Enhanced Mobility Awareness - A Data-Driven Approach to Analyze Traffic under Planned Special Event Scenarios

Von der Fakultät für Elektrotechnik und Informatik der Gottfried Wilhelm Leibniz Universität Hannover zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
(abgekürzt: Dr. rer. nat.)
genehmigte Dissertation

von Dipl.-Inform. Simon Kwoczek
geboren am 13.01.1984
in Berlin, Deutschland

Juni 2018

Leibniz
Universität
Hannover

Tag der Promotion: 08. Juni 2018
Referent: Prof. Dr. Wolfgang Nejdl
Korreferent: Prof. Dr. Sergio Di Martino

Disclaimer

The results, opinions and conclusions expressed in this thesis are not necessarily those of Volkswagen AG.

"Essentially, all models are wrong, but some are useful."

George E. P. Box, 1979

Abstract

Traffic disruptions impose societal costs of billions of dollars every year. A constant increase in mobility demand, combined with ongoing urbanization, exacerbates the problem. Since extensions of the infrastructure are for the most part no longer feasible, researchers are trying to find solutions to increase the efficiency of the road network usage. One key element to meeting that goal is to use smart prediction techniques on as many traffic-influencing factors as possible. With the availability of traffic datasets with high spatial and temporal resolutions, more and more data-driven solutions to predict the impact of these factors have been presented by the community. However, while the impacts of hazards, road accidents, and daily rush hour have been the subjects of intense study and analysis the specific impact of so-called planned special events on traffic remains mostly unexplored. Are the effects of upcoming concerts, sporting events, etc. predictable at all? This is the main question that we address in this thesis. We focus our analysis on three different aspects. First, we analyze the general characteristics of event-caused traffic disruptions around different venues in Germany. The results show, that the impact of events varies strongly, being highly affected by its venue *location*, the *time of day*, and the event *category*. In the second step, we analyze the *spatial impact* of events around different venues. This *spatial impact* describes a set of road segments, that people tend to use to get to and from the venue. To identify those preferred routes, we propose a classification-based technique that measures event influence for each road segment separately. The approach is based on a large scale analysis across many different venues in Germany. Results show impact zones around several soccer venues in Germany that we discuss in detail. In the third part of this thesis we analyze features from online sources (Twitter, Facebook, etc.) in terms of their *explanatory power* towards the expected event impact. We collect a large list of different information sources for major events in different venues. Based on that collection, we present prediction models for various measures of event impact.

Our results show, that these approaches are capable to predict the severity of event impact under certain conditions, which allows decision makers to create traffic management strategies tailored to event caused traffic disruptions.

Keywords: Traffic Prediction, Planned Special Event, Social Media

Zusammenfassung

Verkehrsstaus kosten unsere Gesellschaft jährlich Milliardenbeträge. Ein konstant zunehmendes Mobilitätsbedürfnis, in Kombination mit einer anhaltenden Urbanisierung, verschärfen das Problem zunehmend. Eine Lösung wäre ein weiterer Ausbau der bestehenden Infrastruktur. Vielerorts ist das jedoch aus gesellschaftlichen und ökologischen Gründen nicht mehr möglich. Daher versuchen verschiedene Forschungsgemeinschaften neue Lösungsansätze zu finden, welche die bestehende Infrastruktur effizienter auslasten. Eine Schlüsseltechnologie dafür sind Verkehrsprognosen, die möglichst viele verschiedene Faktoren, welche den täglichen Verkehr beeinflussen, berücksichtigen. Mit steigender Verfügbarkeit von Datensätzen, die ein genaues Abbild der räumlichen und zeitlichen Verkehrssituation liefern, werden immer mehr daten-getriebene Ansätze für derartige Prognosen vorgestellt. Der Forschungsfokus dabei liegt derzeit auf den Einflüssen von Gefahrenstellen, Unfällen und des täglichen Berufsverkehrs. Ein bedeutender Einflussfaktor bleibt jedoch fast unbeachtet: Die Beeinflussung des Verkehrs durch Events wie Konzerte oder Sportveranstaltungen. Dieser wird in der vorliegenden Promotionsarbeit detailliert betrachtet. Die Analyse unterteilt sich dabei in drei verschiedene Aspekte. Im ersten Teil analysieren wir die grundsätzlichen Eigenschaften von veranstaltungsbedingten Verkehrsstaus anhand von verschiedenen ausgewählten Veranstaltungsstätten in Deutschland. Die Ergebnisse zeigen ein starkes Schwanken der beobachteten Einflüsse auf die Verkehrslage. Diese Schwankungen sind vor allem abhängig von dem Ort der Veranstaltungsstätte, dem Zeitpunkt der Veranstaltung sowie der Veranstaltungskategorie. Im zweiten Teil analysieren wir den räumlichen Einfluss von Veranstaltungen. Dieser beschreibt eine Menge von Straßensegmenten, welche von Besuchern vornehmlich zur An- und Abreise benutzt werden. Um diese Routen zu identifizieren, entwickeln wir klassifikationsbasierte Verfahren, welche die Stärke des Einflusses von Events pro Straßensegment separat berechnen. Diese basieren auf großangelegten Studien über verschiedene Veranstaltungsstätten in ganz Deutschland. Wir zeigen die Ergebnisse für ausgewählte Fußballstadien und diskutieren diese detailliert. Im dritten Teil dieser Arbeit beschäftigen wir uns mit zusätzlichen Informationsquellen zu den Veranstaltungen, um den zu erwartenden Einfluss genauer vorhersagen zu können. Dazu werden verschiedene Daten aus Online-Quellen (z.B., Twitter, Facebook) herangezogen. Auf Basis dieser Daten entwickeln wir Prädiktionsmodelle und bewerten und diskutieren die Aussagekraft der unterschiedlichen Quellen. Unserer Ergebnisse zeigen, dass die vorgestellten Ansätze in der Lage sind, in gewissen Situationen die durch Events verursachten Verkehrsbelastungen vorhersagen zu können. Dies ermöglicht Experten zukünftig veranstaltungsbezogene Verkehrsmanagementstrategien zu entwickeln.

Schlagworte: Verkehrsprognosen, Veranstaltungen, Soziale Medien

Acknowledgements

Completing my PhD thesis after all these years of work is a fundamental milestone in my life. Looking back now, I would like to express my deepest gratitude to a list of different people.

First, I would like to express my sincere appreciation to Prof. Dr. Wolfgang Nejdl for his valuable critiques, hours of discussions and patient guidance of my research work. When I met him in April 2013, he was willing to take on that project together with me and I am very thankful for that opportunity.

I got to know Prof. Dr. Sergio di Martino in February 2013 when I started at Volkswagen Group Research in Wolfsburg/Germany. He introduced me to scientific working and fundamentally shaped my attitude and perspective. We worked together on this topic and spent many hours discussing different aspects and possible approaches. He was a true mentor and he is still a good friend to me. For all this he has my deepest gratitude.

I would also like to extend my thanks to my friend Dr. Silviu Homoceanu. His critics, his ideas, his patience, and his constant support were of great help. Although he always has a lot on his plate, he took the time to discuss many parts of this thesis, for what I am very grateful.

Further, I would like to express my sincerest regards to my current and former colleagues at Volkswagen AG. I especially owe a lot to Tatiana Deriyenko, who spent hours on reading and commenting my document (even when being caught in the rain). Also, Felix Richter who was always in for a coffee and a talk about PhD candidate life, and many others that helped me over the years.

Finally, I would like to thank my family for their constant support and motivation. I especially owe my deepest gratitude to Luisa Jordan for her support and patience, even when this work required me to put in late night hours or work on the weekends.

Contents

Chapter 1

Introduction

1.1 Motivation

Billions of dollars are wasted owing to traffic congestion every year [1]. Apart from the economic aspects, these traffic delays are a daily burden to road users. Taking into account the ongoing urbanization and constant growth in mobility demand, smart solutions are needed. Toward those solutions, traffic prediction has become a key technology within the research community of intelligent transportation systems (ITS).

Traffic disruptions are caused by a wide variety of reasons. Daily rush hour, weather phenomena, incidents, construction sites, and many other factors act upon the traffic state. To alleviate the negative effects of congestion by smart prediction, it is necessary to consider the individual effects of as many influencing factors as possible. From the technical perspective, this requires methods and algorithms to learn the specific impact patterns based on real-world observations and create prediction models. Future systems need to incorporate individual influencing factor behavior patterns and provide coherent predictions based on the resulting models. This is visualized in figure 1.1.

One influencing factor that we may all experience in daily life is the impact of planned special events (PSEs) on traffic. Have you ever tried to reach the concert of your favorite band by car and found yourself caught with all the other fans in bumper-to-bumper traffic? According to a study from the U.S. Federal Highway Administration ([3]), there are approximately 24000 PSEs with more

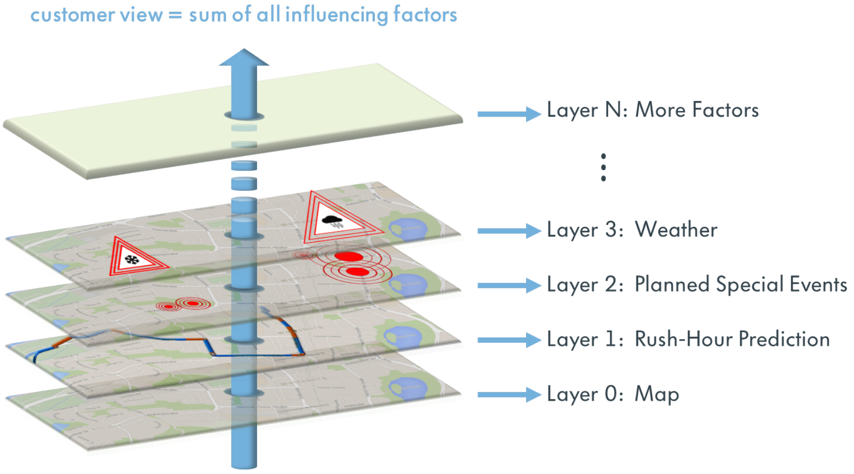

customer view = sum of all influencing factors
Layer N: More Factors
Layer 3: Weather
Layer 2: Planned Special Events
Layer 1: Rush-Hour Prediction
Layer 0: Map

FIGURE 1.1: Example of traffic-influencing factors. Future systems (e.g., navigation systems) need to incorporate as many information sources as possible. Layer 0: Navigation based on a digital map. Layer 1: Include rush-hour prediction. Layer 2: Include prediction of the effects of planned special events. Layer 3: Include prediction of the effects of weather phenomena. Layer N: Include additional influencing factors. Source: [2]

than 10000 attendees per year in the United States. This results in roughly 460 PSEs per week where large crowds try to reach a given location at the same time. According to the same report, the nationwide congestion cost in the United States is between 1.7 and 3.4 billion US dollars per year. These numbers confirm the need to further analyze PSE-specific influences and develop solutions that can be incorporated into future systems.

The term PSE is a very broad definition. PSEs range from very large events (e.g., a festival with 80000 attendees) to concerts of local bands in small corner pubs. Whereas traffic disruptions are a well known problem in transportation planning, only very large events receive special attention (e.g., soccer world championships or Olympic games). For such events, procedures exist whereby organizers, traffic operators, and local authorities work and plan together far ahead of the event [4]. For medium or small events, these procedures are usually not in place [5]. A major problem in handling these events is the poor availability of information. Typically, there is no systematic way to collect information about events in centralized manner. However, even if there is information about upcoming events, little can be done without knowing the traffic demand beforehand. In some of the larger cities, authorities address this problem by manually searching for information about events on local news sites or on the web. While this work is very labor-intensive, it still faces the problem that without knowing the expected impact, it is difficult to plan traffic strategies [4].

With this problem in mind, a review of the current scientific literature shows that there is not much work addressing this specific problem. A detailed overview is given in later chapters of this thesis. For now, we can summarize by stating that most of the related work focuses on demand modeling of public transportation networks (e.g., [4, 6]), and those works that deal with traffic data typically only tease the topic (e.g., [7, 8]). In this work, we focus on this precise phenomenon and analyze PSE-inflicted congestion around venues. The ultimate goal is to better understand PSE-caused traffic behavior and enable future systems to incorporate knowledge of their specific impact in advance.

But why is this difficult? If PSEs have an impact on traffic and if there is a pattern behind it, why do we struggle so much to solve this issue? One major challenge that we faced during our research is event traffic impact variability. Events might act very differently on traffic, depending on a large list of different reasons. In this thesis, we focus on two main categories of those reasons: 1) location-specific characteristics and 2) event-specific characteristics.

Location-specific characteristics describe all influencing factors related to the specific location of a venue. Is it located in the heart of a lively city or in a more rural area? To what extent is the infrastructure capable of handling additional demand due to events? In this thesis, we first analyze these characteristics by creating event-related traffic footprints within different radii around various venues. Based on these footprints we discuss the overall question of whether it is possible to generate a model for event-specific impact on traffic across different venues and locations. Is the impact of events strong enough to obscure location-based characteristics? This topic is the main focus of Chapter 4, but it is also discussed in many other sections of this thesis.

Intuitively, we expect certain road segments to be more affected by event events than others. These road segments are usually part of preferred routes that visitors take to get to the venues or leave them after the event ends. How can we find these segments and how far does the impact of events reach from the venues? This topic is called the *spatial challenge* in this thesis, and it is discussed in Chapter 5 as well as in Chapter 6. We show our approaches and results of finding the spatial impact zone around different venues based on data-driven approaches.

Event-specific characteristics describe all factors that are directly connected to the specific event happening. Which category of event is it? Who is playing? Is the concert a major act with thousands of visitors or a concert of a small local band? How many people are expected? In this thesis, we analyze the relevance of event-specific attributes for traffic impact prediction. On one hand, these attributes include features that we can derive directly from the event description. They include the category of an event, the artist and event time information (e.g., start time and entrance time). On the other hand, we collect additional information about events from online sources. This includes data from social media sites (e.g. Twitter, Facebook), statistics from the search engine Bing[1], and additional information such as the presence of a Wikipedia[2] page for instance. In Chapter 6, we analyze the information value of a large set of these attributes in terms of their explanatory power toward observed traffic disruptions. In the following study, we use a selected attribute set to gain insights about their specific use for event traffic predictions.

In this thesis, we present the chances and challenges of creating data-driven prediction approaches for event-caused traffic disruptions. We draw conclusions about the effect of each mentioned characteristic and show our approaches to handle event traffic variability for prediction approaches.

1.2 Boundaries of the Dissertation

Simulation-based traffic prediction approaches were not handled in the course of this work. Although there is a lot of work within the research community that deals with different simulation environments and frameworks, we focused solely on the data-driven domain.

Another branch that is relevant to this research but was intentionally excluded from this thesis is the field of data collection. There are many sensors available today that could serve as input to derive the traffic state and additional information. For instance, Bluetooth signals (e.g., [9, 10]) or mobile phone data

[1] http://www.bing.com

[2] http://wikipedia.org

(e.g., [11]) has been widely used within the community to derive movement patterns. In this thesis, we focus on the usability of existing commercial traffic information sources.

We have discussed several influencing factors above that affect the traffic state. Of course, this list is far from complete. There are many influencing factors that also interfere with event traffic disruption, such as weather phenomena or seasonal traffic effects [12]. In the course of this work, we focused on the characteristics discussed above and leave others for future work.

1.3 Structure of the Dissertation

This dissertation consists of seven chapters. Chapter 3 has a preparative character: it explains the data sources used within the various experiments presented in this thesis. It also briefly touches the topic of storing and processing massive datasets, the methodology for which has been published in:

- **S. Kwoczek**, S. Di Martino, T. Rustemeyer, and W. Nejdl. An architecture to process massive vehicular traffic data. In 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), pages 515–520, Nov.
- S. Di Martino, **S. Kwoczek**, and W. Nejdl. Smart Sensors Networks: Communication Technologies and Intelligent Applications, chapter Scalable Processing of Massive Traffic Datasets, pages 123-142. Elsevier, 2017.

Chapter 4 gives a first impression of event-caused traffic disruptions. We create event-specific traffic footprints around different venues and discuss the role of several location-specific and event-specific characteristics based on our observations. Parts of this chapter have been published in:

- **S. Kwoczek**, S. Di Martino, and W. Nejdl. Predicting traffic congestion in presence of planned special events. In Proceedings of the Twentieth International Conference on Distributed Multimedia Systems, DMS, pages 357–364, 2014.

- **S. Kwoczek**, S. Di Martino, and W. Nejdl. Predicting and visualizing traffic congestion in the presence of planned special events. Journal of Visual Languages & Computing, 25(6):973–980, 2014

Chapter 5 is focused on the spatial impact of events. A method is shown to identify affected road segments around venues and the findings are discussed. The presented methodology is also published in:

- **S. Kwoczek**, S. Di Martino, and W. Nejdl. Stuck around the stadium? an approach to identify road segments affected by planned special events. In Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on, pages 1255–1260, Sept.

The following chapter 6 focuses on the relevance of different event-specific characteristics. Different event attributes and additional features from online sources are evaluated in terms of their relevance and information value for traffic predictions. At the end of chapter 6 we review the results from chapter 5 and show an alternative approach for the *spatial challenge*.

Finally, in Chapter 7 we present our conclusion on the impact of PSEs on traffic and discuss the perspective for future work.

Before starting with our own analytics and findings, we discuss the current state of the art in this and related fields of research in the following chapter.

Chapter 2

Related Work

Human mobility behavior is, to a certain extent, predictable [13]. A recent study by Song et al. [14] revealed 93% potential predictability in user mobility across their database, based on mobile phone data. However, today's systems are rather far from these numbers. Although human mobility patterns tend to be habitual, the heterogeneity of these patterns is recognized to impose challenges for mobility predictions. This gains special importance whenever large crowds are involved [4]. PSEs and their impact on mobility are a good example of this phenomenon. This chapter is focused on presenting the current state of the art in research related to this domain. In the following sections, we define relevant nomenclature required to follow the rest of the thesis and outline relevant work in the fields of mobility prediction, special treatment of events, and modern event management.

2.1 Definitions

As this work focuses on the impact of PSEs on traffic, a common understanding of these terms is crucial. Here, we define *Traffic Congestion* and *Planned Special Event*, and explain the concept of *Transportation Resilience*.

2.1.1 Traffic Congestion

Generally, traffic congestion can be divided into *recurrent congestion (RC)* and *non-recurrent congestion (NRC)* [15–17]. RC results mostly from influencing factors that act periodically on the network infrastructure, such as daily commute traffic or weekend trips [16]. Because of their recurring nature, the time, location, and duration of RC events are typically known to commuters and traffic management centers. This leaves RC as a burden to drivers but makes it, to a certain extent, predictable. NRC, however, is defined as unusual congestion where the time, location, and duration are mostly unknown beforehand. NRC events are caused by external influencing factors (e.g., accidents, construction zones, PSE, etc.) and their effect on the traffic network highly depends on its local condition, travel demand, and traffic capacity.

In the following, we will refer to the "regular" traffic behavior (including RC events) as *routine* and traffic conditions with NRC events as *non-routine*, following the wording in [18]. Differentiation between *routine* traffic and when a traffic situation exceeds the habitual patterns and turns into *non-routine* is a complex task and different approaches have been reported in the literature. Most of them do not directly define a non-routine situation but infer it based on the deviation of the observed traffic from the routine, which itself is created from historic traffic information. For example, Anbaroglu et al. (2014) [17] used spatiotemporal clustering on Link Journey Time estimates (LJTs) on adjacent links in the route network to detect NRC events. They used a congestion factor as threshold with a proposed value of 1.4, indicating that LJTs that exceed the expected value by more than 40% belong to an NRC event.

Another example is found in Hojati et al. (2006) [19] where they modeled the routine for each link in the route network by applying the quantum-frequency algorithm (proposed in [20]) on loop sensor data. They compared their detected NRC events with information about reported events from traffic management centers. Although it does not deal with road traffic data, the work of Pereira et al. (2015) [4] tackles the same problem. They investigated overcrowding hotspots within public transportation network during PSEs in Singapore. They use the 90th percentile as threshold to identify when a certain station shows a passenger demand higher than the norm.

TABLE 2.1: Event categories from the Traffic Engineering Handbook. Source: [5]

Event	Planning	Advanced Notice	Duration	Hazard	Impact Area	Frequency
Vehicle Crash	Unplanned/ Emergency	None	Minutes to hours	Low	Local to several miles	Frequent
Concert/ Sporting	Planned	Months/ Years	1+ Days	None	Several Miles	Seasonally Frequent
Olympics/ One-Time Events	Planned	Years	1+ Days/ Weeks	None	Several Miles	Infrequent
Parades	Planned	Months/ Years	Hours	Low	Few Miles	Occasional
High-Security Events	Planned	Days to Weeks	Hours to Days	Low	Several Miles	Occasional
Snow/ Ice Storm	Unplanned	Hours to Days	Hours to Days	Medium	Regional	Seasonal. varies by region
Wildfire	Unplanned/ Emergency	Minutes to Days	Hours to Weeks	Medium to High	Regional	Seasonal, varies by region
Hurricane Evacuation	Unplanned/ Emergency	Days to a week	Days	High	Regional	(Seasonal) Infrequent
Bridge Collapse	Unplanned/ Emergency	None	Months	High	Several Miles	Infrequent

2.1.2 Planned Special Events

There are many different definitions of the term *event* in the literature. Although events differ in many ways, they have one thing in common: the potential to put stress on the road infrastructure in terms of capacity, safety, and demand [5]. One commonly used definition is from the Traffic Engineering Handbook [5], where they use three major categories: *planned*, *unplanned*, and *emergency events*. An event can belong to different categories at the same time (e.g., unplanned and emergency). To provide an example, Table 2.1 shows a summary from [5] with selected events and their categorization.

An alternative approach to event classification is shown in Mueller [21]. They developed a methodology to categorize events into three group: Giga-, Mega-, and Major events. They introduced a point scale to classify an event into these three classes based on four attributes: Visitor attractiveness (number of tickets sold), mediated reach (value of broadcast rights), cost, and transformation (capital investment). A similar approach can be found in the project STADIUM (**S**mart **T**ransport **A**pplications **D**esigned for large events with **I**mpact on **U**rban **M**obility) [22]. They developed two ways to categorize an event: Either by choosing from among four predefined event types (shown in Figure 2.1) or by defining the event characteristics by five attributes: magnitude, population, dispersion, frequency, and duration (shown in Figure 2.2). The STADIUM project focuses on PSEs only. From the official perspective, the term PSE is defined by the Federal Highway Administration as an event where the scheduled times,

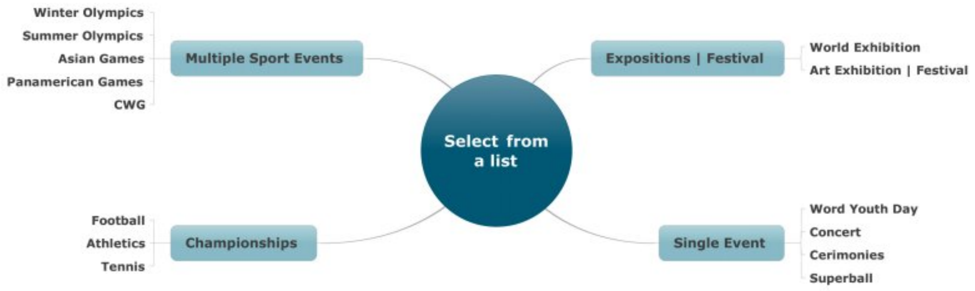


FIGURE 2.1: STADIUM: Selecting Event Types. Source: [22]

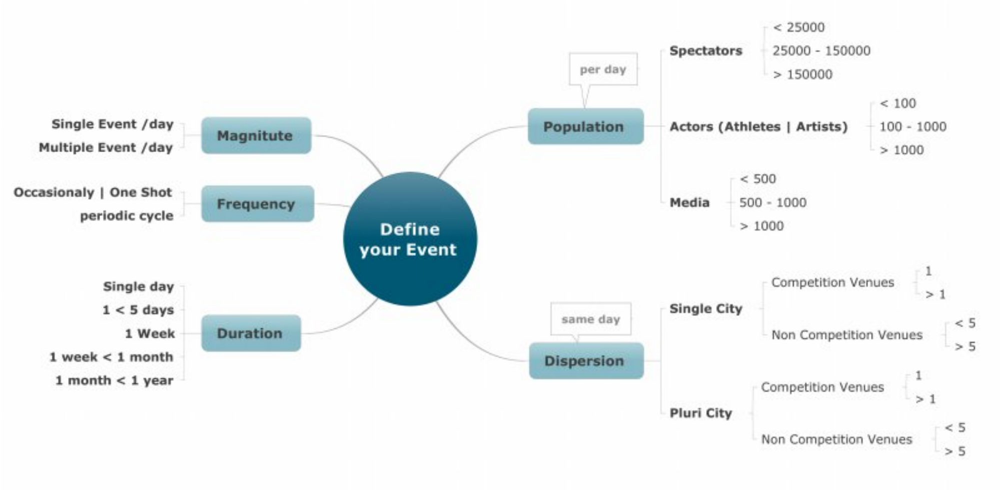
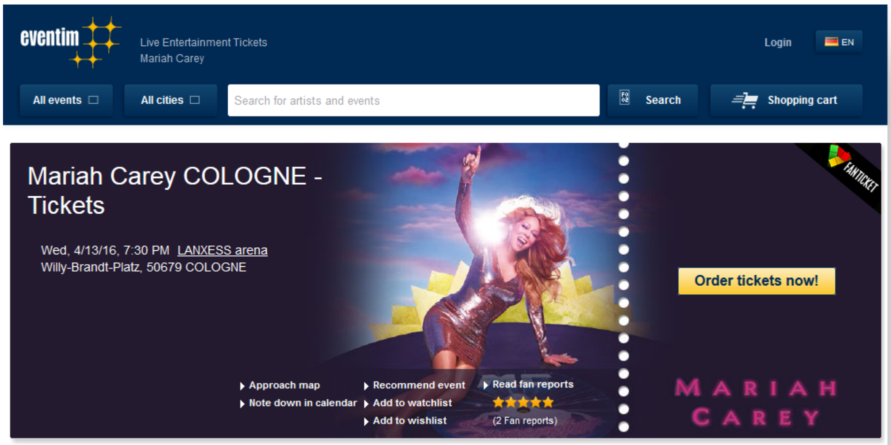


FIGURE 2.2: STADIUM: Selecting Event Attributes. Source: [22]

location, and associated operating characteristics are known in advanced [23]. Under this definition, PSEs include sporting events, festivals, concerts, and conventions occurring at permanent multi-use venues [23].

Information about PSEs can be obtained from different internet sources, such as ticket sales websites or websites from local municipalities. The data usually contain information about the place of the event, the time, the artist, and sometimes additional metadata (text description of the event, opening hours, etc.). Figure 2.3 shows an example event from the Eventim[1] website with information about the artist, date, time, and location of events. The information can either be parsed from the website, received via APIs or, depending on the website, exported into a machine readable format such as *comma separated values (csv)*. However, the data usually require cleaning and preprocessing to actually be useful for analysis purposes. Information about the preprocessing toolchain used in this thesis will be further explained in Chapter 3.

[1] `http://www.eventim.de`



Live Entertainment Tickets
Mariah Carey

Mariah Carey COLOGNE - Tickets

Wed, 4/13/16, 7:30 PM LANXESS arena
Willy-Brandt-Platz, 50679 COLOGNE

FIGURE 2.3: Example event entity from the eventim.de website. Source: `http://www.eventim.de` - Downloaded on 02/12/16

2.1.3 Transportation Resilience

For attending a PSE, many people try to reach the same destination at the same time which causes additional traffic. However, whether this situation results in non-routine congestion highly depends on the network itself. The ability to cope with additional traffic and still maintain a level of service is defined as *resilience* [24, 25]. Instead of a single measure, it can be seen as a concept that describes the network's ability to cope with changes. Although resilience itself is a concept that applies to any type of network, Murray-Tuite [26] assembled a list of ten dimensions to described the resilience of a transportation network: redundancy, diversity, efficiency, autonomous components, strength, adaptability, collaboration, mobility, safety, and recovery. The specific details of the proposed dimensions or further detailed information about transportation resilience is not required to follow this thesis and the interested reader is referred to Murray-Tuite (2006) [26] or Laprie (2008) [24] for further reading. However, understanding the general concept is crucial for the following chapters, as it explains the variation in observed traffic congestion based on the location and the network in the vicinity of the venues where PSEs occur.

2.2 Traffic State Estimation

Within recent years, high temporal and spatial resolution datasets about the traffic state have become available. Usually, these datasets are based either on:

1) readings from sensors along the road network, 2) Floating Car Data (FCD) or 3) a combination of both. In this section, we present a general brief overview without going into specific detail about the generation of these data sources. The interested reader can find more specific information in [27] or the FHWA Traffic Detector Handbook [28].

Sensor readings have been the most prominent data source for years. These datasets are generated from built-in sensors such as loop detectors, cameras, or infrared sensors along the road network. An extensive list of different types of sensors can be found at [27]. This information contains accurate sensor readings at fixed locations with very high temporal resolution (i.e., per minute) and has been heavily used within the research community over the years (e.g., [29–32]). These sensors usually provide speed measurements and traffic flow counts (e.g., number of vehicles/min). On the downside, such built-in sensors impose rather high maintenance costs for the communities and nonfunctioning sensors are usually a significant issue.

FCD is generated from so-called *probe vehicles*. Each vehicle sends information about its position, speed, and heading to a server. The traffic state is derived on the server side from a collection of information from different vehicles for the same road segment. FCD is mostly extended through the use of mobile phone data (called cellular-FCD), whereby movement patterns from mobile phone providers are included. While FCD was preliminarily used as an extension to stationary data, research has shown that it can be successfully used to derive the traffic state [33–35].

Commercial providers of traffic data usually use a combination of the mentioned data sources to derive the traffic state. For instance, TomTom NV published their methodology in [36, 37]. In the presented approach, a central system merges live traffic information from multiple sources. Those include sensor readings from local traffic management centers, and a collection of FCD and GSM probe data from mobile phone companies. The FCD is obtained from users of TomTom navigation devices who opt-in to share their data to improve the traffic state estimation. Each participating user receives information about the current traffic state and in return, sends its position and speed to the central server. For the GSM probe data, TomTom collects information about cellular phone movement between different GSM network cells, which they receive from the mobile phone companies. This information source has also been proven to give valuable information about the traffic state on the roads (e.g., see [38]). A weighted combination of all available information sources allows the fairly accurate deduction of the traffic state with large geographic coverage. Of course, the described methodology from [36] describes TomTom's approach from 2009 and some information is probably outdated today. Still, it gives a fair overview about the general principle of commercial road traffic state deduction.

2.3 Traffic Prediction

Traffic prediction has become as widely studied field within the ITS community. In general, there are two major approaches for traffic prediction: model-based and data-driven. Model-based approaches incorporate sophisticated models (e.g., traffic flow models and traveler behavior models) in a simulation to represent the holistic traffic situation. However, data-driven approaches rely on fitting the current situation to previous observations by modeling the relationships between independent variables (e.g., from temporal and spatial characteristics, information from sensors or other information sources) to a target variable (e.g., traffic flow) [39]. Although these fields are merging more and more, this work focuses on the data-driven domain only.

2.3.1 General Overview

Within the data-driven traffic prediction domain an extensive list of research activities can be found. Searching for the term "road traffic prediction" with Google Scholar[2] yielded approximately 513000 results on December 5th 2016. Approximately 20000 of these results are from 2016. Owing to the vast amount of research, a complete overview about all exiting approaches is difficult to provide in this thesis. Instead, in this section, we will mention a list of papers that are relevant and somewhat representative of their specific domain.

In the past, there have been four major papers that focus on summarizing the current approaches and offer an extensive literature review. In [40], an overview of the literature on short-term traffic prediction up to 2003 is provided. In [41]

[2] `www.scholar.google.com`

and [42], the authors focus on the domain of artificial intelligence (AI) and artificial neural networks (ANNs) within the short-term traffic prediction domain. The fourth paper is [43]. Apart from giving an overview of the current state of the art, they also define ten challenges that research in this domain is currently facing.

When browsing the literature for a set of different traffic prediction methods from the data-driven domain, one finds a large variety of different "schools of thoughts." In this work, we follow the definition from [44], under which methods are separated into two main approaches: *statistical* and/or *Computational Intelligence (CI)* [44].

Statistical approaches for modeling have existed for decades and their mathematical foundations are solid and widely accepted. However, those approaches mostly struggle when dealing with high-dimensional and highly nonlinear datasets [44]. The second class of approaches, CI, is a sub-branch of AI that concentrates on the creation of intelligent behavior in complex and changing environments by combining elements of adaptation, evolution, fuzzy logic, and learning [44–46].

One large subclass of statistical methods for short-term traffic prediction are the so-called autoregressive integrated moving average (ARIMA) methods, which were first introduced to the domain by Ahmed and Cook [47]. Working on stationary time series only, these methods have been widely used within the community for traffic prediction [48–52]. However, these approaches have lately been losing attention, as they are usually not high-performing under unstable traffic conditions or complex road network topologies and settings.

Instead, supported by an increased availability of high-resolution mobility datasets and growing computational resources, CI methods are on the rise. In particular, the growth of ANNs, an extremely popular class of CI models, is noticeable. Among the ANN models, multilayer perceptron (MLP), radial basis function neural networks (RBFNN), and back-propagation neural networks (BPNN) are the most popular [44]. Which of these models shows superior performance is usually highly dependent on the use case [53–56].

Combinations of statistical and CI approaches have also been investigated over the years. Already by 1996, [57] used a self-organizing map as classifier, whereby each defined class had an underlying ARIMA model associated with it. They predicted traffic with time horizons of 30 min and 60 min on French highways. In [58], they used genetic algorithms (GAs) to optimize the structure of a time delay neural network (TDNN). Their model showed good performance on data from the California testbed in Orange County and outperformed other ANN-based approaches. A generic optimization strategy based on GAs that supports structuring of the traffic datasets and choosing the neural network structure has been shown in [59]. In [60] they used a fuzzy–neural model (FNM) for traffic prediction in urban street networks on five test sites in Hong Kong. They proposed a two-module approach, in which the first module groups similar traffic situations and the second module predicts the traffic situation based on the selected clusters. In [61] they proposed a Bayesian combined neural network (BCNN) approach to join different models for different prediction horizons. They used a neural network model to combine the predictions from single-neural-network models based on a credit assignment algorithm.

Apart from the many different algorithms and methods, the common goal stays the same: predicting traffic more accurately and in a scalable and (real-world) applicable manner. The abovementioned list of selected studies is far from complete, and does not capture the entire field. However, it gives a comprehensive overview of the different research directions pursued over the years. Keeping the research question of this thesis in mind, we will close this chapter by focusing on additional work that is seminal to ours.

2.3.2 Traffic Prediction for Special Event Scenarios

As mentioned earlier, the specific problem of predicting traffic during PSEs has not been extensively investigated within the community. However, there is a list of publications that focus on research questions similar to those outlined in this thesis.

The first example is the work presented in [29]. They investigated different methods for time-series-based traffic prediction on highways for RC and NRC. They benchmarked their models with two baseline approaches, namely the ARIMA and historical average model (HAM) approaches. ARIMA used a defined set of past observations to predict the upcoming data, whereas HAM relied on models based on historic information only, without any adjustment to the current traffic situation. As a result, the ARIMA model showed better performance than the HAM for short-term traffic prediction. However, for longer time horizons the performance of the ARIMA decreased, whereas the HAM performance was independent of the time horizon. Using this observation, they build a hybrid model called historical ARIMA (H-ARIMA) that combines both models. This model performed well for day-to-day traffic prediction (RC events). However, owing to missing contextual information about the traffic state, its use for NRC events was limited.

As an extension, focusing on also including NRC events, they proposed a model called H-ARIMA+. It included additional information about traffic incident characteristics in terms of five attributes: *start time*, *location*, *direction* (on the highway), *type* (e.g., road construction, traffic collision) and *affected lanes*. They applied their models to predict two variables: *impact post-mile* and *speed impact*, where the *impact post-mile* was defined as the length of the affected stretch of the highway and *speed impact* was defined as the speed decrease on the sensors.

The incident attributes were used to cluster impact post-mile values. For each cluster the average post-mile value was used to represent the impact of events with similar attributes. In their experimental setup for NRC event prediction, they benchmarked the H-ARIMA, H-ARIMA+, and a MLP ANN. Their results showed that the H-ARIMA+ model outperformed the others, especially at the beginning of NRC events when the other models were incapable of properly reacting to the traffic dynamics.

The presented approach showed good results for the outlined scenario. However, the applicability for PSE prediction in urban areas is rather limited. Although the concept of an *impact post-mile* works for highway scenarios without on- and off-ramps, specifying the impacted area in an inner-city road network is a more complex research task. Moreover, the shown similarity measure of events is limited owing to the very few attributes of which some are highly restricted to highways scenarios (e.g., number of blocked lanes). In terms of the methodology, their approach uses a representative value for each observed event class. This assumes a homogeneous behavior of traffic during events of the same class, which, as we will describe later, is not always the case in our scenarios.

In [30] they used a large dataset about incidents and traffic situations on a highway in Los Angeles, USA to predict the spatiotemporal behavior of incidents. They considered the initial behavior of traffic when the incident occurred in combination with the long-lasting congestion propagation wave that results from the incident. They classified incidents based on selected features: street name, start time, affected number of lanes, and incident type (called level 1) and clustered all incidents belonging to the same group according to their traffic volume and occupancy features (called level 2). Their first prediction algorithm, PAD, takes a new incident, defines its group by the basic feature set (from level 1), finds its containing cluster in level 2 based on its traffic volume and occupancy level, and uses the average propagation behavior of all other incidents from the identified cluster for prediction. Their second approach (PADI) uses the initial behavior of incidents to further refine prediction candidates by clustering candidates of each cluster in level 2 according to their initial behavior (called level 3).

To predict new incidents, PADI selects the right candidates from levels 1 and 2 and uses the observed initial behavior of the new incident to match the cluster centroid that corresponds best based on the Mahalanobis distance between the initial propagation behaviors. Their results show that with the additional information about the initial behavior, they outperform their baseline (using incident features of level 1 only) by up to 45%.

The approach showed a possible way to include additional incident (or event) information in the prediction algorithm. To apply the presented methodology to real-word PSEs, a feature set is required that allows us to cluster similar events. A possible similarity measure for this clustering will be discussed in later chapters.

In [7], they analyzed traffic on a 45 mile stretch of the I-880 in California, USA and showed a method to divide the observed congestion into six different components: 1) congestion caused by incidents, 2) congestion caused by PSEs, 3) congestion caused by lane closures, 4) congestion caused by weather phenomena, 5) congestion caused by nonoptimized metering, and 6) congestion caused by daily commute traffic. Under the assumption of a linear contribution to the overall observed delay of each individual component they built the following model:

$$\begin{aligned} D_{total}(d) &= \beta_0 + \beta_{col}X_{col}(d) + \beta_{event}X_{event}(d) + \beta_{lane}X_{lane}(d) \\ &\quad + \beta_{weather}X_{weather}(d) + \epsilon(d) \end{aligned}$$

where $\epsilon(d)$ is the error term, $X_{col}(d)$ is the number of incidents on day d, $X_{event}(d)$ is the number of PSEs on day d, $X_{lane}(d)$ is the number of lane closures on day d, and $X_{weather}$ a boolean indicator of whether there was an adverse weather situation on day d (source: [7]). They used a linear least squares method to fit the model to their data to retrieve the model parameters β. The results showed the individual contribution of each component to the overall delay, whereby PSEs accounted for (only) 4.5% of the observed daily delay. The approach showed a straightforward way to divide observed delays into their individual component causes. Of course, the individual influence of each component is highly dependent on the location of the study, the time of day, and other factors.

In [62], they analyzed a large set of mobile phone traces within the Boston, USA metropolitan area. They showed that events of the same category resulted in a similar spatial distribution of origin locations from visitors. This outcome also matches the intuition that certain events (e.g., sporting events) are usually attended by the same people. Those findings are of fundamental importance for this work, as they motivate the assumption that the event category has a significant impact on the spatial distribution of event traffic. However, they focused on hand-picked events that were expected to be large enough to attract at least a certain number of people, which drastically reduces the complexity resulting from impact variation across events of different size.

In [63], they used information from the web about PSEs to predict public transport arrival numbers. Event information was obtained from the eventful.com[3] website, including (among other information) information about the title, date, venue, price, start time, popularity, category, and a text description of the artist. They evaluated different models from the Weka framework ([64]) for their regression task and concluded that the proposed ANN model was the superior choice. To the best of our knowledge, this was the most comprehensive analysis of the use of Internet data for event prediction at that time. The results clearly showed that, to a certain extent, received information from publicly available

[3] `http://eventful.com`

Internet sources improved their prediction. In a later publication [4], they analyzed the use of Internet data to explain overcrowding behavior in public transportation during PSEs in Singapore. As already outlined in 2.1.1, they defined overcrowding as those points where the observed arrival numbers exceed the 90% percentile of historic data at a given location. Whenever this criterion holds, they referred to a hotspot impact.

