UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

BIG DATA ANALYTICS IN TRANSPORTATION NETWORKS USING THE NPMRDS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

In
Electrical and Computer Engineering

By

NAIM BITAR
Norman, Oklahoma
2016

BIG DATA ANALYTICS IN TRANSPORATION NETWORKS USING THE NPMRDS

A THESIS APPROVED FOR THE
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

BY

_____
Dr. Hazem Refai, Chair

_____
Dr. William Ray

_____
Dr. Ali Imran

To my parents, the two people who made me who I am today:

Samira and George.

To my beloved sister,

Alfreda.

To my love,

Christen Hammell

I dedicate this humble work to you,

Naim

Acknowledgements

I would like to extend my heartfelt thanks and gratitude to my advisor **Dr. Hazem Refai** for his support, direction and mentorship. This work would not have been possible without his vision and guidance. In addition, I extend my sincere appreciation to my thesis committee members, **Dr. William Ray** and **Dr. Ali Imran** for their time and effort in reviewing this thesis. Appreciation goes as well to **Michelle Farabough** for editing this thesis.

I also wish to express gratitude to my friends, office mates, and the extended OU family of professors, students and staff. My sincere appreciation to you all.

Table of Contents

List of Tables

List of Figures

Abstract

Urban traffic congestion is common and the cause for loss of productivity (due to trip delays) and higher risk to passenger safety (due to increased time in the automobile), not to mention an increase in fuel consumption, pollution, and vehicle wear. The fiduciary effect is a tremendous burden for citizens and states alike. One way to alleviate these ill effects is increasing state roadway and highway capacity. Doing so, however, is cost prohibitive. A better option is improving performance measurements in an effort to manage current roadway assets, improve traffic flow, and reduce road congestion.

Variables like segment travel time, speed, delay, and origin-to-destination trip time are measures frequently used to monitor traffic and improve traffic flow on the state roadways. In 2014, ODOT was given access to the FHWA's National Performance Management Research Data Set (NPMRDS), which includes average travel times divided into contiguous segments with travel time measured every 5 minutes. Travel times are subsequently segregated into passenger vehicle travel time and freight travel time. Both types of time are calculated using GPS location transmitted by way of participating drivers traveling along interstate highways.

This thesis presents research detailing the use of NPMRDS dataset consisting of highway vehicle travel times, for computing performance measurements in the state of Oklahoma. Data extraction, preprocessing, and statistical analysis were performed on the dataset. A comprehensive study of the dataset characteristics, including influencing variables that affect data measurements are presented. A process for identifying anomalies is developed, and recommendations for improving accuracy and alleviating data anomalies are reported. Furthermore, a process for filtering and removing speed data

outliers across multiple road segments is developed, and comparative analysis of raw baseline speed data and cleansed data is performed. Identification and computational comparison of travel time reliability performance measurements is done. A method for improved congestion detection is investigated and developed. Finally, traffic analytics using machine learning is performed to identify and to classify congested segments and a novel approach for identifying non-recurrent congestion sources using Bayesian inference of speed data is also developed and introduced.

Chapter 1: Introduction

Traffic congestion is commonplace in populated cities where most commuters expect delays, especially during peak driving hours. Accordingly, travelers and transportation companies (i.e., shippers) adjust their schedules and budget additional time for unforeseen circumstances that alter travel time. However, unexpected congestion (i.e., traffic delay worse than usual) is even more troublesome for travelers [1] who desire travel time reliability (i.e., consistency or dependability in travel time) based on their typical day-to-day driving experience at various times throughout the day.

Traffic congestion is typically communicated in terms of simple averages. However, most travelers are quick to recall an incident that was much worse than their average travel time. Travel time can vary greatly from day to day, and days when a driver spent time suffering through an unexpected delay often stands out. Figure 1 illustrates this concept. In essence, averages do not tell the full story.

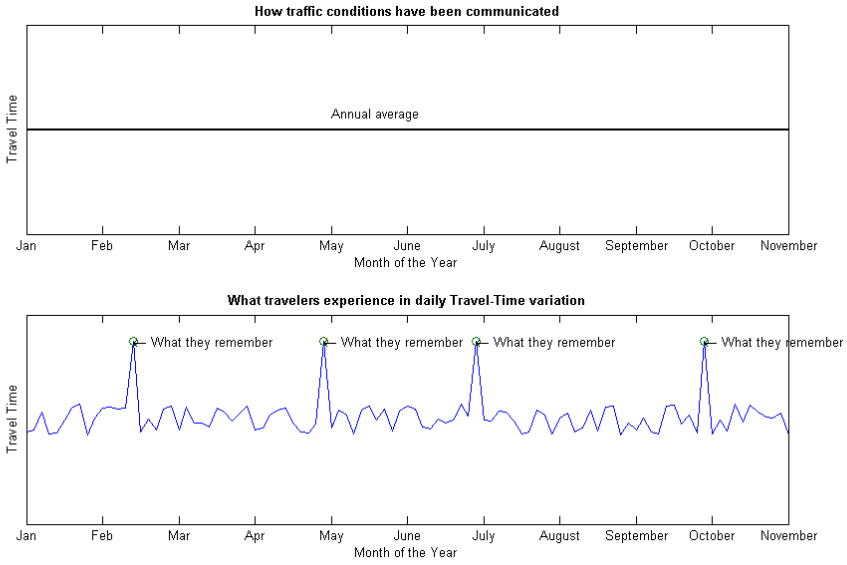


Figure 1 - Theoretical vs. perceived notion of congestion.

1.1. What is travel time reliability?

Work done by the University Of Florida Transportation Research Center in collaboration with Florida Department of Transportation (DOT) [2], provides a comprehensive review of travel time reliability. In an early report they quote Ebling's [3] widely accepted definition of reliability: "*the probability that a component or system will perform a required function for a given period of time when used under stated operating conditions. It is the probability of a non-failure over time.*" Ebling states that travel time reliability must be made specific by providing an unambiguous and observable description of a failure, including the unit of time over which failure will be evaluated. In other words, travel time reliability is the absence of variability in travel times. In a roadway network context, users perceive a reliable system as one in which each traveler or shipper experiences actual time-of-arrival (ATA) that matches desired-time-of-arrival (DTA) within some accepted window of time. This notion is shown in Figure 2.

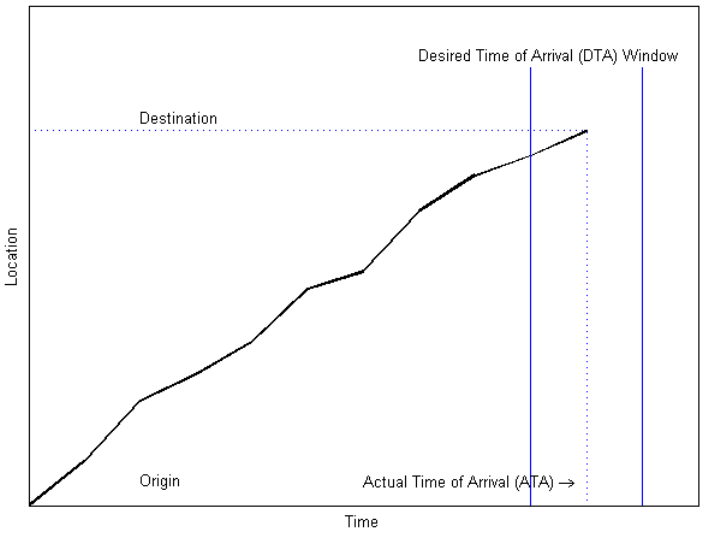


Figure 2 - Desired vs. actual times of arrival in defining travel time reliability.

1.2. What affects travel time reliability?

Researchers in [4], detail seven main causes that affect travel time reliability. These can roughly be grouped into three categories:

Category 1 — Non- Recurrent causes:

1. Traffic incidents. Traffic incidents are defined as events that disrupt the normal flow of traffic. In general, such incidents represent physical impedances in travel lanes on the roadways. Examples include roadway vehicle accidents, vehicle breakdowns, and debris obstructing travel lanes used for commute. In addition to physical, on-road impediments, events that occur on the shoulder or side of the road, even fire or accidents, can also impact traffic flow by distracting drivers, which can cause changes in driver behavior.

2. Work zones: Work zones include construction activity on the roadway that affects traffic flow and results in physical changes to the highway environment (e.g., reduction in the number or width of travel lanes, lane diversions, and temporary roadway closures). Unpredicted delays caused by work zones are one of the most frustrating conditions travelers encounter.

3. Weather: Environmental conditions like high levels of snow or rain precipitation, bright sunlight, fog, or icy roadway surface conditions can cause reduced visibility or hazardous driving conditions. Drivers will often react by lowering their speed and/or increasing their headway.

Category 2 —Recurrent causes:

4. Demand fluctuations. In-demand day-to-day variability in demand leads to higher traffic volume on some days than on others. When superimposed on a system with fixed capacity, such variability results in unreliable travel time.

5. Repetitive events. An out-of-the-ordinary, abnormally large traffic volume (e.g., special events like sporting events or concerts) occasionally occur and cause a surge in traffic demand that often times overwhelms a traffic system.

Category 3 — Continuous causes:

6. Traffic control devices. Intermittent disruption caused by control devices (e.g., poorly timed traffic signals and railroad grade crossings) could contribute to congestion and travel time variability, sometimes causing traffic disruption and changes in driver behavior at disjoint instances of time.

7. Inadequate base capacity. This effect on travel time reliability is defined as the maximum amount of traffic managed by a given highway section. Transportation engineers have long studied and addressed the physical capacity of roadways, which is determined by a number of factors (e.g., number and width of lanes and shoulders; merge areas, such as onramps and off ramps; and roadway alignment, such as grades and curves). Given that congestion occurs when volume is larger than roadway capacity, it can be said that inadequate base capacity creates delay in the same way traffic volume variations and fluctuations do, namely as bottlenecks in areas where section capacity is ineffective at supporting traffic volume.

1.3. Why travel time reliability?

Costs associated with travel time are critical factors when evaluating transportation infrastructure initiatives and investments aimed at minimizing time delay. As mentioned above, travel time reliability is a measure of the extent of unexpected delay. This measure is highly significant to a variety of transportation system users, including vehicle drivers, transit commuters, freight shippers, and air travelers. Personal and business travelers value reliability, as it affords them the utmost use of their time. Shippers and freight carriers require predictable travel times to remain competitive. Reliability is a value-added tangible on privately financed highways (i.e., tollways). The importance of reliability has forced transportation planners and decision-makers to consider travel time reliability a key performance measure.

1.4. National Performance Management Research Data Set (NPMRDS)

The Federal Highway Administration (FHWA) recognizes the importance of travel time reliability and its significance for quantifying the benefits of traffic management and roadway operations. Accordingly, the FHWA offers state DOTs a dataset of travel times for all National Highway System (NHS) roadways as a way of promoting the adoption and use of travel time reliability measures. Such nationwide data is designed to complement existing state DOT's travel time measurements and reports. The relationship of this National Performance Management Research Data Set (NPMRDS) and the Oklahoma Department of Transportation (ODOT) is the focus of this thesis and all work presented herein.

1.4.1. Overview of the NPMRDS

In 2013, the FHWA acquired a national dataset of average travel times for use in its performance measurement reports [5], most notably the Freight Performance Measures (FPM) and the Urban Congestion Report (UCR). The latter leverages data toward developing congestion and reliability measures in the 52 most populated urban areas in the U.S [6]. States and Metropolitan Planning Organizations (MPOs) can utilize the data to meet their Moving Ahead for Progress in the 21st Century Act (MAP-21) performance management requirements. Monthly data reports detail the entire NHS. Observed average travel time measurements are collected 24 hours-a-day in 5-minute intervals for freight truck vehicles and passenger vehicles, as well as for combined vehicles records for both types.

