g = 0 & p \in \mathcal{P}_{g,2} \\ w_p^h \geq 0 & p \in \mathcal{P}_{h,1} \\ w_p^h = 0 & p \in \mathcal{P}_{h,2} \\ \Delta(w^g + w^h) - w^x = 0 & \\ \Lambda w^g = 0 & \\ \Lambda w^h = 0 & \end{array} \right\}.$$

Here, we decompose path set $I(g^0) \subseteq \mathcal{P}$, $I(h^0) \subseteq \mathcal{P}$, as follows

$$\begin{aligned} \mathcal{P}_{g,1} &= \{p \in \mathcal{P}_k, k \in \mathcal{K} \mid p \in I(g_0), (c_p(x^0) + \rho_p^0) - \min_{q \in \mathcal{P}_k} (c_q(x^0) + \rho_q^0) = 0\}, \\ \mathcal{P}_{g,2} &= I(g^0) \setminus \mathcal{P}_{g,1}, \\ \mathcal{P}_{h,1} &= \{p \in \mathcal{P}_k, k \in \mathcal{K} \mid p \in I(h_0), c_p(x^0) - \min_{q \in \mathcal{P}_k} c_q(x^0) = 0, \} \\ \mathcal{P}_{h,2} &= I(h^0) \setminus \mathcal{P}_{h,1}. \end{aligned}$$

$\mathcal{P}_{g,2}, \mathcal{P}_{h,2}$ are the path sets that consist of the paths with an accompanying positive multiplier. Note that $QP(g^0, h^0, r)$ can be interpreted as a traffic assignment problem with a restricted path set (*cf.* Patriksson, 2004). In comparison with $Q(\rho)$, the link cost function is linear, some paths might carry negative flows, and each OD pair has zero demand.

5.5 Algorithm and numerical experiments

Thus far, we proved the existence of the directional derivative of the link flows under perturbations in the parameter, and found a constructive method to calculate it. In this section, we solve optimization problem (BL) using a feasible descent method. The algorithm is so that we solely need to solve convex optimization problems and, thus, it can be implemented in standard optimization toolboxes.

5.5.1 Algorithm

Consider optimization problem (BL). We can reformulate it as (BL'), a *nonsmooth* optimization program in which x is an implicit function of ρ, i.e.,

$$(BL'): \qquad \min_{\rho} \varphi(x(\rho)) \quad \textit{s.t.} \quad \rho \in \Xi.$$

Consider ρ^0 with solution $x^0 = x(\rho^0)$ of lower-level problem $Q(\rho^0)$. We proved that the directional derivative $x'(\rho^0; r)$ into direction r exists, i.e., for $t > 0$ small,

$$\begin{aligned} \varphi(x(\rho^0 + tr)) - \varphi(x^0) &= \nabla_x \varphi(x^0)^T (x(\rho^0 + tr) - x^0) \\ &= t \nabla_x \varphi(x^0)^T x'(\rho^0; r). \end{aligned} \tag{5.21}$$

So, any direction r, $\|r\| = 1$, that satisfies $\nabla_x \varphi(x^0)^T x'(\rho^0; r) < 0$ yields a descent direction for (BL'). This allows us to formulate the necessary optimality conditions for (BL').

The calculation of a steepest descent direction r is difficult and is the optimal solution of a linear-quadratic optimization problem, which can be found using an expensive branch-and-bound technique (Bard, 1998). To reduce computational intensity and to enhance application by traffic engineers, we use an algorithm that assumes that $x(\rho)$ is differentiable at any ρ^0, i.e., $\nabla_\rho x(\rho^0)$ exists (see Josefsson & Patriksson, 2007). Algorithms that explicitly use the nonsmoothness of the objective function in (BL') can be found in Outrata et al. (2013).

At every iteration $i \in \mathbb{N}$, with iteration point $\rho^i \in \Xi$, we find a feasible descent direction by solving convex optimization problem

$$\min_{\nu} \frac{1}{2} \| -\nabla_\rho \varphi(x(\rho^i)) - \nu \|^2 \quad \textit{s.t.} \quad \nu \in D(\rho^i), \tag{5.22}$$

with feasible cone

$$D(\rho^i) = \left\{ \nu \in \mathbb{R}^{|\mathcal{P}|} \;\middle|\; \begin{array}{ll} \nu_p \geq 0 & p \in P^1 = \{p \in \mathcal{P} \mid \rho^i_p = 0\} \\ \nu_p \leq 0 & p \in P^2 = \{p \in \mathcal{P}_k, k \in \mathcal{K} \mid \rho^i_p = \varepsilon_k\} \end{array} \right\}.$$

Summarizing, the algorithm is as follows (based on Faigle et al., 2013; Josefsson & Patriksson, 2007):

Step 0 Initialize $\rho^0 \in \Xi$, $\eta > 0$ small, Armijo line search factor $\tau > 0$ and multiplier κ, set $i := 0$;

Step 1 Solve $Q(\rho^i)$ to obtain $x(\rho^i)$;

Step 2 Construct the *approximate Jacbian*, $\nabla_\rho x(\rho^i)$ by solving for each $p \in \mathcal{P}$:

- (a) let $r = \mathbb{1}_p$;
- (b) find $(g, h) \in SP(r)$, i.e., solve $P(r)$.
- (c) find w that solves $QP(g, h, r)$;
- (d) let $(\nabla x(\rho^i))_p = w^x$.

Step 3 Solve (5.22) to find ν^i;

Step 4 Use the inexact Armijo line search (using κ) to find $m \geq 0$ that satisfies :

$$\varphi(x(p^i)) \leq \varphi(x(\rho^i)) - \tau m((\nu^i)^T(\nabla_\rho \varphi(x(\rho^i))), \tag{5.23}$$

where p^i is the projection of $(\rho^i + m\nu^i)$ onto Ξ, let $\rho^{i+1} = p^i$ and $i := i + 1$, goto Step 1. If there is no such m, terminate.

5.5.2 Implementation and settings

We implemented our method in MATLAB, and adapted a path-based algorithm to solve $Q(\rho)$ for a fixed ρ. Therefore, we used an adapted version of the gradient projection method, with a quadratic approximation line search (Gentile, 2014; Perederieieva et al., 2015). We used the built-in linear programming method of MATLAB to solve $\tilde{P}(r)$ rather than $P(r)$. Here,

$$\tilde{P}(r): \qquad \min r^T g \quad \textit{s.t.} \qquad (g,h) \in \tilde{S}^{(g,h)}(\rho),$$

where $\tilde{S}^{(g,h)}$ is equivalent to $S^{(g,h)}$, except that we replace

$$\rho^T g = \psi(\rho) \qquad \text{with} \qquad \rho^T g \in [\psi(\rho) - \delta, \psi(\rho) + \delta],$$

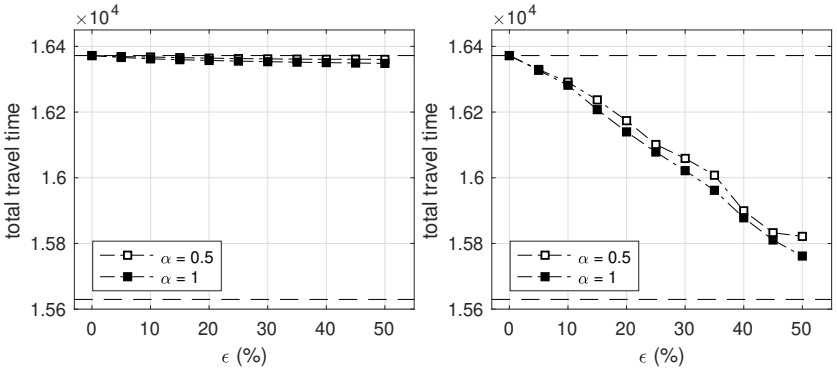
in which $\delta > 0$. To solve $QP(g,h,r)$, given (g,h,r), we use the algorithm as described by Josefsson and Patriksson (2007). In order to apply our algorithm based on sensitivity analysis, one needs to solve $Q(\rho)$ with high accuracy. Therefore, we introduced the following metric to measure accuracy (for simplicity, here assuming $d^n = 0$):

$$Acc = \frac{\sum_{k\in\mathcal{K}} \sum_{p\in\mathcal{P}_k} g_p \left((c_p(f) + \rho) - \min\left(c_p(f) + \rho\right)\right)}{\sum_{k\in\mathcal{K}} \sum_{p\in\mathcal{P}_k} g_p (c_p(f) + \rho)},$$

and stopped when an accuracy of 10^{-12}, or a maximum number of iterations, was achieved. In the remainder, we assumed $\delta = 5 \times 10^{-4}$ in $\tilde{P}(r)$, and used $\tau = 0.1$ and $\kappa = 0.5$ in the backtracking line search.

Two networks are implemented to provide insight into the potential of social routing in practice. We use the network of Nguyen and Dupuis (1984) ($|\mathcal{K}| = 4$), with the settings of Ohazulike et al. (2013) and the demand scenario of the latter paper of 400, 800, 600, and 200, respectively. To assess performance in larger networks, we used the Sioux Falls network (Transportation Networks for Research Core Team, 2019), with $|\mathcal{K}| = 528$. For the first network the path set is known a priori, in the latter network the path set needs to be constructed iteratively while solving the bilevel problem. Therefore, we add (if necessary) the k-shortest paths ($k = 2$) for each commodity every time we accept the Armijo condition (5.23). To initialize the path set, we used the path set generated while solving the user equilibrium and system optimum.

The main computational burden of the presented algorithm - compared to approaches solving NDPs - is the construction of an approximate Jacobian $\nabla_\rho x(\rho^i)$, which requires $P(r)$ and $QP(g,h,r)$ to be solved $|\mathcal{P}|$ times for each *outer iteration*. In particular for dense networks with many OD pairs this might lead to increasing run times. For example, for the Sioux Falls network, we ended with about 2050 paths in the path set. Therefore, we limited the outer iterations to 25. For practical purposes, one might relieve the computation time by aggregating zones.



(a) 25% *of the OD pairs targeted* *(b)* 50% *of the OD pairs targeted*

Figure 5.2: Impact varying social demand, acceptable travel time difference ε, and compliance rate α with respect to system performance in the Nguyen & Dupuis network.

5.6 Results and management implications

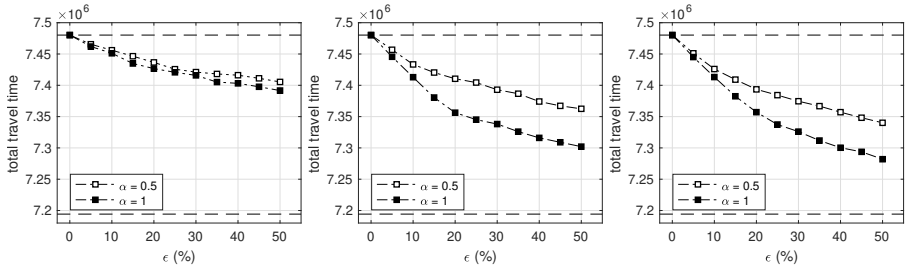
We explore the potential network impacts of a social routing service adopting the proposed strategy: we apply the algorithm (Section 5.5.1) to two test networks (see Section 5.5.2). In Section 5.6.2, we draw some preliminary conclusions about social routing for traffic management purposes.

5.6.1 Network impact

We provide insight in the potential network efficiency, by assuming varying social trip rates d^s, and acceptable travel time differences ε. In these experiments, we assume that only a portion of the travelers is receptive for advice. Receptive drivers might be unequally distributed over the network, and, therefore, we consider for each network eight social demand scenarios. We assume that 25%, 50%, 75% or 100% of the largest OD pairs (in terms of trips) can be reached or targeted by a social routing service. Furthermore, only a portion of this demand is assumed to comply with the advice, hence we assume $d^s = \alpha d$ ($d^n = (1-\alpha)d$)) for these OD pairs, with $\alpha \in \{\frac{1}{2}, 1\}$. To allow comparison with the unfair system optimum, we express the OD-pair dependent maximum detour ε as a percentage of the maximum detour needed in the system optimum (for the same OD pair). For each scenario we determine the distribution of social demand over the network by solving problem (BL').

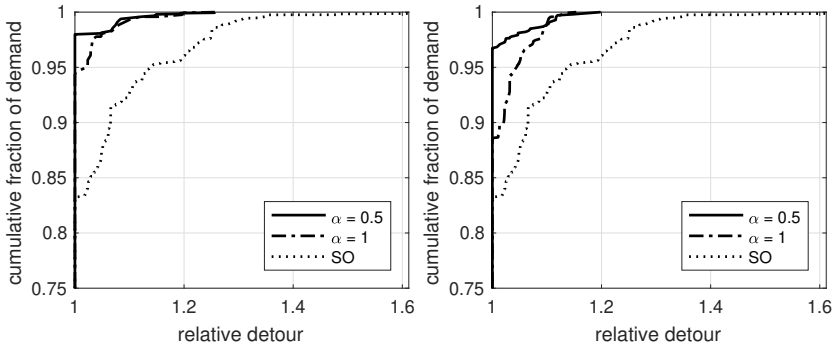
Figures 5.2 and 5.3 show the performance of the routing service (in terms of total travel time) for the Nguyen & Dupuis and Sioux Falls network, respectively, under different scenarios. In each figure, the upper and lower dashed lines depict the total travel time in user equilibrium and system optimum, respectively. In general, a larger share of social trips, and a less equitable (i.e., larger values of ε) routing strategy leads to a better performance in terms of total travel time.

When analyzing the results for the Nguyen & Dupuis network (Figure 5.2), we observe that the routing strategy is able to approach the performance of the system optimum (Figure



(a) 25% of the OD pairs targeted *(b) 50% of the OD pairs targeted* *(c) 75% of the OD pairs targeted*

Figure 5.3: Impact varying social demand, acceptable travel time difference ε, and compliance rate α with respect to system performance in the Sioux Falls network.



(a) 25% of OD pairs targeted *(b) 50% of OD pairs targeted*

Figure 5.4: Cumulative distribution of relative travel time detours compared to the fastest paths in the Sioux Falls network. Figure 5.4a corresponds to the scenario of Figure 5.3a with $\varepsilon = 50\%$, Figure 5.4b corresponds to the demand scenario of Figure 5.3b with $\varepsilon = 50\%$.

5.2b). However, targeting the right (amount of) OD pairs is crucial, since we see in Figure 5.2a almost no travel time improvement with only one OD pair reached. This can be explained by the minor detour in the system optimum for this OD pair. Further increasing the social trip rate to 75% and 100% does not substantially change performance and the corresponding results are therefore not shown. Interestingly, the compliance rate α has only limited impact on the results.

In the Sioux Falls network, the total travel time improvement is 2.7% compared to the user equilibrium (Figure 5.3); the system optimum shows an improvement of 3.8%. With a compliance rate of 50%, the strategy has a maximum improvement of 1.9% in total travel time. The results with 100% of the OD pairs targeted are comparable to the results as depicted in Figure 5.3c and therefore not shown. If only 25% of the largest OD pairs can be targeted by a routing service, improvements drop (Figure 5.3a). Again, we observe only a minor change in total travel time when OD pairs targeted increase above 50% (compare Figure 5.3b and 5.3c).

In Figure 5.4, we depict the cumulative distributions of the detours (in travel time, relative to the shortest path available) in the resulting states (assuming $\varepsilon = 50\%$) for the different demand scenarios. We also show the distribution of detours in the system optimum (SO). We note that in user equilibrium, all travelers take the fastest path (i.e., no detour) - see (5.2b). Here, we see that - although more than 50% of the drivers receive advice - only about 12% of the drivers need to take a small detour to obtain 2.4% total travel time improvement (Figure 5.4b), i.e., a major share of the social travelers is still advised to take the shortest route. At the same time, the detours, if advised, are less than 26% worse compared to the fastest path. For a system-optimal assignment, detours might potentially take 60% longer. Figure 5.4a shows that here only a very small fraction (2.1% of all trips) of social trips is needed to obtain already 1% improvement in total travel time.

5.6.2 Management implications

A real-life implementation of a social routing system adopting the proposed strategy requires a travel information service, using, e.g., a smartphone application. Based on the market penetration rate, (expected) compliance rate, and acceptable travel time differences, a central system calculates the paths for each user by solving (BL). These paths, provided to the drivers, are the best possible ones for the traffic system while meeting user constraints alongside. Based on the results of Section 5.6.1, we provide some preliminary management implications.

The numerical experiments show that a social routing system is a potential powerful measure to improve efficiency, and preserve fairness at the same time. Even if a small portion of travelers can be rerouted onto social routes, the resulting traffic state might show a major improvement in total travel time compared to the user equilibrium.

We note that the spatial distribution of the social travelers, in combination with the maximum acceptable travel time difference of users, might highly impact the strategy's performance. In the experiments, advised detours are usually fairly limited which is expected to lead to high compliance rates. In addition, travelers can be motivated to take a detour, e.g., by providing rewards. Obviously, also autonomous vehicles might be routed onto such paths (Speranza, 2018).

Even for the relatively simple setting we considered in this chapter, finding the optimal solution of the bilevel problem is highly complex. The algorithm as proposed in Section 5.5.1 finds an improving solution over the iterations. This procedure is however time-consuming. Evaluating the potential of the strategy on real-world network instances requires therefore an alternative procedure. The theoretical analysis and algorithm can nonetheless be used to assess the quality of faster heuristics that find a good solution of (BL).

An application of the social routing system in real life requires further research. First, we only considered fairness of the resulting state, but one might also evaluate the inter-state travel time differences, i.e., before and after implementation of the service (see Jahn et al., 2005). Second, we used a relatively simple procedure to construct the path set. In practice, one might consider column generation that further explores the path set while solving the bilevel problem. Finally, we focused ourselves to the equilibrium state in an assignment with static demand. Developing a similar routing strategy for the dynamic case is much more complex, in particular since a range of possible behavioral responses should be accounted for.

5.7 Conclusion

In this chapter, we consider a social routing strategy that explicitly accounts for the route choice behavior of drivers. The routing strategy asks a portion of the travelers to take a small detour for the system's benefit. Recent empirical research proved that such a strategy is implementable in a routing system in real life.

We showed that the best possible routes (with respect to efficiency) to be proposed by a routing system can be found by solving a bilevel optimization problem that anticipates the route choice behavior of compliant and non-compliant travelers. We used parametric analysis to study the behavior of the solution set of the lower-level problem as a function of the upper-level variable. Under mild conditions, we can efficiently calculate the directional derivative of the lower-level link flow solution by solving a convex quadratic optimization problem. A numerical procedure uses this directional derivative to find the paths to be proposed. The numerical experiments show the potential efficiency gain of such a system in practice. Indeed, only a small portion of the travelers need to take a fairly limited detour to achieve a substantial travel time improvement.

