2018

QUANTIFYING NON-RECURRENT DELAY USING PROBE-VEHICLE DATA

Jacob Douglas Keaton Brashear
University of Kentucky, jacob.brashear@gmail.com
Digital Object Identifier: https://doi.org/10.13023/etd.2018.305

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Jacob Douglas Keaton Brashear, Student

Dr. Mei Chen, Major Professor

Dr. Timothy Taylor, Director of Graduate Studies

QUANTIFYING NON-RECURRENT DELAY USING PROBE-VEHICLE DATA

THESIS

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in Civil Engineering
in the College of Engineering
at University of Kentucky

By

Jacob Douglas Keaton Brashear

Lexington, Kentucky

Director: Dr. Mei Chen,

Associate Professor of Civil Engineering

Lexington Kentucky

2018

ABSTRACT

QUANTIFYING NON-RECURRENT DELAY USING PROBE-VEHICLE DATA

Current practices based on estimated volume and basic queuing theory to calculate delay resulting from non-recurrent congestion do not account for the day-to-day fluctuations in traffic. In an attempt to address this issue, probe GPS data are used to develop impact zone boundaries and calculate Vehicle Hours of Delay (VHD) for incidents stored in the Traffic Response and Incident Management Assisting the River City (TRIMARC) incident log in Louisville, KY. Multiple linear regression along with stepwise selection is used to generate models for the maximum queue length, the average queue length, and VHD to explore the factors that explain the impact boundary and VHD. Models predicting queue length do not explain significant amounts of variance but can be useful in queue spillback studies. Models predicting VHD are as effective as the data collected; models using cheaper-to-collect data sources explain less variance; models collecting more detailed data explained more variance. Models for VHD can be useful in incident management after action reviews and predicting road user costs.

Key Words: Probe GPS, Modeling, Multiple Linear Regression, Impact Zone, Vehicle Hours of Delay

Jacob Douglas Keaton Brashear

July 26, 2018

QUANTIFYING NON-RECURRENT DELAY USING PROBE-VEHICLE DATA

By

Jacob Douglas Keaton Brashear

Dr. Mei Chen
Director of Thesis

Dr. Timothy Taylor
Director of Graduate Studies

July 26, 2018
Date

Acknowledgements

First and foremost, without the help of my Lord and Savior Jesus Christ, nothing is possible. He has gifted me with the ability and desire to push forward in my endeavors in transportation research as well as my other passions in life. His guiding hand has been the ultimate motivator in completing this work.

I would like to thank my advisor, Dr. Mei Chen. Her desire to see me excel and her advice during my tenure as a master's candidate in the Department of Civil Engineering has kept me focused and driven. Also, I am eternally grateful for the opportunity as a research and teaching assistant which made returning to school possible.

I would like to thank my committee, Dr. Reginald Souleyrette, Dr. Gregory Erhardt, and Dr. Mei Chen for reviewing my work and providing feedback. I would also like to thank them for their knowledge they bestowed on me during my two years; I am a better version of myself and I feel more confident into entering the workforce thanks to you.

My co-workers, Xu Zhang, Obaidur Kazi, Galen Tanner Smith, and Fahmida Rahman are also a considerable reason this work was possible. Xu, thank you for reviewing my work and calming me down when I felt the weight of the world on my shoulders. Obaidur, thank you for helping me feel welcome when I first arrived on campus and providing insights on what to expect during this process; our discussions on the differences in cricket and baseball are also memorable. Tanner, your friendship, especially before college, has meant the world to me and I want to thank you for convincing me to work for Dr. Chen; it was life-changing. Fahmida, thank you for showing bravery when coming to a new country and jumping straight into research. It has been inspiring watching you overcome adversity and has impacted my work positively.

I want to extend special gratitude to one of my closest colleagues, Alex Mucci, for his continued support and friendship throughout these two years. Endless nights of Halo, studying, and traversing the TRB conference allowed us to develop a friendship that I'll cherish for the rest of my life. I now consider you a brother and I am grateful for your honest advice and endless support. I'm going to miss working with you.

Without the help of my family and friends who have provided emotional support, this work would not be possible. My mom, Brenda Brashear, and my dad, Doug Brashear, are the foundation of my very being. I am the man I am today with their love and support through this process and throughout life. I'd like to thank my brother, James Feltner, as well as his wife and two children, Jessica Feltner, Bentley Feltner, and Ava Feltner, for their continued love and support upon my return to Lexington, KY. Thank you all.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Quantifying Congestion Costs

The 2015 Urban Mobility Scorecard recently conducted a cost analysis of the level of congestion in the United States. It stated that the problem is "very large" and that in 2014 Americans experienced "an extra 6.9 billion hours" of travel and purchased "3.1 billion gallons of fuel" leading to a "congestion cost of $160 billion" (*1*). Based on these data and Americans' love affair with cars, traffic congestion is becoming increasingly worse and must be addressed. It begs the question of what avenues are available leading to limiting or eliminating the damage of congestion. One particular avenue involves addressing non-recurrent congestion.

Non-recurrent congestion is defined as congestion caused by extraneous events outside of normal day-to-day patterns. For example, weather events, crashes, and work zones are all considered as causes of non-recurrent congestion. Non-recurrent congestion, as a result, lead to workplace tardiness and delayed shipments which results in time lost to roadway users and dollars lost to businesses. Examples of current strategies deployed to alleviate non-recurrent congestion include incident management strategies for crashes and road user cost allocation to contractors for work zones. To be effective, delay must be quantified for both types of incidents to know the potential cost-benefit of a given strategy.

Existing methodologies typically use aggregated volume numbers such as AADT and hourly volume distribution factors to estimate hourly traffic flows. These flows are then used to develop capacities, estimate queues, and calculate delay for roadway incidents. This method does not consider the day-to-day conditions present in everyday traffic which can lead to potential inaccuracies in quantifying congestion. Data that can capture these day-to-day fluctuations could possibly improve calculating congestion.

One possible avenue to capture day-to-day conditions is the use of big data applications such as speed data from private vendors like INRIX (*1*). Vendors sell these data from a database of what are known as probe GPS data, or speed data collected from roadway users during every day commutes or other driving activity. These data are a powerful tool in understanding congestion due to its spatial and temporal detail along various roadways and can be used to identify slowdowns and calculate, for example, the number of miles of congested roadway (*2*).

1.2 Research Goals

The primary goal of this research is to explore the relationships between the factors that describe the incident and the delay caused by the incident. It is the hope of this research that the day-to-day fluctuations of traffic can be captured to generate better estimates of delay. This study will pursue this goal by using a combination of probe GPS data purchased by the Kentucky Transportation Cabinet (KYTC) and an incident log monitored by Louisville, KY through their Intelligent Transportation Systems initiative called the Traffic Response and Incident Management Assisting the River City

(TRIMARC) program. First, the TRIMARC incident dataset will be conflated with the probe GPS dataset from KYTC to identify the speed reductions potentially attributed to the incident.

Next, the "impact zone" or spatial and temporal boundary of the incident, will be identified. After that, the speed reductions found within the impact zone will be quantified into Vehicle Hours of Delay (VHD) for the incident. Finally, the study will incorporate regression modeling on the impact zone dimensions and VHD to draw conclusions on the relationships that best describe the impact zone and congestion.

1.3 Organization of Document

The document consists of six chapters. Chapter 1 will introduce the topic and research goals. Chapter 2 will overview the comprehensive literature search prior to and throughout the research process. Chapter 3 will overview the study area and data sources used during the research and outline the data processing steps to generate the impact zone and VHD metric. Chapter 4 will detail the modeling process and present advantages and disadvantages of both modeling procedures. Chapter 5 will conclude the document and summarize the findings.

This section overviewed the goals of the project and outlined the structure of the document. The next section will overview the literatures that helped guide and inspire this work.

Chapter 2 Literature Review

Congestion management is an expansive topic. With multiple approaches and access to more data than in recent memory, various approaches have been implemented to attempt to answer the question of how to quantify and alleviate delay. Methods such as simulation, modeling, and deterministic queueing have been instrumental in the research behind understanding the nature of delay and what strategies best address the delay in concern. Given the scope of this work is to calculate a delay metric for each incident and explore the relationship between the calculated delay metrics as well as given information about the incident, a review in previous practices on how delay is quantified is warranted. Below is a review of these practices.

2.1 Current Practice in Quantifying Congestion

Beginning to understand how delay is quantified required studying the state of the practice of how congestion is identified. When reviewing literatures on this topic, it was noted that methods pertaining to congestion were especially popular in work zone analysis. The traditional approach to modeling congestion is deterministic queuing or using inflows and outflows of vehicles in an area to determine whether queues propagate or dissipate (*3-6*). Some have implemented the use of simulation software, such as VISSIM (*7*), Quickzone (*8*), and Netzone (*9*) to both assess road user costs and develop planning-level routing strategies for roadways. Traffic monitoring has also been adopted for delay management strategies. Most notably, the Washington DOT is using a background volume profile to capture incident-induced delay with data collected from loop detectors (*10*).

The aforementioned methods use aggregated data to calculate delay and measure the impact area. More specifically, tools like Quickzone and the Kentucky User Cost Program (KYUCP) use AADT to generate results (*4; 8*). The main disadvantage of using aggregated data sources is that the day-to-day or even the minute-to-minute fluctuations in traffic are not accounted for. More recent research has focused on quantifying congestion and measuring the impacted area using techniques that both build on using volume data and utilizing new data sources such as probe GPS speed data.

2.2 Research in Quantifying Congestion

Researchers in previous studies have used various modeling techniques in identifying and quantifying congestion which can be categorized by:

- Simulation
- Machine Learning
- Statistical Modeling
- Traffic Monitoring
- Utilization of Background Speed Profile

Park et al. developed a simulation model of incident conditions and non-incident conditions using the INTEGRATION microscopic traffic simulation software (*11*). The

study found that incidents "increase the mean travel time and variability in travel times for the congested condition." Other simulation approaches utilize unique processes such as a binary integer algorithm as proposed by Chung and Recker as well as tools such as VISSIM (*12; 13*).

Some research exists in using machine learning techniques in both predicting traffic flows (*14; 15*) and modeling traffic impacts (*16*). Du et al. developed a multilayer feed forward artificial neural network (ANN) model to "estimate work zone delay using probe-vehicle data" (*16*). It was found that the modeling technique outperformed the traditional deterministic queuing technique in accurately "estimating work zone costs and performance measures." It was also observed that the implementation of probe vehicle speeds were superior to using traffic volumes in this case, given that in some cases it is more difficult to find volume-based data for a specific work zone at a specific time. Edara et al also developed a process to predict travel times in work zones using Random Forests (*17*).

Various statistical modeling and empirical analyses were examined in this review. Multiple linear regression was found to be a choice for two papers that both modeled impact duration and VHD (*18*) as well as work zone speeds (*13*). Other statistical techniques, such as the t-test, were also utilized. Seeherman et al. utilized t-tests to determine the effect of weather at freeway bottlenecks and freeway merge sections (*19*). The study concluded that for the study areas "Discharge flow during rainy days dropped by an average of 12.6% at the lane drop and 13.6% at the merge, with both differences being significant at the 95% confidence level." Wright et al. used an empirical analysis with travel times derived from volumes collected by loop detector data to study the differences in travel time, travel time variability, buffer index, and probability of breakdown for incidents with a shoulder blockage, one lane blocked, or multiple lanes blocked as compared to normal conditions (*20*). It was found that travel time variability and buffer index were highest with incidents with multiple blocked lanes. Also, shoulder incidents were found to produce much higher probabilities of breakdown as compared to normal conditions.

Areas of research have focused on improving traffic monitoring in identifying and quantifying congestion. For example, McNamara et al. developed a "congestion ticker" which utilizes probe GPS data to identify congestion where speeds are less than 45 mph (*2*). The congestion ticker can be used for "after-action review of major events such as ice storms, major crashes, and construction work zones." Traffic monitoring has also been used in quantifying congestion on arterials and investigating Travel Time Reliability metrics (*21; 22*).

The background speed profile is also a valuable tool in quantifying the effect of congestion with respect to a "normal day." However, the definition of a "normal day" can vary from the type of background profile developed. For example, researchers in the past have used the average speed of incident-free days as a benchmark (*23*). Recent research has also explored the use of algorithms such as kalman filter, k-nearest neighbor,

day-of-week matching, clustering, and gaussian mixture models since using the average background speed may not be representative of the speeds for the day of the incident (*18; 23-25*).

2.3 Research in Identifying Impact Zone

It was noted that extensive research is available in identifying and quantifying the spatiotemporal areas of non-recurrent congestion, chiefly, queue length and duration. First, data must be collected and processed to identify and quantify the spatiotemporal impact of an incident whether it be through statistical means such as Sullivan et al.'s standard normal deviate or Li et al.'s delta speed to identify the back of the queue (*26-29*). Modeling is also a popular topic especially when attempting to predict impact duration of crashes (*18; 30-32*). The reason is because impact duration can be used as an indicator of vehicle delay (*18*) and to allow TMCs to choose "the appropriate response to an incident" (*32*). Linear regression modeling as well as analysis that uses linear regression as a basis, such as ANOVA, have been popular in the past (*18; 30; 33*). However, other studies have utilized other methods of modeling such as Logistic Regression, Quantile Regression, and copula-based models (*18; 31*).

One study found to be particularly interesting is Kazi's work in developing impact boundaries to ascertain the "data-driven" effect of crashes in Louisville, Ky (*18*). The study used TRIMARC stationary sensor speed data to develop spatiotemporal impact zones for crashes recorded in the TRIMARC incident log. Crash delay was calculated using a difference between the background speed profile as developed by the kalman filter algorithm and the speeds during the crash incident. Then, the study developed three models for impact duration using multiple linear regression, logistic regression, and quantile regression. Kazi also developed a linear model to estimate impact delay. Kazi found the post-crash mean speed and the weather to be contributing factors to impact duration in the linear model, the logistic model, and the quantile model for incident duration. However, the effects of injury are found significant in the higher quantiles of the quantile regression model. For the delay model, the post-crash mean speed and the impact duration are found as significant factors in explaining delay. The study also conducted a reliability analysis using Cumulative Frequency Diagrams (CDFs) to illustrate different scenarios.

2.4 Emergence of Probe GPS Data

In recent years, the emergence of probe GPS speed data have become increasingly popular in congestion analysis primarily due to the cost of operating and maintaining sensors (*34*). Without the burden of operating and maintaining sensors, efforts can be redirected to develop tools to utilize the data for congestion analysis. Based on this knowledge, it is of great interest to conduct a review of the available data and the current practices in using probe GPS data in congestion management.

On example of data available for use in transportation research is the National Performance Management Research Data Set (NPMRDS), which is available in Traffic

Message Channel (TMC) format for the National Highway System (NHS) (*35*). The Federal Highway Administration (FHWA) purchases these data from vendors and allow MPOs and DOTs free access to the data for various purposes. If a DOT or MPO wanted non-NHS data or data at finer spatial granularity, they would have to purchase the data from the vendors themselves.

In the transportation data marketplace, more than just probe GPS data are available for purchase. Various suites are also available to assist in the calculation of congestion metrics using these data sources. For example, tools such as the Regional Integrated Transportation Information System (RITIS) and the Iteris Performance Measurement System (iPeMS) handle speed data from multiple sources including probe GPS data and fuse them with information such as weather data, traffic incidents, and signal timing to detect bottlenecks, develop travel time reliability measures, and conduct after-action reviews on congestion events (*36; 37*). As an example, the Oregon Department of Transportation (ODOT) used iPeMS to conduct a congestion study around the solar eclipse event in 2017 (*38*). Using the data and tools provided within iPeMS, bottlenecks caused by event traffic were identified and lessons learned were compiled to ensure ODOT would be prepared for the upcoming 2024 total eclipse.

Tools are also being developed by DOTs to handle probe vehicle data. For example, Chien et al developed a tool called Work Zone Interactive Management Application-Planning (WIMAP-P), which is a "work zone lane-closure impact prediction system" used to plan and schedule work zones in New Jersey (*39*).

2.5 Synthesis of Research

The applications of probe GPS speed data in traffic monitoring, congestion management, and reducing data collection costs promise great potential in reshaping how roadway projects are prioritized and how future applications will shape the transportation infrastructure. It is of great interest to research the potential uses of these data and how they can be leveraged to maximize the effectiveness of new projects through congestion reduction and reducing the cost to contractors through more accurate road user cost estimations.

As stated earlier, Kazi developed models for impact duration and delay using stationary sensor data from TRIMARC(*18*). Even though stationary sensor data can capture the day-to-day fluctuations of traffic, it was noted that the average spacing between sensors is 0.4 miles and the data were aggregated to the 15-minute level. These levels of aggregation increase the difficulty of attaining accurate values for impact duration as well as queue length (which was not covered in Kazi's study) so it is a focus of this work to explore data at more detailed levels of aggregation.

Building from the research presented by Kazi, an analysis approach using "third party data" is proposed(*18*). Probe GPS data purchased by KYTC is the selected data source for speed data. These data are presented in 5-min aggregation and at "link" level which is spatially more detailed than the stationary sensor data from TRIMARC. This greater

level of spatial detail will allow the research to explore quantifying queue length for incidents and its association with delay. The analysis will conflate these data with the TRIMARC incident log so weekday incidents can be identified using a combination of a background speed profile comprising of average historic speeds for weekday non-holidays. Once these incidents are identified, a custom python-based search algorithm will be implemented to identify the spatial and temporal boundaries of the impact zone. With this boundary information, the speed and volume data associated with the impact boundary will be used to calculate VHD. Also, linear models will be developed to explore the significant factors explaining the dimensions of the impact zone and the delay of the incident. A challenge moving forward with using this probe GPS dataset is the presence of missing data, or times when data are not collected for a given roadway segment. The study must address these concerns by imputing speed data for where there are missing data within the defined impact boundaries.

This section described the prevailing literature on the subject of congestion and the impact zone. The next section overviews the data sources and details the process of defining the impact zone and calculating VHD.

Chapter 3 Data Sources and Processing

This chapter will overview the data sources used for this study. Then, this chapter will detail the process of how the impact zone and VHD are calculated.

The corridor selected for study is 35 miles of I-64 East and I-64 West in Louisville, KY. The corridor stretches from Shelbyville, KY to the Kentucky/Indiana Borderline on the west side of Louisville. Along this corridor, probe GPS data from a total of 324 links have been extracted of which 164 are on I-64E and 160 are on I-64W. Figure 1 shows an overview of the study area.