The NPMRDS is a probe based traffic data [7] characterized by high spatial-temporal record count variability generated by vehicles (i.e., probes) reporting to a central server via some type of telemetry. Passenger probe data is collected by HERE, and freight probe data is collected by the American Transportation Research Institute (ATRI). HERE data is collected from mobile phones, vehicle navigation systems, and portable navigation devices [8]. Freight data leverages embedded fleet data-collection systems. Combined travel time data is a weighted average of freight and passenger vehicle travel times based on respective traffic volumes. Neither freight nor passenger volumes are reported. The Geographic Information System (GIS) roadway network divides the NHS into directed segments. Time statistics are binned in 5-minute intervals per Traffic Message Channel (TMC) segment per vehicle type. Probe coordinates are based on Global Positioning System (GPS) equipment (e.g., smartphones, navigation devices) located in vehicles.

Recorded data is referenced to segments on a map, and multiple speed records collected from all probes in a single segment during any given 5-minute time bin are used to assign a travel time value to that particular segment. HERE's static files contain all TMC segment information details. The information is only updated when necessary changes are present. Table 1 details information associated with a static NPMRDS file and also provides a description of each entry.

A separate NPMRDS data file reports average travel times for roadways geo-referenced to each of the TMC location codes. Table 2 details a description of associated fields. Given the continuous, large scale, and probe-based nature of traffic data, the number of observations reported in variable traffic conditions can fluctuate significantly. Furthermore, because the FHWA has specified that no smoothing, outlier detection, or imputation of traffic will be performed on the NPMRDS data after it is collected by HERE, the dataset is known to contain unique characteristics that yield traditional processing techniques that are routinely performed by DOT agencies ineffective. This presents several challenges, as well as several opportunities for DOT agencies to make beneficial use of the data.

Table 1 - TMC static file format.

Field Name	Type	Example	Description
TMC	String	111N06515	The TMC code is an industry convention that defines a particular directional segment of the road. In North America, a consortium consisting of HERE (NAVTEQ) and Tele Atlas created and continually maintain the location code table that adheres to the international standard on location referencing (ISO 14819-3:20043) [9]. Traffic Location code in the format of: **CLLDTTTTT** • **C** is the Country Code (1 digit). • **LL** is the Country Code (2 digit). • **D** ('P' Positive or 'N' Negative direction of the TMC). • **TTTTT** is the Country Code (5 digit).
ADMIN_LEVEL_1	String	USA	The Country where the listed TMC is located.
ADMIN_LEVEL_2	String	Oklahoma	The State/Province where the listed TMC is located.
ADMIN_LEVEL_3	String	Osage	The County where the listed TMC is located.
DISTANCE	Float	7.2245	The length of TMC segment measured in Miles to five decimal places.
ROAD_NUMBER	String	US-60	The Route Number of the road.
ROAD_NAME	String	Bartlesville Rd	The Local Name of the route.
LATITUDE	Float	36.74456	WGS84 Latitude coordinate to five decimal places
LONGITUDE	Float	-96.29404	WGS84 Longitude coordinate to five decimal places
ROAD_DIRECTION	String	Westbound	Represents the direction of travel on the road.

Table 2 - Travel time file format.

Field Name	Type	Example	Description
TMC	String	111N06515	Traffic location code
DATE	String	01022014	Day Month Year (DDMMYYYY)
EPOCH	Integer	48	A value from 0 through 287 that defines the 5-minute peruid the average speed applies (local time)
Travel_TIME_ALL_VEHICLES	Integer	44	Travel times calculated in seconds between the segment length and the average speed on the segment. Average segment speed is determined from a combination of the passenger and freight trucks GPS probe speed observations.
Travel_TIME_PASSENGER_VEHICLES	Integer	76	Travel times calculated in seconds between the segment length and the average speed on the segment. Average segment speed is determined from only passenger individual GPS probe speed observations.
Travel_TIME_FREIGHT_TRUCKS	Integer	66	Travel times calculated in seconds between the segment length and the average speed on the segment. Average segment speed is determined from only freight trucks individual GPS probe speed observations.

1.4.2. Existing and related work using NPMRDS

Currently, DOTs, MPOs, and research institutions with some experience concerning analyzing probe data and performing big data analytics are utilizing the NPMRDS data in their performance measurements and reliability reports. Public documentation describing the NPMRDS dataset was first made available via a presentation given by the FHWA Office of Operations and Resource center, HERE, and The Volpe Center in November of 2013 [7]. Soon afterwards, research was reported by academic institutions and other parties who were interested in investigating ways to utilize the dataset. One of the earliest presentations was made by the Wisconsin Traffic Operations and Safety Laboratory during the second quarterly NPMRDS webinar in February, 2014 [10], [11] and [12]. Researchers discussed performance measures, along with a representation of the data on maps. Also, during the same webinar, the University of Maryland highlighted differences in the Traffic Message Channel (TMC) codes and map realizations used by NPMRDS and the I-95 Corridor Coalition's Vehicle Probe Project (VPP). Results indicated that direct comparisons between different sources should be carefully executed to account for differences in segment properties [10]. In March 2014, a collaborative effort by the University of Minnesota and Minnesota Department of Transportation explored the feasibility of using one month of NPMRDS data gathered in Minnesota to compute freight mobility and speed variations along the NHS during AM and PM peak periods [13]. No data filtering was performed prior to analysis and visualization. In April 2014, the ATRI center published work using NPMRDS data to compute congestion and the cost of delay incurred by the trucking industry [14]. Freight truck data from NPMRDS and data from ATRI's Freight Performance Measures database was used in the study.

During the third quarterly NPMRDS webinar in May 2014, Iteris shared their work implementing performance measures for Utah DOT [15]. The presentation indicated data imputation was the result of smoothing, although no filtering was applied to the dataset. A study comparing NPMRDS data with Bluetooth re-identification and VPP probe data was conducted at the University of Maryland and presented at the 2014 ITS World Congress [16]. Results were further expanded and subsequently presented at the 94th Annual Meeting of Transportation Research board in January 2015 [17]. Researchers concluded that congestion measures using the NPMRDS were accurate 95% of the time, and reliability measures were accurate 15% of the time. Researchers stated that "*At this point it is not clear whether the source of this difference is because NPMRDS data is non-filtered and not validated or something more intrinsic is occurring*". In 2015, the University of Maryland published a report [18] discussing the benefits of the NPMRDS dataset. In the report, they addressed how agencies could include travel time reliability as part of a cost-benefit analysis when making decisions about congestion reduction–related project investments. The University of Maryland also published their findings in the Transportation Research Board (TRB) [19]. Researchers discussed their methodology for processing NPMRDS data. In the article, the researchers described the use of 24-hour overlay plots for imputing missing values for any particular epoch. No outlier filtering was applied. The group also demonstrated a case study of comparing NPMRDS data and Bluetooth traffic probe data from INRIX. Researchers recommended investigating NPMRDS fidelity as the basis of performance and basic outlier detection. Researchers at Old Dominion University [20] (in collaboration with the Virginia Center for Transportation Innovation and Research) conducted a study based on data gathered

during a one month time period. Results suggested differences in freight and general traffic characteristics with slightly higher freight travel times and slightly lower reliability. CDMsmith [21] [22], a private engineering solutions firm, presented a study for Oklahoma DOT about using NPMRDS data for analyzing road traffic congestion.

To date, all related, published work had relied on data imputation with no filtering or a process for outlier removal for the NPMRDS specific domain. The University of Wisconsin-Maddison Traffic Operations and Safety (TPOS) [23] however, introduced early work addressing filtering the dataset. Researchers identified outliers with a negligible effect on summary statistics and recommended scanning the dataset for observations several standard deviations above the mean that occurred throughout the analysis period. In July 2015, the University of Washington (in collaboration with the state of Washington DOT) published a more comprehensive report for computing freight performance measures characterized by outliers [24]. Three primary limitations to the NPMRDS dataset were the impetus for researchers to recommend data pre-processing by eliminating speeds below 2 mph, resetting all speeds above the speed limit to the posted speed limit, and implementing an epoch correction phase to reset epochs based on the value of the consecutive epochs of the same segment. Researchers also reported that segments longer than one mile resulted in data that were less accurate and that optimum results are found in segments one mile in length and less. In February 2016, the university of Wisconsin-Maddison published a guidebook for freight transportation planning using truck GPS data [25]. A section of the study included data for one month from the FHWA's NPMRDS dataset, which was used to compute freight mobility and speed variations along Minnesota's NHS. The Upper Midwest Reliability Resource Center maintains an online

Travel Time Reliability Reference Manual [26] where NPMRDS data is compared with probe data from INRIX. Results indicate NPMRDS data has a lower mean for travel time with a higher variance than data from INRIX.

Several academic research communities have developed tools based on NPMRDS probe data. The University of Wisconsin's developed a traffic tool for Wisconsin DOT that featured an interactive map of the interstate system based on NPMRDS data [27]. A working prototype, operations coordination mapping application, namely "The Interstate Mobility Performance Scanning Tool" (IMPST) [28], was developed as part of the Great Lakes Regional Transportation Operations Coalition (GLRTOC), which includes, Illinois Department of Transportation, Illinois State Toll Highway Authority, Indiana Department of Transportation, Indiana Toll Road Concession Company, Iowa Department of Transportation, Kansas Department of Transportation, Kentucky Transportation Cabinet, Michigan Department of Transportation, Ministry of Transportation Ontario, Minnesota Department of Transportation, Missouri Department of Transportation, Ohio Department of Transportation, Skyway Concession Company, and the Wisconsin Department of Transportation. Also, researchers at the University of Maryland at the Center for Advanced Transportation Technology (CATT) laboratory have developed the Regional Integrated Transportation Information System (RITIS), which is an automated data sharing, dissemination, and archiving system that includes many performance measures that are available for agencies use. The CATT Laboratory operates three independent data centers. Most data centers are used, in part, to collect and archive nearly 60 incoming transportation data feeds from agencies across the

country, one of which is the NPMRDS dataset. The RITIS website allows registered public safety and DOT employees to view real-time RITIS data in a browser.

Tools and services offered in the industry sector include HERE-based services such as HERE Real Time Traffic Services [29], INRIX, which provides roadway congestion information in real time and claims to report accurate real time traffic conditions; and Iteris [30], which offers a range of services and software that includes arterial, freeway, and transit route online traffic monitoring tools. Iteris also offers a software solution called IterisPeMS, which is a performance management system for transportation networks. TomTom is another traffic index provider offering traffic congestion information about traffic jams and accidents occurring during rush hour, as well as telematics, maps, and location-based services. The tool relies on data collected from its network of users. Privately owned companies are also beginning to provide solutions for using NPMRDS data.

Nevertheless, an online investigation has proven that few DOT agencies are utilizing the NPMRDS dataset due to the sheer volume of records, which requires big data analytics capabilities. Also, there is significant complexity associated with analyzing and visualizing the datasets in a meaningful way. Although the FHWA provides reports that utilize travel time data from the NPMRDS dataset [31] (e.g., Urban Congestion Report (UCR)), reports are produced on a quarterly basis and reflect only the collective congestion trend of each state. State DOT agencies have been left to their own to develop tools for investigating a more detailed view of intrastate highway conditions, for analyzing types and locations of congestion, and finding methods for mitigating the effects. Previous work has indicated that the shorter the roadway segment, the more

accurate the NPMRDS data. In many cases, however, this finding was contrary to results presented herein for the current NPMRDS data. In fact, shorter segments exhibit an unknowingly problematic anomalous data, as will be shown in subsequent chapters of this thesis. Furthermore, the notion of congestion expanding both in time and space renders scanning for congestion in only the same segment insufficient, as travel times over roadways follow trajectories spanning consecutive segments over time. In short, scanning must be performed for both the selected and the segments subsequent to the selected segment. Thus, further research is required to formulate correct processes for filtering the dataset prior to using it in reliability reports, as the presence of outliers greatly affects results accuracy.

