Assessing the Impact of Game Day Schedule and Opponents on Travel Patterns and Route Choice using Big Data Analytics

Final Report
June 2019

NEBRASKA
Good Life. Great Journey.
DEPARTMENT OF TRANSPORTATION

Disclaimer

About CTRE

This work was performed by the Center for Transportation Research and Education (CTRE) at Iowa State University. CTRE's mission is to develop and implement innovative methods, materials, and technologies for improving transportation efficiency, safety, and reliability while improving the learning environment of students, faculty, and staff in transportation-related fields.

Non-Discrimination Statement

Technical Report Documentation Page

1. Report No. SPR-1(18) M078	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Assessing the Impact of Game Day Schedule and Opponents on Travel Patterns and Route Choice using Big Data Analytics		5. Report Date June 2019
		6. Performing Organization Code
7. Author(s) Anuj Sharma and Vesal Ahsani		8. Performing Organization Report No.
9. Performing Organization Name and Address **Center for Transportation Research and Education** **Iowa State University** **2711 S. Loop Drive, Suite 4700** **Ames, IA 50010-8664**		10. Work Unit No. (TRAIS)
		11. Contract or Grant No.
12. Sponsoring Agency Name and Address Nebraska Department of Transportation 1500 Hwy. 2 Lincoln, NE 68502		13. Type of Report and Period Covered Final Report July 2017 – June 2019
		14. Sponsoring Agency Code SPR-1(18) M078
15. Supplementary Notes		

16. Abstract

The transportation system is crucial for transferring people and goods from point A to point B. However, its reliability can be decreased by unanticipated congestion resulting from planned special events. For example, sporting events collect large crowds of people at specific venues on game days and disrupt normal traffic patterns.

The goal of this study was to understand issues related to road traffic management during major sporting events by using widely available INRIX data to compare travel patterns and behaviors on game days against those on normal days. A comprehensive analysis was conducted on the impact of all Nebraska Cornhuskers football games over five years on traffic congestion on five major routes in Nebraska. We attempted to identify hotspots, the unusually high-risk zones in a spatiotemporal space containing traffic congestion that occur on almost all game days. For hotspot detection, we utilized a method called Multi-EigenSpot, which is able to detect multiple hotspots in a spatiotemporal space. With this algorithm, we were able to detect traffic hotspot clusters on the five chosen routes in Nebraska. After detecting the hotspots, we identified the factors affecting the sizes of hotspots and other parameters. The start time of the game and the Cornhuskers' opponent for a given game are two important factors affecting the number of people coming to Lincoln, Nebraska, on game days. Finally, the Dynamic Bayesian Networks (DBN) approach was applied to forecast the start times and locations of hotspot clusters in 2018 with a weighted mean absolute percentage error (WMAPE) of 13.8%.

17. Key Words game day traffic—hotspot detection—local sporting events—spatiotemporal data		18. Distribution Statement	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 75	22. Price

Assessing the Impact of Game Day Schedule and Opponents on Travel Patterns and Route Choice using Big Data Analytics

Final Report

Anuj Sharma, PhD
Research Scientist and Associate Professor
Center for Transportation Research and Education
Iowa State University

Vesal Ahsani, PhD
Graduate Research Assistant
Center for Transportation Research and Education
Iowa State University

A Report on Research Sponsored by

Nebraska Department of Transportation

June 2019

Table of Contents

List of Figures

List of Tables

Acknowledgments

The authors would like to thank the Nebraska Department of Transportation (formerly the Nebraska Department of Roads) for sponsoring this research and the Federal Highway Administration for the state planning and research funds that were used to help fund this project.

Executive Summary

In recent years, traffic congestion has become a significant issue in urban areas. People in the United States travel an extra one billion hours and consume an extra one billion gallons of fuel due to traffic congestion every year. Therefore, monitoring the performance of the transportation system plays an important role in any transportation operation and planning strategy.

Congestion that is caused by accidents, road work, special events, or adverse weather is called non-recurring congestion. Non-periodic events with an expected large attendance (referred to as planned special events [PSE]), such as concerts, football games, etc., play a major role in transportation delays.

Memorial Stadium in Lincoln, Nebraska, is the home of the Nebraska Cornhuskers football team. With an extended capacity of more than 85,000 people, the stadium is commonly referred as the "third-largest city in Nebraska" on game days. Game days, therefore, typically affect travel patterns in Lincoln and its neighboring regions.

This report documents a study evaluating the relationship between professional sporting events and traffic congestion using INRIX data covering the past five years in Nebraska. The objective of this study was twofold: (1) monitor and evaluate the performance of the transportation system and travel behavior on football game days and (2) detect game day traffic hotspots on five major routes in Nebraska and identify significant factors affecting hotspot size.

This study demonstrates a systematic way to assess travel patterns and identify traffic hotspot clusters on football game days compared to normal days. Five major routes in Nebraska were selected for this study, and the analysis utilized historical and real-time traffic data, including speeds, travel times, and location information, collected through the INRIX traffic message channel (TMC) monitoring platform. The INRIX dataset is currently regarded as the largest crowd-sourced traffic dataset. A comprehensive exploratory analysis of performance monitoring on game days against normal days for the five selected routes in Nebraska was also performed.

Among the different analytical tasks that can be performed on spatiotemporal data, hotspot analysis is an important tool in the transportation field. A realistic scenario involving the application of hotspot detection is in traffic incident detection. A novel method for hotspot detection is proposed in this report. The proposed algorithm uses the spatiotemporal matrix of expected congestion cases as the baseline information. Using the expected congestion case matrix as the baseline information, we can replace the observed cases by the respective expected cases for the previously detected congestion regions in the spatiotemporal space and re-run the algorithm to detect additional hotspot clusters, if they exist.

After detecting hotspots, it is crucial to identify the factors affecting the sizes of the hotspots, their locations, and other possible parameters. The start time of the game and the Cornhuskers' opponent for a given game are two important factors affecting the number of people coming to Lincoln, Nebraska, on game days. The start time of the game can be classified as either noon or evening. The opponent of the Nebraska Cornhuskers also plays a significant role in the

importance of a given game and therefore the size of the crowd that the game draws. Over the last five years, the Cornhuskers' toughest opponents, i.e., the opponents drawing the largest crowds, were (1) Ohio State, (2) Wisconsin, (3) Northwestern, (4) Michigan State, (5) Iowa, and (6) Purdue. Hotspot size can be defined as (1) the number of congested lanes, (2) the number of congested segments, and (3) congestion duration.

Finally, given the start time of the game (noon or evening), the toughness of the opponent, and the specific congested segments on each route, traffic speeds on the following year's game days (2018) were forecast using Dynamic Bayesian Networks, and hotspot clusters were identified based on the dataset of predicted traffic speeds. Data from 2018 were utilized as a validation dataset.

1. Introduction

1.1 Background

Monitoring the performance of the transportation system is a fundamental element of any transportation operation and planning strategy. Traditionally, transportation system performance monitoring was based on average travel times. However, travel time is not capable of representing the quality of service that commuters experience daily and may also inaccurately reflect the actual level of congestion by not accounting for unexpected congestion.

Traffic congestion directly translates into transportation cost and plays a key role in assessing the performance of the transportation systems and the impacts of planning decisions. When a road reaches its capacity, every additional vehicle creates overload, which in turn delays other vehicles. Increased travel times, accidents, unpredictability of arrival times, increased fuel consumption, and increased pollution emissions are some of the impacts of congestion.

Generally, there are two types of congestion: recurring and non-recurring. Recurring congestion is caused by routine traffic in a normal environment and is repetitive in nature and observed during peak periods, whereas non-recurring congestion is unexpected and is most likely caused by an incident. Non-recurring congestion may also result from a variety of other factors, such as lane-blocking crashes, disabled vehicles, work-zone lane closures, and adverse weather conditions. For urban road networks, travel time (and indirectly delay) is the most commonly used indicator to determine whether the congestion is recurring or non-recurring. Since unexpected incidents are the predominant source of travel time unreliability (Hojati et al. 2016), it is crucial to predict the performance of the transportation network during unusual conditions and plan a set of actions to enhance the mobility and safety of travelers.

Daily congestion is common in many US cities, and most travelers expect and plan for some delay, particularly during peak hours. Most commuters modify their schedules or budget extra time to allow for traffic delays. It is the unexpected congestion that worries travelers the most. Travelers want to have a reliable travel time and want to be confident that a trip that takes 30 minutes today will also take 30 minutes tomorrow. Travel time reliability reflects the extent of this unexpected delay. Reliability is formally defined as the consistency or dependability in travel times, as measured from day to day and/or across different times of the day.

1.2 Planned Special Events (PSE)

Non-periodic events with an expected large attendance (known as planned special events [PSE]), such as concerts, football games, etc., play a major role in transportation delays (Kwoczek et al. 2014). Although such events are mostly different from each other, they all have one attribute in common: they impose a non-recurring stress on the transportation network, which leads to safety risk, capacity reduction, and demand surge.

The presence of a professional and college sports team in a city can have a considerable impact on the local economy of that city. Previous research has focused mainly on assessing the benefits of professional and college sports teams to the local economy, without any focus on the direct and indirect costs generated by professional and college sports teams and their games. Direct costs include facility construction; salaries for players, managers, and officials; and the costs associated with public safety at games. Indirect costs come from traffic, crowds, trash and pollution, noise, crime, and other negative aspects of the games. A thorough understanding of both the benefits and costs of professional and college sports teams provides context for understanding the public subsidies provided to professional and college sports teams.

In this report, we empirically analyze the relationship between attendance at National Collegiate Athletic Association (NCAA) Division I Football Bowl Subdivision (FBS) games and traffic in a US metropolitan area, an indirect cost associated with the presence of a college football team.

The FBS is the most competitive subdivision of NCAA Division I, which itself consists of the largest and most competitive schools in the NCAA. As of the 2017 college football season, there are 10 conferences and 130 schools in the FBS. College football is very popular in the US, and the top schools generate tens of millions of dollars in yearly revenue. The top FBS teams attract thousands of fans to games, and the largest American stadiums by capacity all host FBS teams. Football teams typically play at least six home games per season.

Memorial Stadium in Lincoln, Nebraska, is the home of the Nebraska Cornhuskers football team. With an extended capacity of more than 85,000, the stadium is commonly referred as the “third-largest city in Nebraska” on game days. The stadium holds the NCAA record for consecutive sellouts for every game since 1962, a streak of more than 300 games. Game days, therefore, typically affect the travel patterns of Lincoln and its neighboring regions. Most of the existing research on the economic costs associated with professional and college sports has focused on either the financial costs associated with facility construction or the crime associated with events held in sports facilities. However, little research has focused on the direct costs generated by games, such as the costs associated with public safety and sanitation, or indirect costs, such as the opportunity cost of funds used to subsidize the construction and operation of sports facilities. This report focuses on the relationship between professional and college sports events and traffic congestion, another overlooked cost of hosting sporting events.

1.2.1 INRIX Data Sources

In this study, we utilized historical and real-time traffic data, including speeds, travel times, and location information, collected through the INRIX traffic message channel (TMC) monitoring platform. The INRIX dataset is currently regarded as the largest crowd-sourced traffic dataset. With the help of today’s technologies, including connected vehicles and smartphones, INRIX offers a vast amount of historical and real-time data that can be analyzed and investigated to improve the performance of transportation networks. INRIX’s historical traffic flow data includes spatial and temporal data on average speeds for major roadways and arterials across all 50 states. These speeds are determined by algorithms that evaluate multiple years’ worth of data collected using INRIX’s patented Smart Dust Network system, which reports speed values on roads across the country. The speed data are then processed across several different temporal resolutions and reported on a customer-configurable basis for each temporal resolution.

INRIX derives historical flow data using the following:

- Traffic sensors – Sensors put in place by local departments of transportation (DOTs) or private sector companies that report traffic speed or other data from which traffic speed can be inferred. The sensors utilize one of several types of technology:
  - Induction loop sensors embedded in the roadway
  - Radar sensors
  - Toll tag readers along stretches of roadway
- Probe vehicles – The INRIX network includes hundreds of thousands of probe vehicles—trucks, taxis, buses, and passenger cars with onboard global positioning system (GPS) devices and transmitting capability—to relay vehicle speed and location back to a central facility. INRIX has agreements with several fleets to obtain the speed and location data anonymously.
- INRIX Smart Dust Network – This network combines real-time GPS probe data from more than 650,000 commercial vehicles across the US that travel on specific road segments during particular time windows, physical sensor information, and other real-time traffic flow information with hundreds of market-specific criteria that affect traffic, such as construction and road closures, real-time incidents, sporting and entertainment events, weather forecasts, and school schedules. The Smart Dust Network gathers all input points, weights them appropriately based on input quality and latency, and calculates the speeds on a given road segment to a measured degree of accuracy.

1.2.2 INRIX Data Format

All INRIX historical traffic flow data for the state of Nebraska were delivered in comma-separated value (CSV) format. The data provided by INRIX contained the following information:

- TMC ID – the basic spatial unit used by INRIX to report the traffic flow data; INRIX uses a nine-digit TMC ID to define a unique segment
- Time segment – a 19-digit time format used by INRIX to define the year, month, day, hours, minutes, and seconds (e.g., 2014-09-30 23:59:33 for September 30, 2014 at the 23rd hour, 59th minute, and 33rd second) for each TMC
- Speed – the average speed for a given TMC code, calculated using live data from the most current time slice
- Reference speed – an uncongested "free-flow" speed determined for each TMC segment using the INRIX traffic archive
- Average speed – the historical average mean speed for the reporting segment for that time of day and day of the week in miles per hour
- Travel time – an attribute reported by INRIX based on an aggregation of data provided by GPS probes
- Confidence – an attribute reported by INRIX having three levels: 10, 20, and 30. A

confidence of 30 indicates that enough base data were available to estimate traffic conditions in real-time, rather than using either historical speed based on time of day and day of week (indicated by confidence of 20) or free-flow speed for the road segment (indicated by a confidence of 10).

- C_value – the confidence value (ranging from 0 to 100), designed to help agencies determine whether the INRIX value meets their criteria for real-time data

An instance of Nebraska data is shown in Figure 1.1.

```
118+04749,2016-09-01 00:00:22,69.0,66.0,67.0,0.03,30.0,96.0
118-08667,2016-09-01 00:00:22,57.0,57.0,57.0,8.58,20.0,0.0
118N08667,2016-09-01 00:00:22,12.0,12.0,20.0,0.03,20.0,0.0
118+13461,2016-09-01 00:00:22,64.0,64.0,62.0,4.85,20.0,0.0
118-11314,2016-09-01 00:00:22,61.0,61.0,62.0,1.8,20.0,0.0
118-08816,2016-09-01 00:00:22,60.0,60.0,60.0,14.45,20.0,0.0
118-08813,2016-09-01 00:00:22,59.0,59.0,60.0,2.51,20.0,0.0
118P10037,2016-09-01 00:00:22,59.0,59.0,58.0,0.01,20.0,0.0
118+07616,2016-09-01 00:00:22,64.0,64.0,61.0,5.62,20.0,0.0
118-11309,2016-09-01 00:00:22,60.0,60.0,60.0,4.93,20.0,0.0
118+08795,2016-09-01 00:00:22,64.0,64.0,64.0,29.0,20.0,0.0
118P11076,2016-09-01 00:00:22,62.0,62.0,60.0,0.09,20.0,0.0
118-07590,2016-09-01 00:00:22,34.0,34.0,40.0,3.37,20.0,0.0
118+10577,2016-09-01 00:00:22,37.0,60.0,61.0,17.98,30.0,0.0
```

Figure 1.1. An instance of Nebraska INRIX data

1.3 Hotspot Detection

Hotspot detection is used in many disciplines, such as in crime analysis for analyzing where crimes occur with a certain frequency, in fire analysis for studying the phenomenon of forest fires, and in disease analysis for studying the localization and focus of diseases.

In the transportation field, a realistic scenario involving the application of hotspot detection is in traffic incident detection. Suppose that there are several detectors across a city recording the speeds of vehicles passing the detectors, and consider the vehicles' speeds on normal days over multiple years to be the baseline information and the vehicles' speeds on game days over multiple years to be the case dataset. The goal in hotspot detection is to detect those spatiotemporal regions that contain unexpected lower speeds that lead to non-recurring congestion.

In addition to detecting hotspots, this study aims to identify the factors that affect the sizes of hotspots, their locations, and other possible parameters.

1.4 Report Organization

This report is organized as follows. A literature review summarizing previous pertinent studies is provided in Chapter 2. Chapter 3 presents the data used in this study, describes the routes selected, and provides some preliminary analysis. In Chapter 4, the experiments and results are explained in detail, a complete traffic hotspot analysis is presented, a novel hotspot detection method is proposed, and insights into the observed results are provided. The report concludes in Chapter 5 with a summary of the findings of this study and a discussion of recommendations for future research.

2. Literature Review

2.1 Introduction

This chapter provides a review of previous studies conducted on probe data, planned special events and their impact on traffic congestion and travel behavior, and methods for detecting hotspots during special events.

2.2 Planned Special Events (PSE)

Traffic congestion represents a significant problem in many urban areas. Duranton and Turner (2011) note that in 2001, the average American household spent more than 2.5 person-hours each day in a passenger vehicle. They also investigated the effects of road construction and other factors on congestion. Rappaport (2016) extended the standard monocentric city model to include commuting and identified traffic congestion as a critical factor constraining local growth. Another recent study concluded that commuting to and from work is among urban households' least enjoyable activities, suggesting that additional time spent in a car at the end of the day involves substantial psychic costs.

Non-periodic events with large attendance (i.e., PSEs) play a significant role in transportation delays (Kwon et al. 2006). Although such events are mostly different from each other, they all have one feature in common: they impose a non-recurring stress on the transportation network, which leads to safety risk, capacity reduction, and demand surge. Major events are discussed in many studies. They can be recognized by their larger spatio-temporal size compared to recurring congestion, but they are not well defined. Müller (2015) proposed a methodology containing four parameters for defining major events: number of visitors, media coverage, costs, and urban transformation (Müller 2015). The *Handbook for Event Transportation* (Handbuch Eventverkehr) similarly categorizes events according to a substantial list of factors, including but not limited to the number of expected visitors, relative size, open or closed access, location, whether the event is weather dependent, duration, and financing (Amini et al. 2016). As an example of the congestion generated by large events, a concert by Rihanna in South Africa in October 2013 forced people who were trying to reach the stadium to sit in traffic for more than five hours. Similarly, a concert by Robbie Williams in London in 2003 created tailbacks of up to 10 miles on highway A1 towards the stadium. Traffic congestion created by special events has a typical pattern, including two sequential waves of congestion (Leilei et al. 2012). The first wave consists of people going to the event, while the second consists of people leaving the venue. Interestingly, the second wave may be even bigger that the first.

Few studies have been conducted to predict congestion due to special events. At the same time, there is almost no way to predict this kind of non-recurring congestion ahead of time. In this report, we examine the effects of one specific type of special event, football games, on traffic patterns and travel behaviors in the city of Lincoln, Nebraska.

It is worth noting that the relationship between urban vibrancy, traffic congestion, and greenhouse gas emissions has been investigated; the presence of a professional sports team in a city could represent a type of consumer amenity that contributes to urban vibrancy. Professional sporting events attract large numbers of fans attending games in a small area at the same time. The presence of large surface parking lots and parking structures near sports facilities indicates that large numbers of fans drive to games. Many professional sporting events take place on weekend evenings, and many sports facilities are located in the urban core of large cities. Taken together, this suggests that sporting events could have a substantial impact on traffic congestion. Basic “back-of-the-envelope” estimates of annual vehicle miles travelled (VMT) based on National Household Travel Survey (NHTS) data and actual FBS attendance suggest that fan travel to football games could account for as much as one-half of one percent of annual metropolitan area VMT, which could plausibly affect local traffic congestion.

2.3 Professional Sporting Events

Fan attendance represents the key link between sporting events and urban traffic. To attend a sporting event, most fans travel between their home or place of work and the venue where the event takes place. Fan attendance at professional sporting events concentrates economic activity spatially in and around facilities and temporally on game day. This concentration has clear economic impacts.

Humphreys and Zhou (2015) developed a spatial economic model that includes agglomeration effects stemming from increased fan activity in and around professional sports facilities on game day that predicts that the presence of a professional sports team will increase nearby property values and induce other service-providing firms to collocate near the sports facility. Huang and Humphreys (2014) found evidence of increased housing market activity near new sports facilities after the facilities opened, supporting the predictions of the model by Humphreys and Zhou (2015). If this housing market activity reflects the immigration of new residents, the population density near sports facilities will increase. Coates and Humphreys (2003) show that employees in the amusements and recreation industry—the industry that includes athletes and other employees working in sports facilities—earn more in cities with professional sports teams than employees in this industry in cities without professional sports teams; these results support the idea of increased economic activity in and near sports facilities (Coates and Humphreys 2003).