This chapter assumed a static setting, but finding the best possible paths to be proposed to the receptive travelers is already difficult. Nonetheless, the chapter introduces a strategy (and proves it potential) worth considering for application in a general traffic engineering context. For instance, in the case of incidents, authorities can particularly apply a similar routing strategy to mitigate the impact on the network with respect to the total travel time, but at the same time limit the detour of individual drivers.

5.8 Appendix

Proof of Proposition 2

For a single-valued function $f(\rho)$ we define, for $\rho^0 \in \text{dom}(f)$ and r, $\|r\| = 1$,

$$f'_+(\rho^0; r) := \limsup_{t \to 0^+} \frac{f(\rho^0 + tr) - f(\rho^0)}{t},$$

$$f'_-(\rho^0; r) := \liminf_{t \to 0^+} \frac{f(\rho^0 + tr) - f(\rho^0)}{t},$$

where $f'_+(\rho^0; r) = f'_-(\rho^0; r)$ holds if and only if $f'(\rho^0; r)$ exists.

Proposition 2. *The optimal value function $v(\rho)$ is directionally differentiable at each $\rho^0 \in \Xi$ and in each direction $r \in \mathbb{R}^{|\mathcal{P}|}$, $\|r\| = 1$. In fact, $v'(\rho^0; r)$ is the optimal value that corresponds to a solution of the parametric linear program*

$$P(r): \qquad \min r^T g \quad \text{s.t.} \quad g \in S^g(\rho^0).$$

Proof. Assume that $\rho^0 \in \Xi$ and r, $\|r\| = 1$, are given. We first show that $v'_+(\rho^0; r) \leq r^T g^0$ for all $g^0 \in S^g(\rho^0)$. Let $\rho(t) := \rho^0 + tr$ with $t > 0$.

Given t, let $(g^0, x^0) \in S(\rho^0)$ be arbitrary. Then, the following holds:

$$\begin{aligned} v(\rho^0) &= z_0(x^0) + (\rho^0)^T g^0 \\ &= z_0(x^0) + (\rho^0)^T g^0 + (\rho^0 + tr)^T g^0 - (\rho^0 + tr)^T g^0 \\ &\geq v(\rho^0 + tr) - tr^T g^0. \end{aligned}$$

It directly follows that $v(\rho(t)) - v(\rho^0) \leq tr^T g^0$ for any $g^0 \in S^g(\rho^0)$ and any $t > 0$. Hence,

$$v'_+(\rho^0; r) \leq \min_{g \in S^g(\rho^0)} r^T g,$$

for all $t > 0$. Similarly, we can show that $v(\rho(t)) \geq v(\rho^0) + tr^T g^t$, with $g^t \in S^g(\rho(t))$,

$$\frac{v(\rho(t)) - v(\rho^0)}{t} \geq r^T g^t, \quad g^t \in S^g(\rho(t)),\ t > 0.$$

By definition of $v'_-(\rho^0; r)$, there exists sequence $t^v > 0$, $v \in \mathbb{N}$, with $t^v \to 0$, so that

$$v'_-(\rho^0; r) = \lim_{v \to \infty} \frac{v(\rho(t^v)) - v(\rho^0)}{t^v} \geq r^T g^{t^v}, \quad g^{t^v} \in S^g \rho(t^v).$$

By the (uniform) boundedness of $S^g(\rho)$ and the upper semicontinuity of $S^g(\rho)$ at ρ^0, we can assume (for a subsequence of t^v) that $g^{t^v} \to g^0 \in S^g(\rho^0)$ and thus

$$v'_-(\rho^0; r) \geq r^T g^0 \geq \min_{g \in S^g(\rho^0)} r^T g.$$

It follows that

$$v'_-(\rho^0; r) = v'_+(\rho^0; r) = v'(\rho^0; r) = \min_{g \in S^g(\rho^0)} r^T g.$$

Note that for every r, a solution of $P(r)$ exists, and that the corresponding optimal value is finite and unique. □

Proof of Theorem 2

This appendix gives the proof of Theorem 2. Therefore, we *linearize* problem $Q(\rho)$ around *reference point* (ρ^0, x^0), with $x^0 \in S^x(\rho^0)$. The linearized Lagrange dual problem of $Q(\rho)$ (at ρ^0, x^0), with parameter ρ, is

$$\begin{array}{ll} l(x^0) + \nabla_x l(x^0)(x - x^0) - \beta = 0 & g^T \gamma = 0 \\ \Delta^T \beta - \gamma - \Lambda^T \lambda = -\rho & (g, x) \in \mathcal{F}, \end{array} \tag{5.24}$$

and we denote the corresponding mapping by

$$LS_{KKT}(\rho) := \{(g, x, \phi) \mid (g, x, \phi) \text{ satisfies (5.24)}\,, \gamma \geq 0\}.$$

By Assumption 2, $LS_{KKT}(\rho^0) = S_{KKT}(\rho^0)$ (note that (5.24) are the optimality conditions for a strictly convex quadratic program), and the solution x^0 of the KKT system that corresponds to $Q(\rho)$ is also the unique solution (with respect to x) of the KKT system (5.24).

Before we continue, we state a classic result of Mangasarian and Shiau (1987).

Theorem A (Mangasarian and Shiau (1987))**.** *Let the multifunction $F(\varepsilon)$ be defined as*

$$F(\varepsilon) = \{x \in \mathbb{R}^m \mid Ax \leq C_A\varepsilon, Bx \leq C_B\varepsilon\},$$

with matrices A, B, C_A, C_B. Let $\varepsilon^0, \varepsilon^1 \in dom(F)$. Then, there is some constant $K > 0$ so that for each $x^0 \in F(\varepsilon^0)$ exists $x^1 \in F(\varepsilon^1)$ so that

$$\|x^1 - x^0\| \leq K\|\varepsilon^1 - \varepsilon^0\|.$$

Theorem 2. *Let Assumption 2 hold at x^0, the multifunction $S_{KKT}(\rho)$ is upper Lipschitz continuous at ρ^0.*

Proof. We decompose the set $LS_{KKT}(\rho)$ into a finite number of smaller subsets. For each $\mathcal{I} \subseteq \mathcal{P}$, consider (with parameter ρ)

$$\begin{array}{ll} l(x^0) + \nabla_x l(x^0)(x - x^0) - \beta = 0 & g_p = 0, p \in \mathcal{I} \\ \Delta^T\beta - \gamma - \Lambda^T\lambda = -\rho & (g, x) \in \mathcal{F} \\ \gamma_p = 0, p \notin \mathcal{I}, & \end{array} \tag{5.25}$$

with corresponding mapping

$$LS^{\mathcal{I}}_{KKT}(\rho) := \{(g, x, \phi) \mid (g, x, \phi) \text{ satisfies (5.25)}, \gamma \geq 0\}.$$

It is trivial that, for any ρ,

$$LS_{KKT}(\rho) = \bigcup_{\mathcal{I} \subseteq \mathcal{P}} LS^{\mathcal{I}}_{KKT}(\rho).$$

Notice that only a finite number of $LS^{\mathcal{I}}_{KKT}(\rho)$ can occur. Also, for any $(g^0, x^0, \phi^0) \in LS_{KKT}(\rho)$ there is a corresponding $\mathcal{I} \subseteq \mathcal{P}$ so that $(g^0, x^0, \phi^0) \in LS^{\mathcal{I}}_{KKT}(\rho)$. $LS^{\mathcal{I}}_{KKT}(\rho)$ is a Lipschitz continuous multivalued mapping (with parameter ρ) relative to its domain $\text{dom}(LS^{\mathcal{I}}_{KKT})$ (see Theorem A).

The remainder of this proof consists of two parts. First, we show that $LS_{KKT}(\rho)$ is upper Lipschitz continuous at ρ^0. Then, we use this result to prove the claim of the theorem.

$LS_{KKT}(\rho)$ is upper Lipschitz continuous at ρ^0.
Let ρ^0 be given, we prove that there exists $\delta > 0$ so that for any $\mathcal{I} \subseteq \mathcal{P}$

$$\rho^1 \in \text{dom}(LS^{\mathcal{I}}_{KKT}) : \ \|\rho^1 - \rho^0\| < \delta \Rightarrow \rho^0 \in \text{dom}(LS^{\mathcal{I}}_{KKT})$$

holds. Otherwise, there would exists a sequence $\rho^1 \to \rho^0$, with $\rho^1 \in \text{dom}(LS^{\mathcal{I}}_{KKT})$ but $\rho^0 \notin \text{dom}(LS^{\mathcal{I}}_{KKT})$. This contradicts that the domain is closed. Indeed, $\text{dom}(LS^{\mathcal{I}}_{KKT})$ can be considered to be a projection of $LS^{\mathcal{I}}_{KKT}(\rho)$ onto the ρ-space. Since the projection of a polyhedron is a polyhedron, $\text{dom}(LS^{\mathcal{I}}_{KKT})$ is closed.

Let ρ^1 be so that $\|\rho^1 - \rho^0\| < \delta$. Let $z^1 := (g^1, x^1, \phi^1) \in LS_{KKT}(\rho^1)$. Obviously, then there exists $\mathcal{I} \subseteq \mathcal{P}, \mathcal{I} = \mathcal{I}(z^1)$. so that $z^1 \in LS^{\mathcal{I}}_{KKT}(\rho^1)$.

Since $LS^{\mathcal{I}}_{KKT}(\rho)$ is a Lipschitz continuous multifunction for any ρ in its domain, there

exists $z^0 \in LS^{\mathcal{I}}_{KKT}(\rho^0)$ so that

$$\|z^1 - z^0\| \leq K\|\rho^1 - \rho^0\|$$

for some Lipschitz constant $K > 0$, $K = K(\mathcal{I})$. Note that there are only finitely many $\mathcal{I}$. So, we can put $K = \max_{\mathcal{I}\subseteq\mathcal{P}} K(\mathcal{I})$. Since the choice of z^1 was arbitrary and $z^0 \in LS_{KKT}(\rho^0)$, the claim follows.

$S_{KKT}(\rho)$ is upper Lipschitz continuous at ρ^0.
We use arguments from Robinson (1982) in the remainder of the proof.

We make the following observations:

1. $LS_{KKT}(\rho)$ is upper Lipschitz continuous at ρ^0, i.e., exists $\delta^1 > 0$ and $K > 0$ so that

$$LS_{KKT}(\rho) \subseteq LS_{KKT}(\rho^0) + K\|\rho - \rho^0\|\mathbb{B}, \qquad \text{for all } \rho \in U_{\delta^1}(\rho^0);$$

2. $S^x(\rho)$ is in particular upper semicontinuous at ρ^0 (see Theorem 1): for any $\tau > 0$ exists $\delta^2 > 0$ so that

$$S^x(\rho) \subseteq S^x(\rho^0) + \tau\mathbb{B}, \qquad \text{for all } \rho \in U_{\delta^2}(\rho^0);$$

3. $l(x)$ ($(\Delta^T l(x))$ is continuously differentiable at x^0, i.e., for any $\eta > 0$ exists a corresponding $\delta^3 > 0$ so that

$$\|\Delta^T(l(x^1) - l(x^0) - \nabla_x l(x^0)^T(x^1 - x^0))\| \leq \eta\|x^1 - x^0\|, \qquad \text{for all } x^1 \in U_{\delta^3}(x^0).$$

Choose $\eta \leq \min\{\frac{1}{2}, \frac{1}{2K}\}$, $\tau \leq \min\{\delta^1, \delta^3\}$, and ρ^1 so that $\|\rho^1 - \rho^0\| \leq \min\{\frac{1}{2}\delta^1, \delta^2\}$.

Pick an arbitrary $z^1 \in S_{KKT}(\rho^1)$. Then

$$z^1 \in LS_{KKT}(\rho^1 + \zeta^1), \qquad \text{with} \qquad \zeta^1 = \Delta^T(l(x^1) - l(x^0) - \nabla_x l(x^0)^T(x^1 - x^0)),$$

and choose $z^0 \in LS_{KKT}(\rho^0)$ so that $\|z^1 - z^0\|$ is minimized.

From observation 2 and $\|\rho^1 - \rho^0\| \leq \delta^2$ one should notice that $\|x^1 - x^0\| \leq \tau \leq \delta^3$.

From observation 3 and $\|x^1 - x^0\| \leq \delta^3$, we arrive at $\|\zeta^1\| \leq \frac{1}{2}\|x^1 - x^0\|$. Then

$$\|\rho^1 + \zeta^1 - \rho^0\| \leq \|\rho^1 - \rho^0\| + \|\zeta^1\| \leq \frac{1}{2}\delta^1 + \frac{1}{2}\delta^1 = \delta^1,$$

since $\|x^1 - x^0\| \leq \tau \leq \delta^1$. So, $\rho^1 + \zeta^1 \in U_{\delta^1}(\rho^0)$.

Observation 1 tells us the following:

$$\begin{aligned}
\|z^1 - z^0\| &\leq K\|\rho^1 + \zeta^1 - \rho^0\| \\
&\leq K\|\rho^1 - \rho^0\| + K\|\zeta^1\| \\
&\leq K\|\rho^1 - \rho^0\| + K\frac{1}{2K}\|x^1 - x^0\| \\
&\leq K\|\rho^1 - \rho^0\| + K\frac{1}{2K}\|z^1 - z^0\|
\end{aligned}$$

where the third inequality is the result of $\|\zeta^1\| \leq \eta\|x^1 - x^0\|$ with $\eta \leq \frac{1}{2K}$.

Hence $\|z^1 - z^0\| \leq 2K\|\rho^1 - \rho^0\|$. Since $z^1 \in S_{KKT}(\rho^1)$ was chosen arbitrarily, and $z^0 \in LS_{KKT}(\rho^0) \Rightarrow z^0 \in S_{KKT}(\rho^0)$ by Assumption 2, the claim of the proof follows. $\square$

Chapter 6

Conclusion

In this thesis, we investigated the variability in urban traffic systems on various scales. We mainly focused on the variations in traffic volumes measured near signalized intersections. Variations in the order of seconds typically reflect fluctuations in arrivals and departures, while the variations on timescales longer than 5-15min are often used to characterize changing traffic conditions.

Aggregated urban traffic volumes or counts are usually studied using 24h time series, and at a single location these time series show systematic differences over various timescales. A large share of the systematic variations can be expressed using recurrent but latent temporal patterns, and only a few of these patterns express the within-day and day-to-day changes in the 24h time series - even when recurrent events occur. Apart from systematic differences, measurements also exhibit volume-dependent random fluctuations, introducing an inherent uncertainty to decision-making processes partly based on information regarding volumes such as route planning by logistics service providers.

The arrival and departure processes near signalized intersections highly determine the properties of the random variation. A statistical characterization of recorded arrival events indicates that arrival processes show a very different structure compared to the typically-assumed Poisson or renewal processes, mainly due to the periodicities in arrivals introduced by traffic signal control at upstream intersections. Although these variations show patterns on a very high resolution, when the aggregation levels of the volume measurements increase, information on the arrival process is lost in that volume noise shows similarities with random variation from standard renewal processes. In any case, the understanding of these processes is used to forecast volumes for different prediction horizons while accounting for both systematic and random variations. The statistical characterization of the noise is used, together with the underlying temporal patterns, to provide point and density estimates of urban traffic volumes 15min to 24h in advance. Such a prediction scheme is valuable in the context of proactive decision-making processes of different actors utilizing information regarding urban traffic, since volume changes may precede travel time and delay variations. We illustrated an anticipatory decision-making mechanism in the context of the demand-based traffic management measure social rerouting, where overall traffic network performance is improved while explicitly accounting for the behavior of users in response to the measure.

In this chapter, we summarize our conclusions (Section 6.1). Furthermore, we discuss

implications of our research for practice (Section 6.2), and examine topics for further research in Section 6.3.

6.1 Conclusions and discussion

During different stages of the decision-making processes of LSPs, urban traffic managers and individual road users, information regarding the conditions throughout the urban traffic system is utilized. These processes benefit from information about the evolution of the conditions, particularly when accompanied with a dynamic characterization of the uncertainty. Volume measurements of different resolutions collected near signalized intersections support such predictions and enhance anticipatory and uncertainty-aware decision making. The research aim of this thesis was therefore as follows: *Quantifying and understanding variations that occur in urban traffic volumes at different spatio-temporal levels*. In Section 1.6, we formulated a series of research questions. We discuss our findings with respect to these questions one by one.

Research question 1: To what degree do 24h urban traffic volume time series show systematic variations, and how to characterize the random variation in volume measurements?

Traffic volumes measured at regular intervals at a single location show systematic and thus predictable variability within a day and from day to day. Although within-day and between-day systematic differences are widely studied in literature in isolation (see Crawford, 2017), there is less known considering the systematic variation in 24h urban traffic volume time series over time. Indeed, changes in both the shape and height of the time series occur over the days (Crawford, 2017; Weijermars & Van Berkum, 2005), including deviations relative to the 24h pattern due to events, incidents, etc. In this thesis, we showed that a large share of the systematic variation is in fact recurrent and exhibits clear volume patterns. That is, the recurrent variations in the 15min volume measurements can be expressed using a combination of underlying recurrent temporal patterns (profiles), that due to its periodic character in theory can be predicted. Various 24h time series look different whereas they exist of profiles (representing the shape) that are subject to small yet systematic transformations (shift and scaling) changing from day to day. Longer-term profiles are designed to express the shape of a volume pattern occurring on a 24h scale, while short-term profiles represent recurrent deviations on timescales longer than 15min but shorter than 24h. Using two years of volume data collected at almost every signalized intersection throughout the Enschede traffic network, we showed that only a few recurrent and physically-meaningful profiles with natural transformations express almost all systematic variations at a point in the network, indicated by the fact that only weak serial correlation is left in the remaining residuals. Hence, 24h volume time series show a high degree of systematic variation - even in the case of events with various starting times - only revealed when simultaneously assessing the variability over various timescales.

Apart from systematic variations, many fluctuations in the volume time series in the order of several minutes can be considered to occur by chance (noise or random variation) (e.g., Bates et al., 2001; Thomas et al., 2010). A so-called noise level function incorporates knowledge regarding the noise for a wide variety of conditions. This noise level function characterizes the stochastic fluctuations around a deterministic volume estimate and pro-

vides a generic relation between the underlying systematic variations and the variance of the noise on an aggregated level. Adopting this flexible probabilistic framework allows for a full density characterization of the random variation under all conditions. Since an inferred noise level depends on the systematic volumes and vice versa, a joint estimate procedure is required and designed. It was estimated that on a network-wide level the variance of the noise is linearly dependent on the underlying systematic 15min volume with slight overdispersion compared to Poisson noise. In fact, the noise distribution widens when volumes grow and decision making occurs in an increasingly uncertain environment when road usage increases.