They collected information about PSEs from different online sources (e.g., eventful, last.fm, etc.). In addition to structured information about the events (location, start time, artist, etc.), they also extracted event descriptions in text form and applied topic modeling to create topics from these descriptions using latent Dirichlet allocation (LDA). As a popularity indicator for events they collected the number of Facebook[4] likes and the number of results in Google[5] for the event title.

They based their experiment on the assumption that the specific event contribution is latent and that there is going to be an explainable part (due to event-specific) and a nonexplainable part of contributions to the overall hotspot. Thus, the hotspot impact h is defined as:

$$h_{r,j} = a_{r,j} + b_{r,j}$$

where r is the area index, j is the hotspot index, a is the nonexplainable component, and b is the explainable component. They assumed a Gaussian distribution for a and the specific event contribution and modeled the impact using a Bayesian hierarchical additive model, where they defined a and b as:

$$\begin{aligned} a &\sim \mathcal{N}\alpha^T x_\alpha \sigma_\alpha \\ b &= \sum_{k=1}^{K} \mathrm{e}_k \text{ with } e_k \sim \mathcal{N}(\beta^T x_{e_k}, \sigma_k) \end{aligned}$$

where x_α, α, and σ_α are attributes, parameter vectors, and variance of a, respectively. The explainable part b is thereby defined as a sum of individual event contributions e_k, where x_{e_k} corresponds to the individual event attributes of event k and σ_k and β correspond to the event variance and attribute parameters, respectively.

[4] http://facebook.com

[5] http://google.com

They implemented their model into the Infer.NET[6] platform. As the event impact is *latent*, they evaluated their model on synthetic data and ran a quantitative analysis on real data. The results showed good performance of the model for the synthetic data and results that were intuitively plausible for the real-world dataset. The shown approach is very interesting for the purpose of explaining traffic behavior in the vicinity of venues during events. However, the approach worked for stationary locations with continuous time series data, and applying this approach to sparse traffic datasets is challenging.

2.4 Inferring Information about Real-world Events from Online Sources

Apart from traffic analysis and traffic prediction, a critical factor for our research is to gain additional information about PSEs. Within that domain, research has shown that social media can be used to discover and get insights about real-world events. A good (and in the meantime, rather famous) example is [65], where Twitter news was used to detect earthquakes in Japan or [66] and [67], where Twitter was again used to detect real-world events. In general, Twitter data is used in a large variety of different domains, from detecting riots [68] up to small-scale events such as a factory fire [69]. Not only has the detection of real-world events from Twitter been shown, but also the use of that information source to gain detailed contextual information about the events. For example, [70] created a summary of sporting events using Tweets. In [71], information was put onto a timeline to serve as a summary of scheduled events. Another example in which Twitter was used to examine real-world happenings was shown in [72]. They examined the use of Twitter to identify important changes in the city in real time. They identified a change in the overall city behavior on Twitter during the Mobile World Congress 2012 in Barcelona, Spain. Other related examples can be found in [73–75].

For the specific application of gaining information about the traffic state from social media, most efforts can be divided into two main categories: traffic detection and traffic prediction [76].

[6] `http://infernet.azurewebsites.net/`

2.4.1 Traffic Detection

To detect traffic, [77] described a system that used Twitter data in Thailand to extract tweets that contained traffic-related information together with positioning data. They used a dictionary approach to classify relevant tweets and retweeted them to broadcast the information to the public. In [78], they compared traffic-related tweets to an incident database from the California Highway Patrol. From their results, they claimed that tweets can be matched to traffic incidents. Tweets seemed to be posted within 5 h of the incident that they refer to and were sent from a location 10–25 miles away from the incident location. In [79], they analyzed Twitter to find incident-related messages. They used extracted features from part-of-speech tagging to train a classifier that detected relevant messages. In a second step, they applied the classifier to live twitter streams to detect incidents in near-real time. In a similar way, [80] presented a system called Dub-STAR to derive the underlying causes of traffic congestion. They fused information from social media, historical traffic information, DBPedia, event information, information about road works, and data about the road network topology. In [81], they analyzed Twitter messages in terms of transportation-related information using a keyword-based hierarchical annotation schema for message categorization. Also using Twitter, [82] analyzed traffic on Italian highways and used support vector machines to classify Tweets as traffic-related or not. For the set of traffic-related messages, they investigated a multiclass classifier to detect whether the congestion was due to an external event (defined as scheduled event or unexpected event). In [83], they used Twitter to detect road hazards by using sentiment analysis on Twitter messages. In [84], they investigated the correlation between Twitter concentrations and the traffic surge in July 2014 in Virginia, USA. Their results showed promising findings about the correlation between Twitter message concentration and the local traffic surge in the area. In [85], they detected traffic-related tweets and used a combination of language models and hinge-loss Markov random fields to find the traffic incident locations. They evaluated their findings with real-world traffic information from INRIX for Philadelphia and Washington, DC in the USA.

2.4.2 Traffic Prediction using Social Media Information

The literature about research that explicitly used social media information for road traffic prediction is sparse. In [86], they presented an algorithm for long-term traffic prediction that incorporated information from Twitter. They showed that their models that considered traffic intensity, Twitter semantics, and intensity outperformed the models without additional information from social media. In [87], they extracted weather-related information (e.g., snowfall) from Twitter and used the information to improve their linear regression models for highway traffic prediction.

2.5 Summary and Conclusion

The previous sections have shown a brief overview of the current state of the art within the domain of our work.

Apart from the presented approaches within the research community, there is also currently a set of commercial companies that deal with traffic. Whereas some of them have their core business in the navigation domain (e.g., TomTom International BV[7]), others deal with traffic information in other applications such as map data provisioning (e.g. Google[8]). These companies usually collect traffic-related data from their users (e.g., from their own handheld devices as TomTom International BV or Google from their Android users) to process in their analytic components. Of all the known providers, only two have announced a special treatment of PSEs in their traffic engines: INRIX, Inc. and Google.

On their website, INRIX, Inc. claims a consideration of concerts and sporting events in their INRIX XD Traffic stream [88]. In the FAQ section, they write: "INRIX uniquely factors in historical data with information about the traffic impacts of unique local events like weather, concerts, sporting events and school schedules to reliably help drivers know what to expect before they ever get into their car."[89].

[7] `https://www.tomtom.com`

[8] `http://www.google.com/`

Google Maps[9] on Android sends notifications about nearby events that might cause traffic on your route. The website states: "You get notifications about nearby events or road closures when Google Maps thinks it might affect a route that you travel often. If we know about a scheduled event, you'll get an alert ahead of time so you can plan an alternate route. For example, if there is a concert on your way home from work, you might get a notification one day before the concert."[90].

Neither company presents details of their methodologies, and no publication about their algorithms or results are available to the public.

In general, although traffic prediction and the inference of information about real-world events from online sources has been widely studied in recent years, the specific impact of PSEs on traffic remains unknown.

[9] `maps.google.com`

Chapter 3

Data Sources and Preprocessing

Following the current state of the art within the domain of this thesis, this chapter introduces the different information sources used within our research activities: traffic data and event data. In the following sections, we explain the data sources in terms of their information value and original structure. We then briefly outline the required preprocessing steps and close with a discussion of the specific values of each source.

3.1 Traffic Data

Within recent years, many research efforts have dealt with various types of traffic information (see Chapter 2.2). The most prominent data source is probably that generated from sensors built into the infrastructure. While such continuous sensor readings are easily applicable to analysis use cases, they come with a drawback: most of them are only available for small geographic regions or focus solely on highway scenarios. As the focus of this thesis is the analysis of traffic disruptions in dense inner-city scenarios in a variety of different spatial regions, this poses a problem. In addition, in Germany, those datasets are managed by local municipalities and retrieving data for different spatial regions would require access to a vast number of different providers and interfaces. An alternative to fixed sensor information is FCD (see also Chapter 2.2). Whereas FCD provides broader coverage, the problem remains: there is no single point of access to large fleets available today that could cover the geographic regions required.

However, there is a list of companies that currently provide access to high-resolution datasets. These companies usually derive their traffic states from a combination of sensor readings, mobile phone data, and FCD (e.g., see [36]). Companies such as TomTom NV[1], INRIX[2], or HERE[3] cover traffic information for broad geographic regions that they use for their own specific products. Most of them have their core business in the traffic and traffic data provisioning domain. They can provide access to aggregated traffic information for large geographic regions that match our spatial requirements. One disadvantage of using such data is that the structure of these datasets is optimized for their specific use cases (e.g., providing real-time traffic information to navigation devices) and they were never meant to serve as input for analytic tasks. This results in rather high preprocessing efforts, which we will further explain in this chapter.

Although these datasets are available on the market today, to the best of our knowledge, they have not yet been used extensively within the ITS research community. In our scenarios, we used two different types of traffic data: *Incident Data* and *Flow Data*. In the following sections, we will describe these data sources in detail. At the end of this section, we will close with a comparison of them and a discussion about their specific advantages and disadvantages.

3.1.1 Flow Data

The *flow data* received is a collection of speed information for Germany between 2014 and 2015. It contains traffic information for major road segments based on the table-based referencing system called *traffic message channel (TMC)* and is updated twice per minute. Each data update contains information for road segments where the current speed fell below 80% of the speed that could be driven under free-flow conditions. For example, if the free-flow speed on a given road segment is usually 80 km/h, we receive a message as soon as the average speed drops below 64 km/h. The 80% threshold is defined by the industrial provider. As long as the driven speed stays below 80%, we receive continuous updates for this road segment.

[1] https://www.tomtom.com/

[2] http://inrix.com/

[3] https://here.com/

Name	Explanation
TrafficUpdate	the received data package that contains traffic information for all referenced TrafficLocations within Germany for one particular time stamp
TrafficMessage	a subset of the TrafficUpdate, that contains traffic information for one specific TrafficLocation for one particular time stamp
TrafficLocation (TL)	one referenced road segment
FreeFlowSpeed (FFS)	speed that is usually driven on one TrafficLocation when it is not congested
FreeFlowPercentage (FFP)	percentage of the FreeFlowSpeed that is the average speed on a given TrafficLocation at one specific time stamp

TABLE 3.1: *Flow Data* - naming conventions.

Preprocessing

Each data package received (hereafter referred to as *TrafficUpdate*) contains information for all road segments within Germany. The data are divided into subpackages for each referenced road segment that we refer to as *TrafficMessage*. Each *TrafficMessage* contains the TMC code ID for the referenced road segment (hereafter called *TrafficLocation*), information about the *FreeFlowSpeed* (in km/h), and the current traffic state for this *TrafficLocation*. The traffic state is given as the current average *FreeFlowPercentage*. For example, if a road segment with a given *FreeFlowSpeed* of 80km/h is currently congested and the average speed is now 40km/h, the *TrafficMessages* gives 50% as the current *FreeFlowPercentage*. A summary of the naming conventions is listed in Table 3.1. The data are received in the protocol buffer[4] format from an external storage. To preprocess the data, first each update gets decoded into the aforementioned subparts. In the next step, all TMC codes get decoded to spatial objects (called *map matching*). Map matching is done using open-source tools and a digital map based on the navigation data standard (NDS)[5]. Each TMC code ID is resolved to a set of NDS road segments and their geometries are used to build the *TrafficLocation* geometries for this specific TMC code. As TMC is a fixed referencing system, the *TrafficLocation* does not change and the map matching procedure only has to run once for the entire dataset.

[4] https://developers.google.com/protocol-buffers

[5] http://www.nds-association.org/

The data are stored using a NoSQL database. Further details about our processing toolchain can be found in [91, 92].

Traffic Flow Statistics and Examples

An overview of the spatial distribution of our flow dataset is given in Figure 3.1. It shows all approximately 55000 *TrafficLocations* throughout Germany. Their density rises in urban areas and decreases in rural areas. Based on these TLs, one can easily spot large German cities on the map shown in Figure 3.1. At the city scale, the different spatial distributions get more obvious. Figure 3.2 shows a comparison of three big cities in Germany (Berlin, Cologne, and Hamburg). Each *TrafficLocation* (in black) has a driving direction. If the same road segment is drivable in both directions we get two separate *TrafficLocations* resolved. Figure 3.3 shows traffic flow data for two different road segments within the inner city of Berlin, Germany. The *TrafficLocations* represent the same road segment (Hauptstrasse, Berlin-Schoeneberg) but for different driving directions. Both *TrafficLocations* are highly congested on that day. In particular, the location in Figure 3.3iv essentially only reaches *FreeFlowSpeed* at night, and is almost constantly congested from approximately 06:00 until midnight.

With the given update rate, spatial coverage, and time frame, this dataset gives detailed traffic information for major roads all over Germany. The preprocessed data stored in the NoSQL database accounts for approximately 1 TB of required storage space.

3.1.2 Incident Data

We received an incident dataset with a broad coverage of the entirety of Germany between 2013 and 2014. It contains information about severe traffic congestion on major and side roads based on a dynamic georeferencing system called OpenLR[6]. Unfortunately, we cannot further specify the term *severe traffic congestion* here, as we did not receive any information about the threshold from the commercial provider. However, in comparison to the *Flow Data* (explained in the previous section), the *Incident Data* is a small subset of it.

[6] http://www.openlr.org/

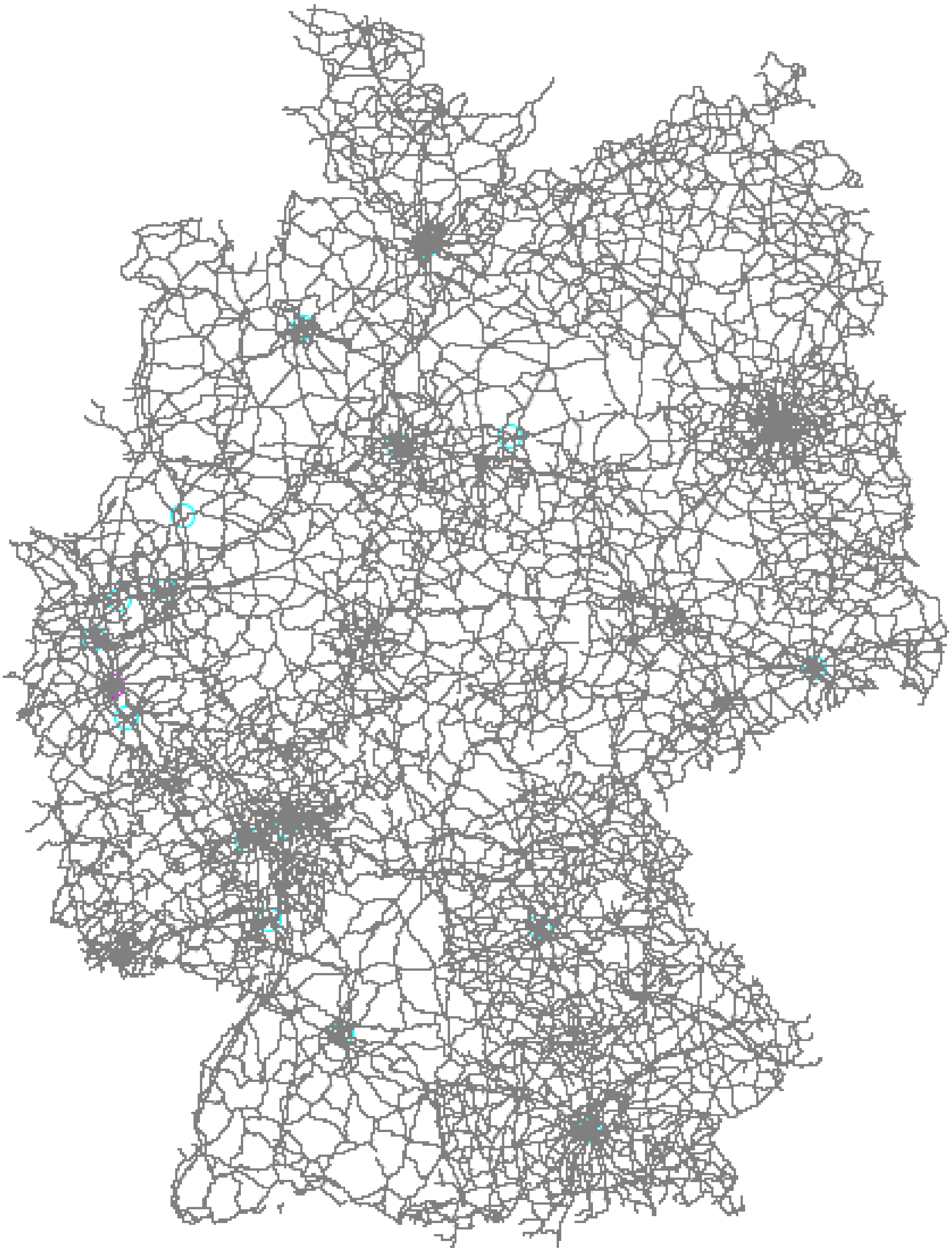
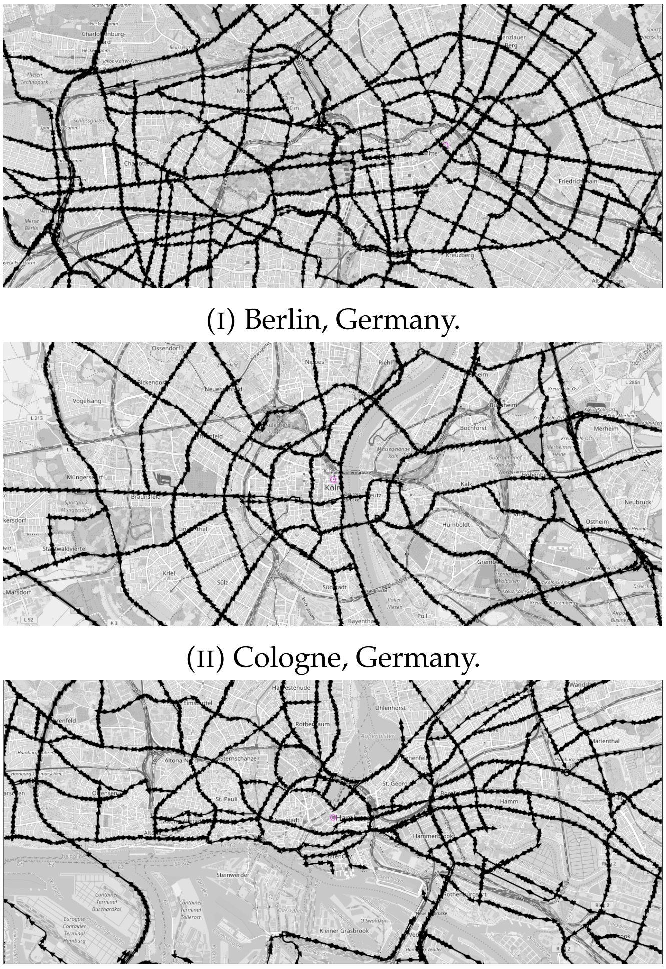


FIGURE 3.1: Overview of all *TrafficLocations* of the *Flow Data* dataset in Germany.



(I) Berlin, Germany.

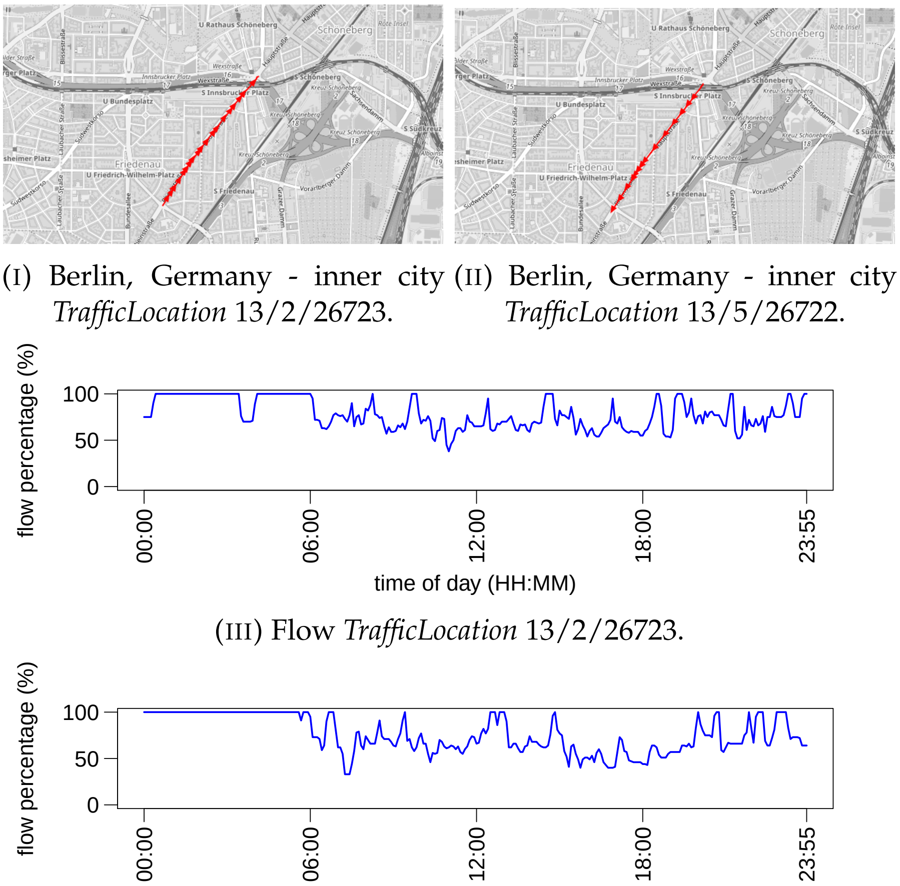
(II) Cologne, Germany.

(III) Hamburg, Germany.

FIGURE 3.2: Flow Data *TrafficLocation* distribution in urban areas. Screenshots of *TrafficLocations* within major cities in Germany.

The OpenLR reference is a binary code that can be translated to a set of road segments and driving directions from an underlying map using open-source libraries.

Each data package received (hereafter referred to as *TrafficUpdate*) contains information for all road segments within Germany. The data is divided into subpackages for each referenced road segment, that we refer to as *TrafficMessage*. Each *TrafficMessage* contains the OpenLR code for the referenced road segments (hereafter called *TrafficLocation*) and the current traffic state for this *TrafficLocation*. For each *TrafficLocation*, we received information about the current average *Speed* (in km/h), *DelayTime* (in s) and *CongestionLevel* (on a scale from 1 to 5). A summary of the naming conventions is listed in Table 3.2.



(I) Berlin, Germany - inner city *TrafficLocation* 13/2/26723.

(II) Berlin, Germany - inner city *TrafficLocation* 13/5/26722.

(III) Flow *TrafficLocation* 13/2/26723.

(IV) Flow *TrafficLocation* 13/5/26722.

FIGURE 3.3: Example traffic flow on selected *TrafficLocations* in Berlin, Germany on 12th of May 2015.

Figure 3.4 shows an example of the data, where two incidents are shown on the map.

The data are received via an online connection using XML files that follow the Datex2[7] schema. After the *TrafficUpdate* has been received and decoded into the specific *TrafficMessages* the data are *map matched* and stored in a relational database.

Map matching is done using open-source tools and a digital map based on NDS[8]. Each OpenLR code ID is resolved to a set of NDS road segments and their geometries are used to build the *TrafficLocation* for this specific OpenLR code. As one OpenLR is a dynamic referencing system that changes over time,

[7] http://www.datex2.eu/

[8] http://www.nds-association.org/

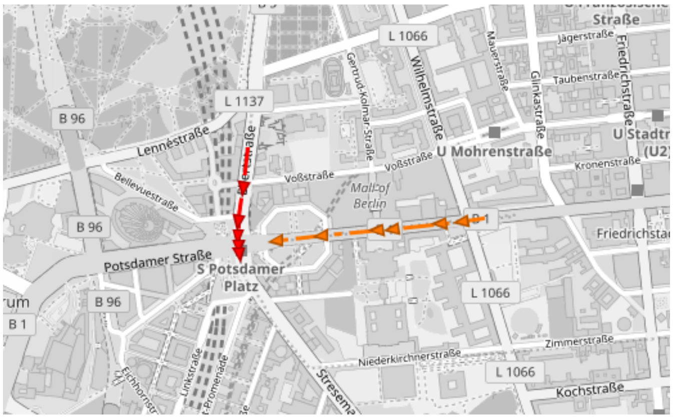


FIGURE 3.4: An example of a traffic congestion received from the Incident Data set. The image shows two messages where severe congestion was observed in Berlin on 08/02/2014 at 18:20.

Name	**Explanation**
TrafficUpdate	the received data package that contains traffic information for all referenced TrafficLocations within Germany for one particular time stamp
TrafficMessage	a subset of the TrafficUpdate, that contains traffic information for one specific TrafficLocation for one particular time stamp
TrafficLocation	one referenced road segment
Speed	information of the average speed (in km/h) that is currently driven on this TrafficLocation
DelayTime	additional time (in s) that is needed to pass the referenced TrafficLocation compared to noncongested conditions
CongestionLevel	index about the severity of the congestion (1: little congestion, 4: highly congested, 5: unknown) based on an internal scale from the data provider

TABLE 3.2: *Incident Data* - naming conventions.

TrafficLocations might have overlapping NDS road segments. The challenges caused by this characteristic are further explained in 3.1.3.

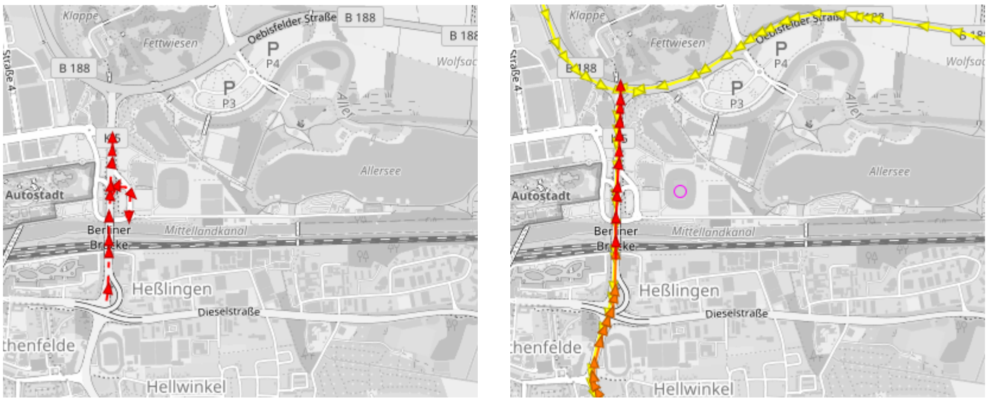
We implemented a solution for storing the dataset, based on a standard relational database management system (RDBMS), using a PostgreSQL[9] database with the PostGIS[10] extension.

[9] `http://www.postgresql.org/`

[10] `http://postgis.net/`

3.1.3 Comparison

Although both traffic information sources provide information about congestion, they differ significantly in their structure, information value, and usability. An example can be seen in figure 3.5 where traffic information is shown from both data sets for the exact same location and time. In terms of spatial coverage the *Incident Data* is clearly the more comprehensive dataset. Using the OpenLR referencing system, this dataset includes many roads that are not covered by TMC. For example, the *Incident Data* in figure 3.5 shows severe traffic congestion on a bridge in Wolfsburg/Germany. While the *Incident Data* also contains the resulting congestion on the off-ramp (the small red cycle on the east side of the bridge) this information is not included in the *Flow Data*.



(I) Incident Data

(II) Flow Data

FIGURE 3.5: Comparison of Incident Data and Flow Data for the same location at the same time. Location: Wolfsburg. Date: 08/02/2014. Time: 15:00.

On the other hand, the data quality of the *Flow Data* is much higher. The dataset contains information with an update interval of 0.5 s whenever speed falls below 80% of the *FreeFlowSpeed*, which is much more than the *Incident Data* contains. Again, that becomes visible in figure 3.5 as the *Incident Data* only covers the congestion on the bridge while the surrounding congestion on the other road segments is only visible in the *Flow Data*.

However, the biggest difference is between the utilized referencing systems. OpenLR is optimized for low bandwidth consumption when sending updates over the air to mobile devices and for being independent of the utilized digital map. These are specific requirements from the mobile navigation domain, where devices receive their updates via a UMTS or LTE connection. This is achieved by using dynamic referencing methods, whereby affected road segments are grouped together into one *TrafficLocation*, as long as they show a similar traffic behavior. While this method fits the requirements for the typical OpenLR use case, it complicates the usage of such data for analysis tasks. As each OpenLR code references not only one but a list of road segments, information from the *TrafficMessage* (e.g., delay time) has to be split up among them. For example, consider three road segments that are connected to each other (segments **A**, **B**, and **C**). Segments **A**, **B**, **C** are 45, 100, and 5 m long, respectively. Now imagine a *TrafficUpdate* with an OpenLR code that references those three segments and gives a *DelayTime* of 20 s. How do you distribute those 20 s among the involved road segments? We used information about the length of the segments and an even distribution. In this example, we would assign segment **A** a delay time of *20(s)×45(m)/150(m) = 6(s)* and for the segments **B** and **C**, *13.7(s)* and *0.7(s)*, respectively. But does this reflect the real situation on the road? Traffic lights, accidents, crosswalks, or other influencing factors might lead to a very different delay distribution.

The example also reveals another issue with OpenLR in this domain. Because there are no fixed *TrafficLocations*, road segments from a digital map have to be used as *TrafficLocation* geometries. Although these digital maps are usually also designed for different use cases, their road schema might not be optimal. One often finds very short road segments (especially at intersections) that complicate the analysis process.

The combination of very sparse information (only if traffic is heavily congested) and a dynamic referencing system makes it very difficult to derive a coherent picture of the traffic behavior of specific *TrafficLocations*. For example, Figure 3.6 shows the traffic data for one specific road segment for 24 h. On this day, the data show two severe congestions by which this *TrafficLocation* was affected (around 07:00–08:00 and 18:00). At times without heavy congestion, we do not have any information about that *TrafficLocation*, but we see drastic peaks during congestion times. Compared to other traffic information formats, for example sensor readings, these data make it very difficult to apply any sort of statistical learning method.

However, when dealing with geographic regions instead of focusing on the road network level, this data source can be used to identify the traffic behavior for this region. This approach has been used, for instance, in [93] and [94] to

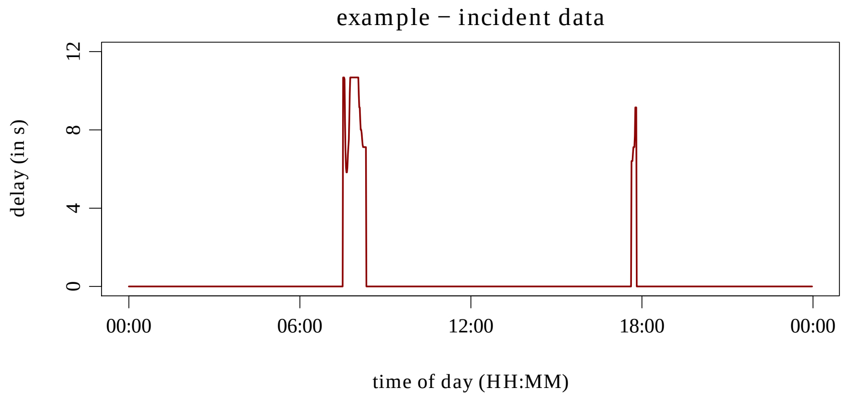


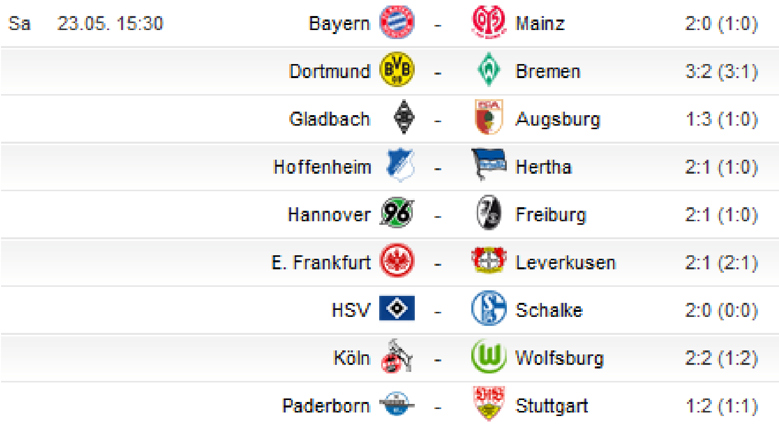
FIGURE 3.6: Example *Incident Data* for one *TrafficLocation* on 05/06/2013 in Berlin, Germany. y-axis: delay time, x-axis: time of day.

Information	Incident Data	Flow Data
Update interval	60 s	30 s
Location referencing	OpenLR	TMC
Supported data format	Datex2	Protocol buffer
Size per *TrafficUpdate*	5.6 MB (XML)	26 MB (Protobuf)
Compressed size (gz)	326 KB	12038 KB

TABLE 3.3: Traffic statistics for the Flow & Incident dataset.

observe the footprint around the LANXESS Arena in Cologne, Germany (which is also shown in Chapter 4).

Although the *Flow Data* set only covers major road segments its information value on these segments is significantly higher compared to the *Incident Data*. Although we do not get any observations of traffic on segments that have a *FreeFlowPercentage* above 80% of the *FreeFlowSpeed* we get a rather coherent picture of the traffic situation when there is congestion (as seen in Figure 3.3). However, the added information leads to a significant growth of the data source in terms of size. We implemented a completely new import toolchain and storage concepts able to handle the large amounts of data (as presented in [91, 92]). Table 3.3 shows a comparison of the statistics of the *Flow Data* and the *Incident Data*



Sa	23.05. 15:30	Bayern	-	Mainz	2:0 (1:0)
		Dortmund	-	Bremen	3:2 (3:1)
		Gladbach	-	Augsburg	1:3 (1:0)
		Hoffenheim	-	Hertha	2:1 (1:0)
		Hannover	-	Freiburg	2:1 (1:0)
		E. Frankfurt	-	Leverkusen	2:1 (2:1)
		HSV	-	Schalke	2:0 (0:0)
		Köln	-	Wolfsburg	2:2 (1:2)
		Paderborn	-	Stuttgart	1:2 (1:1)

FIGURE 3.7: Example event dataset about soccer games from the kicker.de website. Contains information about games of the first soccer league in Germany (time, teams, location, and results).

3.2 Event Data

The Internet provides a vast variety of event information for different venues. In our research, we focused on two different information datasets: One that we collected manually and another, which we received from an industrial provider that focuses on ticket sales.

The manually collected data were mostly retrieved by parsing relevant websites. For example, we collected information about soccer games in 26 venues all over Germany using the kicker.de website[11] (see Figure 3.7). Information retrieved from those websites usually contains data about the event name, start time, location, and occasionally additional information in an unstructured form (e.g., descriptions of the event in text form).

The industrial event dataset was provided by one of the largest ticket retailer companies in Europe. It contains information for over 3000 venues within Germany for the years 2014 and 2015. In total, the dataset contains information for more than 430000 events of different categories. The data are stored in a relational database, shown in 3.8. For each *Venue*, we get a list of *Events*. Each *Event* holds information about the *StartTime*, the *Name* of the event, and a short *Description*. Each *Event* is linked to one *Artist* entity. Each *Artist* holds information about their *Name*. Each *Event* is also part of a *SubCategory* and a *Category*. Examples of a *Category* and *SubCategory* are **Music** and **Rock/Pop**, respectively. Whereas this dataset holds a very comprehensive list of events, it explicitly does

[11] http://www.kicker.de/

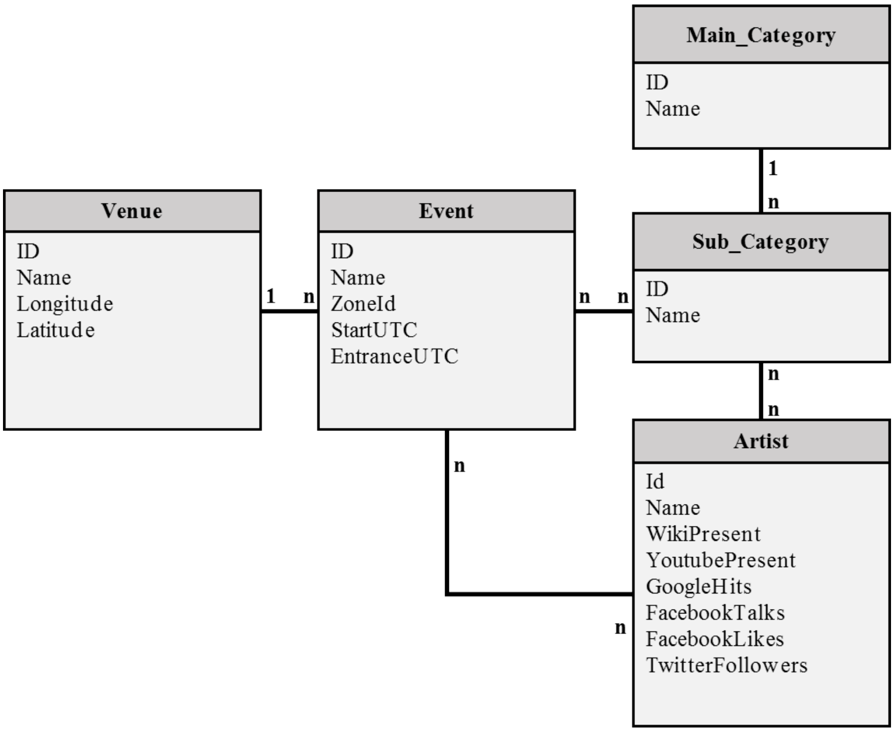


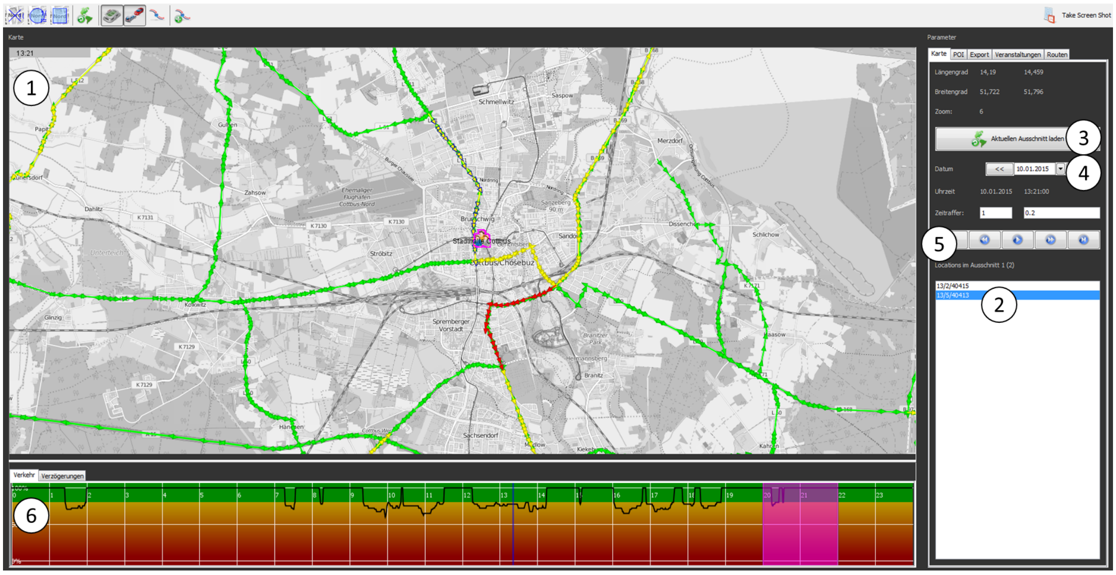
FIGURE 3.8: Data model of the event data from the ticket retail company.

not contain all events at a given venue, but only those for which the company sold tickets.