Despite this evidence of increased economic activity near sporting events, no evidence exists to support the idea that professional sports teams or facilities generate broader economic benefits across metropolitan areas. However, the concentration of fans around sports facilities on game days, along with an increase in the nearby population, has clear consequences for traffic near sports facilities. Most professional sports facilities are located in or near the central business district (CBD) in their respective cities, which also contains many firms employing large numbers of workers who travel to and from their residence on weekdays, often by car. Many fans drive to games and park in dedicated lots surrounding sports facilities or in nearby lots and parking structures that are also used by local workers and residents.

A few papers in the geography literature have examined the effect of sports facilities on local parking and traffic. All are case studies, and most use surveys of local residents to assess the extent to which increased traffic, parking, crowds, and noise on game days are perceived as a “nuisance” externality by local residents. Mason et al. (1983) used household surveys to assess the importance of negative externalities generated by games played in a football stadium in Southampton, England; the paper concluded that traffic and parked cars generated substantially larger “nuisance” externalities on game days than crowds or noise, and the negative effects of traffic and parking extended several miles from the stadium (Mason et al. 1983). Chase and Healey (1995) assessed the importance of negative externalities generated by games played and rock concerts held in a football stadium in Ipswich, England; this paper also concluded that traffic and parked cars were the largest “nuisance” externalities associated with football matches and found a similarly large traffic impact area. Chase and Healey (1995) discussed proposed stadium location decisions in Australia in light of Australian transportation policy initiatives and the existing transportation environments around several rugby and Australian Rules Football stadiums located in the center of larger Australian cities. Although this paper did not gather empirical evidence, the discussion highlights the importance of increased local traffic and parking on game day.

Little research has focused on the direct costs generated by games, such as the costs associated with public safety and sanitation, or indirect costs, such as the opportunity cost of funds used to subsidize the construction and operation of sports facilities. In one such study, Pyun and Hall (2019) reviewed the existing evidence on the relationship between professional sporting events and crime. Nevertheless, case study-based evidence clearly indicates that additional traffic around sports facilities on game days represents an important “nuisance” externality to residents of areas near stadiums in England and Australia. The existing theoretical and empirical evidence on professional sports teams in North America suggests that stadiums and arenas concentrate fans and economic activity in and near sports facilities on game days and may also increase the number of businesses and residents near sports facilities. All of these factors could increase traffic. However, the perceptions of residents near sports facilities about traffic conditions on game days may not reflect outcomes across the broader metropolitan area, and a concentration of fans and economic activity near a sports facility may not increase overall traffic in a metropolitan area. A full understanding of the potential impact of sporting events on traffic in metropolitan areas requires a model that determines realized driving outcomes.

In general, predicting traffic congestion in urban environments is a highly complex task. Early approaches to traffic prediction used simulations and theoretical modeling (e.g. Clark 2003, Chrobok et al. 2004). More recently, thanks to the availability of massive new datasets on traffic, several different statistical and data-driven approaches have been presented. Examples include generalized linear regression (Zhang and Rice 2003), nonlinear time series (Ishak and Al-Deek 2002), Kalman filters (van Lint 2008), support vector regression (Wu et al. 2004), and various neural network models (van Lint 2008, Park et al. 1999, Vanajakshi and Rilett 2004). A combination of some of the latter approaches is used by current commercial navigation solutions, which are able to predict recurring congestion by identifying characteristic traffic flow patterns on street segments based on historical data. These commercial systems can also optimize route planning based on the real-time traffic situation.

In general, traffic congestion can be divided into recurring congestion, usually caused by a mobility demand that exceeds the capacity of the road network (e.g., due to rush hour), and non-recurring congestion (e.g., due to incidents or special events) (Kwon et al. 2006). The effects of non-recurring traffic congestion and the prediction of this type of congestion are widely investigated topics within the research community (e.g., Miller and Gupta 2012, Pan et al. 2012, Pan et al. 2015). Although approaches to predicting non-recurring congestion have improved significantly over time, most use data from stationary loop sensors that are not always capable of reflecting the traffic state at the level of granularity required for urban scenarios. In addition, the focus of these approaches has been on unidirectional street segments such as highways, whereas usually in cities the impact of congestion is multidimensional, evolving in a two-dimensional (2D), more complex route network. Previous studies have highlighted that PSEs are possible influencing factors on non-recurring congestion (Kwon et al. 2006, Ishak and Al-Deek 2002, Horvitz et al. 2005), since they may lead thousands (or even hundreds of thousands) of people to travel towards and then away from the same destination in a very limited time span.

To the best of our knowledge, the only work available that focuses on the influence of PSEs on traffic is presented in Kwoczek et al. (2014). The authors present a generic overview of the influence of PSEs on road networks, derived from an event classification system defined by the Chinese State Council. The authors also introduce management plans for different types of events, but there is no quantifiable solution for predicting traffic.

In the present report, we make use of INRIX probe data to analyze the influence of PSEs on traffic and make planning decisions based on that. However, first it is crucial to explain what probe-sourced data are.

2.4 Widely Available INRIX data

As demand for comprehensive traffic monitoring grows from both travelers and transportation agencies, a new technology that would reduce the installation and maintenance costs of monitoring systems is needed for collecting accurate and real-time traffic details. Probe-based methods of measuring travel time and speed data can easily scale across large networks without the need for deploying any additional infrastructure (Young 2007).

The emergence of probe vehicle technology, the use of which has grown over the past few years, has caused a remarkable change in traffic data collection, processing, analysis, and utilization. The ability to access a huge volume of historical and real-time traffic data without any of the costs of installation, configuration, and maintenance of infrastructure-mounted sensors interests many agencies that want to utilize a single, uniform data source for monitoring traffic conditions across most routes in the US. Traffic information is collected from millions of cell phones, vans, trucks, connected cars, commercial fleets, delivery vehicles and taxis, and other GPS-enabled vehicles. At present, several probe data vendors, such as INRIX, HERE, TomTom, NAVTEQ, and TrafficCast, provide broad and high-quality real-time and historical traffic data around the world.

INRIX provides updates on speed, travel time, incidents, and data quality along each mile-long travel segment at a frequency of once every minute. For the entire Nebraska roadway system, the stream for the INRIX TMCs comprises approximately 9 to 10 GB/month, or more than 100 GB/year, and for XD segments the stream is approximately 45 GB/month, or more than 545

GB/year. With the introduction of greater spatial coverage and resolution, the size of the input streams is expected to increase (Cookson and Pishue 2017).

Many studies have been conducted comparing the accuracy and reliability of probe-sourced data against that of local sensor data, such as data from radar sensors and loop detectors, which are considered the benchmark (Feng et al. 2010, Coifman 2002, Lindveld et al. 2000, Kim and Coifman 2014, Hu et al. 2015, Mudge et al. 2013). Kim and Coifman (2014) showed that INRIX speeds tend to lag behind the speeds measured by loop detectors by almost 6 minutes. Although INRIX reports two measures of confidence, these confidence measures do not appear to reflect this latency or the occurrence of repeated INRIX-reported speeds. Kim and Coifman (2014) used two months of INRIX data against the concurrent loop detector data to evaluate INRIX's performance during both recurrent and non-recurrent events on 14 miles of I-71. To calculate the amount of latency, the authors used a correlation coefficient with several months of continuous data from concurrent detectors while shifting the time-series loop detector with 10 second steps.

The Federal Highway Administration (FHWA) conducted a survey to gather information on (1) products and services offered by private sector data providers and (2) the use of those private sector data products and services by public sector agencies. The FHWA found that agencies are using a range of data sources, including GPS data from fleet vehicles, commercial devices, cell phone applications, fixed sensors installed and maintained by other agencies, fixed sensors installed and maintained by data providers, and cell phone locations. Most providers did not disclose specific quality evaluation results or quality assurance algorithms. INRIX explicitly stated its capability of meeting an availability level of more than 99.9% and an accuracy of greater than 95% (FHWA 2016).

Nanthawichit and Nakatsuji (2003) proposed a method for treating probe vehicle data together with fixed detector data to estimate the traffic state variables of traffic volume, space mean speed, and density. The method uses a macroscopic model along with the Kalman filtering technique and was verified with several sets of hypothetical traffic data. The authors suggested the possibility of using estimated/predicted states to estimate/predict travel time.

Coifman (2002) investigated various means of measuring link travel times on freeways. He used basic traffic flow theory to estimate link travel time using point detector data without the need for any new hardware.

Sadrsadat and Young (2011) worked on the I-95 Corridor Coalition's Vehicle Probe Project (VPP) to determine the probability that traffic data are reported in real-time as a function of hourly volume. The authors compared the VPP data against travel time data collected using Bluetooth traffic monitoring equipment. The VPP provides an indication that traffic data are reported in real-time data by a confidence score attribute equal to 30; the confidence score is provided by INRIX. The study confirmed the increasing availability of real-time data with increasing traffic volume, as measured by the percentage of confidence scores of 30.

Feng et al. (2010) investigated the analytical relationships between travel time prediction/estimation accuracy and sensor spacing by means of two basic travel time prediction/estimation algorithms. The authors also measured probe vehicle penetration rate. Travel times estimated and predicted online using induction loop detectors were evaluated against observed travel times. The findings of the study provide support for detector placement and probe vehicle deployment, especially along freeway corridors with existing detectors.

Lindveld et al. (2000) found reasonably accurate results (10% to 15% root mean square error proportions) for travel time prediction/estimation accuracy across different sites for uncongested to lightly congested traffic conditions. They used various travel time estimators, but only speed-based travel time estimators could be tested under congested conditions.

The Florida Department of Transportation (FDOT) used several metrics, such as absolute average speed error, average speed bias, absolute average travel time error, and travel time bias, to determine the accuracy of vendors' (NAVTEQ, TrafficCast, and INRIX) data. Overall, the data looked consistent with the ground truth and license plate reader data, and no significant differences in data accuracy among the three vendors were observed (FDOT 2012).

Sharma et al. (2017) explored the reliability of probe data for congestion detection and overall performance assessment using an adaptive, data-driven, multiscale data decomposition algorithm called Empirical Mode Decomposition. The authors noted that the cost of deploying large-scale control strategies for traffic networks has increased the need for more reliable real-time traffic condition prediction.

Liu et al. (2016) discussed two approaches for travel time prediction/estimation accuracy : dynamic mode decomposition and spatiotemporal pattern networks. Their results showed that data-driven approaches effectively detected changes in traffic system dynamics during different times of the day.

A technical memorandum published by FDOT (2012) summarizes the various data available for analyzing bottlenecks and congestion on Florida's Strategic Intermodal System. This technical memorandum also makes recommendations concerning the applicability of using existing FDOT data versus vehicle probe data from INRIX.

Schuman and Glancy (2015) discussed how INRIX launched the world's first crowd-sourced traffic monitoring network using sensors in fleet vehicles and mentioned how INRIX XD gives greater traffic detail on any map and a platform for planning, analysis, and operation of road networks.

Matsumoto et al. (2010), using probe data to estimate CO_2 emission reductions, defined three services (traffic flow analysis, improvement of signal control performance, and priority control of bypasses) that enhance traffic flow control. The authors confirmed the detection of a bottleneck without depending on the deployment rate of in-vehicle GPS units by using probe data statistically in traffic flow analysis (Matsumoto et al. 2010).

Different techniques (data assimilation, Newtonian relaxation) to incorporate probe data into macroscopic traffic flow models have been used to solve the optimization problem in urban areas, and these techniques have confirmed the possibility of decreasing the amount of probe data needed to detect congested traffic with negligible degradation of the quality of the traffic status estimation (Chu and Saito 2013). While reducing CO_2 emissions using intelligent traffic control requires many detectors and high installation costs, Nagashima et al. (2014) used probe data collected by vehicles equipped with GPS or other devices and a signal control system that calculated consecutive spatial traffic information (spatial data) such as queue length. The authors showed that it is possible to reduce the number of detectors needed for the calculation.

Haghani et al. (2015) described a novel validation scheme for comparing travel time data from two independent data sources with an emphasis on arterial applications. In addition, a context-dependent-based travel time fusion framework was developed to integrate data from INRIX and Bluetooth datasets to improve data quality. To minimize the impact of random errors that can occur with INRIX data, two new techniques, confidence value and smoothing, were developed by a coalition of the University of Maryland and INRIX. When used together, these techniques reduce both the frequency and severity of the sudden changes in traffic condition that have been observed. Kobayashi et al. (2011) suggested using probe data to collect spatial traffic information in an effort to reduce CO_2 emissions and verified the possibility of detecting bottleneck intersections based on traffic flow analysis utilizing infrared beacon probe data collected from the field.

In the present study, we utilized the historical and real-time traffic data, including speeds, travel times, and location information, collected through the INRIX TMC monitoring platform. With the help of today's technologies, including connected vehicles and smartphones, INRIX offers a vast amount of historical and real-time data that can be analyzed and investigated to improve the performance of transportation networks. INRIX's historical traffic flow data includes spatial and temporal data on average speeds for major roadways and arterials across all 50 states. These speeds are determined by algorithms that evaluate multiple years' worth of data collected using INRIX's patented Smart Dust Network system, which reports speed values on roads across the country. The speed data are then processed across several different temporal resolutions and reported on a customer-configurable basis for each temporal resolution.

2.5 Hotspot Detection

Generally, predicting traffic congestion in urban environments is an extremely complex task. In general, two types of congestion are defined: recurring and non-recurring. Recurring congestion is caused by the usual traffic in a normal environment and is repetitive in nature and observed during peak periods, whereas non-recurring congestion is unexpected and is often caused by weather conditions, work zones, and incidents. While early approaches for traffic forecasting included simulations and theoretical modeling, the massive traffic datasets available today have made several different statistical and data-driven approaches available to the research community, including linear regression, nonlinear time series, Kalman filters, support vector regression, and various neural network models. The effects of traffic congestion and the prediction of these effects have been extensively studied. However, to the best of our knowledge, only one study has focused on the impacts of PSEs on traffic congestion (Kwoczek et al. (2014). The authors of that study present a general theory of the impact of PSEs on road networks, derived from an event classification system defined by the Chinese State Council. The authors also introduce management plans for different types of events, but there are no measurable solutions to predict traffic.

Over the years, many researchers have attempted to utilize mathematical prediction methods for traffic prediction. In the field of traffic flow prediction, traffic flow has always been regarded as a two-dimensional stochastic process (temporal and spatial). Parametric models try to find a mathematical model parameter that describes traffic flow as a time series process. In 1979, the first parameter approach was proposed to predict short-term freeway flow using an autoregressive integrated moving average (ARIMA) model. Many studies have shown the value of the ARIMA model, but the approaches in these studies suffer from a tendency to focus on the average values of the time series and therefore are not able to predict extremes. In order to predict the flow of traffic within a study area, other parametric models, such as the Kalman filtering model and local linear regression, have also been suggested.

Since 1990, researchers have tended to make use of nonparametric instead of parametric models. In order to define the model's structure and the number of parameters, nonparametric models rely on training data. While nonparametric models are promising because of the nonlinear nature of traffic flows, many of the proposed methods only characterize traffic flow temporally in a time series process. This paper investigates Bayesian networks (BN) to predict traffic flows using spatial and temporal information. Dynamic Bayesian Networks (DBN) extend Bayesian networks to model systems that evolve over time. In other words, a DBN is a BN that relates variables to each other over contiguous time stamps.

2.6 Conclusion

This chapter summarized previous studies on the impacts of various kinds of planned special events. Moreover, the impacts of professional sporting events, an example of a PSE, on traffic congestion were examined. Finally, information was presented on INRIX, the source of data for this study. The next chapter presents details on the data used and routes selected for this study and an exploratory analysis.

3. Data

3.1 Introduction

In today's complex global economy, transportation connections enable a business to locate in any region offering the best possible combination of labor, land, tax, and cost while competing worldwide. All state departments of transportation (DOTs) rely on fixed-mounted sensors to collect traffic information such as travel time, traffic speed, volume, etc. Such traffic information can be used by Nebraska Department of Transportation (NDOT) councils to identify which routes are used most and to decide whether to improve those roads or provide alternatives if there is an excessive amount of traffic.

Probe data collection involves a set of relatively low-cost methods for obtaining travel time and speed data for vehicles traveling on freeways and other transportation routes. NDOT has already procured probe data streams through a third-party vendor, INRIX, to augment traffic data collection and assess the performance of its operations. INRIX is maintaining 4,125 traffic management centers to collect traffic information for major freeways and urban areas in Nebraska.

The objective of this study was to assess and explore the impact of University of Nebraska Cornhuskers football game days on travel patterns. Game days attract a significantly high volume of traffic and hence result in congestion and higher travel times for road users. The past several years of INRIX data available through NDOT were used to generate travel time reliability curves and thereby estimate shockwave lengths.

This project provides insights on the impact of game day schedules and the Cornhuskers' opponents on travel patterns and route choices. The insights gained from this study will help NDOT implement active traffic assignment and thereby reduce congestion on game days.

Table 3.1 shows the Nebraska Cornhuskers home game schedule from 2013 to 2017. For all games, the table shows the date and day of the week, the opposing team, the game's result, and the start time of the game.

Table 3.1. Nebraska Cornhuskers home game schedule and results from 2013 to 2017

Date	Day	Opponent	Location	Result	Status	Time
Game Days 2013						
8/2/2013	Fri	Fan Day	Memorial Stadium			
8/31/2013	Sat	Wyoming	Memorial Stadium	W, 37-34		7:00 PM
9/7/2013	Sat	Southern Miss	Memorial Stadium	W, 56-13		5:00 PM
9/14/2013	Sat	UCLA	Memorial Stadium	L, 41-21		11:00 AM
9/21/2013	Sat	South Dakota State	Memorial Stadium	W, 59-20		
10/5/2013	Sat	Illinois	Memorial Stadium	W, 39-19		11:00 AM
11/2/2013	Sat	Northwestern	Memorial Stadium	W, 27-24		
11/16/2013	Sat	Michigan State	Memorial Stadium	L, 41-28		
11/29/2013	Fri	Iowa	Memorial Stadium	L, 38-17		11:00 AM
Game Days 2014						
8/30/2014	Sat	Florida Atlantic	Memorial Stadium	W, 55-7		2:30 PM
9/6/2014	Sat	McNeese State	Memorial Stadium	W, 31-24		11:00 AM
9/20/2014	Sat	Miami FL	Memorial Stadium	W, 41-31		7:00 PM
9/27/2014	Sat	Illinois	Memorial Stadium	W, 45-14	Homecoming	8:00 PM
10/25/2014	Sat	Rutgers	Memorial Stadium	W, 42-24		11:00 AM
11/1/2014	Sat	Purdue	Memorial Stadium	W, 35-14		2:30 PM
11/22/2014	Sat	Minnesota	Memorial Stadium	L, 28-24		11:00 AM
Game Days 2015						
4/11/2015	Sat	Red-White Spring Game	Memorial Stadium	Red 24, White 15		11:00 AM
8/5/2015	Wed	Nebraska Football Fan Day	Memorial Stadium		Presented by US Cellular	
9/5/2015	Sat	Brigham Young	Memorial Stadium	L, 33-28		2:30 PM
9/12/2015	Sat	South Alabama	Memorial Stadium	W, 48-9		7:00 PM

9/26/2015	Sat	Southern Miss	Memorial Stadium	W, 36-28	Homecoming	11:00 AM
10/10/2015	Sat	Wisconsin	Memorial Stadium	L, 23-21		2:30 PM
10/24/2015	Sat	Northwestern	Memorial Stadium	L, 30-28		11:00 AM
11/7/2015	Sat	Michigan State	Memorial Stadium	W, 39-38		6:00 PM
11/27/2015	Fri	Iowa	Memorial Stadium	L, 28-20		2:30 PM
Game Days 2016						
8/3/2016	Wed	Fan Day	Memorial Stadium			
9/3/2016	Sat	Fresno State	Memorial Stadium	W, 43-10		7:00 PM
9/10/2016	Sat	Wyoming	Memorial Stadium	W, 52-17		11:00 AM
9/17/2016	Sat	Oregon	Memorial Stadium	W, 35-32		2:30 PM
10/1/2016	Sat	Illinois	Memorial Stadium	W, 31-16	Homecoming	2:30 PM
10/22/2016	Sat	Purdue	Memorial Stadium	W, 27-14		2:30 PM
11/12/2016	Sat	Minnesota	Memorial Stadium	W, 24-17		6:30 PM
11/19/2016	Sat	Maryland	Memorial Stadium	W, 28-7		11:00 AM
Game Days 2017						
4/15/2017	Sat	Spring Game	Memorial Stadium	Red 55, White 7		
9/2/2017	Sat	Arkansas State	Memorial Stadium	W, 43-36		7:00 PM
9/16/2017	Sat	Northern Illinois	Memorial Stadium	L, 21-17		11:00 AM
9/23/2017	Sat	Rutgers	Memorial Stadium	W, 27-17		2:30 PM
10/7/2017	Sat	Wisconsin	Memorial Stadium	L, 38-17		7:00 PM
10/14/2017	Sat	Ohio State	Memorial Stadium	L, 56-14		6:30 PM
11/4/2017	Sat	Northwestern	Memorial Stadium	L, 31-24		2:30 PM
11/24/2017	Fri	Iowa	Memorial Stadium	W, 56-14		3:00 PM

3.2 Exploratory Analysis

The research team and the technical advisory committee for the project decided to select five major routes to Memorial Stadium in Lincoln, Nebraska. Figure 3.1 indicates these five routes, which included I-80 (No. 1), NE 2 (No. 2), NE 31 (No. 3), US 6 (No. 4), and US 77 (No. 5).