Research question 2: What is the influence of the arrival processes near signalized intersections on the variations in urban volumes and delays?

A major share of the variability in travel times in urban networks is determined by delays induced at signalized intersections. An understanding of the delay-contributing factors is required for decision-making processes explicitly anticipating the evolution and the accumulation of (uncertainty in) delays under various conditions. Although numerous studies have been conducted to model the operations at such intersections, many prefer a mathematically-tractable arrival process over a realistic one. Yet, the structure of the stochastic arrival process should be accurately captured to estimate the delays and the uncertainty therein. Using millions of arrival events collected throughout the Enschede traffic network, theoretical stochastic arrival models are challenged by studying arrivals on different scales. In fact, urban arrival events can be characterized as a sequence of inter-arrival times and as a continuous-time counting process, and analyzed using both a time-domain as well as a frequency-domain approach. The combination of characterizations assesses the burst and memory structure of arrivals, thereby accounting for short-term periodicities related to upstream signals, the formation of platoons, and the changing structure as traffic proceeds.

The systematic variability in network usage provides an indication for the time-varying arrival rate (see Chapter 2 and 4). At the same time, stochastic fluctuations around this estimate occur. These natural yet random variations show volume and location-invariant properties in that on a 10min scale the variability is well captured using a heteroscedastic Gaussian distribution with limited overdispersion compared to Poisson noise. Hence, the typically used aggregation scales can be considered stable in the sense that locations and conditions are easily compared without explicit consideration of the local dynamics. Using shorter aggregation intervals than 5min in this context, however, is delicate since volume measurements collected near but downstream of a signalized intersection may show substantial overdispersion even under stationary demand conditions. Nonetheless, in this case, significant changes in arrival processes can be recognized using the second-order properties of the counts.

Although the 10min stationary volume measurements at a fixed point in the network show similarities with the counts resulting from a Poisson arrival or renewal process, the latter processes fail to reflect the structure in the stochastic inter-arrival times on shorter timescales. When considering individual inter-arrival times by statistically characterizing these intervals in an urban setting, there is a higher probability of medium and high inter-arrival times. This excess probability (compared to headway distributions for freeway traffic) is introduced by traffic signal phase times, and statistically reflects a combination of variable red times and the inter-dependencies with arrival events upstream. In addition,

consecutive inter-arrival times turn out to show (weak) correlations – but this effect accumulates to a significant level when examining a multitude of vehicles.

When looking at the arrivals as a point process in time, clear periodicities appear at some measurement locations close to an upstream signalized intersection. These periodicites were shown to correspond to the cycle times of the traffic signal upstream and can significantly influence the dispersion index. Using Bartlett's spectrum, the dominant frequencies are easily recognized. Such periodicities tend to persist over long distances when volumes grow. For uninterrupted processes, the power spectrum is smooth, and the count dispersion index is a much smoother function of the aggregation level.

Although empirical data regarding delays or speeds were not available in our case, the structure of arrival processes as revealed in Chapter 3 were shown to influence delays in a simulation environment with vehicle-actuated control settings. Indeed, the distribution of delays differs systematically when comparing a real-world mirroring arrival process with the mathematically-tractable Poisson process. Particularly in low volume situations, both the mean as well as the variability in the delays are overestimated using iid exponentially distributed inter-arrival times. Real-world arrivals contain predictive behavior in that conditioning on an arrival event provides additional information about near-future events. Nonetheless, it is complex to accurately capture the structure of the arrivals since regularities appear on many different scales. In any case, failing to capture this structure for the benefit of tractability can underestimate the variations in delays and volumes and thereby have serious implications for tactical and operational decisions.

Research question 3: To what degree can the systematic variations be predicted, and how can the characterization of the random variation be used to provide probabilistic volume forecasts over various timescales?

Systematic variations are the patterns in the measurements, and is the only part of the variability that can be predicted. The presence of noise, however, makes that the realization of a volume measurement cannot be predicted exactly, i.e., many fluctuations occur by chance. Predictions should only capture the systematic changes, but uncertainty remains an integral part due to the stochastic setting. As such, the amount of random variation provides a lower bound regarding the predictability of urban traffic volumes. The quality of the predictions needs to be assessed relative to the predictability of the system – and the difference is the true systematic prediction error. This prediction error can be expressed using a relative error based on the point prediction or by using a coverage difference. The latter indicator expresses the deviation between expected and true coverage of a density forecast. It is important to evaluate densities as a whole to prevent overfitting to a single confidence level (Diebold et al., 1998; Khosravi et al., 2011), e.g., by adopting the absolute coverage difference or the pinball loss function (see Chapter 4).

Not all systematic variations can be predicted a long time in advance using straightforward exogenous variables only (e.g., time of day, day of the week), but many of the variations can be well predicted in a statistical sense, i.e., residuals compared to an initial forecast based on basic exogenous variables were shown to be highly correlated. Hence, a large share of these differences are systematic on timescales longer than 15min. A prediction scheme accounting for such regularities can be used as an indicator for the predictability of the systematic variations over various timescales. In this thesis, longer-term variations in the order of multiple hours are captured in a 24h forecast providing a prediction for a full

day before the start of the day and a remaining-day prediction that gives at any time of day a forecast for the volumes during the remainder of the day. Short-term predictions cover the next 15min to 1.5h.

Considering the 15min predictions, we found a point prediction error of $10 - 15\%$, suggesting that systematic variations can be predicted to a high degree. A large share of the variation was possible to predict well in advance, at the beginning of the day when accounting for the day-dependent characteristics. Interestingly, predictions are substantially improved over the course of the day, but short-term forecasts only provide a minor improvement compared to the remaining-day prediction. That is, many variations in the 15min volume measurements are systematic over timescales in the order of hours. The density forecasts naturally resulting from an estimate of the random variation and the point prediction error are accurate and show an absolute coverage difference of about $2 - 3\%$. Hence, not only the deterministic yet systematic volumes can be predicted, but also the stochastic fluctuations that occur around estimates (subway uncertainties) are predictable over various timescales.

Compared to the mean prediction error, urban traffic counts during the night and weekends are more difficult to predict but are characterized by very low volumes. Systematic prediction errors also increase when considering higher urban roads, for which spatial rather than temporal correlations appear - to be used as input to improve the 15min predictions (see, e.g., Ermagun et al., 2017). In addition, volume variations due to football matches relative to the 24h pattern show a high degree of regularity and can be predicted well in advance. However, when deviations compared to the short-term pattern are recognized, adaptation is difficult since only a few measurements are available to adapt the prediction in a robust manner.

Probabilistic forecasts in the form of predictive densities quantify the remaining uncertainty accompanying a prediction. A perfect prediction method forecasts only the systematic volumes, and the predictive density then accurately accounts for the random yet natural fluctuations that are unpredictable. In practice, however, not only the random variation in the counts is part of the uncertainty. Also the systematic prediction error and limitations and errors of measurements are part of the uncertainty and should therefore be incorporated in probabilistic forecasts. In theory, the systematic error can be reduced and has a significant impact on the width of the distribution.

Research question 4: What is the potential of anticipatory urban traffic management, in particular a social rerouting strategy, while accounting for different user requirements?

A social rerouting strategy improves overall network efficiency by advising socially-beneficial routes to users but needs to explicitly account for the response of travelers to advice. In fact, only some drivers can be assumed to receive and comply with advice that reroutes them onto possibly sub-optimal routes for themselves. Here, behavioral responses influence travel times and should thus be anticipated to advise routes that are acceptable. However, behavioral changes occur from travelers complying with advice but also from those that do not comply but are now confronted with changing conditions on routes as a result of changing behavior of others. Although empirical research has shown that social routing has great potential in real life, there are theoretical challenges to be addressed before a real-world implementation of a service providing social route advice. Under minor assumptions, a bilevel optimization problem can be formulated that implicitly calculates

the best possible paths (from a system's perspective) to be proposed to receptive travelers (those that use and comply with advice) and limits the realized travel time differences in the resulting state.

A critical issue in solving the bilevel problem is that the lower-level optimal solution is non-unique. Nonetheless, the directional derivative of the lower-level link flow exists. This generalized derivative can be efficiently found as a solution of a quadratic optimization problem but requires a suitable route flow solution as parameter which can be found using an auxiliary linear program. Even in a static setting the bilevel problem is difficult to solve in general. This becomes even more pressing in a dynamic environment since a range of feedback effects needs to be anticipated - further increasing the complexity.

Numerical experiments show the potential efficiency gain of a social rerouting system in practice and only a small portion of the drivers need to take a fairly limited detour to achieve a substantial travel time improvement overall. Experiments in the numerical Sioux-Falls network show that if the right origin-destination pairs are targeted by a routing service that accounts for user requirements by limiting the maximum a posteriori travel time detour, 2.7% improvement compared to the user equilibrium can be obtained. If only about 12% of the drivers take a fairly small detour of at most 20% compared to fastest path, 2.4% travel time improvement is obtained. The system optimum, not accounting for user requirements by rerouting travelers onto routes that might take 60% longer than the fastest path, shows an improvement of 3.8%. With a compliance rate of 50%, the strategy has a maximum improvement of 1.9% in total travel time. Hence, anticipatory management systems are a potential powerful measure to improve efficiency while anticipating user responses and therefore worth to be considered in a general traffic engineering context.

Summarizing, traffic volume time series support LOS predictions and show systematic and random differences over time and space. A substantial share of the temporal systematic variation in volumes can be predicted hours in advance. At the same time, the random variation introduces an inherent uncertainty to decision making but needs to be explicitly anticipated in decision problems to assure robust solutions that perform well under a range of scenarios. In an urban context, random variation in aggregated measurements mainly originates from the departure and arrival processes at signalized intersections. These processes show much more regularity on a disaggregated scale than one could expect from the typically used counts. In any case, actors operating in the context of urban traffic potentially benefit from information regarding the dynamics in LOS during their decision-making processes, as illustrated by the improvement in network efficiency using the anticipatory urban traffic management measure social rerouting.

6.2 Implications for decision making

We discuss implications of the results of this thesis for the decision-making processes of the actors under consideration: LSPs, urban traffic managers and road users employing ATIS.

LSPs involved in home delivery aim for accurate and robust predictions regarding the ETA under a range of scenarios with respect to the traffic conditions. However, they design their route plans usually hours, or even days, in advance (Agatz et al., 2008) while not all necessary information regarding the evolution of the LOS of the urban traffic system is available by then. Hence, LSPs are faced with a decision-making process that is charac-

terized by uncertainty. They use travel time predictions to reduce the uncertainty. Volume forecasts support such predictions since the counts may show much more variability and thereby allow the anticipation of travel time changes. In this thesis, we showed that many of the systematic variations including the stochastic fluctuations in urban traffic volumes throughout a traffic network can be well predicted - even in the long term. In fact, the recurrent variations and the remaining uncertainty can be predicted using long and short-term temporal patterns combined with a noise level function, and an estimate of the systematic error. The information regarding the dynamics at signalized intersections, and with volumes being an indicator for the delays under a variety of conditions, support LSPs in their offline route planning process. Furthermore, not only an estimate of the temporal evolution is available, also the accompanying inherent uncertainty can be described. Consequently, LSPs have at the beginning of the day already quite a good indicator for the variability in volumes using probabilistic forecasts. Relatively speaking, volumes during the weekend and the night seem to be more difficult to predict compared to weekdays, and in this sense LSPs should introduce more flexibility in their plan during these time windows although the impact on the delays can be expected to be limited. In any case, provided an appropriate volume-delay function and a framework for the offline routing problem with time-dependent and stochastic travel times, LSPs can use the developed prediction model to facilitate a shift towards anticipatory decision making.

Probabilistic forecasts are improved over the day, reducing the relative prediction error for the remainder of the day. Now, plans can be refined during execution since dispatchers have not only information regarding the recorded travel times of the fleet, but also have access to partial information regarding the realized volumes throughout the traffic network and an updated dynamic characterization of the (uncertainty in the) future volumes. With the largest share of the systematic variations predictable using remaining-day forecasts, LSPs should aim for plans that have sufficient degrees of freedom for adaptations throughout the day using forecasts in the order of hours - even under 'business-as-usual' conditions.

Offline route plans, and dynamic adaptations thereof, can also anticipate recurrent events which show clear patterns in time and space relative to the 24h pattern. Still, a plan must be re-evaluated based on unexpected disruptions. Significant changes in the volumes can be rapidly recognized using a noise model, providing a probabilistic characterization of the random variation over time, useful to filter deviations that are likely to impact operations. In fact, re-optimization of the route plan at every time epoch is not necessarily required when the environment remains stable. Re-planning only needs to be executed or triggered in case a significant deviation from what was expected occurs. A framework for this kind of decision making is considered by Bijl (2016).

Urban traffic managers concerned with decision making on longer temporal scales can employ the profiles that explain almost all intra-day and between-day volume variations to capture trends and thereby find projections about future volumes as input for a junction design process. Also on shorter timescales, typical volume patterns can be extracted from current data sources and serve as an input for policy evaluation and adaptation. If external factors are additionally incorporated, the profiles support policy making since the underlying volume-contributing factors over various timescales can be investigated independently rather than by studying the noisy measurements as such.

Not only typical 24h patterns exist, also recurrent events are shown to have a high degree of regularity. In this thesis, we illustrated this using football matches. Although these matches occur on different days and have varying kick-off times, the location-dependent

impacts on volumes were shown to be comparable – provided one accounts for variable starting times. Hence, these typical patterns support traffic managers in equipping them with quantitative evidence for anticipatory event-based management measures (see CROW, 2008), e.g., event-dependent traffic light control settings. Considering the variations on timescales in the order of 5-15 minutes, it is difficult to adapt predictions in case the event-based impact substantially deviates from what was expected in advance. Indeed, (one-off) events might cover only a limited timescale meaning that only a few measurements are available to update forecasts. Nevertheless, using the generic noise-level model, these type of deviations and other disruptions such as incidents and accidents can be recognized in a timely manner although future impact projections remain difficult without additional information.

A substantial share of the variation in volume time series in the order of 5-15min can be considered random. We can model this noise as being distributed according to a heteroscedastic Gaussian, with the variance as an increasing function of the underlying systematic flow. Specifically, the random variation shows slight overdispersion compared to Poisson noise and the absolute amount of noise increases when volumes grow. For monitoring and decision-making purposes, this complicates issues since actual changes in the conditions could be confused with a random fluctuation. In fact, in particular under high-volume conditions, monitoring is important and management measures might need to be deployed to prevent congestion or to mitigate its negative impact. Hence, the increasing amount of noise when volumes grow makes that decision making occurs with additional uncertainty under high-volume occasions. At the same time, the systematic error of high-volume situations in an urban setting decreases in the relative sense, i.e., one can quite accurately cover the density function of such conditions, and are in that sense well predictable.

In the context of traffic control of a corridor with a series of signalized intersections, anticipatory decisions require very short-term predictions regarding the actual arrivals of all vehicles. Indeed, to minimize consecutive delays, it is beneficial to form platoons of vehicles that face no stopping delay at a string of consecutive traffic lights (e.g., 'green wave'). In an urban setting, arrival events show a high degree of regularity introduced by upstream interruptions and the formation of platoons. Therefore, arrivals can be predicted on a very short timescale but a continuous data stream is required to do so. Control schemes in general can be improved by additionally using predictions of the actual arrival times of vehicles at all arms rather than the realized arrival times only (as is current practice with vehicle-actuated traffic signals). Also changes in arrival processes might be recognized using high-resolution information, e.g., the onset of congestion or the occurrence of incidents. Further, many manuals are still based on relatively old delay estimates based on theoretically-appealing yet naive arrival processes. Our results suggest that these arrival processes are highly volume and location dependent, and that corresponding delay estimates in manuals may need to account for such characteristics as well.

Historically-average conditions do not always provide a reliable estimate of the LOS for users to make optimal travel decisions (Zhang et al., 2007). Indeed, the largest share of the systematic volume variation is predictable using remaining-day forecasts in the order of hours, indicating that the information provided by ATIS as well as the advice by such a system should be adapted not only on a day-to-day basis but also throughout the day. Relatively speaking, low-volume occasions such as nights and weekends are more difficult to predict – although the absolute impact on travel times is likely to be small. In any case, volume information can be incorporated in travel time and speed predictions in particular for

near-saturated conditions – with speeds and travel times showing little variation compared to the volumes. In addition, one can quite accurately cover the density function of volumes over different timescales. Provided that the uncertainty in volumes is used to provide reliability estimates regarding delays and travel times, ATIS can anticipate future uncertainties and thereby communicate robust ETAs, and dynamically adapt estimates as well as advice based on newly-arriving measurements.

To allow proactive rather than reactive decision making, short-term predictions of the network-wide traffic state under all conditions are required. Such forecasts are inherently difficult - particularly in an urban setting - with the traffic system being complex having variations in supply and demand on different timescales. Predictions become even more difficult with limited information regarding traffic management services and incidents in place. In addition, projected measures of all involved parties need to be included, as well as the behavioral responses of travelers and operators to those measures (feedback effects). On a strategic level with limited uncertainty in the predictions, anticipatory traffic management accounting for user responses can improve overall network performance in terms of total travel time. For instance, in the case of incidents, authorities can particularly apply a social rerouting strategy to mitigate the impact on the network with respect to the total travel time, but at the same time limit the a posteriori detour of individual drivers.

6.3 Topics for further research

In this thesis, we mainly considered variations that occur in the volume measurements. For many of the actors in the urban traffic domain, variations in volumes are not of direct interest. In fact, they are mainly concerned with changes that occur in, e.g., travel speeds, queues, and travel times, as a result of the interaction between demand and supply. Data regarding volumes support such estimates, particularly under lower-to-medium saturation levels, but further research is required to accurately capture the interrelations. Moreover, we mainly focused on local variations rather than network-wide dynamics. For freeways, it is well-known that adjacent locations in space show clear (cor)relations, and measurements from neighboring locations can be used to refine predictions. In a complex urban network, these relations are less trivial since flow does not need to be conserved, e.g., due to on-street parking. In any case, data from different sources – all with their own uncertainty – can be used to get a grip on the spatio-temporal variations in a network-wide context. Such variations are particularly interesting to examine after incidents and accidents, but further research is required to capture or model volume fluctuations resulting from non-recurrent events and the accompanying feedback effects.

We have shown that LSPs and ATIS can potentially improve the reliability of their communicated ETAs by incorporating volume data and accounting for local supply-based variations particularly at signalized intersections. Often, private parties do not have access to volume data as measured by induction loop detectors or other roadside measurement devices as both the data and equipment are owned by the local road authority. Although recently significant efforts have been made in practice to change this, in our context, the continuous exchange of data and information between public and private parties happens infrequently. Collaboration between parties means that data and information are shared - so that a full picture arises of the current traffic conditions, making it possible to improve individual services. However, far-reaching collaboration of different parties with typically conflicting

objectives is difficult if there is no benefit for some of the parties involved (Cruijssen et al., 2007; Hrelja et al., 2018; Van Heeswijk, 2017). Initiation of cooperation in practice therefore requires an identification of different collaboration models and the possible benefits for the individual actors.