1.5. Contribution of thesis

This thesis presents research detailing the use of the NPMRDS for computing performance measures in the state of Oklahoma. Data extraction, preprocessing, and statistical exploratory data analysis were performed on the NPMRDS dataset. Baseline historical raw calculations of road segment speed average (including outliers), variance, and standard deviation (STD) across various time scales are shown. A comprehensive study of NPMRDS data characteristics and influencing variables that affect probe data measurements (e.g., segment length, road geometry, and other external factors on speed data) is presented. A process for identifying anomalies is developed, and recommendations for improving accuracy and alleviating data anomalies are reported. Moreover, a process for filtering and removing speed data outliers across multiple road segments is developed, and comparative analysis of raw baseline NPMRDS speed data and cleansed data is presented. Identification and computational comparison of free flow

speed, 85th percentile, and travel time reliability performance measures were computed using both raw and cleansed datasets. A method for improved congestion detection was also investigated and presented. Finally, traffic analytics using machine learning was performed to identify and to classify congested segments. A novel approach for identifying non-recurrent congestion sources using speed data was also developed and introduced.

The main contributions of this thesis are summarized below

- This work applies traffic data analytics, statistical analysis, and machine learning to the NPMRDS to develop models, tools, filtering processes, and performance measures enabling agencies and other users to characterize, understand, and gain insight into actual traffic patterns of NHS roadways using the dataset.
- To the author's knowledge, this work includes a first-of-a-kind analysis incorporating an adapted version of Benford's law, developed to detect inadvertent anomalous data generated in the dataset. Furthermore, models are formulated that alleviate and remove these anomalies.
- This work presents a step-by-step process for filtering and removing outliers from the NPMRDS dataset. The process is highly beneficial for agencies and researchers interested in working with the NPMRDS dataset.
- A novel approach is introduced for identifying non-recurrent congestion sources that affect roadway segments. The proposed method can promptly respond to changes in traffic patterns, proving it is suitable for implementing real-time detection technology.

The balance of this thesis is organized in the following manner. The next chapter presents the framework and tools utilized for NPMRDS data acquisition and preprocessing. It also provides information for a necessary understanding of the unique characteristics of the NPMRDS data, with a focus on challenges associated with probe data. Chapter 3 presents a core process for detecting anomalies/outliers in the dataset and develops models for alleviating said anomalies. An inclusive process for filtering outliers, which caters to the NPMRDS domain, is presented. Chapter 4 is devoted to statistical exploratory analysis of the dataset. A qualitative comparative analysis of both raw and cleansed datasets is presented to aid in determining the effect of outlier removal from final results. The chapter also includes an improved approach for detecting congested segments. Reliability performance measure computations follow in Chapter 5, wherein Free flow, 85^{th} percentile, travel time index, buffer index, and planning time index are identified and computed—separately for each segment and collectively for the overall highway. Chapter 6 presents traffic data analytics applied to the dataset via clustering and classification using a combination of unsupervised and supervised learning techniques. A novel solution for congestion identification is demonstrated at chapter end. Finally, summary findings and a conclusion are presented in Chapter 7. Future directions for research are suggested, as well.

Chapter 2: NPMRDS Acquisition, Characteristics and Processing

NPRMDS data contains travel times for all NHS roadways, including those in the state of Oklahoma. This chapter provides information necessary to develop an understanding of the framework required for processing data collected from one particular interstate highway in Oklahoma, namely Interstate 35 (hereafter, I-35). This chapter discusses limitations and challenges associated with utilizing NPMRDS data. Such information is necessary to arm the reader with knowledge about specific features of this data domain. Once necessary tools and a framework are developed, they can be extended to collectively process travel times for all NHS roadways in Oklahoma.

2.1. Dataset acquisition

Data records were obtained from ODOT following the successful collection of NPRMDS data files from a shared FHWA repository accessible only by state DOTs and MPO agencies. The dataset was composed of large files with the naming convention "FHWA_TASK_201x_xx_OK_TT," where marked x's represent the year and month of data collection. Travel times were recorded monthly per segment on NHS roadways. Figure 3 depicts Oklahoma's NHS roadways and illustrates locations at which travel time data is captured. Figure 4 highlights the three interstate highways which form a crossroad in Oklahoma. According to NPMRDS static file, NHS roadways in Oklahoma are composed of 4,323 defined segments, each generating one epoch every five minutes, which is equivalent to 288 epochs per day, per segment. These figures scale to approximately 1,245,024 records per day, and 448,208,640 records annually. Nationwide, 282,402 segments generate 81,331,776 records daily, which scale to approximately

29,279,439,360 annually. Figure 5 shows NHS roadways for all 50 states, including Puerto Rico [32].

Figure 3 - NHS roadways in Oklahoma.

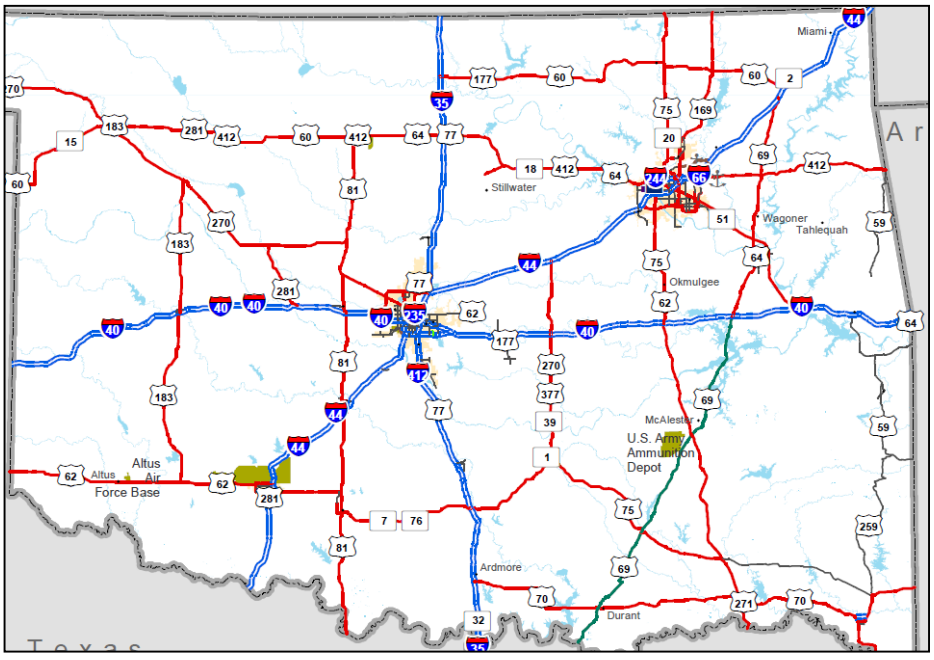


Figure 4 - NHS roadways in Oklahoma magnified.

Figure 5 - NHS for all states.

The amount of travel time data records recorded inhibits the ability of using typical desktop software, which most public agencies rely on, for processing. Handling the files requires knowledge of, and access to, more advanced database or statistical analysis tools.

2.2. Hadoop environment and data extraction

Apache™ Hadoop® is a popular open source tool that enables distributed processing and manipulation of large data sets across clusters of commodity servers [33]. The software is highly scalable from a single server to thousands of machines, with an extremely high degree of fault tolerance. Accordingly, a five-node Hadoop setup was constructed for data pre-processing on the large sets of NPMRDS data. See Figure 6.

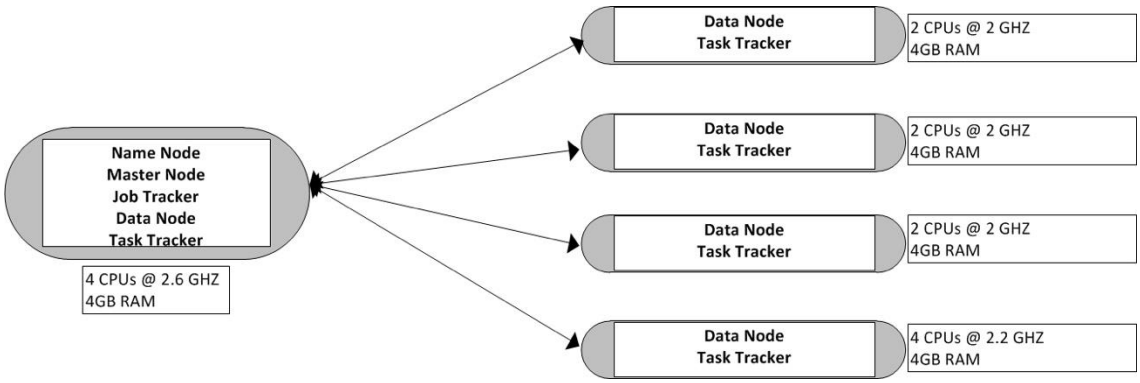


Figure 6 - Illustration of analytics lab 5 node Hadoop setup.

Processing using Hadoop commences with the copying of travel time files from the PC computer to the Hadoop NAMENODE server. Uploading data to the Hadoop File System (HDFS), and then storing it as an accessible, query-able file on the cluster, allows manipulation and processing of data in any order. In turn, the user is granted the flexibility to quickly and efficiently extract a particular record according to a predefined criterion from the millions of available records. The following steps are necessary to achieve this task:

1- Create a new directory in the HDFS to save the files in the Hadoop cluster.

```
hadoop fs -mkdir /user/hadoop/NPMRDS/2014
hadoop fs -ls /user/hadoop/NPMRDS/
```

2- Copy the data to the HDFS

```
hadoop fs -copyFromLocal ~/NPMRDS/*2014*.CSV /user/hadoop/NPMRDS/2014
hadoop fs -ls /user/hadoop/NPMRDS/2014
```

3- Check the contents of a data file using the below command:

```
hadoop fs -tail /user/hadoop/NPMRDS/test/testdata.csv
```

Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis [34]. Apache Hive supports analysis of large datasets stored in Hadoop's HDFS and provides a Structured Query Language (SQL)-like language, namely HiveQL, with schema on read to transparently convert queries to

map/reduce. HIVE was used to query the datasets in HDFS and execute desired map/reduce queries. HIVE-generated customized query commands necessary for the work in this thesis are shown below:

1- Create a searchable internal container for the NPMRDS data

```
CREATE TABLE sampletest_2015(col_value STRING);
LOAD DATA INPATH '/user/hadoop/NPMRDS/2015' OVERWRITE INTO TABLE
sampletest_2015;

CREATE TABLE NP_2015(bef int, aft int, let string,month int, day int, year int, TMC
string, DATE int, EPOCH int, TravelALL int, TravelPass int, TravelFre int);

INSERT OVERWRITE TABLE NP_2015
SELECT
regexp extract(col value,'([0-9]+)[A-Z]')
bef,
regexp extract(col value,'[A-Z]([0-9]+)')
aft,
regexp_extract(col_value,'([A-Z]+)')
let,
regexp_extract(col_value,'[,]([0-9])[0-9]+[,]')
month,
regexp_extract(col_value,'[,][0-9]([0-9][0-9])[0-9]+[,]')
day,
regexp_extract(col_value,'[,][0-9]+([0-9][0-9][0-9][0-9])[,]')
year,
regexp_extract(col_value,'([0-9A-Z]*)[,]')
TMC,
regexp_extract(col_value,'[,]([0-9]*)[,]')
DATE,
regexp_extract(col_value,'([0-9]*)\,([0-9]*)\,([0-9]*)\,([0-9]*)$')
EPOCH,
regexp_extract(col_value,'([0-9]*)\,([0-9]*)\,([0-9]*)$')
TravelALL,
regexp_extract(col_value,'([0-9]*)\,([0-9]*)$')
TravelPass,
regexp_extract(col_value,'\,([0-9]*)$')
TravelFre
from sampletest_2015;
```

2- Query for Oklahoma i35 TMC's, Southbound, in January 2015.

```
CREATE TABLE i3512015(TMC string, DATE int, EPOCH int, TravelALL int, TravelPass int,
TravelFre int);

INSERT OVERWRITE TABLE i3512015
SELECT
TMC,
DATE,
EPOCH,
TravelALL,
TravelPass,
TravelFre
from np 2015 WHERE bef= "111" AND let="N" AND DATE< 2000000 AND ((aft<5638 AND
aft>5619)OR(aft<4932 AND aft>4894)OR(5144<aft AND aft<5160)OR(5481<aft AND
aft<5505)OR(5398<aft AND aft<5404));

hadoop fs -cat  /user/hive/warehouse/i3512015/000000_0 > ~/Results/i3512015
scp Results/i3512015 nbitar@156.110.167.57:~/Dropbox
```

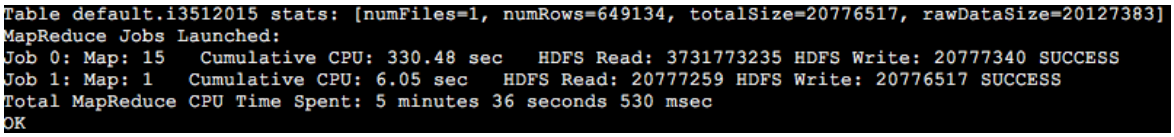


Figure 7 - Output of Hive.