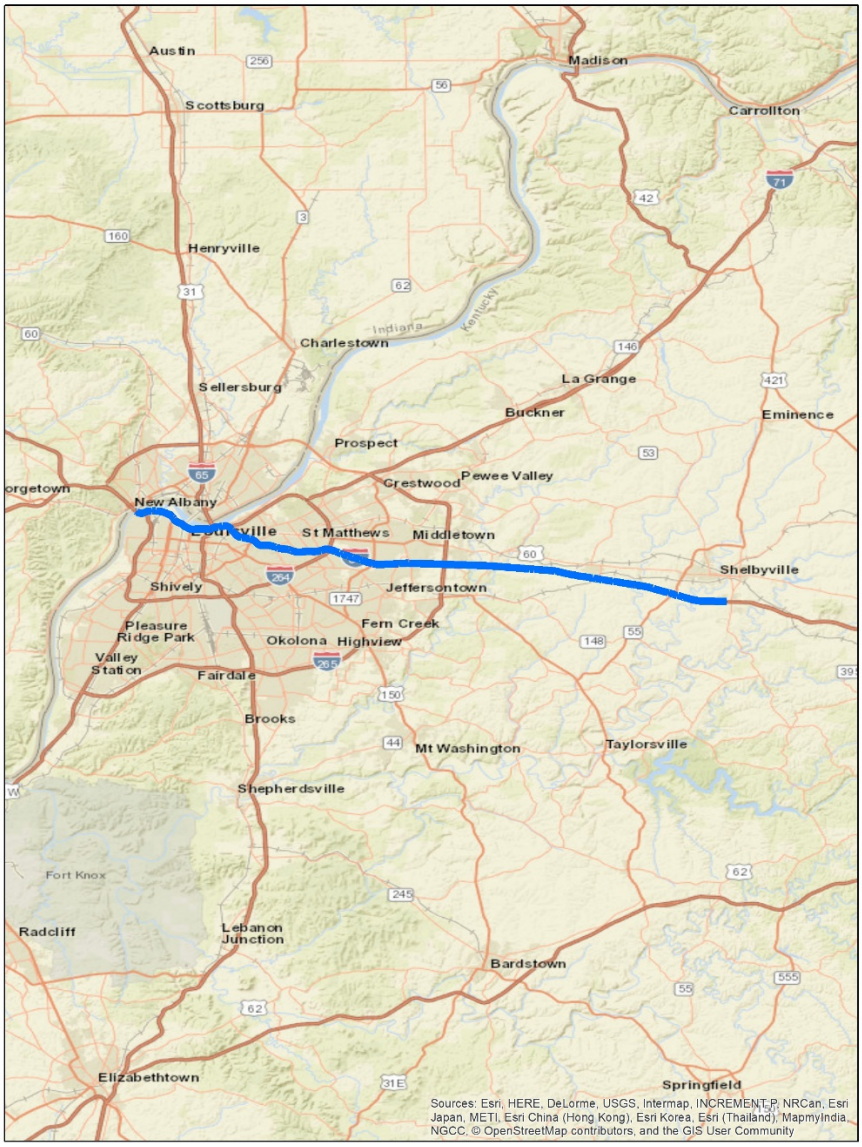


Figure 1 Study Area: I-64 in Louisville, KY

3.1 Probe GPS Data

Probe GPS speed data from a private data vendor are used for this research. Two years of data from January 2013 to December 2014 were extracted for the links along the corridor and the speeds are aggregated to 5-minutes. In terms of spatial detail, the data links along the roadway average 0.2 miles with the longest link measuring 2.3 miles long and the shortest at approximately 34 feet. Details provided by the data are the link the speeds are associated with, the date and time of collection/aggregation, and the mean speed of the 5-min aggregation period. The dataset selected is considered a "probe" dataset, in which there was no smoothing or preprocessing performed by the data vendor.

Since the original data do not contain mile markers, a process was developed to generate the start and end mile markers for each probe GPS link. This is done by starting at the last link on the segment and identifying the mile marker using the KYTC Photo Log. Then, the link lengths can be used to calculate the mile marker location of all the other nodes.

3.2 TRIMARC Incident Log

The initial focus of this research was to use work zone data from TRIMARC for analysis. However, during initial studies, it was found that adjustments were required to proceed with the analysis. For example, work zone data are concentrated during the nighttime when less traffic is present. Given probe GPS data are scarce during the nighttime, not enough information is present to illustrate the delays present during the work zone period. Therefore, crashes are used instead. Also, preliminary analysis showed that little congestion is present within the TRIMARC dataset for incidents that do not at least block a lane. As a result, only lane-blocking crashes are considered for this analysis.

Two years of lane-blocking crashes ranging from January 2013 to December 2014 are extracted from the TRIMARC incident log. Table 1 shows a sample from the TRIMARC incident log, respectively.

Table 1 Sample TRIMARC Incident Log for I64E in 2013

Type	Start Date	Start Time	End Date	End Time	MP	Conditions	blockedLanes	Injury	Weather
Accident	1/21/2013	17:23:00	1/21/2013	18:09:00	1.5	Dry Pavement, Sunny, Non-incapacitating Injury, Rear-End Collision, Vehicle Damage, Car, Pickup Truck	2	Injury	Sunny
Accident	1/23/2013	17:34:00	1/23/2013	18:25:00	17.8	Overcast Cloudy, Unknown Collision, Car, Pickup Truck	1	No Injury	Overcast Cloudy
Accident	1/25/2013	9:09:00	1/25/2013	10:08:00	1.8	Ice, Vehicle Damage, Car, SUV, FSP Service, Ice Pavement	1	Injury	Ice
Accident	1/25/2013	9:30:00	1/25/2013	9:50:00	1.6	Ice, Car, Pickup Truck, SUV, Single Unit Truck, FSP Service, Ice Pavement	2	Injury	Ice
Accident	1/28/2013	17:04:00	1/28/2013	17:40:00	8	Wet Pavement, Rain, Rear-End Collision, Car, Pickup Truck, FSP Service	1	No Injury	Rain
Accident	2/11/2013	16:14:00	2/11/2013	17:13:00	14.8	Chemically Wet Pavement, Sunny, Incapacitating Injury, Vehicle Damage, Car, FSP Service, Vehicle Overturned	1	Injury	Sunny
Accident	2/20/2013	17:54:00	2/20/2013	18:14:00	19	Overcast Cloudy, Rear-End Collision, Car	1	Injury	Overcast Cloudy
Accident	3/5/2013	13:00:00	3/5/2013	14:09:00	17	Wet Pavement, Rain, Possible Injury	1	Injury	Rain
Accident	3/8/2013	7:03:00	3/8/2013	7:22:00	3.7	Overcast Cloudy, Car, FSP Service	1	Unknown	Overcast Cloudy

For this analysis, weather and injury severity were of interest. Weather and injury severity data are present in the TRIMARC incident log. However, the data are incomplete for both categories. To complete the weather dataset, weather underground was used (*40*). To complete the injury category, the injury column was first coded as whether the incident had any injuries. Next, data from the Kentucky State Police (KSP) database were used to code injuries for crashes with unknown injuries with respect to the TRIMARC incident log. Of these that did not have data from either the TRIMARC incident log or the KSP database, these were assumed to have no injury. It was decided that the severity of the injury will not be used because the KSP database does not record injury severity. If injury severity were to be used, either an assumption of the injury severity for the KSP-imputed incidents would be required or an unknown injury variable would be created.

3.3 Traffic Volumes

KYTC has 16 counting stations deployed along I-64 within the study area. These stations keep track of Average Annual Daily Traffic (AADT), K factors, D factors, and truck percentages. AADT was collected during the 2013 calendar year because the difference between 2013 and 2014 AADT were small according to the Automatic Traffic Recorder (ATR) sensor located along the roadway (62769 in 2013 vs 61867 for 2014). If 2013 data are not available, the closest year was collected. Table 2 shows the AADT by milepoint. Figure 2 is an example of a readout of a sensor readout as found in the KYTC traffic database (*41*).

Table 2 AADT By Milepoint

Begin Mile Point	End Mile Point	AADT	Year Collected
0	0.818	78534	2014
0.818	2.74	62769	2013
2.74	4.052	66992	2013
4.052	4.759	89526	2013
4.759	4.995	90900	2010
4.995	5.967	135400	Imputed
5.967	6.303	79991	2013
6.303	7.809	74600	2011
7.809	10.308	79308	2014
10.308	12.275	77818	2013
12.275	15.018	130876	2013
15.018	17.177	92843	2013
17.177	18.956	84200	2011
18.956	27.596	56852	2013
27.596	31.842	50900	2011
31.842	35.163	49662	2013

Historical Traffic Volume Summary

Station Details:

Sta ID:	056P92
Sta Type:	Permanent (ATR)
Map:	MapIt
District:	5
County:	Jefferson
Route:	056-I -0064 -000
Route Desc:	I-64

Begin MP:	0.8180
Begin Desc:	I 264
End Mp:	2.74
End Desc:	22ND ST & NORTHWESTERN PKWY BR
Impact Year:	1994
Year Added:	1979

Newest Count:

AADT:	65301
Year:	2017
% Single:	3.2850
% Combo:	5.5140
K Factor:	10.20
D Factor:	72

Definitions:
Sta. ID - Three digit county number + station number
MP - milepoint
Impact Year – year of significant change to traffic pattern within station segment
AADT – Annual Average Daily Traffic – the annualized average 24-hour volume of vehicles on a segment of roadway
% Single – single unit truck volume as a percentage of the AADT
% Combo – combination truck volume as a percentage of the AADT
K Factor – peak hour volume as a percentage of the AADT
D Factor – percentage of peak hour volume flowing in the peak direction

Year	AADT	Year	AADT	Year	AADT
2018		2008	66400	1998	57600
2017	65301	2007		1997	58100
2016	68131	2006	68300	1996	57300
2015	60628	2005	68200	1995	54400
2014	61867	2004	69600	1994	52400
2013	62769	2003	70200	1993	59000
2012	58146	2002	69300	1992	55200
2011		2001	65900	1991	51600
2010	64600	2000	63200	1990	58700
2009	60400	1999	63700	1989	50000

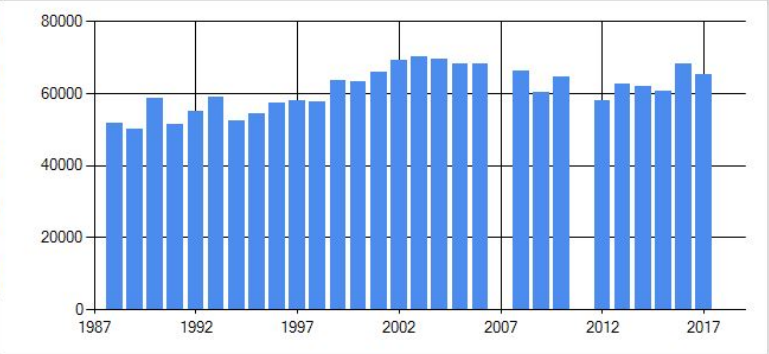


Figure 2 Detail of KYTC Traffic Counts

These counts are used to calculate VHD by first associating the counts with the mile points of the incident locations for the 2013 calendar year or whichever year was closest. The details of this process are outlined further in Section 3.4.3.

Before the delay analysis could begin, one unusual record was investigated as indicated by the “Imputed” denotation in Table 2. It took place between mile marker 5 and mile marker 6 on sensor 056M84. The reading on this sensor was 144000 vehicles in the 2008, which was thought of as strange because it doubled the counts preceding and succeeding the station. Also, it was strange that the sensor ceased counting after 2008. After contacting KYTC, it was determined that this was indeed a mainline count and that data collection past 2008 ceased due to construction along the roadway. Therefore, since counts preceding and succeeding the sensor were similar between 2008 and 2013, it was assumed that the counts for this segment did not drastically change. Therefore, it was necessary to get an imputed value for sensor 056M84 for 2013. It was noticed that in the area, I64 splits and the start of I71 begins as denoted by sensors 056M88 (I71) and 056M86 (I64). Therefore, the combined values of the two sensors would yield a similar count to 056M84. The combined values for 2008 counts for the two sensors was 151800, which is less than 10000 vehicles different from the reported 144000 at 056M84. Therefore, it is likely that these values are comparable since there is no other access for vehicle to enter or leave the roadway. Based on this finding, a ratio of the combined

values over the value at 056M84 was generated and the result is 1.0542. Then, 056M86 and 056M88 were combined again only using 2013 data which is 142756. Using this ratio, an imputed 2013 volume of approximately 135400 was generated for 056M84 by dividing the 2013 combined counts after the split with the ratio. 135400 was used for this sensor for the analysis found in Section 3.4.3.

3.4 Identifying the Impact Boundary and Calculating Delay

This section will overview the process of identifying the impact zone and calculating the delay associated with the incident. It is important to understand what constitutes the impact zone because each incident affects traffic by how long it is present, how far back the queue propagates, and the speeds at which traffic flow through the impacted area. The impact zone is identified with two sources: 1. TRIMARC incident information containing the mile marker, the starting time of the incident, and the end time of the incident, and 2. the speed data considered statistically congested. Beginning at the start time and location of the incident, the impact zone is identified by searching the congested areas for the extents of the impact boundary to generate information on incident duration and queue length. Then, the impact zone is used to calculate delay by imputing speeds that are missing and using volume data to calculate VHD.

3.4.1 Background Speed Profile

First, to address what typically happens along any portion of the roadway at any given time, a background profile must be developed. This profile is used to compare with speed data for a given incident day to determine the difference in speed between what occurred on an incident day and what typically occurs. The data used in this work are weekday data that did not occur on holidays because the incident data selected occur on weekdays. Developing a background profile which reflects the typical commuting patterns of traffic is the most sensible approach to developing a relationship between what is considered normal operation and what is considered abnormal. The background profile is constructed by taking the mean average of the speeds within a certain 5-minute window and for a certain probe GPS link for all time periods and links. These features are called “cells.” The result is mean speeds in a two-dimensional profile of cells where the x axis is the time of day and the y axis is the mile marker along the roadway. An example of a background speed profile is provided in Figure 3.

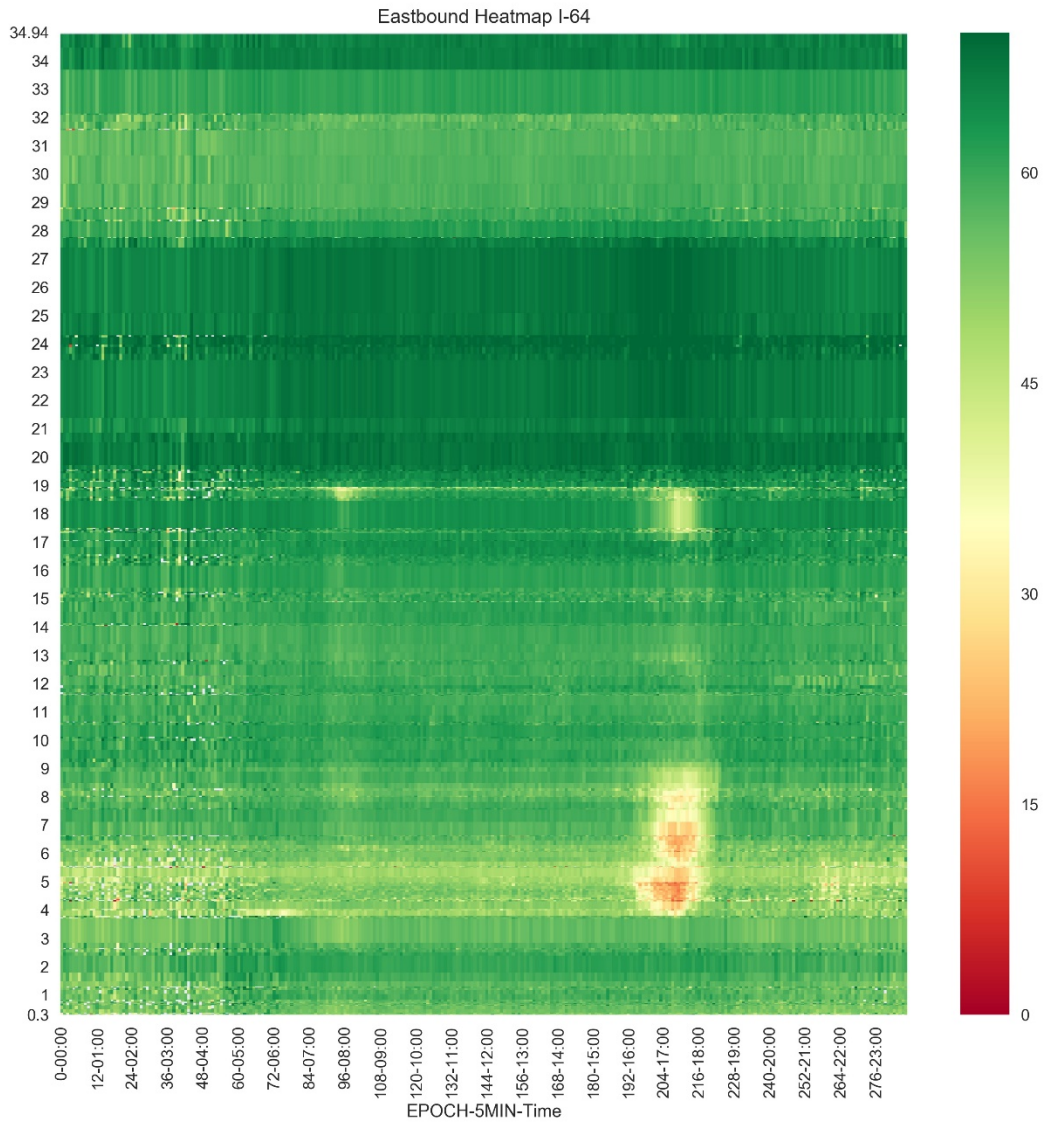


Figure 3 Background Speed Profile for I64E in 2013

3.4.2 Identifying Impact Zone

Identifying the impact zone requires knowledge of when and where the incident occurred as well as the time and places where congestion occurs during the day of the incident. After these are established, a search can be conducted to determine the impact boundary of the incidents. Finally, metrics such as queue length and duration of the incidents can be generated from the impact boundaries. The following will detail how the congested region is defined, how the search was conducted, and how the impact zone characteristics such as queue length and duration are calculated.