3.3 Data Visualization

For data visualization, we developed a viewer (hereafter called *TrafficViewer*) for our data collection. It is a tool based on Java[12] that allows us to get a quick first impression about the traffic situation around venues for events in our database. A screenshot of *TrafficViewer* is shown in Figure 3.9. It is based on a map (marker **1** in Figure 3.9) that shows the *TrafficLocations* on top of an OSM[13] raster image. The user can scroll to any location within Germany and load the specific *TrafficLocations* from the database using the "reload" button (marker **3** in Figure 3.9). The date can be selected from the control panel on the right side (marker **4** in Figure 3.9). Traffic information for one specific *TrafficLocation* is shown at the bottom graph (marker **6** in Figure 3.9) after the user selects the location within the map (the selected location is marked in blue). By clicking on the

[12] `http://java.com`
[13] `https://www.openstreetmap.org/`



Längengrad 14,19 14,459

Breitengrad 51,722 51,796

Zoom: 6

Datum 10.01.2015

Uhrzeit 10.01.2015 13:21:00

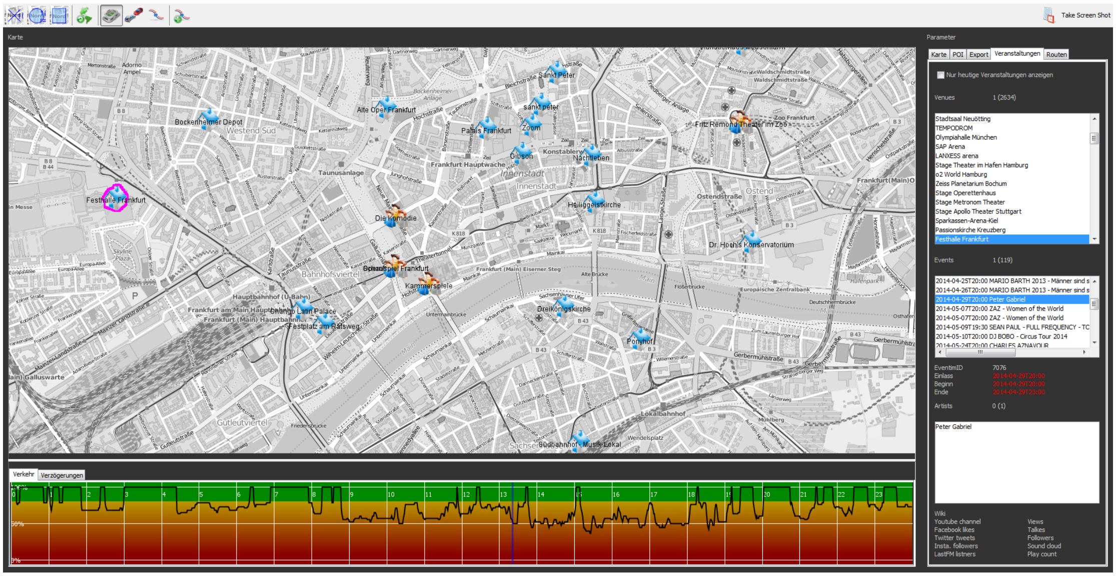
Zeitraffer: 1 0.2

Locations im Ausschnitt 1 (2)

13/2/40415
13/5/40413

FIGURE 3.9: Screenshot of the *TrafficViewer* showing traffic information.

play button in the control panel (marker **5** in Figure 3.9) the traffic situation for all loaded *TrafficLocations* for the entire day is replayed in the *TrafficViewer*. The tool uses a color schema for traffic situations from red (severe congestion) to yellow (medium congestion) to green (no congestion at all), which is also shown in the color graph at the bottom (marker **6** in figure 3.9). Events are shown on



☐ Nur heutige Veranstaltungen anzeigen

Venues 1 (2634)

Stadtsaal Neuötting
TEMPODROM
Olympiahalle München
SAP Arena
LANXESS arena
Stage Theater im Hafen Hamburg
o2 World Hamburg
Zeiss Planetarium Bochum
Stage Operettenhaus
Stage Metronom Theater
Stage Apollo Theater Stuttgart
Sparkassen-Arena-Kiel
Passionskirche Kreuzberg
Festhalle Frankfurt

Events 1 (119)

2014-04-25T20:00 MARIO BARTH 2013 - Männer sind s
2014-04-26T20:00 MARIO BARTH 2013 - Männer sind s
2014-04-29T20:00 Peter Gabriel
2014-05-07T20:00 ZAZ - Women of the World
2014-05-07T20:00 ZAZ - Women of the World
2014-05-09T19:30 SEAN PAUL - FULL FREQUENCY - TC
2014-05-10T20:00 DJ BOBO - Circus Tour 2014
2014-05-24T20:00 CHARLES AZNAVOUR

EventimID 7076
Einlass 2014-04-29T20:00
Beginn 2014-04-29T20:00
Ende 2014-04-29T23:00

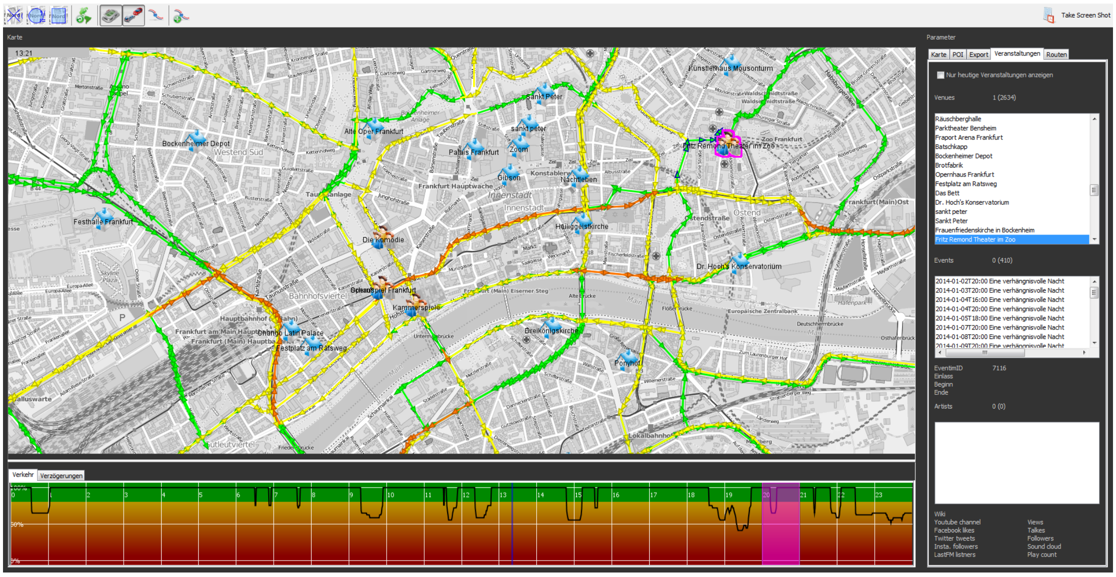
Artists 0 (1)

Peter Gabriel

Wiki
Youtube channel Views
Facebook likes Talkes
Twitter tweets Followers
Insta. followers Sound cloud
LastFM listners Play count

FIGURE 3.10: Screenshot of *TrafficViewer* showing event information.

the map depending on the venue's position (see Figure 3.10). The control panel on the right side shows all planned events for a selected venue, including the title of the event, the artist, and additional information that we collected from the web (used for the analysis in Chapter 6). The *TrafficViewer* shows the begin-



☐ Nur heutige Veranstaltungen anzeigen

Venues 1 (2634)

Räuschberghalle
Parktheater Bensheim
Fraport Arena Frankfurt
Batschkapp
Bockenheimer Depot
Brotfabrik
Opernhaus Frankfurt
Festplatz am Ratsweg
Das Bett
Dr. Hoch's Konservatorium
sankt peter
Sankt Peter
Frauenfriedenskirche in Bockenheim
Fritz Remond Theater im Zoo

Events 0 (410)

2014-01-02T20:00 Eine verhängnisvolle Nacht
2014-01-03T20:00 Eine verhängnisvolle Nacht
2014-01-04T16:00 Eine verhängnisvolle Nacht
2014-01-04T20:00 Eine verhängnisvolle Nacht
2014-01-05T18:00 Eine verhängnisvolle Nacht
2014-01-07T20:00 Eine verhängnisvolle Nacht
2014-01-08T20:00 Eine verhängnisvolle Nacht
2014-01-09T20:00 Eine verhängnisvolle Nacht

EventimID 7116
Einlass
Beginn
Ende

Artists 0 (0)

Wiki
Youtube channel
Facebook likes
Twitter tweets
Insta. followers
LastFM listners

Views
Talkes
Followers
Sound cloud
Play count

FIGURE 3.11: Screenshot of *TrafficViewer* showing event and traffic information.

ning of the event on the timeline, whenever a venue is selected where an event happens on the selected day (see Figure 3.11).

In the background, the *TrafficViewer* is connected to different databases to get access to the *Flow Data*, *Incident Data*, and *Event Data*. It enables us to get a good impression of the impact of specific PSEs on traffic, and is further used for illustration purposes within the remainder of this thesis. An example of the *TrafficViewer* is shown in Appendix A where the traffic situation before and after a soccer game is presented.

Chapter 4

Event Traffic Characteristics

So far, we have examined the state of the art of research activities within our domain and the data sources used in our studies have been outlined. However, what is event traffic and what does it depend on? How do events usually impact the traffic situation around venues? In this chapter, we will focus on these questions and introduce the overall characteristics of event-inflicted traffic situations.

Traffic during events is expected to show variations in its behavior. There are multiple reasons for this. Major roles are played by the *venue location* and the road infrastructure around it. Its capability to cope with additional traffic and to maintain a sufficient level of service (the concept introduced as *traffic resilience* in Chapter 2) is a crucial characteristic. As traffic in general varies over time, *daytime variation* will also play an important role in terms of the observed traffic behavior during events. An event on a Wednesday afternoon during rush hour might result in congestion caused by adding load to a road network that is already at its capacity limits, whereas an event at night might show less congestion, as the route capacity is sufficient. Another known phenomenon of traffic is its different behavior during weekdays and weekends due to commute traffic and rush hour (e.g., see [31, 95]). These *day of week variations* will also interfere with the additional traffic due to an event happening. Whereas a concert on a Sunday might result in less congestion as the route network is generally not under stress, an event on a Monday might have a different impact.

As mentioned earlier the authors of [62] have shown that there is a certain footprint of people attending different event types. These results allow the assumption of specific travel behavior of visitors based on the *event category*, which implies that it is also an influencing factor. Another factor is the *attractiveness* of an event. A concert of Rihanna, for example, is expected (by our intuition) to attract more people to the venue and have a higher impact than a concert of a lesser known band at the same location. Also intuitive is the relevance of the *transport modality choice* of visitors. Whereas a classical concert might attract visitors that prefer to arrive by car, a concert of a teen band probably puts more load on the public transportation network than on the road infrastructure.

The discussion above shows our expectation about the impact factors that are purely based on our intuition. In this chapter, we put our expectations to the test. We first present results of a real-world study that we conducted in Cologne, Germany. It shows the general characteristics of event-inflicted traffic and motivates the following chapters of this thesis. Next, we present a large-scale study that we conducted in a broader geographic region using high-resolution traffic datasets. This *Flow Study* focuses on all the influencing factors that we mentioned earlier, which can be expressed directly: *venue location*, *daytime variation*, *day of week variation*, and *event category*. The other two influencing factors *attractiveness* and *modality choice* are latent measures which cannot be derived directly from our data sources. However, they might be described indirectly by other attributes and we will further analyze them in Chapter 6.

Before presenting the studies and their results, we first introduce a set of required metrics and prerequisites.

4.1 Study Prerequisites

To capture event impact on traffic, spatial and temporal criteria need to be defined. The spatial criteria describes the *impact region* of events depending on the venue locations. The temporal criterion defines the *event time window* during which event traffic is expected to happen before it starts.

For each observed venue, the *impact region* of events might be different and highly dependent on the route network topology (including traffic lights, etc.).

Because the impact region is not known to us, we can only approximate it (the challenge of finding the exact impact region will be further discussed in Chapter 5).

Whereas event contributions are presumably more dominant in the immediate vicinity of the venue, there is a high probability to miss parts of the event traffic when focusing only on a small area. At the same time, capturing traffic in a larger area probably lowers the contribution of the event traffic in the observations and the data get noisier. In the following studies, we focus on different approximations of the impact region depending on the focus of the study. Whereas the *General Event Characteristic* study focuses on a small radius around the venue (500 m) to capture mostly event-related traffic congestion, the *large-scale* study uses different regions (500, 1000, and 2000 m) to analyze their impact on the observed traffic.

As with the impact region, there is no common method to detect the event time window and it can only be approximated. Whereas the *General Event Characteristic* does not require this metric (as we only compare entire days), it is highly relevant for the *large-scale* study. For that, based on our intuition and observations throughout the datasets, we use a window of 2 h before the official event start.

The individual impact of events on traffic cannot be observed directly (i.e., they are latent) but we do know that they contribute to the total observed congestion behavior. A descriptive measure for congestion behavior is *delay time*. A delay is defined as additional time that is required to pass a certain road when it is congested compared to noncongested situations. As we focus on impact areas rather than on specific road segments, we define our measure *sumdelay* as the sum of delay time for all road segments in the observed area for the defined *event time window*.

Formally, *delay d* (in s) for *TrafficLocation (TL)* at time *t* is defined as:

$$d_{TL}(t) = TT(CS(t))_{TL} - TT(FFS(t))_{TL} \tag{4.1}$$

where *TT* is the travel time, *CS* is the current speed at time *t*, and *FFS* is the *FreeFlowSpeed* at the specific TL for time *t*. TT is defined as

$$TT(s)_{TL} = \frac{TTL}{s} * 3600 \tag{4.2}$$

where *s* is the speed at the specific TL at time *t* (in km/h) and TTL is the length of the *TL*.

As the *incident data* used for the *General Event Characteristic* study already contain *delay* information, this attribute is used directly. For the *large-scale* study, the delay time is calculated based on our dataset. Based on the delay information, we calculate our final metric *sumdelay* as:

$$sumdelay = \sum_{1}^{K} d_{TL_k} \tag{4.3}$$

where *d* is the delay information for one specific *TL* at time interval *k*.

We will present the studies and their specific results in the next sections.

4.2 General Event Characteristic

To get a first impression of event-caused traffic, we conduct an initial study in Cologne, Germany. This study explicitly does not have the goal to analyze as many influencing factors as possible, but serves as a starting point to understand basic event traffic characteristics. It is based on the incident dataset described in Chapter 3 that we collected for seven months between June and December 2013. During that timespan (and after cleaning our dataset) we observed traffic for 29 events hosted in the LANXESS Arena in Cologne, Germany.

For these events we collected information about the observed delay time (as received from the data provider) within a fixed region around the venue (500 m radius), as described in the prerequisites. The selection of this radius was based on our observations during the study and allowed us to capture most of the (apparently) event-caused traffic disruptions.

The resulting delay time is summed up for the entire region and days with scheduled events (*event days*) and days without scheduled events (*non-event days*) are separated. Data from *non-event* days are used to build a model for each day of the week as the average delay time throughout all observations. An example is shown in Figure 4.1, where the model for Tuesdays is presented. The graph shows a typical rush-hour phenomenon with a significant traffic disruption in the morning and in the afternoon. As the data used to build the

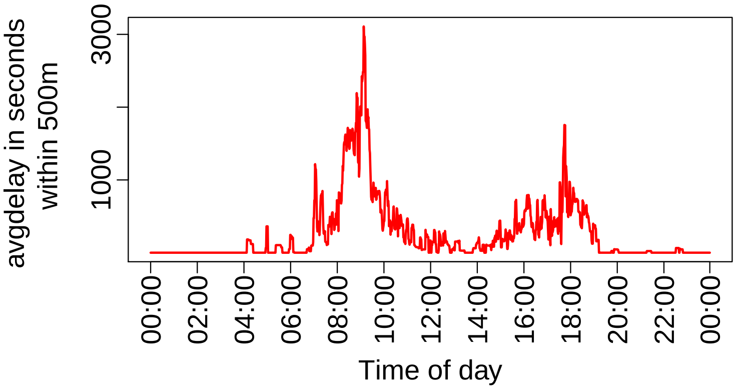


FIGURE 4.1: Average summed delay time around the LANXESS Arena in Cologne, Germany for Tuesdays without scheduled events. avgdelay: average of the summed up delay time within the entire radius, for all considered days. Radius: 500 m. Source:[93, 94]

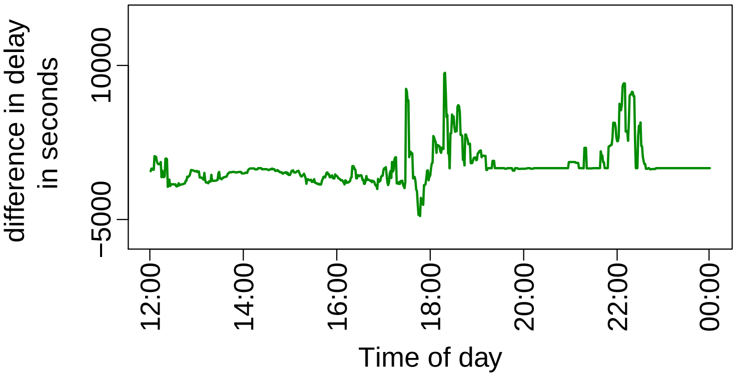


FIGURE 4.2: Additional delay due to the concert of Mark Knopfler around the LANXESS Arena in Cologne, Germany. The delay on top of the historic trend line for Tuesdays on Tuesday the 2nd of July 2013. Radius: 500 m. Source:[93, 94]

trend line only cover days without events, this graph can be seen as the *routine* (the "regular" traffic behavior, as described in Chapter 2.1.1) for that particular radius on that day of the week.

We use this generated trend line to illustrate the traffic disruptions during event days by analyzing the discrepancy of observed traffic to the routine. An example is shown in Figure 4.2. This shows the traffic situation on July 2nd 2013, when *Mark Knopfler*[1] played a concert at the LANXESS Arena. The graph shows the difference in the observed delay time on July 2nd 2013 from the created trend line for Tuesdays (shown in Figure 4.1). A delay above zero means that we observed more traffic than "usual" at that time of the day and below zero means that we observe less traffic than on "regular" Tuesdays at that time of

[1] `http://www.markknopfler.com/`

day. With this interpretation in mind, we observe two major additional traffic congestions: between 18:00 and 20:00, and between 22:00 and 23:00. The concert was scheduled for 20:00. These data lead to the assumption that the observed two peaks are due to people attending the event. The first wave (18:00–20:00) might be caused by visitors going to the venue, whereas the other increase in traffic (22:00–23:00) is probably caused by people leaving the venue after the concert ended. This phenomenon of two subsequent waves of traffic has also been reported in [96] and describes the *incoming* and *outgoing* traffic around a venue.

Traffic during the *Mark Knopfler* concert almost exactly follows the assumption of two waves of traffic. However, further examples from the initial study, as presented in Figure 4.3, show that this is not always the case. Whereas the concert of *Rihanna*[2] on June 26th 2013 showed a similar pattern as seen for the previous concert before, traffic during the *RUSH*[3] concert on June 06th 2013 showed essentially no increase in congestion at all. Although both events were concerts and started around a similar time in the evening, the traffic behavior seems quite different. As a nonconcert example, we see the results of a comedy show of Mario Bart[4] (a German comedian). The traffic on that day again shows no significant increase in congestion.

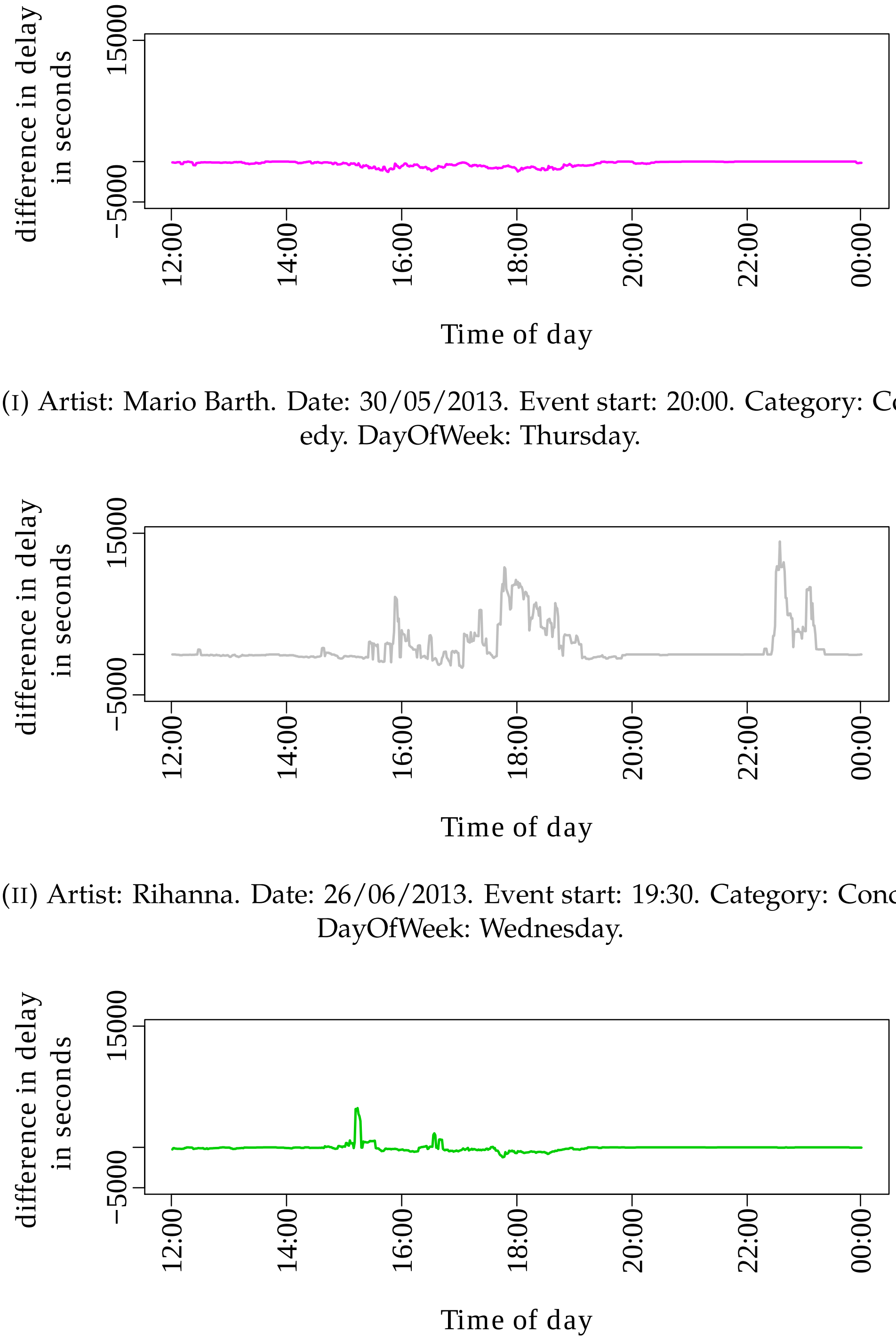
Another interesting fact from this initial experiment is the varying end times of events. Although the start time is usually (at least roughly) known in advance, the end times vary. In our examples, *Rihanna* apparently played until approximately 23:00, whereas *Mark Knopfler* finished around 22:00. Not knowing the end of an event adds more noise to the analysis. In the rest of this thesis, we will therefore mainly focus on the start times of events. Ideas of how to detect the end of an event before it becomes visible in the traffic data are also discussed at the end of this thesis, in Chapter 7.

So far, we have seen different events with different impacts on the infrastructure. Although some of them were of the same category their traffic impacts still varied. They also happened on different days of the week, which also might have an impact on the resulting traffic observation. To get a more concrete idea

[2] `http://www.rihannanow.com/`

[3] `https://www.rush.com/`

[4] `https://www.mario-barth.de/`



(I) Artist: Mario Barth. Date: 30/05/2013. Event start: 20:00. Category: Comedy. DayOfWeek: Thursday.

(II) Artist: Rihanna. Date: 26/06/2013. Event start: 19:30. Category: Concert. DayOfWeek: Wednesday.

(III) Artist: RUSH. Date: 04/06/2013. Event start: 20:00. Category: Concert. DayOfWeek: Tuesday.

FIGURE 4.3: Additional delay on top of the average trend line for that specific day of the week around the LANXESS Arena in Cologne/ Germany. Radius: 500m.

of the impact of different influencing factors, we consider a broader study, the *large-scale study*.

4.3 Large-Scale Study

This study is conducted in different situations around three different *venue locations* in Germany: 1. Mercedes-Benz Arena (Berlin), 2. LANXESS Arena (Cologne), 3. Alte Oper (Frankfurt). Whereas the first two venues usually host all kinds of events, from hockey games to large concerts, the third is mostly focused on classical and jazz events. All of the venues are in inner city environments, but their surrounding traffic networks differ.

We selected a set of attributes that describe our influencing factors to analyze. For *daytime variation*, we defined a manual threshold to separate between daytime and nighttime. In our study, we followed the suggestion from the common literature (e.g., [97–99]) and set this threshold to 19:00. For the *day of week variations*, we followed the classification of days into four categories: Monday till Thursday (Mo–Th), Friday (Fr), Saturday (Sa), and Sunday (Su). This classification is also commonly used in the literature (e.g., [12]). For the *event category*, we need to assign consistent categories throughout our dataset. A categorization can be performed using different metrics of the event (e.g., artist) or, in larger event scenarios (e.g., festivals), can be derived from the overall theme of the event. In our study, we used the labels from the *Main Category* attribute from our industrial event dataset (described in more detail in Chapter 3).

Of course, there are numerous other factors that can affect traffic behavior, which are not directly tied to PSEs (e.g., holiday season, weather phenomena, as already discussed in Chapter 1). We cannot filter out these phenomena from our dataset completely, but we can minimize their impact to a certain extent by using a large time horizon for our study. For each venue, we used the industrial event dataset and selected all events over a time period of 15 months, between 02/2014 and 05/2015. Each event contains information about: the artist, category, date, and start time (for further information about the dataset, please refer to Chapter 3.2). From these attributes, we defined our metrics as described above.

As mentioned in 4.1, we focus on three different impact regions around the venues to approximate event traffic contribution: small, medium, and large, defined as circular areas of radius 500, 1000, and 2000 m.

To calculate the delay times, we took the average *FreeFlowPercentage* (as introduced in Chapter 3) over a time window of five min within the 2 h time window before the event start to retrieve the speed information for this time period for each TL. Finally, we summed up all delays from all TLs within the given radius to derive our final traffic metric *sumdelay* (in s), as defined in 4.3.

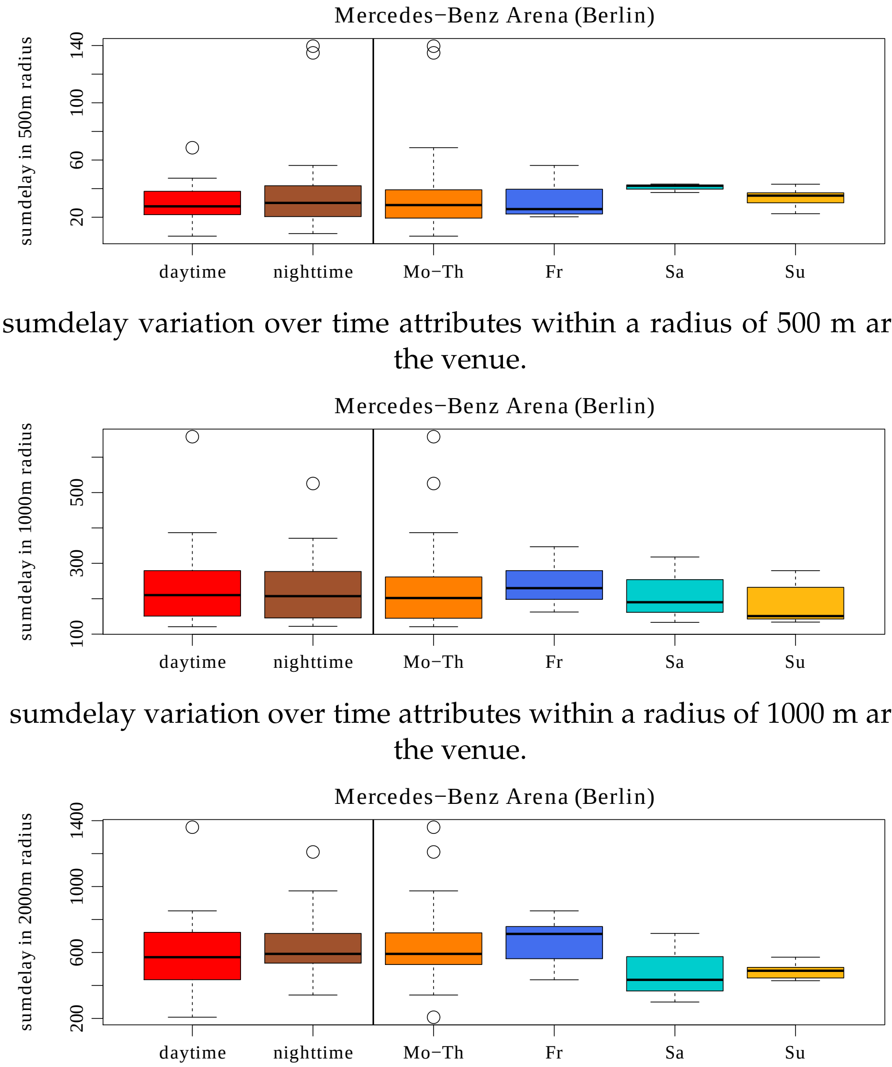
4.4 Results

For each venue, we created an overview of the traffic behavior grouped by the selected metrics: *daytime variation*, *day of week variation*, and *event category*.

4.4.1 Mercedes-Benz Arena Berlin

The Mercedes-Benz Arena is located in the heart of city east in Berlin, Germany. In total, we observed 102 events at this venue between 01/2014 and 05/2015. Of these 102 events, we collected data for 42 events starting during daytime and 60 events starting during nighttime (after 19:00). For the *day of week variation*, we observed 64 events happening between Monday and Thursday (*Mo–Th* class), 22 events on Fridays (*Fr* class), six events on Saturdays (*Sa* class), and 10 events happening on Sundays (*Su* class). The resulting daytime variations can be seen in Figure 4.4, where the delay around the venue in all three selected *impact regions* is shown.

In a 500 m radius, the results show that, although daytime and nighttime have a similar median delay value, traffic during nighttime seems to undergo higher fluctuations than during the day. Interestingly, this observation changes for the other two radii. Traffic variation in a 1000 m radius looks rater similar between daytime and nighttime. For the 2000 m radius, traffic during daytime seems to fluctuate more than during nighttime. A possible explanation could lie in the already discussed strong variation of event impact on traffic. This variation would appear more pronounced in a rather small radius (e.g., 500 m), because



(I) sumdelay variation over time attributes within a radius of 500 m around the venue.

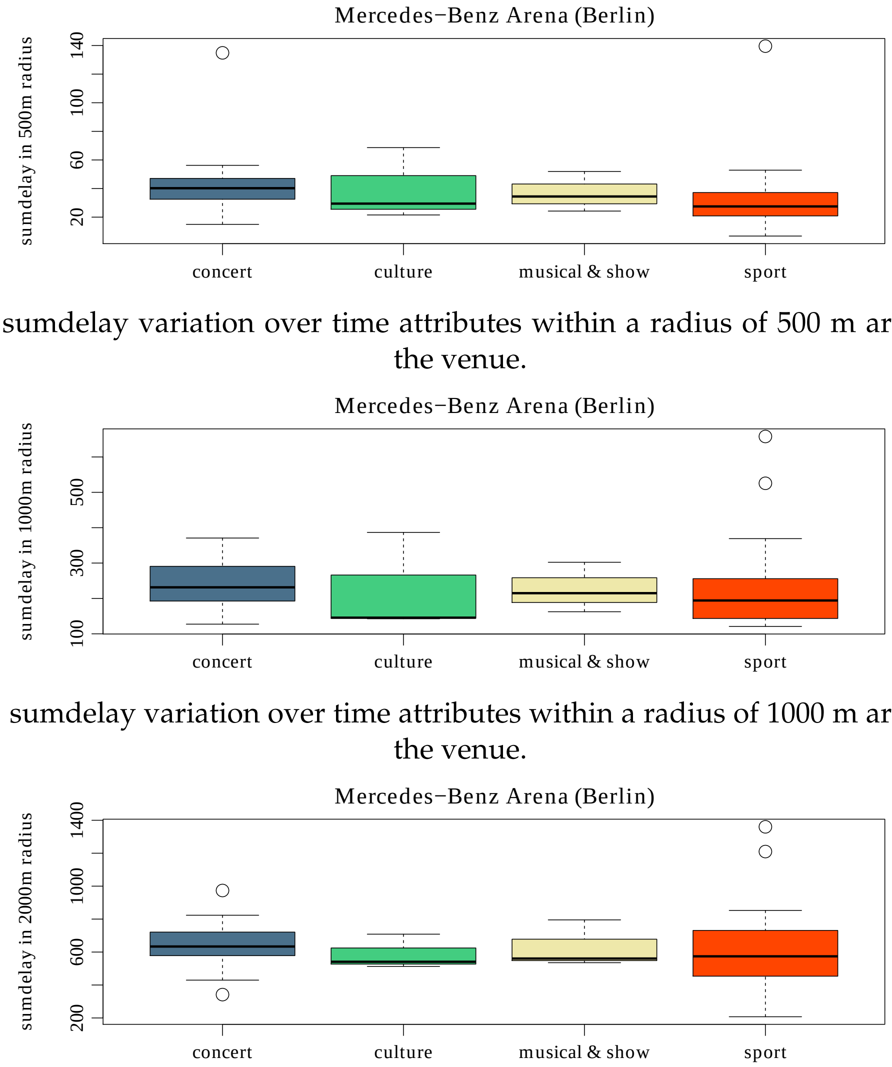
(II) sumdelay variation over time attributes within a radius of 1000 m around the venue.

(III) sumdelay variation over time attributes within a radius of 2000 m around the venue.

FIGURE 4.4: *Daytime variation* and *day of week* variation during events at the Mercedes-Benz Arena in Berlin, Germany.

the small impact area would bring the actual effect of events more into focus. The effect would also gain higher visibility during nighttime, because there is usually less traffic in general and again, event-specific congestion would be emphasized. With this idea, the observations for the 1000 m and 2000 m radii could be explained by too much noise resulting from the broader range of selected TLs and non-event specific traffic phenomena.

In general, it has to be considered that in a dense city such as Berlin, a large radius (e.g., 2000 m) captures a huge number of TLs. Many of these TLs are rather far away from the venue and probably not influenced by events at all.



(I) sumdelay variation over time attributes within a radius of 500 m around the venue.

(II) sumdelay variation over time attributes within a radius of 1000 m around the venue.

(III) sumdelay variation over time attributes within a radius of 2000 m around the venue.

FIGURE 4.5: *Event category variation* during events at the Mercedes-Benz Arena in Berlin, Germany.

This leads to a high probability that the observed traffic variation is caused by other influencing factors and not connected to the observed events.

For the weekday patterns, we see the highest fluctuation in the Mo–Th class, which might be a result of daily rush hours and a generally higher occupancy on the roads. Again, this changes for the 2000 m radius, but we can probably conclude that there are too many disturbances in a 2000 m radius to assume a change in event behavior. A clear trend is the lower delay on weekends, which is probably attributable to less traffic in general. It is interesting to note the variation in Figure 4.4ii for the *Sa* and *Su* classes. The variation increases drastically in the 1000 m radius. One possible reason could be that the 500 m radius does not capture all event-related traffic (at least on weekends) and the 2000 m radius hides the effect owing to too many other influences.

At this venue, we observed four different categories of events: *concert, culture, musical & show, and sport*. For these categories we observed 22 concerts, six culture events, six events belonging to musical & show, and 64 sport events. Four events in our database were not assigned to any category, and we filtered them out. The traffic behavior grouped by category is shown in figure 4.5.

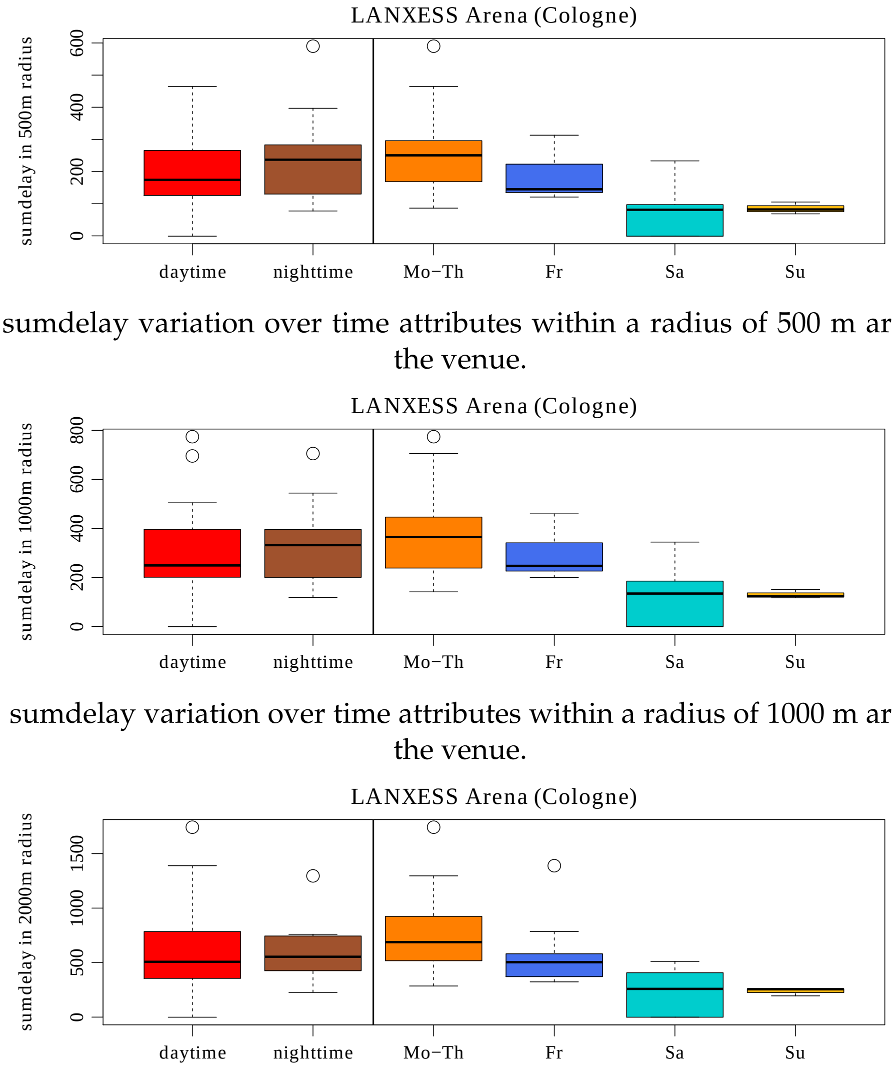
In general, the data show a rather stable pattern across different categories. Categories *culture* and *sport* show slightly higher variations than *concert* and *musical & show*, but only to a certain extent. It is interesting to note the lower median delay at the 1000 m radius for the *culture* category, which indicates a higher fluctuation of event-caused traffic disruptions for that particular category.

4.4.2 LANXESS Arena Cologne

The LANXESS Arena is a multi-event venue in the city center of Cologne, Germany. In total, we observed 101 events at this venue between 02/2014 and 05/2015. Of these 101 events, we collected data for 59 events starting during daytime and 42 events starting during nighttime (after 19:00). For the *day of week variation*, we observed 59 events happening between Monday and Thursday (*Mo–Th* class), 24 events on Fridays (*Fr* class), 10 events on Saturdays (*Sa* class), and eight events happening on Sundays (*Su* class). The time variation for different events is shown in figure 4.6.

For the LANXESS Arena, traffic delay observations vary more for daytime and nighttime than for the first venue, particularly for the 500 and 1000 m radii. A clear trend between weekday/weekend classes is observable. The delay times are in general higher than for the first venue.

At this venue, we observed four different categories of events: *concert, misc, musical & show, and sport*. We also observed two *cultural* events but because two events are insufficient for a representative statement about the traffic for this event category, we filtered them out.



(I) sumdelay variation over time attributes within a radius of 500 m around the venue.