Figure 7 shows Hadoop final output after the map/reduce execution is complete. At this stage, required data had been extracted and rearranged into segment–travel time) matrices. Adequate statistical processing requires a thorough understanding of the characteristics of the data. Accordingly, the following subsection investigates the availability, attributes, and limitations of the NPMRDS dataset, and, in particular, illustrates examples for I-35 southbound.

2.3. Dataset characteristics: challenges and limitations

As aforementioned, NPMRDS data is based on instantaneous GPS data records obtained from vehicles that carry GPS devices reporting location and speed to HERE and ATRI, [19] [23], [24] and [22]. Combined travel time measurements reported in the NPMRDS dataset are computed as a weighted average of both recorded passenger and truck travel times according to the number of available probes for each. However, actual volume of each vehicle type is not reported by HERE/ATRI. Understanding the nature of the NPMRDS dataset is key for effective data post processing (e.g., anomaly and outlier detection, as well as measures for their removal). Challenges and limitations are enumerated below:

2.3.1. Size of the data:

The monthly, HERE-generated NPMRDS dataset size is large. Moreover, the number of records generated per segment for each highway renders conventional tools,

such as Microsoft Excel, ineffective for post processing. Any given typical month can generate data in the order of 30 to 40 million records. This number far exceeds the one million record capability of Excel. Thus, working with NPMRDS data requires database and scripting expertise [23].

2.3.2. High spatial-temporal probe and record data variability:

NPMRDS probe data is based on a variable number of available probes and resulting records generated at any segment location. Data varies considerably depending on time of day and day of the week. Also, variance in the spatial domain is due to variance in the number of probes between consecutive segments at any given time of day. Furthermore, variability is dependent upon the number of probes per vehicle type at the same location and the same time (i.e., passenger vehicle vs. truck probes). For example, Figure 8 shows TMC segment (111N04920) located south of Oklahoma City.

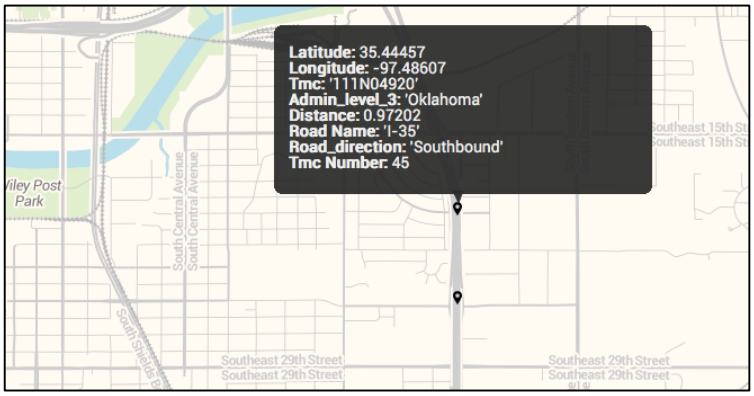


Figure 8 - TMC "111N04920" located south of Oklahoma City.

Figure 9 shows a bar plot for the total number of epochs recorded on TMC segment (111N04920) per day for 31 days during the month of January 2015. Mean value of recorded epochs was 219.5806, and Standard Deviation (STD) was 20.0678. Clearly, the number of epochs for the same segment fluctuates daily.

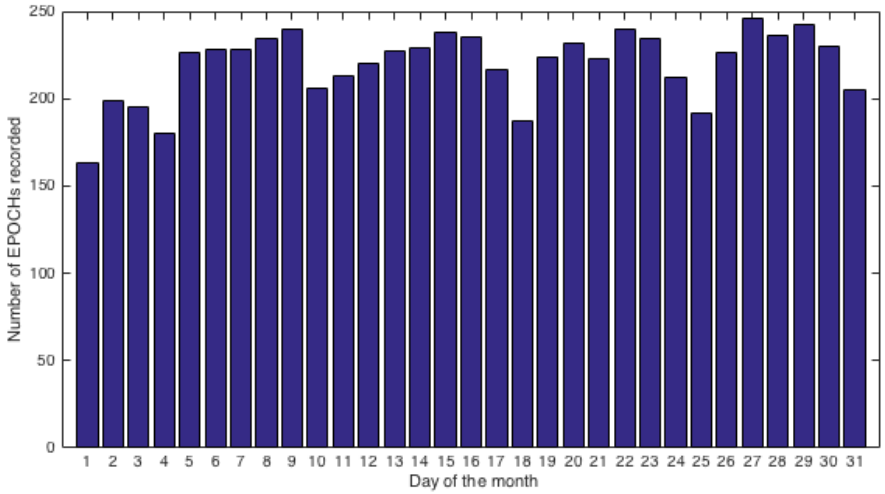


Figure 9 - Daily bar plot of epochs recorded for TMC 45 during Jan 2015.

Figure 10 details the difference in epoch count per day for two bordering segments. For TMC 46, mean was 184.0968 epochs and STD was 24.2918. Epoch count variance between both segments is considerable.

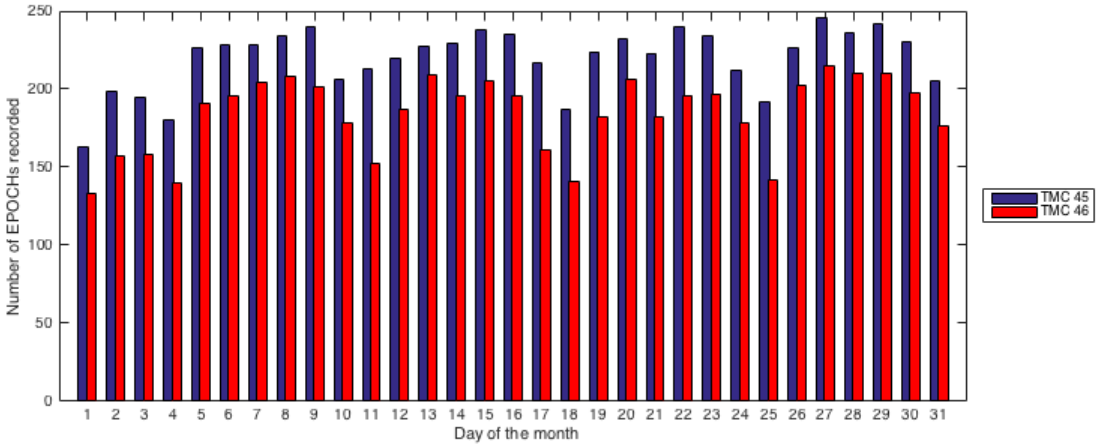


Figure 10 - Bar plot of epochs recorded for segments 45 and 46 during Jan 2015.

Variance per day relative to three time groupings is as follows. Group 1 is indicated by morning hours from 12 a.m. to 8 a.m.; Group 2 indicates afternoon hours between 8 a.m. and 4 p.m.; and Group 3 represents the evening hours from 4 p.m. to 12 a.m. Group 2 (i.e., afternoon) generated the greatest number of epochs; the least number of epochs were

generated during the evening. Table 3 illustrates the mean over 31 days per group for segment TMC 45.

Table 3 - Probe epochs available per time of the day for segment 45.

Group	Group (1): 12am – 8am	Group (2): 8am – 4pm	Group (3): 4pm – 12am
Mean	56.3508%	93.9180%	78.4610%
STD	8.8708	6.0338	6.8185

When inspecting the number of epochs recorded per vehicle type per day, a difference between probe types was evident. As count per probe type varies, the combined travel time computed as the weighted average is highly influenced. Table 4 shows the mean percentage of epochs per probe type, as well as the percentage of combined travel time mean.

Table 4 - Mean number of epochs per probe type for segment 45.

Group	Combined	Passenger Vehicles	Trucks
Mean	76.2433%	57.1909%	56.5076%
STD	20.0678	30.4961	19.5703

Given the average across all segments of highway I-35, we get similar results, as shown in Table 5 and Table 6.

Table 5 - Probe epochs available per time of the day for i-35 (98 segments).

Group	Group (1): 12am – 8am	Group (2): 8am – 4pm	Group (3): 4pm – 12am
Mean	58.1135%	87.8185%	76.6424%
STD	8.6746	4.4879	5.8671

Table 6 - Mean number of epochs per probe type for i-35 (98 segments).

Group	Combined	Passenger Vehicles	Trucks
Mean	74.1915%	49.8046%	60.9439%
STD	16.4836	25.1715	16.4760

2.3.3. Missing data:

A special case of spatial and temporal variance in epochs was reported for segments per probe type when probe data was unavailable for any type of vehicle. The result is a gap in travel time, as HERE fails to generate any record data for such special cases. This phenomenon was evident on certain rural NHS roadways in Oklahoma when probe number was very low on average and resulted in an extremely small number of epochs. The result was large data gaps for several hours, which made characterizing travel time for a particular segment highly skewed. This problem was found to a lesser extent on interstate highways and large arterial roadways, where the number of probes is higher on average. A comparison between the number of epochs generated on I-35 during January 2014 and January 2015 can be drawn by looking at Table 7 and Table 8, the number of probes increased for both types of vehicles, particularly for trucks. This phenomenon is reflected in an increase in combined travel time epochs, from 54% to approximately 73%.

Table 7 - Number of epochs recorded per probe type.

Group	Combined	Passenger Vehicles	Trucks
January 2014	481338	388040	234403
January 2015	649134	435762	533225

Table 8 - Percentage of total epochs per probe type.

Group	Combined	Passenger Vehicles	Trucks
January 2014	53.913306%	43.463262%	26.254816%
January 2015	72.707661%	48.808468%	59.725022%

2.3.4. Bias toward Lower speeds:

Travel time data in NPMRDS is probe data based on GPS records reported at fixed rates of time. Hence, the slower the probe vehicle speed, the larger the number of samples generated as the vehicle travels the length of the roadway segment. Consequently, a slow

vehicle will report more records than a fast vehicle. Since travel time reported for a segment is the average of all probe travel times calculated during a fixed time period and since slow moving vehicles report a higher number of records, average travel time is biased toward slower moving vehicle speeds. This limitation can be overcome by implementing a weighted average, where each vehicle is weighted according to the number of samples generated prior to computing travel time average of the segment. Doing so increases data collection complexity, but it also eliminates the effect of bias toward slower moving vehicles.

2.3.5. Variability of segment lengths:

TMC segments defined for use in NHS roadways vary considerably in length. This variability entails several effects on travel time reliability and measurement accuracy. On one hand, shorter segments exhibit a smaller number of samples. Figure 11 illustrates Oklahoma I-35 southbound between the Kansas and Texas borders, per segment, per day. Several factors are at play, one being that the shorter the length of the segment, the less the density of vehicles contained in any unit of time. Moreover, because probe vehicles traverse the length of a short segment faster than they do a long segment, they generate a smaller number of samples in the shorter segment. In some cases, it is possible that probe vehicles could pass through an entire segment without reporting any record, especially if the sample time for instantaneous data being reported is larger than the time required to traverse the segment.