Determining the congested regions requires knowledge of the speed difference between the incident day speeds and the speeds in the background profile. Although knowing the speed difference is useful for calculating delay, it is not as useful when defining what

constitutes the impact boundary. Because of the dynamic nature of traffic, speed fluctuations exist but may not be considered congestion. Therefore, a threshold must be identified that constitutes congestion outside of normal traffic fluctuations so that a clear boundary can be identified. One method to do this is to use the coefficient of variation, which is defined as the ratio between the standard deviation and the mean of the data. It is defined in the FHWA Travel Time Data Collection Handbook as follows (*42*).

$$c.v. = \frac{\sigma}{\mu} \; or \frac{s}{x}$$

Where:

c.v. = coefficient of variation

σ = population standard deviation

μ = population mean

s = sample standard deviation

x = sample mean

The typical c.v. value for freeways is 15-25%. Therefore, the assumption of a 20% speed threshold will be used for the project, in that, if speeds for a given time and section drop below 80% of the typical background speed, then it will be considered congested.

Although the 20% speed threshold is used for identifying impact boundaries, it is not used when aggregating delay because there is the possibility that "uncongested" cells exist within the congested region but still exhibit a speed drop. This speed drop must be aggregated because, despite not statistically being congested, is still a speed reduction resulting in a delay; not aggregating this delay will cause a systematic underestimation of the total delay. There are also instances within the impact boundary that speeds are greater than the background speed. These are assumed to have zero delay and are not used in calculation. The heatmap in Figure 4 shows an example of the end-result of this process. Red cells denote cells that are operating under normal conditions, green cells represent cells that are considered statistically congested, and gray cells are cells where there are no data for speeds on the day in question.

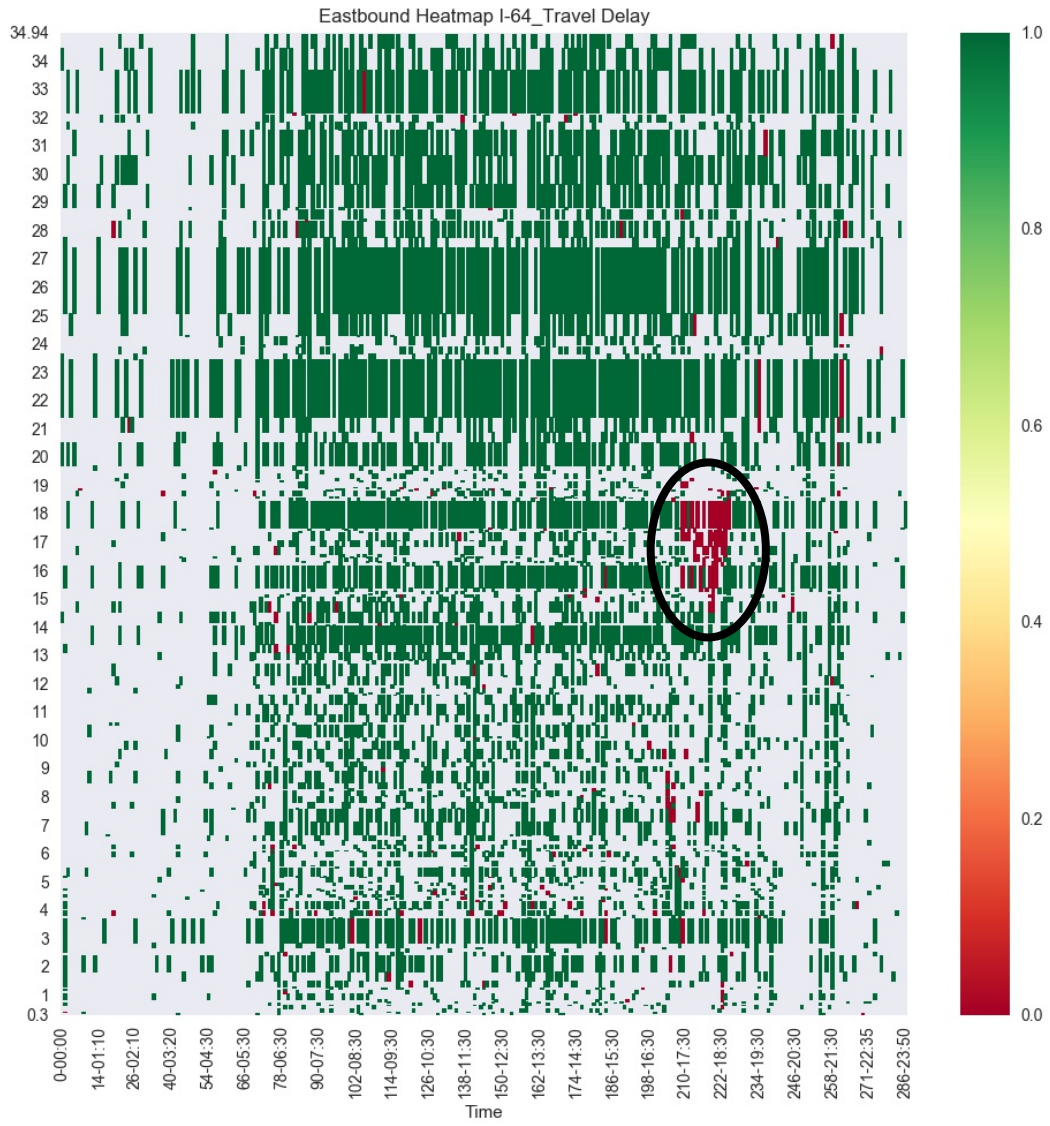


Figure 4 Congestion for Incident on 01/23/2013 at Mile Marker 18

After the congested cells have been identified, they are used to calculate the impact zone boundaries.

To determine the impact zone boundary, an algorithm was written in Python script and implemented to identify the temporal congestion bounds along the roadway. The algorithm takes information about the start time and the starting mile marker from the TRIMARC incident log and begins a search based upon the calculated or assumed duration of the incident. One challenge of this process was that the probe GPS data are aggregated to the nearest 5 minutes. Since the TRIMARC incident log times are in more detail, they are truncated to the nearest 5 minutes to allow the start time and the probe GPS speed data to be directly comparable.

The algorithm identifies the start and end of congestion for each affected link by searching the probe GPS speed data beginning at the start time and mile marker of the incident. With respect to space, the algorithm searches for 5 miles upstream from the point the incident occurs assuming: 1. the algorithm does not run outside of the bounds of the study zone or 2. not enough data are present to determine the end of the queue has been reached. Details on how the algorithm stops by detecting the end of the queue based on the given data is explained in the following paragraphs. With respect to time, the algorithm iterates over each 5-minute period for each link and searches 1.25 hours past the TRIMARC-reported end time but no longer than 6 hours past the TRIMARC-reported start time for each link. When the search is complete for the link in question, the link will begin to search the adjacent upstream link unless the algorithm determines it should stop. These constraints do not mean that all incidents will be properly identified based on these constraints because there are, for example, incidents that are longer than 5 miles. This is addressed in the manual update phase of this report explained later. This algorithm is merely to serve as a high-level tool to generate a baseline for all the impact zones to reduce the manual workload.

As mentioned earlier, the algorithm uses data within each link to determine whether the algorithm should stop. During the algorithm search, the count of three data items are identified: the number of congested cells, the number of uncongested cells, and the number of empty cells for each link. Using these data, a determination is made on whether the algorithm should stop based on the logic provided in Figure 5.

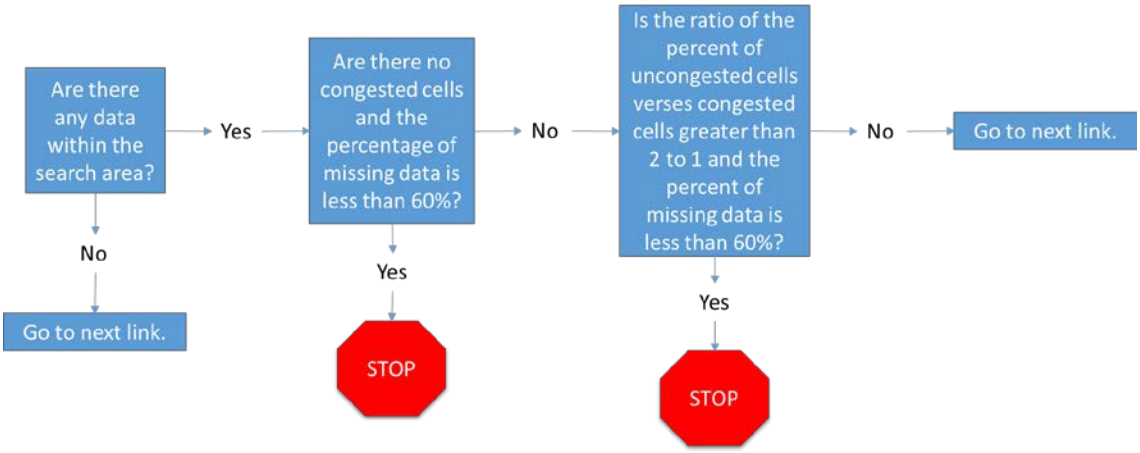


Figure 5 Algorithm for Detecting the End of the Queue

As indicated by Figure 5, a ratio of the percent of uncongested verses congested cells is used alongside a percent of missing data. The ratio determines whether the number of uncongested cells warrants the end of the queue has been reached. Too high of a ratio and the algorithm runs the risk of overestimating the maximum queue length; too low and

the algorithm could stop prematurely. Also, a percent of missing data metric is used to determine whether enough data are present to stop the algorithm. This is to ensure the determination made is reasonably confident given enough data are present.

This study employed a heuristic approach to determine the ratio and percent of missing data. Five different ratios and five different percent of missing data values were selected. The ratios selected are 1:1, 2:1, 3:1, 4:1, and 5:1 and the percent missing data values selected are 40%, 50%, 60%, 70%, and 80%. All 25 possible combinations of ratios and percent missing data are tested using the algorithm and the resulting maximum queue lengths and end times are compared to the manually verified end times and queue lengths as shown on the developed heatmaps. The smallest amount of variance in maximum queue length and end time would determine the optimum combination of ratio and percent missing data. The test was carried out on I64E 2013 data. After commencement of the test, a 2 to 1 ratio and a percent of missing data less than 60% was determined to produce the least error and was used for the algorithm for both years and both directions of incident data.

Once the algorithm was applied, a manual check was performed to ensure the accuracy of the impact boundaries. The first was to ensure the maximum queue lengths and end times were representative of the manually verified maximum queue length and end times taken from the incident heat maps. The second involved addressing impossible shifts, or when the data-driven start time shifts to earlier in time upstream of the incident.

To ensure the accuracy of the end times and maximum queue lengths, the following criteria was used to check for inaccurate data: 1. if maximum queue lengths varied more than 0.5 miles from the visually verified maximum queue length or 2. if the end time varied greater than 30 minutes compared to the visually verified end time. If maximum queue lengths or end times did not meet the criteria, a correction was needed.

As stated earlier, it was noticed that some of the impact boundaries were exhibiting impossible behavior. The chief concern was the realization that as the impact boundary was developed upstream, the left bound would shift back in time, which is an impossible movement because queues cannot propagate in negative time; it violates the assumption that traffic queues forward in time. An example of this phenomenon is shown in Figure 6 which is for an incident occurring on 05/14/2014 at 11:58 AM on mile point 17.3 of I64E. Mile point 17.3 rests on top of the congestion bounds, but as the queue moves upstream, a shift can be seen in the top-left corner which is impossible.

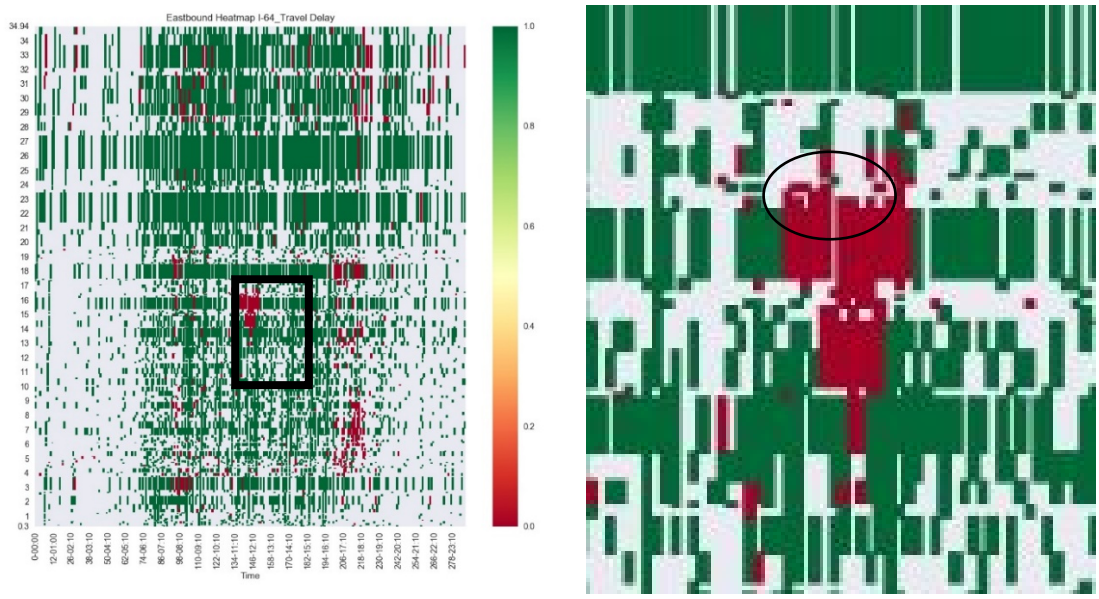


Figure 6 Example of Left Shift

Addressing this required developing another algorithm which searched downstream of the impossibly shifted cells for uncongested cells. If cells are found along any link up to the top of the impact zone, or where the incident occurred, then the start time will be shifted to 5 minutes past the temporal location of the downstream uncongested cell. After the implementation of the algorithm, the results were inspected for any abnormalities. It was found that in some cases the algorithm shifted start times too far forward. In these cases, the start times were manually adjusted to correct the issue. In the case where no shift occurred, it was assumed there was not enough data to perform the shift. Therefore, these incidents were left unaltered.

After the identification of the impact zone, queue length and duration can be calculated. These are not only helpful in describing the spatial and temporal characteristics of the impact zone, but also can potentially be used to model VHD as described in chapter 4. The list of values for queue length and duration generated for this analysis are as follows:

- Average Duration is the average of the difference between the data-derived end and start times.
- Maximum Duration is the identified longest duration between the data-driven end time and the TRIMARC-reported start time of an incident.
- Site Duration is the difference between the data-driven end time and the TRIMARC-reported start time at the incident site (or probe GPS link in which the incident was reported to occur).
- Average Queue Length is the average of the aggregate queue length for each time period in the impact zone.
- Maximum Queue Length is the largest identified aggregate queue length in the impact zone.
- Zone Area is the product of the Average Queue Length and the Average Duration.

Once the impact boundaries are defined and the queue length and duration metrics are calculated, they are used to calculate VHD, which is then used alongside the available categorical, continuous, and spatiotemporal data to develop linear regression models.

3.4.3 Calculating VHD

The final step after generating the impact zone is to calculate VHD from a combination of the impact zone, speed data, and volume data for each incident. However, one challenge still exists; each incident has missing data within the impacted boundaries which must be addressed. Since VHD is calculated using the sum of the delays for each cell within the impact zone, not addressing the missing data cells will cause a systematic underestimation of delay for each incident. Therefore, data imputation is required to properly estimate delay.

A review of literatures on imputing missing data was conducted to investigate the available options. Methods used to impute traffic data to increase sample sizes include multiple linear regression, adaptive smoothing method, k-nearest neighbor method, cokriging via GIS, and local least squares (*43-47*). The methods presented are sophisticated and can potentially impute speed data with great accuracy. However, Smith et al. states that using methods such as the "weighted average of surrounding detectors…" produce results that are fast and "generally reliable" (*48*). Given the speed data are disaggregate, the conditions influencing traffic and the impact itself are unlikely to change. Therefore, an average of the data links within a time period within an impact zone is used for this study.

Data are imputed by using the speed ratio calculated by dividing the congested speed and the background speed. Then, the ratios are averaged for all links within a specific time period of the incident. Finally, new speeds based on these imputed ratios are calculated. The succeeding example illustrates the process of imputing speed ratios and calculating new speeds.

Say an incident occurred on mile marker 0.8 between 7:30-8:15 AM along a stretch of freeway traveling in the cardinal direction. The historic average background speed is assumed to be 60 mph for the entire segment. The recorded speeds for the day of the incident as well as the speed differences for each cell are aggregated to a 15-minute level for the purposes of this example and can be seen in Table 3.

Table 3 Speeds and Speed Differences for Hypothetical Incident

Legend	Congested	Uncongested	Missing	Imputed

MM	Recorded Speeds			
1	50	50	50	55
0.9	50	55	50	55
0.8	20			20
0.7	20			
0.6	15	25		45
0.5	15	20	20	50
0.4	20	10		60
0.3	20	10	20	60
0.2	50	10	40	60
0.1	55	40	60	60
0	60	60	60	60
Time	7:30	7:45	8:00	8:15

MM	Speed Difference			
1	10	10	10	5
0.9	10	5	10	5
0.8	40			40
0.7	40			
0.6	45	35		15
0.5	45	40	40	10
0.4	40	50		0
0.3	40	50	40	0
0.2	10	50	20	0
0.1	5	20	0	0
0	0	0	0	0
Time	7:30	7:45	8:00	8:15

Using these data points, speed ratios for each cell must be calculated. This is shown in Table 4.

Table 4 Speed Ratios for Each Cell

Legend	Congested	Uncongested	Missing	Imputed

MM	Speed Ratio			
1	0.83	0.83	0.83	0.92
0.9	0.83	0.92	0.83	0.92
0.8	0.33			0.33
0.7	0.33			
0.6	0.25	0.42		0.75
0.5	0.25	0.33	0.33	0.83
0.4	0.33	0.17		1.00
0.3	0.33	0.17	0.33	1.00
0.2	0.83	0.17	0.67	1.00
0.1	0.92	0.67	1.00	1.00
0	1.00	1.00	1.00	1.00
Time	7:30	7:45	8:00	8:15

Notice the white cells which indicate missing data. These must be filled using the data within a specific time period for every link missing data by taking the average speed ratio for all of the segments within a specific time period. The reason for averaging within specified time periods is because traffic is subject to similar conditions during the same time period; averaging within the same roadway link over different time periods may result in averaging speeds during periods when different traffic conditions exist. Using the imputed ratios, new congested speeds can be calculated. The results of these calculations are shown in Table 5.