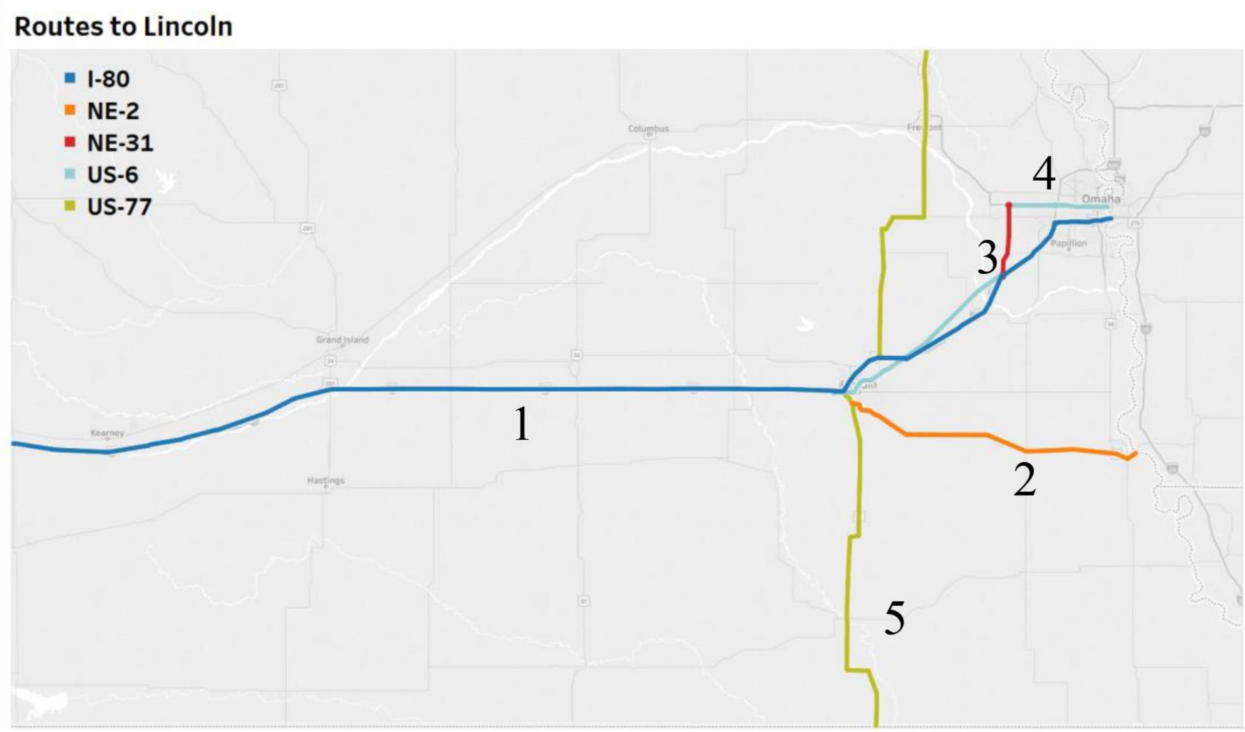
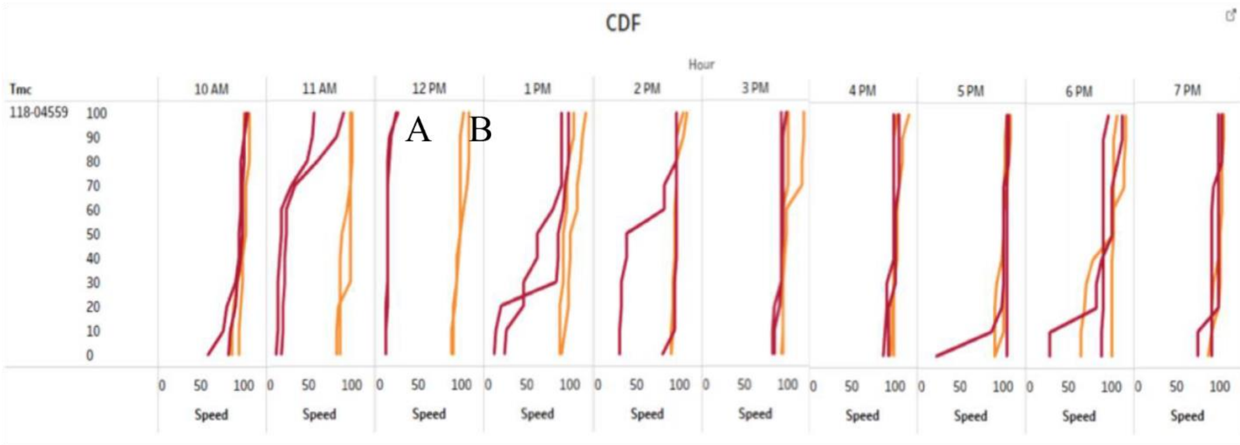


Figure 3.1. Five routes selected for this study

Raw data files received from the INRIX server were parsed using Hadoop technology and then processed using tools like Tableau and Python programming to visualize all routes and detect the mostly congested locations on each of the routes on game days. In this report, each of the five routes is separately analyzed for all game days over five years, from 2013 through 2017.

Figure 3.2 illustrates the inspiration for examining traffic speeds on game days before the start time of each game until after the end of the game.



Orange and red represent normal and game days, respectively

Figure 3.2. Hourly CDFs of speeds on two game days and two normal days for a sample game starting at 2:30 PM

The horizontal axis in the figure shows speed in mph, and the vertical axis represents the cumulative distribution functions (CDFs) of the speeds. The CDF is the probability that a variable takes a value less than or equal to x. The horizontal axis represents the allowable domain for the given probability function. Because the vertical axis reflects probability, it must fall between 0 and 1; it increases from 0 to 1 from left to right on the horizontal axis.

As can be seen in Figure 3.2, the CDFs of the speeds for two normal days and two game days (orange and red, respectively) start to shift in the hours before the start time of the games (11 a.m., 12 p.m., 1 p.m., 2 p.m.) and after the end of the games (5 p.m. and 6 p.m.). Take, for example, games with start times of 2:30 p.m. Point A in Figure 3.2 indicates two red lines, the CDFs of the speeds on two separate game days at 12 p.m. It can clearly be seen that the CDFs (point A) are well below 45 mph, showing congestion at 12 p.m. (almost two hours before the start time of the games), which can be contrasted to the orange lines (point B), which represent the CDFs of speeds on two separate normal days. A similar scenario is observed at 11 a.m., 1 p.m., 2 p.m., 5 p.m., and 6 p.m.

In the following sections, each route is thoroughly analyzed in terms of the congested zones identified from a couple of hours before the start time of the games to a few hours after the end of the games.

3.2.1 Route 1: I-80

First route is I-80, which, in Nebraska, runs east from the Wyoming state border across the state to Omaha. Nebraska has over 80 exits along I-80. Figure 3.3 shows I-80 in the state of Nebraska.

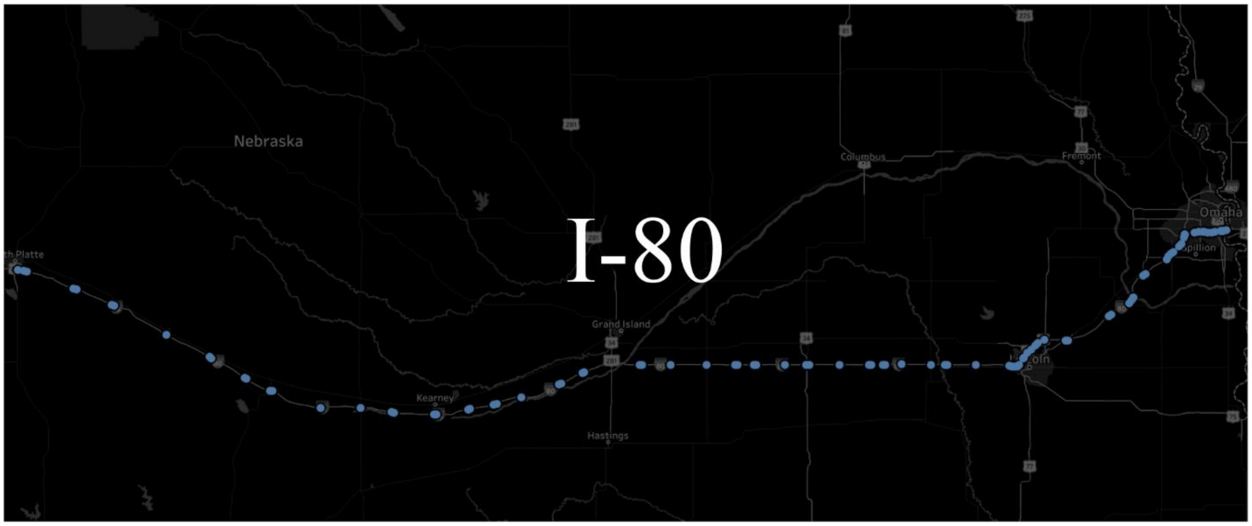


Figure 3.3. Route I-80 in Nebraska, with blue points representing INRIX TMC segments

There are several points on I-80 eastbound (EB) showing congestion during game days (from exits 353 to 369 in Figure 3.4).

Figure 3.4. Route I-80 EB, with red points representing INRIX TMC segments showing congestion on game days

When the start time of the game is 11:00 a.m. or 2:30 p.m., there is congestion on I-80 westbound (WB) from Omaha to Lincoln (Figure 3.5). However, when the start time of the game is 6:30 p.m. or 7:00 p.m., there is almost no congestion on I-80 WB from Omaha to Lincoln.

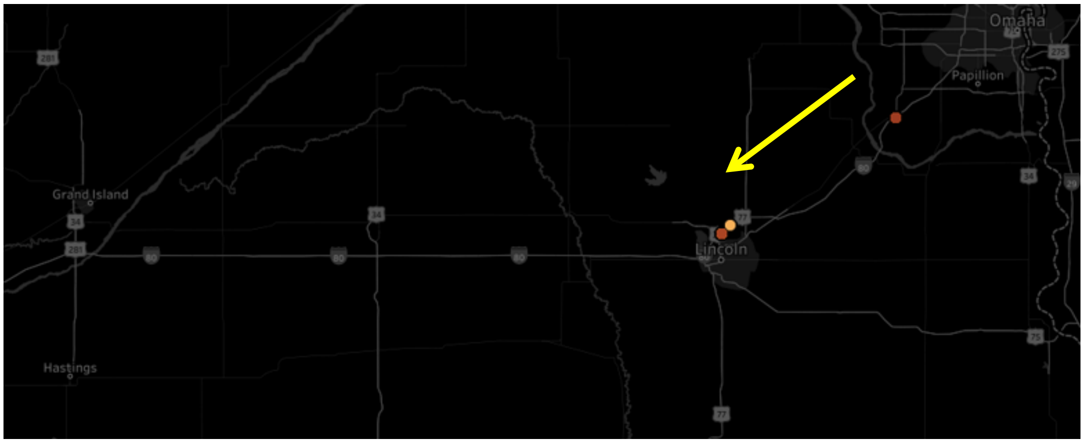
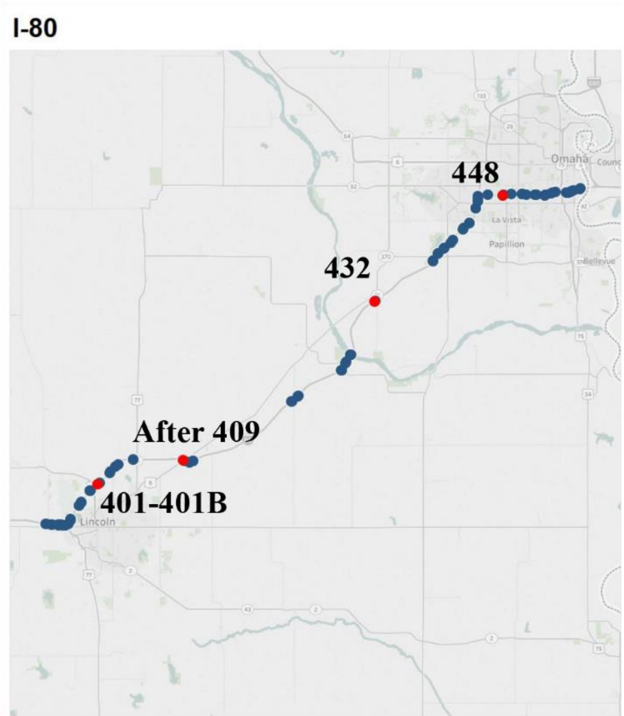
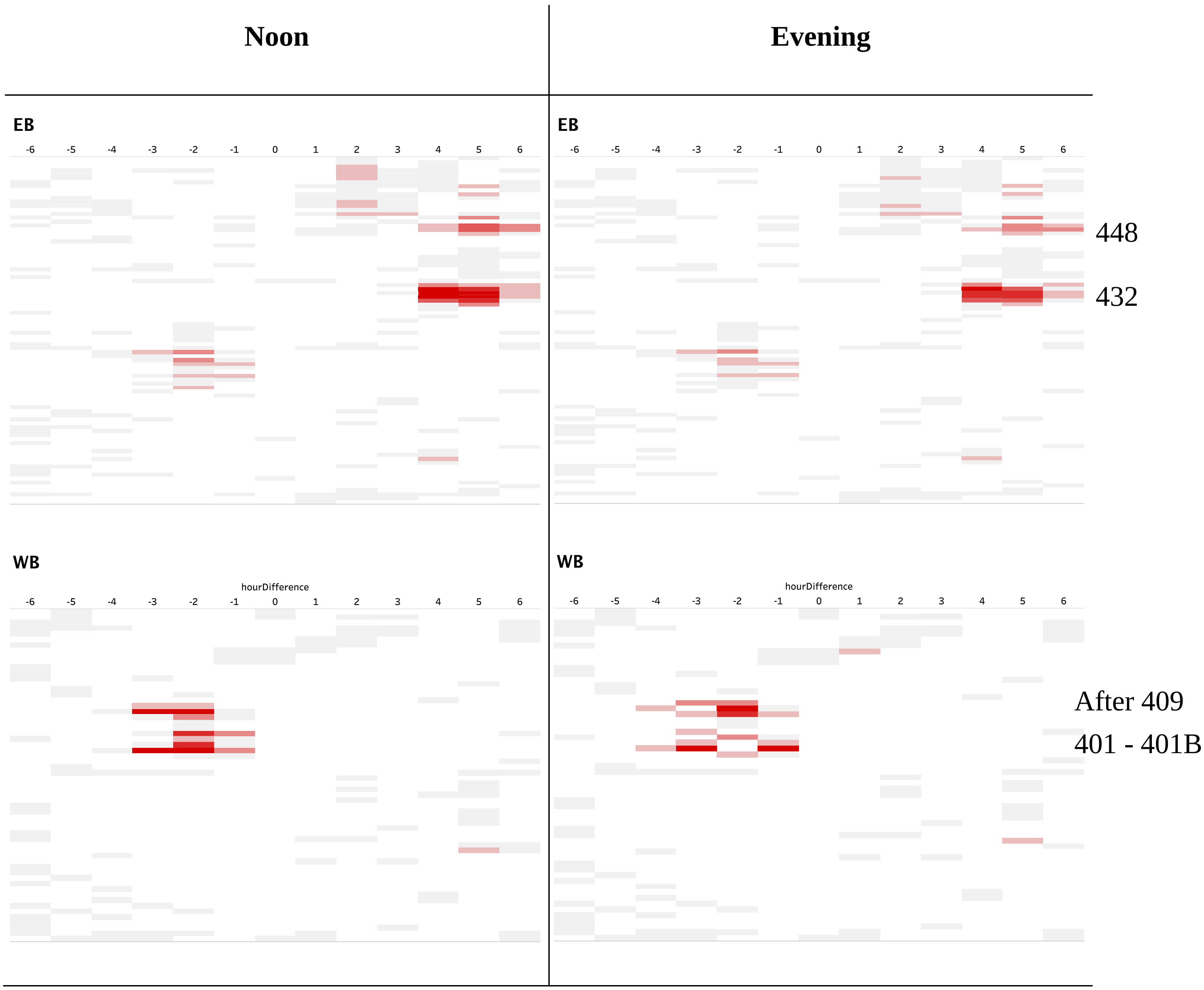


Figure 3.5. Route I-80 WB, with red points representing INRIX TMC segments showing congestion on game days

For hotspot detection, a very thorough exploratory analysis is conducted on each route. All significant speed drops from 2013 to 2017 for each segment is analyzed. If the proportion of significant speed drops to total number of game days is greater than 0.5 the segment is classified as a hotspot. For instance, if the total number of game days are 40 over the five years (2013 to 2017) and segment A experienced traffic congestion for 20 times or more during this period, that segment will be classified as a hotspot. Figure 3.6(a) shows all segments from Omaha to Lincoln (I-80 WB) as blue points. In general, blue points represent all segments on each route. Red points represent hotspot segments.


I-80
448
432
After 409
401-401B
Omaha
La Vista
Papillion
Lincoln

a)



b)

Figure 3.6. (a) Hotspots indicated by red points and (b) heat maps for I-80 EB and WB for noon and evening games

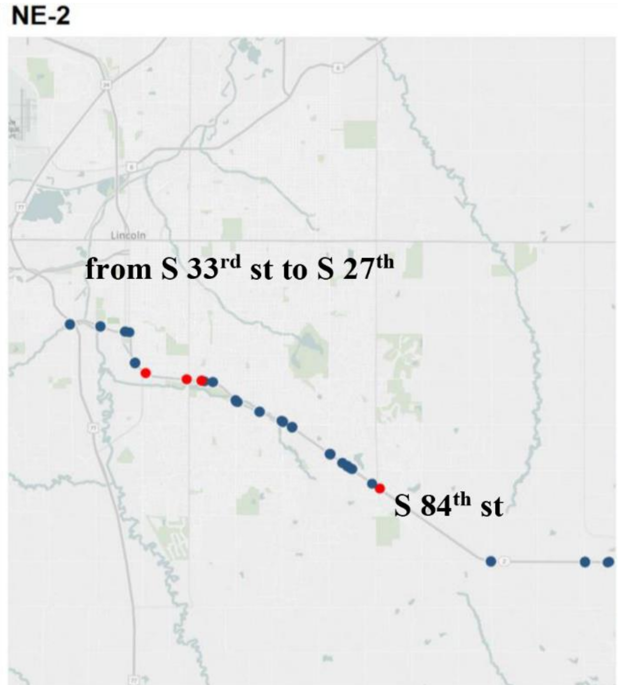
Figure 3.6(b) shows heat maps for I-80 EB and WB for noon and evening game days. Each heat map shows 0 as the start time of each game. The heat maps also show six hours before and after the start time of the games. Red point are also annotated by name of exit number or street name in the figure. Before the games, considerable congestion is evident for both noon and evening games starting from three hours before the games on I-80 WB. On I-80 EB, the heat maps show traffic congestion after the end of each game, which starts from three and a half hours after the start time of the game. The red points in Figure 3.6(a) correspond to the segments on the heat maps that show congestion. Those red points correspond to exits 448, 432, after 409, and 401-401B.