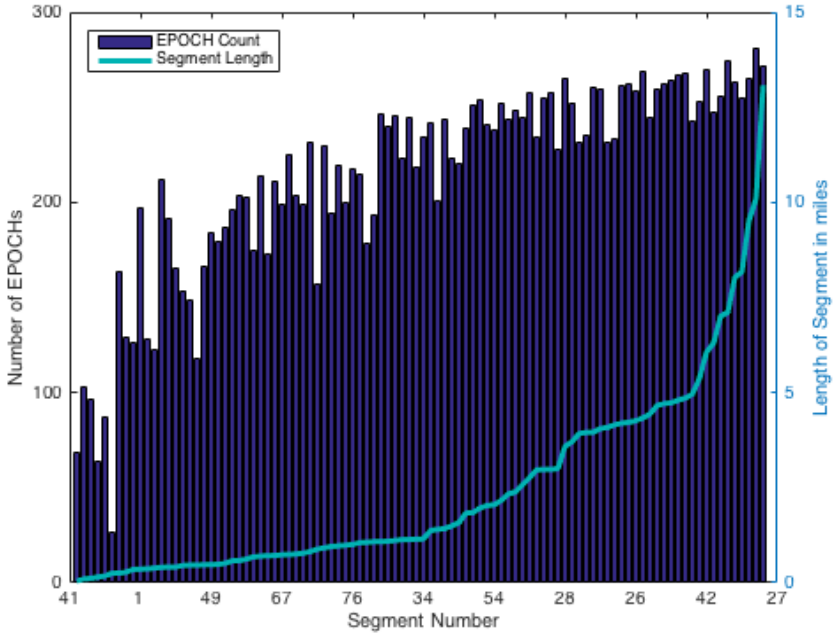


Figure 11 - Trend plot for number of epochs recorded versus length of segment.

Consequently, this affects the number of samples recorded per segment for any roadway. Figure 12 illustrates the variability of average number of epochs recorded per day for I-35 southbound.

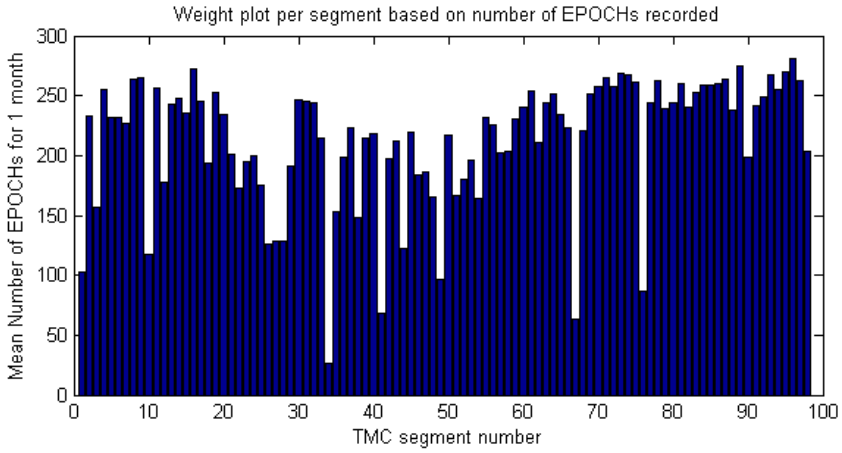


Figure 12 - Average number of epochs recorded per day reported per segment.

Long segments could experience different travel times across different parts of the segment, rendering average travel time an inaccurate representation of actual travel time across the entire segment.

2.3.6. Vehicle performance and roadway geometry effect:

In particular cases, truck-reported travel times were higher than passenger vehicle-reported travel times. Inversely, this means that trucks traveling those particular segments are moving slower on average than passenger vehicles. Truck-reported travel times are prone to what is known as the Power-to-Weight ratio model [13], [24], which adversely affects truck speed. Trucks with heavier cargo tend to slow their speed for precautionary measures. In addition, traversing steep or elevated roads could also cause trucks to reduce their speeds. In such cases, reported travel time would model vehicle performance or roadway geometry characteristics rather than traffic conditions.

2.3.7. Instantaneous speed reporting increases variability:

Given a small number of probes, average speed for all vehicles on the roadway might not be accurately represented by the average of the probe samples. Moreover, because travel time is derived from instantaneous speeds reported by GPS devices, resulting captured values could project higher variability than might actually be occurring on the roadway. As vehicles maintain an average speed when traversing a roadway during these periods, it is possible that vehicles might continually increase and/or decrease at speeds above and below that average. Reporting instantaneous speeds results in travel time variation that might indicate variation that is different from that actually occurring on the

roadway. Figure 13 illustrates the variation in speed for segment 45 for one entire, non-congested day. Clearly, there is significant variation between each consecutive epoch.

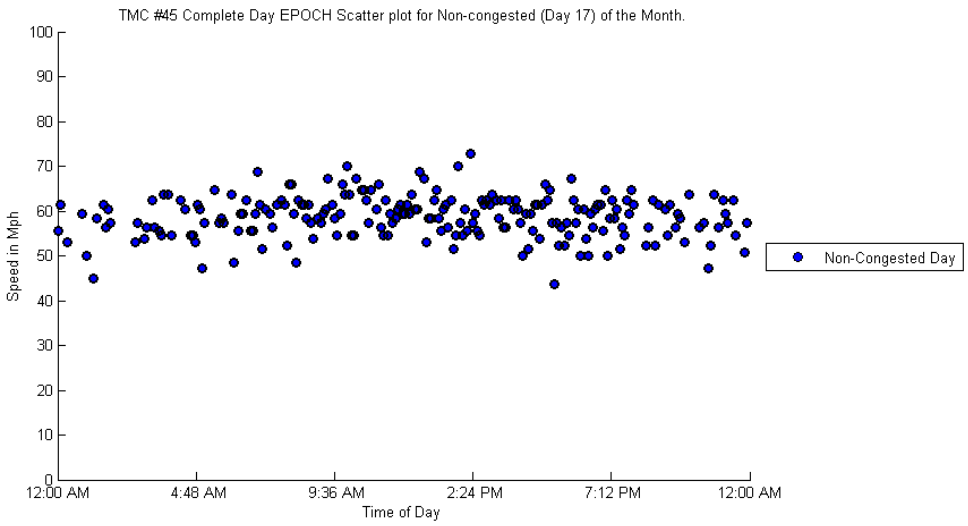


Figure 13 - TMC 45 complete day epoch scatter plot for non-congested day of January.

2.3.8. GPS in-acccuracy:

In some cases, GPS coordinates of NHS roadways could match coordinates of non-NHS roadways. Consequently, vehicles traveling non-NHS roadways could be mistakenly accounted as those traveling NHS roads and, as a result, distort collected travel time measurements. For example, bridges, tunnels, and parallel roadways cause NHS and non-NHS roadways to be located at the same geographical coordinate. If directionality is not provided or if the accuracy of GPS positioning is not precise, a traveler can easily be mistaken on an NHS roadway, even though he/she is actually traveling a non-NHS roadway, adjacent or near the NHS road. At an intersection, GPS location is associated with directionality, thus the error can be detected. Ultimately, the result of miscounted data is an increase in the variability of road travel times.

Figure 14 shows TMC 47 characterized by 0.5m of roadway crossing SE Grand Blvd road, which happens to be a major arterial. The satellite view depicted in Figure 15 shows that the NHS passes under the roadway. If directionality was not reported as a function of GPS measurement, vehicles on SE Grand Blvd could be miscounted as traveling I-35. Figure 15 also shows two parallel non-NHS roadways adjacent to I-35 southbound and northbound. If GPS positioning is not completely accurate, an erroneous count is possible as a result of vehicles traveling on either road.

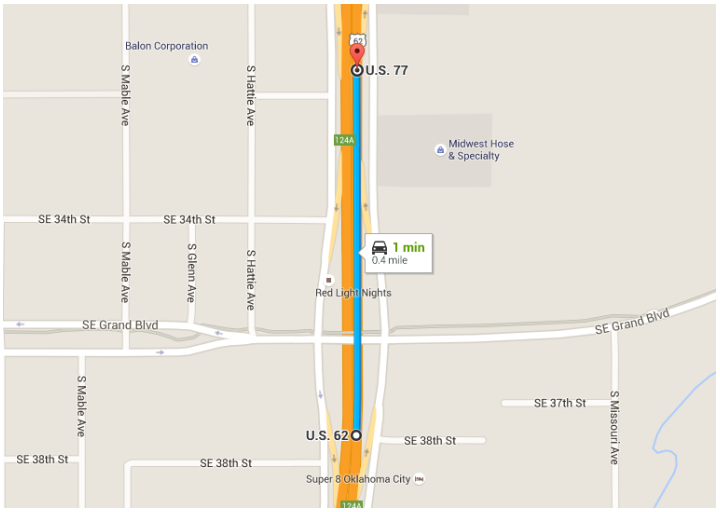


Figure 14 - Map view of TMC 47 crossroads with a major arterial.

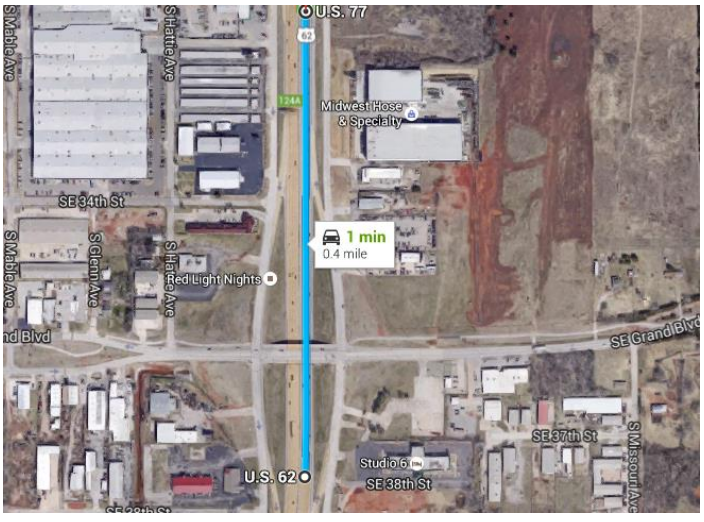


Figure 15 - Satellite view of TMC 47 cross with a major.

Chapter 3: Dataset Cleansing: Anomalies and Outliers

In the previous chapter, limitations and challenges inherit in the NPMRDS dataset were described and discussed. Despite the aforementioned limitations, the NPMRDS dataset has important advantages that make it a valuable tool for crafting traffic performance measures. For example, because NPRMDS is a probe data set, travel times can be easily collected from different geographic regions. Compared to traditional fixed location detectors, NPRMDS data has higher granularity without the confines of location or forced infrastructural physical constraints. Moreover, NPRMDS data is continuously generated, enabling DOT agencies to look beyond separate periodic surveys of unusual highway conditions. However, capturing this information requires developing the right tools to extract, manipulate, and process NPRMDS data. A thorough understanding of the domain characteristic is necessary for accurate and effective statistical processing. Accordingly, the aforementioned limitations serve as guidelines for further anomaly detection and outlier removal procedures. These accommodations are presented in the next sections.

A report published by CDMSmith—a private consulting company shows a procedure reportedly adopted by HERE (Provider of the NPMRDS) for dataset validation and quality assurance. a summary of which is shown in Figure 16. Details of this can be found in [22].

Speed records acquired by HERE and ATRI can be affected by anomalies and outliers, which collectively affect the accuracy of travel time reported in NPRMDS, as well as other performance measures that rely on travel time accuracy. See Figure 17. In short, the study begins analyzing data anomalies present in NPMRDS data, and then further

presents recommendations to alleviate and remove them. Moreover, the study continues to address outliers present in the dataset, offering suitable techniques to detect and remove outlier points from the data.