Table 5 Speed Ratios after Imputation

Legend	Congested	Uncongested	Missing	Imputed

MM	Speed Ratio Imputed			
1	0.83	0.83	0.83	0.92
0.9	0.83	0.92	0.83	0.92
0.8	0.33	0.32	0.44	0.33
0.7	0.33	0.32	0.44	0.54
0.6	0.25	0.42	0.44	0.75
0.5	0.25	0.33	0.33	0.83
0.4	0.33	0.17	0.44	1.00
0.3	0.33	0.17	0.33	1.00
0.2	0.83	0.17	0.67	1.00
0.1	0.92	0.67	1.00	1.00
0	1.00	1.00	1.00	1.00
Time	7:30	7:45	8:00	8:15
Average		0.32	0.44	0.54

MM	New Congested Speeds			
1	50	50	50	55
0.9	50	55	50	55
0.8	20	19	27	20
0.7	20	19	27	33
0.6	15	25	27	45
0.5	15	20	20	50
0.4	20	10	27	60
0.3	20	10	20	60
0.2	50	10	40	60
0.1	55	40	60	60
0	60	60	60	60
Time	7:30	7:45	8:00	8:15

For visual representation of the actual results of the imputation, Figure 7 shows the same incident as seen in Figure 4 but with imputed speed data.

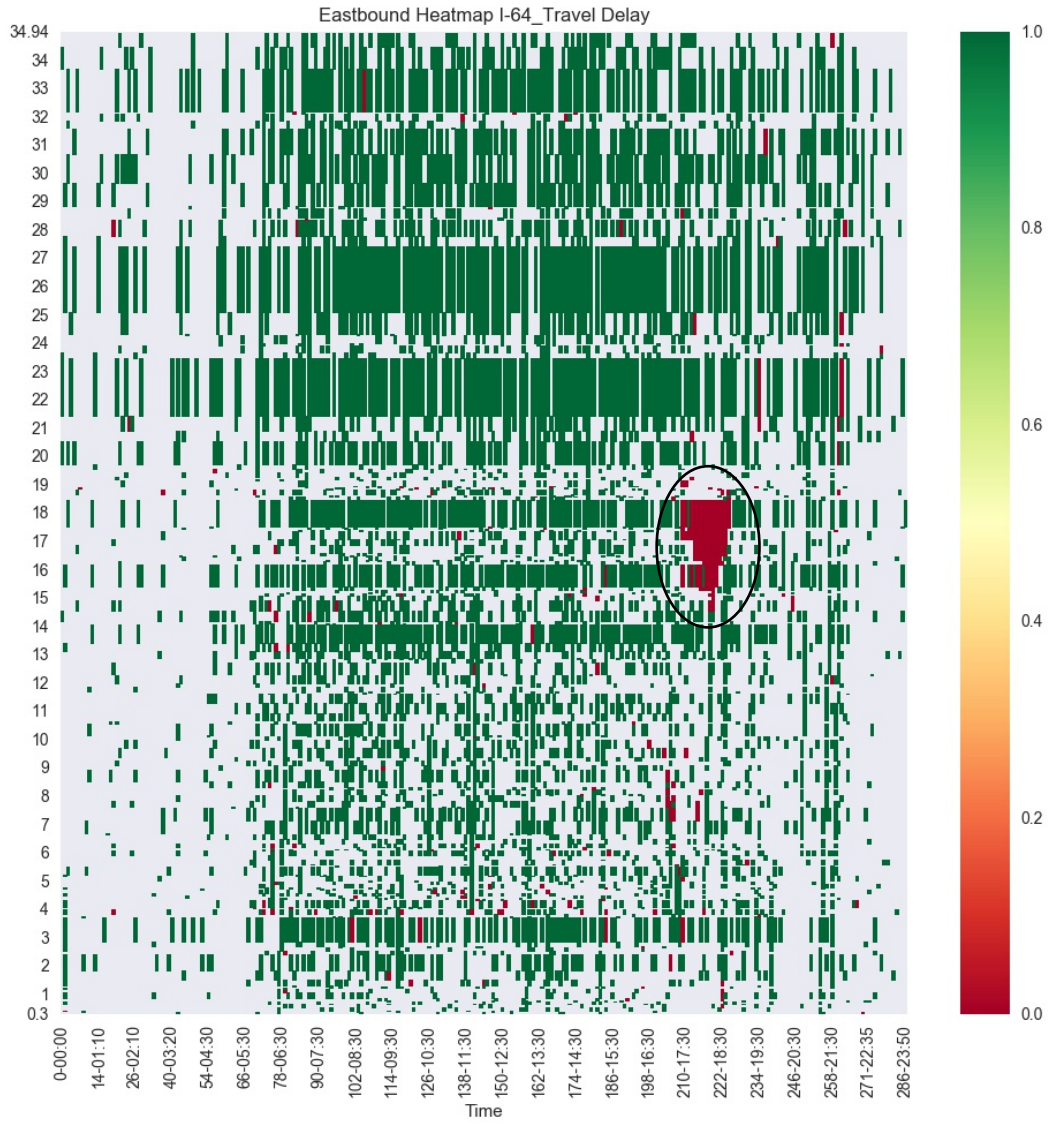


Figure 7 01-23-2013 Incident After Imputation

Once the dimensions of each incident's impact zone are defined and the proper imputation is performed, the cells within the boundary are used to calculate VHD. This can be derived by using the speed data that exhibit lower speeds, both recorded and imputed as discussed in section 3.4.2, versus the background speeds to generate a travel delay. Then, volume data can then be applied to calculate the VHD for each cell and then aggregated to get the total VHD for the incident.

Hourly volumes are derived from the KYTC traffic volume data as described in section 3.3. To generate hourly volumes, since hourly data were not available for the segment, hourly factors for functional class 11 highways in the state of Kentucky, which are urban

freeways, were used to assume the hourly volume pattern. After the hourly volumes were calculated, it is assumed that the directional split of the roadway is 50%, so the volumes are divided by two. Finally, 5-minute volumes must be calculated from the directional hourly volumes by dividing by 12 5-min time periods.

The calculation for travel delay is as follows.

$$TD_{ij} = \frac{L_{ij}}{C_{ij}} - \frac{L_{ij}}{F_{ij}}$$

Where:

TD_{ij} is the travel delay calculated for each link i and for each time j

L_{ij} is the probe GPS link length

C_{ij} is the congested speed for each link i and for each time j

F_{ij} is the background speed for each link i and for each time j

VHD for the incident is calculated as follows.

VHD = $\sum \max\{TD_{ij}V_{ij}, 0\}$

Where:

VHD = Vehicle Hours of Delay

TD_{ij} is the travel delay calculated for each link i and for each time j

V_{ij} is the volume for each link i and for each time j

In some impact zones, cells exhibiting speeds greater than the background speed exist as explained in section 3.4.2. The formula for VHD reflects this need to consider negative delays as zero.

This chapter overviewed the sets of data and described the process in which the impact zones were defined and VHD was calculated. The next chapter will overview the model selection process and discuss the outcomes and different considerations when comparing the process to work zone analysis.

Chapter 4 Modeling the Impact Zone and Delay

This chapter focuses on modeling VHD as well as the impact zone dimensions. The purpose of modeling these data is to explore the relationships between the variables that describe incidents and the delay associated with the incidents. Also, it is important to infer if variables will be collinear with each other.

The variables used for the analysis are the impact zone dimensions previously defined, AADT, hourly volume, weather, the number of blocked lanes (shortened as blocked lanes for this work), time period, and injury. Table 6 showcases details on each factor.

Table 6 Variables Selected

Factors Considered	Definition
Impact Zone Dimensions	Average Length, Max Length, Average Duration, Maximum Duration, Site Duration, and Zone Area
AADT	Average Annual Daily Traffic
Hourly Volume	Bi-Directional Hourly Volume
Weather	Either Clear or with Precipitation
Number of Blocked Lanes	One Lane Blocked or More than One Lane Blocked
Time Period	Either Peak (6AM-9AM or 3PM-6PM) or Non-Peak
Injury	Either Injury or Non-Injury

4.1 Data Exploration

Before diving into deeper modeling and analysis, it is important that the relationships between the variables be explored. The reason being is because understanding how the factors relate to each other will assist in interpreting the results of the models for the queue length, duration, and VHD. To begin, a high-level look at the continuous variables in the analysis will be conducted using Pearson correlation coefficients. Next, the relationship between the continuous and categorical variables will be studied using ANOVA analysis. Finally, the findings from the Correlation and ANOVA analyses will be explored deeper by using Welch's t-test to examine how the categorical variables interact with each other when comparing the means of VHD as well as the other continuous variables.

4.1.1 Correlation

In this step, it is important to see how the continuous variables correlate with each other before exploring any deeper. It is important to view the strength of correlation to discover which factors are strongly related to each other. For example, what factors are strongly correlated to VHD? Average length? Site duration? A correlation plot is provided in Table 7.

Table 7 Correlation Plot

	VHD	*AADT*	*Hourly Volume*	*Avg Length*	*Max Length*	*Avg Duration*	*Site Duration*	*Max Duration*	*Zone Area*
VHD	1.00								
AADT	-0.05	1.00							
Hourly Volume	-0.04	0.80	1.00						
Avg Length	0.76	-0.05	-0.03	1.00					
Max Length	0.75	-0.04	-0.01	0.97	1.00				
Avg Duration	0.74	-0.03	0.00	0.53	0.49	1.00			
Site Duration	0.70	-0.03	0.03	0.48	0.47	0.90	1.00		
Max Duration	0.73	-0.05	0.00	0.54	0.57	0.92	0.89	1.00	
Zone Area	0.91	-0.09	-0.07	0.85	0.80	0.80	0.71	0.75	1.00

The variables correlated with VHD are average length, maximum length, average duration, maximum duration, incident duration, and zone area. Zone area has the highest direct correlation and site duration has the lowest direct correlation with VHD, respectively. Hourly volume and AADT are not correlated with VHD or the impact zone dimensions based on this study area which is surprising because one would assume that increased volume would cause traffic breakdowns. One potential reason is that the data used for this analysis are derived from AADT for the course of a year as well as hourly factors derived for functional class 11 highways (Interstates). Therefore, an assumption is being made for the number of vehicles that are traversing each impact zone which is most likely not representative of the actual number of vehicles traversing the impact zone. If further investigation is required, more detailed volume data would be required. However, in this case, hourly volume or AADT will not enter the modeling process for VHD described in section 4.2.

It was observed that the duration variables are correlated with the length variables. This is logical given a longer duration of an incident will more than likely result in longer queues. Both duration and length variables are directly correlated with zone area since zone area is a derivative of duration and length and with increased length or duration comes a potentially larger zone area. Hourly volume and AADT are not correlated with any of the spatiotemporal variables. It is expected that the reasons presented in the previous paragraph are also a significant factor in the lack of relationship between volume and the impact zone. Therefore, hourly volume and AADT will not be used to model the impact zone dimensions.

This section only serves as a high-level analysis of the continuous variables. When modeling these variables, a more detailed picture of the relationship between the variables can be realized. However, before that can be conducted, the relationship between the categorical variables must be evaluated. In the next section, ANOVA will be conducted on the categorical variables to detect their level of significance as well as any signs of interaction between the variables.

4.1.2 ANOVA

In this section, ANOVA is conducted to determine whether a categorical variable or a pair of categorical variables such as injury and weather have a significant effect in explaining continuous variables such as VHD, duration, and queue length. ANOVA can also determine if the relationships are linear or non-linear as well as detect if categorical variables interact with each other.

ANOVA uses a difference in variance to test whether two categorical variables or pairs of categorical variables are significant in explaining a selected response variable which, in this case, are VHD, average length, maximum length, site duration, and zone area.

The ANOVA procedure is based in using a linear model as shown below.

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i$$

Where:

y is the response variable

β_i is the coefficients

x_1 is the value of the first categorical variable (0 or 1)

x_2 is the value of the second categorical variable (0 or 1)

x_3 is the value of the interaction term between the x_1 and x_2 variable (0 or 1)

ε_i is the error term

The goal is to test if there is enough evidence to suggest that any of the categorical variables explain error in the response variable (or that the slopes are not zero). The null hypothesis for One-Way ANOVA and Two-Way ANOVA are as follows.

One-Way ANOVA

$$H_0 = \beta_1 = 0$$

Two-Way ANOVA

$$H_0 = \beta_1 = \beta_2 = \beta_3 = 0$$

To test this hypothesis, the F-Statistic for both the overall model and each coefficient is used. A significant F-Statistic can either show the overall model has at least one value that explains the variance or, in the case of a coefficient F-Statistics, at least one group within the variable when controlling for all other groups is significant. The resulting p-value from the F-Statistic will be used to determine significance for this work.

Since the experimental designs in this work are unbalanced, an unbalanced ANOVA approach will be adopted. This means that the F-Statistic and the p-values will be gathered from Type 3 Sum of Squares error, which "test a function of the underlying parameters that is independent of the number of observations per treatment combination" (*49*).

For this analysis, the continuous variables selected are VHD, average length, maximum length, site duration, and zone area. Average duration and maximum duration were not selected because site duration is more representative of the recorded incident duration in the TRIMARC incident log, or the difference between the reported start and end times of the incident. Maximum duration can be tricky to collect because the end time of the incident may not occur at the mile marker the incident occurred. Also, average duration is accounted for in the zone area metric.

Upon investigation of the preliminary models, it was seen that the residual distributions are right-skewed which indicates heteroscedasticity. According to Tastan, log-transformations can be used to treat heteroscedasticity and ensure the model meets the assumptions of normality (*50*). Based on this knowledge, it was determined in analysis that the log-transformed versions of the continuous variables should be used because the relationship between the continuous variables and categorical variables are not normal as shown in Figure 8, which shows the residuals of modeling VHD as a function of time period.

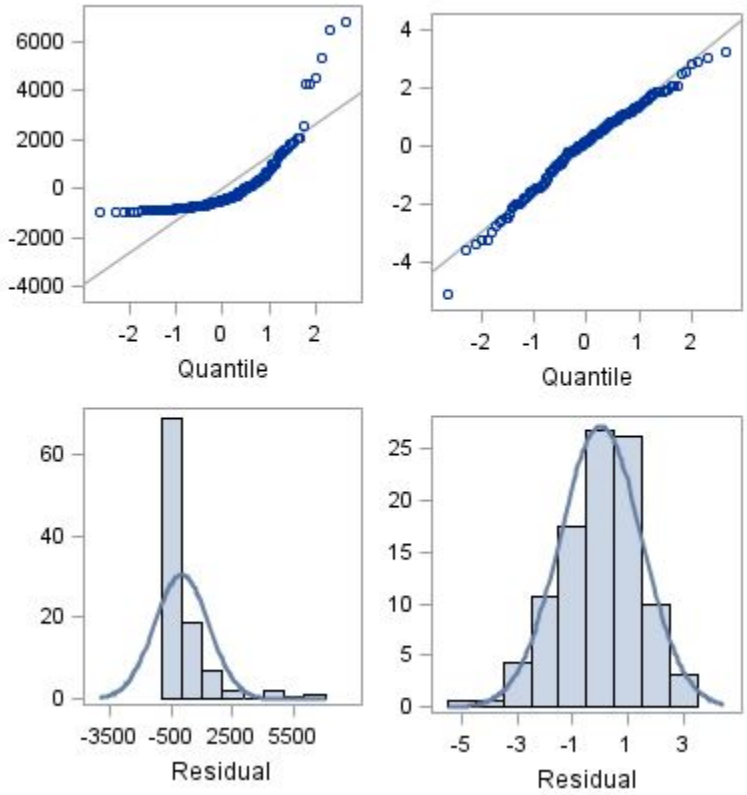


Figure 8 Time Period Model Residuals Before and After Transformation of VHD

As can be seen, before transformation, the QQ plot (top-left plot) is not a straight line and the residual histogram (bottom-left plot) is skewed right. Therefore, before transformation, the assumptions of a linear model are likely not being met. With transformation, the QQ Plot (top-right plot) is linear and the histogram (bottom-right plot) both look more normally distributed. This is also true for all other continuous variables modeled. Figure 9 shows a histogram of VHD and the entire catalog of histograms can be found in Appendix A1.

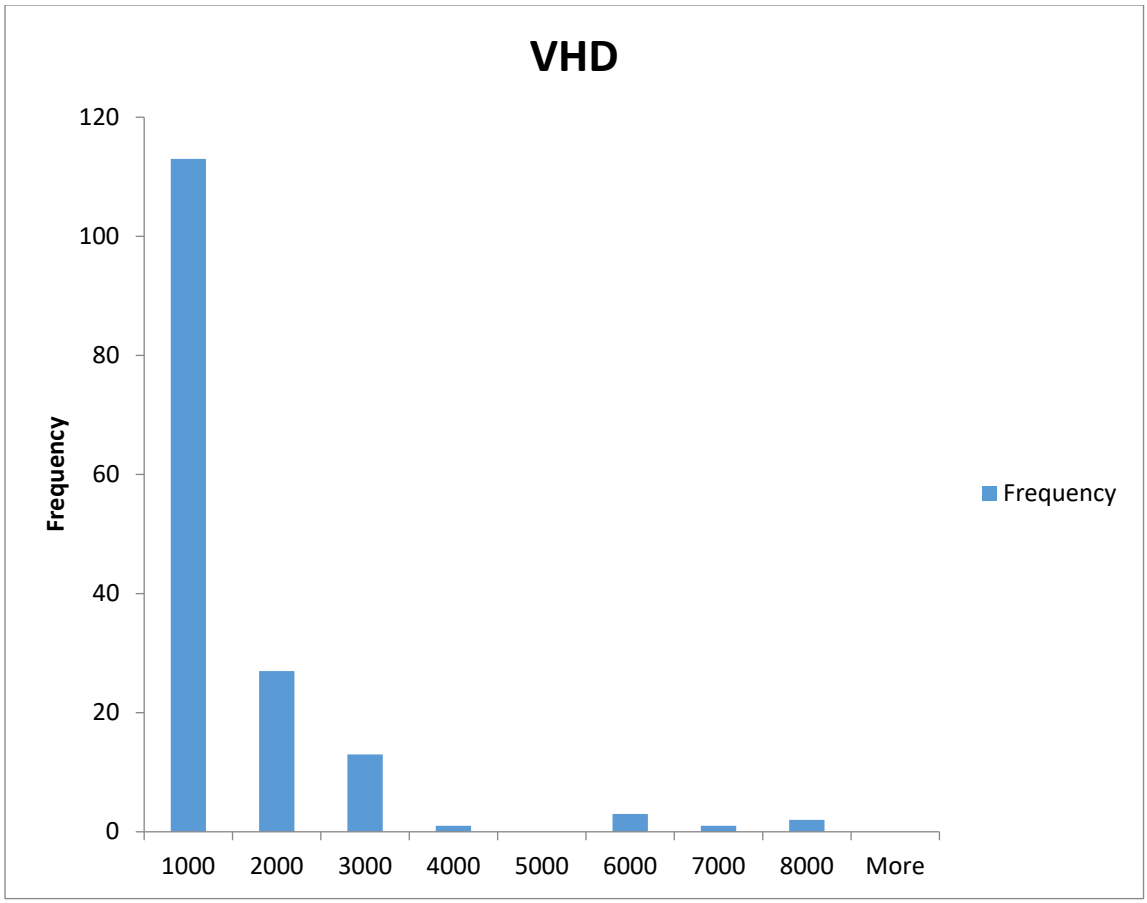


Figure 9 Histogram of VHD

Table 8, Table 9, Table 10, Table 11, and Table 12 show the ANOVA regression F-Statistics and coefficient p-values for VHD, average length, maximum length, site duration, and zone area, respectively.