3.2.2 Route 2: NE 2

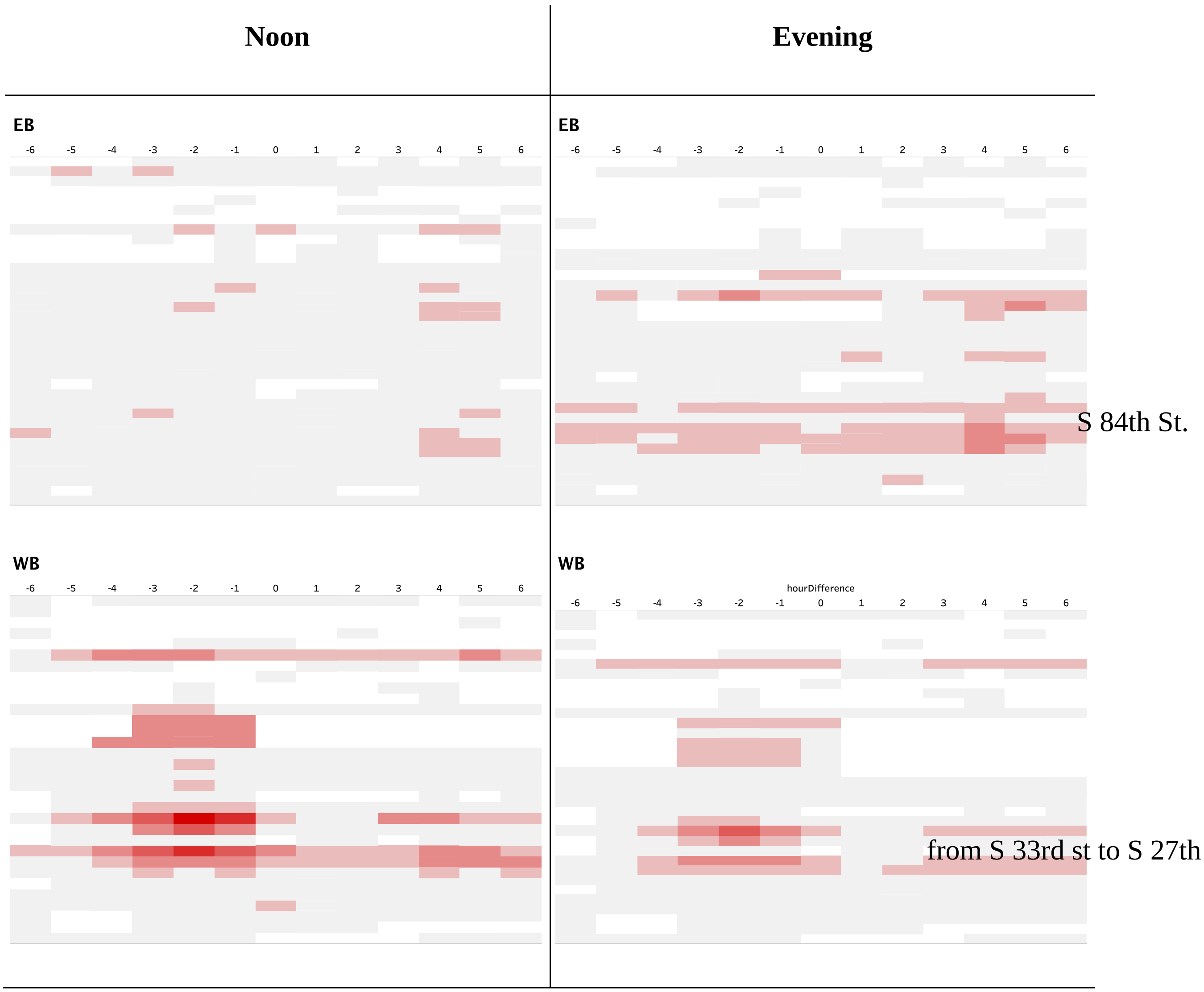
NE 2 is a highway in Nebraska with two segments. The western segment begins at the South Dakota border northwest of Crawford and ends southeast of Grand Island at the intersection with I-80. The eastern segment begins in Lincoln and ends at the Iowa border at Nebraska City. In this study, the eastern part of NE 2 is examined. Figure 3.7 shows the eastern part of NE 2.

Figure 3.7. Route NE 2, with blue points representing INRIX TMC segments

As can be seen in Figure 3.8(b), there is considerable congestion on four segments on NE 2 WB for noon games. There is no considerable congestion on NE 2 EB at all.



a)



b)

Figure 3.8. (a) Hotspots indicated by red points and (b) heat maps for NE 2 EB and WB for noon and evening games

This means that people prefer to choose an alternative route to travel east (for example, to Iowa) after the game. However, there is significant congestion on NE 2 WB before the games begin at noon, which means that people from Iowa or regions around Nebraska's eastern border prefer to use this route to travel to Lincoln for noon games. The red points in Figure 3.8(a) correspond to S 84th Street and a segment between S 33rd Street to S 27th Street.

3.2.3 Route 3: NE 31

NE 31 is a highway in Nebraska. The southern terminus is near Louisville at the intersection with NE 50. The northern terminus is near Kennard at the intersection with US 30. The highway serves as a main north-south highway in the western portion of the Omaha Metropolitan Area (Figure 3.9).

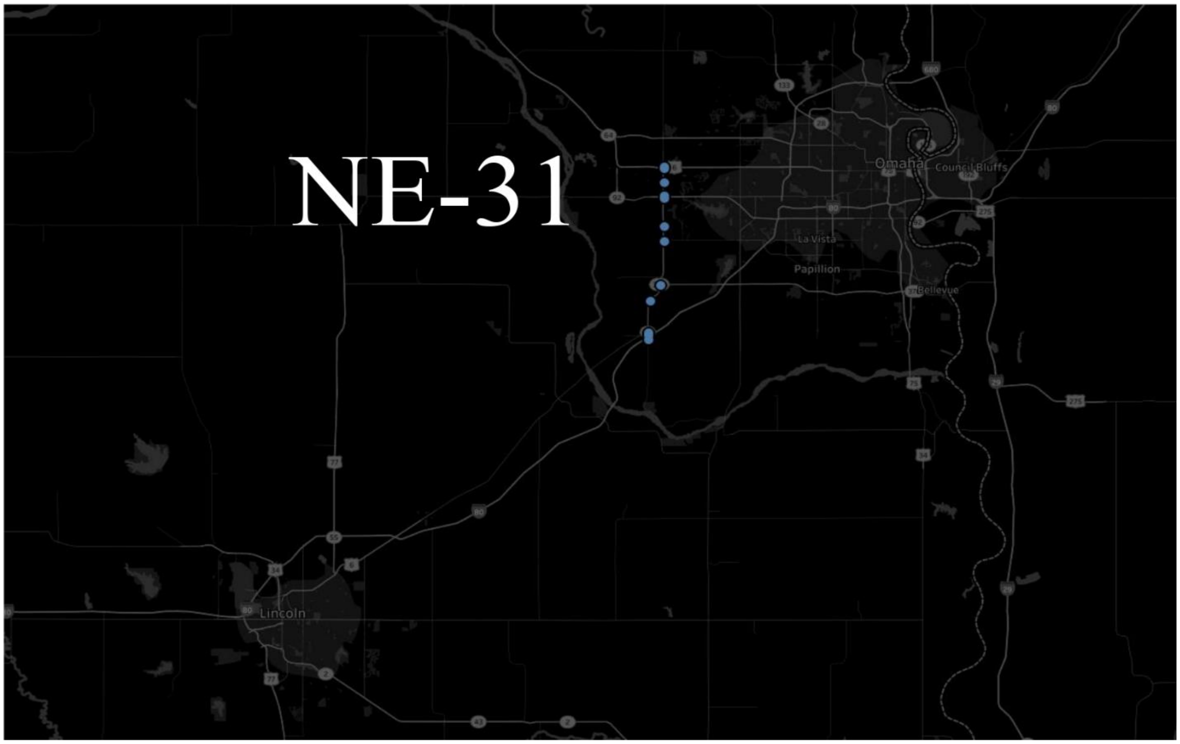
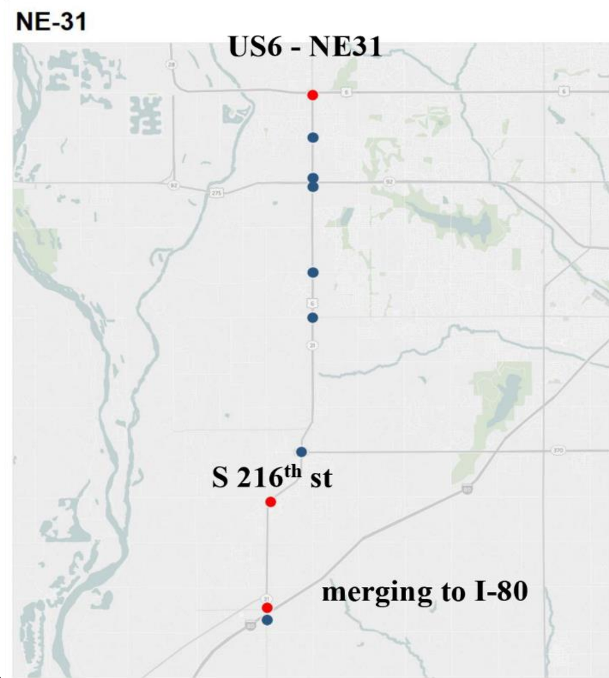
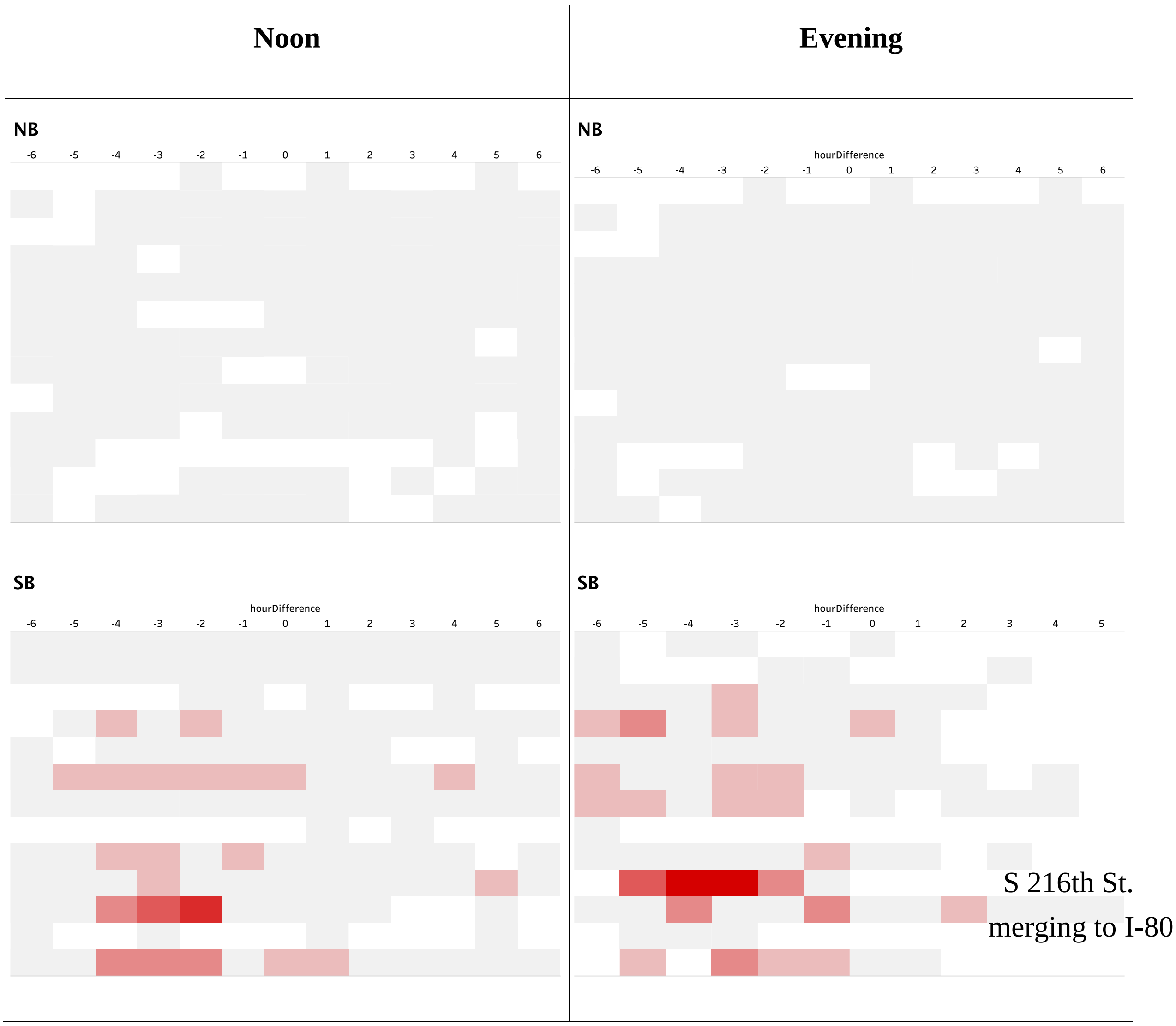


Figure 3.9. Route NE 31, with blue points representing INRIX TMC segments

As can be seen in Figure 3.10(b), there is considerable congestion on three segments on NE 31 southbound (SB) for both noon and evening games. There is no congestion on NE 31 northbound (NB) neither noon nor evening games.

a)





b)

Figure 3.10. (a) Hotspots indicated by red points and (b) heat maps for NE 31 NB and SB for noon and evening games

This means that people prefer to choose an alternative route for traveling north after the game. However, there is significant congestion on NE 31 SB before both noon and evening games. The red points in Figure 3.10(a) correspond to the intersections between NE 31 SB and US 6, S 216th Street, and the merging point to I-80.

3.2.4 Route 4: US 6

US 6 in Nebraska is a highway that goes from the Colorado border west of Imperial to the Iowa border in the east at Omaha. In Lincoln, US 6 comes into the city on West O Street, portions of which are divided highway. At Cornhusker Highway, which is a divided highway, US 6 turns east with a short urban connection to I-180 in the west along Cornhusker Highway. US 6 then follows Cornhusker Highway northeast out of the city. At the east end of Cornhusker Highway (near Waverly), US 6 meets I-80. As can be seen in Figure 3.11, US 6 finally merges onto the freeway, the West Dodge Expressway, and turns due east towards downtown Omaha.

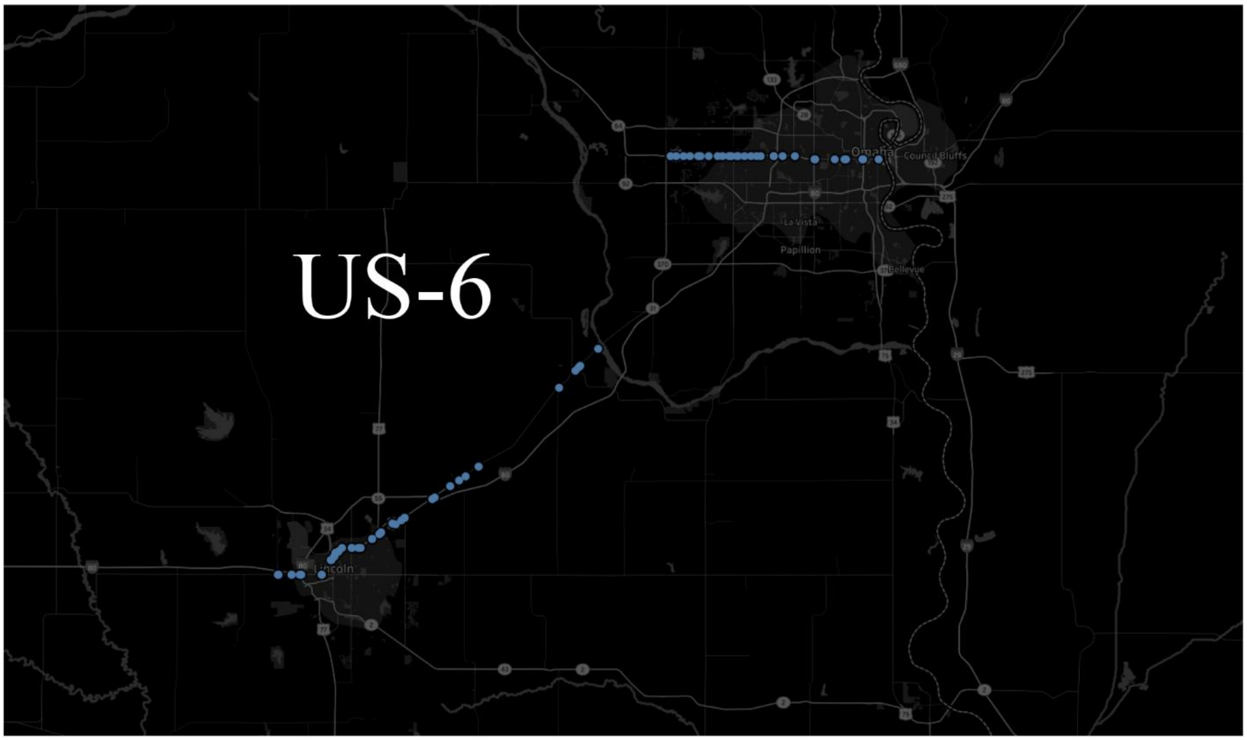
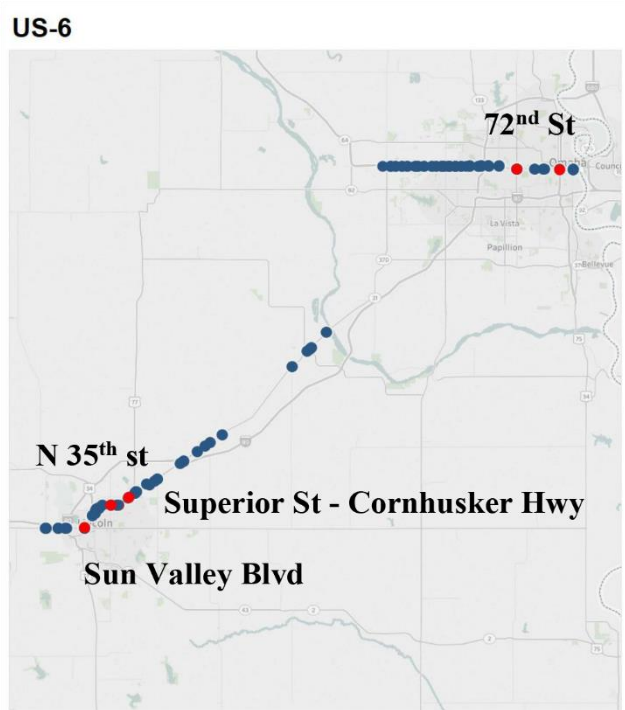
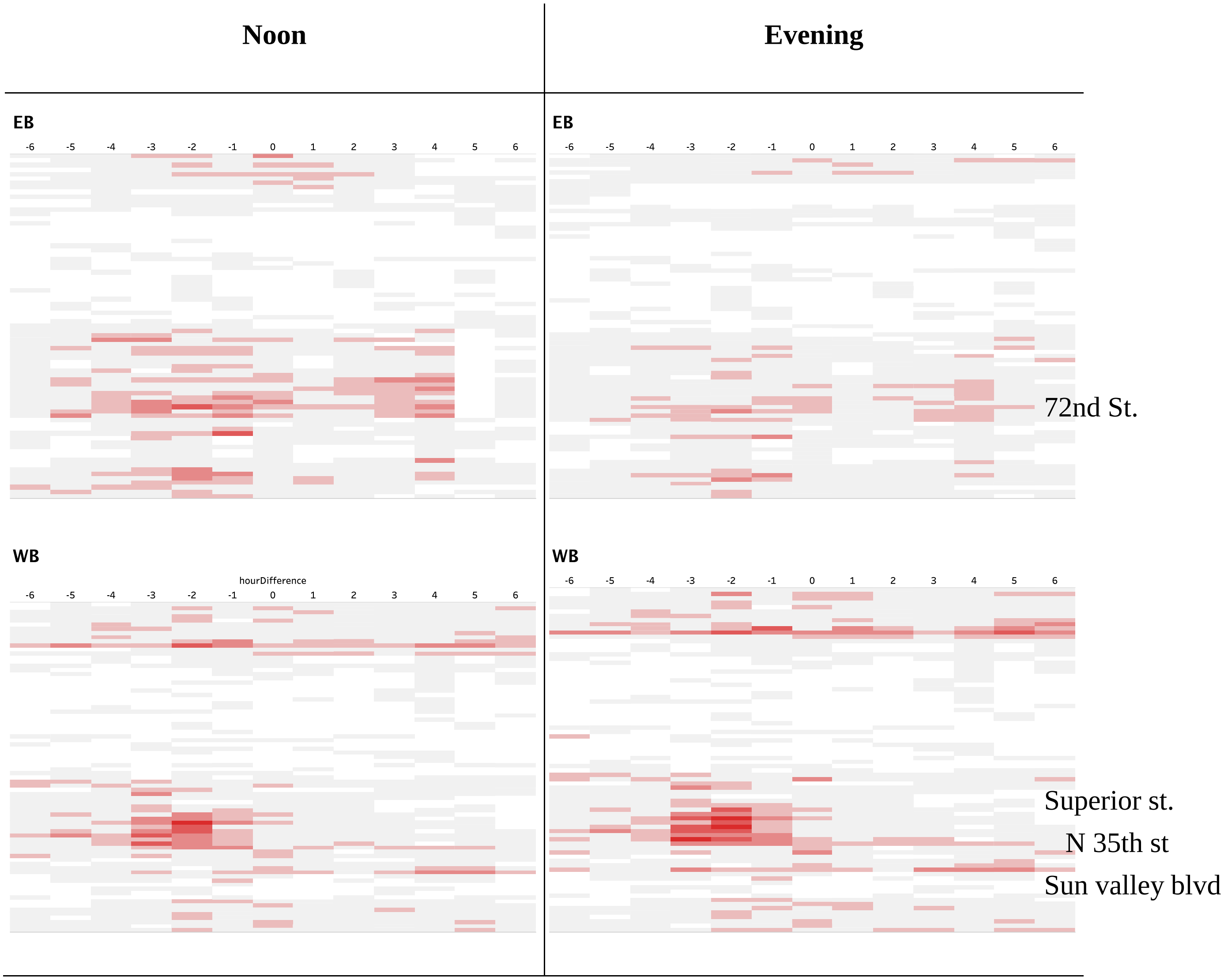


Figure 3.11. Route US 6, with blue points representing INRIX TMC segments

As can be seen in Figure 3.12(b), there is considerable congestion on eight segments of US 6 WB for both noon and evening game days. There is almost no significant congestion on US 6 EB.