In this thesis, we focused on identifying and quantifying sources of uncertainty in urban traffic conditions in the context of decision-making processes using information systems. In addition, we considered an anticipatory traffic management service in which drivers are rerouted onto paths for the system's benefit. Here, we formulated an optimization problem and made simplifying assumptions regarding the range of feedback effects. It is particularly interesting to incorporate the identified uncertainties in decision-making mechanisms and the associated optimization problems. These problems are often stochastic in nature and aim to find robust solutions that perform well under a range of scenarios, e.g., by incorporating an objective function aiming to reduce risk (Rockafellar & Uryasev, 2000) or by anticipating stochastic feedback effects (Rockafellar & Wets, 2017).

Bibliography

Adler, J. L., & Blue, V. J. (2002). A cooperative multi-agent transportation management and route guidance system. *Transportation Research Part C: Emerging Technologies*, *10*(5-6), 433–454.

Agatz, N., Campbell, A. M., Fleischmann, M., & Savels, M. (2008). Challenges and opportunities in attended home delivery. In B. Golden, S. Raghavan, & E. Wasil (Eds.), *The vehicle routing problem: Latest advances and new challenges* (pp. 379–396). Springer US.

Ahmed, M. S., & Cook, A. R. (1979). Analysis of freeway traffic time-series data by using box-jenkins techniques. *Transportation Research Record*, (722), 1–9.

Akcelik, R. (1980). *Time-Dependent Expressions for Delay, Stop Rate And Queue Length at Traffic Signals* (tech. rep.).

ALICE. (2014). *Urban Freight Research & Innovation Roadmap* (tech. rep.).

Angelelli, E., Arsik, I., Morandi, V., Savelsbergh, M., & Speranza, M. G. (2016). Proactive route guidance to avoid congestion. *Transportation Research Part B: Methodological*, *94*, 1–21.

Angelelli, E., Morandi, V., Savelsbergh, M., & Speranza, M. G. (2021). System optimal routing of traffic flows with user constraints using linear programming. *European Journal of Operational Research*, *293*(3), 863–879.

Angelelli, E., Morandi, V., & Speranza, M. G. (2018). Congestion avoiding heuristic path generation for the proactive route guidance. *Computers and Operations Research*, *99*, 234–248.

Angelelli, E., Morandi, V., & Speranza, M. G. (2020). Minimizing the total travel time with limited unfairness in traffic networks. *Computers and Operations Research*, *123*, 105016.

Aravkin, A., Burke, J. V., Ljung, L., Lozano, A., & Pillonetto, G. (2017). Generalized Kalman smoothing: Modeling and algorithms. *Automatica*, *86*, 63–86.

Aravkin, A., & Van Leeuwen, T. (2012). Estimating nuisance parameters in inverse problems. *Inverse Problems*, *28*(11).

Babu, G. J., & Feigelson, E. D. (2006). Astrostatistics: Goodness-of-fit and all that! *Astronomical Data Analysis Software and Systems XV*, *351*, 127.

Bagloee, S. A., Sarvi, M., Patriksson, M., & Rajabifard, A. (2017). A Mixed User-Equilibrium and System-Optimal Traffic Flow for Connected Vehicles Stated as a Complementarity Problem. *Computer-Aided Civil and Infrastructure Engineering*, *32*(7), 562–580.

Bank, B., Guddat, J., Klatte, D., Kummer, B., & Tammer, K. (1983). *Non-linear parametric optimization*. Springer.

Banks, J. H. (1999). Investigation of some characteristics of congested flow. *Transportation Research Record*, *1678*(1), 128–134.

Bard, J. F. (1998). *Practical bilevel optimization: algorithms and applications*. Springer Science & Business Media.

Bartlett, M. S. (1963). The spectral analysis of point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, *25*(2), 264–281.

Bascol, K., Emonet, R., Fromont, E., & Odobez, J. M. (2016). Unsupervised interpretable pattern discovery in time series using autoencoders. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10029 LNCS*, 427–438.

Bates, J., Polak, J., Jones, P., & Cook, A. (2001). The valuation of reliability for personal travel. *Transportation Research Part E: Logistics and Transportation Review*, *37*(2-3), 191–229.

Beckmann, M., McGuire, C. B., & Winsten, C. B. (1956). *Studies in the Economics of Transportation* (tech. rep.).

Bell, B. M., Burke, J. V., & Pillonetto, G. (2009). An inequality constrained nonlinear Kalman-Bucy smoother by interior point likelihood maximization. *Automatica*, *45*(1), 25–33.

Bell, B. M., Burke, J. V., & Schumitzky, A. (1996). A relative weighting method for estimating parameters and variances in multiple data sets. *Computational statistics & data analysis*, *22*(2), 119–135.

Ben-Akiva, M., De Palma, A., & Isam, K. (1991). Dynamic network models and driver information systems. *Transportation Research Part A: General*, *25*(5), 251–266.

Ben-Elia, E., Di Pace, R., Bifulco, G. N., & Shiftan, Y. (2013). The impact of travel information's accuracy on route-choice. *Transportation Research Part C: Emerging Technologies*, *26*, 146–159.

Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, *66*(3), 1025–1044.

Bezembinder, E. M. (2021). *Junction design rules: Improving junction design choices in urban traffic networks* (Doctoral dissertation). University of Twente.

Bijl, P. (2016). *En-route rescheduling of home deliveries: A case study on the home delivery operation of a large retail organization in the netherlands* (Master's thesis). University of Twente.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer New York, NY.

Boon, M. A., & Van Leeuwaarden, J. S. (2018). Networks of fixed-cycle intersections. *Transportation Research Part B: Methodological*, *117*, 254–271.

Branston, D. (1976). Models of single lane time headway distributions. *Transportation Science*, *10*(2), 125–148.

Breiman, L., Lawrence, R., Goodwin, D., & Bailey, B. (1977). The statistical properties of freeway traffic. *Transportation Research*, *11*(4), 221–228.

Breiman, L., & Lawrence, R. L. (1973). Time scales, fluctuations and constant flow periods in uni-directional traffic. *Transportation Research*, *7*(1), 77–105.

Briedis, P., & Samuels, S. (2010). The accuracy of inductive loop detectors. *ARRB Conference, 24th, 2010ARRB*.

Brillinger, D. R. (1975). The identification of point process systems. *The Annals of Probability*, 909–924.

Brillinger, D. R. (2008). Extending the volatility concept to point processes. *Journal of statistical planning and inference*, *138*(9), 2607–2614.

Buckley, D. J. (1967). Road traffic counting distributions. *Transportation Research*, *1*(2), 105–116.

Buckley, D. J. (1968). A Semi-Poisson Model of Traffic Flow. *Transportation Science*, *2*(2), 107–133.

Bureau of Public Roads. (1964). *Traffic Assignment Manual*.

Caceres, N., Romero, L. M., & Benitez, F. G. (2012). Estimating traffic flow profiles according to a relative attractiveness factor. *Procedia-Social and Behavioral Sciences*, *54*, 1115–1124.

Carrion, C., & Levinson, D. (2012). Value of Travel Time Reliability: A Review of Current Evidence. *SSRN Electronic Journal*.

Chatfield, C. (2001). Prediction intervals for time-series forecasting. In *Principles of forecasting* (pp. 475–494). Springer.

Chen, C., Wang, Y., Li, L., Hu, J., & Zhang, Z. (2012). The retrieval of intra-day trend and its influence on traffic prediction. *Transportation Research Part C: Emerging Technologies*, *22*, 103–118.

Chen, M., Yu, G., Chen, P., & Wang, Y. (2017). A copula-based approach for estimating the travel time reliability of urban arterial. *Transportation Research Part C: Emerging Technologies*, *82*, 1–23.

Chen, P., Zheng, F., Lu, G., & Wang, Y. (2016). Comparison of Variability of Individual Vehicle Delay and Average Control Delay at Signalized Intersections. *Transportation Research Record*, *2553*(1), 128–137.

Chen, Y. D., Li, L., Zhang, Y., Hu, J. M., & Jin, X. X. (2008). Fluctuations in urban traffic networks. *Modern Physics Letters B*, *22*(2), 101–115.

Cheng, C., Du, Y., Sun, L., & Ji, Y. (2016). Review on theoretical delay estimation model for signalized intersections. *Transport Reviews*, *36*(4), 479–499.

Chrobok, R., Kaumann, O., Wahle, J., & Schreckenberg, M. (2004). Different methods of traffic forecast based on real data. *European Journal of Operational Research*, *155*(3), 558–568.

Chung, B. D., Cho, H.-J., Friesz, T. L., Huang, H., & Yao, T. (2014). Sensitivity analysis of user equilibrium flows revisited. *Networks and Spatial Economics*, *14*(2), 183–207.

Chung, E. (2003). Classification of traffic pattern. *Proc. of the 11th World Congress on ITS*, 687–694.

Coogan, S., Flores, C., & Varaiya, P. (2017). Traffic predictive control from low-rank structure. *Transportation Research Part B: Methodological*, *97*, 1–22.

Cox, D. R., & Lewis, P. A. W. (1966). *The statistical analysis of series of events*. Springer.

Cox, D., & Miller, H. (1977). *The theory of stochastic processes* (Vol. 134). CRC Press.

Crawford, F., Watling, D. P., & Connors, R. D. (2017). A statistical method for estimating predictable differences between daily traffic flow profiles. *Transportation Research Part B: Methodological*, *95*, 196–213.

Crawford, F. (2017). *Methods for analysing emerging data sources to understand variability in traveller behaviour on the road network* (Doctoral dissertation). University of Leeds.

Crawford, F. (2020). Segmenting travellers based on day-to-day variability in work-related travel behaviour. *Journal of Transport Geography*, *86*, 102765.

Crovella, M. E., & Bestavros, A. (1997). Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on networking*, *5*(6), 835–846.

CROW. (2006). *Handboek verkeerslichtenregelingen*.

CROW. (2008). *Verkeersmaatregelen bij evenementen*.

Cruijssen, F., Dullaert, W., & Fleuren, H. (2007). Horizontal cooperation in transport and logistics: A literature review. *Transportation journal*, *46*(3), 22–39.

Daley, D. J., & Vere-Jones, D. (2003). *An introduction to the theory of point processes: Volume i: Elementary theory and methods*. Springer.

De Jong, G., Daly, A., Pieters, M., Miller, S., Plasmeijer, R., & Hofman, F. (2007). Uncertainty in traffic forecasts: Literature review and new results for The Netherlands. *Transportation*, *34*(4), 375–395.

De Jong, M. (2020). Enschede slibt dicht: 'Combinatie van werkzaamheden niet ideaal'. https://www.tubantia.nl/enschede/enschede-slibt-dicht-combinatie-van-werkzaamheden-niet-ideaal%7B~%7Dae42f3b6/

De Jong, M. (2021). Enschede slibt wéér dicht: 'Rondwegen kunnen grote stroom verkeer niet aan'. https://www.tubantia.nl/enschede-e-o/enschede-slibt-weer-dicht-rondwegen-kunnen-grote-stroom-verkeer-niet-aan%7B~%7Da66d5b23/

Dempe, S. (1989). On the directional derivative of the optimal solution mapping without linear independence constraint qualification. *Optimization*, *20*(4), 401–414.

Dempe, S., & Vogel, S. (2001). The generalized Jacobian of the optimal solution in parametric optimization. *Optimization*, *50*(5-6), 387–405.

Dempe, S. (1993). Directional differentiability of optimal solutions under Slater's condition. *Mathematical Programming*, *59*(1), 49–69.

Dempe, S. (2002). *Foundations of bilevel programming*. Springer Science & Business Media.

Dempe, S., & Zemkoho, A. B. (2012). Bilevel road pricing: theoretical analysis and optimality conditions. *Annals of Operations Research*, *196*(1), 223–240.

Di, X., Liu, H. X., Pang, J.-S., & Ban, X. J. (2013). Boundedly rational user equilibria (BRUE): Mathematical formulation and solution sets. *Transportation Research Part B: Methodological*, *57*(100), 300–313.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review*, *39*(4), 863–883.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*(1), 269–271.

Dion, F., Rakha, H., & Kang, Y.-S. (2004). Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. *Transportation Research Part B: Methodological*, *38*(2), 99–122.

Djavadian, S., Hoogendoorn, R. G., Van Arem, B., & Chow, J. Y. (2014). Empirical evaluation of drivers' route choice behavioral responses to social navigation. *Transportation Research Record*, *2423*, 52–60.

Dutreix, M., & Coogan, S. (2017). Quantile forecasts for traffic predictive control. *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 5666–5671.

Ehmke, J. F., Campbell, A. M., & Urban, T. L. (2015). Ensuring service levels in routing problems with time windows and stochastic travel times. *European Journal of Operational Research*, *240*(2), 539–550.

Eikenbroek, O. A. L., Still, G. J., Van Berkum, E. C., & Kern, W. (2018). The boundedly rational user equilibrium: a parametric analysis with application to the network design problem. *Transportation Research Part B: Methodological*, *107*, 1–17.

Ermagun, A., Chatterjee, S., & Levinson, D. (2017). Using temporal detrending to observe the spatial correlation of traffic. *PLoS ONE*, *12*(5), 1–21.

Ermagun, A., & Levinson, D. (2019). Spatiotemporal short-term traffic forecasting using the network weight matrix and systematic detrending. *Transportation Research Part C: Emerging Technologies*, *104*, 38–52.

Faigle, U., Kern, W., & Still, G. (2013). *Algorithmic principles of mathematical programming* (Vol. 24). Springer Science & Business Media.

Fendick, K. W., Saksena, V. R., & Whitt, W. (1989). Dependence in packet queues. *IEEE Transactions on Communications*, *37*(11), 1173–1183.

Ferrucci, F., & Bock, S. (2014). Real-time control of express pickup and delivery processes in a dynamic environment. *Transportation Research Part B: Methodological*, *63*, 1–14.

Ferrucci, F., & Bock, S. (2015). A general approach for controlling vehicle en-route diversions in dynamic vehicle routing problems. *Transportation Research Part B: Methodological*, *77*, 76–87.

Fiacco, A. V., & Ishizuka, Y. (1990a). Sensitivity and stability analysis for nonlinear programming. *Annals of Operations Research*, *27*(1), 215–235.

Fiacco, A. V., & Ishizuka, Y. (1990b). Suggested research topics in sensitivity and stability analysis for semi-infinite programming problems. *Annals of Operations Research*, *27*(1), 65–76.

Fleischmann, B., Gietz, M., & Gnutzmann, S. (2004). Time-varying travel times in vehicle routing. *Transportation Science*, *38*(2), 160–173.

Foi, A., Trimeche, M., Katkovnik, V., & Egiazarian, K. (2008). Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, *17*(10), 1737–1754.

Fu, L., & Hellinga, B. (2000). Delay variability at signalized intersections. *Transportation Research Record*, *1710*(1), 215–221.

Gendreau, M., Ghiani, G., & Guerriero, E. (2015). Time-dependent routing problems: A review. *Computers and Operations Research*, *64*, 189–197.

Gentile, G. (2014). Local user cost equilibrium: a bush-based algorithm for traffic assignment. *Transportmetrica A: Transport Science*, *10*(1), 15–54.

Gerlough, D. L., & Huber, M. J. (1976). *Traffic flow theory* (tech. rep.).

Ghosh, B., Basu, B., & O'Mahony, M. (2010). Random Process Model for Urban Traffic Flow Using a Wavelet-Bayesian Hierarchical Technique. *Computer-Aided Civil and Infrastructure Engineering*, *25*(8), 613–624.

Gmira, M., Gendreau, M., Lodi, A., & Potvin, J.-Y. (2021). Managing in real-time a vehicle routing plan with time-dependent travel times on a road network. *Transportation Research Part C: Emerging Technologies*, *132*, 103379.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 243–268.

Goh, K.-I., & Barabási, A.-L. (2008). Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, *81*(4), 48002.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goodwin, P., & Wright, G. (2010). The limits of forecasting methods in anticipating rare events. *Technological Forecasting and Social Change*, *77*(3), 355–368.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Guardiola, I. G., Leon, T., & Mallor, F. (2014). A functional approach to monitor and recognize patterns of daily traffic profiles. *Transportation Research Part B: Methodological*, *65*, 119–136.

Guo, J., Huang, W., & Williams, B. M. (2014). Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. *Transportation Research Part C: Emerging Technologies*, *43*, 50–64.

Guo, J., Huang, W., & Williams, B. M. (2015). Real time traffic flow outlier detection using short-term traffic conditional variance prediction. *Transportation Research Part C: Emerging Technologies*, *50*, 160–172.

Guo, J., & Williams, B. M. (2010). Real-time short-term traffic speed level forecasting and uncertainty quantification using layered Kalman filters. *Transportation Research Record*, *2175*(1), 28–37.

Guo, J., & Williams, B. M. (2012). Integrated Heteroscedasticity Test for Vehicular Traffic Condition Series. *Journal of Transportation Engineering*.

Guo, J., Williams, B. M., & Smith, B. L. (2007). Data collection time intervals for stochastic short-term traffic flow forecasting. *Transportation Research Record*, *2024*, 18–26.

Gusella, R. (1991). Characterizing the variability of arrival processes with indexes of dispersion. *IEEE Journal on Selected Areas in Communications*, *9*(2), 203–211.

Gwiggner, C., & Nagaoka, S. (2014). Data and queueing analysis of a Japanese air-traffic flow. *European Journal of Operational Research*, *235*(1), 265–275.

Ha, D.-H., Aron, M., & Cohen, S. (2012). Time headway variable and probabilistic modeling. *Transportation Research Part C: Emerging Technologies*, *25*, 181–201.

Habtemichael, F. G., & Cetin, M. (2016). Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transportation Research Part C: Emerging Technologies*, *66*, 61–78.

Hamilton, A., Waterson, B., Cherrett, T., Robinson, A., & Snell, I. (2013). The evolution of urban traffic control: Changing policy and technology. *Transportation planning and technology*, *36*(1), 24–43.

Hansen, P. C. (2010). *Discrete inverse problems: insight and algorithms*. SIAM.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *58*(1), 83–90.

Haykin, S. (2004). *Kalman filtering and neural networks* (Vol. 47). John Wiley & Sons.

Heim, G. R., & Sinha, K. K. (2001). Operational Drivers of Customer Loyalty in Electronic Retailing: An Empirical Analysis of Electronic Food Retailers. *Manufacturing and Service Operations Management*, *3*(3), 264–271.