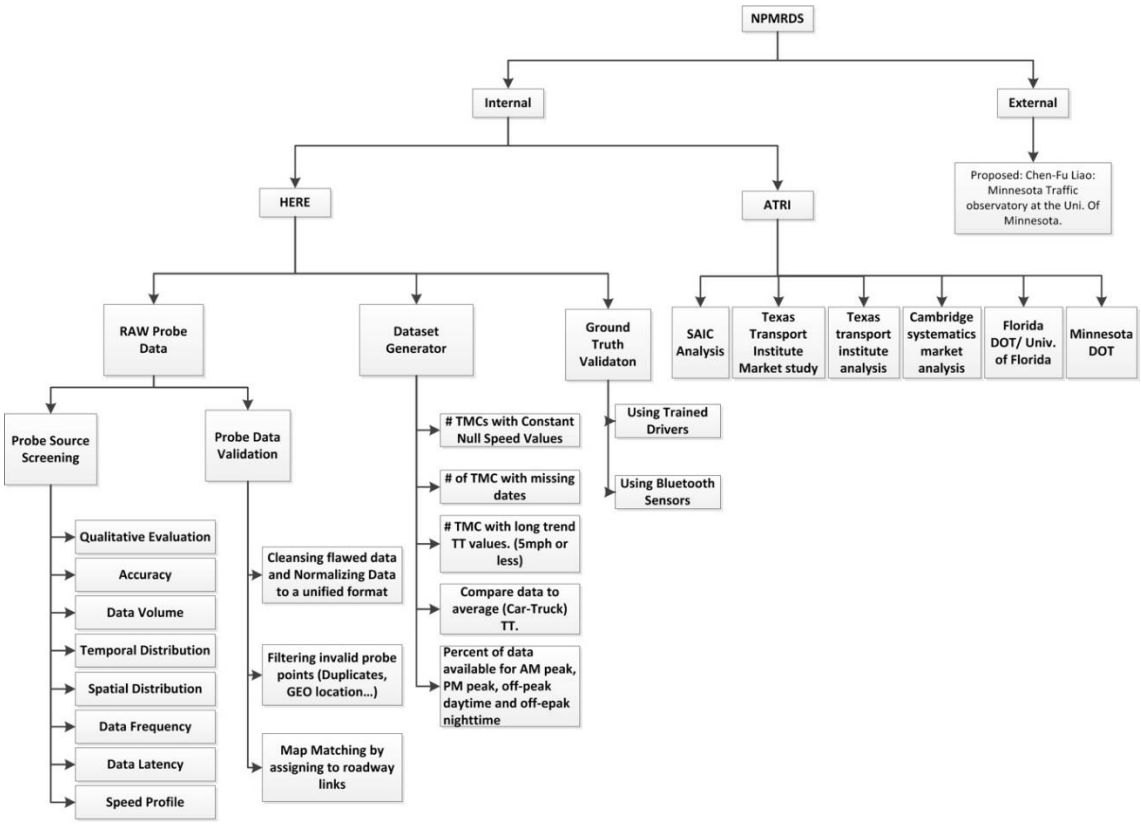


Figure 16 - NPMRDS procedure for probe data validation and quality assurance

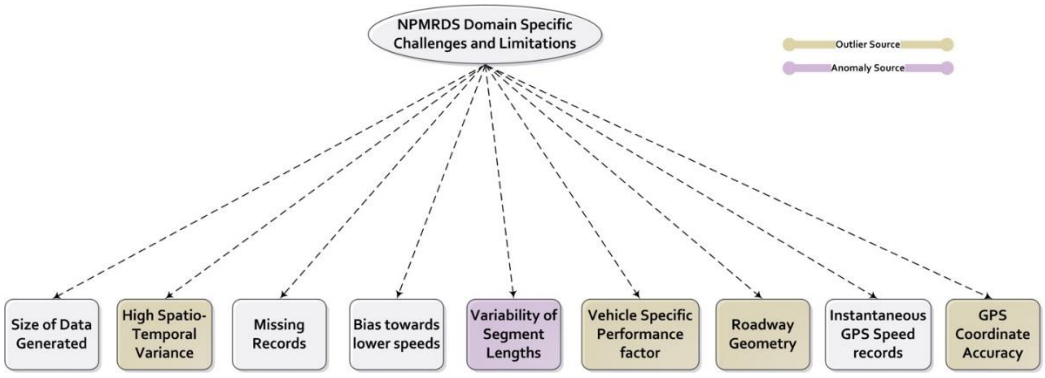


Figure 17 – Summary of limitations generating outliers and anomalies in the NPMRDS

3.1.Data Anomalies:

Data anomalies refer to erroneous, illegitimate data points present in a dataset that are caused by pre-processing, incorrect filtering, or other external processes or procedures irrelevant to the phenomena under measure. Anomalies threaten statistical soundness of a quantitative dataset.

A prominent approach for evaluating statistical soundness of a quantitative dataset commonly applied in forensics and admissible in U.S. courts, is to check the digit distribution of a measured quantity. This stems from a famous law described by Benford in 1938 [35], and proved mathematically by T. Hill in 1995 [36]. Benford's law is applicable to occurrences of natural events [37]. Simply stated, it is the principle that in any large, randomly produced set of natural numbers, there exists an expected distribution for digits in numerical data that deviates from the uniform, commonly known as Benford's distribution. One limitation for this law is when a digit is capped by a maximum or minimum. Nevertheless, applying a similar approach, as a digit count process for the second digit of the speed converted time data recorded gives an understanding of the statistical distribution of measured speeds and provides insight to the statistical soundness of the data. Then, taking the variance of the distribution, instead of the actual histogram values, yields a prominent indicator for the occurrence of natural randomness in the events. The significance of this test is that the variance of the digits will not be heavily affected by sample outliers that might occur in particular days due to external factors such as weather, incident, or other causes. On the contrast, taking speed opposed to digits as a measure would be heavily influenced by such outliers in any variance measurement.

Consider a vector used to represent a set of measured speeds for consecutive vehicles traveling on a road. Let ψ_1=[71 62 73 64 67 29 65 68 66] be the vector. Statistical analysis demonstrates that vector speed has a mean of 62.77 mph and a variance of 171.994. These are an inadequate indicator for anomalies. In the example, high variance was the result of a recorded outlier speed of 29 mph. Intuitively, speeds such as those reported in the vector could be expected for consecutive vehicle speeds, as they tend to be random in nature. However, the proposed distribution digit test for this same vector has a variance of zero, mainly because each second digit occurred only once. In this way, the test indicated that in spite of the outlier, data was not anomalous because recorded samples were random enough to represent actual natural occurrence. Given ψ_2=[65 65 65 65 65 65 65 65 64], it is logical to assume the probability of eight consecutive vehicles traveling the exact speed is highly unlikely. Applying the speed variance statistical test results in a very small variance of 0.11, which inconclusively indicates vector speed data is natural. On the other hand, the proposed digit variance test reports a variance of 7, indicating the data exhibits abnormality in speed records recorded.

Accordingly, a matrix of second digit distribution per segment for I-35 Southbound was constructed. Normalized variance was computed, and variance versus segments with decreasing length was plotted. The variance of Benford's law for the second digit was calculated and can be found in [38] equal to 0.0011. Figure 18 illustrates the results with the Benford variance plotted in red. Clearly, a trade-off exists between segment length and the variability of second digit distribution. In other words, as segment length is reduced there exists a higher repetition in recorded consecutive speed. This means that recorded samples tend to deviate from the randomness expected in any natural occurrence.

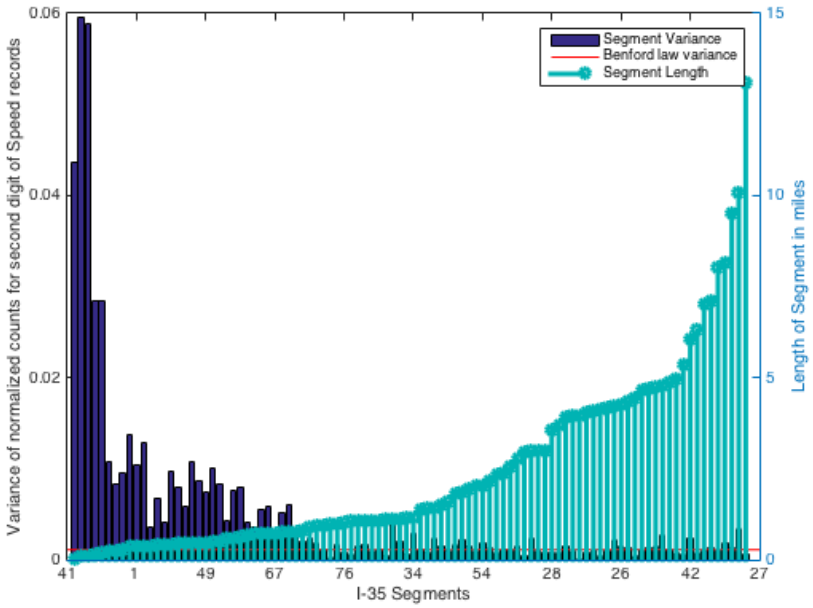


Figure 18 - Variance between percentages of digits vs length of segment on I-35.

This fact gives insight that the NPMRDS data contains anomalous entries being generated by HERE unknowingly. The reason we say unknowingly, is that we are sure the process is of natural occurrence and should always exhibit the random statistical soundness all natural occurrences generate. This is not the case in the NPMRDS data for smaller segments as Figure 18 shows. Further investigation reveals the cause of this anomaly. The reason is an inherent trade-off between segment length, system time granularity and the speed of vehicles traveling the segment. Assume a segment is of length 0.0426 miles. If the vehicle were traveling at the speed limit of 65 mph, it should traverse the entire segment in 2.3627 seconds. Because HERE reports epochs with a time granularity of integer seconds, the value will be rounded to 2 seconds, effectively translating speed to 76.6 MPH. Furthermore, if a vehicle were traveling slower than 65 mph, for instance 62 mph, then that time would be rounded to 3 seconds, effectively translating speed to

51.1920 mph. Thus, the range of actual speed suffers from a quantization error when reported. The error quantifying the range of ambiguous speeds, including actual vehicle speed measured, will hereafter be referred to as the Error Range ($\boldsymbol{E_r}$) of speed for a particular segment. $\boldsymbol{E_r}$ for the example described above is 40 mph. According to the theory, speeds between 62.3 and 102 mph would be rounded off to 76.6 mph. The ramifications of this on accuracy and reliability are severe. Figure 19 shows such effects on segment 41, which has a length of length 0.0426 miles. By plotting measurements in the NPMRDS data, it is clear that exactly 2 speeds were reported.

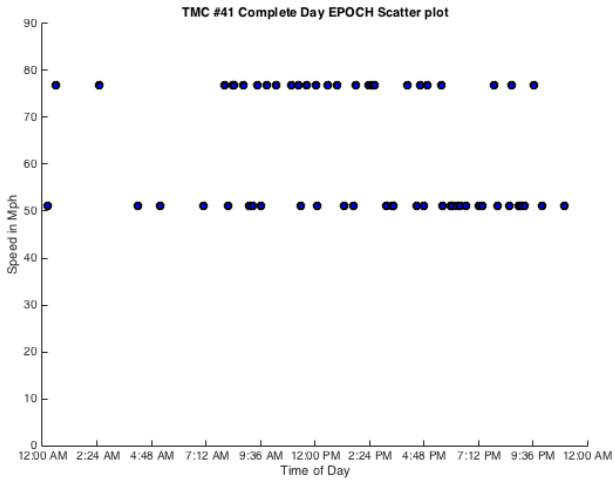


Figure 19 – Segment 41 daily epoch plot

Accordingly, interaction between time granularity and segment length should be modeled to provide E_r for reported vehicle speed, given segment length and reported time granularity of the system.