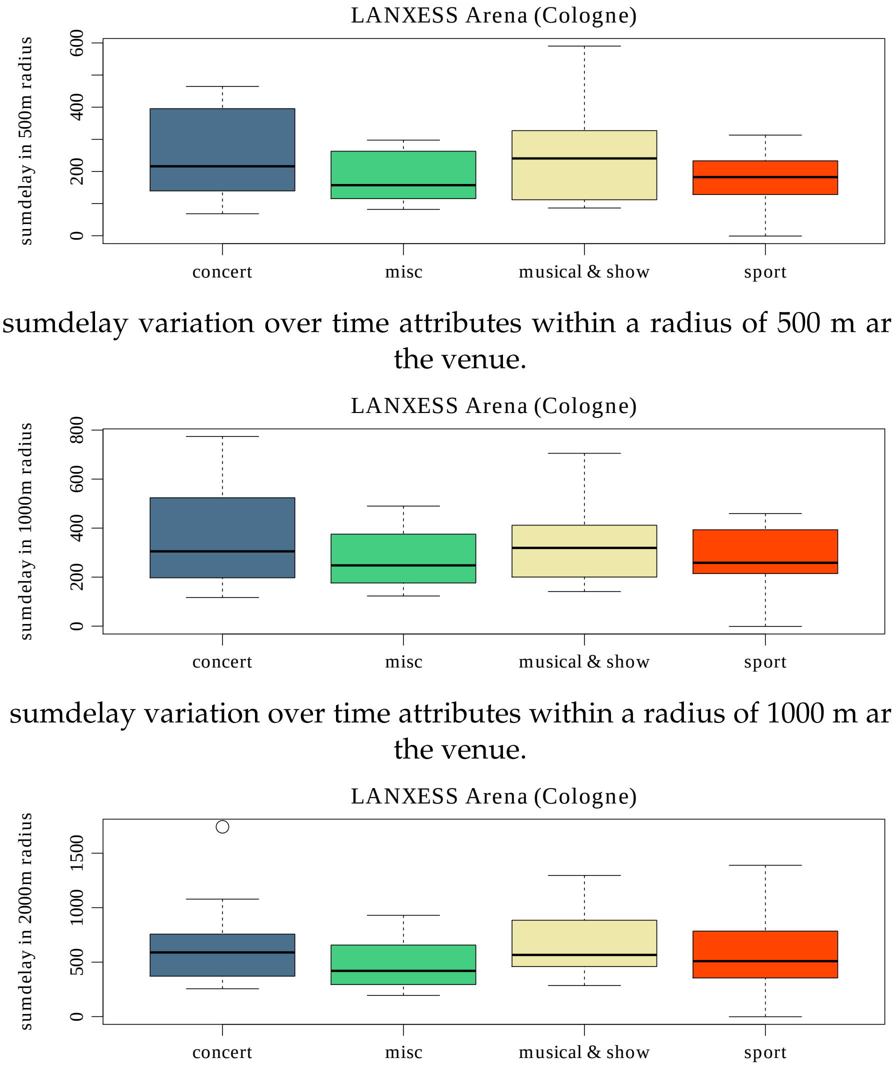
(II) sumdelay variation over time attributes within a radius of 1000 m around the venue.

(III) sumdelay variation over time attributes within a radius of 2000 m around the venue.

FIGURE 4.6: *Daytime variation* and *day of week* variation during events at the LANXESS Arena in Cologne, Germany.

For the remaining categories, we observed 25 concerts, nine events belonging to musical & show, and 50 sport events. The *misc* category holds all events that do not match into one of the others. In this case, events of this category included 11 carnival events ("Lachende Kölnarena") and one gaming convention.

The categories show a different behavior than for the first venue in Berlin (see Figure 4.7). For the LANXESS Arena, the *concert* category shows by far the most variation for the 500 and 1000 m radii, and their impact range differs significantly from the other categories. A possible explanation could lie in the mixture of different concerts of different popularity. The *misc* category is also



(I) sumdelay variation over time attributes within a radius of 500 m around the venue.

(II) sumdelay variation over time attributes within a radius of 1000 m around the venue.

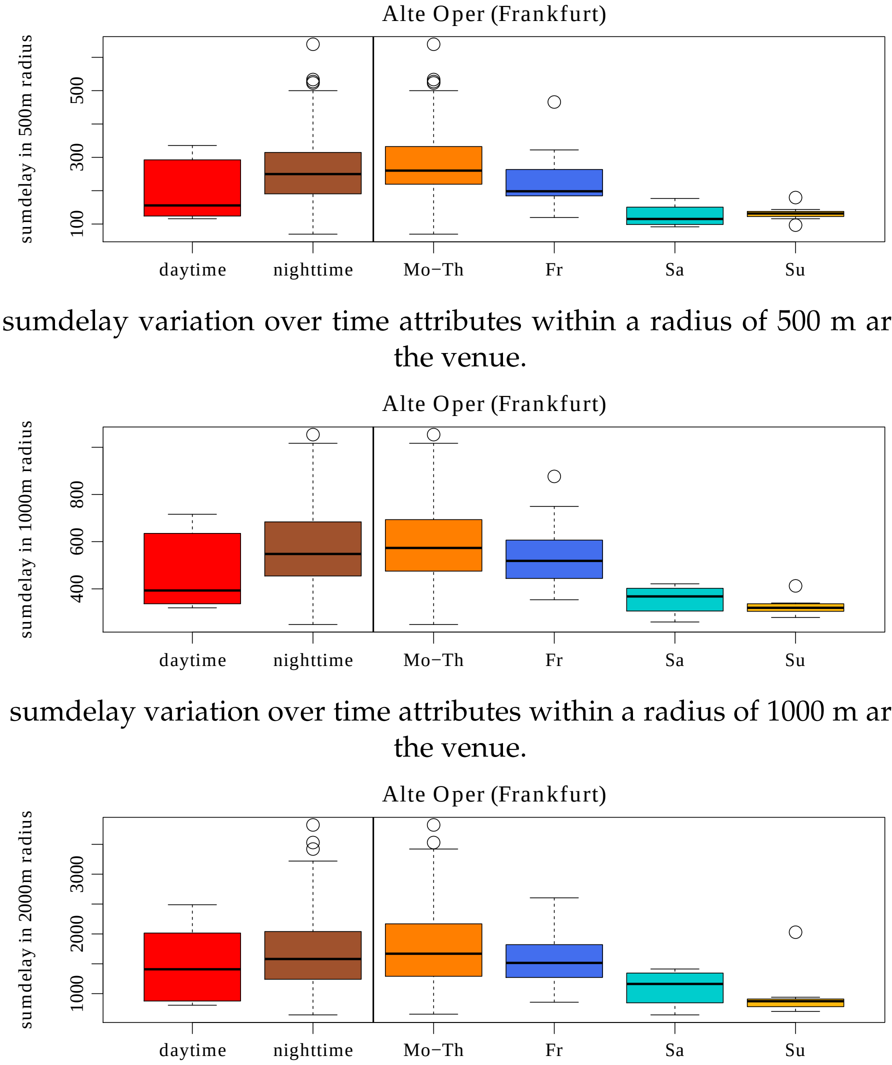
(III) sumdelay variation over time attributes within a radius of 2000 m around the venue.

FIGURE 4.7: *Event category variation* during events at the LANXESS Arena in Cologne, Germany.

interesting. In our example, all events in this category are essentially the same event, but happening at different times. Although one would expect a more stable behavior because the target audience is exactly the same for all events, it shows rather large fluctuations. However, the carnival season is a very special time, especially in Cologne. Many people celebrate on the streets and there are organized and non-organized street parades and parties. With that in mind, results for that particular category have to be treated with caution.

4.4.3 Alte Oper Frankfurt

The *Alte Oper Frankfurt* is an opera hall in the inner city of Frankfurt (Main), Germany. In contrast to the other two venues this one is focused more on classical events and jazz concerts. In total, we observed 257 events at this venue between 02/2014 and 05/2015. Of these 257 events we collected data for 19 events starting during daytime and 238 events starting during nighttime (after 19:00). For the *day of week variation*, we observed 166 events happening between Monday and Thursday (*Mo–Th* class), 69 events on Fridays (*Fr* class), eight events on Saturdays (*Sa* class), and 14 events happening on Sundays (*Su* class). As shown



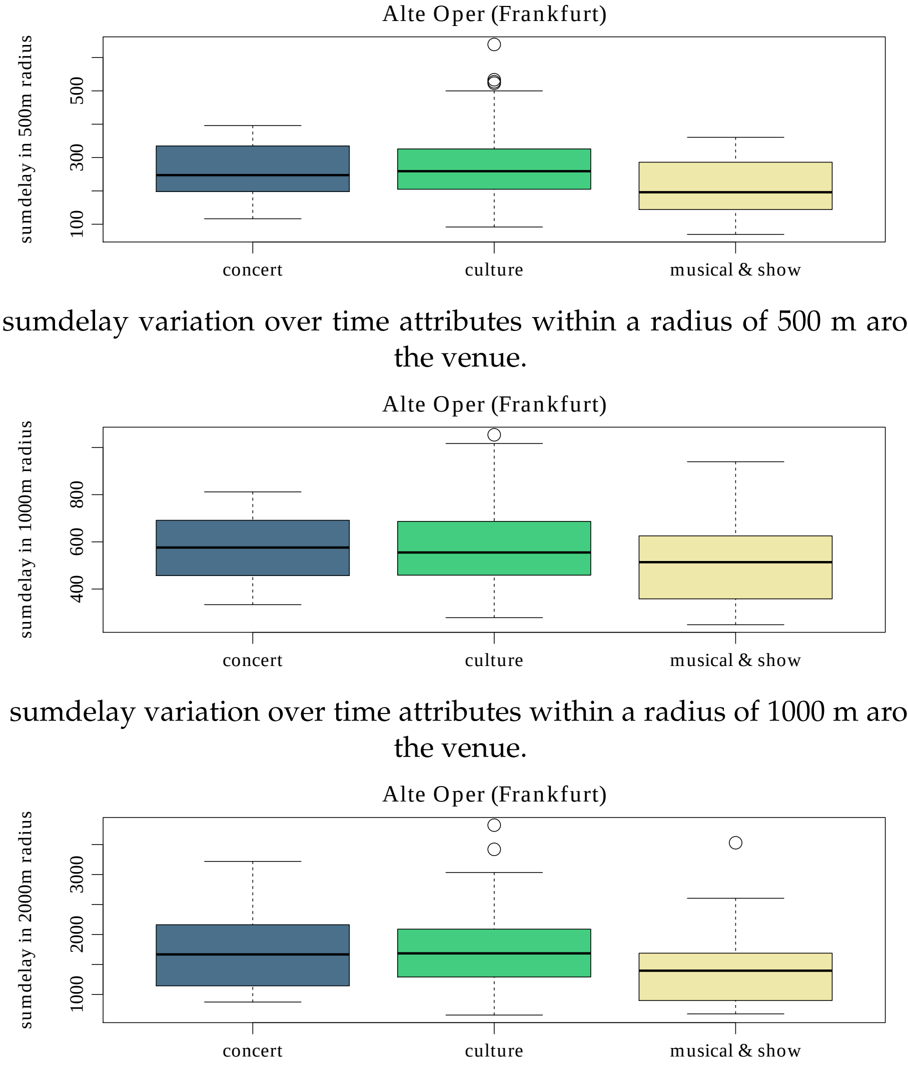
(I) sumdelay variation over time attributes within a radius of 500 m around the venue.

(II) sumdelay variation over time attributes within a radius of 1000 m around the venue.

(III) *Daytime variation* and *day of week* variation during events at the Alte Oper in Frankfurt, Germany.

FIGURE 4.8: *Daytime variation* and *day of week* variation during events at the Alte Oper in Frankfurt, Germany.

in Figure 4.8 this venue shows a higher variation in traffic during nighttime than daytime. Traffic around the venue shows distinct weekday/weekend traffic patterns, where traffic disruptions between *Mo–Th* vary the most, followed by *Fr*, *Sa*, and finally *Su*, where almost no traffic disruptions are observed.



(I) sumdelay variation over time attributes within a radius of 500 m around the venue.

(II) sumdelay variation over time attributes within a radius of 1000 m around the venue.

(III) sumdelay variation over time attributes within a radius of 2000 m around the venue.

FIGURE 4.9: *Event category variation* during events at the Alte Oper in Frankfurt, Germany.

Although the number of events at this venue is by far the highest, we observed events of only three different categories: *concert, culture, and musical & show*. For these categories we observed 52 concerts, 143 culture events, and 58 events belonging to the category musical& show. Four events were filtered from our dataset because they did not have any category assigned to them.

Despite the high variation in daytime and nighttime delay, the Alte Oper shows by far the most stable behavior for the event categories throughout all different radii, as shown in Figure 4.9.

Whereas *concert* and *culture* events show very similar behaviors, *musical & show* shows nearly the same variations in traffic but less delay than the other two categories. A possible explanation lies in the different event focus of the venue. The target audience of classical events and jazz concerts probably differ from the group of visitors that attend the events at the other two venues. The specific target group could be a reason for the observed stable behavior in traffic impact. Of course, the road infrastructure and the availability of public transportation also plays a major role.

4.5 Discussion and Conclusion

This chapter has shown that many of the discussed intuitions about event traffic are valid. The *venue location* is obviously of major importance, as we have seen huge differences in the observed delay times among the venues. Additionally, our intuitions regarding the *daytime variations* seem to be correct, because for all venues, we have seen very different behaviors between traffic on weekdays and weekends. An interesting observation is the difference between the *day of week variations* for the venues. Whereas around the Mercedes-Benz Arena in Berlin, traffic on Fridays shows the highest delay time, traffic around the other two venues show maximum median delay for the Mo–Th class. For all venues, we have seen a major importance of the *category* of events. However, again, the same categories of events act differently at different venues.

The presented results lead to the assumption that a generalization of event impact is probably not feasible (beyond a certain accuracy). A local discussion for each specific venue is recommended and in the following chapters, we will always focus on individual locations separately.

Another result is the impact of the selected geographic region. Many plots show a significantly different behavior for the 2000 m radius than for the other radii. For a detailed discussion of event traffic, a method is required to detect the precise spatial impact zone around venues, instead of the presented radius-based approach. Our approaches in that direction will be further discussed in the next chapter.

Chapter 5

Spatial Impact of Events

The previous chapter mainly focused on fixed *impact regions* to approximate the event traffic around venues. Although this approach allows a general discussion about traffic behavior, it does not accurately capture the reality of event traffic. As discussed earlier in this thesis, we expect event traffic on specific road segments rather than an entire region around a venue. These road segments will be part of routes that people prefer to use to get to the venue before the event starts and routes that people tend to use after the event ends. Road segments that are part of these routes are called:

1. *Inbound segments*: composed of the roads that drivers tend to use to get to the venue, and
2. *Outbound segments*: composed of the roads that drivers tend to use after the PSE is over, in order to leave the venue.

This chapter describes our approaches, experiments, and results of finding the specific impact region around venues in an automatic manner. We present an approach that works similarly for *Inbound* and *Outbound* segments. Owing to a lack of information about the end of PSEs (as previously discussed in Chapter 4), we focus in this chapter on experiments and results for the *Inbound segments* only.

The chapter is structured as follows: First, we present and discuss our general approach. After that, we explain the experimental setup and show results of a large-scale analysis based on traffic *Flow Data* (hereafter called *flow study*). The next part of this chapter focuses on a similar study based on the less detailed *incident data* set (hereafter called *incident study*). The chapter closes with a general discussion of the presented results, conclusions, and next steps.

5.1 Proposed Approach

For finding the spatial impact region of PSEs, the most straightforward approach is to observe the traffic behavior and identify road segments that are affected during the times of the PSEs. However, this approach has two main flaws: First, road segments that represent bottlenecks in the infrastructure and are frequently congested might be wrongly classified as affected by PSEs. Second, this approach does not give any information about the spatial region to observe, which would still force us to choose a region manually.

Our approach is based on the assumption that road segments that are affected by the presence of PSEs show a different behavior on event days than on non-event days.

With that definition in mind, the problem of finding the spatial impact region around a venue is reduced to finding those road segments that show a different behavior during PSEs. With this goal, our approach follows the assumption that, given the traffic state for all road segments around a venue, a classifier should exist that can successfully classify the road segments into positive (those that show a PSE-specific behavior) and negative (those that do not show any different behavior) classes. The problem therefore becomes a binary classification task.

The discussed method avoids both problems discussed earlier for the straightforward approach. We intentionally exclude road segments that are "always" congested. We also avoid the manual selection of the spatial region. By observing all road segments in a (too) big radius (one that makes sure we capture all road segments that are affected) the classifier returns those that are actually affected by PSEs.

However, the approach also comes with a great challenge. As it is based on the comparison between behaviors on event and non-event days for road segments, the time plays a major role. In our scenarios, the time can be described in terms of four variables: *daytime, day of week, month of year*, and *event time*.

Daytime and day of week describe the normal fluctuation of traffic over time and we will refer to these two variables as the *traffic time*. As discussed earlier, traffic changes during the day (e.g., owing to rush hour) and shows different patterns during weekdays and weekends. Seasonal effects are also known, whereby traffic behaves differently according to the *month of year*. However, our datasets did not allow us to observe any seasonal affects, because it was too small for such observations and we excluded them from the research focus. For the same reasons, we also did not include other potential influencing factors, such as holiday seasons.

Event time window describes the timespans before and after an event during which traffic behavior is changed owing to the event happening. The exact moments of these time windows depend on a large variety of different influencing factors (e.g., popularity of an event and weather conditions). In this work, we use a simplification of these time windows by manually defining them based on our observations. A discussion of possible strategies to find the *event time windows* from other information sources is presented in Chapter 7.

To compare traffic on event and non-event days, we need a strategy to deal with the mentioned time variables. In theory, one must eliminate as many varying time attributes as possible until finally, the observed traffic is solely influenced by the specific event. To put that into practice, we developed different strategies and concepts called the *Timespans of Interest*.

5.1.1 Timespans of Interest

The concept of the *Timespans of Interest (TOI)* results from a simple question: How do we compare traffic from a Monday afternoon 15:00–17:00 to a Saturday 18:00–20:00? For that question, we implemented two different concepts: the *Absolute Timespan of Interest* and the *Relative Timespan of Interest*.

5.1.1.1 Absolute Time Span of Interest

For the example question above, the *Absolute Time Span of Interest* returns the simple solution that we do not compare those timespans at all. Instead, the approach focuses on finding events that happen at comparable times rather than expending effort to normalize traffic situations.

For each *day of week* we define one particular timespan (our *time span of interest*), during which we have the most information about events in our dataset. Those TOIs are defined separately for the incoming and outgoing traffic.

For identifying the TOI, we look for the longest consecutive timespan that holds the most information about events. As an example, let us assume that we observed three different events on the same *day of week*. The entrance times might be 12:00–14:00, 12:30–14:30, and 14:00–16:00. The TOI for that venue would be selected to be between 12:30 and 14:30, because that time interval holds the most information about event traffic (this example was partly taken from [2]).

We implemented this approach by splitting up each *day of week* into minute-long time steps (e.g., Monday:14.42, Monday:14:43, ...). For each time step, we count the number of events happening in our dataset and choose the longest consecutive timespan as our TOI. The method is further visualized in Figure 5.1. For our research, we defined the TOIs for each venue for each day of week separately. We defined a minimum TOI of 30 min. This restriction was made to prevent excessively short TOIs, which would result in insufficient traffic samples to compare.

The Absolute Timespan of Interest focuses on a single time span for each weekday and thereby minimizes the *TrafficTime* fluctuations as much as possible. This makes it easy to compare traffic for this particular timespan on event and non-event days. However, this approach has the disadvantage that it forces us to limit the number of events to those that actually occur within the defined timespan. In our analysis, this drastically limited the number of observed events. A different problem lies in the fact that this approach is highly focused on the *traffic time* variability by ignoring the *event time window* to a certain extent. For example, assume we define our TOI for Saturday as 13:30–14:30. For event days, we would always capture traffic between 13:30 and 14:30, disregarding the fact that the event on one day would start at 14:30 and the other event

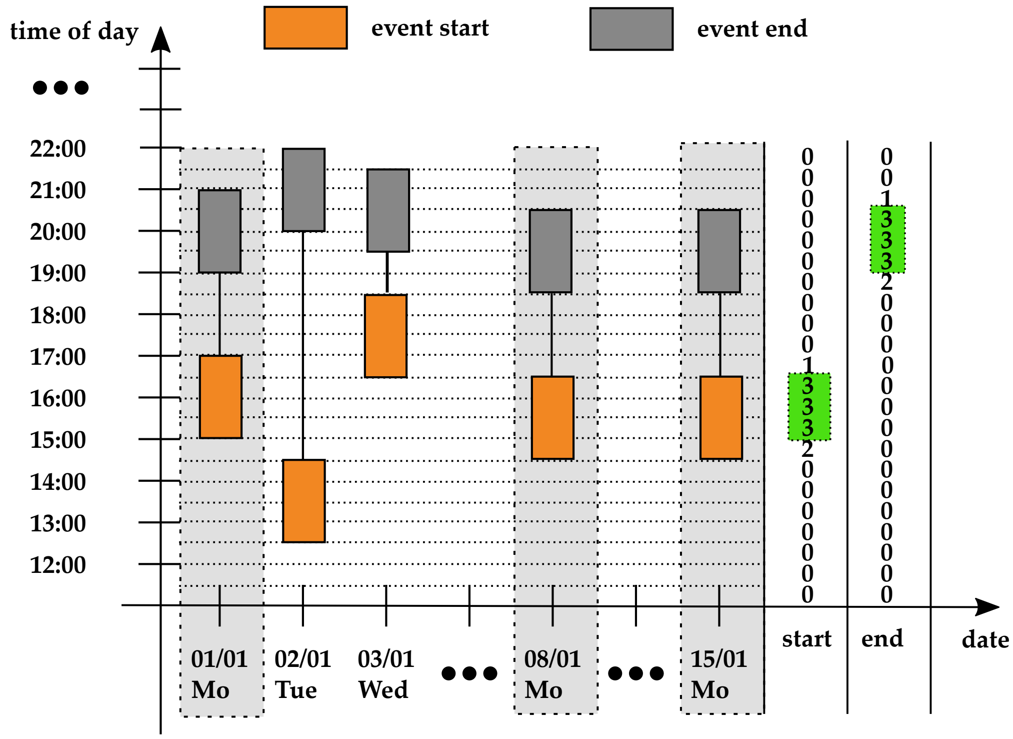


FIGURE 5.1: Example for calculating the *Absolute Timespan of Interest* for different events on the same day of week (Monday in the example). The same days of the week are marked in grey and the resulting TOIs for the incoming and outgoing traffic are marked in green.

might start at 14:00. This leads to variations in traffic behavior, with which the approach must cope.

5.1.1.2 Relative Timespan of Interest

The Relative Timespan of Interest is based on the idea that traffic flow follows a regular pattern that can be modeled for certain day of week categories. By creating a model for these categories, we assume that we capture the daily traffic *routine*. Consequently, the difference in traffic on a specific day captures the *non-routine* traffic behavior. For the goal to capture the traffic on event an non-event days, we build a model for each location within our traffic location network for all mentioned day of week categories (the same ones used in Chapter 4). Using this model, we captured the traffic caused by PSEs by observing the difference in traffic between event days and the model.

While the Absolute Timespan of Interest is focused on the *TrafficTime* variables, the Relative Timespan of Interest focuses on the *TimeOfEvent* variability. By normalizing traffic and by working with the difference in traffic behavior from the model, we can analyze all events for a given weekday, which highly increases the observations in our analysis. However, a clear drawback lies of course in the traffic modeling that is applied. Unknown events or other influencing factors on traffic (e.g., heavy snow) might influence the created models such that they do not represent the regular behavior anymore, which could lead to a misinterpretation of additional traffic. Another challenge comes with the task of comparing traffic on event and non-event days. Let us assume we collected a set of time windows before and after a list of events and collected the normalized traffic states. How do we compare those timespans from different days of the week and different times of the day to traffic on non-event days? Whereas the *Absolute Timespan of Interest* simply defined a time window per day that can be used for event and non-event days, handling those for the *Relative Timespan of Interest* requires different strategies.

5.1.2 Binary Classification Task

For each defined Timespan of Interest from the different approaches described above, we collect traffic information for each road segment within a given radius around a venue. The collection of traffic data is thereby aimed at providing a balanced dataset between traffic on event and non-event days for each road segment in a given radius. The resulting dataset contains traffic data normalized over defined time periods (we used 5 min and 15 min slots) for the selected TOIs in combination with the information of whether an event happened at the specific time. The fact of the occurrence of an event (false = no event, true = event) is used as a predictor for our analysis. In the end, we obtain the following dataset for each traffic location in a specified radius around the venues to analyze:

$$X_1, ...X_n, E \tag{5.1}$$

where $X_1, ..., X_n$ are our traffic information and E is the event label.

We perform a cross-validation on these datasets of the performance of different classifiers with different parameter optimization techniques. As a result, we assume that those road segments where the tested classifiers perform best are those that show different behaviors on event non-event days and are, according to our definition, especially affected by the occurrence of PSEs.

At this point, we have presented our approach to collect information about traffic situations during event occurrences based on different time alignment strategies, and explained the classification-based approach and the dataset used for our studies. In the following sections, we will present our two mentioned studies and discuss their results and findings.

5.2 Flow Study

The *flow study* was conducted in 2016 and focused on a broad range of venues to allow further discussions of the findings. During that time, we had access to the traffic flow dataset as described in Chapter 3.1.1. As discussed in Chapter 4, we observed different behaviors for event impacts on traffic for different venues. One major influencing factor observed was the event type focus of a venue. In Germany, we find essentially two different types of venues: those that are specialized on sporting events and occasionally host also different event types (e.g., large-scale soccer arenas) and those that are focused on entertainment events and sometimes also host sporting events (e.g., the LANXESS Arena in Cologne). In our studies, we focus on a mixture of these venue types using the approach outlined in 5.1.

5.2.1 Experimental Setup

In total, we observed 12 different venues, of which nine were focused mainly on soccer games and three hosted different events of all types (see Table 5.1). To allow a comparison with previous observations, we selected the same "mixed" venues as presented in Chapter 4. For all venues, traffic information was derived in 5 min intervals for all *TrafficLocations (TLs)* within a radius of 4000 m. The *TrafficLocations* were based on the TMC referencing system (see Chapter 3.1.1), which guaranteed a sufficient length of the road segments and avoided

TABLE 5.1: List of venues.

Name	**City**	**Type**
Olympic Stadium Berlin	Berlin	Soccer
Commerzbank Arena	Frankfurt (Main)	Soccer
Volkswagen Arena	Wolfsburg	Soccer
Benteler Arena	Paderborn	Soccer
Bay Arena	Leverkusen	Soccer
Mercedes-Benz Arena Stuttgart	Stuttgart	Soccer
Borussia-Park	Moenchengladbach	Soccer
Signal-Iduna-Park	Dortmund	Soccer
Schwarzwald-Stadion	Freiburg	Soccer
Mercedes-Benz Arena Berlin	Berlin	Mixed
LANXESS-Arena	Cologne	Mixed

the need for further cleaning or filtering. In total, we observed one season of soccer games, which ended up with 12 games per venue. As there were also days included when our traffic dataset was incomplete, we ended up with 11–12 games per venue. For the "mixed" venues, we covered a broad range of events during the time period from 05/2014 to 05/2015 and for each venue, we captured between 50 and 60 events.

As mentioned earlier, a critical requirement for the proposed approach is a dataset that allows comparison of traffic behavior on event and non-event days for each road segment. For that, time plays an important role and we presented two different approaches to handle time variations earlier, the *Absolute Timespan of Interest* and the *Relative Timespan of Interest*. As the high frequency the traffic dataset used in this study allowed us to create a model for each *TrafficLocation* based on historic information, we implemented the *Relative Timespan of Interest* concept.

With this concept we were able to analyze all the event days in our database for the given venues without further loss. A major challenge of this concept was to define a time window to compare traffic during event and non-event days. Finding the time window on event days is straightforward, as we can simply put a time window around the event start time. But how do we find that time window on non-event days? In our approach, we defined so-called *day-pairs*. A *day-pair* is a set of two days, of which one is an event day and the other is a non-event day. The event time on the event day defines the time of interest for that *day-pair*, and we select the same time window from both days. As traffic behaves differently on different days of the week, we need to ensure that both days of the *day-pair* are comparable. To this end, we followed the same *day of week classification* as already showed in Chapter 4: Monday to Thursday (Mo–Th), Friday (Fr), Saturday (Sa), and Sunday (Su). To create the *day-pairs* and required models for the *Relative Timespan of Interest*, we divide our dataset into three parts for each of these *day of week (DOW)* classes:

- Event set (E)
- Non-event set (N)
- Model set (M)

where $E_{DOW} \cup N_{DOW} \cup M_{DOW} = D_{DOW}$. The event set *E* contains all event days for that specific venue that we analyze. The non-event set *N* contains a list of randomly selected non-event days with the same number of days as the *E* set. The model set *M* contains all non-event days of the database that are not part of *N*.

First, we created a model for the *day of week class* by calculating the mean traffic behavior over all non-event days in M_{DOW}. For each event day in E_{DOW}, we captured traffic within the *event time window* of 120 min before the event start. That time window was then also used for one randomly selected non-event day in N_{DOW}. We calculated the difference between the captured traffic within the *event time window* for event and non-event days from the model and put the resulting *day-pair* into our dataset. For example, we assume that we have an event starting on a Saturday at 18:00. We select the times depending on the *event time window* of 120 min, which results in Saturday 16:00–18:00. Our approach would add that specific time window to the dataset, together with a randomly selected day from the Saturday *day of week class* from 16:00–18:00 when no event happened (derived from the *N* set). The final *day-pair* would contain difference in traffic to from Saturday model for the event day from 16:00–18:00 and for a randomly selected non-event day on Saturday from 16:00–18:00.

We used a set of different classifiers (ANNs, K-Nearest Neighbor, and SVMs) and benchmarked them using a cross-validation approach with the *F1 measure* (hereafter called the *f-measure*) as a metric. For more details about these classifiers, we can strongly recommend referring to [100] or [101].

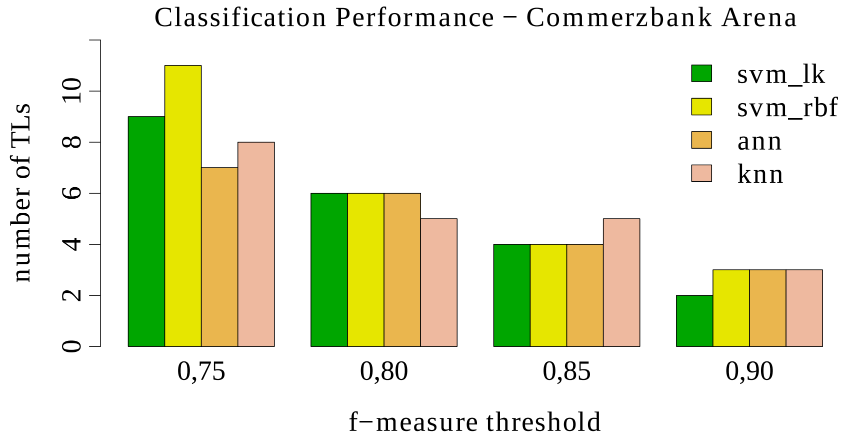


FIGURE 5.2: Performance after the cross-validation for different classifiers. Venue: Commerzbank Arena in Frankfurt.

For the benchmark, we analyzed traffic around the Commerzbank Arena in Frankfurt, Germany. It is explicitly emphasized that the classification method is not the main finding of this thesis, but rather the methodology. We therefore do not claim any sort of completeness of the benchmarked algorithms. The crucial point is the applicability of the method to real-world examples. We used the different classification methods in combination with a simple parameter optimization technique. Figure 5.2 shows the results for the following classifiers: ANN, K-Nearest Neighbor, SVM with a linear kernel, and SVM with a rbf kernel.

The results show similar performances of all the classification methods. Based on these results and the requirement to run the analysis for a large-scale study, we implement the method that takes the least training time: K-Nearest Neighbor. In the following, we report on the results of a K-Nearest Neighbor approach. By experimenting with different distance measures, the *Dynamic Time Warp (DTW)* distance measure performed best on our data. The reasons for that is the shift in time of traffic influence by events, which are captured better using DTW than, for instance, Euclidean measures. Therefore, in the following, we report on the results of a K-NN optimization method using DTW as a distance metric.

5.2.2 Results

Throughout all venues, we find very different numbers of affected TLs. Because we compared venues in very different areas (some in dense inner cities, some in more rural areas), these results are expected. In the following subsections, we present the results for all venues from Table 5.1 in detail.

5.2.2.1 Olympic Stadium Berlin

For the Olympic Stadium in Berlin, we captured 211 *Traffic Locations* in total within a selected radius of 4000 m (see Figure 5.3). From these 211 TLs, 23 showed significant correlation to the occurrence of soccer games in the stadium. Figure 5.4 shows the results for the *K-NN* model for different thresholds. The road segments with the highest *f-measure* (≥0.9) are the one directly in front of the stadium (see **A** in figure 5.4i). Those show a huge discrepancy between their behavior on event and non-event days, which is intuitive as they are frequently

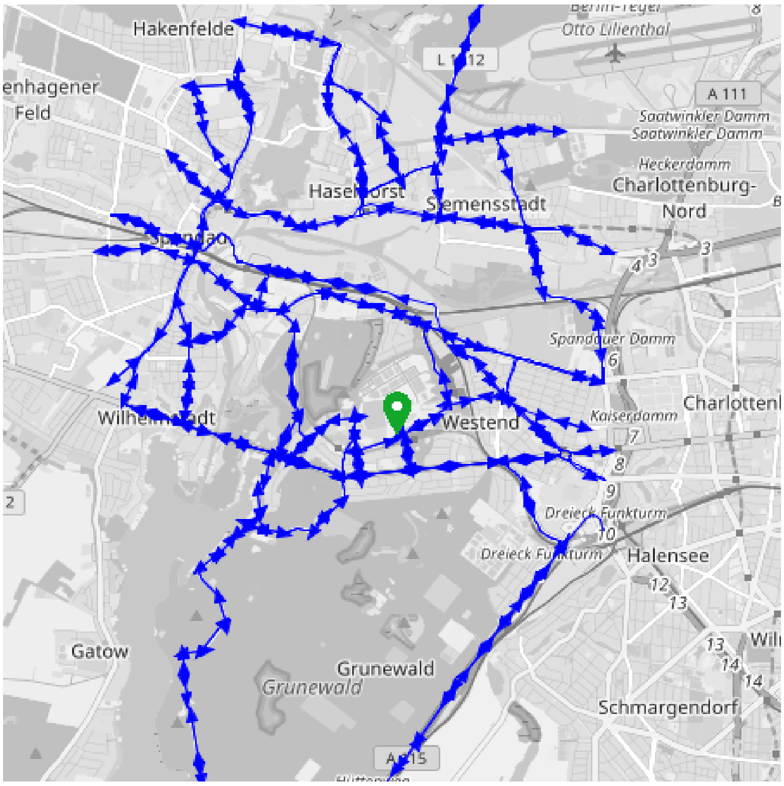
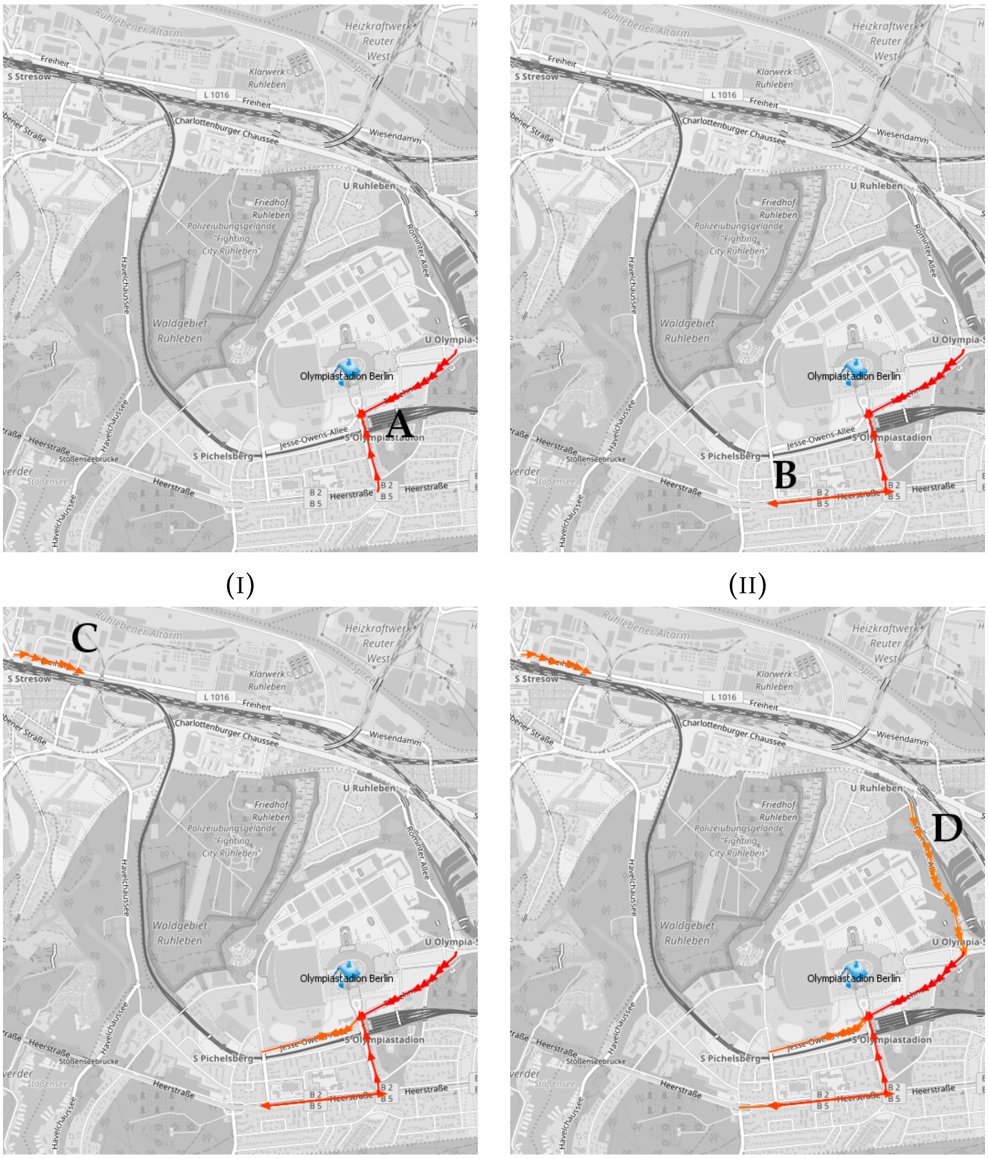


FIGURE 5.3: Selected TLs within a radius of 4000 m around the Olympic Stadium Berlin (green marker).



(I)

(II)

(III)

(IV)

FIGURE 5.4: Venue: Olympic Stadium Berlin. (5.4i) *f-measure* ≥ 0.90. (5.4ii) *f-measure* ≥ 0.85. (5.4iii) *f-measure* ≥ 0.80. (5.4iv) *f-measure* ≥ 0.75.

used to reach the parking lots of the stadium. Comparing their traffic behavior on event and non-event days (see Figure 5.5) supports that intuition.

The graph shows the traffic behavior for one exemplary TL (TMC code:13/5/32571). The mean line on event days shows a completely different behavior than on non-event days, which matches our definition of being usually affected by PSEs. When lowering the threshold to an *f-measure* ≥ 0.85, additional road segments are involved (see **B** in Figure 5.4ii). They show parts of the “Heerstraße” at the intersection to the stadium. A possible explanation is the upcoming congestion

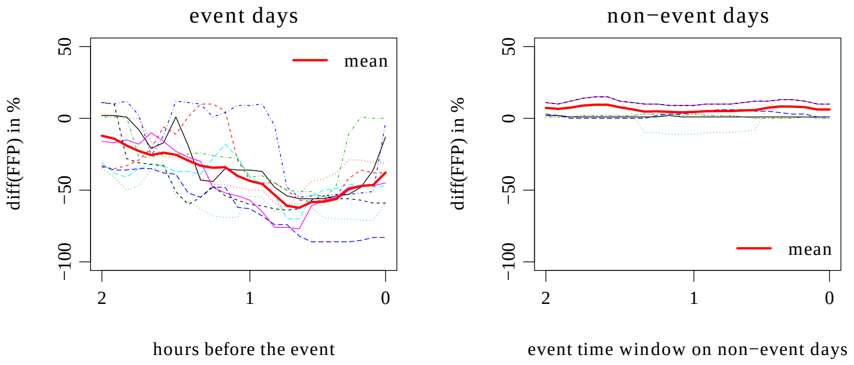


FIGURE 5.5: Venue: Olympic Stadium Berlin. Traffic during event and non-event days for the TL at marker **A** in Figure 5.4i (TL 13/5/32571). y-axis: difference in *FreeFlowPercentage* compared to the model on non-event days.