US-6
72nd St
N 35th st
Superior St - Cornhusker Hwy
Sun Valley Blvd

a)



b)

Figure 3.12. (a) Hotspots indicated by red points and (b) heat maps for US 6 EB and WB for noon and evening games

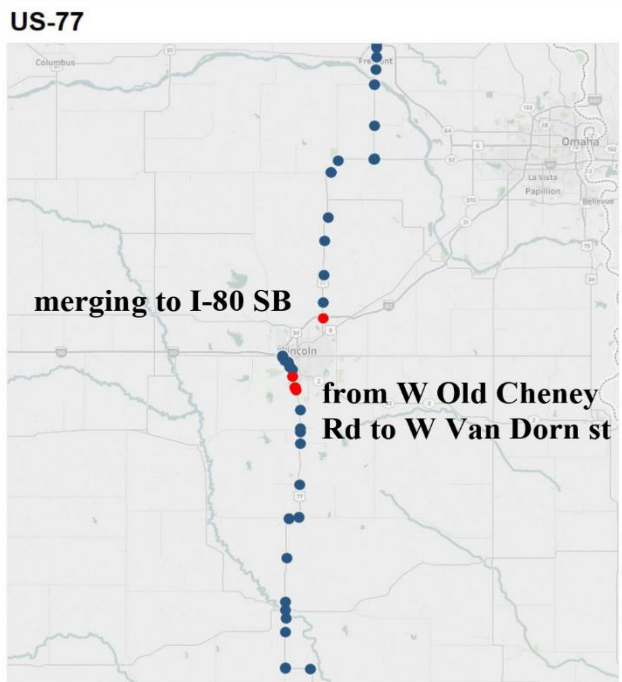
This means that people prefer to choose an alternative route for traveling to Omaha after the game, specifically I-80 EB. However, there is significant congestion on US 6 WB before the game. The red points in Figure 3.12(a) correspond to 72nd Street in Omaha, Superior Street–Cornhusker Highway, N 35th Street, and Sun Valley Boulevard in Lincoln.

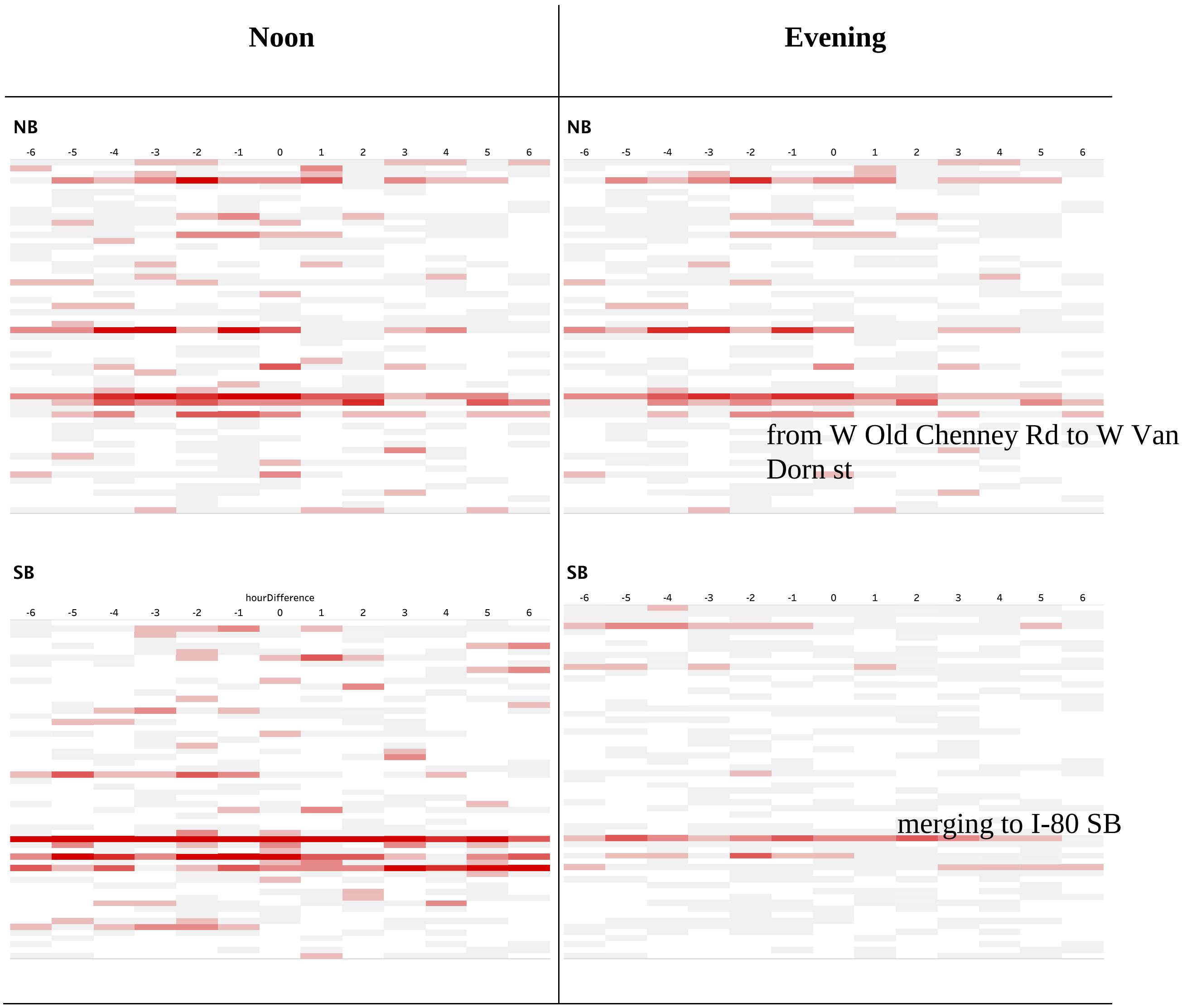
3.2.5 Route 5: US 77

US 77 in Nebraska runs south to north across the eastern portion of the state, emerging from Kansas in Gage County south of Wymore and ending in Dakota County north of South Sioux City before making a brief entrance into Iowa. Figure 3.13 shows the portion of US 77 analyzed in this study.

Figure 3.13. Route US 77, with blue points representing INRIX TMC segments

Figure 3.14(a) shows that there is no significant congestion far from Lincoln. All considerable traffic congestion is at the entry to Lincoln for both the north and south directions.


US-77
merging to I-80 SB
from W Old Cheney
Rd to W Van Dorn st
Omaha
Lincoln
a)



b)

Figure 3.14. (a) Hotspots indicated by red points and (b) heat maps for US 77 NB and SB for noon and evening games

Figure 3.14(b) also indicates significant congestion for the whole game day, especially noon games, on the SB entry into Lincoln (the merging point with I-80) and the NB entry into Lincoln (from W Old Cheney Road to W Van Dorn Street).

3.3 Conclusion

This chapter provided a detailed, visualized description of all five routes covered in this study and the most frequently congested locations on each route before and after games for noon and evening game days separately. Table 3.2 below summarizes all hot spots of noon and evening game days. In the next chapter, the reliability and accuracy of real-time INRIX data using different performance measures for the selected locations is discussed.

Table 3.2. Summary of all hot spots of noon and evening game days

Hot spots	Route	Location	Noon/Evening	
1	I-80	Exit 448	N	E
2		Exit 432	N	E
3		After 409	N	E
4		401-401B	N	E
5	NE-2	S 84th St	N	--
6		from S 33rd St to S 27th	N	--
7			--	E
8			N	E
9	NE-31	US6-NE31	--	E
10		S 216th St	N	E
11		Merging to I-80	N	E
12	US6	72nd St	N	E
13			--	E
14		Superior St - Cornhusker Hwy	N	E
15		N 35th St.	N	E
16		Sun Valley blvd	N	E
17	US77	Merging to I-80 SB	N	E
18		from W Old Cheney Rd to W Van Dorn St	N	E
19			N	E

4. Traffic Hotspot Analysis

4.1 Introduction

4.1.1 Incident Detection

Researchers and engineers have long been motivated to improve traffic safety and operations. Traffic congestion, especially that due to traffic accidents and special event traffic, is of great importance because of the delays and costs to the local community. Traffic delays can be attributed to certain events, including but not limited to traffic accidents and adverse weather conditions. These incidents may also have other effects, such as secondary collapse and delays in emergency medical services, which may result in additional costs. As a result, in the area of traffic management, the monitoring of the transportation network and the ability to detect and report abnormalities in real time is very important.

4.1.2 Data Stream and Pre-Processing

Most of the time in real-world scenarios, raw traffic data are incomplete, highly susceptible to noise, and inconsistent for many reasons, such as sensor failures, measurement technique errors, the large size of datasets, etc. Data pre-processing can be used to try to detect and correct corrupt and erroneous traffic data. However, the storage and analysis of massive amounts of INRIX data are impossible using traditional methods because they require the processing of more than 500 GB of data, which would be prohibitively time intensive on a traditional machine. For this study, a high performing cluster was used for data processing. The Hadoop Distributed File System (Apache Software Foundation 2018a) was used for storage of the data, and map-reduce was used for processing. Pig Latin (Apache Software Foundation 2018b) was used as the language to implement map-reduce algorithms.

4.2 Hotspot Detection

4.2.1 Introduction

Hotspot detection is utilized in many disciplines, such as in crime analysis for identifying where crimes occur with a certain frequency, in fire analysis for studying the phenomenon of forest fires, and in disease analysis for examining the localization and focus of diseases. Today, there is great interest in spatiotemporal data analysis because of the availability of huge amounts of data. Among the different analysis tasks that can be performed on spatiotemporal data, hotspot analysis is an important tool in security informatics and bio-surveillance. For instance, in applications for detecting crime hotspots, an outcome such as the identification of an increase in criminal activity at a specific shopping mall between the hours of 5:00 to 8:00 p.m. would be a spatiotemporal hotspot. Outcomes such as identifying crime hotspots at shopping mall or city centers would be strictly spatial hotspots, and identifying crime hotspots between 5:00 and 8:00 p.m. and between 9:00 and 11:00 a.m. are examples of temporal hotspots. The goal of hotspot analysis consists of detecting certain spatiotemporal regions among datasets. For example, in facial recognition the specific set of the largest eigenvectors can be used to approximate images of the human face. In structural engineering, both the eigenvalues and eigenvectors are used to assess the vibration of structures. In control engineering, the eigenvalues of a linear system are used to evaluate the stability and response of the system.

In the transportation field, a realistic scenario involving the application of the hotspot detection is in traffic incident detection. Suppose that there are several detectors across a city recording the speeds of vehicles passing the detectors, and consider the vehicles' speeds on normal days over multiple years to be the baseline information and the vehicles' speeds on game days over multiple years to be the case dataset. The goal in hotspot detection is to detect those spatiotemporal regions that contain unexpected lower speeds that lead to non-recurring congestion. For instance, an outcome that identifies certain activity on segments S1, S2, and S3 during the years Y1 to Y3 might be considered a spatiotemporal hotspot. The detection of such hotspots enables transportation agencies to better understand their target of interest and thereby provide essential interventions and preventive measures.

Hotspot analysis involves a spatiotemporal count matrix for the cases needed for the detection of those spatiotemporal regions (hotspots) that seem unexpected, given the baseline spatiotemporal matrix. Each cell in each matrix represents a count corresponding to a specific region and time. In particular, for traffic incident detection, each cell in the baseline matrix represents the speed corresponding to a segment at a specific time period on normal days. Each cell in the case matrix also represents the value of the reported speed on a specific segment within a given time period, but on game days. The goal of the analysis is to determine those subgroups of the spatiotemporal space whose reported cases are unexpected.

In this research, we aimed to develop a method that (1) does not require any input parameters and (2) weighs all the possible hotspots based on a standard metric like statistical significance (p-value). The alpha threshold should also be easy to estimate (usually alpha = 0.15). Addressing this problem, an Eigenspace-based algorithm called EigenSpot was recently proposed to detect space-time clusters with no restriction on the distribution and quality of the data or the shape of the cluster. However, this method can detect single hotspots only and is unable to detect multiple clusters. In traffic incident detection, when one cluster (incident) is detected, it is of interest to know whether there are additional clusters of high-risk regions present in the spatiotemporal space.

This research aimed to utilize an extension of the EigenSpot algorithm called Multi-EigenSpot to allow the detection of multiple clusters in a spatiotemporal space. The proposed algorithm uses the spatiotemporal matrix of expected congestion cases as the baseline information. Using the expected case matrix as the baseline information, we can replace the observed cases by the respective expected cases for the previously detected regions in the spatiotemporal space and re-run the algorithm to detect additional clusters, if they exist.

Eigenspace-based algorithms identify space-time incident hotspots by tracking changes in the space-time occurrence structure instead of conducting an exhaustive search over the space, as in traditional methods. Traditional methods are more useful for sensitive applications when the assumptions about the distribution of the data and the nature of the clusters are satisfied. However, for some nontraditional data sources, where these assumptions are not met, Eigenspace-based methods are an ideal solution for detecting potential clusters in a spatiotemporal space with no restrictions on the distribution and quality of the data or the shape of the cluster. Eigenspace-based methods detect clusters of homogeneous regions in terms of a congestion occurrence structure and do not restrict the regions in a cluster to being spatial neighbors. Since our proposed algorithm is based on the EigenSpot method, the following section briefly reviews the EigenSpot method before we present the proposed algorithm.

4.2.2 EigenSpot Algorithm

The inputs of the EigenSpot algorithm are two spatiotemporal m × n matrices, C, the game day's speeds, and, B, the baseline information (normal day's speeds), where m represents the number of segments and n represents the number of temporal instances. Each cell in each matrix represents a count (vehicle speed) corresponding to a specific segment and time. Given these matrices, the EigenSpot algorithm aims to identify a subgroup of regions in the spatiotemporal space where the reported cases (speeds) are unexpected with respect to the baseline information. Each matrix is decomposed using a one-rank singular value decomposition (SVD) to obtain the principal left and principal right singular vectors. The SVD is applied to each matrix, B and C, and the left and right singular vectors are calculated. The singular value decomposition of a spatiotemporal m × n matrix, M, is of the form $M = UDV^T$, where the columns of U are the left singular vectors corresponding to the spatial dimension and the columns of V are the right singular vectors corresponding to the temporal dimension. D is a diagonal matrix whose diagonal entries are the Eigenvalues of matrix M.

For instance, imagine that the baseline matrix (normal day) and case matrix (game day) are as follows:

$$B = \begin{bmatrix} 65 & 65 & 60 & 60 & 65 & \dots \\ 67 & 68 & 65 & 60 & 58 & \dots \\ 70 & 70 & 70 & 70 & 73 & \dots \end{bmatrix} \quad C = \begin{bmatrix} 65 & 35 & 30 & 39 & 49 & \dots \\ 27 & 30 & 40 & 65 & 69 & \dots \\ 55 & 55 & 50 & 40 & 35 & \dots \end{bmatrix}$$

As explained above, the SVD of B and C each equals UDV^T.

$B = U_B D_B V^T_B$

Where,

$$U_B = \begin{bmatrix} -0.5527 & 0.2357 & -0.7994 \\ -0.5581 & -0.8170 & 0.1450 \\ -0.6189 & 0.5263 & 0.5831 \end{bmatrix}$$

$$D = \begin{bmatrix} 255.0036 & 0 & 0 & 0 & 0 \\ 0 & 8.1791 & 0 & 0 & 0 \\ 0 & 0 & 3.5015 & 0 & 0 \end{bmatrix}$$

$$V_B \begin{bmatrix} -0.4574 & -0.3149 & -0.4075 & -0.1711 & -0.7045 \\ -0.4596 & -0.4148 & -0.3661 & 0.3156 & 0.6189 \\ -0.4422 & -0.2593 & 0.6512 & -0.5372 & 0.1569 \\ -0.4313 & 0.2402 & 0.4441 & 0.7029 & -0.2550 \\ -0.4450 & 0.7771 & -0.2806 & -0.2974 & 0.1761 \end{bmatrix}$$

Similarly,

$C = U_C D_C V^T_C$

Where,

$$U_C = \begin{bmatrix} -0.5574 & 0.3050 & -0.7722 \\ -0.5920 & -0.7981 & 0.1121 \\ -0.5821 & 0.5196 & 0.6254 \end{bmatrix}$$

$$D = \begin{bmatrix} 177.5876 & 0 & 0 & 0 & 0 \\ 0 & 42.6109 & 0 & 0 & 0 \\ 0 & 0 & 22.5603 & 0 & 0 \end{bmatrix}$$

$$V_C = \begin{bmatrix} -0.4743 & 0.6302 & -0.5658 & 0.2389 & 0.0252 \\ -0.3901 & 0.3593 & 0.4759 & -0.5490 & -0.4368 \\ -0.3914 & 0.0752 & 0.5581 & 0.2751 & 0.6737 \\ -0.4702 & -0.4506 & 0.0971 & 0.5403 & -0.5239 \\ -0.4985 & -0.5149 & -0.3639 & -0.5233 & 0.2830 \end{bmatrix}$$

The elements of the principal left singular vectors and right singular vectors are associated with the spatial dimension and temporal dimension, respectively. In the next step, the distances between the corresponding elements of the pair singular vectors are calculated. If the spatial singular vector for the normal day's (baseline) matrix is represented by ($sb_1, sb_2, . . ., sb_n$) and for the game day's (case) matrix by ($sc_1, sc_2, . . ., sc_n$), then the subtract vector is calculated as follows:

$ds = [ds_1 = sc_1 - sb_1 \; ds_2 = sc_2 - sb_2 \; . . . \; ds_n = sc_n - sb_n]$

Similarly, for the temporal dimension, the subtract vector is given by the following:

$dt = [dt_1 = tc_1 - tb_1 \; dt_2 = tc_2 - tb_2 \; . . . \; dt_m = tc_m - tb_m]$

A z-score control chart is applied to vectors ds and dt with a significance level of α to identify the out-of-control spatial and temporal components. The locations of hotspot regions in the spatiotemporal space are approximated by the joint combination of the out-of-control spatial and temporal components.

For instance, assume that sb = (0.25, 0.10, 0.75, 0.20) is the spatial singular vector of the baseline and sc = (0.30, 0.90, 0.80, 0.15) is the spatial singular vector of the cases. Each element in the spatial singular vector corresponds to a specific region. For instance, 0.30 and 0.25 in the first element corresponds to region 1. Similarly, the second, third, and the fourth elements correspond to regions 2, 3, and 4, respectively. The angle between the two singular vectors sb and sc is equal to 68° in this example. This angle does not tell us what elements of the singular vector have contributed to this difference. However, if, in the aforementioned example, we remove region 2 from the system, we have two vectors, sb = (0.25, 0.75, 0.20) and sc = (0.30, 0.80, 0.15), where the angle between them is equal to 0.09, which is almost equal to zero. Region 2 in this example is equivalent to the spatial component of the hotspot. In order to identify region 2 in this example, a z-score control chart is applied to the subtract vector ds = (0.25 - 0.30, 0.10 - 0.90, 0.75 - 0.80, 0.20 - 0.15) = (0.05, -0.80, -0.05, 0.05). Afterwards, we compute the standardized z-scores of the subtract vector, which in this case is zds = (0.4119, -1.4893, 0.4119, 0.6654). As shown, a z-score of -1.4893 is equivalent to the left-tailed p-value of 0.06. If we define α = 0.10, region 2 would be identified as a hotspot spatial component. This is because its p-value is lower than 0.10.