Let E_r represent the Error range for any given segment of length D at speeds V_i where $i \in \{1,2,3, \ldots.\}$. E_r encompasses all speeds that when rounded due to time granularity, report identical time. Thus, the difference between two speeds that yield the same time can be expressed as.

$$V_2 - V_1 = E_r$$

$$= 3600\left[\frac{D}{T_{time} - \frac{T_{gran}}{2}} - \frac{D}{T_{time} + \frac{T_{gran}}{2}}\right]$$

$$= D \cdot 3600\left[\frac{T_{time} + \frac{T_{gran}}{2} - T_{time} + \frac{T_{gran}}{2}}{T_{time}^2 - \frac{T_{gran}^2}{4}}\right]$$

$$= D * 3600\left[\frac{4T_{gran}}{4T_{time}^2 - T_{gran}^2}\right]$$

Substituting $\beta = 3600*4 = 14400$, $T_{time} = \frac{D\ (distance)}{S\ (Speed)}$ yields

$$E_r(S, D, T_{gran}) = \frac{D \cdot \beta \cdot T_{gran} \cdot S^2}{\beta \cdot D^2 \cdot 3600 - T_{gran}^2 \cdot S^2} \quad (1)$$

where D is given in miles (M); T_{gran} is the reported time granularity in seconds (s); T_{time} is the travel time reported by HERE in seconds (s); and S is the reported speed of vehicles in mph.

Agencies can utilize equation (1) to validate speed accuracy reported by HERE. Figure 20 plots E_r vs. speed for segment 41.

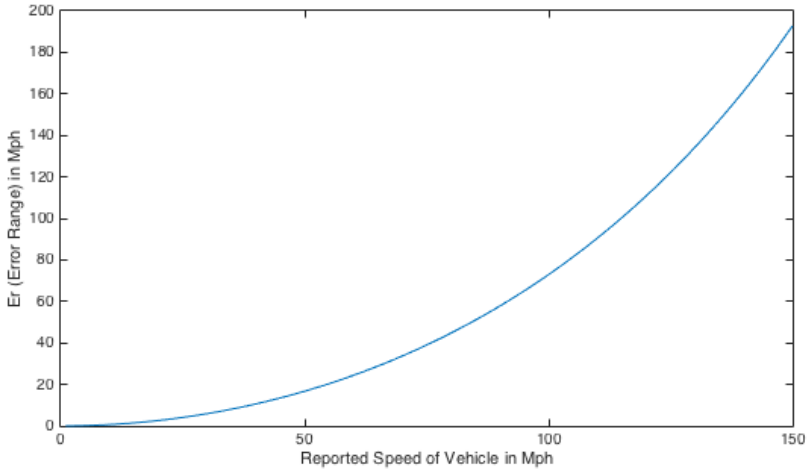


Figure 20 - Plot of vehicle speed vs. error range in mph for Segment 41.

Notably, the faster the vehicle speed, the larger the E_r. Section 41 was identified as the segment with the worst speed accuracy among all sections tested. Vehicles traveling at faster speeds create a larger bin of lumped speeds that confirm the same rounded-off second. Figure 20 demonstrates that even at moderate speeds of 50-60 miles, variance of 20 to 40 mph is possible. Two critical questions and equations to solve them are presented below:

1. Given segment length and maximum speed limit, what is the optimum time granularity for a system to achieve desired E_r? After solving equation (1), executing equation (2) can provide the solution to the question:

$$T_{gran}(D, S, E_r) = -\frac{1}{2}\left[\left(\frac{D \cdot \beta}{E_r}\right) - \sqrt{\frac{D^2 \cdot \beta^2}{E_r^2} + 16 \cdot \left(\frac{D}{S} \cdot 3600\right)^2}\right] \qquad (2)$$

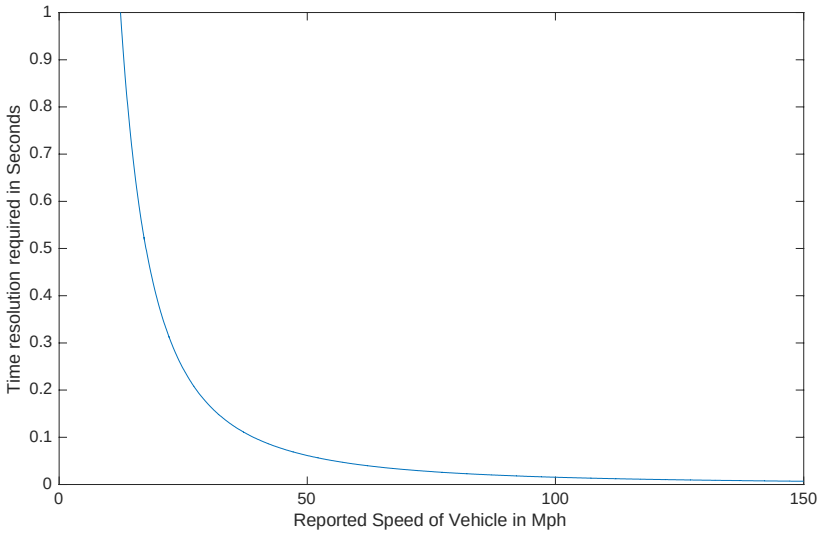


Figure 21 - Plot of Vehicle Speeds vs. Time resolution for Segment 41.

Figure 21 shows a plot diagram for equation 2 for segment 41. Recorded time must be increased to 2 decimal points in order to achieve a 1 mph E_r. DOT agencies are advised to apply this equation to a road according to the highest speeds expected and smallest segment lengths to ensure that any data reported is correct for all segments of any roadway.

2. Given a maximum speed limit and system capability for time granularity, what is the minimum acceptable length of a segment to achieve desired E_r for a particular speed? Equation (3) provides the solution:

$$D\left(S, E_r, T_{gran}\right) = \frac{\beta \cdot T_{gran} + \sqrt{\left(\beta \cdot T_{gran}\right)^2 + \frac{16 \cdot E_r^2 \cdot T_{gran}^2 \cdot 3600^2}{S^2}}}{8 \cdot E_r \cdot \frac{3600^2}{S^2}} \quad (3)$$

The benefit calculating the answer to Equation 3 is twofold. First, for currently deployed systems, engineers are able compute minimal segment length to ensure a desired E_r.

Meaning that they are able to detect the number of segments falling below a threshold E_r and flag those particular segments as less reliable data sources. Second, Equation 3 allows researchers interested in constructing a new travel time reporting system to properly plan placement of capture devices to insure segment length achieves the desired speed accuracy. In short, Equation 3 can be used by DOT agencies and interested parties during the development phase of a system when segment length is a factor.

When applying Equation 3 to Interstate I-35, results show that to achieve E_r of 1 mph, the smallest segment with average speed limit of 65 mph and time-capture granularity of 1 sec must be 1.1736 miles in length. In Oklahoma, there are 50 segments shorter than this distance, meaning that 50 out of 98 segments are affected by this anomaly. Statistical analysis using NPMRDS data in these segments will be affected. Measurements such as detecting free flow speeds, 85th percentile, and others can be skewed by this error. Figure 22 shows speeds recorded for another segment, #91, an example of a segment which is of length 1.373 miles; longer than the minimum distance calculated.

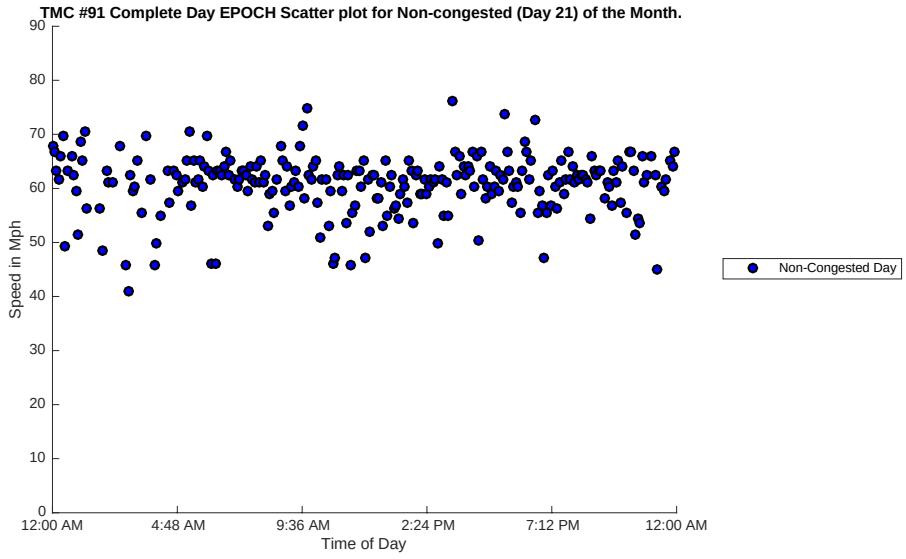


Figure 22 - Segment 91 reported speed scatter plot

We observe the natural occurrence of randomness in speeds to be present in this segment. Moreover, for the purposes of congestion detection, most of the shorter segments can still be used if the extent of quantization error is acceptable at lower speeds, which could indicate congestion. Figure 23 illustrates this effect for segment 49. Applying Equation 1 to a speed limit of 65 mph and time granularity of 1 sec, E_r is calculated at 10.296 mph. The blue scatter plot illustrates the original, uncleansed data points and shows that a step size of approximately 10 mph occurs between 60 and 70 mph as a result of calculated E_r. The step size increases to 13 mph when a vehicle surpasses 70 mph. This error does not come into effect at lower speeds. For example, at a speed of 40 mph, the error becomes 3.89 mph, and at speeds of 30 mph, the error reaches 2.19 mph. Thus, congestion detection algorithms could be applied at speeds of 40 mph and below.

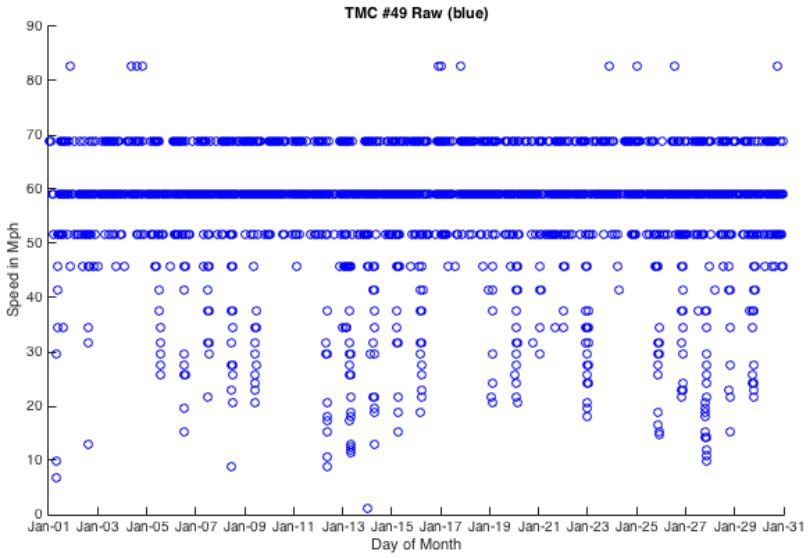
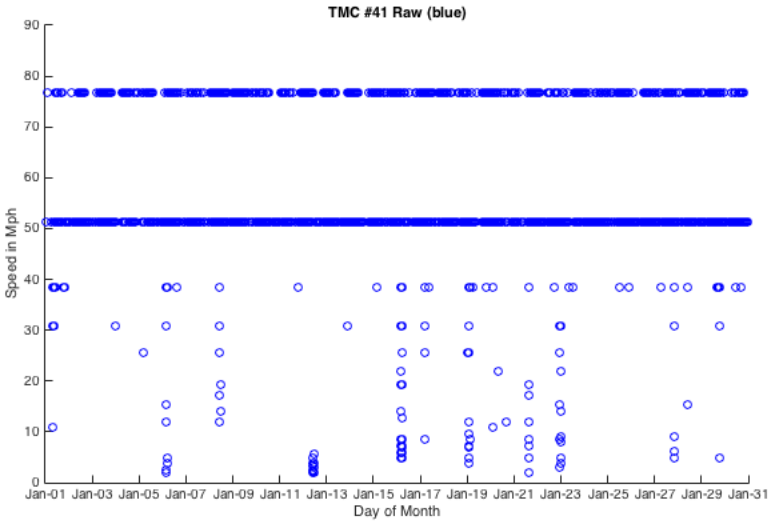


Figure 23 - TMC 49, January 2015 monthly speed plot illustrating the E_r at different speeds.