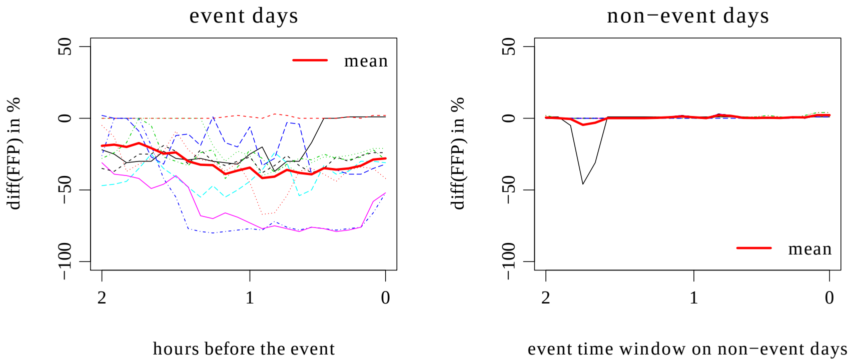


FIGURE 5.6: Venue: Olympic Stadium Berlin. Traffic during event and non-event days for the TL at marker **B** in Figure 5.4ii (TL 13/5/21534). y-axis: difference in *FreeFlowPercentage* compared to the model on non-event days.

at the intersection, which also influences traffic on these segments. Again, manual inspection of the traffic behavior on these links (shown as an example for TL id: 13/2/21534 in Figure 5.6) supports this case. However, the event day behavior on this TL varies much more than for that in Figure 5.5. Whereas the road segment in 5.5 shows severe congestion on every single game day, traffic on this road segment was barely influenced on 23/08/2014, when Herta BSC (home team) played against Bremen (red dotted line). However, other games showed a severe impact on this road segment as, for instance, traffic dropped up to -80% of the *FreeFlowPercentage* during the soccer match on 28/02/2015, when Hertha BSC played against FC Augsburg (blue dotted line). Possible explanations for this could be a higher popularity of the game, a higher traffic density in general on this day, or other external influencing factors (e.g., road constructions). The last two phenomena could also be the reason for the outlier during non-event days (black line), when FFP drop by to -43%.

When further reducing the *f-measure* threshold, additional road segments become involved (see **C** and **D** in Figure 5.4iii and 5.4iv). However, the TL at **C** seems particularly like a misinterpretation, as it is far away from the venue. As there might be an explanation that a local expert could give, it is also quite likely that these are simply misinterpretations and that segments below a threshold of 0.8 might be influenced by other phenomena. Additionally, analyzing traffic on TL 13/2/41077 (**C**) shows that a differentiation between event and non-event behavior is rather difficult to achieve (see Figure 5.7). A possible conclusion is that for this specific venue, a threshold ≥ 0.80 is suitable to identify impacted areas in future analysis.

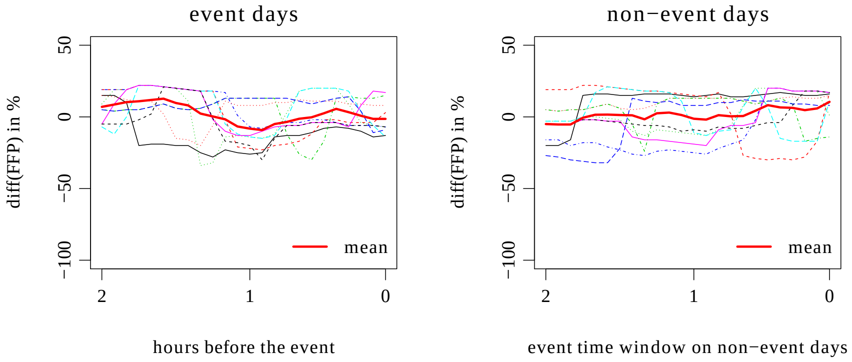


FIGURE 5.7: Venue: Olympic Stadium Berlin. Traffic during event and non-event days for the TL at marker **C** in Figure 5.4iii (TL 13/5/41077). y-axis: difference in *FreeFlowPercentage* compared to the model on non-event days.

5.2.2.2 Commerzbank Arena Frankfurt

Around the Commerzbank Arena in Frankfurt, we captured at total of 158 *Traffic Locations* within a selected radius of 4000 m (see Figure 5.8). Of these TLs, 22 showed significant correlation to the occurrence of soccer games at the stadium. Figure 5.9 shows the results for the *K-NN* model for different thresholds.

P1–P3 in Figure 5.9i are official parking lots and people are advised to use them to get to the stadium (as written on the official website[1]). The road segments with the highest *f-measure* (≥ 0.90 shown in Figure 5.9i) are those directly leading to the stadium on the B44, a federal highway within Frankfurt area (between **P1**, **P2**, and **P3** in the image). As all of these TLs directly lead toward the recommended parking lots and the venue itself, this result seems intuitive. An

[1] `http://www.commerzbank-arena.de/`

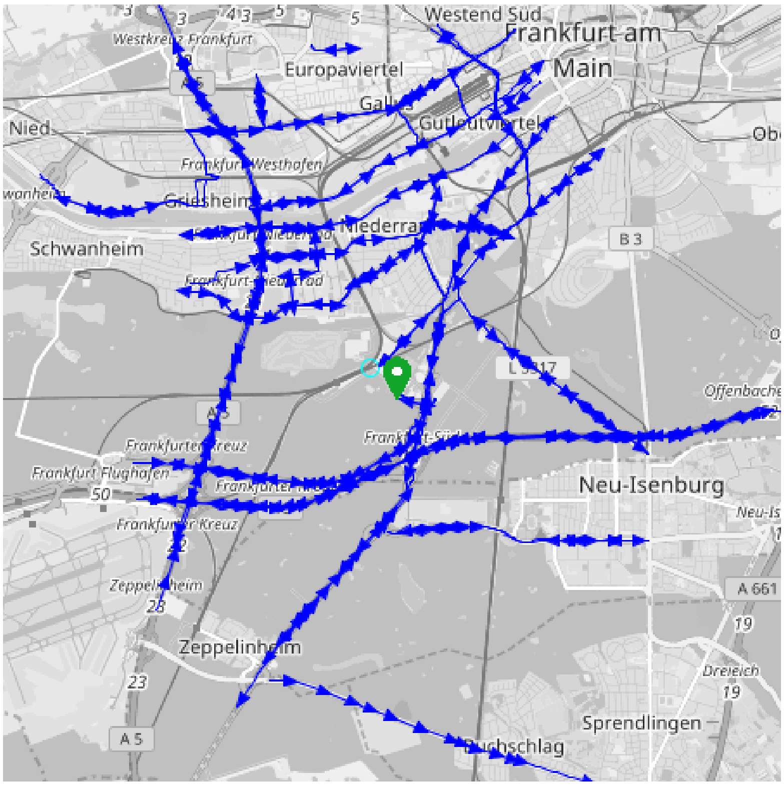
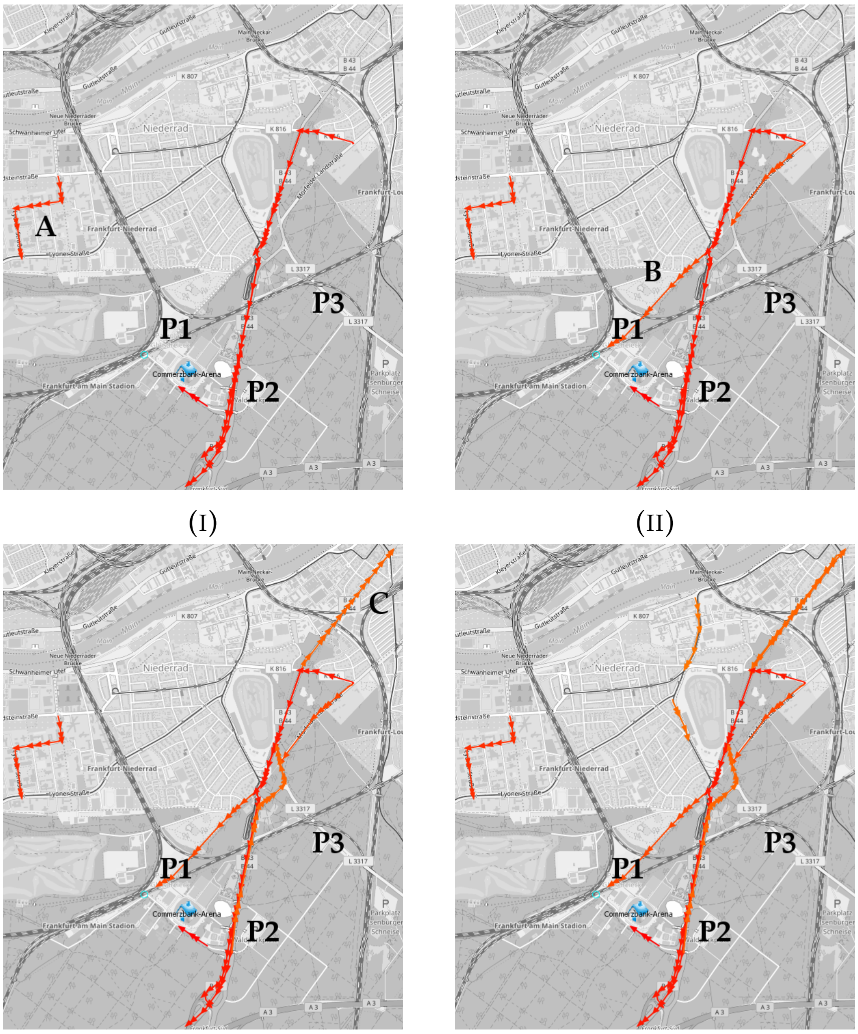


FIGURE 5.8: Selected TLs within a radius of 4000 m around the Commerzbank Arena Frankfurt (green marker).

exception is the road segments around marker **A**. They seem oddly far away from the venue and do not seem to be part of any *incoming* route. Are they wrongly classified? A manual analysis of the traffic behavior shows a rather clear impact of events on these locations (see Figure 5.10). Although the mean line is less obviously different than in the previous examples, the specific event impact is visible in the data. Because we used the DTW as a distance measure, the event and non-event days are separable by the classifier. An explanation is found on the website of the arena, where they give advice on how to reach the stadium by car. They mention three official parking lots (marker **P1**, **P2**, and **P3**) and refer to additional parking space within the area around **A**. From the website (own translation): "...within the area of Lyonerstraße, Herriotstraße and Hahnstraße, you can find additional parking...". The impacted road segments are part of the Lyonerstraße, which explains their behavior.

When reducing the threshold to ≥ 0.85 (Figure 5.9ii), additional road segments get involved (marker **B**). Those lead directly to the **P1** parking lot, which is a possible explanation why people would use those road segments to before the game starts.



(I)

(II)

(III)

(IV)

FIGURE 5.9: Venue: Commerzbank Arena Frankfurt. (5.9i) *f-measure* ≥ 0.90. (5.9ii) *f-measure* ≥ 0.85. (5.9iii) *f-measure* ≥ 0.80. (5.9iv) *f-measure* ≥ 0.75.

Figure 5.9iii and 5.9iv show further reduced thresholds (≥ 0.80 and ≥ 0.75, respectively). As already seen with the results above for the Olympic Stadium in Berlin, the further we reduced the thresholds, the more likely get misclassification. The TL at marker **C** in Figure 5.9iii points directly away from the venue. Its traffic behavior is shown in Figure 5.11. The TL shows similar behavior during event and non-event days, and a clear distinction based on traffic observation is difficult to perceive. A possible explanation for why it shows event-specific behavior could lie in the side effects from traffic on other incoming road segments due to, for example, blocked intersections. However, more specific conclusions would require a domain expert that knows the road segments and their behavior during soccer games in Frankfurt.

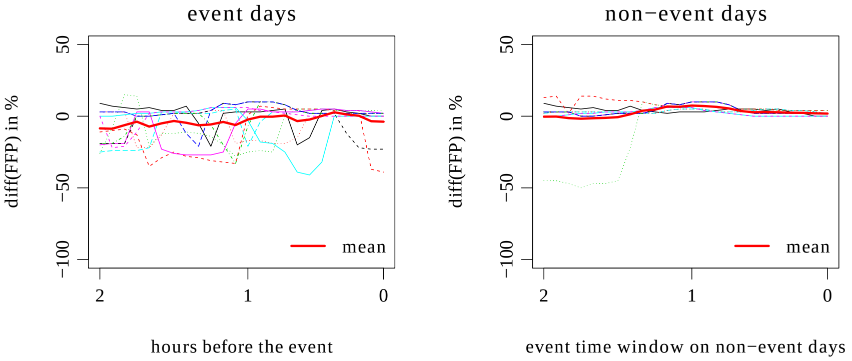


FIGURE 5.10: Venue: Commerzbank Arena Frankfurt. Traffic during event and non-event days for the TL at marker **A** in Figure 5.9i (TL 13/5/41077). y-axis: difference in *FreeFlowPercentage* compared to the model on non-event days.

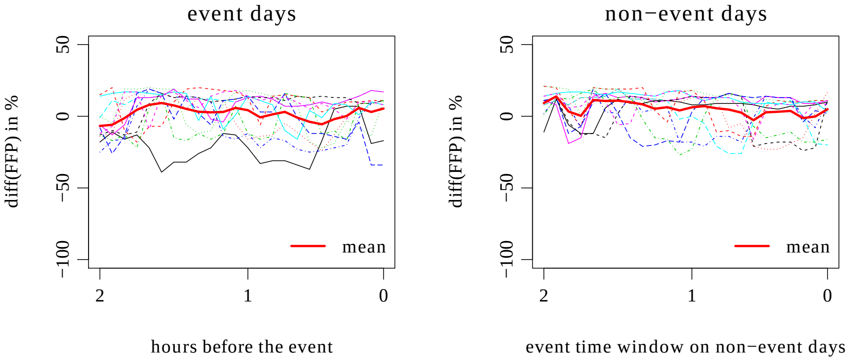


FIGURE 5.11: Venue: Commerzbank Arena Frankfurt. Traffic during event and non-event days for the TL at marker **C** in Figure 5.9iii (TL 13/2/23489). y-axis: difference in *FreeFlowPercentage* compared to the model on non-event days.

5.2.2.3 Signal Iduna Park Dortmund

Around the Signal Iduna Park in Dortmund, we captured a total of 131 *Traffic Locations* within a selected radius of 4000 m (see Figure 5.12). Of these TLs, 18 showed a significantly different behavior during soccer games. Figure 5.13 shows the results for the K-NN model for different thresholds. In this example, the road segments with the highest *f-measure* are those that are not directly in front of the venue, but point toward it. To get to the stadium, drivers have to cross the intersection at marker **A** and head toward the venue. As this intersection is highly used by incoming traffic, these results seem reasonable. The result for TL 13/2/24373 at marker **B** is less intuitive. Its traffic behavior is shown in Figure 5.14. Although the traffic behavior varies significantly during event days, its different behavior compared to non-event days is obvious in the

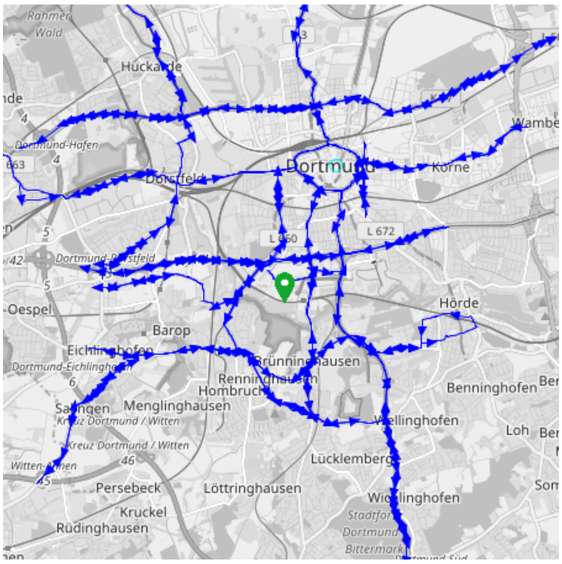
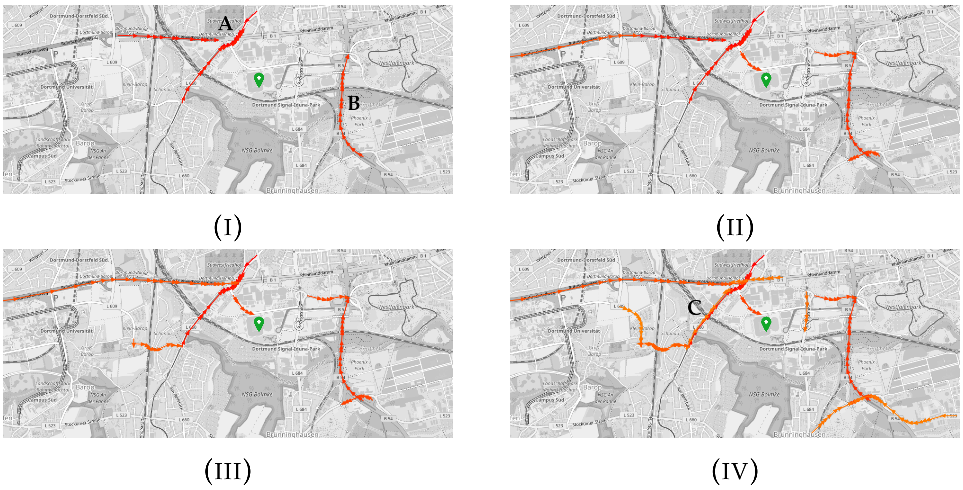


FIGURE 5.12: Selected TLs within a radius of 4000 m around the Signal Iduna Park Dortmund (green marker).



(I)

(II)

(III)

(IV)

FIGURE 5.13: Venue: Signal Iduna Park Dortmund. (5.13i) *f-measure* ≥ 0.90. (5.13ii) *f-measure* ≥ 0.85. (5.13iii) *f-measure* ≥ 0.80. (5.13iv) *f-measure* ≥ 0.75.

data. An explanation is, again, found on the venue website[2] where the B54 (TL 13/2/24373 is a part of it) is suggested as an incoming route for people arriving from the south. The road segment ends at an intersection where people leave the B54 to get to the venue. The results up to a threshold of 0.80 look reasonable, knowing that the parking lots for this stadium are located north and east of the venue, which explains the traffic in that area. Below a threshold of 0.80, we see the first questionable results at marker **C** in Figure 5.13iv, where TLs are identified that seem to point away from the venue. This means that, compared to the other venues, we see intuitively reasonable results up to a threshold of 0.80,

[2] `https://www.signal-iduna-park.de/`

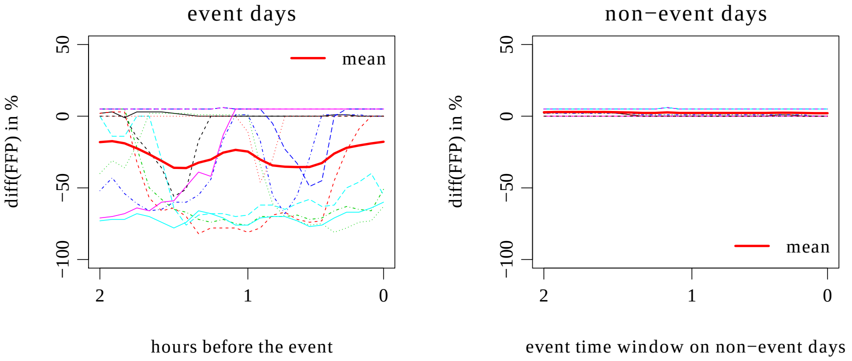


FIGURE 5.14: Venue: Signal Iduna Park Dortmund. Traffic during event and non-event days for the TL at marker **B** in Figure 5.13i (TL 13/2/24373). y-axis: difference in *FreeFlowPercentage* compared to the model on non-event days.

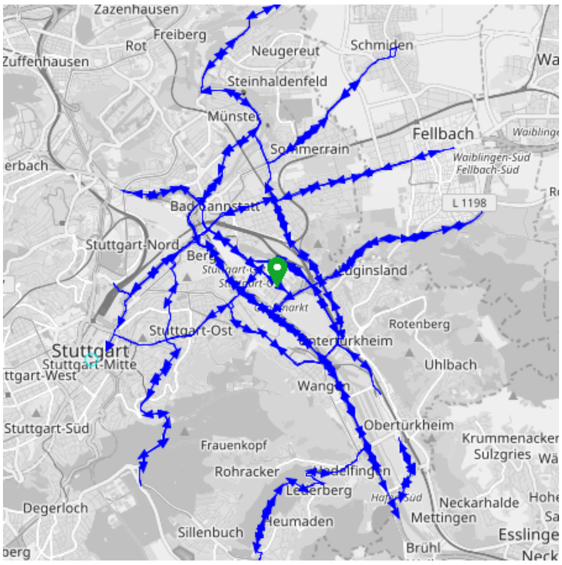
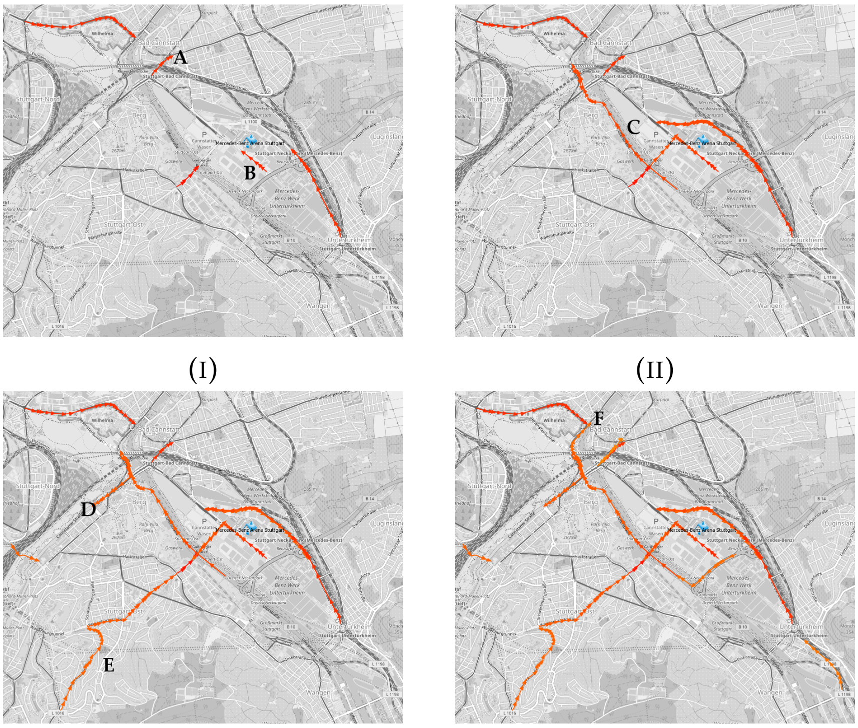


FIGURE 5.15: Selected TLs within a radius of 4000 m around the Mercedes-Benz Arena (green marker).

which is more than for the other venues. One could argue that the Signal Iduna Park in Dortmund is the biggest German first league soccer venue in Germany, with a capacity of approximately 81000 visitors. This could be a reason why event traffic is more explicitly noticeable in this area.

5.2.2.4 Mercedes-Benz Arena Stuttgart

Around the Mercedes-Benz Arena in Stuttgart, we captured 152 TLs in total (see Figure 5.15) within the given radius. Of these, 27 showed a significantly different behavior on event days compared to non-event days. Figure 5.16 shows the results for the K-NN model for different selected thresholds. Again, similar to



(I)

(II)

(III)

(IV)

FIGURE 5.16: Venue: Mercedes-Benz Arena Stuttgart. (5.16i) *f-measure* ≥ 0.90. (5.16ii) *f-measure* ≥ 0.85. (5.16iii) *f-measure* ≥ 0.80. (5.16iv) *f-measure* ≥ 0.75.

the results for Signal Iduna Park the road segments with the highest f-measure (Figure 5.16i) are those that are obviously part of incoming routes to the venue. The only questionable results are those at marker **A** in Figure 5.16i, as the segment is pointing away from the venue and does not seem to be of any obvious incoming route. It still results in an f-measure of 0.91, a recall of 0.83, and precision of 1.0. These results are remarkable, because the high precision means that, all the times that the TL showed a significantly different behavior, a soccer game happened. Figure 5.17 shows the traffic behavior on this specific TL (TMC code: 13/2/54812). The data reveal, on average, a lower FFP on event than on non-event days (red mean line). During event days, the mean never reaches the 0-line which means that it always shows more congestion than normally, which probably leads to the classification result. However, compared to other TLs that have been analyzed so far (see above), this road segment also shows severe fluctuations during non-event days. Is this a common phenomenon in that venue? For comparison, we analyze the *TrafficLocation* directly in front of the venue (marker **B** in Figure 5.16i) in Figure 5.18.

The TL shows an *f-measure* of 0.92, a recall of 1.0, and a precision of 0.86. As

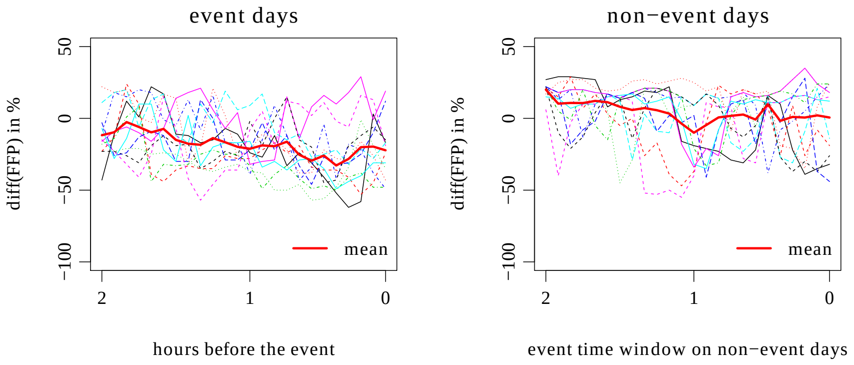


FIGURE 5.17: Venue: Mercedes-Benz Arena Stuttgart. Traffic during event and non-event days for the TL at marker **A** in Figure 5.16i (TL 13/2/54812). y-axis: difference in *FreeFlowPercentage* compared to the model on non-event days.

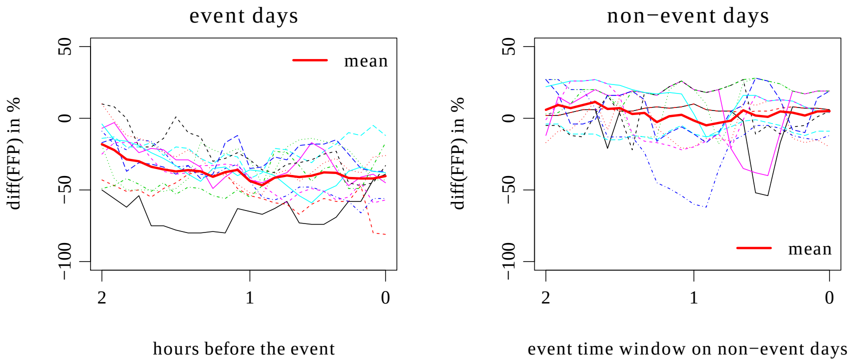


FIGURE 5.18: Venue: Mercedes-Benz Arena Stuttgart. Traffic during event and non-event days for the TL at marker **B** in Figure 5.16i (TL 13/2/54812). y-axis: difference in *FreeFlowPercentage* compared to the model on non-event days.

expected, the TL shows a clearly distinguishable behavior on event and non-event days. However, it also shows rather large fluctuations on non-event days. A possible explanation is given by the INRIX scorecard[3]. It states that (own translation): "... Stuttgart experiences the biggest increase in traffic congestion in Europe and is now officially the city with the highest traffic demand in Germany..."[102]. As Stuttgart is one of the most congested cities in Germany, fluctuations, independent of events, seem to be the norm.

Further lowering the *f-measure* threshold for selecting road segments in Figure 5.16 shows similar results as seen in the examples above. Additional road segments get selected (see **C** in Figure 5.16ii and **D** and **E** in Figure 5.16iii). Their data allow an interpretation of being affected by events. However, a clear decision (especially for **D**) can only be made by a domain expert from the area.

[3] `http://inrix.com/press-releases/scorecard-de/`

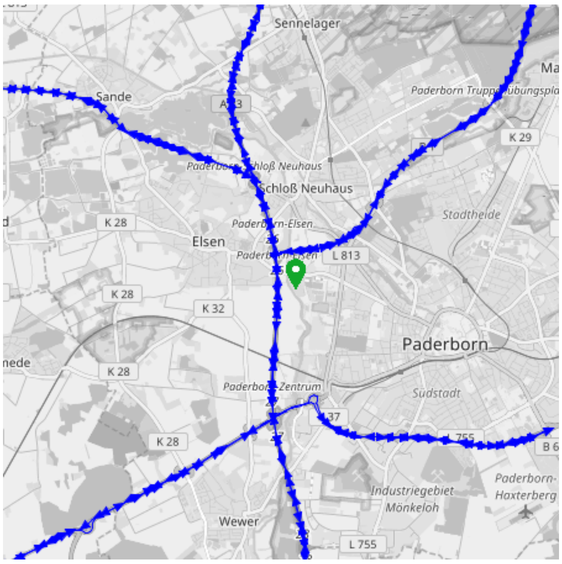


FIGURE 5.19: Selected TLs within a radius of 4000 m around the Benteler-Arena Paderborn (green marker).

Further reducing the threshold to ≥ 0.75 lead to misclassification (see **F** in Figure 5.16iv) of road segments that are obviously not involved in any incoming routes to the stadium. In general, results for this venue show the applicability of the approach in dense inner city scenarios with very high traffic fluctuations.

5.2.2.5 Benteler-Arena Paderborn

Earlier discussions have shown that the impact on traffic highly depends on the venue. A good example of this is the Benteler-Arena in Paderborn. The arena is in a more rural area than those above. In total, we captured 27 *Traffic Locations* within the selected radius of 4000 m (see Figure 5.19). Of these 27 TLs, none show a significant congestion behavior over the time of our analysis. Figure 5.20 shows the *f-measure* for all selected TLs (ordered by *f-measure*) showing that none of them reach a value ≥ 0.75. As an example, Figure 5.21 shows the traffic

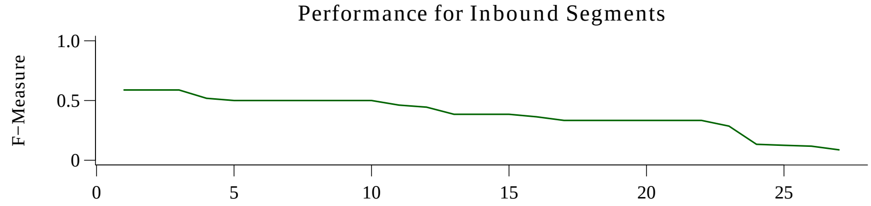


FIGURE 5.20: Venue: Benteler-Arena Paderborn. *f-measure* of all TLs within the radius (sorted by f-measure).

behavior of the TL right in front of the venue on the direct path to the venue parking lot. As seen in the graph, there is no significant traffic congestion on event or non-event days. Additionally, manually analyzing traffic in the area reveals that there are no noticeable traffic congestion events in the area around the venue. The capacity of the road network seems to be capable to cover an (eventual) additional event traffic. The most intuitive explanation probably lies in the size of the arena. Whereas the Signal Iduna Park accommodates approximately 81000 visitors, the Benteler-Arena can accept (only) 15000 visitors.

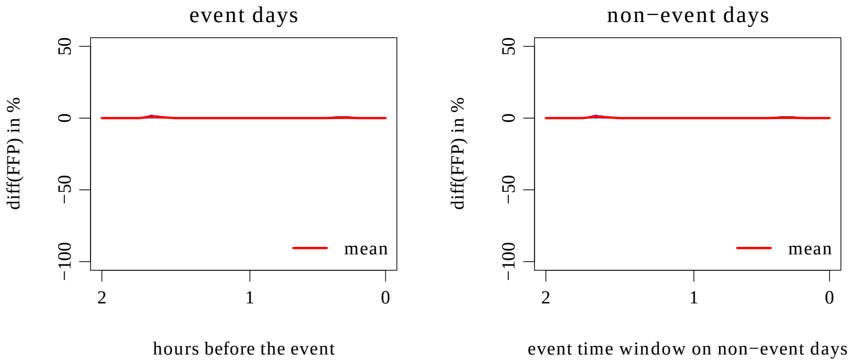


FIGURE 5.21: Venue: Benteler-Arena Paderborn. Traffic during event and non-event days for a TL directly in front of the venue (TL 13/5/41077). y-axis: difference in *FreeFlowPercentage* compared to the model on non-event days.

5.2.2.6 Mercedes-Benz Arena Berlin

As the first "mixed" venue in our analysis, we analyzed traffic around the Mercedes-Benz Arena in Berlin. During the time of our analysis we observed 66 events happening. Owing to gaps in our traffic database, we ended up with 48 events for which both event and traffic information were available. Although we lost around one quarter of all the events in our final dataset, this amount is still significantly higher compared to 11–12 games observed at the soccer venues. In total, we captured 285 *Traffic Locations* within a radius of 4000 m (see Figure 5.22). The maximum *f-measure* observed is 0.72 for one specific TL. That result is significantly lower than for most of the other venues discussed above. Further reducing the threshold to 0.65 results in five TLs with an *f-measure* above that threshold (see figure 5.23). The results show that the most correlated TL (marker **A** in Figure 5.23i) is directly in front of the venue, which is similar to our observations at the other venues. Although its final *f-measure*

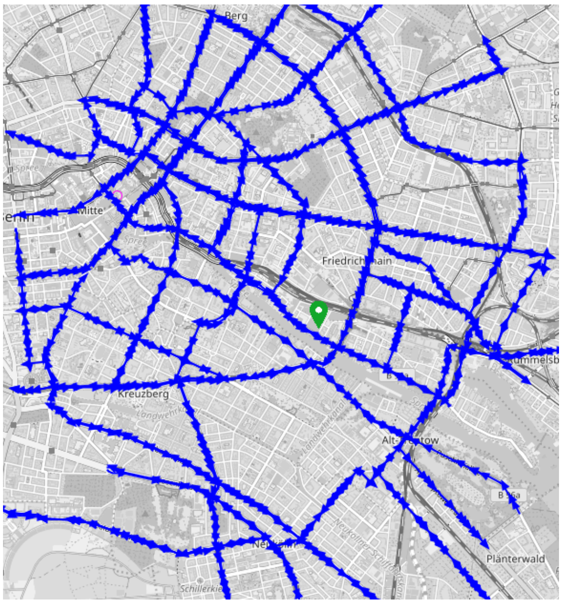
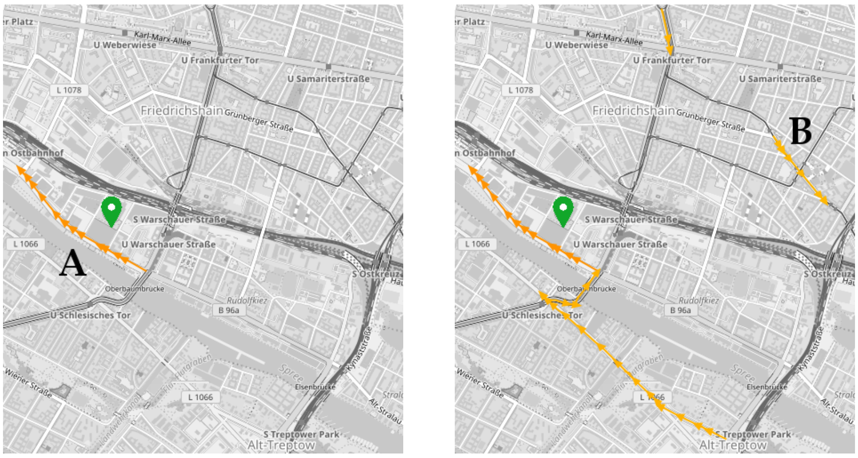


FIGURE 5.22: Selected TLs within a radius of 4000 m around the Mercedes-Benz Arena in Berlin (green marker).



(I) (II)

FIGURE 5.23: Venue: Mercedes-Benz Arena Berlin. (5.23i) *f-measure* ≥ 0.65. (5.23i) *f-measure* ≥ 0.70.

value is much lower than for similar TLs in the previous examples, its selection as the most impacted seems plausible. The additional TLs by further reducing the threshold do not seem to be far off the expectations for incoming route segments either, except for the TL at marker **B** in Figure 5.23ii. Below a threshold of 0.65, the results become inconsistent with the expectations and obvious misclassifications appear. However, analyzing traffic behavior in detail explains the weak classification result. Figure 5.24 shows the traffic behavior at the road segment with the highest classification result (marker **A** in Figure 5.23i). Although the mean line is slightly lower on event than non-event days, both day

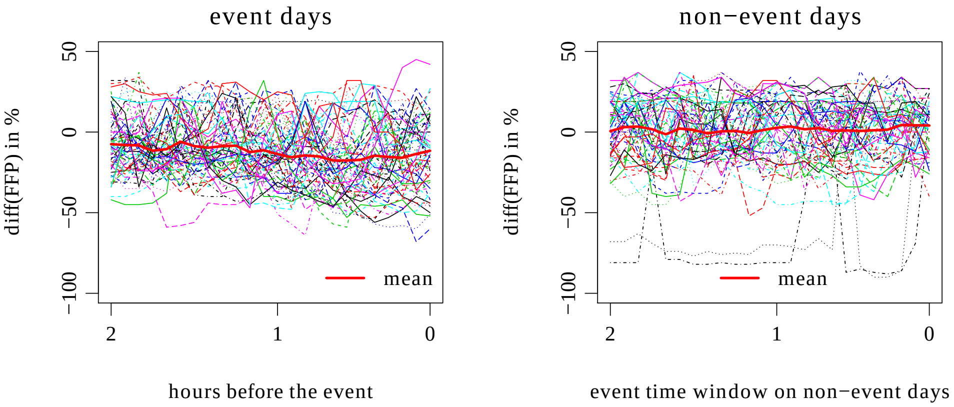


FIGURE 5.24: Venue: Mercedes-Benz Arena Berlin. Traffic during event and non-event days for the TL at marker **A** in Figure 5.23i (TL 13/5/26927). y-axis: difference in *FreeFlowPercentage* compared to the model on non-event days.

categories show very high variations in traffic.

A possible explanation is the road infrastructure around the venue. We have seen drastic influences of the infrastructure, especially its ability to cope with varying traffic load scenarios, on the event traffic observations. Although possible, it seems unlikely that these results could be explained by the road infrastructure only. The suggested approach worked in the previous examples in dense inner cities scenarios and even in cities with very high traffic volume (see Olympic Stadium Berlin or the Mercedes-Benz Arena in Stuttgart above).

Another explanation lies in other venues within the radius of the analysis that might interfere with the observations for this particular one. Especially in Berlin, where many PSEs are hosted, the density of venues is very high. Figure 5.25 gives an impression of venues in the area (green markers are venues, and the Mercedes-Benz Arena is marked in magenta). In total, we list 77 venues within a radius of 4000 m around the Arena. This list is derived from our industrial event dataset and only shows those venues for which the company sells tickets. In reality, there might be even more venues. Events happening at the same time in different venues could result in additional load on the infrastructure at times that are counted as non-event times. That leads to a reduced performance of the classifier. For example, on Thursday 28/05/2015, which is counted as a non-event day for the analysis, we observed severe traffic congestion for the TL shown in figure 5.24 (dashed black line, starting at -81%). On that day three different events happened in the direct vicinity (closer than 1000 m radius) of the Mercedes-Benz Arena. Two of them were concerts (Rhodes played

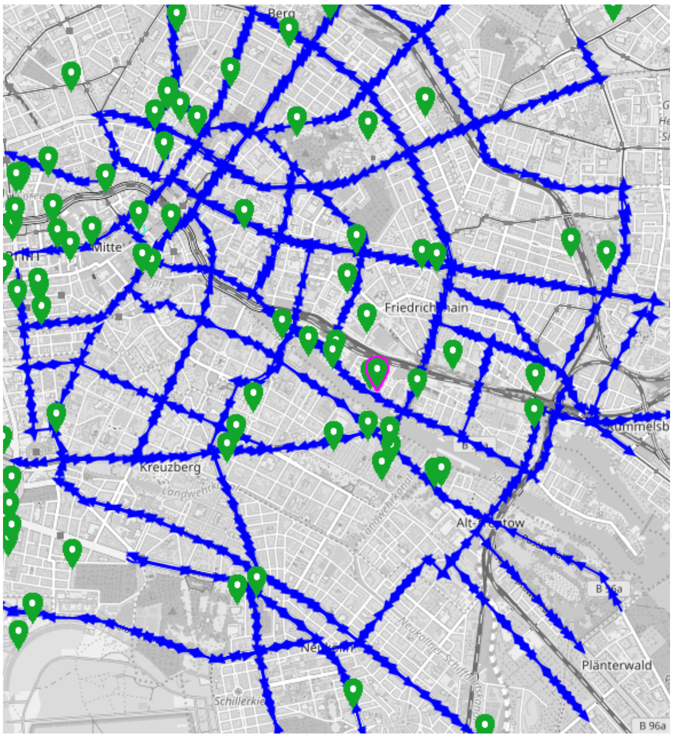


FIGURE 5.25: Venues around the Mercedes-Benz Arena in Berlin. Green marker: venues. Magenta: Mercedes-Benz Arena Berlin.