4.2.3 Multi-EigenSpot Algorithm

For the proposed algorithm, we consider the situation in which the vehicle speed data on game days and normal days are aggregated for different sub-regions over a time period range. In the proposed algorithm, the vehicle speeds on game days and normal days are arranged in the form of identical m × n spatiotemporal matrices, C and B, respectively, where m denotes the number of components in the spatial dimension (sub-regions) and n denotes the number of components in the temporal dimension (time points).

Given the spatiotemporal matrices, C (game day's speeds) and B (normal day's speeds), two spatiotemporal matrices, E (expected speeds) and R (relative risks), are calculated. For the expected traffic congestion, if no cluster exists in the spatiotemporal space, we use the formula proposed in Kulldorff et al. (2005), which assumes the reported cases to be distributed over the spatiotemporal space proportional to the respective normal day's speeds. The risk measure, RR, is also calculated as the proportion of the C to the E. The SVD is applied to each matrix, C and E, and the left and right singular vectors are calculated. The SVD of a spatiotemporal m × n matrix, M, is of the form $M = UDV^T$, where the columns of U are the left singular vectors corresponding to the spatial dimension and the columns of V are the right singular vectors corresponding to the temporal dimension. D is a diagonal matrix whose diagonal entries are the Eigenvalues of matrix M. For the purposes of comparison, only the singular vectors corresponding to the largest Eigenvalue were considered because these principal vectors explain or extract the largest part of the inertia of the data table.

If we assume that C and E are identical, their principal left and right singular vectors are identical as well, i.e., the corresponding elements in the pair singular vectors stay at a zero distance. If some change occurs in C, this change can be detected from the changes in the principal singular vector's elements. In such cases, some distances between the corresponding elements of the pair singular vectors become abnormal for the components corresponding to the affected areas in both the spatial and temporal dimensions.

Our approach uses the z-control chart for monitoring the distances between the corresponding elements of the pair singular vectors. The corresponding elements of the pair left singular vectors showing abnormal differences are associated with the spatial components of a cluster, and the corresponding elements of the pair right singular vectors are associated with the temporal components. If abnormal components are found in both dimensions, matrix C is upgraded by replacing the elements (game day's lower speeds) corresponding to the out-of-control spatial and temporal components by the respective expected cases. In addition, matrix R is upgraded by replacing the elements corresponding to the out-of-control components by their average values to further visualize these elements on the heat map with the same color as a hotspot. The process is repeated with the upgraded matrix C and the original matrix of the expected cases, E. The matrices C and R are upgraded iteratively until no out-of-control component is found in either spatial or temporal dimension. Since the upgraded elements in matrix R are used to approximate the hotspots, the elements in the upgraded matrix R, other than the average values, are replaced by 1 to differentiate the upgraded elements from the non-upgraded ones. The resulting matrix R is then visualized on the heat map to show the different average relative risks using different colors. In cases where no space-time cluster exists, the resulting heat map has all elements equal to 1, indicated by a dark blue color only. Colors on the heat map other than dark blue approximate different space-time clusters. The sub-regions in a cluster are homogeneous with respect to the space-time occurrence structure and are represented by the same color on the heat map.

The Multi-EigenSpot algorithm requires three types of tools: (1) an SVD for finding the singular vectors of a non-square matrix, (2) a statistical process control tool for monitoring distances between the corresponding elements of the pair singular vectors and (3) a visualization tool (heat map) for visualizing the detected clusters. The detailed stepwise process and the ways these tools are deployed in the algorithm are given below.

- Step 1: Calculate the spatiotemporal matrices of the expected vehicle speeds and relative risks.

$$E_{ij} = \frac{C_{.j}}{B_{.j}} \times B_{ij}$$

$$E = [E_{11} \cdots E_{1n} \vdots \ddots \vdots E_{m1} \cdots E_{mn}]$$

Where,

$C_{.j}$ is the j_{th} column-average of matrix C
$B_{.j}$ is the j_{th} column-average of matrix B
B_{ij} is the speed in the i_{th} sub-region over the j_{th} time-point

$$R_{ij} = \frac{C_{ij}}{E_{ij}}$$
$$R = [R_{11} \cdots R_{1n} \vdots \ddots \vdots R_{m1} \cdots R_{mn}]$$

Having matrix R, we are able to visualize hotspot clusters on the heatmap.

- Step 2: Calculate the SVDs of matrices C and E.
- Step 3: Calculate the subtract vectors.
- Step 4: Identify abnormally high distances in the corresponding elements of the pair singular vectors.
- Step 5: Upgrade matrices C and R.
- Step 6: Find additional abnormal components in the spatial and temporal dimensions. Repeat Steps 2 through 5 until no abnormal component is found in each dimension.
- Step 7: Replace the elements in the last upgraded matrix R, corresponding to the components (spatial/temporal) that were not found to be abnormal, with 1.
- Step 8: Visualize the resulting matrix R on a heat map on which the average RR values are represented by different colors.

Figure 4.1 is an example output of the proposed method.

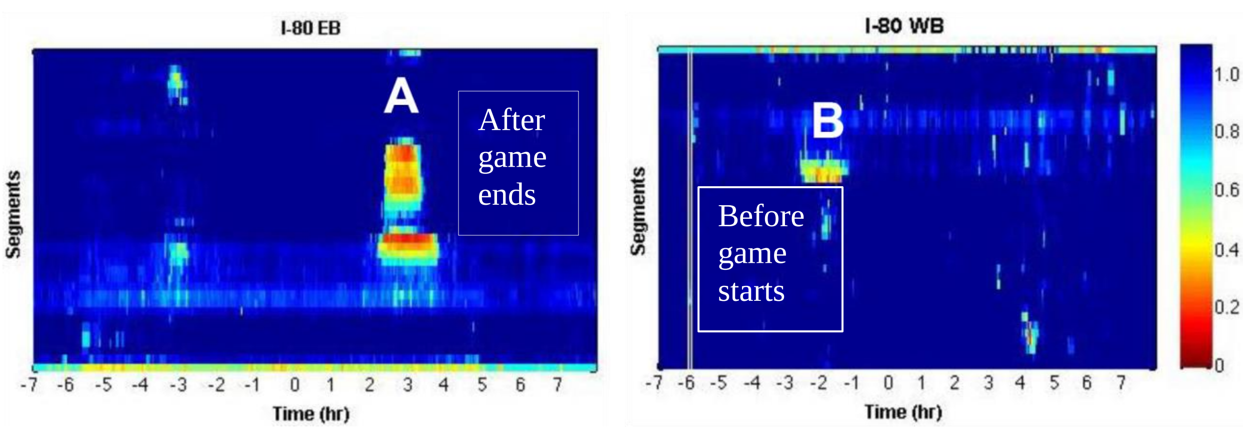


Figure 4.1. Sample result of the proposed algorithm showing a spatiotemporal matrix for I-80

The colored regions on the heat map (points A and B) corresponding to different average RR values (less than 1) show multiple space-time hotspots. Point A indicates a traffic hotspot cluster after a game ends on route I-80 EB. Point B shows congestion before the start time of the game on I-80 WB. If no cluster exists, then all the data values on the heat map are equal to 1, indicated dark blue color.

After analyzing normal days against game days, it was observed that congestion mostly occurs a few hours before the start time and immediately after the end of each game. It is important to compare a normal day against a typical congested day (i.e., a congested non-game day) and assess congestion trends and the differences between the traffic patterns observing on game-day versus typical congested day. In order to do that, two speed contour maps are plotted in Figure 4.2 that show speeds over each minute of a normal day and those on a typical congested day.

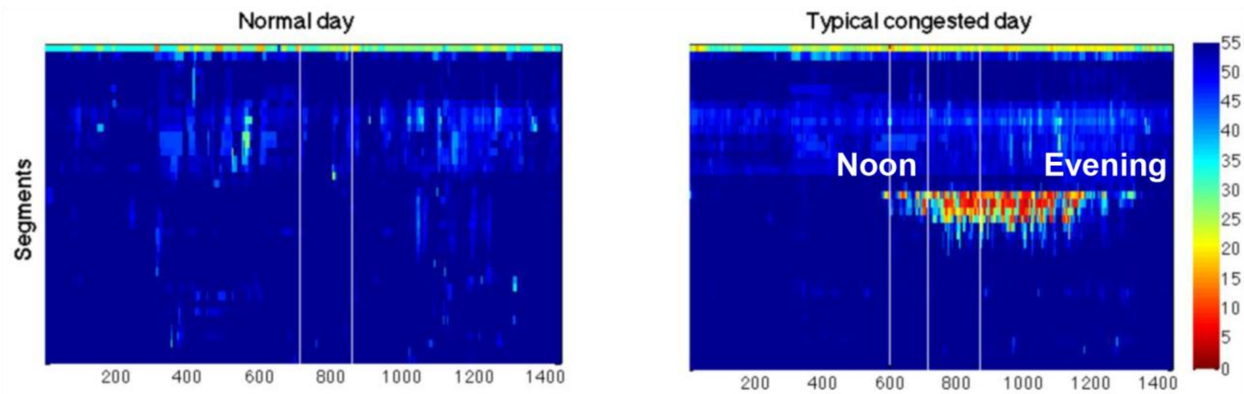


Figure 4.2. Sample results showing speed contour maps for a normal day and a typical congested day

As can be seen, the normal day's heat map (left) is almost completely blue, indicating speeds above 45 mph, which means no congestion occurred. The heat map showing a typical congested day (right) shows speeds less than 45 mph for a few hours starting at noon and lasting until the evening. That congestion could be recurring and/or non-recurring.

According to a sample analysis of a normal day against a typical congested day, there is no similarity between the congestion that occurs on game days and the congestion that occurs on typical congested days. On game days, congestion occurs in a specific period of time before or after the start time of the game and in specific zones, while on a typical congested day traffic congestion can occur anywhere on a route and at any hour of the day.

4.3 Hotspot Parameters

After detecting hotspots on game days, it is crucial to identify the factors affecting the sizes of the hotspots, their locations, and other possible parameters. The start time of the game and the Cornhuskers' opponent on a given day are two important factors affecting the number of people coming to Lincoln, Nebraska, on game days. The start time of the game can be classified as noon or evening. The noon category contains games starting at 11:00 a.m. or 2:30 p.m. Similarly, the evening category contains all games starting at 6:30 or 7:00 p.m. Moreover, the Cornhuskers' opponent on a given day can significantly influence the importance of the game and therefore the size of the crowd that the game draws. For example, the Cornhuskers' toughest opponents in

2018, i.e., the opponents drawing the largest crowds, were (1) Ohio State, (2) Wisconsin, (3) Northwestern, (4) Michigan State, and (5) Iowa. Therefore, it is important to assess the impacts of these two factors (start time and opponent) on the sizes of the hotspot. Hotspot size can be defined as (1) the number of congested lanes, (2) the number of congested segments, and (3) congestion duration. The number of congested lanes itself is divided to three categories: one lane, two lanes, and three lanes.

4.3.1 Start Time of the Game

Using the proposed method for hotspot detection, it is possible to determine the number of consecutive segments in each hotspot. Having the segments' lengths and the number of consecutive segments for each hotspot, it is possible to approximate the length of each congestion region (hotspot). Figures 4.3 and 4.4 illustrate how the length and duration of congestion vary between noon and evening games with respect to the time the congestion occurred, whether before the start time of the game (negative values) or after that (positive values).

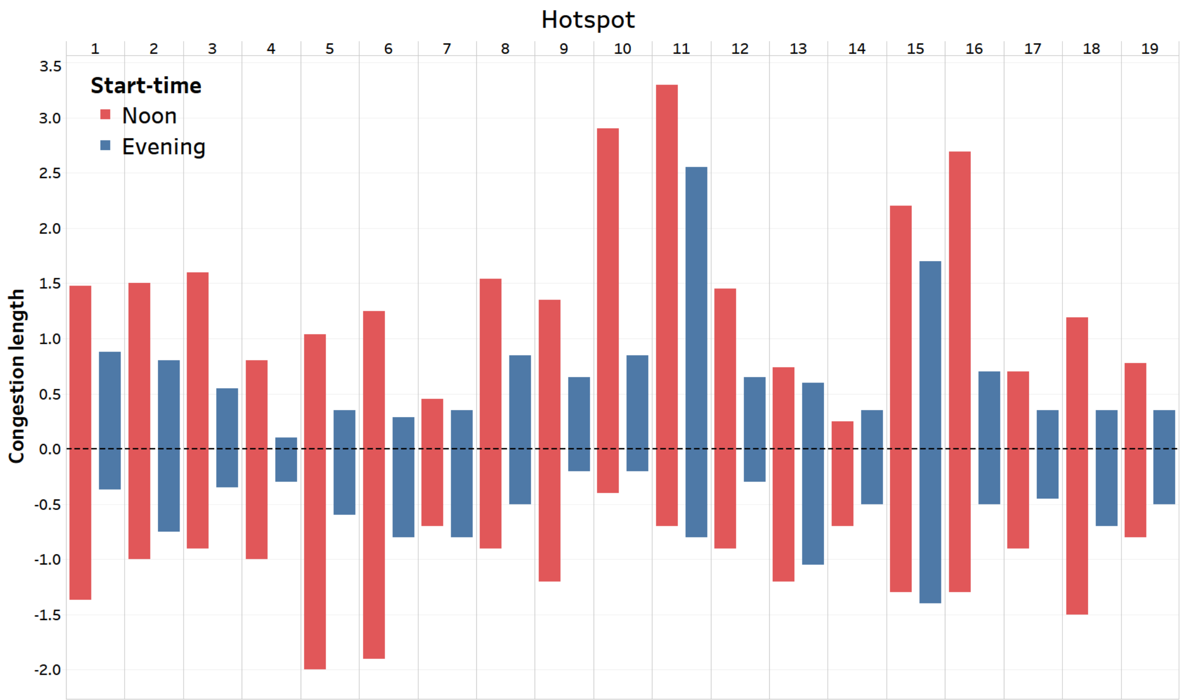


Figure 4.3. Impact of the start time of the game on congestion (hotspot) length

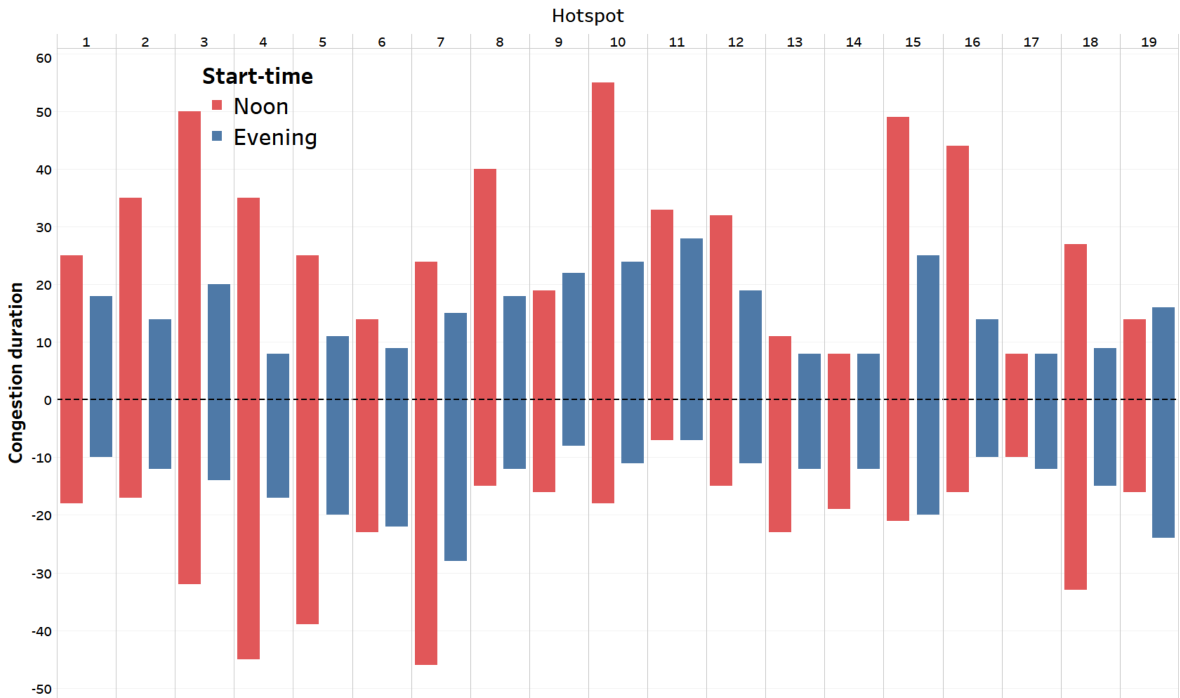


Figure 4.4. Impact of the start time of the game on congestion (hotspot) duration

The vertical axis in each figure is divided into positive and negative values. Positive represents congestion that occurs after the end of the game, while negative represents congestion that occurs before the start of the game. Among the 19 hotspots (horizontal axis) identified in the case study by the proposed hotspot detection algorithm, it is clear that for most of the hotspots, the congestion length when the start time of the game is noon (red bars) is higher than that of evening games (blue bars), no matter whether the congestion occurred before (negative y-values) or after the start of the game (positive y-values).

The Multi-EigenSpot algorithm is also capable of identifying the duration of each hotspot cluster. As shown in Figure 4.1 as an example, the time (congestion duration) and number of segments (congestion length) of each traffic hotspot cluster are easily visible. Thus, we estimated congestion duration using the proposed method. Figure 4.4 shows the average congestion duration of 19 hotspots during noon and evening football games with respect to the time the congestion occurred, either before the start time of the game (negative values) or after that (positive values). As can be seen in the figure, all games that started in the evening (blue bars) had a lower average congestion duration than noon games (red bars), no matter whether the congestion occurred before or after the start of the game.

4.3.2 Opponent

The Nebraska Cornhuskers' opponent on a given day play a significant role in the importance of the game and therefore the size of the crowd that the game draws. Over the last five years, the Cornhuskers' toughest opponents, i.e., the teams drawing the largest crowds, were (1) Ohio

State, (2) Wisconsin, (3) Northwestern, (4) Iowa, (5) Michigan State, and (6) Purdue. It is therefore important to evaluate the effect of the Cornhuskers' opponent on game day traffic congestion.

Using the proposed method for hotspot detection, we are able to determine the length and duration of traffic hotspots. The goal of this section is to determine the impact of the Cornhuskers' opponents on the length and duration of traffic congestion. For instance, is there a longer duration or length of congestion in the Lincoln area when the game is between the Cornhuskers and Ohio State compared to when the game is between the Cornhuskers and Rutgers?

In Figures 4.5 and 4.6, the horizontal axis shows the top five opponents of the Nebraska Cornhuskers over the past five years.