Table 8 ANOVA Analysis for Variables Modeling for VHD

Note: p-vales in order, respectively, to row names	F-Stat	p-value 1	p-value 2	p-value interaction
Independent				
Weather	0.95	0.95		
Blocked Lanes	0.08	0.08		
Time Period	0.20	0.20		
Injury	0.04	0.04		
Time Period and Injury	0.03	0.06	0.07	0.38
Blocked Lanes and Injury	0.13	0.25	0.17	0.96
Blocked Lanes and Time Period	0.05	0.02	0.29	0.20
Weather and Time Period	0.62	0.87	0.42	0.74
Weather and Blocked Lanes	0.30	0.60	0.34	0.46
Weather and Injury	0.22	0.97	0.09	0.83

For the ANOVA with respect to VHD, there is a significant interaction on VHD when considering blocked lanes and injury as indicated by the small p-value. For time period and injury, time period is significant in explaining VHD when injury is held constant, and injury is significant in explaining VHD when time period is held constant. For blocked lanes and time period, blocked lanes is significant in explaining VHD when time period is held constant.

Table 9 ANOVA Analysis for Variables Modeling for Average Length

Note: p-vales in order, respectively, to row names	F-Stat	p-value 1	p-value 2	p-value interaction
Independent				
Weather	0.91	0.91		
Blocked Lanes	0.93	0.93		
Time Period	0.04	0.04		
Injury	0.46	0.46		
Time Period and Injury	0.10	0.02	0.39	0.56
Blocked Lanes and Injury	0.88	0.72	0.45	0.87
Blocked Lanes and Time Period	0.07	0.40	0.20	0.11
Weather and Time Period	0.14	0.63	0.32	0.29
Weather and Blocked Lanes	0.97	0.75	0.73	0.65
Weather and Injury	0.88	0.85	0.74	0.74

For the ANOVA with respect to the average length, there is a significant interaction on average length when considering time period as indicated by the small p-value. For time period and injury, time period is significant in explaining average length when injury is held constant. For blocked lanes and time period, although the overall F-Statistic reveals an effect present, neither blocked lanes or time period are significant. However, the interaction parameter p value is low despite not being significant to a 90% confidence level.

Table 10 ANOVA Analysis for Variables Modeling for Maximum Length

Note: p-vales in order, respectively, to row names	F-Stat	p-value 1	p-value 2	p-value interaction
Independent				
Weather	0.99	0.99		
Blocked Lanes	0.79	0.79		
Time Period	0.04	0.04		
Injury	0.45	0.45		
Time Period and Injury	0.11	0.02	0.35	0.65
Blocked Lanes and Injury	0.89	0.99	0.47	0.81
Blocked Lanes and Time Period	0.06	0.23	0.21	0.09
Weather and Time Period	0.15	0.55	0.36	0.26
Weather and Blocked Lanes	0.95	0.77	0.90	0.58
Weather and Injury	0.88	0.92	0.73	0.74

For the ANOVA with respect to the maximum length, there is a significant interaction on maximum length when considering time period as indicated by the small p-value. For time period and injury, the F-Statistic does not indicate significance at the 90% confidence level. However, the p-value of the time period coefficient is low enough to indicate further analysis into the relationship especially given time period when injury is held constant is significant with respect to average length. For blocked lanes and time period, there is a significant overall effect based on the F-statistic but neither blocked lanes or time period are significant. However, the interaction parameter p value is significantly low to the 90% confidence level.

Table 11 ANOVA Analysis for Variables Modeling for Site Duration

Note: p-vales in order, respectively, to row names	F-Stat	p-value 1	p-value 2	p-value interaction
Independent				
Weather	0.82	0.82		
Blocked Lanes	0.15	0.15		
Time Period	0.89	0.89		
Injury	0.23	0.23		
Time Period and Injury	0.63	0.88	0.41	0.61
Blocked Lanes and Injury	0.31	0.34	0.24	0.35
Blocked Lanes and Time Period	0.30	0.08	0.72	0.21
Weather and Time Period	0.53	0.68	0.42	0.14
Weather and Blocked Lanes	0.54	0.99	0.33	0.79
Weather and Injury	0.66	0.90	0.51	0.74

For the ANOVA with respect to site duration, the blocked lanes and time period analysis indicate a significant effect of blocked lanes when time period is held constant despite an insignificant overall F-Statistic.

Table 12 ANOVA Analysis for Variables Modeling for Zone Area

Note: p-vales in order, respectively, to row names	F-Stat	p-value 1	p-value 2	p-value interaction
Independent				
Weather	0.93	0.93		
Blocked Lanes	0.63	0.63		
Time Period	0.20	0.20		
Injury	0.14	0.14		
Time Period and Injury	0.14	0.09	0.19	0.49
Blocked Lanes and Injury	0.49	0.96	0.16	0.64
Blocked Lanes and Time Period	0.28	0.26	0.44	0.21
Weather and Time Period	0.58	0.80	0.53	0.55
Weather and Blocked Lanes	0.93	0.74	0.91	0.65
Weather and Injury	0.52	0.87	0.34	0.83

For the ANOVA with respect to zone area, the time period and injury analysis indicate a significant effect of time period when injury is held constant despite an insignificant overall F-Statistic.

The highlighted values in the previous tables denote significance or values worth noting. The summarization of the findings can be found below.

- Injury is significant in explaining VHD by itself.
- Blocked lanes is significant in explaining VHD by itself.
- Time period is not significant in explaining VHD by itself but is when Injury is held constant.
- Time period is significant in explaining both average length and maximum length by itself.
- Blocked lanes and injury are not significant when explaining VHD together.
- Weather removed from analysis due to lack of significance to any continuous variable.

In this analysis, ANOVA is used to test whether there are significant factors explaining the continuous variables. Based on this analysis, the most important interactions to consider are the relationship between time period and injury, blocked lanes and injury, and blocked lanes and time period. Using the results of this section, a detailed analysis using t-tests will be used to detect exactly which groups differ from each other.

4.1.3 Detailed Investigation: T-Tests

In the previous section, ANOVA is used to determine which factors are significant in describing VHD and the impact zone boundaries. However, ANOVA cannot show exactly which pairs of categorical variables are significantly different. For example, in the previous section, time period was found to be significant when injury is held constant with respect to VHD. It is not known if time period is significant when only considering injury crashes or non-injury crashes. Therefore, t-tests shall be used to explore these factors in more detail.

In this analysis, the relationship between time period and injury, blocked lanes and injury, and blocked lanes and time period will be explored with respect to the continuous variables. Given the uneven sample sizes found in this study, Welch's t-test will be used for this analysis. Welch's t-test is identical in interpretation to the Student's t-test, so results will be familiar. In this analysis, the two-tailed test is used verses the one-tailed test because it was not desired to assume how the effect will change the effect of a given continuous variable. Results in this analysis will be considered significant if the p-value is below 0.1. Table 13 shows the results of the t-tests with respect to the continuous variables.

Table 13 Summary of T-Tests

Response Variable	Pair	p-value
Log VHD	Injury Peak vs No Injury Peak	0.00
	Injury and Peak vs Injury and Non-Peak	0.10
	1 Blocked Lane and Peak vs 1 Blocked Lane and Non-Peak	0.10
	1 Blocked Lane vs More than 1 Blocked Lane	0.05
	1 Blocked Lane and Non-Peak vs More than 1 Blocked Lane and Non-Peak	0.03
	Injury vs No Injury	0.04
Log Average Length	Injury and Peak vs Injury and Non-Peak	0.05
	Peak vs Non-Peak	0.04
	1 Blocked Lane and Peak vs 1 Blocked Lane and Non-Peak	0.02
Log Zone Area	Injury Peak vs No Injury Peak	0.03
Log Maximum Length	Injury and Peak vs Injury and Non-Peak	0.06
	Peak vs Non-Peak	0.05
	1 Blocked Lane and Peak vs 1 Blocked Lane and Non-Peak	0.02
	1 Blocked Lane and Non-Peak vs More than 1 Blocked Lane and Non-Peak	0.06
Log Site Duration	1 Blocked Lane and Non-Peak vs More than 1 Blocked Lane and Non-Peak	0.06

This analysis revealed that, for the combination of time period and injury and blocked lanes and time period, the factors are dependent on each other in explaining VHD and the impact zone dimensions. For example, when considering zone area, injury is significant for peak period incidents but not for non-peak incidents. However, there was no such relationship when considering blocked lanes and injury.

Ultimately, the purpose of these investigations is to infer, when performing modeling, how the categorical and continuous variables will interact with each other and to determine if collinearity will be an issue. Based on the significant relationships between multiple categorical variables to VHD and the impact zone dimensions, this is likely problematic. To combat collinearity, techniques such as stepwise regression will be used to develop the most parsimonious models. For example, if a model for VHD is developed using average length as an explanatory variable, categorical variables significant in explaining average length will be removed leaving those not related to explaining average length yet explaining VHD. For this analysis, stepwise selection using the Schwarz Bayesian Criterion (SBC) will be implemented. The next section overviews this process and issues models for both the spatiotemporal boundaries as well as VHD.

4.2 Modeling

In the previous section, Welch's t-tests were used to determine the relationships categorical variables or pairs of categorical variables had with respect to the continuous variables associated with describing incident delay and the impact zone.

The chosen method for performing models in this analysis is multiple linear regression. The general form for Multiple Linear Regression is as follows

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon_i$$

For this analysis, stepwise selection is used to determine parsimonious models based on the values being modeled and their significance in explaining the model. In SAS, the SBC is used by default to select the optimum model based on the input data.

The next section discusses modeling based on the selected dimensions of the impact zone.

4.2.1 Modeling the Impact Zone

In this section, it is of interest to investigate possible models for queue length using site duration. Understanding how the duration of an incident affects queue length for a given area can reveal potential operational issues that may arise, such as queue spillback onto side roads or through interchanges. Also, site duration is easier to collect data-wise because only knowledge of the incident area is required verses monitoring upstream to detect the back of the queue.

Models for maximum length and average length are developed in this section because both can illustrate various magnitudes of disruption when considering spillback. For

example, if an interchange is found within the average queue length, the interchange may be affected for much longer versus if it were found within the maximum queue length but outside of the average queue length.

The models developed prior to stepwise selection are as follows:

$$\ln(Max\ Length) = \beta_0 + \beta_1 \ln(site\ duration) + \beta_2(time\ period) + \beta_3(injury) + \beta_4(blocked\ lanes) + \beta_5(interaction\ time\ period\ and\ injury) + \beta_6(interaction\ blocked\ lanes\ and\ time\ period)$$

$$\ln(Avg\ Length) = \beta_0 + \beta_1 \ln(site\ duration) + \beta_2(time\ period) + \beta_3(injury) + \beta_4(blocked\ lanes) + \beta_5(interaction\ time\ period\ and\ injury) + \beta_6(interaction\ blocked\ lanes\ and\ time\ period)$$

Only one continuous variable was modeled at a time given the strong correlations between the spatiotemporal variables as seen in section 4.1.1.

Performing stepwise selection, the resulting model for maximum length is given below:

$$\ln(Max\ Length) = -1.06 + 0.487 \ln(Site\ Duration) - 0.292 time\ period$$

When performing an exponential transformation of the model, the following equation is given:

$$e^{\ln(Max\ Length)} = e^{-1.06 + 0.487 \ln(Site\ Duration) - 0.292 time\ period}$$ or

$$Max\ Length = 0.346 e^{-0.292 time\ period} Site\ Duration^{0.487}$$

The model states that when the incident occurs during the off-peak period, the maximum length should be approximately $e^{-0.292}$ or 0.75 times smaller than incidents during the peak period. For every minute increase in site duration, the site duration multiplier for maximum length will increase according to a power function raised to the 0.487 power. Since the queue length should decrease during the night time and increase with increasing site duration, the model makes sense.

The transformed equation for average length is presented below:

$$Avg\ Length = 0.251 e^{-0.286 time\ period} Site\ Duration^{0.468}$$

The model states that when the incident occurs during the off-peak period, the maximum length should be approximately $e^{-0.286}$ or 0.75 times smaller than incidents during the peak period. For every minute increase in site duration, the site duration multiplier for maximum length will increase according to a power function raised to the 0.468 power. Also, the coefficient of 0.251 is less than the coefficient for maximum queue length,

which is 0.346 which makes sense given maximum length should be larger than average length.

In summary, it is found that the models for maximum length and average length are better explained when time period is included in the model. This is consistent with the significance found in the ANOVA analysis in section 4.1.2. The R^2 values for maximum length and the average length are 0.253 and 0.246, respectively. Therefore, based on this study area, site duration does explain some variance in maximum length and average length, but results should be taken with caution.

4.2.2 Modeling VHD

In this section, models will be developed for VHD. It is not cheap to constantly monitor the impact area and develop VHD based on collected data. Building models using what is known about the impact zone and conditions of the incident can allow practitioners to both reduce the data collection required and effectively estimate the impact of incidents for traffic impact analyses..

The four VHD models before stepwise selection are presented below:

$$\ln(VHD) = \beta_0 + \beta_1 \ln(site\ duration) + \beta_2(time\ period) + \beta_3(injury) + \beta_4(blocked\ lanes) + \beta_5(interaction\ time\ period\ and\ injury) + \beta_6(interaction\ blocked\ lanes\ and\ time\ period)$$

$$\ln(VHD) = \beta_0 + \beta_1 \ln(avg\ length) + \beta_2(time\ period) + \beta_3(injury) + \beta_4(blocked\ lanes) + \beta_5(interaction\ time\ period\ and\ injury) + \beta_6(interaction\ blocked\ lanes\ and\ time\ period)$$

$$\ln(VHD) = \beta_0 + \beta_1 \ln(max\ length) + \beta_2(time\ period) + \beta_3(injury) + \beta_4(blocked\ lanes) + \beta_5(interaction\ time\ period\ and\ injury) + \beta_6(interaction\ blocked\ lanes\ and\ time\ period)$$

$$\ln(VHD) = \beta_0 + \beta_1 \ln(zone\ area) + +\beta_2(time\ period) + \beta_3(injury) + \beta_4(blocked\ lanes) + \beta_5(interaction\ time\ period\ and\ injury) + \beta_6(interaction\ blocked\ lanes\ and\ time\ period)$$

Where the coefficients are previously defined in section 4.2.1.

The explanatory variables are selected using the stepwise regression procedure and transformed to exponential form as described earlier. The four resulting models are shown below:

$VHD = 2.11 site\ duration^{1.29}$ $R^2 = 0.407$

The model states that for every minute increase in site duration, the site duration multiplier for VHD will increase according to a power function raised to the 1.29 power.

$VHD = 166e^{0.527 blocked\ lanes} avg\ length^{1.71}$ $R^2 = 0.713$

The model states that for every mile increase in average length, the average length multiplier for VHD will increase according to a power function raised to the 1.71 power. Also, when more than one blocked lane is present, the VHD will be $e^{0.527}$ or approximately 1.69 times larger.

$$VHD = 90.9e^{0.442 blocked\ lanes} max\ length^{1.64} \qquad R^2 = 0.726$$

The model states that for every mile increase in maximum length, the maximum length multiplier for VHD will increase according to a power function raised to the 1.64 power. Also, when more than one blocked lane is present, the VHD will be $e^{0.442}$ or approximately 1.56 times larger.

$$VHD = 3.67e^{0.384 blocked\ lanes} zone\ area^{1.11} \qquad R^2 = 0.773$$

The model states that for every mile-minute increase in zone area, the zone area multiplier for VHD will increase according to a power function raised to the 1.11 power. Also, when more than one blocked lane is present, the VHD will be $e^{0.384}$ or approximately 1.47 times larger.

In reviewing the models, it is seen that the model using zone area explains VHD the best with an R^2 of 0.773. The model using site duration is the worst in explaining VHD with an R^2 of 0.407. The models using maximum length and average length perform admirably well in explaining VHD with R^2 values above 0.7. It is also interesting to see blocked lanes result in significance in three of the four models even though injury, according to the ANOVA analysis, appears as the most significant factor in explaining VHD. One potential reason is because injury also explains the queue length variables, which, in this case, causes collinearity.

Although zone area produces the best results, it is also the most data-intensive. As stated before calculating zone area requires knowledge of the average duration and the average length of the incident. Site duration is the lowest performer but the easiest to collect data for since it only requires knowledge of the duration at the crash site.

This section developed models for the VHD, average length, and maximum length. The next section will discuss the results. Details of all model outputs can be found in Appendix A2.

4.3 Discussion

First and foremost, the results of this study are specific to the roadway in question and are most likely not transferrable to other roadways. If results are desired for other roadways, then the process must be repeated for the roadway in question to ensure accuracy. Also, this methodology is applied to a freeway. It is not recommended this methodology be applied to arterials without further research.