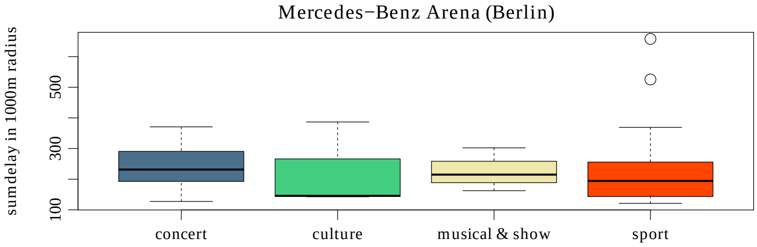


FIGURE 5.26: Avgdelay variation over time attributes within a radius of 1000 m around the Mercedes-Benz Arena Berlin.

at *Berghain* and Leslie Clio at the *Postbahnhof Club*) and one cultural event at the *Kriminaltheater*. Each one of these could have caused the observed delay. Owing to such interferences, a clear separation between event and non-event days solely by traffic observations is difficult to achieve.

A third possible explanation lies in the event type diversity. In contrast to the first venues that mostly focused on soccer games, this venue hosts a mixed type of event categories (concert, culture, musical&show, and sport). We have seen that those categories of events show a different impact on traffic. For the discussion, we take the example from Chapter 4, Figure 4.5ii (shown here again in Figure 5.26 for the convenience of reader). The graphs show that among the categories, the observed traffic impact varies. As their specifics have been discussed earlier, the general observation has a huge impact on the performance of the classifier: The only information available to the classifier is the presence of an event, independent of its category. That implies that its performance is highly dependent on the consistency of the event impact. The more permanent is the impact of events on traffic, the better the classification result becomes. To further analyze this assumption, we discuss the other two *mixed* venues in the following.

5.2.2.7 LANXESS Arena Cologne

Similar to the Mercedes-Benz Arena in Berlin, the LANXESS Arena in Cologne shows comparably weak classification results. In this venue, we observed traffic on 160 TLs within a radius of 4000 m for 33 events. The highest observed *f-measure* for one TL is 0.71. Figure 5.27 shows the results for setting the threshold to 0.65. The road segments at marker **A** are the two TLs with an *f-measure* ≥ 0.7 and all others are in the range 0.65–0.70. In total, we find five TLs ≥ 0.65. From these five TLs, at least those at marker **B** seem (to an observer who is not a local expert) to be wrongly classified as an incoming route because they point away from the venue. Traffic on the TL with the highest *f-measure* looks similar to the example of the Mercedes-Benz Arena Berlin, shown in Figure 5.28. Additionally, in this example, traffic varies significantly on event and non-event

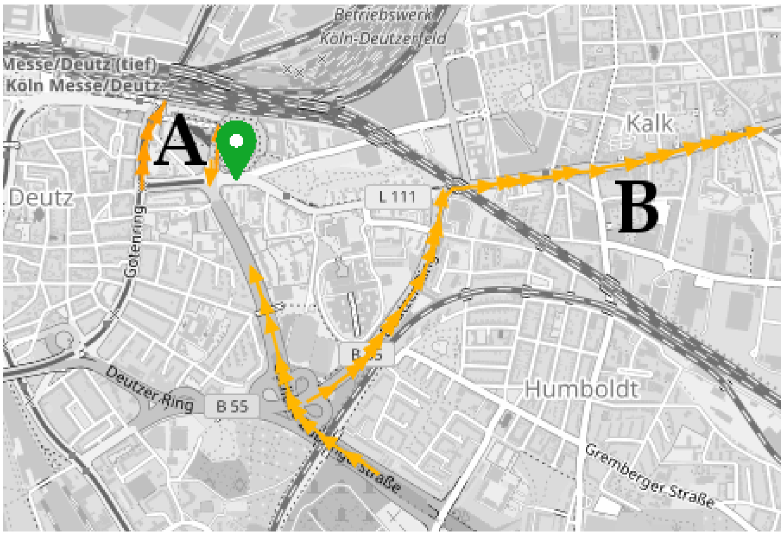


FIGURE 5.27: Venue: LANXESS-Arena Cologne (green marker). *F1-measure* ≥ 0.65.

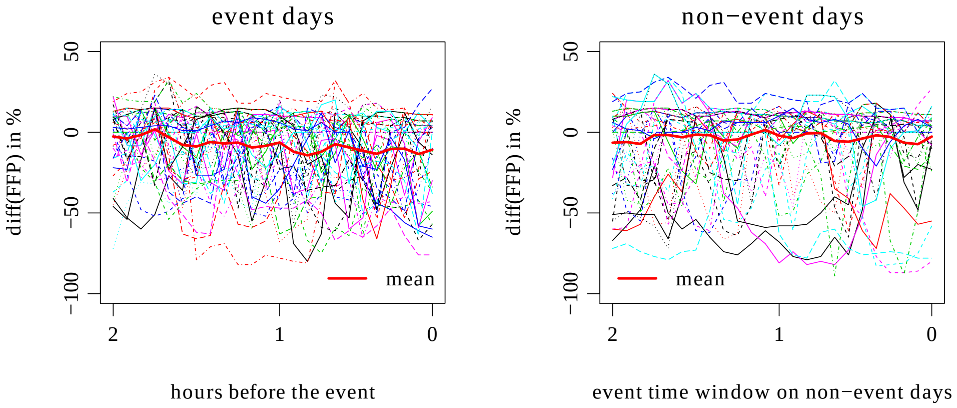


FIGURE 5.28: Venue: LANXESS Arena Cologne. Traffic during event and non-event days for the TL at marker **A** in Figure 5.27. TL with the highest *f-measure* (TL 13/5/40103). y-axis: difference in *FreeFlowPercentage* compared to the model on non-event days.

days and a separation is difficult to achieve. In Cologne, the venue density is significantly high. Figure 5.29 shows the spatial impact radius of the analysis (TLs in blue, green markers are venues, LANXESS Arena marked in magenta).

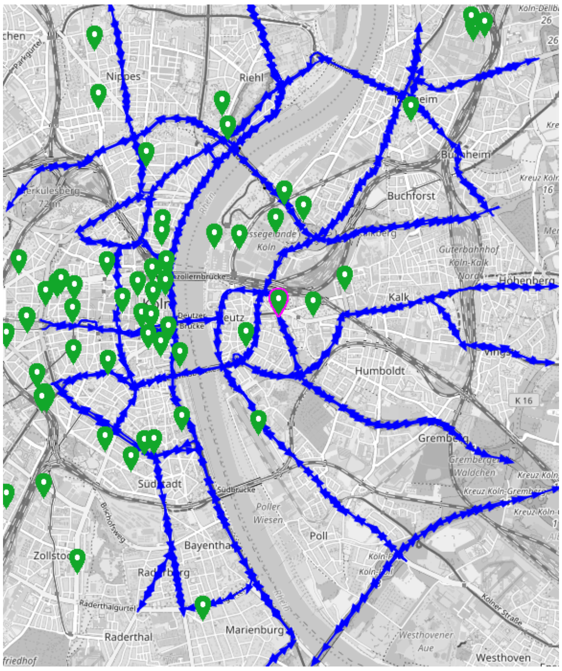


FIGURE 5.29: Selected TLs within a radius of 4000 m around the LANXESS Arena in Cologne. LANXESS Arena: green marker with magenta. Venues: green markers. Selected TLs: blue.

We find in this example, compared to the Berlin example, fewer venues in the immediate vicinity. However, north of the LANXESS Arena, there is the *Messegelände Köln*, where many events are hosted. In addition, most of the

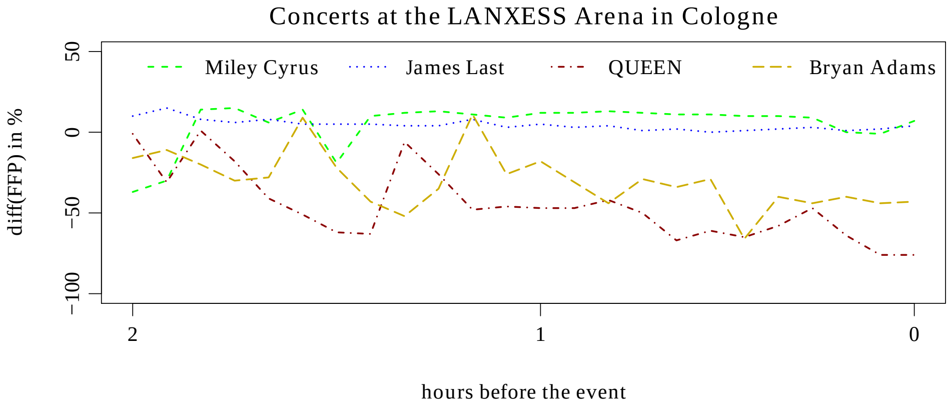


FIGURE 5.30: Venue: LANXESS Arena Cologne. Traffic during four exemplary concerts at marker **A** in Figure 5.27. Miley Cyrus (26/05/14), James Last (26/04/15), QUEEN (29/01/15), Bryan Adams (09/12/14).

venues are in the inner city (west of the arena). As there are only a number of bridges to cross the river to get to the inner city, events happening there might also influence traffic in a wider area.

This arena hosted events of four different categories (concert, misc, musical&show, and sport). Whereas all of them showed severe variations in the observed traffic behavior, the *concert* category showed the highest fluctuations. Figure 5.30 shows traffic on the same TL as in Figure 5.28 for four selected concerts. Whereas traffic before *Miley Cyrus* on 26/05/14 and before the *James Last* on 26/04/15 do not show any significant congestion behavior on this specific TL, traffic before the concert of *QUEEN* on 29/01/15 and Bryan Adams on 09/12/14 is strongly affected. There are many possible reasons for this phenomenon, as we have discussed before. In Chapter 6 we will analyze this phenomenon in detail. At this point, these observed variations between events drastically reduce the performance of the classifier and explain the poor results for this venue.

5.3 Incident Study

We have seen that the shown approach works for soccer games based on traffic flow information. Unfortunately, this type of dataset is usually not available to the public. What is usually available is *incident data* (as described in Chapter

3.1.2), which is used in everyday navigation devices[4]. To get an idea of its applicability for analytics use cases, we conducted the study again using the *Incident* dataset for two soccer venues in Germany.

During the study we observed traffic around the Moenchengladbach Arena and the Wolfsburg Arena in Germany. As a traffic measure, we collected *delay time* information in 15 min intervals for all *Traffic Locations* within a radius of 4000 m around the venues. The *Traffic Locations* were based on the NDS digital map (as described in 3.1.2). As some *Traffic Locations* in this map format are very short, they tend to show high fluctuations in their traffic behavior. We therefore decided to filter out road segments that are shorter than 30 m. To handle varying times of events, we implemented the *Absolute Timespan of Interest* concept (see 5.1.1.1) to limit *TrafficTime* variances. Unfortunately, this concept reduced our event database significantly. However, as the *incident data* does not allow a coherent time information for each *Traffic Location* a more data-preserving method such as the *Relative Timespan of Interest* was not applicable in this setup. The resulting TOIs for both venues were Saturday 13:30–15:30 and 17:30–19:30 for the start and end time, respectively.

In total, we observed 37 events over a time period of one year and ran the experiment on more than 1000 road segments in total within the vicinity of both venues. The classification was performed using an ANN in combination with a simple brute force parameter optimization technique. The results were evaluated by a leave-one-out cross-validation approach resulting in the following metrics: *precision*, *recall*, and *f-measure* for each road segment. The study has been published in [2]. In Wolfsburg, the road segments with the highest *f-measure* are those on the bridge directly in front of the venue (marker **1** and **2** in Figure 5.31). As this is the shortest way to reach the venue from the southbound direction, these results seem intuitively reasonable. What looks odd are the road segments at level **4**. They seem to point away from the venue. Are they wrongly classified? Marker **3** shows a parking lot that people frequently use to reach the venue. What is not visible in the image is the fact that during games, the area between markers **3** and **4** can also be used to park cars. Those road segments at marker **4** are the last possible way to reach those parking spots and are frequently used by visitors. The results therefore seem reasonable.

[4] e.g. `https://www.tomtom.com/de_de/drive/maps-services/live-services/`

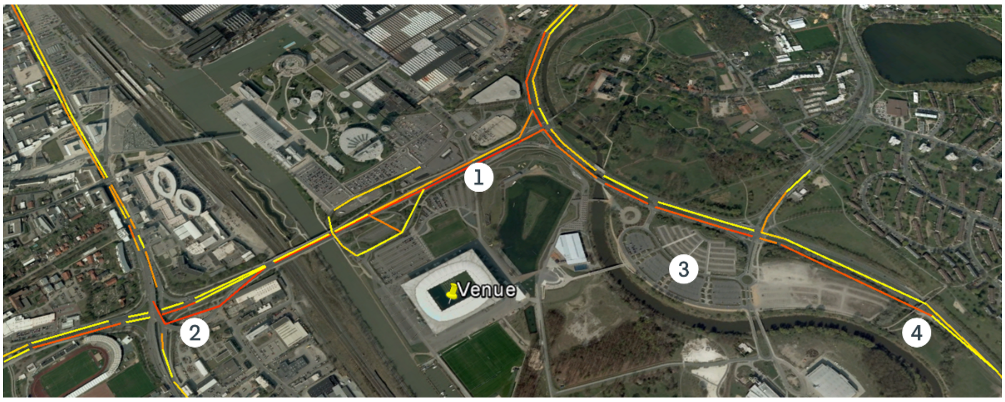


FIGURE 5.31: Venue: Volkswagen Arena in Wolfsburg. *f-measure* for a subset of the roads around the Volkswagen Arena on a color scale from yellow (*f-measure*=0) to red (*f-measure*=1). Source:[2]

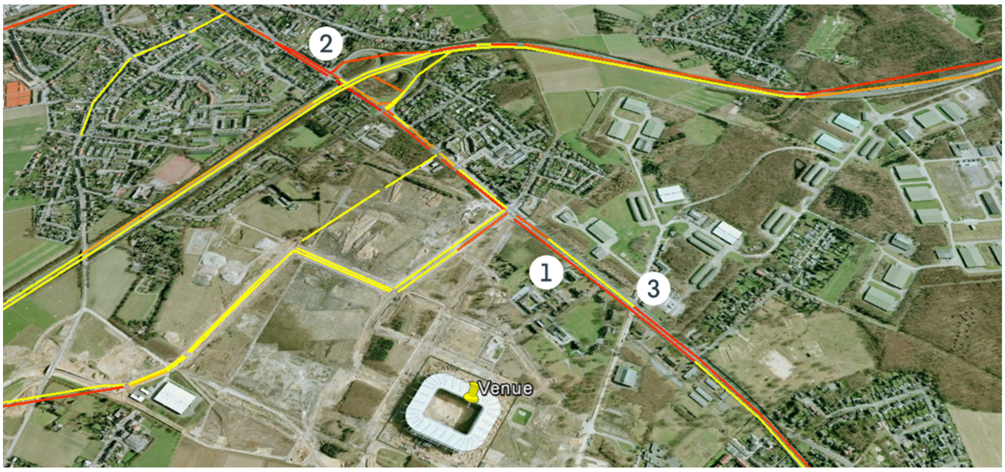


FIGURE 5.32: Venue: Borussia-Park in Moenchengladbach. *f-measure* for a subset of the roads around the Borussia-Park on a color scale from yellow (*f-measure* =0) to red (*f-measure*=1). Source:[2]

For Moenchengladbach the situation is similar. Figure 5.32 shows the results using the *f-measure* as metric. The road segments at marker **1** and **2** are pointing directly toward the venue and are rather obviously part of the *incoming route*. At marker **3**, we find an intersection that connects the incoming routes to a rural road that leads directly to the venue.

We have seen that *Incident* data can successfully be used for data analytics tasks. However, the data source itself requires significant preprocessing effort. The inability to retrieve coherent information about the traffic state at a road segment in combination with dynamic referencing methods make it extremely difficult to use.

5.4 Discussion and Conclusion

From the results above, we have seen that finding the spatial impact zone of events is possible by applying the presented classification approach on traffic data. For most of the road segments that were identified by the classifier as potentially affected by PSEs, we found an intuitive explanation that justified the result.

However, we have also seen the limitations of this approach. Although it shows good results for soccer games in the *flow study*, its performance lowered drastically for the "mixed" venues. For those venues, the results were less straightforward to interpret. None of the observed road segments showed an *f-measure* that was comparable to the results from the soccer stadia. However, at the same time, we have also not seen any false positives. Whenever a road segment got an *f-measure* above a reasonable threshold, we were able to find an explanation by which it could actually be affected by PSEs in that area. The threshold itself was dependent on the venue location, which was expected because the traffic network plays an important role. For each soccer venue, we were able to define a threshold that seemed reasonable for the given results. In future studies, we could reuse that threshold and build an automatic procedure to detect the affected spatial region.

Another interesting result is the performance of the *Incident study*. Although the information value of the data source is much lower, we were still able to detect affected road segment.

In summary, we can conclude three different statements from this chapter:

1. Traffic incident data is useful for analysis, up to a certain extent.
2. Soccer games are the most stable event category in our database.
3. PSEs vary drastically in their specific impact on traffic.

Statement 1 is of major importance. Throughout the literature, we find a huge collection of articles that describe the use of traffic information for a long list of use cases. So far, to the best of our knowledge, none of them (except [2]) has shown the use of *Incident data* for data analysis. At the same time, *Incident* *data* is used in for all types of navigation devices today and it is possible to retrieve this data in vast amounts. However, it comes with great challenges: 1) Its information value is limited (you only get information about drastically severe congestion). 2) As it is used in modern navigation devices, it mostly uses dynamic referencing formats such as OpenLR, which makes it impossible to retrieve consecutive information for fixed *Traffic Locations*. We have shown in this chapter that this data source can be used in certain cases, but that its limitations are severe.

Statement 2 is a direct result of our analysis. The more stable the event impacts are, the better the presented approach works. Our results for soccer games are, were congestion was observed at all, satisfyingly independent of the road network infrastructure. This implies a very consistent behavior of soccer games regarding their impact on traffic. Possible reasons for this are the same target group of people, probably similar mobility choices of those people, and a similar number of visitors for most of the games. This observation also possibly explains why current literature about the impact of PSEs on traffic mostly focus either on megaevents or soccer games only (e.g., [103, 104]).

Statement 3 is an extension of what we have already seen in previous chapters. In Chapter 4, we observed that the event impact of different categories varies . However, from our spatial data analysis, we can conclude that those variations are much more drastic on the road segment level. This behavior cannot be explained by the category of events only, as events of the same category might also show very different behavior (see the soccer examples shown in Figure 5.6). To further analyze their impact on specific road segments, more information about the event and/or the road network is required. In the following chapters, we will analyze the value of additional information about events in detail.

Chapter 6

Social Media for Event Impact Explanation

In the previous chapters, we have seen that the impact of PSEs on traffic varies drastically. Whereas soccer games tend to show a rather stable behavior in this regard, other categories highly fluctuate in the amount of related traffic congestion. But what is the describing attribute that allows us to predict whether an event is going to have a huge impact on traffic or not? How can we know in advance that a concert of *QUEEN* results in more traffic congestion than a concert of *Miley Cyrus* (example taken from Figure 5.30)?

Possible intuitive explanations for varying traffic impact have been discussed in Chapter 4. Although that chapter mostly focused on influencing factors that can be expressed directly, such as *venue location*, *daytime variation*, *day of week variation*, and *event category* we now focus on latent measures of events. Such measures describe the overall popularity of an event (*attractiveness*), the choice of modality of visitors, and much more. The focus of this chapter is to analyze the possibilities to find a describing measure for the expected *traffic impact size* of events.

Unfortunately, there is no such measure in existence today. What does exist (in vast amounts) is information about the *online popularity* of events [63]. The literature has shown that *online popularity* measures, such as number of results in Google or number of likes on Facebook, can be used to describe the attractiveness of events to a certain extent, at least in other domains (see [4, 6]). Although this approach seems appealing, the question remains as to whether the obtained popularity measures are suitable predictors for the impact on traffic as well.

This part of the thesis aims to answer this question in detail by analyzing traffic and event data around the two mixed venues from the previous chapter:

1. Mercedes-Benz Arena Berlin, Germany
2. LANXESS Arena Cologne, Germany

This chapter is structured as follows: We first present our data collection process and describe the resulting dataset in detail. We then evaluate the relevance of the collected online metrics for the observed traffic impact in the next section. Based on these results, in the next section, we build different prediction models and discuss our results. In the following section we propose an alternative approach for the *spatial challenge* based on our findings. This chapter closes with a discussion of our results, lessons learned and open issues.

6.1 Dataset Description

The dataset used in this research consists of two distinct sets of information: event information from online sources and traffic measures during event happenings. Both are presented in the following sections.

6.1.1 Online Metrics

For our research, we applied a very similar strategy as that in [4, 6] and collected six different online metrics: existence of a Wikipedia page *HasWikipediaPage*, number of likes on Facebook *FacebookLikes*, number of Facebook talks from the Bing API[1] *FacebookTalks*, number of results in Bing *BingHits*, number of followers on Twitter *TwitterFollower*, and the existence of a Youtube page *YoutubePresent*. We collected these metrics via different APIs from the web for all events that we observed between 05/2014 and 05/2015 (the timespan for which we have traffic data).

[1] https://www.microsoft.com/cognitive-services/en-us/bing-web-search-api

Although this data collection was done automatically, we had to check parts of the datasets manually for quality assurance. Occasionally, the APIs did not result in any meaningful values or did not return any data at all. The reasons for that were mostly ambivalent artist names or spelling mistakes in our datasets.

Events with more than one artist need special preprocessing. A possible way to handle those situations could be to only collect information for the main act or the most famous artist. As our dataset contains many sporting events, in which usually two teams compete against each other, and the most prominent team is hard to identify, we decided to handle such cases differently. We collected online metrics for all teams or artists and merged the information using the *mean* of the collected values. Using the *mean* as aggregation instead of the *sum* reduces the chance to overestimate the size of an event, especially for sporting events where only limited capacities for the guest team are available anyway. The shown approach, nevertheless, could be easily adapted to use different aggregation functions.

A challenge arises from the potentially time-dependent variably of event popularity. A singer might be highly popular at a certain point in time, but different occurrences might affect this popularity in a positive or negative way. One possibility to alleviate these affects would be to collect data right before the event happens. This would of course not consider that tickets are usually bought far ahead of an event, at least for larger events. Another possibility would be to track the popularity over time and give it a standardized value over the entire timespan. Because many of the collected attributes (e.g. likes on Facebook) do not come with a history, this approach would necessitate tracking all events in real time. For that, prior knowledge of all upcoming events to be analyzed would be needed. As we ran our analysis on historic traffic data, none of the mentioned approaches were feasible for us. In our case, we collected all information in a single shot in June 2016 (past our study time frame). This gives us a snapshot of the popularity of events at this particular point in time, but of course does not allow us to react to ups and downs of artist popularity. We leave it to future work to re-run this study based on more real-time online sources.

Until this point, all the described metrics were collected from social media sources. We added two additional features based on our findings in Chapter 4: the category of an event *EventCategory* and a categorization of weekdays into five distinct classes (Mo–Th, Fr, Sa, and Su) *DayCluster*, as introduced in Chapter 4.

6.1.1.1 Overview

For the Mercedes-Benz Arena we collected information for 48 events for which event and traffic information was available to us. From these events, 33 artists had their own public Youtube[2] account and all of them were mentioned at Wikipedia[3]. To get a rough estimate about the distribution of the data figure 6.1 gives an overview about the collected metrics. They show large differences between the collected datasets. For example, the *TicketSales* range from 0 for some sporting events, where the industrial provider apparently did not sell any tickets for, up to 11774 for a concert of *Herbert Grönemeyer*. At the same time, *FacebookLikes* range from 0 up to almost 30 million (a *WWE* wrestling match).

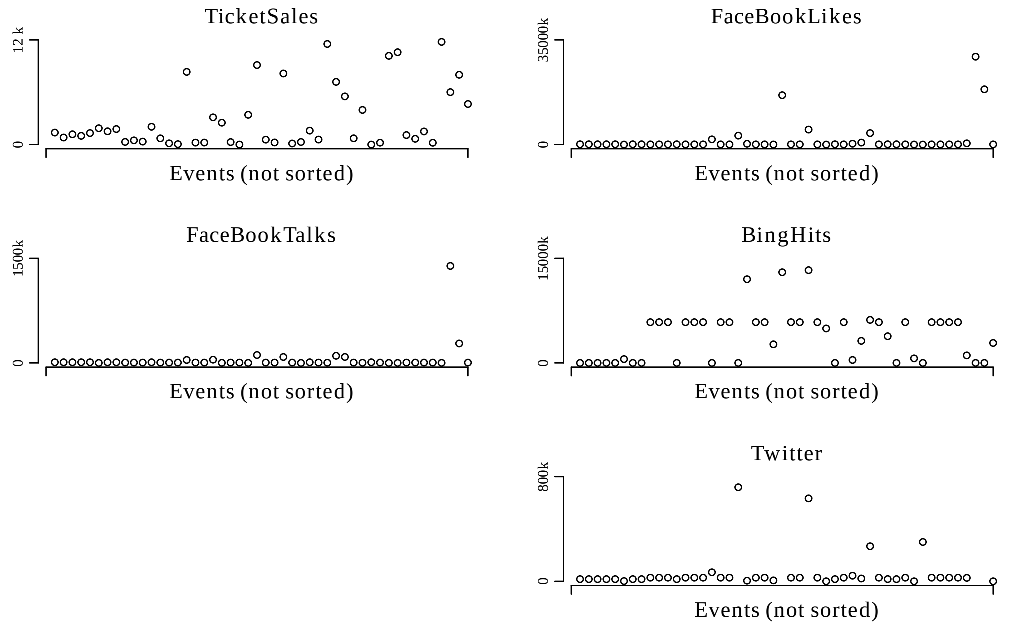


FIGURE 6.1: Venue: Mercedes-Benz Arena. Online metric overview.

Similar behavior can be found for the LANXESS Arena. For this venue we collected data for 33 events. For 29 events we found a *Wikipedia* page, and 23 events were represented on *Youtube*. The overview of the other collected metrics is shown in figure 6.2. The online metrics at this venue also vary significantly. Es-

[2] http://www.youtube.com
[3] https://www.wikipedia.org/

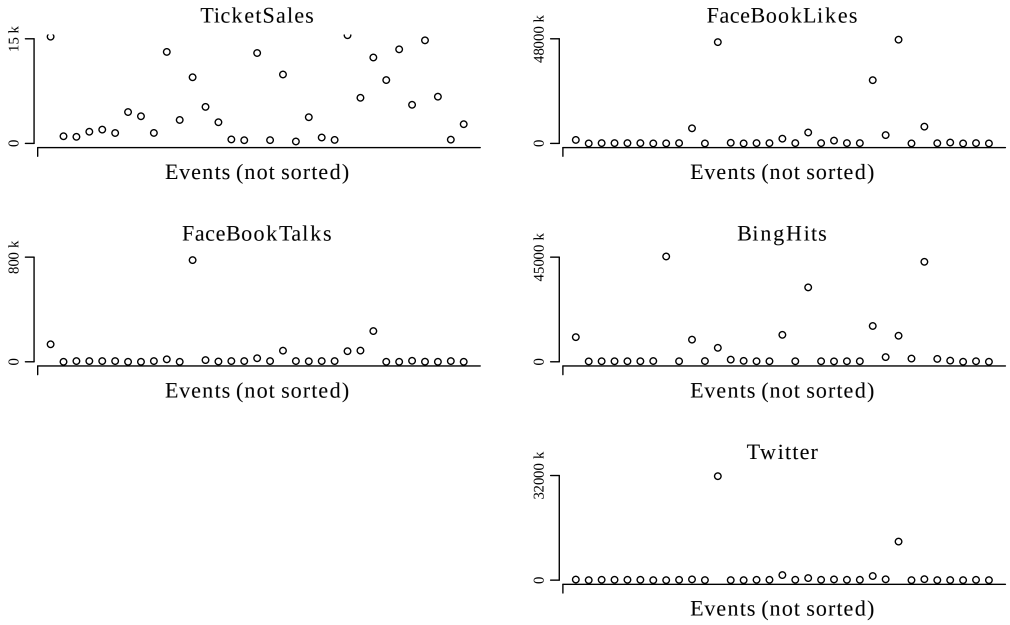


FIGURE 6.2: Venue: LANXESS Arena. Online metric overview.

pecially *Twitter* ranges from 8000 for a sporting event and approximately 32000 k for a *Miley Cyrus* concert.

6.1.2 Traffic Measures

The impact on traffic can be measured in various ways. In the previous chapters, we mostly focused on one single metric to describe the traffic state on the road. Following this approach, in this study we defined the impact on traffic using two different measures: the *mean* and *variance* of the free flow percentage (FFP) over a predefined timespan. We followed the same approximation of the event time window that we introduced in Chapter 4 and analyzed the traffic during the 2 h time window before the event start.

6.2 Feature Relevance

As a first step in this study we analyzed the relevance of our collected event metrics for the observed traffic states. For that kind of study general *feature selection algorithms (FSA)* are applicable (see [105] for an extensive overview about FSA in general).

FSAs are usually separable into those that address two distinctive subproblems: *minimal optimal problem* and *all relevant problem*. A *minimal optimal problem* in feature selection focuses on finding the minimum set of attributes that result in the best classifier for the problem state. An all relevant problem, however, focuses on identifying all relevant features that describe a given problem state [106]. Our application belongs to the second category.

There is a long list of different algorithms that focus on finding all relevant features in the literature [107–110]. Many of them are based around ensembles of decision trees. They mostly work quite similarly to each other, differing in their performance and the implementation of the statistical test [106].

For our analysis, we picked the Boruta (see [111, 112]) algorithm using the Boruta R package [113]. Boruta is based on a random forest classifier and its essential idea is rather straightforward: For each attribute, a *shadow* attribute is generated by permuting its values randomly. These *shadow* attributes get merged to the system. If the importance of a variable exceeds the importance of its *shadow* variable after building the classifier the original attribute gets considered as important.

In our scenario, we used all attributes described in 6.1.1 and ran the study once for each traffic measure described in 6.1.2.

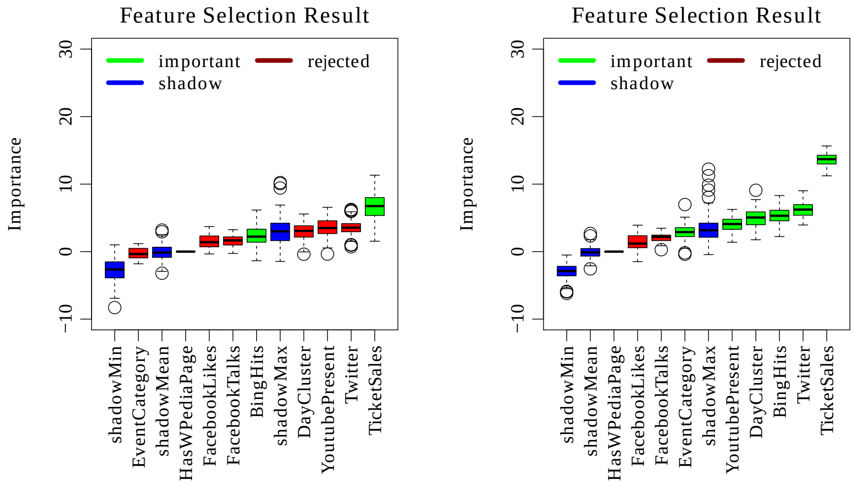
6.2.1 Experimental Setup & Results

We split our analysis into two separate parts: First, we analyzed the importance of the selected attributes on traffic observations within three fixed radii (500, 1000, and 2000 m) around the venues. Second, we took the results from the spatial study (see Chapter 5) into consideration and applied the same method using the *FreeFlowPercentage* (FFP) on *TrafficLocations* (TL) that showed the highest event-related impact around the venues.

For all the experiments, we created a dataset containing all online metrics and one of the traffic measures as a target variable. Because the used attributes show very different distributions, we normalized all of them to a [0,1] range before running the Boruta algorithm. The results are presented per venue in the next sections. We also collected information about feature relevance for a venue in Hamburg, Germany. These results are shown in Appendix B.

6.2.1.1 Mercedes Benz Arena Berlin

Some characteristics of the Mercedes-Benz Arena in Berlin have already been given in the previous chapter. We have seen that it is located directly in the inner city of Berlin (city east), surrounded by other venues, and shows high variations of observed traffic during events. At first, we ran the study for the mentioned three different radii. The results are shown in Figures 6.3, 6.4, and 6.5.



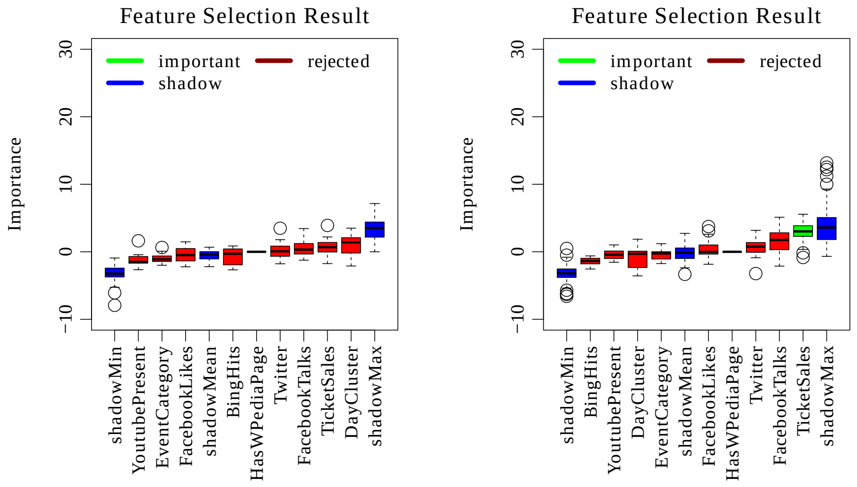
(I) Traffic metric=*mean*.

(II) Traffic metric=*variance*.

FIGURE 6.3: Venue: Mercedes-Benz Arena Berlin. Attribute relevance ranking for traffic observations within 500 m radius. y-axis: Boruta importance measure.

The results for the 500 m radius using the *mean* as traffic measure (see Figure 6.3i) show that only two features that were confirmed as important: *TicketSales* (median *Z score*: 6.74) and *BingHits* (median *Z score*: 2.22). However, the overall median *Z score* of *TwitterFollower*, *YoutubePresent*, and *DayCluster* are rather similar to the selected measures. For the *variance* as a traffic measure (see Figure 6.3ii), the results change. *TicketSales* is by far the one with the highest median *Z score* of 13.69, and in total, five other features also get marked as important by the algorithm: *Twitter* (median *Z score*: 6.23), *BingHits* (median *Z score*: 5.31), *DayCluster* (median *Z score*: 5.01), *YoutubePresent* (median *Z score*: 4.08), and *EventCategory* (median *Z score*: 2.88).

In total, the overall median *Z score*s of all attributes, except *TicketSales* for the *variance* as a traffic measure, are rather low. However, it stands out that most of the attributes that show any significant importance to the observed traffic variables are event-related. In particular, *TicketSales* shows a noticeable predictive power for the traffic situation that is higher than attributes that describe the regular traffic variations (e.g., *DayCluster*). This leads to the assumption that in the selected radius around this venue, events show an impact and change traffic behavior above its regular fluctuations.

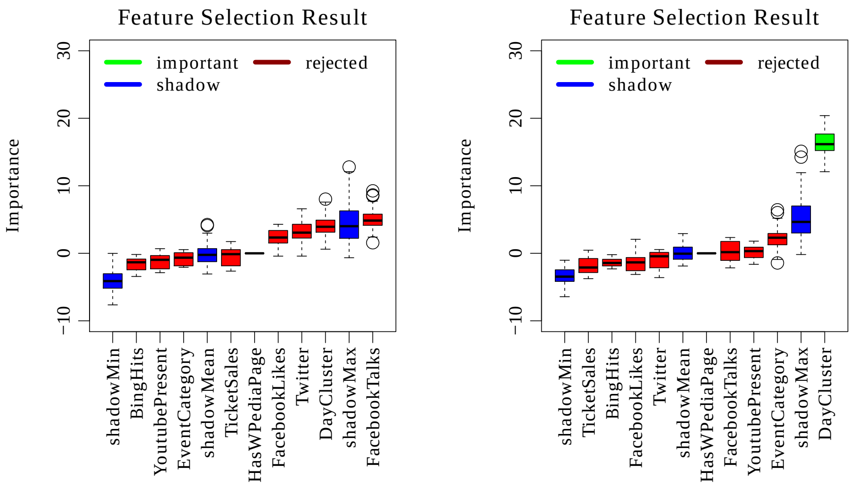


(I) Traffic metric = *mean*. (II) Traffic metric = *variance*.

FIGURE 6.4: Venue: Mercedes-Benz Arena Berlin. Attribute relevance ranking for traffic observations within 1000 m radius. y-axis: Boruta importance measure.

Within the 1000 m (see Figure 6.4) radius, all analyzed attributes lose significance. A possible explanation for this is that the radius includes more segments that are not affected by events, which would lead to a reduced significance of event-related attributes. At the same time, it seems that the radius also contains insufficient non-affected road segments to emphasize day-to-day traffic variables. This effect changes for the 2000 m radius up, to a certain extent (see Figure 6.5). For the *variance* as a traffic measure, we see a strong increase in the *DayCluster* attribute, which indicates that the regular behavior during different days of the week becomes dominant.

In summary, it seems that there is event-related traffic around the Mercedes-Benz Arena in Berlin, which can at least partly be described by some of the selected event-specific attributes. To alleviate other influencing factors and to focus on the event-specific parts, we include the results from our previous study in Chapter 5 in the analysis. We have seen that some road segments showed

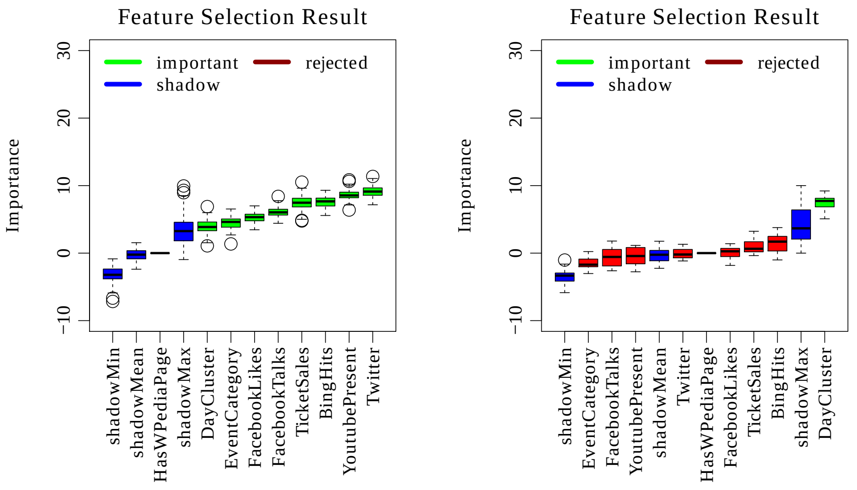


(I) Traffic metric = *mean*.

(II) Traffic metric = *variance*.

FIGURE 6.5: Venue: Mercedes-Benz Arena Berlin. Attribute relevance ranking for traffic observations within 2000 m radius. y-axis: Boruta importance measure.

a higher tendency to change their traffic patterns during events than others. Those segments have been discussed in Chapter 5.2.2.6. For our analysis, we picked the most highly correlated road segment from the previous study (TL 13/5/26927), as shown at marker **A** in Figure 5.23i. For this TL we collected traffic data for all events and ran the same approach as presented in the section above. The results are shown in Figure 6.6.



(I) Traffic metric: *mean*.

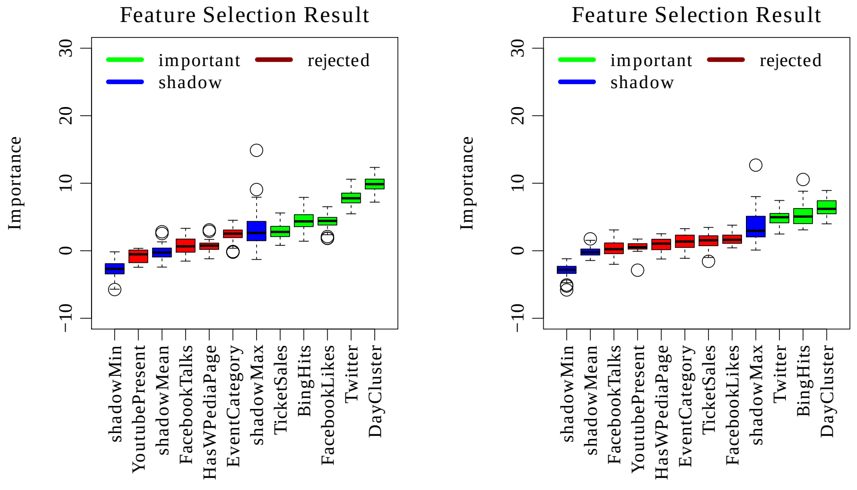
(II) Traffic metric: *variance*.