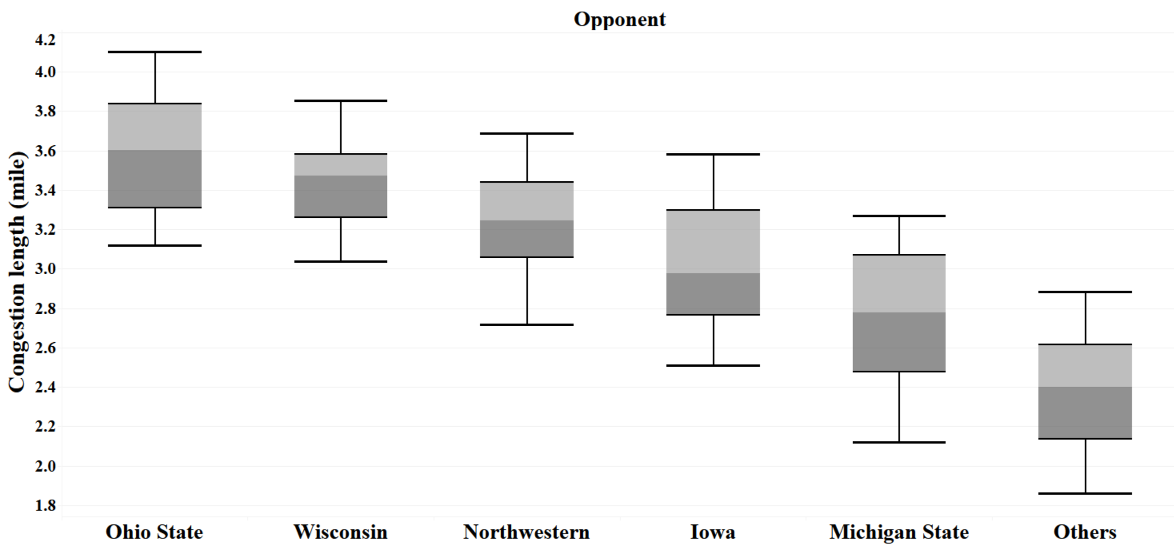


Figure 4.5. Impact of Cornhuskers' opponents on the congestion length

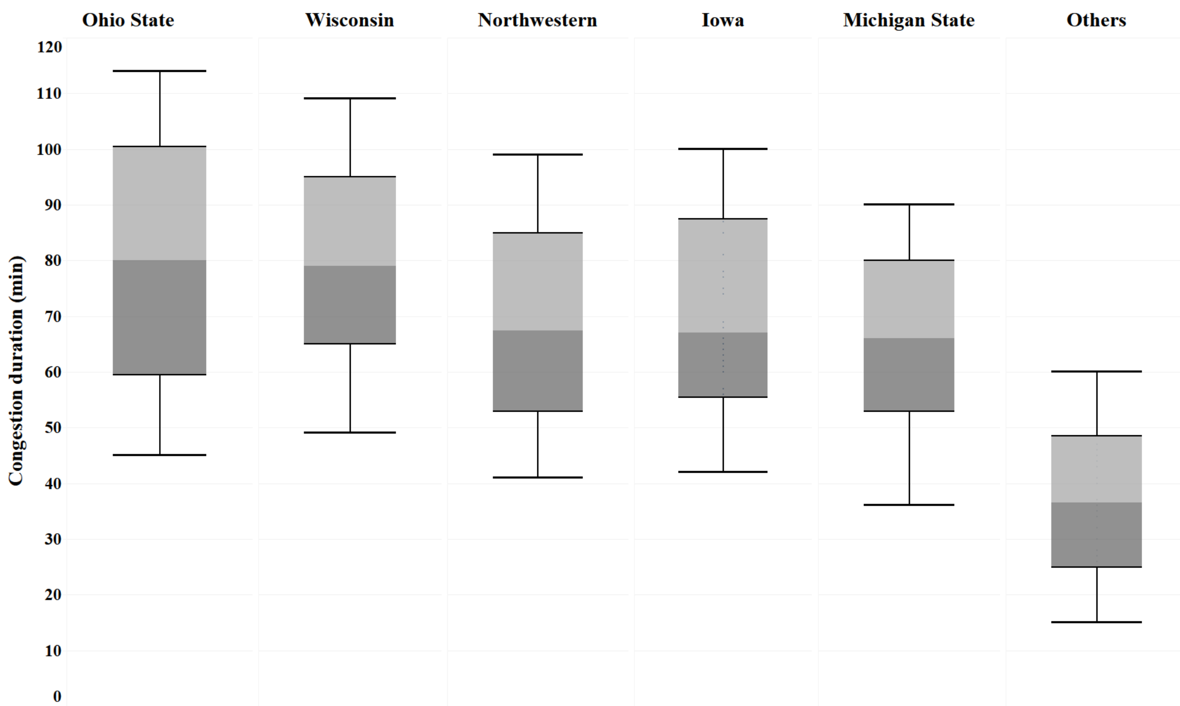


Figure 4.6. Impact of Cornhuskers' opponents on the congestion duration

In the figures, the "Others" category includes all other teams, which usually have little chance of winning against the Cornhuskers. As can be seen in the boxplots, there is a decreasing trend from Ohio State (the toughest opponent) to Others (the weaker teams), implying that the length and duration of traffic hotspots are influenced by the Cornhuskers' opponents.

According to Figure 4.5, the median length of congestion that occurred in the Lincoln area when the opponent was among the top five is more than three miles, while it decreases to around two miles when Cornhuskers faced weaker teams. Similarly, based on Figure 4.6, the median duration of congestion that occurred on game days against stronger teams is around 80 minutes while it drops to 40 minutes on games days against weaker teams.

Based on the exploratory analysis described above, the traffic hotspot size is influenced by the start time of the game and the toughness of the Cornhuskers' opponent. In the next step, this research aims to predict traffic congestion based on the available variables and identify hotspot clusters for the year 2018 based on the predicted dataset. Given the start time of the game (noon or evening), the toughness of the opponent, and the specific congested segments for each route, it is possible to forecast speeds on game days for the following year (2018) using Dynamic Bayesian Networks and identify hotspot clusters based on the predicted dataset. The data from 2018 are utilized as a validation dataset.

4.4 Dynamic Bayesian Networks

Pearl introduced Bayesian networks as probabilistic graphic models that explain the dependence and independence of random variables on conditions. These dependencies are represented by a directed acyclic graph and measured by a joint probability distribution that breaks down into a product of local conditional distributions:

$$p(X1, \ldots, Xn) = \prod_{i=1}^{n} p(Xi|Pa(Xi))$$

Where Pa(Xi) is the set of parents of Xi. Bayesian networks' flexibility allows different sources of information to be combined. For example, you can use your own knowledge to set a part of the model and the other part can be learned automatically from data (Faour et al. 2006). Additionally, inferences (the forecasting process) can be made using the information propagation mechanism, even in the case of incomplete data. This feature is especially useful in real-time applications where it can be harmful to implement a further imputation process.

Dynamic Bayesian Networks are extended to models evolving over time (Dean and Kanazawa 1989). Each node X_i^((t)) represents the instantiation of the variable X_i^ at time slice t. The parents of X_i^((t)) can belong to t, t-1, …, t-r, where r is the order of the Dynamic Bayesian Network (Ghahramani 2006).

Due to the limited number of available observations, 10-fold cross-validation is evaluated in the forecasting performance (Kohavi 1995). This method involves dividing the dataset X randomly into 10 subsets X1, ..., X10 of (approximately) the same size. The model is trained on X\Xk and tested on X for each K ∈ {1, ...,10}. The final performance is estimated by an average of 10 measurements of accuracy. The weighted average absolute percentage error (WMAPE) is adopted in this report:

$$WMAPE(x, \hat{x}) = \frac{\sum_{t=1}^{N} \left| x^{(t)} - \hat{x}^{(t)} \right|}{\sum_{t=1}^{N} x^{(t)}}$$

Where $\hat{x}$ is the estimate of x and N is the number of comments in the dataset. The WMAPE is easy to interpret, like the mean absolute percentage error (MAPE). However, it favors models that predict high values effectively.

4.4.1 Learning with Incomplete Data

In this study of several routes in Nebraska over five years, the missing data are too dispersed to delete list-wise. Sun et al. (2006) proposed that the parents of the Dynamic Bayesian Network can be replaced with variables whose values are missing. Unfortunately, this method is hardly applicable to this study because it means that parents are complete, which in many situations does not necessarily apply.

The expectation-maximization (EM) algorithm, proposed by Dempster et al. (1977), is a method for iteratively estimating the maximum likelihood of parameters when values are missing (or hidden) in the training dataset. This method involves two steps at each iteration, starting from an initial parameter estimation. It completes the dataset of observed data and current estimates of the expectation (E) parameters. This completed dataset is used in the M step to update parameters by maximizing the probability of logging. As Dempster et al. (1977) shows, the log likelihood increases at each iteration until the maximum local convergence is achieved.

The next time slice forecasting process can be considered a problem of inference in the Dynamic Bayesian Network. A comprehensive review of inference methods can be found in Murphy's thesis (2002) or in Koller and Friedman's recent book (2009) on Dynamic Bayesian Networks. The approximate inference methods normally take less time than the exact methods. For the present study, they seem to be a better choice to ensure real-time forecasts given the complexity of our model. The bootstrap filter (Gordon et al. 1993), also known as the fittest survival, is a stochastic simulation algorithm that can be inferred in real-time. It generates weighted sample sequences by sampling unobserved values. These sequences are time collected-multiplied in proportion to their weight, which reflects their probability of time.

4.4.2 Experimental Method

The data for this study were collected during 41 game days and 41 normal days over five years from 2013 to 2017. As explained earlier, the start time of the game and the toughness of the Cornhuskers' opponent are two significant factors affecting the hotspot clusters detected by the Multi-EigenSpot algorithm. The start time of the game can be classified as noon or evening. The noon category contains games that start at 11:00 a.m. or 2:30 p.m. Similarly, the evening category contains all games starting at 6:30 or 7:00 p.m. Moreover, the Cornhuskers' opponent on a given day can significantly influence the importance of the game and therefore the size of the crowd that the game draws. For example, the Cornhuskers' toughest opponents in 2018, i.e., the opponents drawing the largest crowds, were (1) Ohio State, (2) Wisconsin, (3) Northwestern, (4) Michigan State, and (5) Iowa. The variable describing the toughness of the Cornhuskers' opponent can have one of two values: tough opponent or normal opponent. The prediction algorithm is applied to each route separately. To predict speeds and thereby identify hotspots, which are the locations always experiencing congestion on game days, the start time of the game and the opponent's toughness are two discrete variables. In the model's structure, time windows of 15 minutes are considered. In other words, each frame of the model contains 15 minutes of traffic speed as a vector.

In this study, the Dynamic Bayesian Networks approach was found to perform well on each route. The corresponding WMAPE for each route is provided in the Table 4.2. The average WMAPE for all routes is 13.8 %.

Table 4.1. Average forecasting errors (WMAPE in %)

Route	WMAPE
I-80	12.2
NE 31	10.6
US 6	11.4
US 77	18.1
NE 2	16.5

Note that the effect of toughness of the opponent in the model does not change significantly when there are five classes of toughness rather than two. The accuracy of the model is well illustrated by Figure 4.7, which shows the actual and predicted values on I-80 as an example.

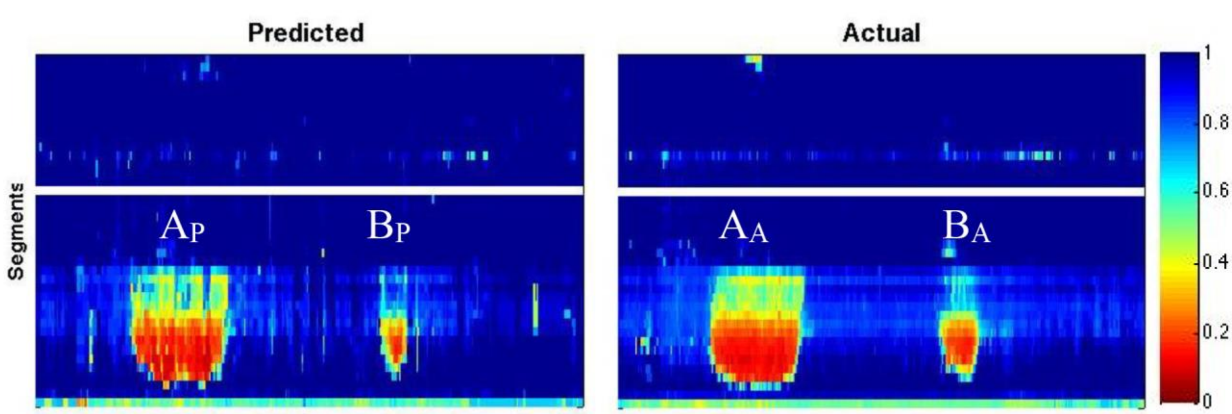


Figure 4.7. Predicted and actual hotspot clusters showing traffic congestion on game days on I-80 in 2018

Heat maps of other routes are provided in Figure A.1 in Appendix A. Game days in 2018 were utilized as the validation dataset. After forecasting speeds using DBNs, we utilized the Multi-EigenSpot algorithm again to find the hotspot clusters, both for the predicted scenario and the actual scenario. As can be seen in Figure 4.7, the predicted and actual values are nearly same, indicating the high accuracy of the proposed prediction method. Points A_P (predicted) and A_A (actual) correspond to each other, as is the case for points B_P (predicted) and B_A (actual).

4.5 Conclusion

In recent years, traffic congestion has become a significant issue in urban areas. People in the United States travel an extra one billion hours and consume an extra one billion gallons of fuel due to traffic congestion every year. Therefore, monitoring the performance of the transportation system plays an important role in any transportation operation and planning strategy. Non-recurring congestion includes congestion caused by accidents, road work, special events, or adverse weather conditions. Non-periodic events with a high expected attendance, including

PSEs such as concerts, football games, etc., play an important role in delays in everyday transportation.

Memorial Stadium in Lincoln, Nebraska, is the home of the Nebraska Cornhuskers football team. With an extended capacity of more than 85,000 people, the stadium is commonly referred as the “third-largest city in Nebraska” on game days. Game days, therefore, typically affect travel patterns in Lincoln and its neighboring regions.

This chapter evaluated the relationship between professional sports events and traffic congestion using INRIX data for the past five years in Nebraska. This study demonstrates a systematic way to assess travel patterns and traffic hotspot clusters on football game days compared to normal days.

Five major routes in Nebraska were selected for this study, and the analysis utilized historical and real-time traffic data, including speeds, travel times, and location information, collected through the INRIX traffic message channel (TMC) monitoring platform. The INRIX dataset is currently regarded as the largest crowd-sourced traffic dataset.

For the detection of hotspots, the Multi-EigenSpot algorithm, which is an extension of the EigenSpot algorithm, was utilized. The spatiotemporal analysis of real-world congestion case data demonstrated that the proposed method addresses the two main limitations of the existing EigenSpot algorithm: multiple cluster detection and visualization.

Ultimately, the DBN approach was proposed to forecast traffic congestion (hotspots) on game days. This approach is designed to provide predictions in real-time even when incomplete data are present. In the presence of incomplete data, the structural EM algorithm is used both to reduce the structure’s dimensions and find the parameters’ maximum probability estimates. The bootstrap filter is then used to make predictions. The experiment was carried out on all game days and corresponding normal days from 2013 to 2017. Data from 2018 was used for validation.

5. Summary and Conclusions

In recent years, traffic congestion has become a significant issue in urban areas. People in the United States travel an extra one billion hours and consume an extra one billion gallons of fuel due to traffic congestion every year. Therefore, monitoring the performance of the transportation system plays an important role in any transportation operation and planning strategy.

Traffic monitoring using wide-area probe-sourced data is increasingly becoming a viable means of comprehensive traffic monitoring without the need for a large investment in the deployment of physical assets in rights-of-way and the associated cost and maintenance burdens. Congestion that is caused by accidents, road work, special events, or adverse weather is called non-recurring congestion. Non-periodic events with an expected large attendance, including PSEs such as concerts, football games, etc., play a major role in delays in everyday transportation.

Memorial Stadium in Lincoln, Nebraska, is the home of the Nebraska Cornhuskers football team. With an extended capacity of more than 85,000 people, the stadium is commonly referred as the "third-largest city in Nebraska" on game days. Game days, therefore, typically affect travel patterns in Lincoln and its neighboring regions.

The study described in this report evaluated the relationship between professional sports events and traffic congestion using INRIX data over the past five years in Nebraska. The study demonstrates a systematic way to assess travel patterns and traffic hotspot clusters on football game days compared to normal days.

Five major routes in Nebraska were selected for this study, and the analysis utilized historical and real-time traffic data, including speeds, travel times, and location information, collected through the INRIX traffic message channel (TMC) monitoring platform. The INRIX dataset is currently regarded as the largest crowd-sourced traffic dataset. A comprehensive exploratory analysis of performance monitoring on game days against normal days for the five selected routes in Nebraska was also performed.

The following observations were drawn from this study:

- There are several points on I-80 EB showing congestion during game days (from exit 353 to exit 369).
- When the start time of the game is 11:00 a.m. or 2:30 p.m., there is congestion on I-80 WB from Omaha to Lincoln. However, when the start time of the game is 6:30 or 7:00 pm, there is almost no congestion on I-80 WB from Omaha to Lincoln.
- There is considerable congestion on four segments on NE 2 WB, while there is no considerable congestion on NE 2 EB. This means that people prefer to choose an alternative route to travel east (for example, to Iowa) after the game. However, there is significant congestion on NE 2 WB before the game, which means that people from Iowa or regions around Nebraska's eastern border prefer to use this route to come to Lincoln.
- There is considerable congestion on three segments on NE 31 SB. However, there is no congestion on NE 31 NB.

- There is congestion on eight segments on US 6 WB, while there is almost no significant congestion on US 6 EB. This means that people prefer to choose an alternative route to travel to Omaha after the game, namely I-80 EB. However, there is significant congestion on US 6 WB before the game.
- On US 77, there is no significant congestion far from Lincoln. All considerable traffic congestion can be seen only at the entry to Lincoln for both the north and south directions.

In the transportation field, a realistic scenario involving the application of hotspot detection is in traffic incident detection. A novel method for hotspot detection was proposed in this report. The proposed algorithm uses the spatiotemporal matrix of expected congestion cases as the baseline information. Using the expected case matrix as the baseline information, the observed cases can be replaced by the respective expected cases for the previously detected regions in the spatiotemporal space, and then the algorithm can be re-run to detect additional clusters, if they exist.

After detecting hotspots, it is crucial to identify the factors affecting the sizes of the hotspots, their locations, and other possible parameters. The start time of the game and the Cornhuskers' opponent for a given game are two important factors affecting the number of people coming to Lincoln, Nebraska, on game days. The start time of the game can be classified as either noon or evening. The opponent of the Nebraska Cornhuskers also plays a significant role in the importance of a given game and therefore the size of the crowd that the game draws. Over the last five years, the Cornhuskers' toughest opponents, i.e., the opponents drawing the largest crowds, were (1) Ohio State, (2) Wisconsin, (3) Northwestern, (4) Michigan State, (5) Iowa, and (6) Purdue. Hotspot size can be defined as (1) the number of congested lanes, (2) the number of congested segments, and (3) congestion duration.

The DBN approach was proposed to forecast traffic congestion (hotspots) on game days. This approach is designed to provide predictions in real-time even when incomplete data are present. In the presence of incomplete data, the structural EM algorithm is used both to reduce the structure's dimensions and find the parameters' maximum probability estimates. The bootstrap filter is then used to make predictions. The experiment was carried out on all game days and corresponding normal days from 2013 to 2017. Data from 2018 was used for validation.

Table 4.3 summarizes the important findings of this report.

Table 5.1 Summary of important findings from this study

Finding	Comments
Hotspot detection	● The Multi-EigenSpot algorithm was used to detect traffic hotspot clusters.
Impact of the start time of the game	● The start time of the game was classified as either noon or evening. ● Among the 19 hotspots identified in the case study by the proposed hotspot detection algorithm, it is clear that for most of the hotspots, the congestion length when the start time of the game is noon is higher than that of evening games, no matter whether the congestion occurred before or after the start of the game. ● All games that started in the evening had a lower average congestion duration than noon games, no matter whether the congestion occurred before or after the start of the game. ● When the start time of game is noon (11:00 a.m. or 2:30 p.m.), there is congestion on I-80 WB from Omaha to Lincoln.
Impact of opponent	● The Cornhuskers' opponent on a given game day affects the length and duration of traffic congestion. ● A decreasing trend in the length and duration of traffic congestion was observed from Ohio State (the toughest opponent) to the Others category (the weaker teams), implying that the duration and length of traffic hotspots are influenced by the Cornhuskers' opponents.
Hotspot prediction	● The DBN approach was proposed to forecast traffic congestion (hotspots) on game days. ● The average WMAPE for all routes was 13.8%. ● The DBN approach performed well on each route.

Finally, it should be mentioned that many tests, analyses, and experiments were not conducted in this study due to a lack of sufficient data. Because this study focused mainly on freeways, future work should be focused on a deeper analysis of arterials and urban areas. This would be possible through the deployment of additional sensor infrastructure on both freeways and arterials.

References

Amini, S., E. Papapanagiotou, and F. Busch. 2016. Traffic Management for Major Events. In *Digital Mobility Platforms and Ecosystems: State of the Art Report*. Eds. A. Faber, F. Matthes, and F Michel. Technical University of Munich, Munich, Germany. 187–197. https://doi.org/10.14459/2016md1324021.

Apache Software Foundation. 2018a. Apache Hadoop. http://hadoop.apache.org/. Last accessed May 2019.

Apache Software Foundation. 2018b. Apache Pig. https://pig.apache.org/. Last accessed May 2019.

Chase, J. and M. Healey. 1995. The Spatial Externality Effects of Football Matches and Rock Concerts: The Case of Portman Road Stadium, Ipswich, Suffolk. *Applied Geography* 15 (1): 18–34. https://doi.org/https://doi.org/10.1016/0143-6228(95)91060-B.

Chrobok, R., O. Kaumann, and J. Wahle. 2004. Different Methods of Traffic Forecast Based on Real Data. *European Journal of Operational Research* 155 (3): 558–68. https://doi.org/doi.org/10.1016/j.ejor.2003.08.005.