Heinen, E., & Chatterjee, K. (2015). The same mode again? an exploration of mode choice variability in great britain using the national travel survey. *Transportation Research Part A: Policy and Practice*, *78*, 266–282.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, *32*(3), 896–913.

Hoogendoorn, S. P. (2005). Unified approach to estimating free speed distributions. *Transportation Research Part B: Methodological*, *39*(8), 709–727.

Hoogendoorn, S. P., & Botma, H. (1997). Modeling and estimation of headway distributions. *Transportation Research Record*, *1591*(1), 14–22.

Hrelja, R., Rye, T., & Mullen, C. (2018). Partnerships between operators and public transport authorities. working practices in relational contracting and collaborative partnerships. *Transportation Research Part A: Policy and Practice*, *116*, 327–338.

Hu, Y., & Hellendoorn, J. (2013). Uncertainty Modeling for Urban Traffic Model Predictive Control Based on Urban Patterns. *Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems (ITSC 2013)*, (ITSC), 845–850.

Huang, W., Jia, W., Guo, J., Williams, B. M., Shi, G., Wei, Y., & Cao, J. (2018). Real-time prediction of seasonal heteroscedasticity in vehicular traffic flow series. *IEEE Transactions on Intelligent Transportation Systems*, *19*(10), 3170–3180.

Hunt, B. R., Kostelich, E. J., & Szunyogh, I. (2007). Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, *230*(1-2), 112–126.

Hutchinson, T. P. (1972). Delay at a fixed time traffic signal—II: Numerical comparisons of some theoretical expressions. *Transportation Science*, *6*(3), 286–305.

Jahn, O., Möhring, R. H., Schulz, A. S., & Stier-Moses, N. E. (2005). System-optimal routing of traffic flows with user constraints in networks with congestion. *Operations Research*, *53*(4), 600–616.

Ji, Y., & Geroliminis, N. (2012). On the spatial partitioning of urban transportation networks. *Transportation Research Part B: Methodological*, *46*(10), 1639–1656.

Jiang, S., Wang, S., Li, Z., Guo, W., & Pei, X. (2015). Fluctuation similarity modeling for traffic flow time series: A clustering approach. *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, 848–853.

Jiang, X., & Adeli, H. (2004). Wavelet packet-autocorrelation function method for traffic flow pattern analysis. *Computer-Aided Civil and Infrastructure Engineering*, *19*(5), 324–337.

Jin, X., Zhang, Y., Wang, F., Li, L., Yao, D., Su, Y., & Wei, Z. (2009). Departure headways at signalized intersections: A log-normal distribution model approach. *Transportation Research Part C: Emerging Technologies*, *17*(3), 318–327.

Josefsson, M., & Patriksson, M. (2007). Sensitivity analysis of separable traffic equilibrium equilibria with application to bilevel optimization in network design. *Transportation Research Part B: Methodological*, *41*(1), 4–31.

Kaas, R., Goovaerts, M., Dhaene, J., & Denuit, M. (2008). *Modern Actuarial Risk Theory*. Springer Berlin, Heidelberg.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Kamarianakis, Y., Kanas, A., & Prastacos, P. (2005). Modeling traffic volatility dynamics in an urban network. *Transportation Research Record*, *1923*, 18–27.

Karlin, S., & Taylor, H. E. (2012). *A First Course in Stochastic Processes: Second Edition*.

Kay, S. M. (1993). *Fundamentals of statistical signal processing: Estimation theory*. Prentice-Hall, Inc.

Khan, M. A., Khalique, A., & Abouammoh, A. (1989). On estimating parameters in a discrete weibull distribution. *IEEE Transactions on Reliability*, *38*(3), 348–350.

Khosravi, A., Mazloumi, E., Nahavandi, S., Creighton, D., & van Lint, J. W. (2011). Prediction intervals to account for uncertainties in travel time prediction. *IEEE Transactions on Intelligent Transportation Systems*, *12*(2), 537–547.

KiM. (2020). *Kerncijfers Mobiliteit 2020*.

Kim, S.-H., & Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management*, *16*(3), 464–480.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.

Klatte, D., & Kummer, B. (2006). *Nonsmooth equations in optimization: regularity, calculus, methods and applications* (Vol. 60). Springer Science & Business Media.

Klein, I., Levy, N., & Ben-Elia, E. (2018). An agent-based model of the emergence of cooperation and a fair and stable system optimum using ATIS on a simple road network. *Transportation Research Part C: Emerging Technologies*, *86*, 183–201.

Koen, C. (2003). The analysis of indexed astronomical time-series - VIII. Cross-correlating noisy autoregressive series. *Monthly Notices of the Royal Astronomical Society*, *344*(3), 798–808.

Koen, C., & Lombard, F. (1993). The analysis of indexed astronomical time series - I. Basic methods. *Monthly Notices of the Royal Astronomical Society*, *263*(2), 287–308.

Kok, A. L., Hans, E. W., & Schutten, J. M. (2012). Vehicle routing under time-dependent travel times: The impact of congestion avoidance. *Computers & operations research*, *39*(5), 910–918.

Koot, M., Mes, M. R., & Iacob, M. E. (2021). A systematic literature review of supply chain decision making supported by the internet of things and big data analytics. *Computers & Industrial Engineering*, *154*, 107076.

Köster, F., Ulmer, M. W., Mattfeld, D. C., & Hasle, G. (2018). Anticipating emission-sensitive traffic management strategies for dynamic delivery routing. *Transportation Research Part D: Transport and Environment*, *62*, 345–361.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems*, *25*.

Kumar, S. V., & Vanajakshi, L. (2015). Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *European Transport Research Review*, *7*(3), 1–9.

Kwon, J., Coifman, B., & Bickel, P. (2000). Day-to-day travel-time trends and travel-time prediction from loop-detector data. *Transportation Research Record*, *1717*(1), 120–129.

Laña, I., Villar-Rodriguez, E., Etxegarai, U., Oregi, I., & Ser, J. D. (2019). A question of trust: Statistical characterization of long-term traffic estimations for their improved actionability. *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 1922–1928.

Lancia, C., & Lulli, G. (2020). Predictive modeling of inbound demand at major European airports with Poisson and Pre-Scheduled Random Arrivals. *European Journal of Operational Research*, *280*(1), 179–190.

Laporte, G. (1992). The vehicle routing problem: An overview of exact and approximate algorithms. *European journal of operational research*, *59*(3), 345–358.

Law, A. M., Kelton, W. D., & Kelton, W. D. (2007). *Simulation modeling and analysis* (Vol. 3). Mcgraw-hill New York.

Lecluyse, C., Van Woensel, T., & Peremans, H. (2009). Vehicle routing with stochastic time-dependent travel times. *4or*, *7*(4), 363–377.

Lee, G. M., Tam, N. N., & Yen, N. D. (2005). *Quadratic programming and affine variational inequalities: a qualitative study*. Springer.

Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, *50*, 57–70.

Levering, N., Boon, M., Mandjes, M., & Núñez-Queija, R. (2022). A framework for efficient dynamic routing under stochastically varying conditions. *Transportation Research Part B: Methodological*, *160*, 97–124.

Lewis, P. (1970). Remarks on the theory, computation and application of the spectral analysis of series of events. *Journal of Sound and Vibration*, *12*(3), 353–375.

Li, B. (2017). Stochastic modeling for vehicle platoons (ii): Statistical characteristics. *Transportation Research Part B: Methodological*, *95*, 378–393.

Li, L., & Chen, X. M. (2017). Vehicle headway modeling and its inferences in macroscopic/microscopic traffic flow theory: A survey. *Transportation Research Part C: Emerging Technologies*, *76*, 170–188.

Li, L., Su, X., Zhang, Y., Lin, Y., & Li, Z. (2015). Trend modeling for traffic time series analysis: An integrated study. *IEEE Transactions on Intelligent Transportation Systems*, *16*(6), 3430–3439.

Li, R., Liu, X., & Nie, Y. (2018). Managing partially automated network traffic flow: Efficiency vs. stability. *Transportation Research Part B: Methodological*, *114*(2018), 300–324.

Li, R., & Rose, G. (2011). Incorporating uncertainty into short-term travel time predictions. *Transportation Research Part C: Emerging Technologies*, *19*(6), 1006–1018.

Li, W., Yang, C., & Jabari, S. E. (2022). Nonlinear traffic prediction as a matrix completion problem with ensemble learning. *Transportation science*, *56*(1), 52–78.

Li, Y., Chai, S., Wang, G., Zhang, X., & Qiu, J. (2022). Quantifying the uncertainty in long-term traffic prediction based on pi-convlstm network. *IEEE Transactions on Intelligent Transportation Systems*, *23*(11), 20429–20441.

Lighthill, M. J., & Whitham, G. B. (1955). On kinematic waves II. A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *229*(1178), 317–345.

Lin, L., Handley, J. C., Gu, Y., Zhu, L., Wen, X., & Sadek, A. W. (2018). Quantifying uncertainty in short-term traffic prediction and its application to optimal staffing plan development. *Transportation Research Part C: Emerging Technologies*, *92*, 323–348.

Lippi, M., Bertini, M., & Frasconi, P. (2013). Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems*, *14*(2), 871–882.

Liu, X., Tanaka, M., & Okutomi, M. (2014). Practical Signal-Dependent Noise Parameter Estimation from a Single Noisy Image. *IEEE Transactions on Image Processing*, *23*(10), 4361–4371.

Liu, Y., & Gupta, H. V. (2007). Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, *43*(7), 1–18.

Lopez, C., Leclercq, L., Krishnakumari, P., Chiabaut, N., & Van Lint, H. (2017). Revealing the day-to-day regularity of urban congestion patterns with 3D speed maps. *Scientific Reports*, *7*(1), 1–11.

Lou, Y., Yin, Y., & Lawphongpanich, S. (2010). Robust congestion pricing under boundedly rational user equilibrium. *Transportation Research Part B: Methodological*, *44*(1), 15–28.

Lu, S. (2008). Sensitivity of static traffic user equilibria with perturbations in arc cost function and travel demand. *Transportation science*, *42*(1), 105–123.

Lu, S., & Nie, Y. (2010). Stability of user-equilibrium route flow solutions for the traffic assignment problem. *Transportation Research Part B: Methodological*, *44*(4), 609–617.

Luo, X., Wang, D., Ma, D., & Jin, S. (2019). Grouped travel time estimation in signalized arterials using point-to-point detectors. *Transportation Research Part B: Methodological*, *130*, 130–151.

Luo, Z.-Q., Pang, J.-S., & Ralph, D. (1996). *Mathematical programs with equilibrium constraints*. Cambridge University Press.

Luttinen, R. (1996). *Statistical Properties of Vehicle Time Headways* (Doctoral dissertation).

Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F. Y. (2015). Traffic Flow Prediction with Big Data: A Deep Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, *16*(2), 865–873.

Ma, D., Song, X., & Li, P. (2021). Daily Traffic Flow Forecasting through a Contextual Convolutional Recurrent Neural Network Modeling Inter- And Intra-Day Traffic Patterns. *IEEE Transactions on Intelligent Transportation Systems*, *22*(5), 2627–2636.

Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., & Wang, Y. (2017). Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. *Sensors (Switzerland)*, *17*(4).

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, *28*.

Mahalel, D., & Hakkert, A. (1983). Traffic arrival patterns on a cross section of a multilane highway. *Transportation Research Part A: General*, *17*(4), 263–270.

Mahmassani, H. S. (2001). Dynamic Network Traffic Assignment and Simulation Methodology for Advanced System Management Applications. *Networks and Spatial Economics*, *1*(3/4), 267–292.

Makridakis, S., Hogarth, R. M., & Gaba, A. (2009). Forecasting and uncertainty in the economic and business world. *International Journal of Forecasting*, *25*(4), 794–812.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE*, *13*(3), e0194889.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74.

Mangasarian, O. L., & Shiau, T.-H. (1987). Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems. *SIAM Journal on Control and Optimization*, *25*(3), 583–595.

Manolakis, D., & Bosowski, N. (2019). Count Time-Series Analysis. *IEEE Signal Processing Magazine*, *36*(3), 64–81.

Markovich, N. M., & Krieger, U. R. (2010). Statistical analysis and modeling of skype voip flows [Special Issue: Heterogeneous Networks: Traffic Engineering and Performance Evaluation]. *Computer Communications*, *33*, S11–S21.

Maze, T. H., Agarwal, M., & Burchett, G. (2006). Whether weather matters to traffic demand, traffic safety, and traffic operations and flow. *Transportation Research Record*, *1948*, 170–176.

McNeil, D. R. (1968). A solution to the fixed-cycle traffic light problem for compound Poisson arrivals. *Journal of Applied Probability*, *5*(3), 624–635.

Miller, A. J. (1963). Settings for fixed-cycle traffic signals. *Journal of the Operational Research Society*, *14*(4), 373–386.

Miller, A. J. (1970). An Empirical Model for Multilane Road Traffic. *Transportation Science*, *4*(2), 164–186.

Miller-Hooks, E. D., & Mahmassani, H. S. (2000). Least expected time paths in stochastic, time-varying transportation networks. *Transportation Science*, *34*(2), 198–215.

Mirchandani, P., & Head, L. (2001). A real-time traffic signal control system: Architecture, algorithms, and analysis. *Transportation Research Part C: Emerging Technologies*, *9*(6), 415–432.

Mohajerpoor, R., Saberi, M., & Ramezani, M. (2019). Analytical derivation of the optimal traffic signal timing: Minimizing delay variability and spillback probability for undersaturated intersections. *Transportation Research Part B: Methodological*, *119*, 45–68.

Mordukhovich, B. S. (2018). *Variational analysis and applications* (Vol. 30). Springer.

Muralidharan, A., Coogan, S., Flores, C., & Varaiya, P. (2016). Management of intersections with multi-modal high-resolution data. *Transportation Research Part C: Emerging Technologies*, *68*, 101–112.

Nakagawa, T., & Osaki, S. (1975). The discrete weibull distribution. *IEEE transactions on reliability*, *24*(5), 300–301.

Nantes, A., Ngoduy, D., Bhaskar, A., Miska, M., & Chung, E. (2016). Real-time traffic state estimation in urban corridors from heterogeneous data. *Transportation Research Part C: Emerging Technologies*, *66*, 99–118.

Nantes, A., Ngoduy, D., Miska, M., & Chung, E. (2015). Probabilistic travel time progression and its application to automatic vehicle identification data. *Transportation Research Part B: Methodological*, *81*(P1), 131–145.

Nellore, K., & Hancke, G. P. (2016). A survey on urban traffic management system using wireless sensor networks. *Sensors*, *16*(2), 157.

Nguyen, S., & Dupuis, C. (1984). An efficient method for computing traffic equilibria in networks with asymmetric transportation costs. *Transportation Science*, *18*(2), 185–202.

Oh, C., Ritchie, S. G., & Oh, J.-S. (2005). Exploring the relationship between data aggregation and predictability to provide better predictive traffic information. *Transportation Research Record*, *1935*(1), 28–36.

Ohazulike, A. E., Still, G., Kern, W., & Van Berkum, E. C. (2013). An origin–destination based road pricing model for static and multi-period traffic assignment problems. *Transportation Research Part E: Logistics and Transportation Review*, *58*, 1–27.

Olabarrieta, I. I., & Laña, I. (2020). Effect of Soccer Games on Traffic, Study Case: Madrid. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems, ITSC 2020*, 1–5.

Olszewski, P. S. (1990). Traffic signal delay model for nonuniform arrivals. *Transportation Research Record*, *1287*.

OpenStreetMap. (2019). *OpenStreetMap*. Retrieved July 14, 2019, from https://www.openstreetmap.org/

OpenStreetMap. (2021). *OpenStreetMap*. Retrieved November 6, 2021, from https://www.openstreetmap.org/

Outrata, J. V. (1997). On a special class of mathematical programs with equilibrium constraints. In *Recent advances in optimization* (pp. 246–260). Springer.

Outrata, J., Kocvara, M., & Zowe, J. (2013). *Nonsmooth approach to optimization problems with equilibrium constraints: theory, applications and numerical results* (Vol. 28). Springer Science & Business Media.

Pang, J.-S., & Ralph, D. (1996). Piecewise smoothness, local invertibility, and parametric analysis of normal maps. *Mathematics of operations research*, *21*(2), 401–426.

Partnership Talking Traffic. (2022). *Talking Traffic*. Retrieved July 1, 2022, from https://www.talking-traffic.com/en/

Patriksson, M. (2004). Sensitivity analysis of traffic equilibria. *Transportation Science*, *38*(3), 258–281.

Patriksson, M., & Rockafellar, R. T. (2002). A mathematical model and descent algorithm for bilevel traffic management. *Transportation Science*, *36*(3), 271–291.

Patriksson, M., & Rockafellar, R. T. (2003). Sensitivity analysis of aggregated variational inequality problems, with application to traffic equilibria. *Transportation Science*, *37*(1), 56–68.

Paxson, V., & Floyd, S. (1995). Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on networking*, *3*(3), 226–244.

Perederieieva, O., Ehrgott, M., Raith, A., & Wang, J. Y. T. (2015). A framework for and empirical study of algorithms for traffic assignment. *Computers & Operations Research*, *54*, 90–107.

Pérez Rivera, A. E., & Mes, M. R. (2017). Anticipatory freight selection in intermodal long-haul round-trips. *Transportation Research Part E: Logistics and Transportation Review*, *105*, 176–194.

Pillac, V., Gendreau, M., Guéret, C., & Medaglia, A. L. (2013). A review of dynamic vehicle routing problems. *European Journal of Operational Research*, *225*(1), 1–11.

Polson, N. G., & Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, *79*, 1–17.

Powell, W. B. (2014). Clearing the jungle of stochastic optimization. In *Bridging data and decisions* (pp. 109–137). Informs.

Powell, W. B., Jaillet, P., & Odoni, A. (1995). Stochastic and dynamic networks and routing. *Handbooks in Operations Research and Management Science*, *8*(100), 141–295.

Qiu, Y., & Magnanti, T. L. (1989). Sensitivity analysis for variational inequalities defined on polyhedral sets. *Mathematics of Operations Research*, *14*(3), 410–432.

Rakha, H., & Van Aerde, M. (1995). Statistical analysis of day-to-day variations in real-time traffic flow data. *Transportation Research Record*, 26–34.

Ralph, D., & Dempe, S. (1995). Directional derivatives of the solution of a parametric nonlinear program. *Mathematical programming*, *70*(1), 159–172.

Ramezani, M., & Geroliminis, N. (2012). On the estimation of arterial route travel time distribution with markov chains. *Transportation Research Part B: Methodological*, *46*(10), 1576–1590.

Rijkswaterstaat. (2021). *Rapportage Rijkswegennet 1e periode 2021* (tech. rep.).

Ritchie, S. G. (1983). Application of counting distribution for high-variance urban traffic counts. *Transportation research record*, *905*, 27–33.