FIGURE 6.6: Venue: Mercedes-Benz Arena Berlin. Attribute relevance ranking for TL 13/5/26927 directly in front of the venue (see marker A in Figure 5.23i). y-axis: Boruta importance measure.

For this TL, nearly all selected attributes become relevant when considering the *mean* as a traffic measure. The only exception is the presence of a Wikipedia page. This can easily be explained, as we found a Wikipedia page for all events that happened in this venue during the timeframe of our analysis. It therefore does not contain any relevant information and can be ignored. For all the other attributes, event-specific ones dominate over the non-event-specific ones. The event-specific attributes with the highest median *Z score* is *TwitterFollowers*, with a score of 9.11. The first non-event-specific attribute is *DayCluster*, with a mean *Z score* of 4.6. For the *variance* as a traffic measure, only one attribute was declared as relevant: *DayCluster*. This finding is interesting, as it indicates that the overall traffic density on this road segment is connected to the events, while its variation seems to be based more on daily traffic pattern fluctuation.

6.2.1.2 LANXESS Arena Cologne

Similar to the Mercedes-Benz Arena in Berlin, the LANXESS Arena in Cologne is close to the city center and has similar properties. For this arena, we analyzed 47 events that happened within the timespan of our analysis. The results of our studies for the mentioned three different radii are shown in Figure 6.7, 6.8, and 6.9.

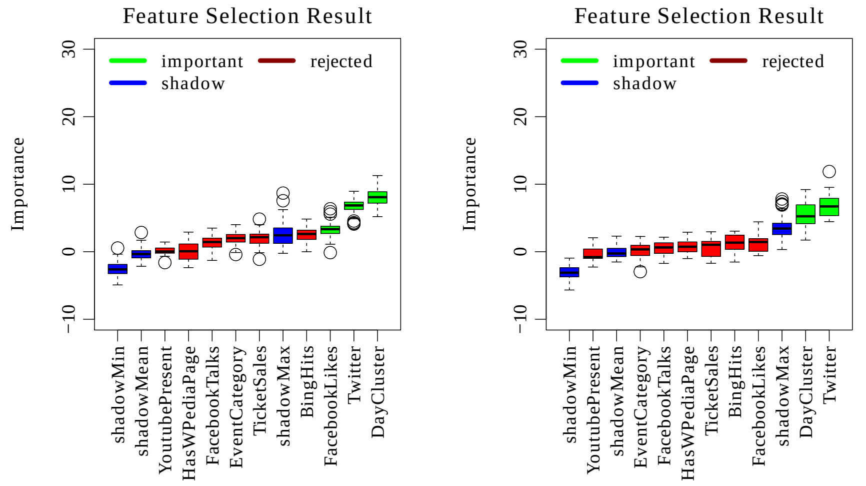


(I) Traffic metric = *mean*.

(II) Traffic metric = *variance*.

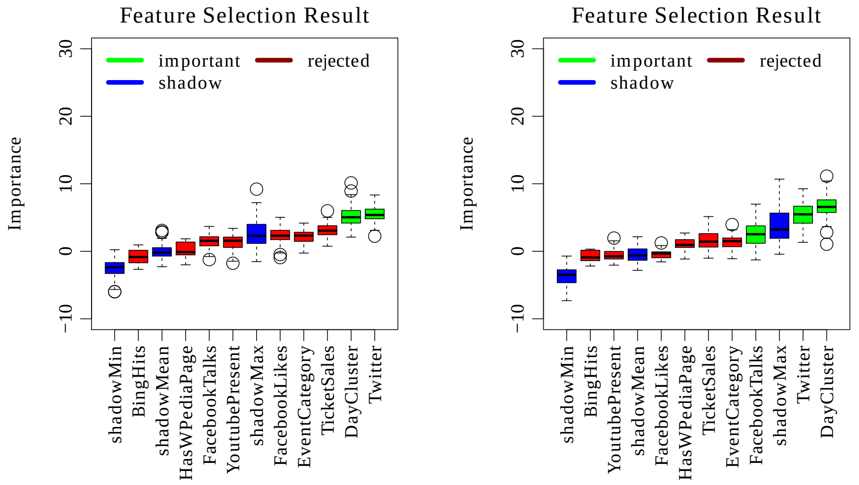
FIGURE 6.7: Venue: LANXESS Arena Cologne. Attribute relevance ranking for traffic observations within 500 m radius. y-axis: Boruta importance measure.

For the 500 m radius using the *mean* traffic measure, five attributes were identified as relevant: *DayCluster*, *TwitterFollower*, *FacebookLikes*, *BingHits*, and *TicketSales*. Of these *DayCluster*, with a mean *Z score* of 9.87 and *TwitterFollower*, with a mean *Z score* of 7.78 are the most relevant. For the *variance* as a traffic measure (see Figure 6.7ii) only three variables are marked as relevant: *DayCluster*, *BingHits* and *TwitterFollower*. It is interesting to observe that the results between the *mean* and *variance* traffic measure do not look drastically different. Compared to the results for the 500 m radius around the Mercedes-Benz Arena, where the relevant attributes differ significantly, the choice of traffic measure does not significantly affect the result. A possible explanation could be that events may lead to a higher variance in traffic than around the Mercedes-Benz Arena. Another interesting fact is the dominance of the *DayCluster* for both traffic measures. It seems that around this particular venue, the traffic variation due to day-to-day traffic is more severe than around the Mercedes-Benz Arena. Within the 1000 m radius (see Figure 6.8), *DayCluster* and *TwitterFollower* remain dominant for both traffic metrics and most of the other measures become rejected (except for *FacebookLikes* but with a very low median *Z score*). This trend persists for the 2000 m radius (see Figure 6.9). These results lead to interesting observations: 1) *TicketSales* are not that relevant compared to the Mercedes-Benz Arena. 2) The two dominant features stay the same for all radii. 3) *mean* and *variance* traffic measures do not differ as much as for the Mercedes-Benz Arena.



(I) Traffic metric = *mean*. (II) Traffic metric = *variance*.

FIGURE 6.8: Venue: LANXESS Arena Cologne. Attribute relevance ranking for traffic observations within 1000 m radius. y-axis: Boruta importance measure.

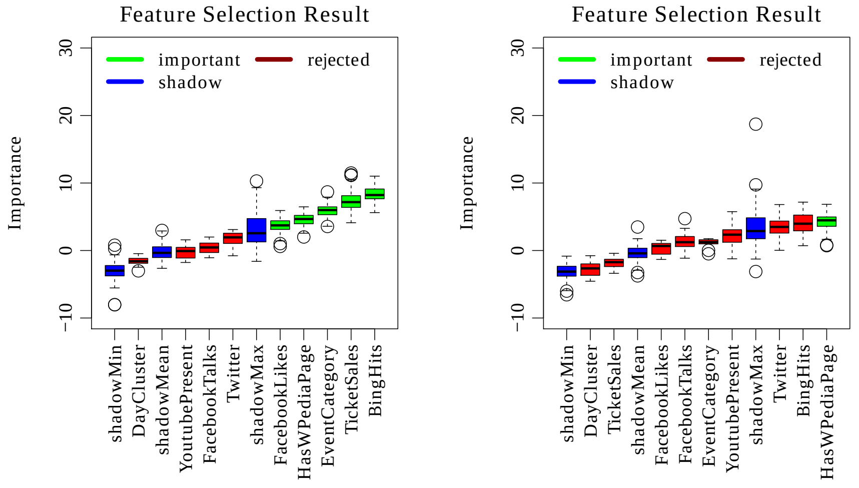


(I) Traffic metric = *mean*.

(II) Traffic metric = *variance*.

FIGURE 6.9: Venue: LANXESS Arena Cologne. Attribute relevance ranking for traffic observations within 2000 m radius. y-axis: Boruta importance measure.

Again, we ran the study for the most affected TL from the previous chapter. In this case, we picked TL at marker **A** in Figure 5.27. The results are shown in Figure 6.10. Similar to the TL example in Berlin, this one shows significant importance of most of the event-related attributes. An interesting observation is the lower importance of *TwitterFollowers*. Whereas this attribute was dominant for the *mean* traffic measure for all different radii, it is rejected by the Boruta algorithms for this particular TL.



(I) Traffic metric: *mean*.

(II) Traffic metric: *variance*.

FIGURE 6.10: Venue: LANXESS Arena Cologne. Attribute relevance ranking for TL 13/2/39689 directly in front of the venue (see marker A in Figure 5.27). y-axis: Boruta importance measure.

6.3 Impact Prediction

The presented feature relevance results show a significance in event-related attributes. The remaining question is, how important are these attributes to explain the observed congestions? To gain initial insights to this question we present, in this section, two different prediction approaches.

The first uses simple linear regression to evaluate the usefulness of different sets of information for the prediction result. The goal is to evaluate if the information sets, that have been identified as relevant, hold enough information to create reliable predictions.

The second approach focuses on classification-based impact prediction based on the same types of information sets. The goal for this approach is to evaluate if the given information sets are useful to classify events into certain severity categories.

In both studies, we limit our datasets to publicly available online information; therefore, *TicketSales* information is excluded.

With both approaches we ran experiments for the same TLs around the venues as in the studies presented earlier in this chapter. The experimental setups and results are described in the following sections. The focus was on generating a first impression of the usefulness of the mentioned information sources. Detailed studies using various modeling strategies will be left for future work.

6.3.1 Regression

Following a similar approach as presented in [7, 63], we used linear regression (LR) models to predict the *mean FreeFlowPercentage* for the selected road segments. To evaluate the explanatory power of different information sources, we developed models that differ significantly in the information they employ.

The *DayToDay* model focuses on information about the specific day of the week only. It predicts the observed *mean(FFP)* by a combination of *day of week* attribute classes. These classes are the same as described in previous chapters: Monday–Thursday (Mo–Th), Friday (Fr), Saturday (Sa), and Sunday (Su). For the LR, the model is given as:

$$\begin{aligned} meanFFP(e) &= \beta_0 + \beta_{mo-th}X_{mo-th}(e) + \beta_{fr}X_{fr}(e) + \beta_{sa}X_{sa}(e) + \\ &\quad \beta_{su}X_{su}(e) + \epsilon(e) \end{aligned}$$

where e is the event, $\epsilon(e)$ is the error term, $X_{mo-th}(e)$ is 1 if e is on a weekday between Monday and Thursday, $X_{fr}(e)$ is 1 if e is on a Friday, $X_{sa}(e)$ is 1 if e in happening on a Saturday, and $X_{su}(e)$ is 1 if e is happening on a Sunday.

The second model, the *CategoryModel* is based on the event category only. Depending on the location, we model all event categories as individual attributes. The model is given by:

$$meanFFP(e) = \beta_0 + \beta_{c1}X_{c1}(e) + ... + \beta_{c(N)}X_{c(N)}(e) + \epsilon(e)$$

where X_{c1} is 1 if event e is of that specific $c1$ category (e.g., concert), N is the number of categories, and $\epsilon(e)$ is the error term.

The third model, the *SocialMediaModel*, includes the social media metrics discussed earlier in this chapter. We adapt our model of each TL separately to those social media metrics that were declared as relevant from the previous study. The model is given as:

$$meanFFP(e) = \beta_0 + \beta_{sm1}X_{sm1}(e) + ... + \beta_{smM}X_{smM}(e) + \epsilon(e)$$

where X_{sm1} is the first relevant social media attribute (e.g., Twitter), M is the number of selected social media sources, and $\epsilon(e)$ is again the error term.

To evaluate the performance of the presented models, we also developed a baseline model called *Baseline*. That model simply used the average of all observed event traffic situations in our dataset as a prediction.

6.3.1.1 Experimental Setup & Results

We ran the study for both TLs from the previous section. $TL_{MercedesBenz}$ is the one shown in Figure 6.6 with the TMC code 13/5/26927. For Cologne, we selected the $TL_{LANXESS}$ with the TMC code 13/2/39689 (shown in Figure 6.10 at marker A) as it is one of the most affected TLs around the LANXESS Arena.

For both TLs, we selected the same dataset as for the previous study in this chapter. A leave-one-out cross-validation was performed for each road segment separately. To fit the model to our data, we followed the method presented in [7] and selected a fitting via linear least squares. Using the specific models, we predicted the *mean(FFP)* from a combination of the shown parameters for each event. The performance was evaluated using the mean absolute error (MAE).

For the $TL_{LANXESS}$, the event categories were *concert*, *sport*, and *musical & show*. The social media attributes that have been selected as being important from the previous study were *BingHits*, *HasWikipediaPage*, and *FacebookLikes*. However, for the LR, our analysis showed that the best model was derived using *BingHits* only.

For the $TL_{MercedesBenz}$, the event categories were *concert*, *culture*, *sport*, and *musical & show*. This time, the social media attributes that have been selected as being important from the previous study were *Twitter*, *YoutubePresent*, *BingHits*, *FacebookLikes*, and *FacebookTalks*. For this TL, our analysis showed the best performance for the top three attributes *BingHits*, *Twitter*, and *YouTubePresent* and those were selected for our following results.

The results for the LANXESS Arena are shown in Figure 6.11. They show a

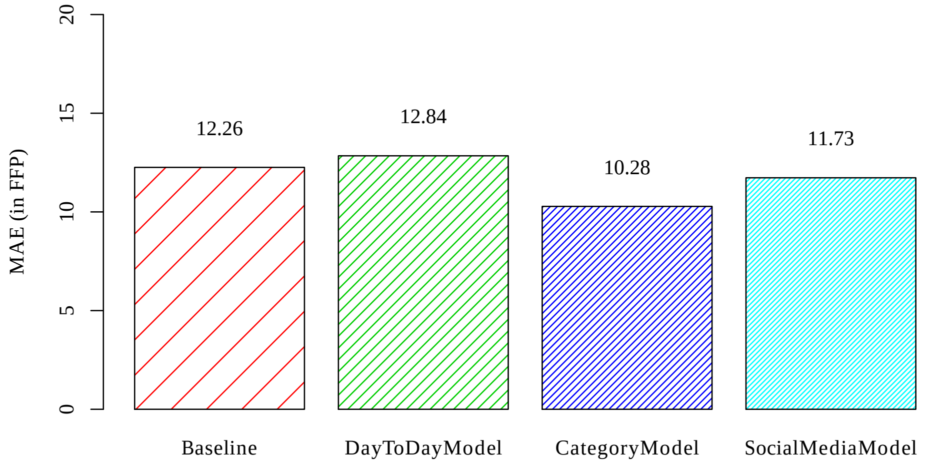


FIGURE 6.11: Model benchmark for the $TL_{LANXESS}$.

.

similar performance for all models, although the *CategoryModel* and the *SocialMedia* model showed marginally better performance than the *Baseline* and the *DayToDayModel*.

The regression results for the different models are shown in Table 6.1. All models show significantly low multiple R^2 values, especially the *DDM* and *SM*. For the *DDM* information about the *day of week* did not seem to be relevant for the resulting traffic situation, shown by the very high *p-value*. The same applies for the *isSport* category for the CM. Apparently, the use of day classes, event categories and information from online metrics was insufficient to create an accurate prediction model for the resulting traffic disruptions.

TABLE 6.1: Regression results for the $TL_{Lanxess}$.

Model	Factor	Estimate	Std. Error	p-value	Multiple R^2
DDM	(Intercept)	-22.5833	6.6363	0.0023	2.83e$-$05
	Mo–Th	-0.1925	7.3842	0.9794	
CM	(Intercept)	-18.6944	4.9004	0.0009	0.373
	IsConcert	-13.9419	6.0920	0.0316	
	IsSport	5.3565	6.3264	0.4059	
SM	(Intercept)	-19.4475	2.9820	0.0000	0.1849
	BingHits	-25.4437	10.9034	0.0283	

DDM: DayToDayModel, CM:CategoryModel, SM:SocialMediaModel

A similar effect can be seen for the *Mercedes-Benz Arena* in figure 6.12. Again, the performance of all models seems very similar, of which the *DayToDay* model and the *SocialMedia* model perform slightly better than the others.

The regression results for the different models are shown in Table 6.2. Again, the regression results show similar Multiple R^2 values as for the $TL_{LANXESS}$, and the overall attribute relevance for the different event specific information sources is rather low. For the CM all attributes show *p-values* that indicate no statistical significance. For the *SM* only the presence of a Youtube page seems to yield valuable information.

The results show that, for both TLs, additional information about the events did not increase the performance of the selected model significantly. Does this mean that the phenomenon is not predictable at all?

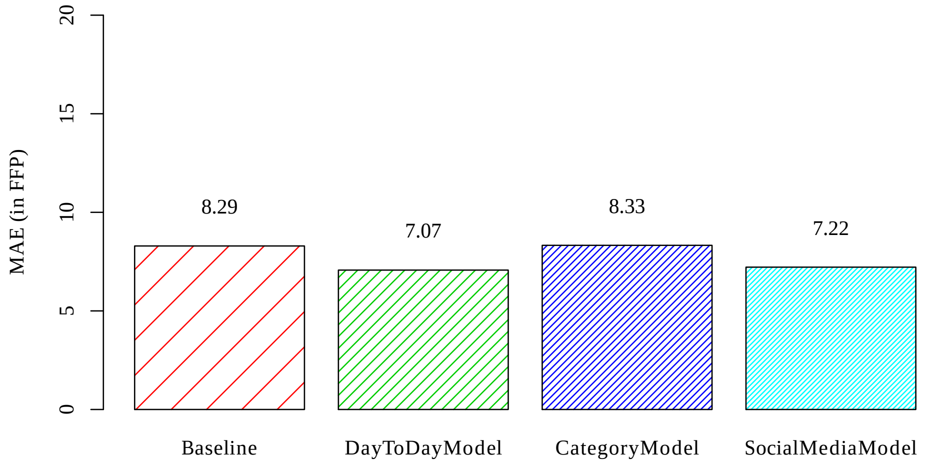


FIGURE 6.12: Model benchmark for the $TL_{Mercedes_Benz}$.

TABLE 6.2: Regression results for the $TL_{Mercedes-Benz}$.

Model	Factor	Estimate	Std. Error	p-value	Multiple R^2
DDM	(Intercept)	−15.0750	3.9052	0.0004	0.2843
	Mo–Th	−1.9207	4.2285	0.6521	
	Fr	11.7990	4.9782	0.0226	
	Sa	−3.5917	6.3772	0.5764	
CM	(Intercept)	−17.1806	5.5498	0.0035	0.1328
	IsConcert	−3.5028	6.3277	0.5829	
	IsSport	5.1028	5.8206	0.3858	
	IsCulture	2.3056	8.7749	0.7941	
SM	(Intercept)	−5.4865	2.4774	0.0324	0.3107
	Twitter	−0.6060	8.1965	0.9414	
	BingHits	−11.4913	6.8525	0.1012	
	YoutubePresent	−11.1108	2.8728	0.0004	

DDM: DayToDayModel, CM: CategoryModel, SM: SocialMediaModel

Of course, different modeling approaches could possibly improve the performance. Nonetheless, the predicted attribute *mean(FFP)* also poses great challenges. The FFP is already the result of an aggregation of traffic information that has been preprocessed from the traffic data provider. During that aggregation process, an information loss is very likely. Additionally, the *meanFFP* metric has been calculated over varying time windows, which also adds noise to the attribute. As a conclusion, the low prediction performance is probably also highly affected by the prediction metric itself.

6.3.2 Classification

To analyze whether we can use the collected event information to classify events into certain severity categories, we first need to create those severity labels. One possibility would be a manual classification based on observations, but this solution would be very time-consuming if the approach were employed for a larger study. To ensure a degree of generalizability of our approach, we create those labels automatically using a clustering technique. We cluster the observed *meanFFP* into three different clusters, and assign a severity level depending on the cluster centers.

For the modeling part, we define the same classes of models based on the same information source as for the regression part: *DayToDay*, *Baseline*, *Category*, and *SocialMedia* model. In this study, the models are trained using standard classification trees.

6.3.2.1 Experimental Setup & Results

We create the clusters for the severity level assignment using the *k-means* clustering algorithm (see [114, 115]). As most of today's state-of-the art navigation systems use three classes to describe traffic situations (e.g., high, medium, and low) we follow that convention and select a *k-value* of 3. This results in three separated groups that we associate with one of the severity levels, based on the cluster centers. The result of the clustering for both TLs can be seen in Figure 6.13 and 6.14. They show the artist name and the *meanFFP* of the event, together with a color code for the assigned level. As an example, Figure 6.13 shows that the concerts of *QUEEN*, *Bryan Adams*, and *Usher* belong to the same cluster of events with a high impact on the traffic situation on that road segment.

For the classification part, we trained a classification tree on the same input vectors as for the LR models in the previous sections, but picked as a dependent variable the resulting cluster ID from the *k-means* algorithm.

The results are shown as the confusion matrices of the different models in Table 6.3 and Table 6.4. For the Mercedes-Benz Arena in Berlin, the results show a

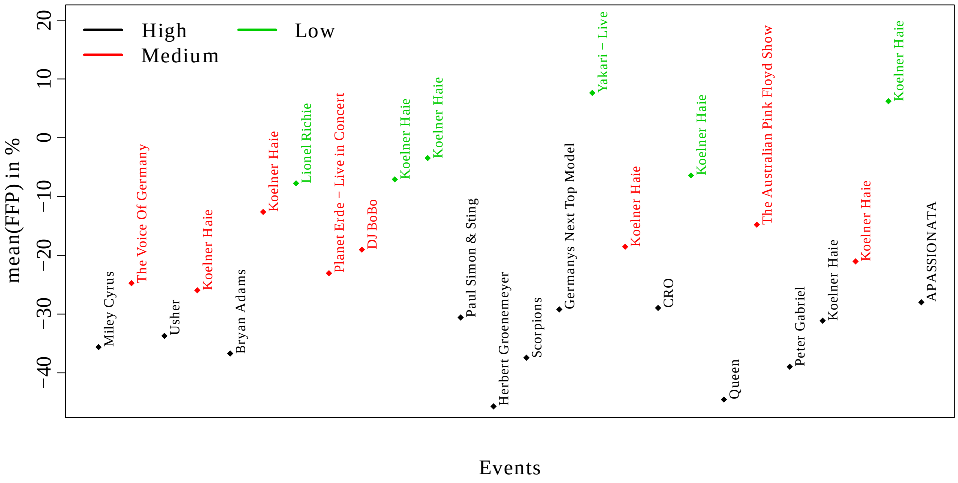


FIGURE 6.13: Clustering of the mean *FreeFlowPercentage* into three clusters at the $TL_{LANXESS}$. High: high traffic disruption, Medium: medium traffic disruption, Low: low traffic disruption.

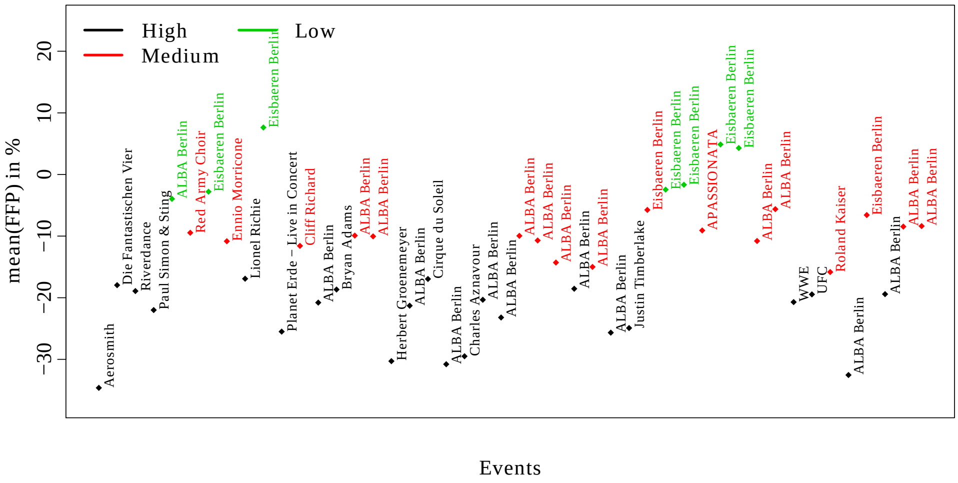


FIGURE 6.14: Clustering of the mean *FreeFlowPercentage* into three clusters at the $TL_{Mercedes-Benz}$. High: high traffic disruption, Medium: medium traffic disruption, Low: low traffic disruption.

similar performance of the *DDM* and *SM* models. Both models result in similar *f-measures* for the *High* class of 0.67 and 0.69, respectively. The *CM* model underestimates the traffic disruptions of events (belonging to the *High* category) almost half of the time and shows an *f-measure* of 0.34 for that specific category. At the same time, it wrongly classifies all events belonging to the *Low* category as events with *High* impact. This behavior can be explained by considering Figure 6.14. It shows that events of the same category (which is the only input for the CM model) belong to very different impact classes. For instance, games of *Alba Berlin* (basketball) appear in all three clusters.

TABLE 6.3: Classification Results for $TL_{Mercedes-Benz}$

Model	Cluster	High	Medium	Low	Precision	Recall	F1
Baseline	High	0	0	0	0	0	NaN
	Medium	22	7	17	0.15	1	0.26
	Low	0	0	0	0	0	NaN
DDM	High	17	3	9	0.59	0.77	0.67
	Medium	0	0	0	0	0	NaN
	Low	5	4	8	0.47	0.47	0.47
CM	High	8	0	17	0.32	0.36	0.34
	Medium	0	0	0	0	0	NaN
	Low	14	7	0	0	0	NaN
SM	High	15	2	4	0.71	0.68	0.69
	Medium	2	5	6	0.38	0.71	0.5
	Low	5	0	7	0.58	0.41	0.48

DDM: DayToDayModel, CM: CategoryModel, SM: SocialMediaModel

TABLE 6.4: Classification Results for $TL_{LANXESS}$

Model	Cluster	High	Medium	Low	Precision	Recall	F1
Baseline	High	0	0	0	0	0	NaN
	Medium	12	6	8	0.23	1	0.375
	Low	0	0	0	0	0	NaN
DDM	High	12	6	8	0.46	1	0.63
	Medium	0	0	0	0	0	NaN
	Low	0	0	0	0	0	NaN
CM	High	9	1	1	0.82	0.75	0.78
	Medium	0	0	0	0	0	NaN
	Low	3	5	7	0.46	0.88	0.61
SM	High	8	0	0	1	0.67	0.8
	Medium	0	0	0	0	0	NaN
	Low	4	6	8	0.44	1	0.62

DDM: DayToDayModel, CM: CategoryModel, SM: SocialMediaModel

However, the fact that the *SM* model performed better is an indicator that information from *SocialMedia* helps to differentiate those events around that venue, at least up to a certain extent.

For the LANXESS Arena, the results show better *f-measures* for the *CM* and *SM* models compared to the other models, and compared to the models at the Mercedes-Benz Arena. The *CM* model shows a *precision* of 0.82 for the *High* class and a *recall* of 0.75. The *SM* model even shows a *precision* of 1.0 for the *High* class but at 0.67, its *recall* is lower than that of the *CM*. In general, for this venue, all event related-models noticeably outperform the *Baseline* and the *DDM* model.

The presented results show that for the $TL_{LANXESS}$, a classification-based approach to predict the severity of event-caused traffic seems feasible. The *SM* model shows an *f-measure* of 0.8, which outperforms the baselines significantly. For the $TL_{Mercedes-Benz}$ the results are less easily interpretable. One explanation could lie in different events in that venue, including other artists and event categories. Another reason that could explain the results, which has already been discussed for that venue, is the overlapping events and in general, a more variant traffic situation in that area of Berlin. These factors would also interfere with the other models, which could explain the overall weak results for that TL.

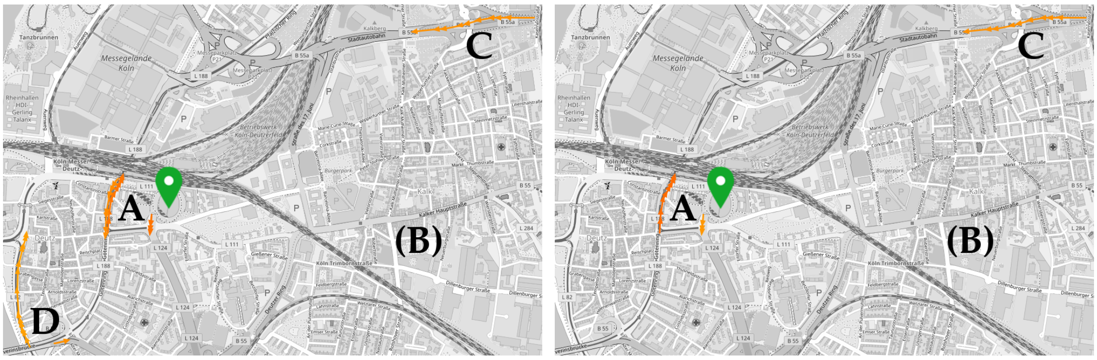
6.4 Review - Spatial Challenge

In Chapter 5, we defined road segments that show a different behavior on event and non-event days as being affected. We have also shown that a classification-based approach works for some event types but the results for mixed venues were less easy to interpret.

What we have seen in this chapter is that in certain scenarios, we were able to build models that were, to a certain extent, capable of predicting the severity of event-caused traffic disruptions. Can we combine these ideas? What does it mean if information from event-specific sources helps to build a model that works better than the one without them? Although not working perfectly, as long as event-specific information lowers the prediction errors, that particular link is probably affected by PSEs.

With that idea in mind, we re-run the classification study for all road segments within a radius of 4000 m around the venues. As in Chapter 5, we plot the road segments according to the resulting *f-measure*. The figures in 6.15 show all affected TLs with an f-measure > 0.7 for the LANXESS Arena in Cologne, Germany for the CM and SM separately.

The results show that we captured all road segments that were also shown in Figure 5.27 at marker A, which showed the highest *f-measure* from the spatial classification part. In this study, these road segments resulted in a *f-measure* of 0.74–0.78 for the CM and 0.73–0.8 for the SM, which is significantly higher than the *f-measures* in the previous studies, ranging between 0.65–0.71. Additionally,



(I) CategoryModel (CM) results.

(II) SocialMediaModel (SM) results.

FIGURE 6.15: Spatial region around the LANXESS Arena in Cologne based on the classification f-measures from the CM and SM.

the CM also marked these road segments leading southwards, left from marker A, with an *f-measure* of 0.74 as possibly affected. As discussed in the previous chapter, these results seem plausible.

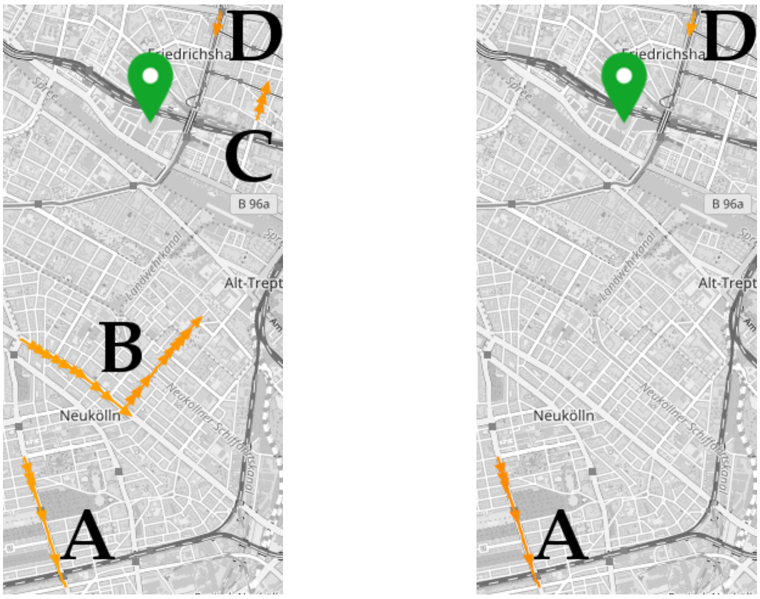
The results also show that those segments around marker B in Figure 5.27 (shown in Figure 6.15 at marker (B)), that have been discussed as possible misclassifications, did not show significantly good results in the classification task for the SM or CM.

However, we can also see new road segments: marker C in Figure 6.15i and 6.15ii and those segments around marker D in Figure 6.15i. The road segments at marker C result in an *f-measure* of 0.73 for the CM and SM. The CM returned an *f-measure* of 0.70 for the road segments at marker D. Thus, on those road segments, information about the category or social media information about events was partly sufficient to identify events with a high impact on traffic on those segments. Again, owing to missing ground truth information, that interpretation is solely intuitive, but the observed behavior allows the assumption that those road segments are at least partly affected by large events in the LANXESS Arena.

Another interesting phenomenon is the observation that the road segments that show a high *f-measure* for the SM represent a subset of road segments where a classification based on the category performs best. For most of the road segments, the resulting *f-measures* are also comparable.

In summary, the results for the LANXESS Arena support the results from the previous study in Chapter 5. We were able to identify road segments that were intuitively expected as being affected by PSE traffic. In addition, the CM and SM also reduced the number of misclassifications for this venue and we found new road segments that showed event-related traffic behavior.

The results for the second venue, the Mercedes-Benz Arena, are shown in Figure 6.16. These figures show the results for the CM and SM models, where



(I) CategoryModel (CM) results. (II) SocialMediaModel (SM) results.

FIGURE 6.16: Spatial region around the Mercedes-Benz Arena in Berlin, Germany based on the classification *f1-measures* from CM and SM.

the *f-measure* ≥ 0.7. Comparing these road segments to those from Figure 5.23 shows that none of the segments with a high *f-measure* in the spatial classification study also showed good results using the SM or CM. As already seen, the road segment right in front of the venue (marker A in Figure 5.23) that was analyzed as $TL_{Mercedes-Benz}$ does not return results ≥ 0.7.

However, there are road segments that have not been selected by the spatial region classification task in the previous chapter. Those at marker A show an *f-measure* of 0.70 for the CM and 0.74 for the SM for the *High* class. These road segments are surprisingly far away from the venue. They point directly to a highway on-ramp that can also be used to reach the venue, but we have not found any evidence supporting this assumption. There is also the possibility that people leave their cars and take public transport from that location.

The road segments at marker D show a *f-measures* of 0.73 and 0.70 for the CM and SM, respectively. Again, these segments are not part of the results of the study in Chapter 5. They point directly to the venue location and the fact they show certain PSE patterns seems plausible.

The CM also classified the road segments at markers B and C, both with an *f-measure* of 0.7. Again, intuitively, they could be affected but we did not find any evidence for that.

In general, the results for the Mercedes-Benz Arena are less intuitive than those for the LANXESS Arena. Compared to those results in Chapter 5, the current approach however showed results that could be explained by our intuition and there are no road segments in the result set that are obviously misclassifications. Still, this approach was also probably highly affected by the difficult traffic situation in this area of Berlin.

6.5 Discussion & Conclusion

We have seen that neither information about the event category nor from social media sources was sufficient to predict traffic reliably using a LR approach. The results have shown that traffic, including among categories and intuitively "similar" events still varies too strongly for reliable predictions.

Whereas a prediction of the expected reduced *FFP* gave insufficient results to use it in driver information systems, a classification of events based on event attributes showed better results. We have seen that we were able to create a model based on online data sources that, at least for some road segments, predicted the size of events correctly. Again, we leave it for future work to develop other models or apply different strategies, but with this initial attempt, we have shown that it is at least partially feasible.

Another interesting take-away of this study is the difference between results for the two venues. For the LANXESS Arena, we found that the event-specific information was useful to classify the severity of events, whereas for the Mercedes-Benz Arena, regular day-to-day traffic was dominant. In conclusion, we argue that around the LANXESS Arena, event-specific disruptions are more distinguishable than for the Mercedes-Benz Arena. Venues in the neighborhood or a generally stronger impact of habitual traffic in Berlin could be a possible explanation.

For the task of identifying affected road segments, the approach shown in this chapter worked slightly better for the mixed venues than that shown before. However, without ground truth information, we can only evaluate the performance by our intuition.

Chapter 7

Conclusions and Future Work

The data-driven traffic prediction domain has been fundamentally boosted by the high availability and coverage of traffic datasets. However, the dynamic of traffic eco systems is influenced by many different factors and deriving a stable prediction model that takes them all into account is still an open challenge. For the specific topic of predicting traffic caused by PSEs, we have analyzed and shown some relevant aspects, which we conclude in this section. As part of this section, we also discuss our general findings and the lessons learned. This chapter closes with a list of challenges and questions that could be addressed in future work.

For the *location-specific characteristics*, as introduced in Chapter 1, we discussed many results and findings in the previous chapters. In this section, we want to emphasize two of them:

1.1) The impact of PSEs on traffic is venue-specific. We have seen very different traffic situations around different venues. A major distinguishing factor is the topology of the road infrastructure, but of course the observed results are not caused solely by that. Different types of people with different interests, different public transportation networks, and many other reasons lead to varying traffic situations. In Chapter 4, we analyzed traffic during events for different venues and concluding from our results, we can assume that a specific traffic footprint exists that is characteristic for each particular venue. This conclusion is, from our perspective, fundamental. Following our argumentation and results, it is not possible to generate a single model that can be applied to multiple venues. Particularly for data-driven approaches, this has severe implications. For each venue, there are only a limited number of event observations per year. As data has to be separately analyzed for each venue, modeling approaches requiring massive amounts of training data (e.g., deep learning strategies) are not applicable.

For future work, an alternative solution could lie in grouping similar venues to exploit more data observations together. Another possibility could be to find specific road segments around different venues that are comparable in their behavior during incoming or outgoing traffic, and use those segments to aggregate observations.

1.2) Traffic impact zones can be found using data-driven approaches. To analyze the behavior of a traffic-influencing factor, one first needs to locate it. This is why we analyzed the spatial impact region of traffic disruptions due to events in Chapter 5. We found that for soccer games, a classification-based method leads to plausible results, identifying road segments that tend to be affected during those games. We validated our results with information from different stadiums and found compelling arguments for the identified road segments. However, we learned that other event categories behave differently and show a less stable behavior. For those categories, we have shown an alternative approach in Chapter 6, using different sets of event-specific prediction models. Again, we found arguments to explain the observed results. However, the method can still be improved. Future work should focus on acquiring more detailed information sources in terms of traffic and event information. Additionally, based on those detailed information sets, more sophisticated prediction models are expected to increase the performance of the method.

In Chapter 4 and Chapter 6, we analyzed *event-specific characteristics*. Again, we want to focus here on two major conclusions:

2.1) Specific data from online sources contains information that is relevant for event-impact prediction. In Chapter 6, we analyzed measures from online sources in terms of their explanatory power toward the final goal of predicting event-specific disruptions around venues. Whereas regression methods did not return applicable results, we have seen that those sources were partially usable to classify the expected impact of events into severity classes. The presented results allow the assumption that, to a certain extent, information from social media is a predictor for the traffic disruptions from events. However, again, those phenomena seem to be venue-specific. We observed that the information value from the analyzed online sources changed among different radii and different venues.

2.2) Soccer games show the most stable behavior of all observed event categories. In Chapter 5, we analyzed traffic around different venues that focused on different types of events. From our results, we can conclude that traffic disruptions due to soccer games have a higher tendency to follow a repetitive pattern than for other event categories. For future work, this information could be crucial to benchmark new prediction models.

A general lesson learned from this work, although not directly connected to event traffic, is that incident traffic data has limited use for data analytics use cases. Such datasets are usually available to the public from commercial providers. However, we found only a few use cases where the limited information from those datasets was actually sufficient. We discussed the difficulties of those datasets in previous chapters and for future work, these issues should be critically evaluated.

We have seen that PSEs have the potential to change the traffic situation around venues drastically. Understanding these phenomena and being able to predict the expected traffic disruptions are crucial steps to ease the problem. In this thesis, we have shown our approaches toward that goal. However, for the time being, that problem is not yet entirely solved. In this thesis, we have identified the various challenges in that domain.

For **future work**, we outline some aspects that are still open issues and that should be analyzed more in detail:

Event Timing. In the presented studies, we always relied on fixed event time windows for the beginning and the end of events. Of course, this is a simplification. Different events probably show very different behavior when the attendees arrive or leave the event. Without knowing those times, even the best prediction would not be of much help.