Chu, K.C. and K. Saitou. 2013. Optimization of Probe Vehicle Deployment for Traffic Status Estimation. IEEE International Conference on Automation Science and Engineering (CASE), August 17–20, Madison, WI. https://ieeexplore.ieee.org/document/6654046.

Clark, S. 2003. Traffic Prediction Using Multivariate Nonparametric Regression. *Journal of Transportation Engineering* 129 (2): 161–68. https://doi.org/10.1061/(ASCE)0733-947X(2003)129:2(161).

Coates, D. and B.R. Humphreys. 2003. *Professional Sports Facilities, Franchises and Urban Economic Development*. UMBC Economics Department Working Papers 03-103. University of Maryland, Baltimore County, Baltimore, MD. https://ideas.repec.org/p/umb/econwp/03103.html.

Coifman, B. 2002. Estimating Travel Times and Vehicle Trajectories on Freeways Using Dual Loop Detectors. *Transportation Research Part A: Policy and Practice* 36 (4): 351–64. https://doi.org/10.1016/S0965-8564(01)00007-6.

Cookson, G., and B. Pishue. 2017. *INRIX Global Traffic Scorecard*. INRIX, Kirkland, WA.

Dean, T. and K. Kanazawa. 1989. A Model for Reasoning about Persistence and Causation. *Computational Intelligence* 5 (2): 142–50. https://doi.org/10.1111/j.1467-8640.1989.tb00324.x.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal Of The Royal Statistical Society, Series B* 39 (1): 1–38.

Duranton, G. and M.A. Turner. 2011. The Fundamental Law of Road Congestion: Evidence from US Cities. *American Economic Review* 101 (6): 2616–52. https://doi.org/10.1257/aer.101.6.2616.

Faour, A., P. Leray, and B. Eter. 2006. A SOM and Bayesian Network Architecture for Alert Filtering in Network Intrusion Detection Systems. In *Proceedings of the 2nd International Conference on Information & Communication Technologies: From Theory to Applications, 24–28 April 2006, Damascus, Syria, Volume 2*. IEEE, Piscataway, NJ. 3175–3180. https://doi.org/10.1109/ICTTA.2006.1684924.

FDOT. 2012. *Technical Memorandum, Probe Data Analysis: Evaluation of NAVTEQ, TrafficCast, and INRIX Travel Time System Data in the Tallahassee Region*. Florida Department of Transportation, Tallahassee, FL. https://fdotwww.blob.core.windows.net/sitefinity/docs/default-source/content/traffic/its/projects_deploy/2012-03-26_probe_data_analysis_v2-0.pdf?sfvrsn=ed757b5e_0.

Feng, W., A. Bigazzi, S. Kothuri, and R. Bertini. 2010. Freeway Sensor Spacing and Probe Vehicle Penetration: Impacts on Travel Time Prediction and Estimation Accuracy. *Transportation Research Record* 2178: 67–78. https://doi.org/10.3141/2178-08.

Mudge, R., H. Mahmassani, R. Haas, A. Talebpour, and L. Carroll. 2013. *Work Zone Performance Measurement Using Probe Data*. FHWA-HOP-13-043. Federal Highway Administration, Washington, DC. https://ops.fhwa.dot.gov/wz/resources/publications/fhwahop13043/fhwahop13043.pdf.

FHWA. 2017. *2016 Urban Congestion Trends*. FHWA-HOP-17-010. Federal Highway Administration, Washington, DC. https://ops.fhwa.dot.gov/publications/fhwahop17010/fhwahop17010.pdf.

Ghahramani, Z. 2006. Learning Dynamic Bayesian Networks. In *Adaptive Processing of Sequences and Data Structures*. Eds. C.L. Giles and M. Gori. Springer, Berlin, Heidelberg. 168–197. https://doi.org/10.1007/bfb0053999.

Gordon, N.J., D.J. Salmond, and A.F.M. Smith. 1993. Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. *IEE Proceedings F - Radar and Signal Processing* 140 (2): 107. https://doi.org/10.1049/ip-f-2.1993.0015.

Haghani, A., X. Zhang, and M. Hamedi. 2015. *Validation and Augmentation of Inrix Arterial Travel Time Data Using Independent Sources*. Research Summary. Maryland Department of Transportation, State Highway Administration, Hanover, MD. http://www.roads.maryland.gov/OPR_Research/MD-15-SHA-UM-3-6_INRIX-Arterial_Summary.pdf.

Horvitz, E.J., J. Apacible, R. Sarin, and L. Liao. 2005. Prediction, Expectation, and Surprise: Methods, Designs, and Study of a Deployed Traffic Forecasting Service. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*. Eds. F. Bacchus and T. Jaakkola. AUAI Press, Arlington, VA. 275–83.

Hu, J., M.D. Fontaine, B.B. Park, and J. Ma. 2015. Field Evaluations of an Adaptive Traffic Signal—Using Private-Sector Probe Data. *Journal of Transportation Engineering* 142 (1): 1–9. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000806.

Huang, H., and B.R. Humphreys. 2014. New Sports Facilities and Residential Housing Markets. *Journal of Regional Science* 54 (4): 629–63. https://doi.org/10.1111/jors.12120.

Humphreys, B.R., and L. Zhou. 2015. Sports Facilities, Agglomeration, and Public Subsidies. *Regional Science and Urban Economics* 54: 60–73. https://doi.org/doi.org/10.1016/j.regsciurbeco.2015.07.004.

Ishak, S. and H. Al-Deek. 2002. Performance Evaluation of Short-Term Time-Series Traffic Prediction Model. *Journal of Transportation Engineering* 128 (6): 490–98. https://doi.org/10.1061/(ASCE)0733-947X(2002)128:6(490).

Kim, S. and B. Coifman. 2014. Comparing INRIX Speed Data against Concurrent Loop Detector Stations over Several Months. *Transportation Research Part C: Emerging Technologies* 49: 59–72. https://doi.org/10.1016/j.trc.2014.10.002.

Kobayashi, M., K. Suzuki, and S. Nishimura. 2011. *Utilization of Probe Data for Traffic Flow Control*. Sumito Electric Industries, Ltd., Osaka, Japan. http://global-sei.com/its/common/pdf/2011_ITSWC_TS132-3167.pdf.

Kohavi, R. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence, Volume 2*. Morgan Kaufmann Publishers Inc., San Francisco, CA. 1137–1143.

Koller, D. and N. Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA. https://doi.org/10.1186/1471-2105-13-S15-S14.

Kulldorff, M., R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari. 2005. A Space-Time Permutation Scan Statistic for Disease Outbreak Detection. *PLoS Medicine* 2 (3): e59. https://doi.org/10.1371/journal.pmed.0020059.

Kwoczek, S., S. Di Martino, and W. Nejdl. 2014. *Predicting Traffic Congestion in Presence of Planned Special Events.* KSI Research Inc., Pittsburgh, PA. https://ksiresearchorg.ipage.com/seke/dms14paper/paper17.pdf.

Kwon, J., M. Mauch, and P. Varaiya. 2006. Components of Congestion: Delay from Incidents, Special Events, Lane Closures, Weather, Potential Ramp Metering Gain, and Excess Demand. *Transportation Research Record* 1959, 84–91. https://doi.org/10.1177/0361198106195900110.

Leilei, D., S. Zheng-Liang, and G. Jin-Gang. 2012. Study on Traffic Organization and Management Strategies for Large Special Events. In *Proceedings of the International Conference on System Science and Engineering (ICSSE), Dalian, Liaoning*. IEEE, Piscataway, NJ. 432–36. https://doi.org/10.1109/ICSSE.2012.6257222.

Lindveld, C.D.R., R. Thijs, P.H.L. Bovy, and N.J. der Zijpp. 2000. Evaluation of Online Travel Time Estimators and Predictors. *Transportation Research Record* 1719: 45–53.

Liu, C., B. Huang, M. Zhao, and S. Sarkar. 2016. Data Driven Exploration of Traffic Network System Dynamics Using High Resolution Probe Data. In *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, Piscataway, NJ. 7629–7634. https://doi.org/10.1109/CDC.2016.7799448.

Mason, C.M., D.C. Humphreys, and S.P. Pinch. 1983. The Externality Fields of Football Grounds: A Case Study of the Dell, Southampton. *Geoforum* 14 (4): 401–11. https://doi.org/10.1016/0016-7185(83)90037-4.

Matsumoto, H., M. Kobayashi, and K. Kamata. 2010. Traffic Flow Control Using Probe Vehicle Data. In Proceeding of 17th ITS World Congress, October 25–29, Busan, South Korea.

Miller, M. and C. Gupta. 2012. Mining Traffic Incidents to Forecast Impact. UrbComp ’12, August 12, Beijing, China. https://www.cs.uic.edu/~urbcomp2012/papers/UrbComp2012_Paper32_Miller.pdf.

Müller, M. 2015. What Makes an Event a Mega-Event? Definitions and Sizes. *Leisure Studies* 34 (6): 627–42. https://doi.org/10.1080/02614367.2014.993333.

Murphy, K.P. 2002. Dynamic Bayesian Networks: Representation, Inference and Learning. MS thesis. University of British Columbia, Vancouver, BC. https://www.cs.ubc.ca/~murphyk/Thesis/thesis.pdf.

Nagashima, Y., O. Hattori, and M. Kobayashi. 2014. Improvement of Traffic Signal Control Using Probe Data. *SEI Technical Review* 78: 44–47.

Nanthawichit, C. and T. Nakatsuji. 2003. Application of Probe-Vehicle Data for Real-Time Traffic-State Estimation and Short-Term Travel-Time Prediction on a Freeway. *Transportation Research Record* 1855: 49–59. https://doi.org/10.3141/1855-06.

Pan, B., U. Demiryurek, and C. Shahabi. 2012. Utilizing Real-World Transportation Data for Accurate Traffic Prediction. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*. IEEE Computer Society, Washington, DC. 595–604. https://doi.org/10.1109/ICDM.2012.52.

Pan, B., U. Demiryurek, and C. Shahabi. 2015. Forecasting Spatiotemporal Impact of Traffic Incidents on Road Networks. *Journal of Knowledge and Information Systems* 45 (1): 75–104. https://doi.org/10.1007/s10115-014-0783-6.

Park, D., L.R. Rilett, and G. Han. 1999. Spectral Basis Neural Networks for Real-Time Travel Time Forecasting. *Journal of Transportation Engineering* 125 (6). https://doi.org/10.1061/(ASCE)0733-947X(1999)125:6(515).

Pearl, J. and S. Russell. 2003. Bayesian Networks. In *The Handbook of Brain Theory and Neural Networks*, 2nd Ed., Ed M.A. Arbib. MIT Press, Cambridge, MA. http://ftp.cs.ucla.edu/pub/stat_ser/r216-reprint.pdf.

Pyun, H. and J.C. Hall. 2019. Does the Presence of Professional Football Cause Crime in a City? Evidence from Pontiac, Michigan. *Applied Economics*: 1–13. https://doi.org/10.1080/00036846.2019.1588942.

Rappaport, J. 2016. *Productivity, Congested Commuting, and Metro Size*. Federal Reserve Bank of Kansas City Working Paper No. RWP 16-03. Federal Reserve Bank of Kansas City, Kansas City, MO. https://doi.org/10.2139/ssrn.2727441.

Sadrsadat, H. and S.E. Young. 2011. Probability of Real-Time Data as a Function of Hourly Volume, Assessment of the I-95 Vehicle Probe Project Data. 18th World Congress on Intelligent Transport Systems, October 16–20, Orlando, FL.

Schuman, R. and R. Glancy. 2015. How Freight Probe Data Is Revolutionizing the Industry. 101st Purdue Road School, March 10–12, West Lafayette, IN. http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1502&context=roadschool.

Sharma, A., V. Ahsani, and S. Rawat. 2017. *Evaluation of Opportunities and Challenges of Using INRIX Data for Real-Time Performance Monitoring and Historical Trend Assessment*. 11-2017. Iowa State University, Ames, IA. https://lib.dr.iastate.edu/ccee_reports/24/.

Sun, S., C. Zhang, and G. Yu. 2006. A Bayesian Network Approach to Traffic Flow Forecasting. *IEEE Transactions on Intelligent Transportation Systems* 7 (1): 124–132. https://doi.org/10.1109/TITS.2006.869623.

van Lint, J.W.C. 2008. Online Learning Solutions for Freeway Travel Time Prediction. *IEEE Transactions on Intelligent Transportation Systems* 9 (1): 38–47. https://doi.org/10.1109/TITS.2008.915649.

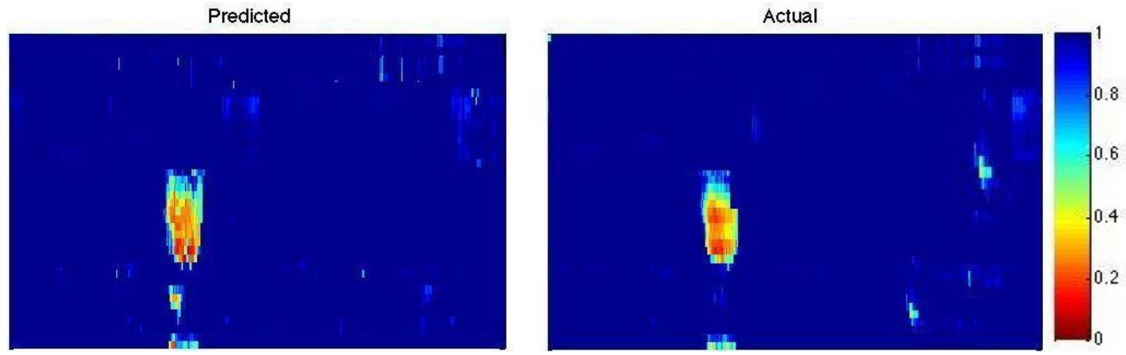
Vanajakshi, L. and L.R. Rilett. 2004. A Comparison of the Performance of Artificial Neural Networks and Support Vector Machines for the Prediction of Traffic Speed. In *IEEE Intelligent Vehicles Symposium, 2004*. IEEE, Piscataway, NJ. 194–199. https://doi.org/10.1109/IVS.2004.1336380.

Wu, C.H., J.M. Ho, and D.T. Lee. 2004. Travel-Time Prediction with Support Vector Regression. *IEEE Transactions on Intelligent Transportation Systems* 5 (4): 276–281. https://doi.org/10.1109/TITS.2004.837813.

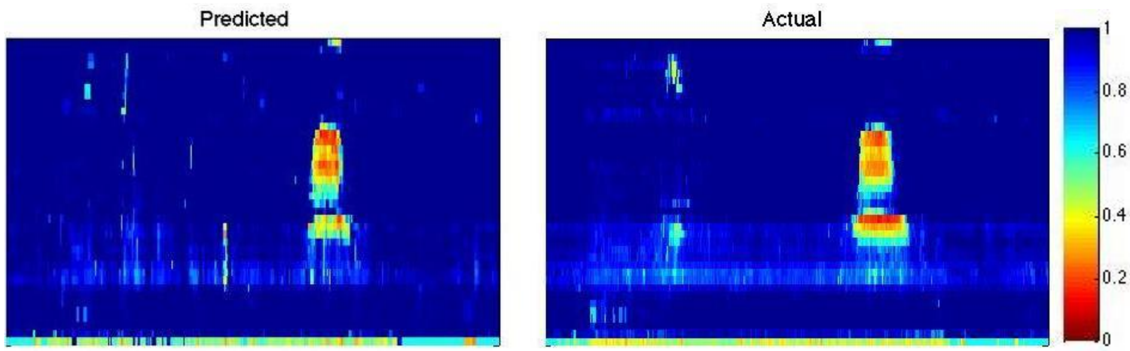
Young, S. 2007. Real-Time Traffic Operations Data Using Vehicle Probe Technology. 2007 Mid-Continent Transportation Research Symposium, August 23–24, Ames, IA.

Zhang, X. and J.A. Rice. 2003. Short-Term Travel Time Prediction. *Transportation Research Part C: Emerging Technologies* 11 (3–4): 187–210. https://doi.org/10.1016/S0968-090X(03)00026-3.

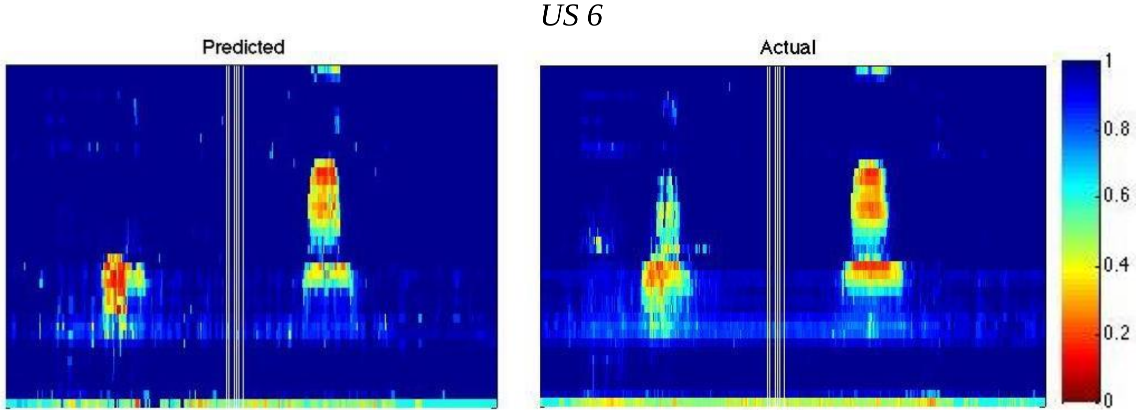
Appendix A: Heat Maps for NE 31, US 6, US 77, and NE 2



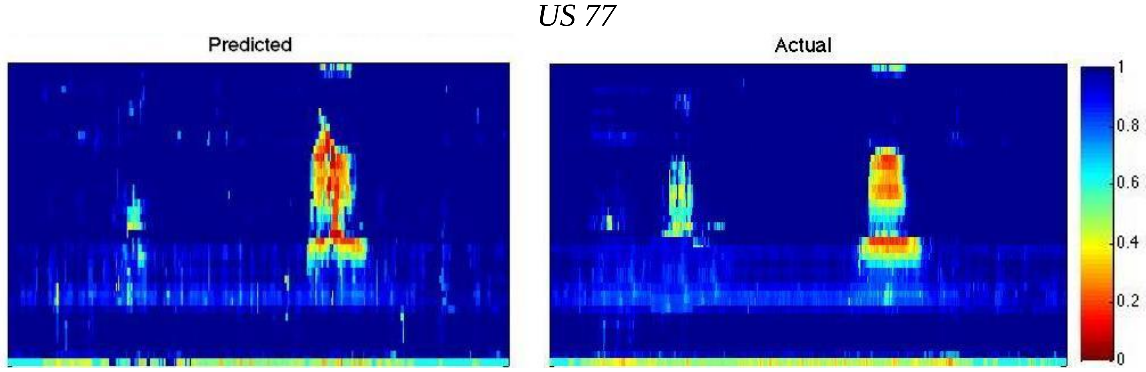
NE 31



US 6



US 77



NE 2

Figure A.1 Predicted and actual hotspot clusters showing traffic congestion on game days in 2018 on four routes: NE 31, US 6, US 77, and NE 2

Appendix B: Length and Duration of Hot Spots of Noon and Evening Game Days

Table B.1 Congestion length and duration of hot spots of noon and evening game days

Hot spot	Congestion length (mile)				Congestion duration (min)			
	Noon		Evening		Noon		Evening	
	Before start-time	After start-time	Before start-time	After start-time	Before start-time	After start-time	Before start-time	After start-time
1	1.37	1.48	0.37	0.88	18	24	9	18
2	1	1.5	0.75	0.8	17	34	11	14
3	0.9	1.6	0.35	0.55	32	50	13	20
4	1	0.8	0.3	0.1	46	35	18	8
5	2	1.04	0.6	0.35	39	25	17	10
6	1.9	1.25	0.8	0.29	22	13	20	9
7	0.7	0.45	0.8	0.35	46	23	27	13
8	0.9	1.54	0.5	0.85	13	41	9	17
9	1.2	1.35	0.2	0.65	20	18	7	21
10	0.4	2.9	0.2	0.85	18	55	10	25
11	0.7	3.3	0.8	2.55	8	33	8	28
12	0.9	1.45	0.3	0.65	12	32	9	19
13	1.2	0.74	1.05	0.6				
14	0.7	0.25	0.5	0.35	20	10	10	8
15	1.3	2.2	1.4	1.7	20	49	19	25
16	1.3	2.69	0.5	0.7	13	43	8	13
17	0.9	0.7	0.45	0.35	8	7	10	7
18	1.5	1.19	0.7	0.35	32	23	11	8
19	0.8	0.78	0.5	0.35	17	15	23	17