Robertson, D. I., & Bretherton, R. D. (1991). Optimizing networks of traffic signals in real time-the scoot method. *IEEE Transactions on vehicular technology*, *40*(1), 11–15.

Robinson, S. M. (1982). Generalized equations and their solutions, part II: applications to nonlinear programming. In *Optimality and stability in mathematical programming* (pp. 200–221). Springer.

Robinson, S. M. (2006). Strong regularity and the sensitivity analysis of traffic equilibria: A comment. *Transportation Science*, *40*(4), 540–542.

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of risk*, *2*, 21–42.

Rockafellar, R. T., & Wets, R. J. B. (2017). Stochastic variational inequalities: single-stage to multistage. *Mathematical Programming*, *165*(1), 331–360.

Rockafellar, R. T., & Wets, R. J.-B. (2009). *Variational analysis* (Vol. 317). Springer Science & Business Media.

Roughgarden, T. (2005). *Selfish routing and the price of anarchy*. MIT press.

Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., & Lenzen, F. (2009). *Variational methods in imaging*. Springer.

Shi, G., Guo, J., Huang, W., & Williams, B. M. (2014). Modeling seasonal heteroscedasticity in vehicular traffic condition series using a seasonal adjustment approach. *Journal of Transportation Engineering*, *140*(5).

Shiftan, Y., Bekhor, S., & Albert, G. (2011). Route choice behaviour with pre-trip travel time information. *IET Intelligent Transport Systems*, *5*(3), 183–189.

Shone, R., Glazebrook, K., & Zografos, K. G. (2021). Applications of stochastic modeling in air traffic management: Methods, challenges and opportunities for solving air traffic problems under uncertainty. *European Journal of Operational Research*, *292*(1), 1–26.

Simon, H. A. (1997). *Models of bounded rationality: Empirically grounded economic reason* (Vol. 3). MIT press.

Simroth, A., & Zähle, H. (2010). Travel time prediction using floating car data applied to logistics planning. *IEEE Transactions on Intelligent Transportation Systems*, *12*(1), 243–253.

Smith, B. L., & Ulmer, J. M. (2003). Freeway traffic flow rate measurement: Investigation into impact of measurement time interval. *Journal of Transportation Engineering*, *129*(3), 223–229.

Soeffker, N., Ulmer, M. W., & Mattfeld, D. C. (2022). Stochastic dynamic vehicle routing in the light of prescriptive analytics: A review. *European Journal of Operational Research*, *298*(3), 801–820.

Son, S., Cetin, M., & Khattak, A. (2014). Exploring bias in traffic data aggregation resulting from transition of traffic states. *Transportation Research Record*, *2443*(1), 78–87.

Song, X., Li, W., Ma, D., Wang, D., Qu, L., & Wang, Y. (2018). A Match-Then-Predict Method for Daily Traffic Flow Forecasting Based on Group Method of Data Handling. *Computer-Aided Civil and Infrastructure Engineering*, *33*(11), 982–998.

Sparks, G. A. (1976). The unpublished schedule of urban peak period traffic. *Transportation Science*, *10*(3), 300–315.

Speranza, M. G. (2018). Trends in transportation and logistics. *European Journal of Operational Research*, *264*(3), 830–836.

Stathopoulos, A., & Karlaftis, M. G. (2001). Temporal and spatial variations of real-time traffic data in urban areas. *Transportation Research Record*, *1768*, 135–140.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, *69*(347), 730–737.

Still, G. (2018). *Lectures on parametric optimization: An introduction* (tech. rep.).

Sun, P., AlJeri, N., & Boukerche, A. (2018). A fast vehicular traffic flow prediction scheme based on fourier and wavelet analysis. *2018 IEEE Global Communications Conference (GLOBECOM)*, 1–6.

Taale, H., Birnie, J., & Bloemkolk, F. (2018). Stedelijk verkeersmanagement: Wat weten we van effecten? *Colloquium vervoersplanologisch speurwerk 2018*.

Tampère, C. M. J., & Immers, L. H. (2007). An extended Kalman filter application for traffic state estimation using CTM with implicit mode switching and dynamic parameters. *2007 IEEE Intelligent Transportation Systems Conference*, 209–216.

Tan, H., Wu, Y., Shen, B., Jin, P. J., & Ran, B. (2016). Short-term traffic prediction based on dynamic tensor completion. *IEEE Transactions on Intelligent Transportation Systems*, *17*(8), 2123–2133.

Tang, J., Wang, Y., Wang, H., Zhang, S., & Liu, F. (2014). Dynamic analysis of traffic time series at different temporal scales: A complex networks approach. *Physica A: Statistical Mechanics and its Applications*, *405*, 303–315.

Tang, K., Chen, S., Liu, Z., & Khattak, A. J. (2018). A tensor-based bayesian probabilistic model for citywide personalized travel time estimation. *Transportation Research Part C: Emerging Technologies*, *90*, 260–280.

Taş, D., Dellaert, N., Van Woensel, T., & De Kok, T. (2014). The time-dependent vehicle routing problem with soft time windows and stochastic travel times. *Transportation Research Part C: Emerging Technologies*, *48*, 66–83.

Tettamanti, I., Varga, I., Peni, T., Luspay, T., & Kulcsar, B. (2011). Uncertainty Modeling and Robust Control in Urban Traffoc. *Proceedings of the 18th World Congress The International Federation of Automatic Control*, 14910–14915.

Thomas, T., & Van Berkum, E. C. (2009). Detection of incidents and events in urban networks. *IET Intelligent Transport Systems*, *3*(2), 198–205.

Thomas, T., Weijermars, W., & Van Berkum, E. C. (2008). Variations in urban traffic volumes. *European Journal of Transport and Infrastructure Research*, *8*(3), 251–263.

Thomas, T., Weijermars, W., & Van Berkum, E. C. (2010). Predictions of Urban volumes in single time series. *IEEE Transactions on Intelligent Transportation Systems*, *11*(1), 71–80.

Tobin, R. L., & Friesz, T. L. (1988). Sensitivity analysis for equilibrium network flow. *Transportation Science*, *22*(4), 242–250.

Toth, P., & Vigo, D. (2002). *The vehicle routing problem*. SIAM.

Touhbi, S., Babram, M. A., Nguyen-Huu, T., Marilleau, N., Hbid, M. L., Cambier, C., & Stinckwich, S. (2018). Time headway analysis on urban roads of the city of marrakesh. *Procedia computer science*, *130*, 111–118.

Transportation Networks for Research Core Team. (2019). *Transportation Networks for Research*. Retrieved September 11, 2020, from https : / / github . com / bstabler / TransportationNetworks/

Transportation Research Board. (2000). *Highway capacity manual*.

Transportation Research Board. (2010). *Highway capacity manual*.

Tsekeris, T., & Stathopoulos, A. (2006). Real-time traffic volatility forecasting in urban arterial networks. *Transportation Research Record*, *1964*(1), 146–156.

Turochy, R. E., & Smith, B. L. (2002). Measuring variability in traffic conditions by using archived traffic data. *Transportation Research Record*, *1804*(1), 168–172.

Van As, S. C. (1991). Overflow delay in signalized networks. *Transportation Research Part A: General*, *25*(1), 1–7.

Van der Drift, S., Wismans, L., & Olde Kalter, M.-J. (2022). Changing mobility patterns in the netherlands during covid-19 outbreak. *Journal of location based services*, *16*(1), 1–24.

Van Essen, M., Eikenbroek, O. A. L., Thomas, T., & Van Berkum, E. C. (2020). Travelers' Compliance with Social Routing Advice: Impacts on Road Network Performance and Equity. *IEEE Transactions on Intelligent Transportation Systems*, *21*(3), 1180–1190.

Van Essen, M., Thomas, T., Van Berkum, E. C., & Chorus, C. (2016). From user equilibrium to system optimum: a literature review on the role of travel information,

bounded rationality and non-selfish behaviour at the network and individual levels. *Transport Reviews*, *36*(4), 527–548.

Van Heeswijk, W. J. A. (2017). *Consolidation and coordination in urban freight transport* (Doctoral dissertation). University of Twente.

Van Leeuwaarden, J. S. H. (2006). Delay analysis for the fixed-cycle traffic-light queue. *Transportation science*, *40*(2), 189–199.

Van Lint, J. W. C., & Van Hinsbergen, C. (2012). Short-term traffic and travel time prediction models. *Artificial Intelligence Applications to Critical Transportation Issues*, *22*(1), 22–41.

Van Lint, J. W., & Van Zuylen, H. J. (2005). Monitoring and predicting freeway travel time reliability: Using width and skew of day-to-day travel time distribution. *Transportation Research Record*, *1917*(1), 54–62.

Van Lint, J. W., Van Zuylen, H. J., & Tu, H. (2008). Travel time unreliability on freeways: Why measures based on variance tell only half the story. *Transportation Research Part A: Policy and Practice*, *42*(1), 258–277.

Vázquez, A., Oliveira, J. G., Dezsö, Z., Goh, K.-I., Kondor, I., & Barabási, A.-L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E*, *73*(3), 36127.

Viti, F. (2006). *The Dynamics and the Uncertainty of Delays at Signals* (Doctoral dissertation). Delft University of Technology. TRAIL Research School.

Viti, F., & Van Zuylen, H. J. (2010a). A probabilistic model for traffic at actuated control signals. *Transportation Research Part C: Emerging Technologies*, *18*(3), 299–310.

Viti, F., & Van Zuylen, H. J. (2010b). Probabilistic models for queues at fixed control signals. *Transportation Research Part B: Methodological*, *44*(1), 120–135.

Vlahogianni, E. I., Golias, J. C., & Karlaftis, M. G. (2004). Short-term traffic forecasting: Overview of objectives and methods. *Transport reviews*, *24*(5), 533–557.

Vlahogianni, E. I., & Karlaftis, M. G. (2011). Temporal aggregation in traffic data: Implications for statistical characteristics and model choice. *Transportation Letters*, *3*(1), 37–49.

Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, *43*, 3–19.

Vogel, K. (2002). What characterizes a "free vehicle" in an urban area? *Transportation Research Part F: traffic psychology and behaviour*, *5*(1), 15–29.

Vreeswijk, J. D., Landman, R. L., Van Berkum, E. C., Hegyi, A., Hoogendoorn, S. P., & Van Arem, B. (2015). Improving the road network performance with dynamic route guidance by considering the indifference band of road users. *IET intelligent transport systems*, *9*(10), 897–906.

Wagner-Muns, I. M., Guardiola, I. G., Samaranayke, V. A., & Kayani, W. I. (2018). A Functional Data Analysis Approach to Traffic Volume Forecasting. *IEEE Transactions on Intelligent Transportation Systems*, *19*(3), 878–888.

Walker, W., Harremoës, P., Rotmans, J., Sluijs, J., Van Der Asselt, M., Van Janssen, P., & M.P., K. (2003). Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment*, *4*(1), 5–17.

Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, *176*, 98–110.

Wang, J., Shang, P., Zhao, X., & Xia, J. (2013). Multiscale entropy analysis of traffic time series. *International Journal of Modern Physics C*, *24*(02), 1350006.

Wang, X. B., Yin, K., & Liu, H. X. (2018). Vehicle actuated signal performance under general traffic at an isolated intersection. *Transportation Research Part C: Emerging Technologies*, *95*, 582–598.

Wang, Y., & Papageorgiou, M. (2005). Real-time freeway traffic state estimation based on extended Kalman filter: A general approach. *Transportation Research Part B: Methodological*, *39*(2), 141–167.

Wardrop, J. G. (1952). Road paper. some theoretical aspects of road traffic research. *Proceedings of the institution of civil engineers*, *1*(3), 325–362.

Wasielewski, P. (1974). An integral equation for the semi-poisson headway distribution model. *Transportation Science*, *8*(3), 237–247.

Wasielewski, P. (1979). Car-Following Headways on Freeways Interpreted By the Semi-Poisson Headway Distribution Model. *Transportation Science*, *13*(1), 36–55.

Webster, F. V. (1958). *Traffic signal settings* (tech. rep.).

Weijermars, W. (2007). *Analysis of urban traffic patterns using clustering* (Doctoral dissertation). University of Twente.

Weijermars, W., & Van Berkum, E. C. (2005). Analyzing highway flow patterns using cluster analysis. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2005*, 308–313.

Wright, G., & Goodwin, P. (2009). Decision making and planning under low levels of predictability: Enhancing the scenario method. *International Journal of Forecasting*, *25*(4), 813–825.

Xing, X., Zhou, X., Hong, H., Huang, W., Bian, K., & Xie, K. (2015). Traffic Flow Decomposition and Prediction Based on Robust Principal Component Analysis. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2015-Octob*, 2219–2224.

Yang, H., Zhang, X., & Meng, Q. (2007). Stackelberg games and multiple equilibrium behaviors on networks. *Transportation Research Part B: Methodological*, *41*(8), 841–861.

Yang, M., Liu, Y., & You, Z. (2010). The reliability of travel time forecasting. *IEEE Transactions on Intelligent Transportation Systems*, *11*(1), 162–171.

Yang, Q., & Shi, Z. (2018). The evolution process of queues at signalized intersections under batch arrivals. *Physica A: Statistical Mechanics and its Applications*, *505*, 413–425.

Yang, S., Wu, J., Xu, Y., & Yang, T. (2019). Revealing heterogeneous spatiotemporal traffic flow patterns of urban road network via tensor decomposition-based clustering approach. *Physica A: Statistical Mechanics and its Applications*, *526*, 120688.

Yang, Y., Yang, H., & Fan, Y. (2019). Networked sensor data error estimation. *Transportation Research Part B: Methodological*, *122*, 20–39.

Yin, Y., Madanat, S. M., & Lu, X.-Y. (2009). Robust improvement schemes for road networks under demand uncertainty. *European Journal of Operational Research*, *198*(2), 470–479.

Yin, Y., & Shang, P. (2016). Multivariate multiscale sample entropy of traffic time series. *Nonlinear Dynamics*, *86*(1), 479–488.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European conference on computer vision*, 818–833.

Zhang, K., & Nie, Y. M. (2018). Mitigating the impact of selfish routing: An optimal-ratio control scheme (orcs) inspired by autonomous driving. *Transportation Research Part C: Emerging Technologies*, *87*, 75–90.

Zhang, P., Ma, W., & Qian, S. (2022). Cluster analysis of day-to-day traffic data in networks. *Transportation Research Part C: Emerging Technologies*, *144*, 103882.

Zhang, W., Medina, A., & Rakha, H. (2007). Statistical analysis of spatiotemporal link and path flow variability. *2007 IEEE Intelligent Transportation Systems Conference*, 1080–1085.

Zhang, Y., Zhang, Y., & Haghani, A. (2014). A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model. *Transportation Research Part C: Emerging Technologies*, *43*, 65–78.

Zheng, F., & Van Zuylen, H. J. (2010). Uncertainty and predictability of urban link travel time: Delay distribution–based analysis. *Transportation Research Record*, *2192*(1), 136–146.

Zheng, F., Van Zuylen, H. J., & Liu, X. (2017). A methodological framework of travel time distribution estimation for urban signalized arterial roads. *Transportation science*, *51*(3), 893–917.

Zhong, R. X., Xie, X. X., Luo, J. C., Pan, T. L., Lam, W. H., & Sumalee, A. (2020). Modeling double time-scale travel time processes with application to assessing the resilience of transportation systems. *Transportation Research Part B: Methodological*, *132*, 228–248.

Zhou, B., Xu, M., Meng, Q., & Huang, Z. (2017). A day-to-day route flow evolution process towards the mixed equilibria. *Transportation Research Part C: Emerging Technologies*, *82*, 210–228.

Zhu, S., & Levinson, D. (2015). Do people use the shortest path? an empirical test of wardrop's first principle. *PloS one*, *10*(8), e0134322.

Summary

Variations in urban traffic

In many urban areas, the traffic network is operating close to capacity. In such networks, unexpected and small fluctuations in traffic flow can result in a disruption in the level of service (LOS), e.g., travel speeds, delays and travel times. In fact, accumulated local and short-term fluctuations pose a serious risk to actors operating in the urban traffic domain who aim for decisions with stable performance under all conditions. Robust decisions anticipate the uncertainty in the sense that the potential effects of local, yet natural fluctuations are incorporated during the decision-making process. Albeit the increase in available traffic data sources, still very little is known about the dynamics and the uncertainty in urban traffic networks compared to freeways. In this thesis, we therefore investigate urban traffic variations on different scales and explore the potential of information regarding the variations on anticipatory decision making.

We distinguish three illustrative actors using urban traffic information during their decision making processes: logistics service providers (LSPs), urban traffic managers, and individual road users. LSPs concerned with home delivery use, e.g., travel time predictions with different time horizons to construct robust offline route plans that can be dynamically refined over time. Urban traffic managers mainly use typical volume patterns based on historical data for policy making and use near real-time data to trigger management scenarios. Individual road users employ advanced traveler information systems (ATIS), e.g., navigation devices, to support them in their travel decisions before departure and while being en route. These decision-making processes benefit from information regarding the development of the urban traffic conditions. Estimates about the accompanying dynamics in the uncertainty are often not considered but are also important for anticipatory decision making and the limitations thereof. Hence, the inter-relations between the systematic (predictable) variability in traffic and the uncertainty on various spatio-temporal scales should be understood and quantified.

In this thesis, we use historical data to get a grip on the systematic and random variations in urban traffic measurements. Since the conditions that occur in an urban network are for a major share determined by the dynamics near signalized intersections, we particularly focus on the variations there. Estimates regarding future travel times and delays are typically of interest for the actors under consideration and traffic volumes (or: flows, counts) throughout the network are an important source for explaining and predicting driving times and delays. Urban volume fluctuations are typically monitored and analyzed using measurements in the

order of minutes but express only a share of the actual variations. Hence, volume fluctuations need to be studied on various scales to not only account for the spatio-temporal variability in network usage, but also to incorporate the dynamics on a detailed level that introduce uncertainty on an aggregated scale. Therefore, our research objective is a follows: *Quantifying and understanding variations that occur in urban traffic volumes at different spatio-temporal levels*.

In Chapter 2, we examine urban traffic volume time series that explain a share of the dynamics occurring in an urban network. By eye, these time series show clear patterns, many of which are recurrent and can therefore in principle be predicted. Apart from systematic variations, a portion of the fluctuations in the measurements shows no pattern and should therefore not be predicted (noise). For monitoring purposes, it is important to separate the systematic from the random variability in the volumes to recognize changing conditions in a situation where high-frequency fluctuations occur in parallel.

24h traffic volume time series show systematic differences within a day and between days, and time of day and day of week are important predictors for network usage. We examine the changes in the 24h volume time series over the days, thereby considering the variability in volumes within the day but also the changing time-of-day volumes over the days. This simultaneous consideration supports one in revealing trends in the width and height of the peak and to accurately assess the impact of shorter-term systematic variations such as events and incidents. In particular the short-term deviations provide valuable information for management decisions but are more variable in their frequency of occurrence and the accompanying magnitude.