Especially for the time before the event, the challenge lies in finding an information source that warns about the event start early enough before the traffic disruptions occur. One option is to focus on additional information from other sensors (e.g., mobile phone data) that could hold valuable information about crowd movements before the event.

For the end of events, we believe that information from live stream social media sources (e.g., Twitter) could provide useful information. Tweets such as "that concert was so cool..." could be used to detect the end of an event. Of course, such an approach would first have to overcome the many hurdles that come with this type of social media information. Still, though, event detection using Twitter data is an ongoing research topic (e.g., see [116] for a short summary) that, in our opinion, holds a high potential for this use case as well.

Live Traffic Information. Another interesting subject for future work is the use of more detailed traffic information. In our work, we used different traffic metrics (e.g. mean or variance) from historic collections only. For future work, an integration of live traffic information could yield significant benefits. Especially for areas such as Berlin, which have proven to show large variation in their daily traffic routines, those information sets would probably be very helpful.

Accurate and Regional Event Information. In Chapter 6, we discussed the collection of online information about events. In this work, for the explained reasons, we collected those datasets during one specific timespan for all artists. Of course, this approach cannot reflect sudden changes in popularity of artists (e.g., the release of a new album). Future studies should focus on the collection of near-real-time online information. The challenges for this approach have been already discussed in the previous chapter. Still, collecting information about the current popularity of artists from online sources could bring an advantage.

Apart from the up-to-dateness of online information, we also observed a region-specific bias in our dataset. For instance, in Figure 6.13, we can observe large traffic disruptions during concerts of *Herbert Groenemeyer* and *The Scorpions*. As both bands are from Germany, their popularity might be higher in their home country than in other regions. The collected online information sets, however, are international and do not reflect local popularity at all. Possible strategies to include such information in the datasets could focus on regional information sources. However, as those sources are rather limited, it would probably be challenging to collect sufficient data.

Another possible strategy could be to add region-specific weights to the dataset. These weights could also be learned from historic data. As promising as this approach sounds, it would require many observations of the same artist in different locations to be able to learn a certain pattern.

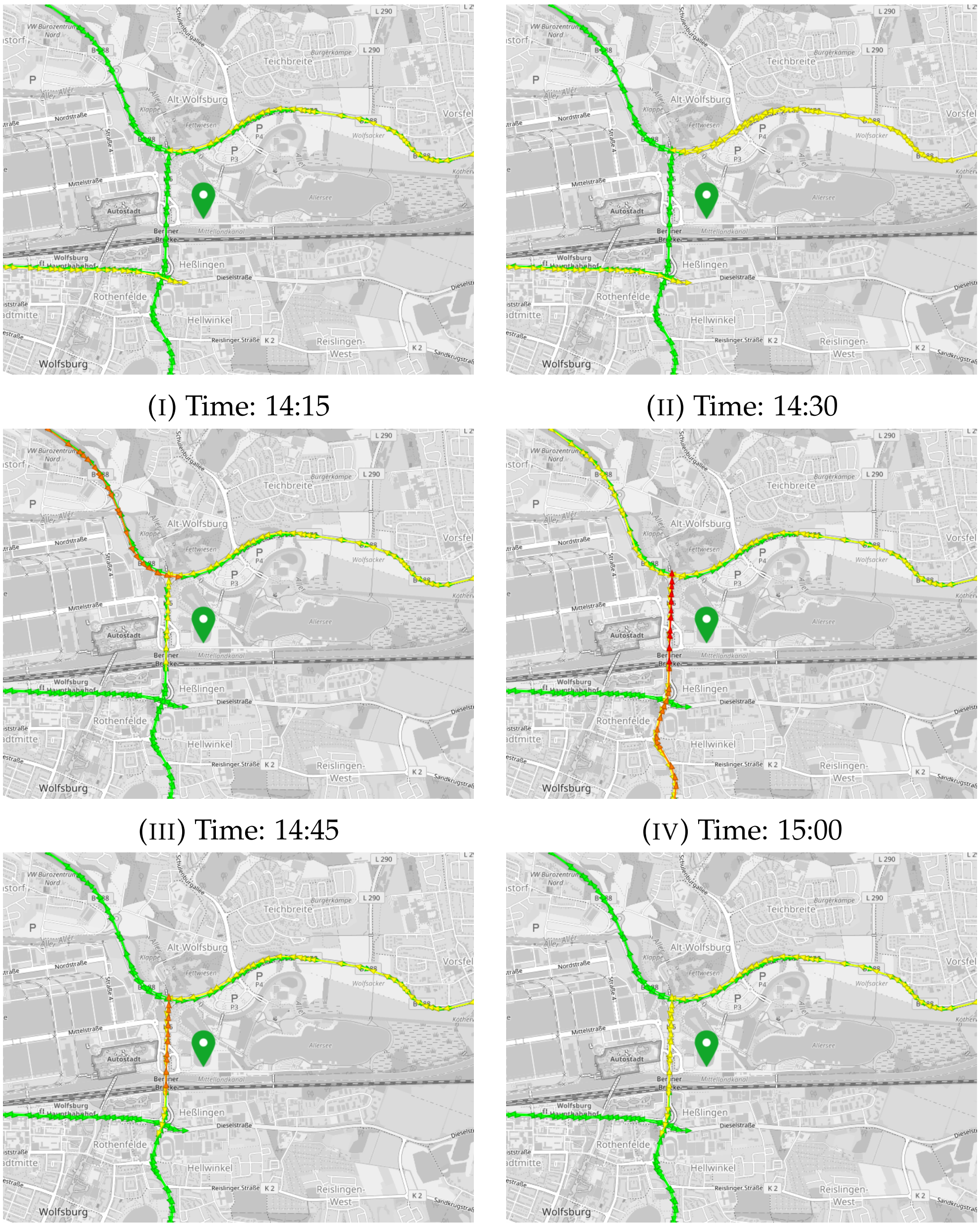
In this thesis, we have taken the initial steps toward a reliable PSE traffic prediction. To the best of our knowledge, research in this domain has been limited so far. We have shown different approaches to the various challenges that come alongside the ultimate goal of predicting these traffic phenomena. However, we have also identified many additional challenges and open issues for future work.

Appendix A

Event Traffic Example

This thesis focuses on the impact of *planned special events* on traffic. The following figures give an example of a traffic situation before and after a soccer game. The shown traffic situations were captured in Wolfsburg, Germany on 08/02/2014 when the *VfL Wolfsburg* played against *1. FSV Mainz 05*. On that day, the *VfL Wolfsburg* won three to zero.

The game was scheduled for 15:30. The following figures show the *incoming* and *outgoing* traffic in A.1 and A.2



(I) Time: 14:15

(II) Time: 14:30

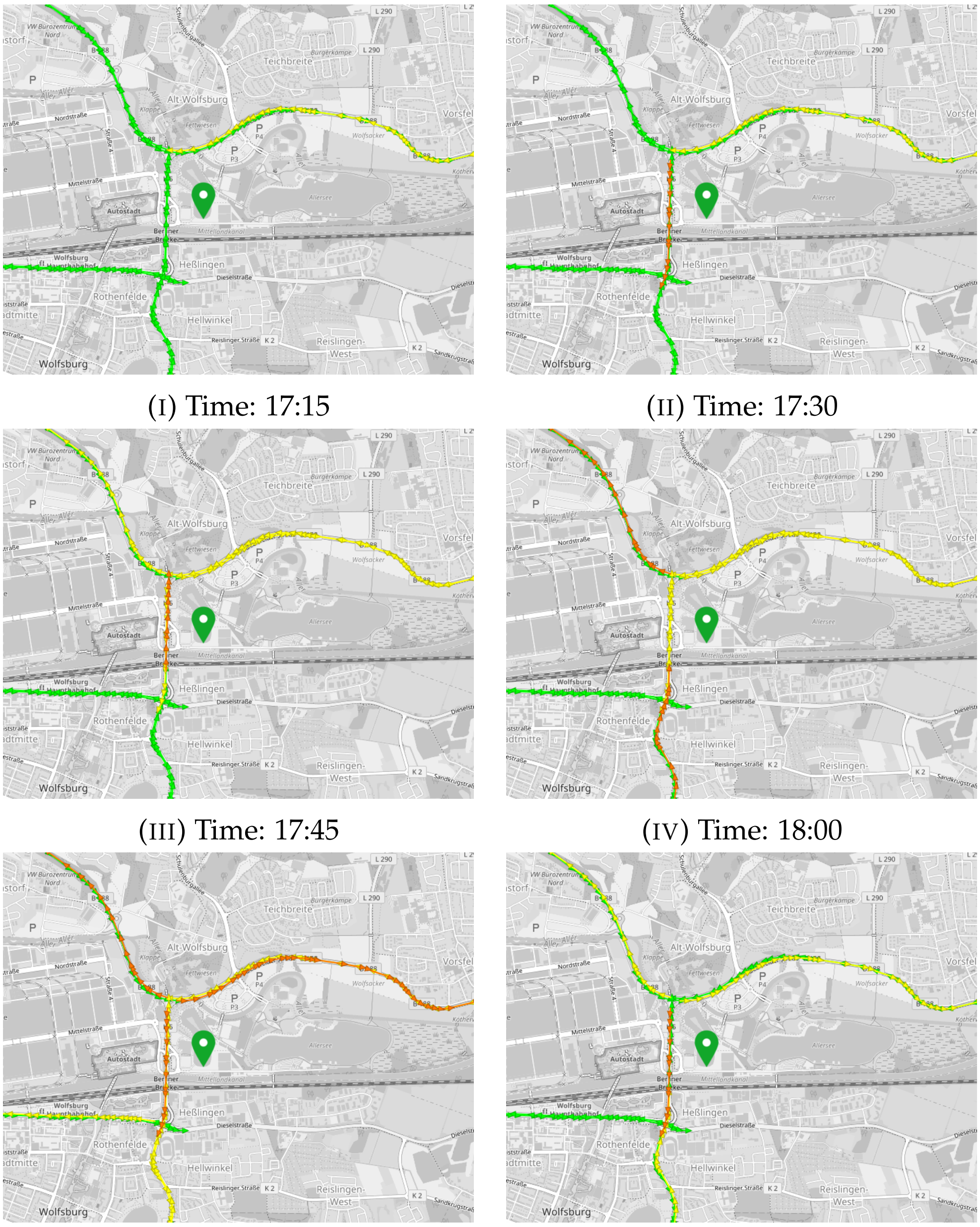
(III) Time: 14:45

(IV) Time: 15:00

(V) Time: 15:15

(VI) Time: 15:30

FIGURE A.1: *Incoming* traffic before a soccer game in Wolfsburg on 08/02/2014. Green marker: soccer venue.



(I) Time: 17:15

(II) Time: 17:30

(III) Time: 17:45

(IV) Time: 18:00

(V) Time: 18:15

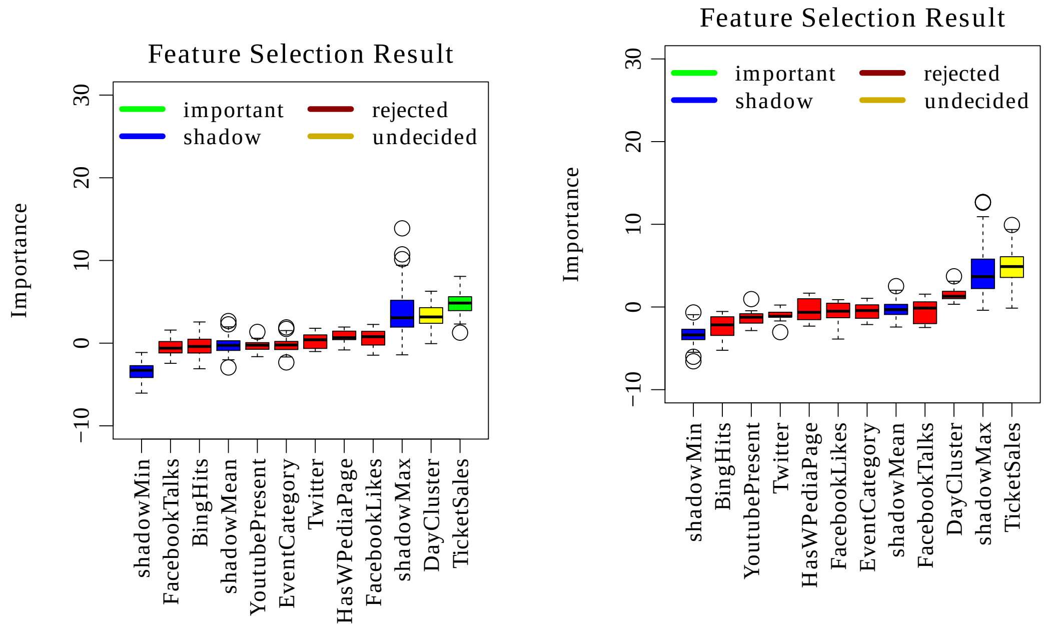
(VI) Time: 18:30

FIGURE A.2: *Outgoing* traffic after a soccer game in Wolfsburg on 08/02/2014. Green marker: soccer venue.

Appendix B

Online Metrics Hamburg

In Chapter 6.2 results of a feature relevance study have been presented for venues in Cologne, Germany and Berlin, Germany. In the following we show the results for the *O2-World* in Hamburg, Germany.



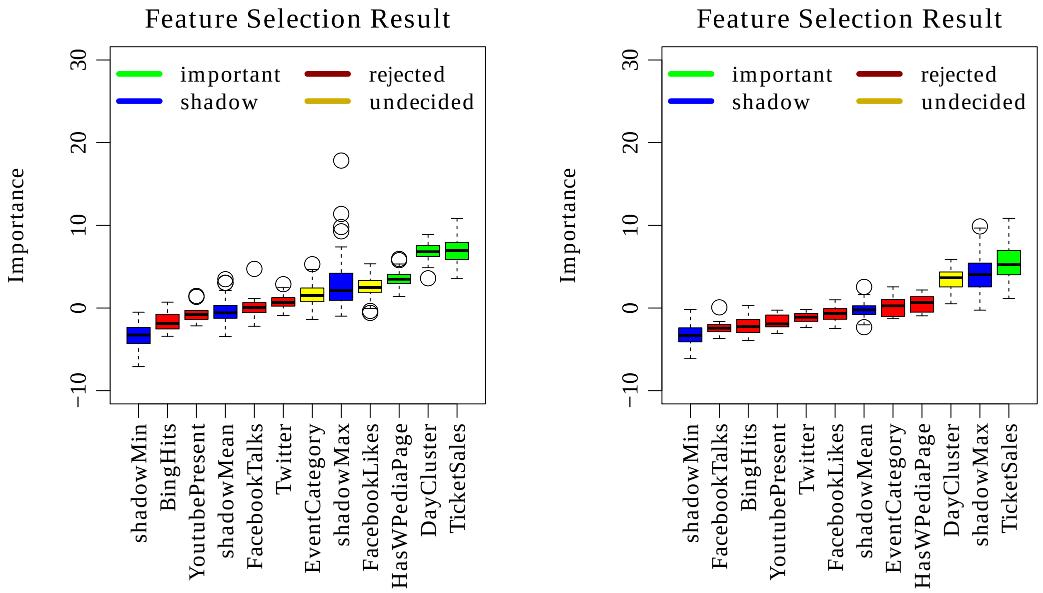
(I) O2-World Hamburg. Metric = mean.

(II) O2-World Hamburg. Metric = variance.

FIGURE B.1: Venue: O2-World Hamburg. 500 m radius.

In a 500 m radius around the O2-World in Hamburg, with the *mean* as a traffic measure (see Figure B.1i), the only feature that has been selected as relevant is the *ticket sales* with a median *Z score* of 7.77. All the other features got rejected by the algorithm. These results stay nearly the same when taking the *variance* as the traffic measure, as shown in Figure B.1ii. Again, only the *ticket sales* got confirmed in this scenario, with a *Z score* of 7.56.

By enlarging the radius to 1000 m (see Figure B.2) around the venue, the overall *Z score* for the *ticket sales* increases to 10.0 for the *mean* and 8.84 for the *variance* as a traffic measure. In addition, for the *mean* traffic measure, the features *FacebookLikes*, *LastFMListeners*, and *LastFMPlayCounts* become more relevant and are no longer rejected by the feature selection algorithm.

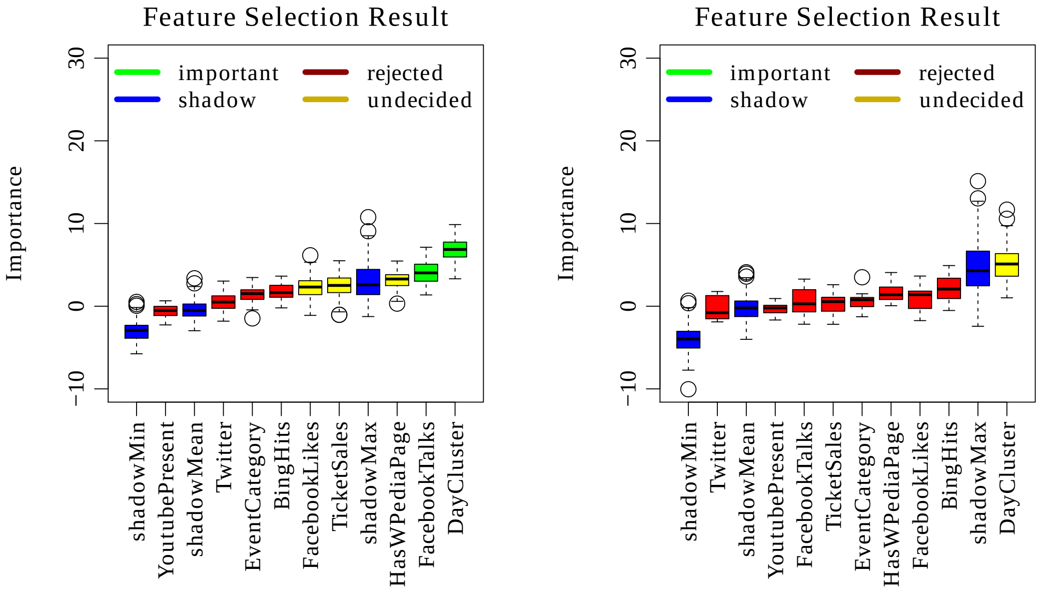


(I) O2-World Hamburg. 1000 m radius. Metric = mean.

(II) O2-World Hamburg. 1000 m radius. Metric = variance.

FIGURE B.2: Venue: O2-World Hamburg. 1000 m radius.

Further increasing the radius to 2000 m (see Figure B.3) shows similar patterns. For the *mean* traffic measure, *BingHits*, *FacebookTalks*, *LastFMListeners*, *TicketSales*, and *LastFMPlayCounts* become relevant. However, their mean *Z score* lowers compared to the 500 and 1000 m radii, and the most relevant feature for 2000 m (*LastFMPlayCounts*) only shows a *Z score* of 4.62. For the *variance metric*, no feature is selected as relevant at a radius of 2000 m, although the *Z scores* for the attributes are similar to those for the *mean* study.



(I) O2-World Hamburg. 2000 m radius. Metric = mean.

(II) O2-World Hamburg. 2000 m radius. Metric = variance.

FIGURE B.3: Venue: O2-World Hamburg. 2000 m radius.

Appendix C

Curriculum Vitae

Born on 13/01/1984 in Berlin/Germany.

EDUCATION

2013 - Today	**Leibniz Universität Hannover** Distributed Systems Institute / Knowledge Based Systems PhD Candidate	Hannover, Germany
2004 - 2011	**Technische Universität Berlin** Institute for Telecommunications Systems / Open Communication Systems Diploma degree in Computer Science	Berlin, Germany

PROFESSIONAL EXPERIENCE

2016 - Today	**Volkswagen AG** Researcher	Wolfsburg, Germany
2013 - 2016	**Volkswagen AG** PhD Candidate - Mobility Solutions	Wolfsburg, Germany

Appendix D

List of Publications

Journals & Book Chapters

S. Di Martino, **S. Kwoczek**, and W. Nejdl. Smart Sensors Networks: Communication Technologies and Intelligent Applications, chapter Scalable Processing of Massive Traffic Datasets, pages 123-142. Elsevier, 2017.

S. Kwoczek, S. Di Martino, and W. Nejdl. Predicting and visualizing traffic congestion in the presence of planned special events. Journal of Visual Languages & Computing, 25(6):973–980, 2014

Conferences

S. Kwoczek, S. Di Martino, T. Rustemeyer, and W. Nejdl. An architecture to process massive vehicular traffic data. In 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), pages 515–520, Nov.

S. Kwoczek, S. Di Martino, and W. Nejdl. Stuck around the stadium? an approach to identify road segments affected by planned special events. In Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on, pages 1255–1260, Sept.

S. Kwoczek, S. Di Martino, and W. Nejdl. Predicting traffic congestion in presence of planned special events. In Proceedings of the Twentieth International Conference on Distributed Multimedia Systems, DMS, pages 357–364, 2014.

Relevant Patents

S. Kwoczek, S. Di Martino and A. Sasse. Verfahren und Vorrichtung zur Ermittlung von Informationen über Mobilitätssituationen , in Deutsches Patent- und Markenamt, Number DE102014213350 (A1), 2016.

Bibliography

[1] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A.H. Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, 2011.

[2] S. Kwoczek, S. Di Martino, and W. Nejdl. Stuck around the stadium? an approach to identify road segments affected by planned special events. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pages 1255–1260, Sept .

[3] U.S. Department of Transportation. Planned special events - economic role and congestion effects. Technical Report FHWA-HOP-08-022, Federal Highway Administration, Aug 2008.

[4] F.C. Pereira, F. Rodrigues, E. Polisciuc, and M. Ben-Akiva. Why so many people? explaining nonhabitual transport overcrowding with internet data. *Intelligent Transportation Systems, IEEE Transactions on*, 16(3):1370–1379, June 2015.

[5] D. Matherly, P. Murray-Tuite, B. Wolshon, A. Pande, and B. Wolshon. *Traffic Engineering Handbook*, chapter Traffic Management for Planned, Unplanned, and Emergency Events, pages 599–636. John Wiley & Sons, Inc, 2015. ISBN 9781119174738.

[6] S. Borysov, M. Lourenço, F. Rodrigues, A. Balatsky, and F. Pereira. Using internet search queries to predict human mobility in social events. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1342–1347, Nov 2016.

[7] J. Kwon, M. Mauch, and P. Varaiya. Components of congestion: Delay from incidents, special events, lane closures, weather, potential ramp metering gain, and excess demand. *Transportation Research Record: Journal of the Transportation Research Board*, 1959(1):84–91, 2006.

[8] E. Horvitz, J. Apacible, R. Sarin, and L. Liao. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. In *In Twenty-First Conference on Uncertainty in Artificial Intelligence*, 2005.

[9] A. Haghani, M. Hamedi, K. Sadabadi, S. Young, and P. Tarnoff. Data collection of freeway travel time ground truth with bluetooth sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2160: 60–68, 2010.

[10] J. Barceló, L. Montero, L. Marqués, and C. Carmona. Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation Research Record: Journal of the Transportation Research Board*, 2175:19–27, 2010.

[11] N. Caceres, J.P. Wideberg, and F.G. Benitez. Review of traffic data estimations extracted from cellular networks. *IET Intelligent Transport Systems*, 2:179–192(13), September 2008.

[12] R. Chrobok, O. Kaumann, J. Wahle, and M. Schreckenberg. Different methods of traffic forecast based on real data. *European Journal of Operational Research*, 155(3):558–568, 2004.

[13] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[14] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[15] R. Dowling, A. Skabardonis, M. Carroll, and Z. Wang. Methodology for measuring recurrent and nonrecurrent traffic congestion. *Transportation Research Record: Journal of the Transportation Research Board*, 1867:60–68, 2004.

[16] OECD. Managing urban traffic congestion. OECD Publishing, 2007.

[17] B. Anbaroglu, B. Heydecker, and T. Cheng. Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks. *Transportation Research Part C: Emerging Technologies*, 48:47 – 65, 2014.

[18] S. Amini, E. Papapanagiotou, and F. Busch. Traffic management for major events. *Digital Mobility Platforms and Ecosystems*, page 187, 2016.

[19] A. T. Hojati, L. Ferreira, S. Washington, P. Charles, and A. Shobeirinejad. Modelling the impact of traffic incidents on travel time reliability. *Transportation Research Part C: Emerging Technologies*, 65:49 – 60, 2016.

[20] R. Venkatanarayana, B. Smith, and M. Demetsky. Quantum-frequency algorithm for automated identification of traffic patterns. *Transportation Research Record: Journal of the Transportation Research Board*, 2024:8–17, 2008.

[21] M. Mueller. What makes an event a mega-event? definitions and sizes. *Leisure Studies*, 34(6):627–642, 2015.

[22] Stadium: Its for large events. `http://www.largeevents.eu`, 2013. Accessed: 01/12/2016.

[23] S. P. Latoski, W. M. Dunn Jr, B. Wagenblast, J. Randall, and M. D. Walker. Managing travel for planned special events. In *Institute of Transportation Engineers (ITE) 2003 Technical Conference and Exhibit*, number CD-020, 2003.

[24] J.-C. Laprie. From dependability to resilience. In *38th IEEE/IFIP Int. Conf. On Dependable Systems and Networks*, pages G8–G9. Citeseer, 2008.

[25] S. P. Hoogendoorn, V. L. Knoop, H. van Lint, and H. L. Vu. Applications of the generalized macroscopic fundamental diagram. In *Traffic and Granular Flow'13*, pages 577–583. Springer, 2015.

[26] P. M. Murray-tuite. A comparison of transportation network resilience under simulated system optimum and user equilibrium conditions. In *Proceedings of the 2006 Winter Simulation Conference*, pages 1398–1405, Dec 2006.

[27] G. Leduc. Road traffic data: Collection methods and applications. *Working Papers on Energy, Transport and Climate Change*, 1(55), 2008.

[28] U.S. Department of Transportation. Traffic detector handbook: Third edition - volume ii. Technical Report FHWA-HRT-06-139, Federal Highway Administration, Oct 2006.

[29] B. Pan, U. Demiryurek, and C. Shahabi. Utilizing real-world transportation data for accurate traffic prediction. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 595–604. IEEE, 2012.

[30] B. Pan, U. Demiryurek, C. Shahabi, and C. Gupta. Forecasting spatiotemporal impact of traffic incidents on road networks. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 587–596, Dec .

[31] R. Chrobok. *Theory and application of advanced traffic forecast methods*. PhD thesis, Universität Duisburg-Essen, Fakultät für Physik, 2005.

[32] J. Z. Zhu, J. X. Cao, and Y. Zhu. Traffic volume forecasting based on radial basis function neural network with the consideration of traffic flows at the adjacent intersections. *Transportation Research Part C: Emerging Technologies*, 2014.

[33] B. S. Kerner, C. Demir, R. G. Herrtwich, S. L. Klenov, H. Rehborn, M. Aleksic, and A. Haug. Traffic state detection with floating car data in road networks. In *Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005.*, pages 44–49, Sept 2005.

[34] C. de Fabritiis, R. Ragona, and G. Valenti. Traffic estimation and prediction based on real time floating car data. In *2008 11th International IEEE Conference on Intelligent Transportation Systems*, pages 197–203, Oct 2008.

[35] R.-P. Schäfer, K.-U. Thiessenhusen, and P. Wagner. A traffic information system by means of real-time floating-car data. In *ITS world congress*, volume 2, 2002.

[36] TomTom International. White paper - how tomtom's hd traffic and iq routes data provides the very best routing. Technical report, TomTom International, 2009.

[37] R.-P. Schäfer. Iq routes and hd traffic: Technology insights about tomtom's time-dynamic navigation concept. In *Proceedings of the the 7th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on The Foundations of Software Engineering*, ESEC/FSE

'09, pages 171–172, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-001-2.

[38] H. Bar-Gera. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel. *Transportation Research Part C: Emerging Technologies*, 15(6):380 – 391, 2007.

[39] Y. Lu, R. Seshadri, F. Pereira, A. O'Sullivan, C. Antoniou, and M. Ben-Akiva. Dynamit2.0: Architecture design and preliminary results on real-time data fusion for traffic prediction and crisis management. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 2250–2255, Sept 2015.

[40] E. I. Vlahogianni, J. C. Golias, and M. G. Karlaftis. Short-term traffic forecasting: Overview of objectives and methods. *Transport Reviews*, 24(5): 533–557, 2004.

[41] H. Adeli. Neural networks in civil engineering: 1989-2000. *Computer-Aided Civil and Infrastructure Engineering*, 16(2):126–142, 2001.

[42] J. W. C. Van Lint and C. P. I. J. Van Hinsbergen. Short-term traffic and travel time prediction models. *Artificial Intelligence Applications to Critical Transportation Issues*, 22:22–41, 2012.

[43] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias. Short-term traffic forecasting: Where we are and where we?re going. *Transportation Research Part C: Emerging Technologies*, 2014.

[44] M. G. Karlaftis and E. I. Vlahogianni. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387 – 399, 2011.

[45] A. P. Engelbrecht. *Computational intelligence: an introduction*. John Wiley & Sons, 2007. ISBN 978-0470035610.

[46] A. Sadek, G. Spring, and B. Smith. Toward more effective transportation applications of computational intelligence paradigms. *Transportation Research Record: Journal of the Transportation Research Board*, 1836:57–63, 2003.

[47] M. S. Ahmed and A. R. Cook. Analysis of freeway traffic time-series data by using box-jenkins techniques. *Transportation Research Record*, (722):1–9, 1979.

[48] A. Stathopoulos and M.G. Karlaftis. A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C: Emerging Technologies*, 11(2):121–135, 2003.

[49] M. Levin and Y.-D. Tsao. On forecasting freeway occupancies and volumes (abridgment). *Transportation Research Record*, (773), 1980.

[50] G. A. Davis, N. L. Nihan, M. M. Hamed, and L. N. Jacobson. Adaptive forecasting of freeway traffic congestion. *Transportation Research Record*, (1287), 1990.

[51] M. M. Hamed, H. R. Al-Masaeid, and Z. M. B. Said. Short-term prediction of traffic volume in urban arterials. *Journal of Transportation Engineering*, 121(3):249–254, 1995.

[52] B. M. Williams and L. A. Hoel. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of transportation engineering*, 129(6):664–672, 2003.

[53] B. Park, C. Messer, and T. Urbanik II. Short-term freeway traffic volume forecasting using radial basis function neural network. *Transportation Research Record: Journal of the Transportation Research Board*, (1651):39–47, 1998.

[54] S.M. Amin, E.Y. Rodin, A.-P. Liu, K. Rink, and A. García-Ortiz. Traffic prediction and management via rbf neural nets and semantic control. *Computer-Aided Civil and Infrastructure Engineering*, 13(5):315–327, 1998.

[55] A.W. Jayawardena, D. Achela, and K. Fernando. Use of radial basis function type artificial neural networks for runoff simulation. *Computer-Aided Civil and Infrastructure Engineering*, 13(2):91–99, 1998.

[56] H. Chen, S. Grant-Muller, L. Mussone, and F. Montgomery. A study of hybrid neural network approaches and the effects of missing data on traffic forecasting. *Neural Computing & Applications*, 10(3):277–286, 2001.

[57] M. Van Der Voort, M. Dougherty, and S. Watson. Combining kohonen maps with arima time series models to forecast traffic flow. *Transportation Research Part C: Emerging Technologies*, 4(5):307 – 318, 1996.

[58] B. Abdulhai, H. Porwal, and W. Recker. Short term freeway traffic flow prediction using genetically-optimized time-delay-based neural networks. *California Partners for Advanced Transit and Highways (PATH)*, 1999.

[59] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias. Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach. *Transportation Research Part C: Emerging Technologies*, 13 (3):211 – 234, 2005.

[60] H. Yin, S.C. Wong, J. Xu, and C.K. Wong. Urban traffic flow prediction using a fuzzy-neural approach. *Transportation Research Part C: Emerging Technologies*, 10(2):85 – 98, 2002.

[61] W. Zheng, D. Lee, and Q. Shi. Short-term freeway traffic flow prediction: Bayesian combined neural network approach. *Journal of Transportation Engineering*, 132(2):114–121, 2006.

[62] F. Calabrese, F. C. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti. *The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events*, pages 22–37. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-12654-3.

[63] F. C. Pereira, F. Rodrigues, and M. Ben-Akiva. Using data from the web to predict public transport arrivals under special events scenarios. *Journal of Intelligent Transportation Systems*, 19(3):273–288, 2015.

[64] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[65] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[66] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010*

Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 181–189. Association for Computational Linguistics, 2010.

[67] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *ICWSM*, 2011.

[68] K. Radinsky and E. Horvitz. Mining the web to predict future events. In *WSDM*, pages 255–264, 2013.

[69] P. Agarwal, R. Vaithiyanathan, S. Sharma, and G. Shroff. Catching the long-tail: Extracting local news events from twitter. In *The 7th International AAAI Conference on Weblogs and Social Media*, ICWSM12, 2012.

[70] J. Nichols, J. Mahmud, and C. Drews. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, IUI '12, pages 189–198, New York, NY, USA, 2012. ACM.

[71] O. Alonso and K. Shiells. Timelines as summaries of popular scheduled events. In *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion, pages 1037–1044, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[72] D. Villatoro, J. Serna, V. Rodríguez, and M. Torrent-Moreno. The tweet-beat of the city: Microblogging used for discovering behavioural patterns during the mwc2012. In *Citizen in Sensor Networks*, pages 43–56. Springer, 2013.

[73] H. Becker, D. Iter, M. Naaman, and L. Gravano. Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 533–542. ACM, 2012.

[74] K. Radinsky, S. Davidovich, and S. Markovitch. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 909–918, New York, NY, USA, 2012. ACM.

[75] K. Radinsky, S. Davidovich, and S. Markovitch. Learning to predict from textual data. *J. Artif. Intell. Res. (JAIR)*, 45:641–684, 2012.

[76] Z. Zhang, M. Ni, Q. He, and J. Gao. Mining transportation information from social media for planned and unplanned events. Technical report, University at Buffalo, The State University of New York, 2016.

[77] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom, and P. Chaovalit. Social-based traffic information extraction and classification. In *ITS Telecommunications (ITST), 2011 11th International Conference on*, pages 107–112, 2011.

[78] E. Mai and R. Hranac. Twitter interactions as a data source for transportation incidents. In *Transportation Research Board 92nd Annual Meeting*, number 13-1636, 2013.

[79] A. Schulz, P. Ristoski, and H. Paulheim. I see a car crash: Real-time detection of small scale incidents in microblogs. In *Proceedings of Social Media and Linked Data for Emergency Response (SMILE) Workshop at ESWC 2013*, 2013.

[80] E.M. Daly, F. Lecue, and V. Bicer. Westland row why so slow?: Fusing social media and linked data sources for understanding real-time traffic conditions. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 203–212, New York, NY, USA, 2013. ACM.

[81] A. Gal-Tzur, S.M. Grant-Muller, T. Kuflik, E. Minkov, S. Nocera, and I. Shoor. The potential of social media in delivering transport policy goals. *Transport Policy*, 32:115 – 123, 2014.

[82] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni. Real-time detection of traffic from twitter stream analysis. *Intelligent Transportation Systems, IEEE Transactions on*, 16(4):2269–2283, Aug 2015.

[83] A. Kumar, M. Jiang, and Y. Fang. Where not to go? detecting road hazards using twitter. In *Proceedings of the 37th International ACM SIGIR Conference on Research; Development in Information Retrieval*, SIGIR '14, pages 1223–1226, New York, NY, USA, 2014. ACM.

[84] Z. Zhang, M. Ni, Q. He, J. Gao, J. Gou, and X. Li. An exploratory study on the correlation between twitter concentration and traffic surge. *To appear in Transportation Research Record*, 35:36, 2016.

[85] P. T. Chen, F. Chen, and Z. Qian. Road traffic congestion monitoring in social media with hinge-loss markov random fields. In *2014 IEEE International Conference on Data Mining*, pages 80–89, Dec 2014.

[86] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence. Improving traffic prediction with tweet semantics. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 1387–1393. AAAI Press, 2013. ISBN 978-1-57735-633-2.

[87] L. Lin, M. Ni, Q. He, J. Gao, and A.W. Sadek. Modeling the impacts of inclement weather on freeway traffic speed: Exploratory study with social media data. *Transportation Research Record: Journal of the Transportation Research Board*, (2482):82–89, 2015.

[88] Inc INRIX. Inrix xd traffic. `http://inrix.com/xd-traffic`, . Accessed: 01/12/2016.

[89] Inc INRIX. Inrix xd traffic - frequently asked questions. `http://inrix.com/inrix-traffic-4-0-faq`, . Accessed: 01/12/2016.

[90] Google maps help. `https://support.google.com/maps/answer/6149565`. Accessed: 01/12/2016.

[91] S. Kwoczek, S. Di Martino, T. Rustemeyer, and W. Nejdl. An architecture to process massive vehicular traffic data. In *2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*, pages 515–520, Nov .

[92] S. Di Martino, S. Kwoczek, and W. Nejdl. *Smart Sensors Networks: Communication Technologies and Intelligent Applications*, chapter Scalable Processing of Massive Traffic Datasets, pages 123–142. Elsevier, 2017.

[93] S. Kwoczek, S. Di Martino, and W. Nejdl. Predicting traffic congestion in presence of planned special events. In *Proceedings of the Twentieth International Conference on Distributed Multimedia Systems*, DMS, pages 357–364, 2014.

[94] S. Kwoczek, S. Di Martino, and W. Nejdl. Predicting and visualizing traffic congestion in the presence of planned special events. *Journal of Visual Languages & Computing*, 25(6):973–980, 2014.

[95] K. Born, S. Yasmin, D. You, N. Eluru, C. Bhat, and R. Pendyala. Joint model of weekend discretionary activity participation and episode duration. *Transportation Research Record: Journal of the Transportation Research Board*, 2413:34–44, 2014.

[96] D. Leilei, S. Zheng-liang, G. Jin-gang, and Q. Hong-tong. Study on traffic organization and management strategies for large special events. In *2012 International Conference on System Science and Engineering (ICSSE)*, pages 432–436, June 2012.

[97] F. Bai and B. Krishnamachari. Spatio-temporal variations of vehicle traffic in vanets: Facts and implications. In *Proceedings of the Sixth ACM International Workshop on VehiculAr InterNETworking*, VANET '09, pages 43–52, New York, NY, USA, 2009. ACM.

[98] R. Simmons, B. Browning, Y. Zhang, and V. Sadekar. Learning to predict driver route and destination intent. In *2006 IEEE Intelligent Transportation Systems Conference*, pages 127–132, Sept 2006.

[99] J. Kwon and K. Murphy. Modeling freeway traffic with coupled hmms. Technical report, Technical report, Univ. California, Berkeley, 2000.

[100] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006. ISBN 978-0-387-31073-2.

[101] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009. ISBN 978-0-387-84858-7.

[102] Inrix scorecard germany. `http://inrix.com/press-releases/scorecard-de/`. Accessed: 09/01/2017.

[103] J.M. Frantzeskakis and M.J. Frantzeskakis. Athens 2004 olympic games: Transportation planning, simulation and traffic management. *Institute of Transportation Engineers. ITE Journal*, 76(10):26, 2006.

[104] M. Bonert, E. Brockfeld, I. Ernst, D. Krajzewicz, M. Ruhé, and P. Wagner. Soccer verkehrslageerfassung und –prognose während der fußball-wm. *Informationssysteme für mobile Anwendungen*, 2006.

[105] H. Liu and H. Motoda. *Computational methods of feature selection.* CRC Press, 2007. ISBN 978-1584888789.

[106] W.R. Rudnicki, M. Wrzesień, and W. Paja. *All Relevant Feature Selection Methods and Applications*, pages 11–28. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015. ISBN 978-3-662-45620-0.

[107] A.L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1):245 – 271, 1997.

[108] G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16 – 28, 2014. 40th-year commemorative issue.

[109] C. Diamantini and M. Panti. An efficient and scalable data compression approach to classification. *SIGKDD Explor. Newsl.*, 2(2):49–55, December 2000.

[110] J. Go and C. Lee. Analytical decision boundary feature extraction for neural networks. In *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings*, volume 7, pages 3072–3074 vol.7, 2000.

[111] W.R. Rudnicki, M. Kierczak, J. Koronacki, and J. Komorowski. *A Statistical Method for Determining Importance of Variables in an Information System*, pages 557–566. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-49842-1.

[112] M.B. Kursa, A. Jankowski, and W.R. Rudnicki. Boruta - a system for feature selection. *Fundam. Inf.*, 101(4):271–285, December 2010.

[113] M. Kursa and W.R. Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software*, 36(1):1–13, 2010.

[114] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297. University of California Press, 1967.

[115] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[116] F. Atefeh and W. Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.

List of Figures

List of Tables