In the modeling sections of this report, it was found that the probe GPS data-driven site duration is the worst performer in explaining VHD in comparison to the other models proposed. However, it can be argued that site duration is an easier metric to collect

because it only requires knowledge of the incident site instead of the entire impact zone. TRIMARC currently collects similar data within the incident log since most incidents identify a start and an end time. Given that information, an investigation in the comparability of the TRIMARC incident log was conducted against the data-driven site duration using Welch's t-test. This is to determine if TRIMARC's current data collection practices are sufficient to use as the site duration versus using the data-driven site duration as found in the probe GPS dataset. It was found that there is enough statistical evidence to suggest that the means are not the same given the small p-value in the two-tail test. Also, given the mean of site duration is larger than the mean of the TRIMARC duration, it can be verified that, given the presented data, the TRIMARC durations underestimate the data-driven site duration. This means that TRIMARC is not capturing the full impact of the incident verses probe GPS speed data. The results of Welch's t-test are shown in Table 14.

Table 14 Welch's T-Test Between TRIMARC Duration and Site Duration

	TRIMARC Duration	*Site Duration*
Mean	55.3	70.7
Variance	1305	2079
Observations	155	155
Hypothesized Mean Difference	0.00	
df	293	
t Stat	-3.29	
P(T<=t) one-tail	0.00	
t Critical one-tail	1.65	
P(T<=t) two-tail	0.00	
t Critical two-tail	1.97	

Despite the results of the t-test, it would be naive to think all incident durations were underestimated by TRIMARC; it is entirely possible TRIMARC overestimated the impact of incidents. For the dataset used in Welch's t-test, TRIMARC is found to overestimate the impact of incidents 52 times and by as much as 70 minutes. To better understand how and when TRIMARC underestimates the data-driven incidents, an investigation using the reported TRIMARC incident lengths, the categorical factors describing the incidents, and the location of the incidents was conducted using Welch's t-test. It was found that TRIMARC under reported site duration worse during incidents reported to have lasted less than on hour according to TRIMARC versus incidents lasting more than 1 hour. It was also noticed that incidents with an injury and incidents occurring during the peak period underestimated data-driven delay. Also, it was noticed that the level of under reporting and over reporting could be different based on location.

Appendix A3 contains the details of the results of the t-tests for factors that showed significance.

This study also utilized a 20% speed threshold as described in section 3.4.2 to identify the congested cells and define the impact zone. Although this can successfully identify incident congestion assuming the TRIMARC incident log as well as the data are correct, it does not guarantee that the congestion identified is entirely related to the incident. Procedures such as Kazi's kalman filter have the potential to further isolate the effects of the incident (*18*).

It is also desirable to use metrics related to the incident type. For example, injury is a metric related to crashes. In analysis for work zones, work zone speed limit and work zone barrier type could be considered (*51*).

Chapter 5 Conclusion

5.1 Summary of Work

The goal of this work is to explore the relationships between the data describing incident conditions, the spatiotemporal features of the impact zone, and the delay associated with incidents to assist in congestion management and the determination of road user costs. The utilization of probe GPS speed data can describe the day-to-day conditions of traffic which yield more detail in explaining the impact and delay caused by incidents. Using this methodology, practitioners can analyze other roadway segments and potentially combine them to allow for network examination. For example, the methodology could be applied to both I-64 and I-65 in Louisville together or separately depending on needs. This work can also serve as a stepping stone into future research using different methods and different algorithms to utilize probe GPS speed data to measure and predict congestion for roadways.

The study began with an overview of the recent research in the field of non-recurrent congestion. Then, it transitioned into discussing the data collection and processing required to generate the metrics for an incident's impact zone. This process was carried out using a combination of Python-based algorithms as well as manual updates to ensure the quality of the collected data. Next, the modeling process explored the relationship between the calculated VHD, the TRIMARC recorded metrics such as the number of blocked lanes, and the spatiotemporal dimensions of the impact zone such as queue length and duration.

This work proposed six linear models: two of which model average length and maximum length and four that model VHD. They are as follows:

$$Max\ Length = 0.346e^{-0.292 time\ period} site\ duration^{0.487}$$

$$Avg\ Length = 0.251e^{-0.286 time\ period} site\ duration^{0.468}$$

$$VHD = 2.11 site\ duration^{1.29}$$

$$VHD = 166e^{0.527 blocked\ lanes} avg\ length^{1.71}$$

$$VHD = 90.9e^{0.442 blocked\ lanes} max\ length^{1.64}$$

$$VHD = 3.67e^{0.384 blocked\ lanes} zone\ area^{1.11}$$

5.2 Applications

The models presented in this analysis can be used in multiple applications in incident management. For example, the two models predicting queue length can be used in estimating queue spillback for incidents. This is valuable information because queue spillback, especially when affecting side streets and interchanges, can greatly impact the network. In a work zone context, if crews know the expected queue length of the incident given the work time allotted, measures can be enacted to reduce the likelihood of queue spillback onto side streets. For models predicting VHD, these are valuable in

estimating costs for incidents for use in after action reviews and incident management program reviews. Before and after studies on roadway treatments can be conducted using quickly estimated VHD in lieu of a direct calculation. These estimates can shed light on the potential cost savings of roadway treatments assuming the treatment reduced congestion. Also, quality VHD estimations can improve the road user cost estimation process by generating road-specific results.

These methods can also be used to calibrate traditional methods, like those from the Highway Capacity Manual (*51*). In the work zone methodology outlined in chapter 10, speed adjustment factors and capacity adjustment factors are calculated and can be used to estimate queue length. A comparison between the HCM-derived queue lengths and the data-driven queue lengths can be conducted for a work zone to determine if the HCM assumptions match what is seen in the data or the data model.

5.3 Future Work

The main goal of this work is to explore the different relationships between the factors that describe congestion and the factors that describe the impact boundary. Multiple Linear Regression is not the only method in which this can be accomplished. A study of other methods and their applications to this study area would be recommended. Examples include Logistic Regression and Quantile Regression as presented by Kazi (*18*).

Also, this study explored the use of Python to conflate probe GPS speed data with the TRIMARC incident log. Given the amount of manual work required to update the impact boundaries and impute missing data, a more sophisticated approach is desired. This approach would better handle the exceptions presented in this report and reduce the amount of manual labor required to successfully process the data while maintaining accuracy when calculating impact boundaries and imputing missing data. Ultimately, this would allow DOTs and MPOs to process larger datasets.

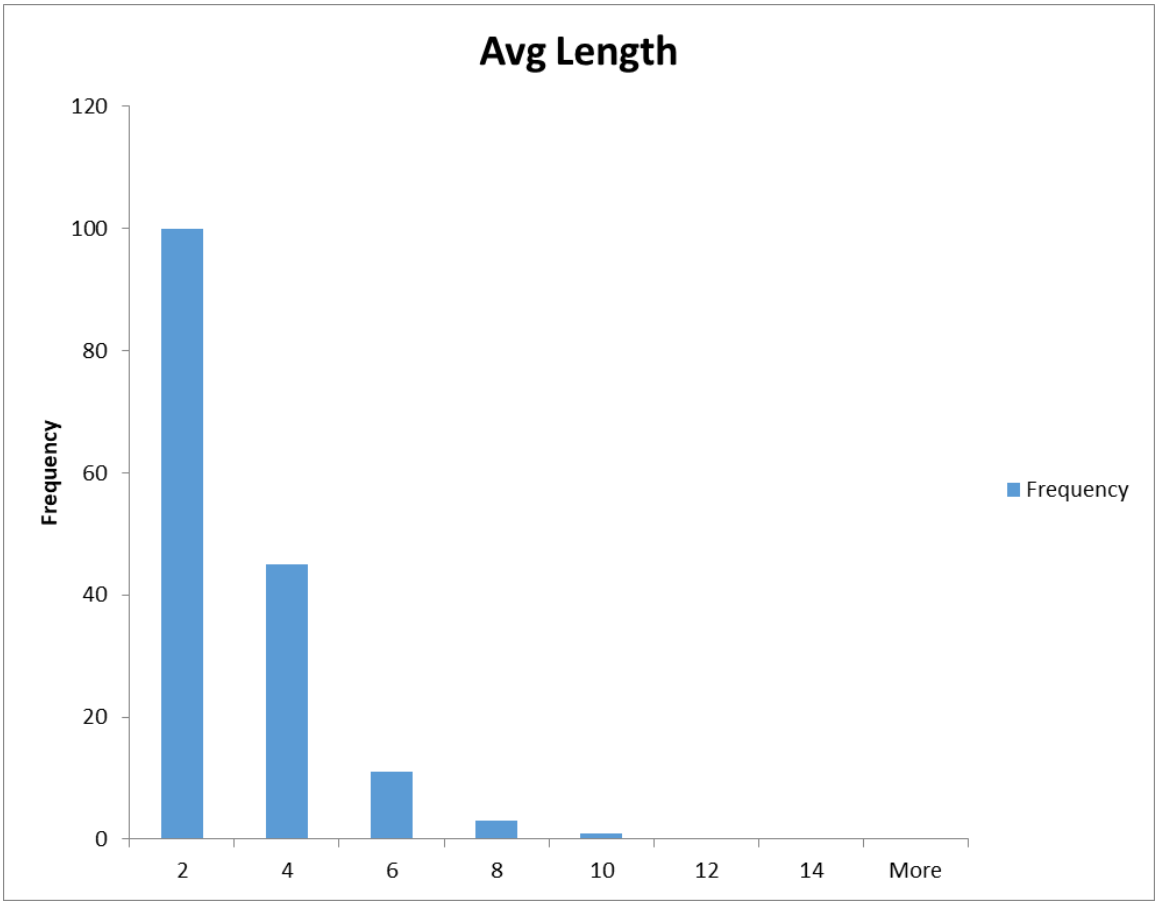
This study presented multiple models in which explain queue length and delay. For DOTs and MPOs to decide which model works best, a new set of data would be required to justify the accuracy of each model using metrics such as Root Mean Square Error (RMSE). With this information, DOTs and MPOs will have more information to properly select the correct model especially if it is desired to use a simpler-to-collect data source such as site duration; if site duration does not greatly reduce the accuracy of prediction to where the model is useless, it may be more effective versus models using zone area or queue length.

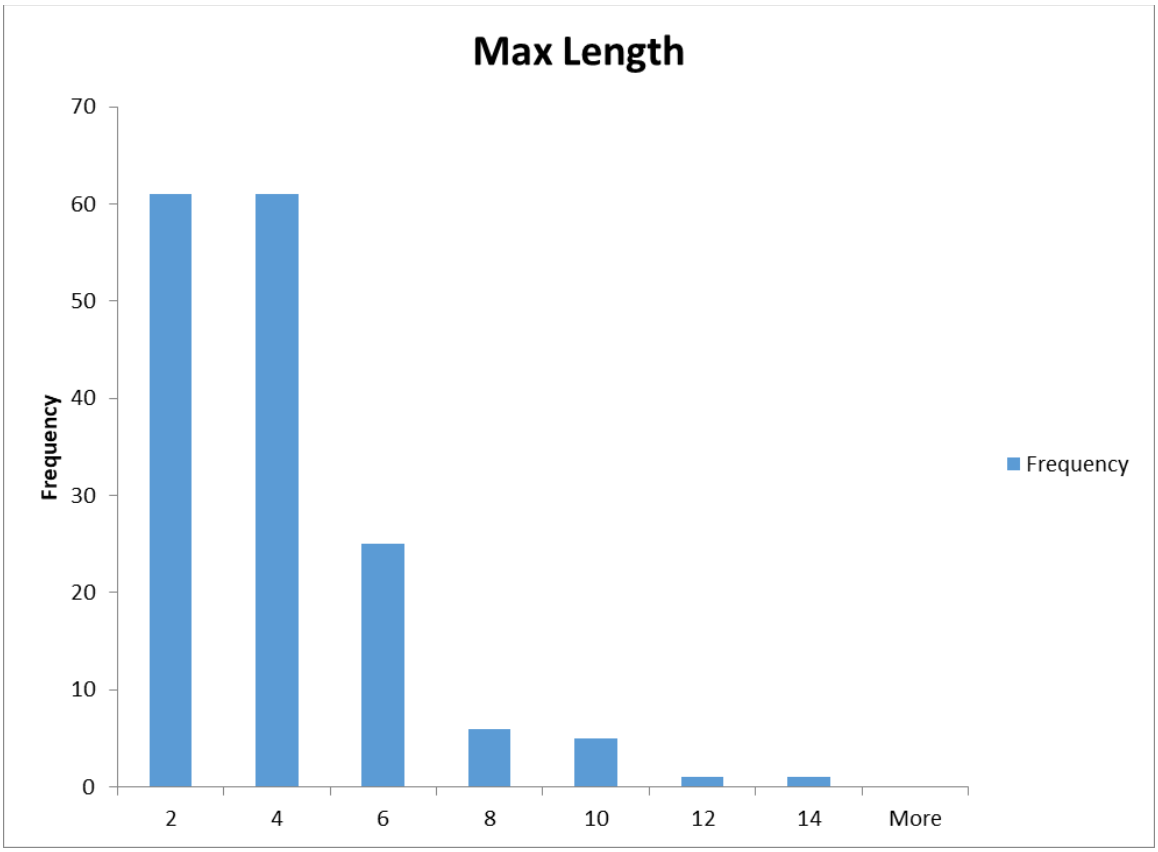
The results of this work are influenced not only by the real data but an imputation process that replaces missing probe GPS speed data as described in section 3.4.2. This could potentially introduce bias into the given results. More complete datasets could potentially yield more accurate representations of congestion and reduce estimation bias. Data Fusion techniques such as spatiotemporal cokriging could impute missing data with potentially more accurate speed data (*44*)

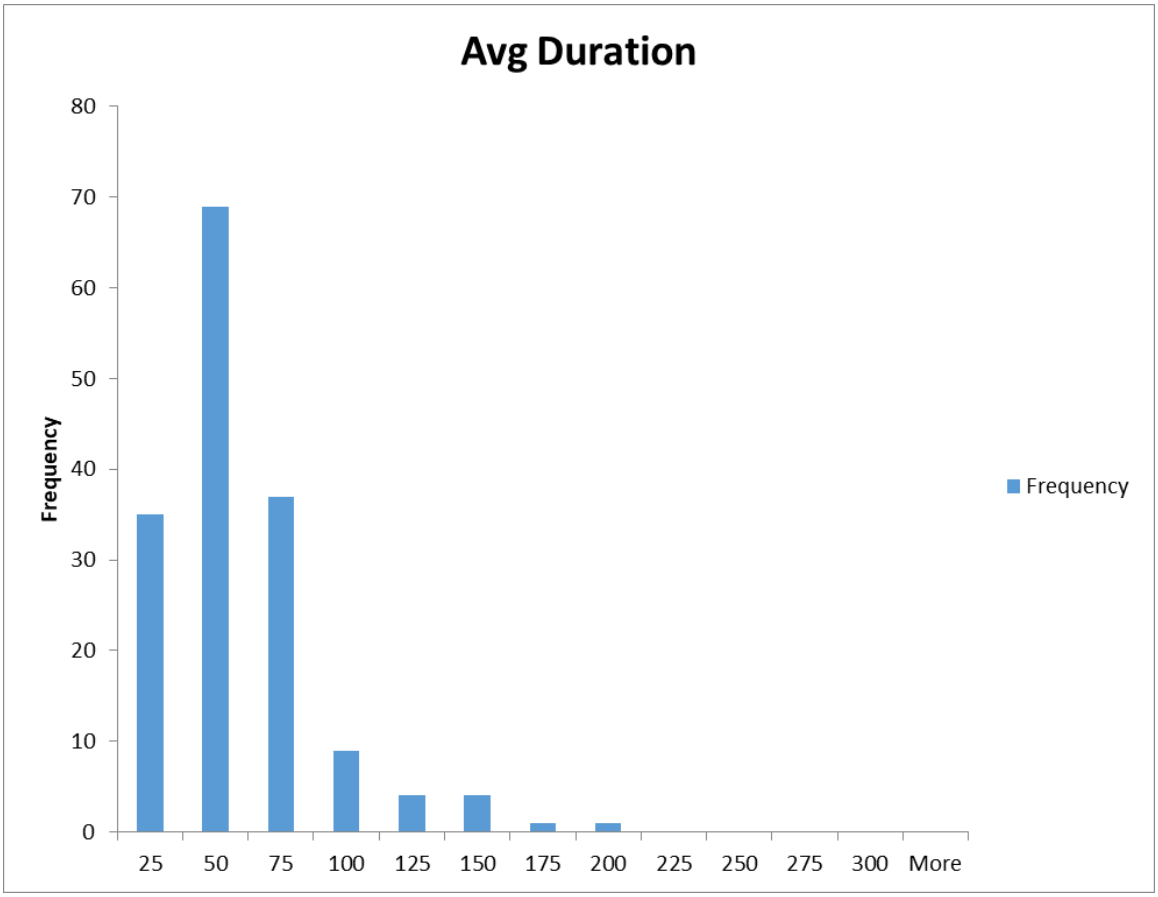
Finally, an analysis approach using arterial streets is recommended for future development. The signalization of an arterial roadway provides new challenges to identify what is considered typical speeds since arterial speeds are so volatile. Along with the freeway methodology, an arterial methodology, especially with comparable VHD and impact zone boundaries, can provide information not only for the freeway network, but for the entire roadway network further providing insight on the nature of congestion.