24h time series at a single point in the network basically consist of a combination of underlying recurrent temporal patterns or profiles. Distinct time series look different since they exist of latent profiles that are subject to small transformations changing over time. Extracting the underlying profiles is a challenging task since the profiles are not known in advance and the measurements are corrupted by noise. Moreover, what is considered systematic depends on a priori assumptions regarding the random variation and vice versa. In any case, many of the systematic variations are recurrent and, therefore, we develop a neural network architecture that infers long and short-term profiles together with a noise level estimate. Longer-term profiles express a volume shape occurring on a 24h scale, while short-term profiles represent the systematic differences compared to an underlying intra-day pattern. The random variation is captured using a so-called noise level function, expressing the probabilistic character of the fluctuations around a deterministic systematic pattern. The generic relation between the variance of the random variation and the underlying pattern allows for a full density characterization of the natural stochastic fluctuations.

Using two years of volume data collected throughout the Enschede traffic network, we show that only a few recurrent and physically-meaningful profiles are needed to express almost all systematic variations. Hence, 24h volume time series show a high degree of systematic variability - even in the case of events with variable starting times - only revealed when assessing variations over various timescales. It was estimated that the variance of the noise is linearly-dependent on the underlying systemic volume with slight overdispersion compared to Poisson noise. In fact, the noise distribution widens when volumes grow and decision making occurs in an increasingly uncertain environment when network usage increases.

In Chapter 3, we study urban arrival processes at signalized intersections. In fact, a large share of the fluctuations in the delays at signalized intersections can be traced back to the arrivals of vehicles at the approaches. Because of the importance of the dynamics in the delays for decision makers, there is a range of models and simulation methods that aim to capture the interactions at intersections. We use millions of recorded arrival events to statistically characterize arrival patterns and thereby assess the empirical consistency of the existing models.

Changes in the arrival patterns on a fixed location can be measured on different temporal scales. The underlying demand or arrival rate is assumed to be slowly varying and to change on timescales exceeding 5-10 minutes. Very short-term fluctuations, in the order of tenths of seconds, describe the stochastic (random) fluctuations in the actual arrival events. However, the two timescales are related and the point process describing the random occurrences of events over time is typically aggregated for monitoring and prediction purposes. Although on a 10min level arrival volumes show strong similarities with a Poisson or a renewal process, the latter processes fail to reflect the true structure in the arrivals. In fact, a stochastic arrival model in an urban setting should capture the non-stationarity in the demand over time and space, the marginal distribution of inter-arrival times accounting for both physical interactions as well as excess probabilities due to traffic signal control, and the periodicities in the arrival events because of upstream interruptions and platoon formation. In general, arrivals show bursts: periods with many arrivals alternate with periods in which no arrivals occur.

We develop a statistical framework to study arrivals as both a sequence of inter-arrival times as well as a counting process using a time-domain and a frequency-domain approach. When considering the distribution of the inter-arrival times, there is an excess probability of medium and high inter-arrival times, introduced by traffic lights upstream, statistically reflecting a combination of variable cycle times and the interaction with arrival events upstream. While consecutive inter-arrival times show only a weak serial correlation coefficient, this effect accumulates to a significant level when looking at a multitude of vehicles. The Bartlett power spectrum corresponding to the sequence of arrival events reveals dominant frequencies corresponding to the periodicities in the traffic signal cycles upstream. These dominant frequencies introduce dispersion in the counts using lower aggregation levels. Nonetheless, different arrival processes are indistinguishable when aggregation levels increase beyond 4-5min.

In a simulation setting, real-world mirroring arrival processes were shown to influence the variability in delays compared to the Poisson process. With vehicle-actuated traffic signal control, delay estimates obtained using a Poisson process overestimate both the mean as well as the variability in the delay particularly under lower volume occasions. The regularity in the real-world arrivals can be used to optimize vehicle-actuated signal control settings since arrivals contain predictive information about future events - in some cases even up to minutes in advance. In any case, not accounting for the interrupted characteristics of urban arrivals for the benefit of tractability overestimates the variations in delays while underestimating variations in volumes in the short-term, and thereby impacts decision-making processes.

Volume predictions support the decision-making processes of the considered actors. Many decisions of the actors operating in the traffic domain face decision problems characterized by uncertainty covering longer timescales, conflicting with the fact that most existing

prediction methods consider short-term point forecasts. Therefore we develop in Chapter 4 a volume forecasting methods that (i) offers reliable forecasts for different urban network conditions, (ii) provides predictions for both the long and short term and (iii) incorporates uncertainty in predictions in the form of probabilistic forecasts.

Traffic volume time series were shown to have a high degree of regularity, which can be well-expressed using latent profiles of different temporal scales. We use these flexible profiles for our prediction method since almost all systematic variability within a day and between days can be explained by using a few profiles. We constructed a prediction method to forecast systematic variations. The 24h forecast provides a prediction for a full day before the start of the day and the remaining-day prediction gives at any time of day a forecast for the volumes for the remainder of the day. Short-term predictions cover the next 15min to 1.5h. Since not all systematic variations can be expressed using basic exogenous variables only, we update initial forecasts based on the systematic differences in the residuals over various timescales. The inferred noise-level function is used to construct full density forecasts and to update the prediction based on the error of previous predictions relative to the inherent variability. Updating the prediction is rather difficult since noise makes it difficult to recognize changing conditions. Therefore, we apply smoothing by means of error aggregation and state-space filtering.

The quality of predictions is tested relative to the predictability of the system - and this difference is the true prediction error. The prediction error is expressed using a relative error based on the point prediction and using the coverage difference - reflecting the difference between the expected and true coverage of a density forecast. Considering 15min predictions, we found a point prediction error of 10- 15% suggesting that systematic variations for a major share can be predicted. A large share of these variations was actually possible to predict well in advance, at the beginning of the day when accounting for day-dependent characteristics. Although predictions are improved over the course of the day, many variations are systematic over timescales longer than hours. The density forecasts anticipating natural fluctuations are accurate and have an absolute coverage difference of 2-3%.

Traffic management measures such as rerouting under higher penetration levels suffer from feedback effects in the sense that current decisions influence future developments. Forecasts need to anticipate the emergent behavior of travelers so that intended outcomes are achieved. In Chapter 5, we investigate the potential and complexity of anticipatory traffic management by means of a social rerouting strategy.

Various traffic management measures have been proposed to reroute drivers towards socially desired paths. The main goal of these measures is to achieve the system optimum: the traffic state with minimum total travel time. The behavioral response to route advice needs to be anticipated since drivers are likely to ignore advice if the strategy reroutes them onto substantially longer paths for the system's benefit. In essence, any social routing strategy should anticipate user responses and persuade travelers to comply with socially oriented advice.

We propose a social routing strategy that explicitly anticipates the behavioral responses to a routing service so that an upper bound on the realized detour can be guaranteed. Compliance can be expected to be much higher when the advised route is only slightly longer than the shortest route. However, the realized travel time depends on the responses with respect to route choice that may occur from travelers that comply with the advice but also from those that do not comply but are now confronted with altered travel times on routes

because of behavioral changes by others.

The developed social routing strategy steers the traffic network towards an efficient but also fair, and therefore achievable and maintainable, traffic state. We show that the best possible paths with explicit a posteriori detour bounds to be proposed by a social routing service can be found by solving a bilevel optimization problem. A critical issue in solving the bilevel problem is that the lower-level optimal solution is not unique. We use techniques from parametric optimization to show that the directional derivative of the lower-level link flow nonetheless exists. This generalized derivative is used in a descent method and can be efficiently found as a solution of a quadratic optimization problem but requires a suitable route flow solution as parameter.

Numerical experiments show that a social routing system is a potential powerful measure to improve efficiency and preserve fairness at the same time. Even if only a small portion of travelers can be rerouted onto social routes, the resulting traffic state shows a major improvement in total travel time compared to the user equilibrium. In fact, only about 12% of the drivers need to take a small detour to obtain 2.4% of the maximum-possible 3.8% total travel time improvement.

Summarizing, in this thesis, we studied urban traffic volume variations on different scales. Although aggregated volume measurements show systematic variations over time and space, a substantial portion of the highly variable volume time series can be considered to occur by chance. Where a large share of the systematic variation is predictable, typically hours in advance, the noise induces an inherent and inevitable uncertainty to decision making. The random variation in aggregated volumes is highly related to fluctuations in the underlying departure and arrival events at signalized intersections, which show much more regularity than can be expected from the variability in the counts. Robust decisions can anticipate the variability in volumes using probabilistic forecasts and our findings substantiate the potential of anticipatory decision-making in the context of urban traffic management.

Samenvatting

Variaties in stedelijk verkeer

In veel stedelijke verkeersnetwerken is de capaciteit niet toereikend om aan de verkeersvraag te blijven voldoen. Een kleine verandering in de intensiteit kan dan een significante verstoring in *level of service* (LOS) veroorzaken, normaliter uitgedrukt in termen van snelheden, vertragingen of reistijden. Een opeenstapeling van opeenvolgende maar kleine en lokale fluctuaties in LOS vormt een risico voor de actoren die zich in het verkeersdomein begeven. Immers, zij verlangen naar een betrouwbare dienstverlening ongeacht de verkeerssituatie. Een robuust beslismechanisme anticipeert daarom op de inherente onzekerheid van het verkeerssysteem. Echter, in vergelijking met snelwegen en ondanks een toename van het aantal databronnen in dit domein, is er nog relatief weinig bekend over de dynamiek en de onzekerheid in het stedelijke verkeersnetwerk. Dit proefschrift bestudeert variaties in stedelijk verkeer en verkent de potentiële bijdrage van informatie over deze variaties aan het nemen van beslissingen die anticiperen op de dynamiek en onzekerheid van het stedelijk verkeer.

Dit werk identificeert drie illustratieve actoren die potentieel gebruik maken van stedelijke verkeersinformatie: logistieke dienstverleners, stedelijke verkeersmanagers en individuele weggebruikers. Logistieke dienstverleners die producten aan huis leveren gebruiken bijvoorbeeld reistijdvoorspellingen met verschillende tijdshorizonten voor hun offline transport plan die gedurende de uitvoering aangepast kan worden. Stedelijke verkeersmanagers maken veelal gebruik van intensiteitspatronen voor het ontwikkelen van beleid, en gebruiken real-time data voor het uitvoeren van vooraf gespecificeerde verkeersmanagementscenario's, bijvoorbeeld in het geval van incidenten. Veel individuele weggebruikers hebben tegenwoordig toegang tot *advanced traveler information systems* (ATIS), bijvoorbeeld navigatiesystemen, ter ondersteuning van hun routekeuze - zowel voor vertrek als onderweg. Alhoewel in al deze beslisprocessen informatie over de variaties in de verkeerssituatie wordt meegenomen, is er maar zelden aandacht voor de onzekerheid op het moment van beslissen. Echter, de mate van onzekerheid dient te worden meegenomen indien men robuuste beslissingen wil nemen. Dit werk richt zich daarom op het verklaren en kwantificeren van de systematiek in de variabiliteit van stedelijk verkeer en de dynamiek in de bijbehorende onzekerheid.

In dit proefschrift gebruiken we historische data om een beter beeld te krijgen van de systematische en willekeurige (*random*) variaties in stedelijk verkeer. Er is in het bijzonder aandacht voor de dynamiek nabij geregelde kruispunten aangezien de afwikkeling in het

stedelijke netwerk grootendeels door verkeersregelinstallaties (VRI's) wordt bepaald. Alhoewel de geïdentificeerde actoren met name zijn geïnteresseerd in voorspellingen betreffende reistijden en vertragingen, kunnen variabiliteit en onzekerheid hierin deels verklaard worden door de variatie in intensiteiten, en daarom bestuderen we voornamelijk intensiteitsgegevens. Intensiteiten worden vaak geaggregeerd tot een resolutie van enkele minuten waarmee een deel van de informatie potentieel verloren gaat. Waar een resolutie van enkele minuten voldoende is om de variatie in netwerkgebruik te beschrijven, is een veel gedetailleerder niveau benodigd om de dynamiek op lokaal niveau te begrijpen. Het onderzoeksdoel is daarom *op verschillende spatiële en temporele resoluties variaties in stedelijke verkeersintensiteiten te kwantificeren en te begrijpen*.

Hoofdstuk 2 beschouwt de variaties in tijdreeksen met verkeersintensiteiten. Deze tijdreeksen beschrijven een deel van de dynamiek in het stedelijke netwerk en hebben op het oog duidelijke patronen. Sterker, een groot deel van de patronen is terugkerend en kan daardoor worden opgenomen in een voorspelmechanisme. Naast de temporele systematiek in verkeersintensiteiten kan een substantieel deel van de variatie als ruis gekenmerkt worden, die niet te voorspellen is. Voor verkeersmonitoring is het belangrijk om systematische en random variaties te onderscheiden zodat veranderende situaties tijdig herkend kunnen worden.

24-uurs tijdreeksen bestaande uit een reeks van 15min intensiteitsmetingen laten een duidelijke systematiek zien, zowel binnen een dag als tussen de dagen. In beginsel zijn dag van de week en tijdstip belangrijke voorspellers voor de verkeersintensiteit, maar een groot deel van de systematische variabiliteit is geleidelijker van aard. Dit werk bestudeert daarom de variabiliteit in 24-uurs tijdreeksen over de dagen, daarmee rekening houdend met de veranderingen die optreden in zowel de vorm van de tijdreeks als in het totale volume. Door meerdere resoluties simultaan te beschouwen, zijn we in staat trends te herkennen, bijvoorbeeld in de piekintensiteit over de verschillende dagen. Daarnaast kan ook de invloed van evenementen en incidenten op de intensiteit ten opzichte van het 24-uurs patroon worden beschreven. Deze afwijkingen ten opzichte van het 24-uurs patroon zijn in het bijzonder waardevol voor verkeersmanagers, maar laten desalniettemin veel variabiliteit zien en zijn daardoor in het algemeen moeilijker te voorspellen.

De patronen in intensiteitsmetingen bestaan in wezen uit een combinatie van onderliggende (latente) maar terugkerende temporele patronen (profielen). 24-uurs tijdreeksen van verschillende dagen zien er anders uit omdat ze zijn opgebouwd uit een klein aantal profielen die onderhevig zijn aan kleine transformaties. Het extraheren van de onderliggende profielen is een uitdagende opgave, aangezien de profielen vooraf niet bekend zijn en de intensiteitsmetingen zijn aangetast door ruis. Bovendien, wat als systematisch wordt gekwantificeerd hangt af van a priori aannames met betrekking tot de ruis en vice versa. In dit proefschrift ontwikkelen we een neural network om korte- en lange-termijnprofielen te extraheren van de tijdreeksen en om de statistische eigenschappen van de ruis te schatten. Een lange-termijnprofiel beschrijft de vorm van een 24-uurs patroon, terwijl een kortetermijnprofiel de systematische verschillen ten opzichte van een onderliggend dagpatroon beschrijft. De random variatie is te karakteriseren middels een ruisniveau-functie, die de verdeling van ruis als een functie van een deterministisch en systematische intensiteit uitdrukt. De generieke relatie tussen de variantie in de ruis en het onderliggende patroon maakt een volledige dichtheidskarakterisering van de natuurlijke stochastische intensiteitsfluctuaties mogelijk.

Voor Hoofdstuk 2 is gebruik gemaakt van verkeersintensiteitsgegevens verzameld in Enschede. De resultaten na het toepassen van de ontwikkelde methode laten zien dat een paar terugkerende - en duidelijk interpreteerbare - profielen een groot deel van alle systematische variabiliteit in de geaggregeerde intensiteitsmetingen verklaart. Over het algemeen laten 24-uurs tijdreeksen veel systematiek zien, zelfs in het geval van evenementen met variërende aanvangstijden. De variantie in de ruis is lineair afhankelijk van de onderliggende intensiteit, met geringe overdispersie ten opzichte van de ruis resulteren van een Poisson proces. De verdeling van de ruis wordt breder naarmate de intensiteit toeneemt, een indicatie dat de onzekerheid in besluitvorming groeit naarmate het netwerkgebruik toeneemt.

In Hoofdstuk 3 bestuderen we aankomstprocessen van voertuigen bij geregelde kruispunten. Een groot deel van de fluctuaties in de vertragingen is immers terug te voeren op de aankomsten. Vanwege het belang van de variabiliteit in vertragingen voor de geïdentificeerde actoren is er een reeks aan (wachtrij)modellen beschikbaar die de interactie en de dynamiek rond kruispunten beschrijft. In dit proefschrift karakteriseren we de aankomsten geregistreerd door inductielussen in Enschede, om vervolgens de empirische consistentie van bestaande modellen te toetsen.

Fluctuaties in het aankomstproces op een vast punt in het netwerk kan men op verschillende resoluties bekijken. Normaalgesproken treden er relatief langzame veranderingen op in de vraag of in het verwachte aantal aankomsten per tijdseenheid (*aankomstintensiteit*), vaak bestudeerd op een resolutie in de orde van 5-10min. Zeer korte-termijnfluctuaties, met een resolutie in de orde van tienden van seconden, kunnen worden beschreven middels een stochastisch proces. De twee temporele resoluties zijn fundamenteel aan elkaar gerelateerd. Desalniettemin, waar op 10min niveau een aankomstproces sterke overeenkomsten vertoont met een Poisson proces, weerspiegelen vernieuwingsprocessen niet de ware structuur in de aankomsten. In beginsel dient een stochastisch aankomstmodel namelijk rekening te houden met de niet-stationaire vraag in ruimte en tijd, de verdeling van tussenaankomsttijden, en de periodiciteiten in de aankomsten als het gevolg van bovenstroomse verstoringen en pelotonvorming. Als gevolg vertonen aankomstprocessen duidelijke *bursts*: periodes met veel aankomsten worden afgewisseld met periodes waarin geen voertuigen aankomen.

We ontwikkelen een statistisch raamwerk om de structuur van de aankomstprocessen in het tijdsdomein en in het frequentiedomein te analyseren. Dit raamwerk wordt gebruikt om aankomstprocessen te beschrijven als een opeenvolging van tussenaankomsttijden en als een telproces. In vergelijking met een typische volgtijdverdeling voor snelwegen laat de verdeling van tussenaankomsttijden in een stedelijke context een grotere kans op gemiddelde en lange tussenaankomsttijden zien, voornamelijk geïntroduceerd door de VRI's stroomopwaarts. Terwijl opeenvolgende tussenaankomsttijden slechts een zwakke correlatiecoefficient hebben, accumuleert dit effect tot op een significant niveau als men een veelvoud aan tussenaankomsttijden bekijkt. Spectraalanalyse van het puntproces laat zien dat aankomstprocessen een duidelijke periodiciteit kunnen hebben die direct gerelateerd is aan de fasecycli van de VRI's bovenstrooms. Als gevolg hiervan laten intensiteitsmetingen op lagere aggregatieniveaus duidelijke overdispersie zien in vergelijking met ononderbroken aankomstprocessen. Echter, verschillende aankomstprocessen zijn nauwelijks meer te onderscheiden wanneer de temporele resolutie toeneemt tot meer dan 4-5 minuten.