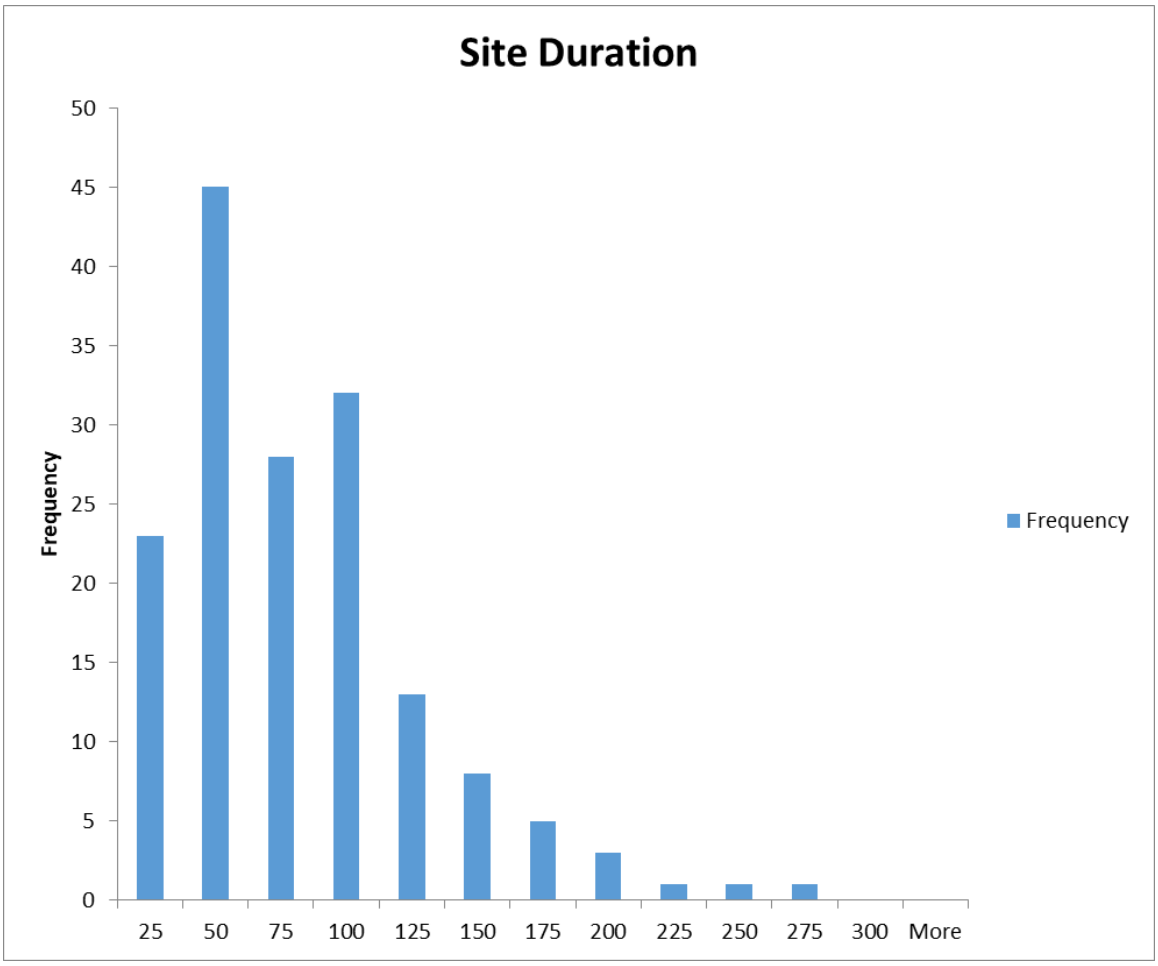
Appendix

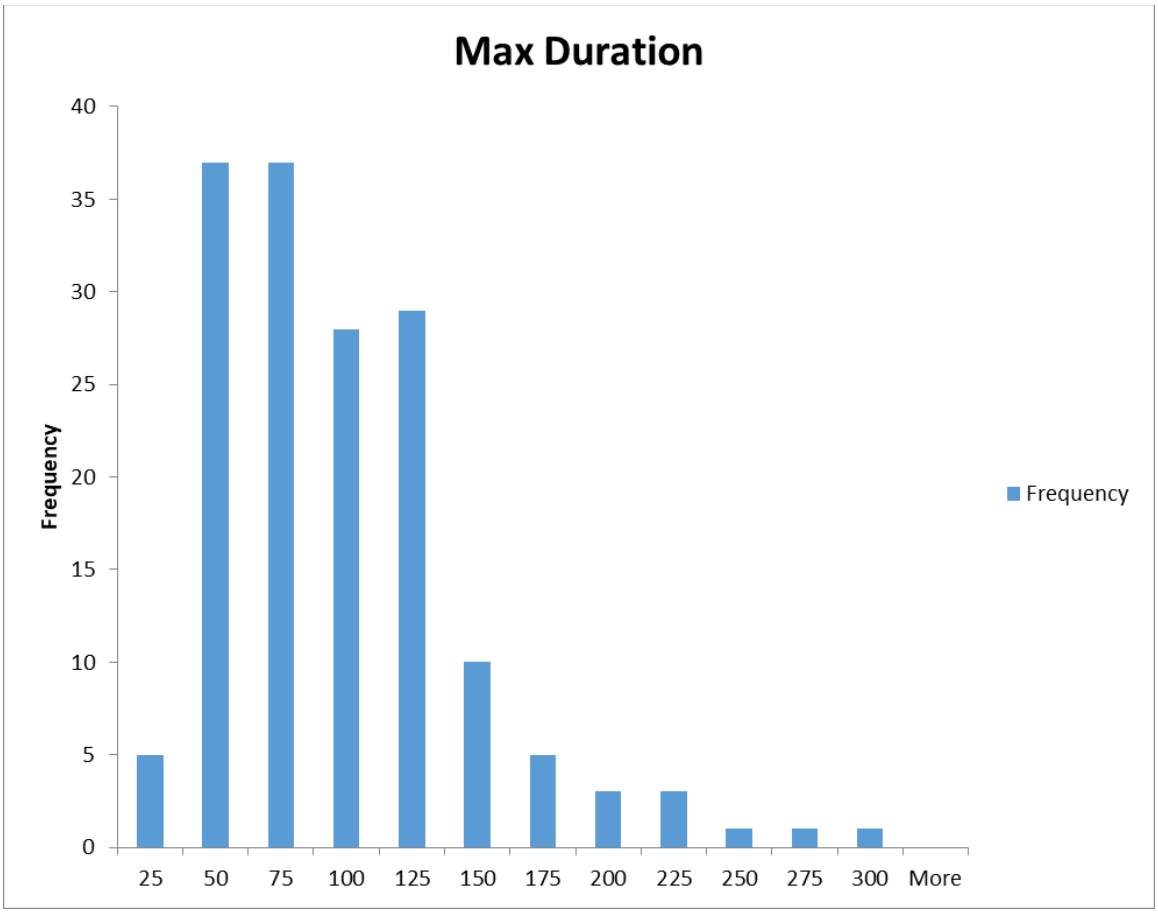
A1 Histograms

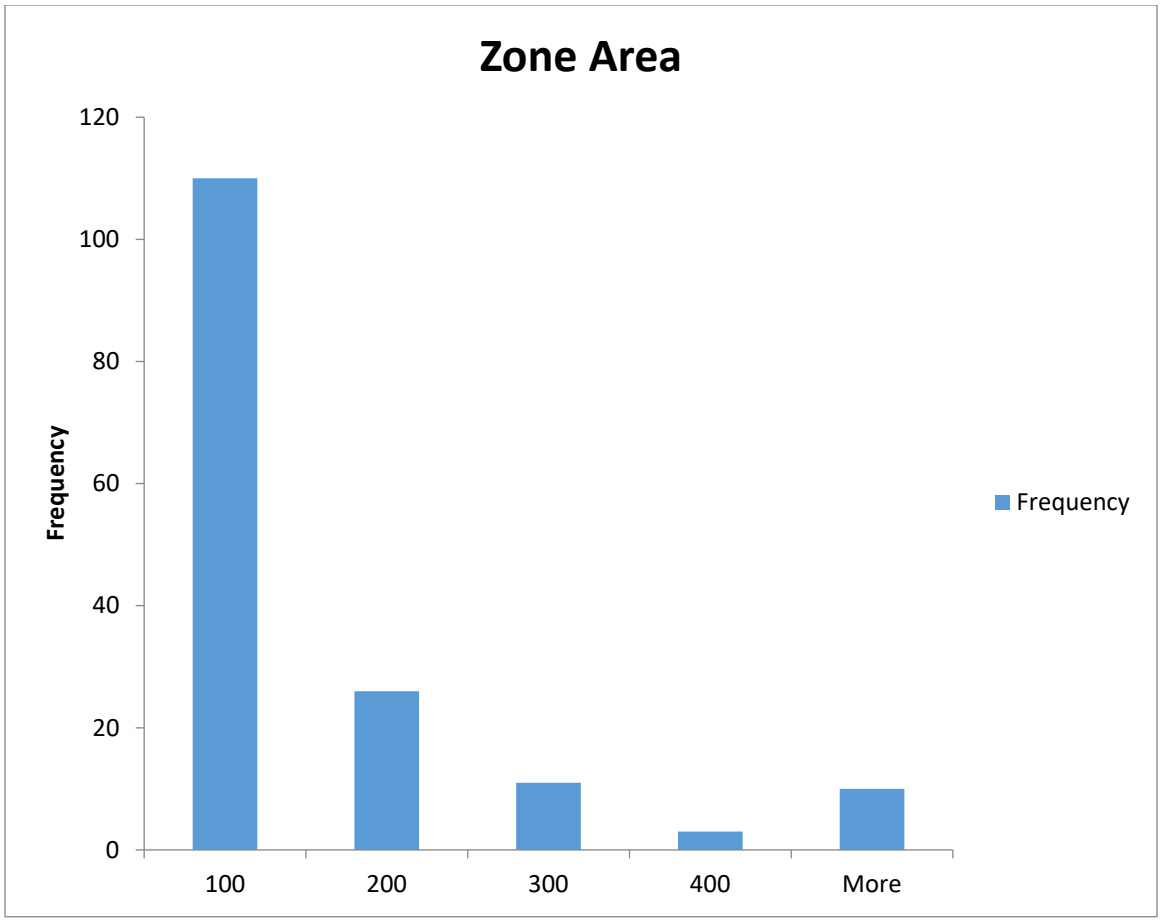



Max Length
70
60
50
40
30
20
10
0
Frequency
2
4
6
8
10
12
14
More
Frequency


Avg Duration
Frequency
80
70
60
50
40
30
20
10
0
25
50
75
100
125
150
175
200
225
250
275
300
More
Frequency


Site Duration
Frequency
50
45
40
35
30
25
20
15
10
5
0
25
50
75
100
125
150
175
200
225
250
275
300
More
Frequency


Max Duration
Frequency
40
35
30
25
20
15
10
5
0
25
50
75
100
125
150
175
200
225
250
275
300
More
Frequency


Zone Area
120
100
80
60
40
20
0
Frequency
100
200
300
400
More
Frequency

A2 Model Outputs

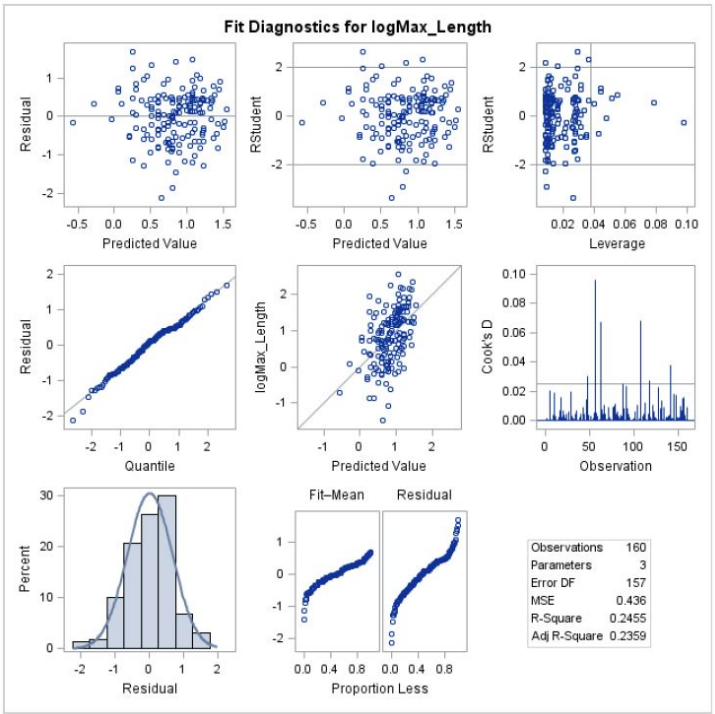
1. Average Queue Length and Maximum Queue Length

R-Square	Coeff Var	Root MSE	logMax_Length Mean
0.245541	78.82957	0.660317	0.837651

Source	DF	Type I SS	Mean Square	F Value	Pr > F
logSite_Duration	1	19.79266847	19.79266847	45.39	<.0001
Time_Period	1	2.48622672	2.48622672	5.70	0.0181

Source	DF	Type III SS	Mean Square	F Value	Pr > F
logSite_Duration	1	19.94619076	19.94619076	45.75	<.0001
Time_Period	1	2.48622672	2.48622672	5.70	0.0181

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-1.064423917	B	0.29725937	-3.58	0.0005
logSite_Duration	0.487370342		0.07205789	6.76	<.0001
Time_Period Non-Peak	-0.292944518	B	0.12267821	-2.39	0.0181
Time_Period Peak	0.000000000	B	.	.	.

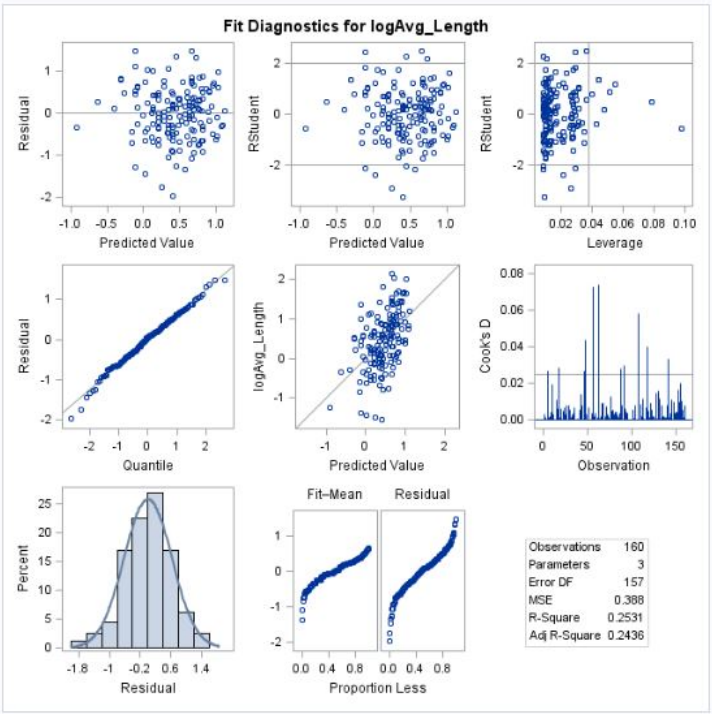


R-Square	Coeff Var	Root MSE	logAvg_Length Mean
0.253091	140.1184	0.622862	0.444526

Source	DF	Type I SS	Mean Square	F Value	Pr > F
logSite_Duration	1	18.27740727	18.27740727	47.11	<.0001
Time_Period	1	2.36182077	2.36182077	6.09	0.0147

Source	DF	Type III SS	Mean Square	F Value	Pr > F
logSite_Duration	1	18.42123376	18.42123376	47.48	<.0001
Time_Period	1	2.36182077	2.36182077	6.09	0.0147

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-1.382444295	B	0.28039823	-4.93	<.0001
logSite_Duration	0.468369353		0.06797062	6.89	<.0001
Time_Period Non-Peak	-0.285521278	B	0.11571966	-2.47	0.0147
Time_Period Peak	0.000000000	B	.	.	.



2. VHD

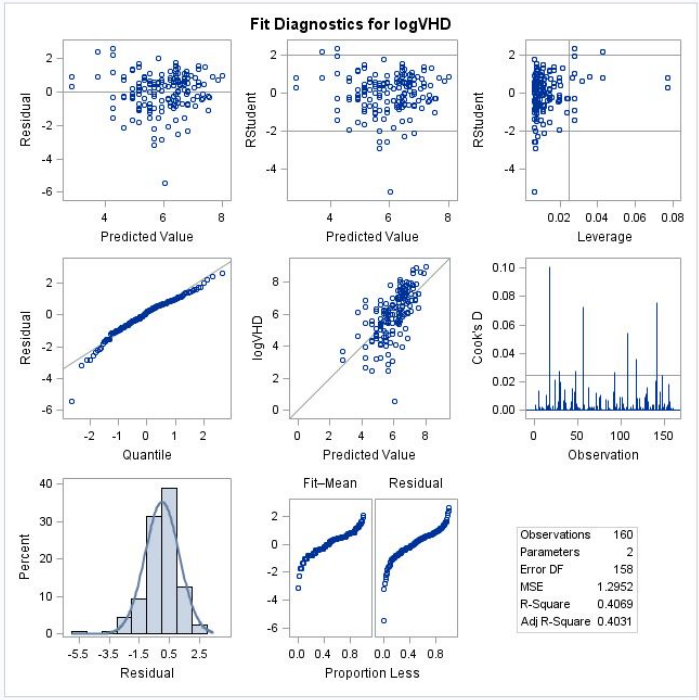
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	140.3950484	140.3950484	108.39	<.0001
Error	158	204.6493017	1.2952487		
Corrected Total	159	345.0443501			

R-Square	Coeff Var	Root MSE	logVHD Mean
0.406890	19.04132	1.138090	5.976948

Source	DF	Type I SS	Mean Square	F Value	Pr > F
logSite_Duration	1	140.3950484	140.3950484	108.39	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
logSite_Duration	1	140.3950484	140.3950484	108.39	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.746379674	0.51039302	1.46	0.1456
logSite_Duration	1.292939715	0.12418781	10.41	<.0001



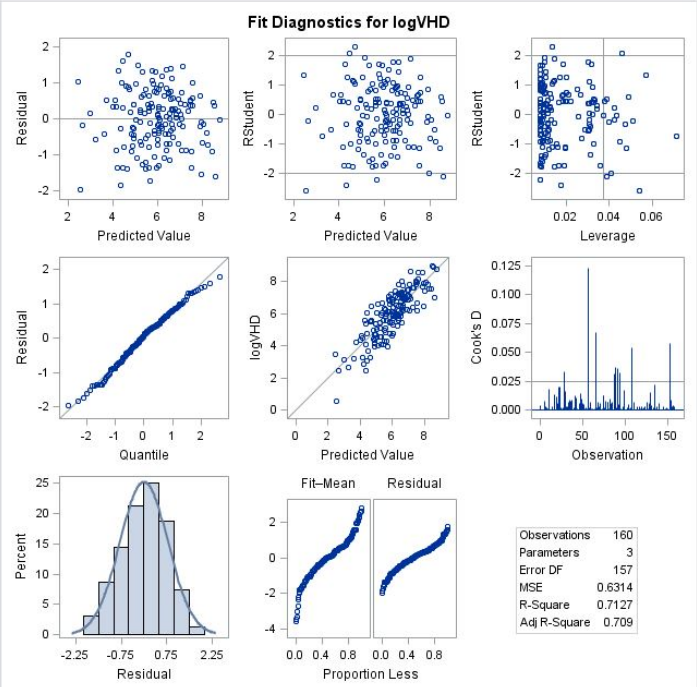
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	245.9072274	122.9536137	194.72	<.0001
Error	157	99.1371226	0.6314466		
Corrected Total	159	345.0443501			

R-Square	Coeff Var	Root MSE	logVHD Mean
0.712683	13.29501	0.794636	5.976948

Source	DF	Type I SS	Mean Square	F Value	Pr > F
logAvg_Length	1	238.6246429	238.6246429	377.90	<.0001
Blocked_Lanes	1	7.2825846	7.2825846	11.53	0.0009

Source	DF	Type III SS	Mean Square	F Value	Pr > F
logAvg_Length	1	239.1690122	239.1690122	378.76	<.0001
Blocked_Lanes	1	7.2825846	7.2825846	11.53	0.0009

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	5.106902583	B	0.08073958	63.25	<.0001
logAvg_Length	1.712591894		0.08799738	19.46	<.0001
Blocked_Lanes More than 1	0.527294861	B	0.15526698	3.40	0.0009
Blocked_Lanes 1	0.000000000	B	.	.	.