In een simulatieomgeving wordt aangetoond dat vertragingsschattingen door wachtrijmodellen worden beïnvloed door de aannames met betrekking tot het aankomstproces. In een situatie met een voertuigafhankelijke verkeersregeling overschat een wachtrijmodel met

Poisson aankomsten zowel het gemiddelde als de variabiliteit in de vertragingen – in het bijzonder met lagere intensiteiten. In beginsel kan de periodiciteit in de aankomsten worden gebruikt om VRI's te optimaliseren aangezien aankomsten voorspellende informatie bevatten over toekomstige gebeurtenissen. In elk geval, door rekening te houden met de kenmerken van aankomstprocessen kan men de variabiliteit in intensiteitsmetingen en vertragingen nauwkeuriger inschatten.

Stedelijke intensiteitsvoorspellingen ondersteunen de besluitvormingsprocessen van de geïdentificeerde actoren. In tegenstelling tot wat veel van de beschikbare voorspelmechanismen aannemen, hebben de actoren te maken met door onzekerheid gekenmerkte beslissingsproblemen in de complexe context van het stedelijk verkeer die een langere tijdsperiode behelzen. In Hoofdstuk 4 ontwikkelen we daarom een intensiteit-voorspelmethode voor kansdichteidsvoorspellingen voor meerdere tijdshorizonten en verschillende stedelijke verkeerssituaties.

Tijdreeksen met intensiteitsmetingen hebben een hoge mate van regulariteit, en de systematiek in de 24-uurs reeksen kan goed worden uitgedrukt met behulp van latente profielen met verschillende tijdschalen. We gebruiken deze flexibele profielen voor het voorspelmechanisme aangezien vrijwel alle systematische variabiliteit op een dag en tussen de dagen hiermee kan worden verklaard. De voorspelmethode voorziet in een 24-uurs voorspelling aan het begin van de dag, een resterende-dag voorspelling gedurende de dag, en een kortetermijn voorspelling voor het komende anderhalf uur. Omdat niet alle systematische variabiliteit in de intensiteiten kan worden verklaard middels elementaire exogene variabelen, construeren we eerst een initiële *baseline* voorspelling die vervolgens verfijnd wordt op basis van de systematische fout in de eerdere voorspellingen. De invloed van de ruis op de voorspellingen wordt beperkt door smoothing en state-space filtering. Het ruisniveau gecombineerd met de systematische fout wordt gebruikt voor de dichtheidsvoorspellingen, daarmee rekening houdend met de inherente variabiliteit in intensiteitsmetingen.

De kwaliteit van de voorspellingen wordt uitgedrukt met inachtneming van de onvoorspelbaarheid van het verkeerssysteem. Het ruisniveau geeft een ondergrens voor de kwaliteit van een puntvoorspelling, en het verschil met deze grens is de systematische voorspelfout. Het verschil tussen de dichtheidsvoorspelling en de kansdichtheid van de ruis wordt daarnaast gebruikt als foutstatistiek. Het voorspelmechanisme heeft een punt-voorspelfout van 10-15% op 15min niveau, daarmee suggererend dat een substantieel deel van de systematische variaties voorspelbaar is. Het grootste deel van de systematische variabiliteit in de intensiteitsmetingen kan zefs aan het begin van de dag voorspeld worden. Hoewel de voorspellingen in de loop van de dag worden verbeterd, zijn veel variaties systematisch over grotere tijdschalen dan in de orde van enkele uren. De dichtheidsvoorspellingen anticiperend op de natuurlijke fluctuaties zijn nauwkeurig met een absoluut dekkingsverschil van 2 tot 3%.

Verkeersmanagement, bijvoorbeeld het her-routeren van verkeer, dient in geval van hoge penetratiegraden rekening te houden met terugkoppeleffecten aangezien een beslissing de toekomstige verkeerssituatie beïnvloedt. Voorspellingen moeten dan anticiperen op de reactie van de weggebruikers op de maatregel om ervoor te zorgen dat het beoogde effect behaald wordt. In Hoofdstuk 5 onderzoeken we de potentie en de complexiteit van anticiperend verkeersmanagement.

Er bestaan verschillende verkeersmanagementmaatregelen die het systeem optimum als

doel hebben: een efficiënte verdeling van verkeer over het netwerk met minimale totale reistijd. Om deze netwerkstaat te bereiken dienen sommige reizigers een omweg te nemen. Echter, dit route advies wordt waarschijnlijk genegeerd als het reistijdoffer te groot is. Sociaal routeadvies-systemen dienen in beginsel dus te anticiperen op het opvolggedrag van de reizigers.

In dit proefschrift ontwikkelen we een strategie voor sociaal routeadvies waarbij er expliciet rekening wordt gehouden met het opvolggedrag van reizigers. Naar verwachting zijn reizigers eerder geneigd om route advies op te volgen als de geadviseerde route slechts iets langer duurt dan de snelste route. De gerealiseerde reistijd is echter niet alleen afhankelijk van het opvolggedrag, maar ook van andere reizigers die geen advies krijgen maar hun route aanpassen in reactie op de verandering in reistijd door het opvolggedrag van anderen. De ontwikkelde strategie houdt rekening met deze terugkoppeleffecten en kan daarmee een maximaal reistijdoffer garanderen. Het optimale advies kan worden bepaald door een wiskundig *bi-level* optimalisatieprobleem op te lossen. Een complicerende factor bij het oplossen van dit probleem is dat de optimale oplossing van het *lower-level* probleem niet uniek is. Echter, de richtingsafgeleide van de lower-level *link flows* bestaat en kan worden gevonden als een oplossing van een kwadratisch optimalisatieprobleem. Deze richtingsafgeleide wordt vervolgens gebruikt in een oplossingsmethode die als doel heeft een lokaal optimum te vinden.

Numerieke experimenten laten zien dat sociaal routeadvies de potentie heeft de efficiëntie in het verkeersnetwerk te verbeteren. Zelfs als slechts een fractie van de reizigers wordt omgeleid naar sociale maar acceptabele routes, zijn de waargenomen netwerkeffecten significant in termen van totale reistijdverbetering ten opzichte van het gebruikersevenwicht. Om 2,4% van de maximaal haalbare 3,8% vermindering in totate reistijd te behalen moet ongeveer 12% van de reizigers een kleine omweg nemen.

Samenvattend, dit proefschrift bestudeert de variaties in stedelijke verkeer op meerdere temporele en spatiele resoluties. De variatie in intensiteitsmetingen kan worden geclassificeerd als systematisch of random. Waar een groot deel van de systematiek in de variabiliteit in de intensiteiten ver vooruit kan worden voorspeld, zorgt de ruis ervoor dat beslismethoden met betrekking tot stedelijk verkeer een inherente onzekerheid bevatten. Desalniettemin kan de random variatie in geaggregeerde metingen voor een groot deel worden verklaard door de aankomst- en vertrekprocessen bij geregelde kruispunten. Anticiperende beslismethoden houden expliciet rekening met deze stochastische fluctuaties bijvoorbeeld door gebruik te maken van probabilistische voorspellingen over verschillende tijdschalen. De resultaten van dit proefschrift laten zien dat anticiperende beslismethoden in het stedelijk verkeersdomein profijt hebben van voorspellingen van de variaties in the LOS van het verkeerssyteem inclusief de bijbehorende onzekerheid.

About the author

Oskar Eikenbroek was born in Emmen, the Netherlands, on December 4, 1991. In 2009, he moved to Enschede to study Civil Engineering at the University of Twente. After obtaining a BSc. degree in 2013, he continued at the same university to study both Civil Engineering and Management (specialization Transport Engineering and Management) and Applied Mathematics (specialization Operations Research). He obtained MSc. degrees for these studies in October 2016. After graduation, Oskar started a PhD research at the department of Transport Engineering and Management, University of Twente, in cooperation with the department of Industrial Engineering and Business Information Systems. His research was part of the ADAPTATION project, a collaboration between University of Twente, Erasmus University Rotterdam, DPD, Albert Heijn Online and Simacan. In September 2020, Oskar was a visiting researcher at the Institute for Transport Planning and Systems, ETH Zürich, Switzerland. Currently, he is a researcher at the department of Transport Engineering and Management, University of Twente.

Publications

Journal publications

Eikenbroek, O. A. L., Thomas, T., Mes, M.R.K. & Van Berkum, E. C. (2023). Pattern-based probabilistic prediction of urban traffic volumes. *Under review.*

Eikenbroek, O. A. L., Still, G. J., & Van Berkum, E. C. (2022). Improving the performance of a traffic system by fair rerouting of travelers. *European journal of operational research*, 299(1), 195-207.

Gkiotsalitis, K., Eikenbroek, O. A. L., & Cats, O. (2020). Robust network-wide bus scheduling with transfer synchronizations. *IEEE transactions on intelligent transportation systems*, 21(11), 4582-4592.

Schasfoort, B. B. W., Gkiotsalitis, K., Eikenbroek, O. A. L., & Van Berkum, E. C. (2020). A dynamic model for real-time track assignment at railway yards. *Journal of Rail Transport Planning & Management*, 14, 100198.

Van Essen, M., Eikenbroek, O. A. L., Thomas, T., & Van Berkum, E. C. (2019). Travelers' compliance with social routing advice: Impacts on road network performance and equity. *IEEE transactions on intelligent transportation systems*, 21(3), 1180-1190.

Eikenbroek, O. A. L., Still, G. J., van Berkum, E. C., & Kern, W. (2018). The boundedly rational user equilibrium: a parametric analysis with application to the network design problem. *Transportation Research Part B: Methodological*, 107, 1-17.

Conference proceedings and presentations

Eikenbroek O.A.L., Luan, X., Corman, F., & van Berkum, E.C. (2022). Social rerouting in public transport networks. Presented at EURO 2022, Espoo, Finland.

Trivella A,., Corman, F., & Eikenbroek, O.A.L. (2021). Decision making under uncertainty in public transport networks - The case of stochastic railway traffic control. Presented at International Conference on Optimization and Decision Sciences 2021, Rome, Italy.

Corman, F., Trivella, A., & Eikenbroek, O.A.L. (2021). Including Stochasticity in Railway Traffic Management Models. Informs Annual Meeting 2021, Anaheim, United States.

Eikenbroek, O. A. L., Still, G. J., & van Berkum, E. C. (2021). How Route Guidance with Small Detours for Only a Fraction of All Travelers Can Significantly Improve Network Performance. 8th International Symposium on Dynamic Traffic Assignment, DTA2021, Seattle, United States.

Eikenbroek, O. A. L., & Gkiotsalitis, K. (2020). Robust rescheduling and holding of autonomous buses intertwined with collector transit lines. In 2020 IEEE 23rd International Conference on Intelligent Transportation Systems, ITSC 2020 (pp. 1-7).

Schasfoort, B. B. W., Gkiotsalitis, K., Eikenbroek, O. A. L., & van Berkum, E. C. (2020). A dynamic model for real-time track assignment at railway yards. Paper presented at 99th Transportation Research Board (TRB) Annual Meeting 2020, Washington, United States.

Gkiotsalitis, K., Eikenbroek, O. A. L., & Cats, O. (2020). An exact method for real-time rescheduling after disturbances in metro lines. Paper presented at 99th Transportation Research Board (TRB) Annual Meeting 2020, Washington, United States.

Eikenbroek, O. A. L., Still, G.J., & van Berkum, E. (2020). Improving the Performance of a Traffic System by Fair Rerouting of Travelers. Presented at 11th International Conference on Computational Logistics, ICCL 2020, Enschede, The Netherlands.

Gkiotsalitis, K., Eikenbroek, O. A. L., & Cats, O. (2020). Robust bus scheduling considering transfer synchronizations. Paper presented at 99th Transportation Research Board (TRB) Annual Meeting 2020, Washington, United States.

Eikenbroek, O. A. L., Mes, M. R. K., & van Berkum, E. C. (2019). Online route planning in response to non-recurrent traffic disturbances. Presented at 30th European Conference on Operational Research, EURO 2019, Dublin, Ireland.

Eikenbroek, O. A. L., Mes, M. R. K., & van Berkum, E. C. (2019). Pattern Recognition in Urban Traffic Flows. Paper presented at 98th Transportation Research Board (TRB) Annual Meeting 2019, Washington, United States.

Eikenbroek, O. A. L., van Berkum, E. C., Still, G. J., & Kern, W. (2016). Computing Boundedly Rational User Equilibria and Implications for the Network Design Problem. Paper presented at TRISTAN IX 2016, Oranjestad, Aruba

TRAIL Thesis Series

The following list contains the most recent dissertations in the TRAIL Thesis Series. For a complete overview of more than 300 titles see the TRAIL website: www.rsTRAIL.nl.

The TRAIL Thesis Series is a series of the Netherlands TRAIL Research School on transport, infrastructure and logistics.

Eikenbroek, O., *Variations in Urban Traffic*, T2023/2, February 2023, Thesis Series, the Netherlands

Wang, S., *Modeling Urban Automated Mobility on-Demand Systems: an Agent-Based Approach*, T2023/1, January 2023, Thesis Series, the Netherlands

Szép, T., *Identifying Moral Antecedents of Decision-Making in Discrete Choice Models*, T2022/18, December 2022, Thesis Series, the Netherlands

Zhou, Y., *Ship Behavior in Ports and Waterways: An empirical perspective*, T2022/17, December 2022, Thesis Series, the Netherlands

Yan, Y., *Wear Behaviour of A Convex Pattern Surface for Bulk Handling Equipment*, T2022/16, December 2022, Thesis Series, the Netherlands

Giudici, A., *Cooperation, Reliability, and Matching in Inland Freight Transport*, T2022/15, December 2022, TRAIL Thesis Series, the Netherlands

Nadi Najafabadi, A., *Data-Driven Modelling of Routing and Scheduling in Freight Transport*, T2022/14, October 2022, TRAIL Thesis Series, the Netherlands

Heuvel, J. van den, *Mind Your Passenger! The passenger capacity of platforms at railway stations in the Netherlands*, T2022/13, October 2022, TRAIL Thesis Series, the Netherlands

Haas, M. de, *Longitudinal Studies in Travel Behaviour Research*, T2022/12, October 2022, TRAIL Thesis Series, the Netherlands

Dixit, M., *Transit Performance Assessment and Route Choice Modelling Using Smart Card Data*, T2022/11, October 2022, TRAIL Thesis Series, the Netherlands

Du, Z., *Cooperative Control of Autonomous Multi-Vessel Systems for Floating Object Manipulation*, T2022/10, September 2022, TRAIL Thesis Series, the Netherlands

Larsen, R.B., *Real-time Co-planning in Synchromodal Transport Networks using Model Predictive Control*, T2022/9, September 2022, TRAIL Thesis Series, the Netherlands

Zeinaly, Y., *Model-based Control of Large-scale Baggage Handling Systems: Leveraging the theory of linear positive systems for robust scalable control design*, T2022/8, June 2022, TRAIL Thesis Series, the Netherlands

Fahim, P.B.M., *The Future of Ports in the Physical Internet*, T2022/7, May 2022, TRAIL Thesis Series, the Netherlands

Huang, B., *Assessing Reference Dependence in Travel Choice Behaviour*, T2022/6, May 2022, TRAIL Thesis Series, the Netherlands

Reggiani, G., *A Multiscale View on Bikeability of Urban Networks*, T2022/5, May 2022, TRAIL Thesis Series, the Netherlands

Paul, J., *Online Grocery Operations in Omni-channel Retailing: opportunities and challenges*, T2022/4, March 2022, TRAIL Thesis Series, the Netherlands

Liu, M., *Cooperative Urban Driving Strategies at Signalized Intersections*, T2022/3, January 2022, TRAIL Thesis Series, the Netherlands

Feng, Y., *Pedestrian Wayfinding and Evacuation in Virtual Reality*, T2022/2, January 2022, TRAIL Thesis Series, the Netherlands

Scheepmaker, G.M., *Energy-efficient Train Timetabling*, T2022/1, January 2022, TRAIL Thesis Series, the Netherlands

Bhoopalam, A., *Truck Platooning: planning and behaviour*, T2021/32, December 2021, TRAIL Thesis Series, the Netherlands

Hartleb, J., *Public Transport and Passengers: optimization models that consider travel demand*, T2021/31, TRAIL Thesis Series, the Netherlands

Azadeh, K., *Robotized Warehouses: design and performance analysis*, T2021/30, TRAIL Thesis Series, the Netherlands

Chen, N., *Coordination Strategies of Connected and Automated Vehicles near On-ramp Bottlenecks on Motorways*, T2021/29, December 2021, TRAIL Thesis Series, the Netherlands

Onstein, A.T.C., *Factors influencing Physical Distribution Structure Design*, T2021/28, December 2021, TRAIL Thesis Series, the Netherlands

Olde Kalter, M.-J. T., *Dynamics in Mode Choice Behaviour*, T2021/27, November 2021, TRAIL Thesis Series, the Netherlands

Los, J., *Solving Large-Scale Dynamic Collaborative Vehicle Routing Problems: an Auction-Based Multi-Agent Approach*, T2021/26, November 2021, TRAIL Thesis Series, the Netherlands

Khakdaman, M., *On the Demand for Flexible and Responsive Freight Transportation Services*, T2021/25, September 2021, TRAIL Thesis Series, the Netherlands

Wierbos, M.J., *Macroscopic Characteristics of Bicycle Traffic Flow: a bird's-eye view of cycling*, T2021/24, September 2021, TRAIL Thesis Series, the Netherlands

TRAIL

Summary

Traffic volume time series show patterns on various scales. However, a substantial portion of the fluctuations in the urban traffic volumes is random and is related to the arrival processes at signalized intersections. This thesis investigates the variations in urban traffic and explores the potential of using forecasts to improve the decision-making processes of logistics service providers, traffic managers and individual car users.

About the Author

Oskar Eikenbroek holds degrees in Applied Mathematics and Civil Engineering and Management. His PhD research was carried out from 2017 until 2022 at the University of Twente. His research interests include multi-objective and multi-level optimization.

TRAIL Research School ISBN 978-90-5584-321-3

Erasmus University Rotterdam · Radboud University · rijksuniversiteit groningen · TU Delft Delft University of Technology · University of Twente. · TU/e Technische Universiteit Eindhoven University of Technology