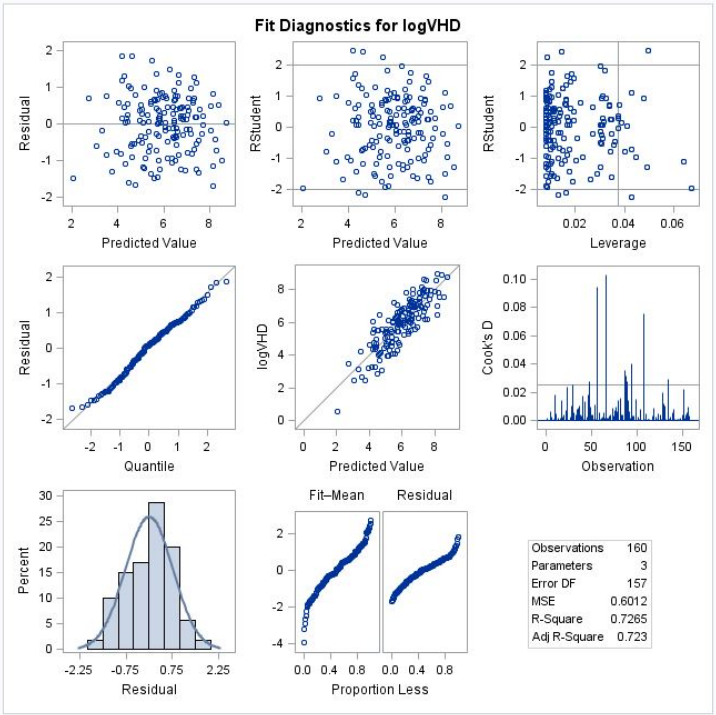
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	250.6623139	125.3311569	208.48	<.0001
Error	157	94.3820362	0.6011595		
Corrected Total	159	345.0443501			

R-Square	Coeff Var	Root MSE	logVHD Mean
0.726464	12.97225	0.775345	5.976948

Source	DF	Type I SS	Mean Square	F Value	Pr > F
logMax_Length	1	245.5280478	245.5280478	408.42	<.0001
Blocked_Lanes	1	5.1342661	5.1342661	8.54	0.0040

Source	DF	Type III SS	Mean Square	F Value	Pr > F
logMax_Length	1	243.9240987	243.9240987	405.76	<.0001
Blocked_Lanes	1	5.1342661	5.1342661	8.54	0.0040

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	4.511880547	B	0.09641069	46.80	<.0001
logMax_Length	1.639983229		0.08141546	20.14	<.0001
Blocked_Lanes More than 1	0.442829889	B	0.15152789	2.92	0.0040
Blocked_Lanes 1	0.000000000	B	.	.	.



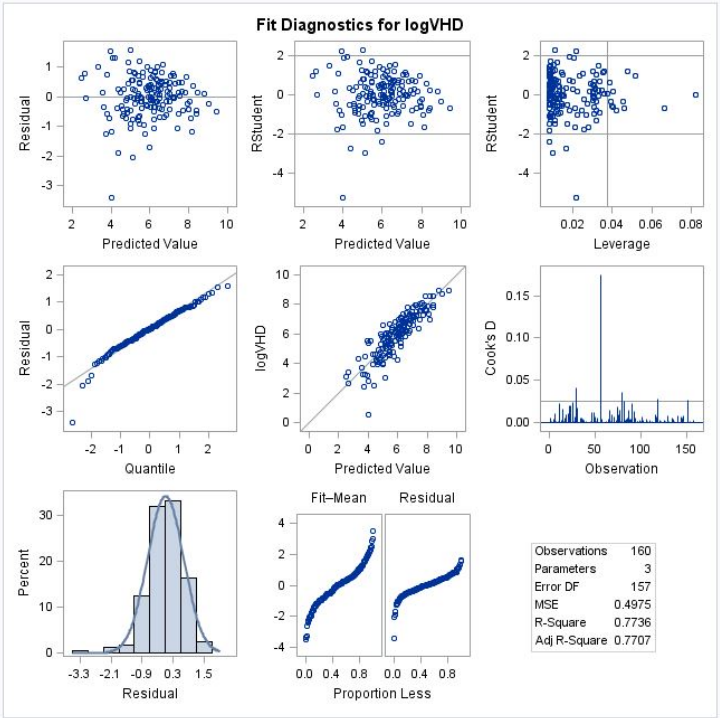
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	266.9324312	133.4662156	268.26	<.0001
Error	157	78.1119189	0.4975281		
Corrected Total	159	345.0443501			

R-Square	Coeff Var	Root MSE	logVHD Mean
0.773618	11.80129	0.705357	5.976948

Source	DF	Type I SS	Mean Square	F Value	Pr > F
logAvg_Zone_Area	1	263.0631586	263.0631586	528.74	<.0001
Blocked_Lanes	1	3.8692726	3.8692726	7.78	0.0059

Source	DF	Type III SS	Mean Square	F Value	Pr > F
logAvg_Zone_Area	1	260.1942159	260.1942159	522.97	<.0001
Blocked_Lanes	1	3.8692726	3.8692726	7.78	0.0059

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	1.304453941	B	0.20932149	6.23	<.0001
logAvg_Zone_Area	1.113793471		0.04870401	22.87	<.0001
Blocked_Lanes More than 1	0.384630704	B	0.13792345	2.79	0.0059
Blocked_Lanes 1	0.000000000	B	.	.	.



A3 T-Tests for TRIMARC Duration vs Site Duration

Note: Negative mean denotes under reporting. Positive mean denotes over reporting

1. Length of Incident

	under 1 hour	*More than 1 hour*
Mean	-24.2	-1.4
Variance	1561	1083
Observations	95	60
Hypothesized Mean Difference	0.0	
df	142	
t Stat	-3.9	
P(T<=t) one-tail	0.0	
t Critical one-tail	1.7	
P(T<=t) two-tail	0.0	
t Critical two-tail	2.0	

2. Injury Type

	Injury	*No Injury*
Mean	-8.7	-19.1
Variance	911	1791
Observations	56	99
Hypothesized Mean Difference	0.0	
df	145	
t Stat	1.8	
P(T<=t) one-tail	0.0	
t Critical one-tail	1.7	
P(T<=t) two-tail	0.1	
t Critical two-tail	2.0	

3. Time Period

	Peak	*Non-Peak*
Mean	-20.2	2.1
Variance	1682	441.1
Observations	121	34
Hypothesized Mean Difference	0.0	
df	108	
t Stat	-4.3	
P(T<=t) one-tail	0.0	
t Critical one-tail	1.7	
P(T<=t) two-tail	0.0	
t Critical two-tail	2.0	

4. Examples of Location Tests

I64W

	MP 0-8	*MP 8-16*
Mean	-32.0	-11.5
Variance	2426	1310
Observations	45.0	34.0
Hypothesized Mean Difference	0.0	
df	77.0	
t Stat	-2.1	
P(T<=t) one-tail	0.0	
t Critical one-tail	1.7	
P(T<=t) two-tail	0.0	
t Critical two-tail	2.0	

I64E

	MP 8-16	*MP 16-24*
Mean	5.0	-17.0
Variance	377.8	895.6
Observations	10	10
Hypothesized Mean Difference	0.0	
df	15	
t Stat	1.9	
P(T<=t) one-tail	0.0	
t Critical one-tail	1.8	
P(T<=t) two-tail	0.1	
t Critical two-tail	2.1	

References

[1] Schrank, D., Eisele, B., Lomax, T., Bak, J. 2015 Urban Mobility Scorecard. Texas Transportation Institute, 2015.
[2] McNamara, M., Li, H., Remias, S., Horton, D., and Cox, E. Real-Time Probe Data Dashboards for Interstate Performance Monitoring During Winter Weather and Incidents. Purdue University, 2016. p. 21.
[3] Schroeder, B., Rouphail, N., Sajjadi, S., Fowler, T. Corridor-Based Forecasts of Work-Zone Impacts for Freeways. North Carolina Department of Transportation, 2011. p. 138.
[4] Rister, B., Graves, C. The Cost of Construction Delays and Traffic Control for Life-Cicle Analysis of Pavements. University of Kentucky, Kentucky Transportation Center, 2002. p. 46.
[5] Jenkins, J., McAvoy, D. Evaluation of Traffic Flow Analysis and Road User Cost Tools Applied to Work Zones. The Ohio Department of Transportation, 2015. p. 97.
[6] Batson, R., Turner, D., Ray, P., Wang, M., Wang, P., Fincher, R., Lanctot, J., Cui, Q. Work Zone Lane Closure Analysis Model. University Transportation Center for Alabama, 2009. p. 105.
[7] Martin, P., Chaudhuri, P. Development of a Statewide User Cost Manual for Rural Work Zones. University of Utah, Utah Department of Transportation, 2010. p. 39.
[8] Abdel-Rahim, A., Cooley, H., Khanal, M. Synthesis of Research on Work Zone Delays and Simplified Application of QuickZone Analysis Tool. Boise State University, Idaho Department of Transportation, 2010. p. 52.
[9] Zhang, M. Integrated Work Zone Traffic Management. University of California Davis, State of California Department of Transportation, 2011. p. 102.
[10] Wang, Y., Cheevarunothai, C. Quantifying Incident-Induced Travel Delays on Freeways Using Traffic Sensor Data. Washington State Department of Transportation, 2008. p. 76.
[11] Park, S., Rakha, H., Guo, F. Multi-state Travel Time Reliability Model: Impact of Incidents on Travel Time Reliability. IEEE, 2011. p. 6.
[12] Chung, Y., Recker, W. A Methodological Approach for Estimating Temporal and Spatial Extent of Delays Caused by Freeway Accidents. No. 13, IEEE, 2012. p. 8.
[13] Edara, P., Rahmani, R., Brown, H., Sun C. Traffic Impact Assessment of Moving Work Zone Operations. Federal Highway Administration, 2017. p. 112.
[14] Bae, J., Choi, K., Oh, J. Multi-Contextual Approach to Modeling Traffic Impact of Urban Highway Work Zones. Transportation Research Board, 2017. p. 17.
[15] Hou, Y., Edara, P., Sun, C. Traffic Flow Forecasting for Urban Work Zones. IEEE, 2015. p. 10.
[16] Du, B., Chien, S., Lee, J., Spasovic, L., Mouskos, K. Artificial Neural Network Model for Estimating Temporal and Spatial Freeway Work Zone Delay Using Probe-Vehicle Data. Transportation Research Board, 2016. p. 7.
[17] Edara, P., Chang, Y., Sun, C., Brown, H. Data-Driven Traffic Impact Assessment Tool for Work Zones. Institute for Transportation Iowa State University, 2017. p. 58.
[18] Kazi, O. A Data Driven Approach to Quantifying the Impact of Crashes. Civil Engineering, No. Masters in Civil Engineering, University of Kentucky, 2016. p. 50.

[19] Seeherman, J., Skabardonis, A. Quantification of Weather Influences on Freeway Bottlenecks. University of Californis Berkeley, Transportation Research Board, 2012. p. 18.
[20] Wright, B., Zou, Y., Wang, Y. Impact of Traffic Incidents on Reliability of Freeway Travel Times. Transportation Research Board, 2015. p. 9.
[21] Edwards, M., Fontaine, M. Investigation of Travel Time Reliability in Work Zones with Private-Sector Data. Transportation Research Board 2012. p. 10.
[22] Mathew, J., Krohn, D., Li, H., Day, C. Bullock, D. Implementation of Probe Data Performance Measures. Purdue University, Commonwealth of Pennsylvania Department of Transportation, 2017. p. 138.
[23] Chung, Y., Kim, H., Park M. Quantifying Non-Recurrent Traffic Congestion Caused by Freeway Work Zones Using Archived Work Zone and ITS Traffic Data. Transportmetrica, 2012. p. 15.
[24] Yildirimoglu, M., Geroliminis, N. Experienced Travel Time Prediction for Congested Freeways. Transportation Research, 2013. p. 9.
[25] Habtemichael, F., Cetin, M., Anuar, K. Incident-Induced Delays on Freeways Quantification Method by Grouping Similar Traffic Patterns. Transportation Research Board, 2015. p. 10.
[26] Sullivan, A., Sisiopiku, V., Kallem, B. Measuring Non-Recurring Congestion in Small to Medium Sized Urban Areas. The University of Alabama Huntsville, University Transportation Center for Alabama, 2013. p. 66.
[27] Anbaroğlu, B., Cheng, T., Heydecker, B. Non-Recurrent Traffic Congestion Detection on Heterogeneous Urban Road Networks. Transportmetrica, 2015. p. 19.
[28] Li, H., Remias, S., Day, C., Mekker, M., Sturdevant, J., Bullock, D. Shock Wave Boundary Identification Using Cloud-Based Probe Data. Transportation Research Board, 2015. p. 10.
[29] Pesti, G., Brydia, R. Work Zone Impact Assessment Methods and Applications. Texas A&M University, 2016. p. 12.
[30] Khattak, A., Wang, X., Zhang, H. Incident Management Integration Tool: Dynamically Predicting Incident Durations, Secondary Incident Occurrence and Incident Delays. Old Dominion University, Intelligent Transportation Systems, 2012. p. 12.
[31] Zou, Y., Ye, X., Henrickson, K., Tang, J., Wang, Y. Jointly analyzing freeway traffic incident clearance and response time using a copula-based approach. Transportation Research, 2018. p. 11.
[32] Boyles, S., Fajardo, D., Waller, S. A Naïve Bayesian Classifier for Incident Duration Prediction. University of Texas, 2006. p. 12.
[33] Hojati, A., Ferreira, L., Charles, P., Kabit, M. Analysing Freeway Traffic-Incident Duration Using an Australian Dataset. Road & Transport Research, 2012. p. 14.
[34] Young, S. Real-Time Traffic Operations Data Using Vehicle Probe Technology. Mid-Continenet Transportation Research Symposium, 2007. p. 8.
[35] NPMRDS and You: Leveraging Statewide Travel Time Data for Projects, Operations, and Planning. WisDOT and TOPS Lab Webinar Series, 2016.
[36] University of Maryland CATT Lab. RITIS Platform: Features & Applications Overview. 2015.
[37] iPeMS. iPeMS: Empower your Agency to Improve System performance and Demonstrate Success, Automatically.

[38] Perley, S. Oregon's Smart Solution to Total Eclipse Traffic. No. 2018, 2017.
[39] Chien, S. Feasibility of Lane Closures Using Probe Data. New Jersey Institute of Technology, New Jersey Department of Transportation, 2017. p. 105.
[40] Weather Underground. https://www.wunderground.com. 2018.
[41] Kentucky Transportation Cabinet. Traffic Counts | KYTC. https://transportation.ky.gov/Planning/Pages/Traffic-Counts.aspx. 2018.
[42] Turner, S., Eisele, W., Benz, R., Holdener, D., and Holdener. Travel Time Data Collection Handbook. Federal Highway Administration, 1998.
[43] Chang, G., Zhang, Y., Yao, D. Missing Data Imputation for Traffic Flow Based on Improved Local Least Squares. Tsinghua National Laboratory for Information Science and Technology Department of Automation, Tsinghua University, Tsinghua Science and Technonoly, 2012. p. 6.
[44] Bae, B., Kim, H., Lim, H., Liu, Y., Han, L., Freeze, P. Missing Data Imputation for Traffic Flow Speed Using Spatiotemporal Cokriging. University of Tennessee Knoxville, Transportation Research, 2017. p. 17.
[45] Treiber, M., Kesting, A. Reconstructing the Traffic State by Fusion of Heterogeneous Data. Computer-Aided Civil and Infrastructure Engineering, 2011. p. 13.
[46] Li, Z., Kluger, R., Hu, X., Wu, Y., Zhu, X. Reconstructing Vehicle Trajectories to Support Travel Time Estimation. Trasnportation Research Board, 2018. p. 21.
[47] Shan, Z., Zhao, D., Xia, Y. Urban Road Traffic Speed Estimation for Missing Probe Vehicle Data Based on Multiple Linear Regression Model. IEEE, 2013. p. 6.
[48] Smith, B., Scherer, W., and Conklin, J. Exploring Imputation Techniques for Missing Data in Transportation Management Systems. Transportation Research Board, 2003. p. 11.
[49] SAS. PROC GLM for Unbalanced ANOVA.
[50] Tastan, H. Heteroscedasticity. Yildiz Technical University, 2012.
[51] Transportation Research Board. Highway Capacity Manual, Sixth Edition: A Guide for Multimodal Mobility Analysis.

Vita

- Jacob Douglas Keaton Brashear
- Place of Birth: Lexington, KY
- Educational Institutions:
    - University of Kentucky
        - Bachelors of Science in Mining Engineering
        - Masters in Business Administration
- Professional Positions Held
    - University of Kentucky
        - Graduate Research Assistant
        - Graduate Teaching Assistant
        - ITE Student Chapter President
        - ASCE Member
    - Ceramic Technology Incorporated
        - Sales Engineer
    - Patriot Coal Corporation
        - Plant Technician
        - Coal Preparation Engineer
        - Operations Management Trainee
    - Alpha Natural Resources
        - Student Intern
- Scholastic Honors
    - University of Kentucky
        - Graduate Research Assistantship
        - Chi Epsilon
        - MBA Scholarship
        - Alpha Natural Resources Scholarship
        - Mining Engineering Scholarship
        - Kentucky River Scholarship
        - KEES Scholarship
        - Commonwealth Scholarship
- Publications and Presentations
    - Zhang, X., Kazi, O.R., Brashear, J., and Chen, M., *A Data-Driven Approach to Quantify the Impact of Crashes* Poster Presented at the Transportation Research Board of the National Academies, Washington, D.C., 2018.