Figure 24 demonstrates that a speed of 50 mph in segment 41 has an E_r of 16.7 mph. As such, congestion detection could not be considered accurate at this level. However, at a

speed of 30 mph, E_r becomes 5.9 mph. For both plots, we find that there exists cases of extreme congestion where cars come to an almost complete stop.



.

Figure 24 - TMC 41, January 2015 monthly speed plot illustrating the *Er* at different speeds.

In conclusion, the aforementioned study indicates that the FHWA should recommend to HERE changing time granularity of NPMRDS data reported according to Equation (2) which should alleviate inherent errors in the nationwide NPMRDS dataset.

3.2.Data Outliers:

Congestion on segmented roadways is a function of both time and space. In space, a shock wave starts at the observed segment and then ripples to subsequent segments lagging behind the observed segment. The result is increased reported travel time. In the time domain, the aforementioned shockwave manifests at the observed segment with an increased travel time for a recorded epoch, and then expands to later epochs of the same segment as congestion continues. At a certain point of time—given that the duration of congestion is long enough—spill over to epochs of segments behind the observed

segment occurs and expands congestion in space. Consequently, congestion can first be detected in time in the observed segment, and then stretch in space to adjacent segments. Given the observed segment is short in length, time and space can expand nearly simultaneously, meaning epoch travel time duration simultaneously increases in the observed and lagging segments when sampling time is long enough to allow congestion spillover to adjacent segments. In light of this understanding, we proceed to analyze outliers and formulate procedures for removing them from the NPMRDS dataset.

3.2.1. Effect of high spatial-temporal variance

As aforementioned, there exists high spatial-temporal variance in the number of epoch records in the NPMRDS data for the NHS roadway segments. The chief cause for this variance is the varying number of probe vehicles present on any segment at any instance of time. A particular case occurs when the sample size is very low. The small sample size could result in outliers' non-representative of actual travel times for vehicles on the segment. These outliers can either be high or low valued points. Cases where sampled data points exhibit extreme unrealistic values could also be caused by a system related error during data acquisition or conditioning. Detecting these outliers is achieved by checking for data points that are too extreme to be realistic in the dataset. Researchers at Wisconson Madison in [23] pointed to this type of outlier, and recommended scanning for observations that are several standard deviations above the mean of the analysis time period, or setting the data as panel observations and flagging points that are significantly different from their lagging and leading neighbors. In the Wisconsin study, researchers detected points that were 73 standard deviations above the mean. In the work presented in this thesis, average speed above 3 mean standard deviations from the speed limit (e.g.,

speed equal to 20.8 mph) is considered an outlier. This equates to approximately 90 mph on a roadway with a speed limit of 70 mph. Reported speed represents averages. Thus, it is unrealistic for all cars traveling on the roadway to be averaging 90 mph or above. If such findings would occur, results could be indicative of a very small sample size. Values for I-35 southbound were first threshold above 90 mph. Results were plotted per segment in ascending order for combined travel time, as shown in Figure 25. Figure 26 shows similar results for passenger car travel time, and Figure 27 shows the same for freight truck travel time.

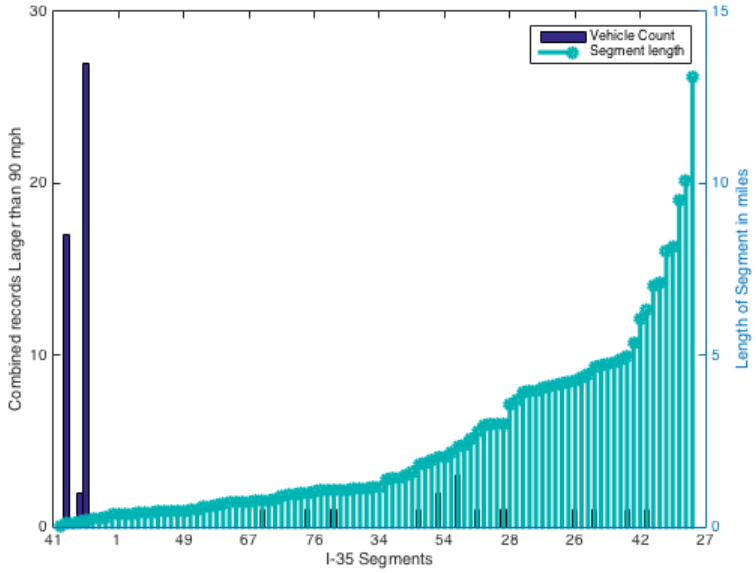


Figure 25 - Combined vehicle count plot for number of epochs with speeds greater than 90 mph for i-35 southbound segments.

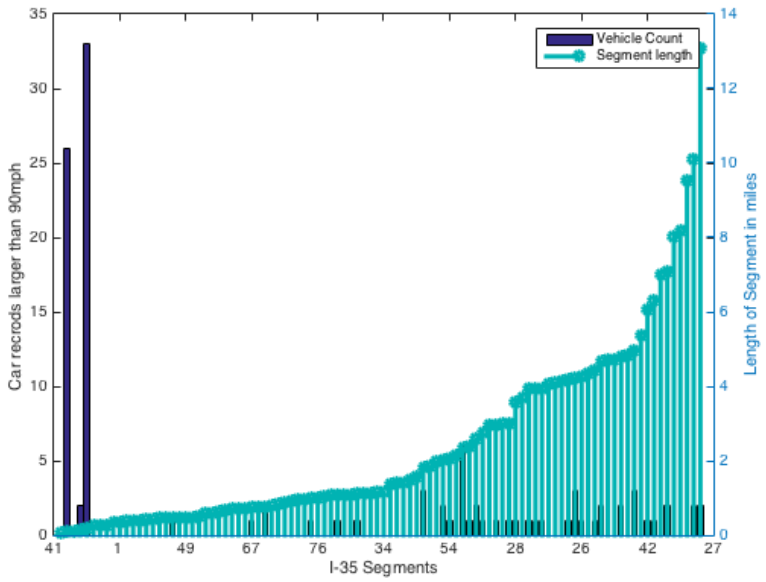


Figure 26 - Passenger vehicle count plot for number of epochs with speeds greater than 90 mph for i-35 southbound segments.

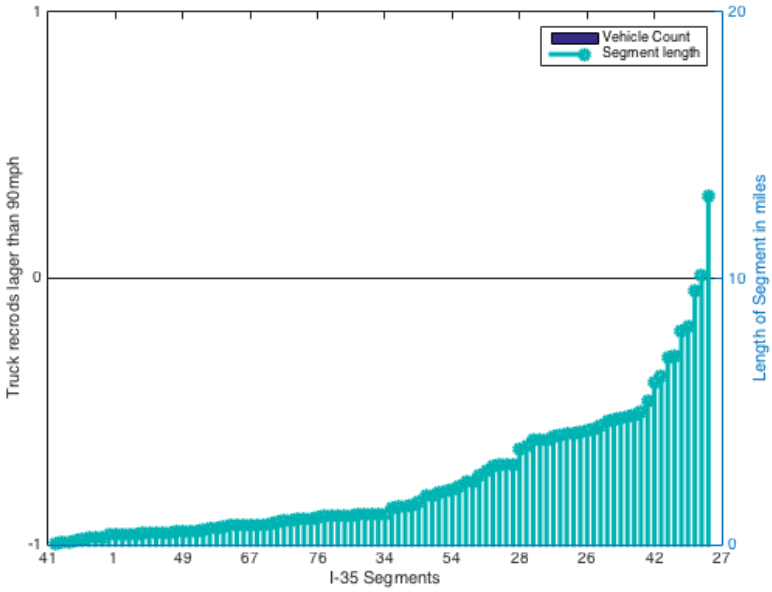


Figure 27 - Truck vehicle count plot for number of epochs with speeds greater than 90 mph for i-35 southbound segments.

Figure 25 demonstrates that 111 records were detected for passenger vehicles traveling I-35 southbound at speeds higher than 90 mph. Speeds were reduced for the combined car-

truck matrix when averaging with truck speed records. Notably, samples were collected on shorter segments of I-35. It is obvious two phenomena were at play.

1- Shorter segments have smaller densities, which in turn affects sample size. Thus, a fast traveling vehicle might be the only sample present at a particular instant of time, making its speed not representative of average vehicle speed. Nevertheless, if the high speed is considered an accurate value of vehicle speed, it could be surmised that vehicles can travel at free flow speed with no obstruction or congestion regardless of actual free flow speed. If the outlier were to remain in the dataset, it would cause problems when performance metrics were calculated. For statistical analysis integrity, the outlier must be removed.

2- Speed quantization error related to the variability of segment length. The fifth spike observed in Figure 25 demonstrates this for segment 76, which has an E_r of 13 mph for speed 91.5 mph.

In the case of congestion analysis, we can set all these points to the speed limit, as they are merely indicative that no congestion is present and cars have the ability to travel at free flow. Three matrices were generated: 1) Combined values matrix with speeds above 90 mph reset to the speed limit; 2) passenger vehicle speed-corrected matrix; and 3) truck speed-corrected matrix. Collectively, there are six matrices: three original and three corrected. Speeds slower than 2 mph were not excluded as in [23], because there were instances when probes reported 0 mph, indicating traffic had come to a complete stop.

3.2.2. Vehicle specific performance data points (Power-to-Weight)

In order to detect outliers that might be caused by vehicle specific characteristics on the road as explained in the power-weight phenomena occurring in heavier vehicles, we build on the assumption that trucks recording slow speeds in correlation with passenger cars recording faster speeds is indicative that the faster speed characterized by a car represents a better approximation to the true speed of the road, and the slower truck speeds represent characteristics of the truck itself, or what is termed as vehicle specific performance data. In this case we set the speed of the combined (car-truck) data matrix, to the speed of the highest of the car or the truck and remove the outlier. Thus, detection is done by correlating speeds of trucks and passenger vehicles for the same epoch and segment, and removal is done by replacing speed entries with the higher of the two speeds.

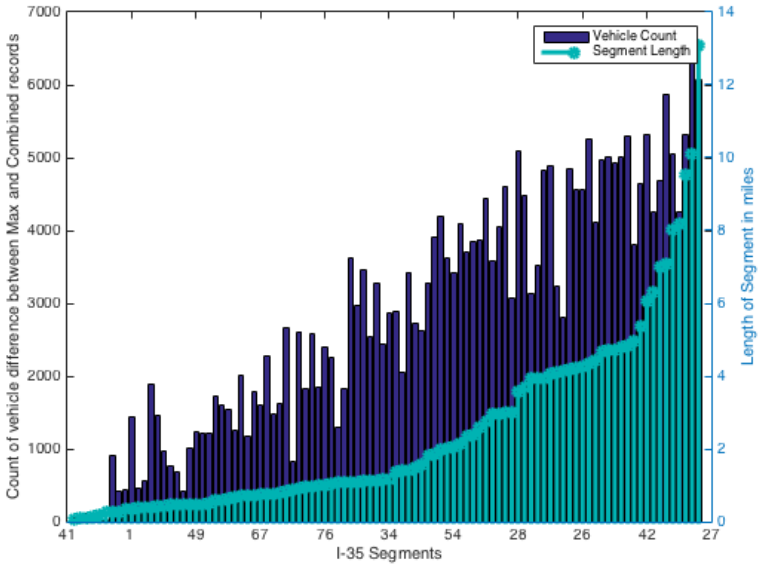


Figure 28 - Epoch record count for difference of max (truck, car) matrix to combined matrix.

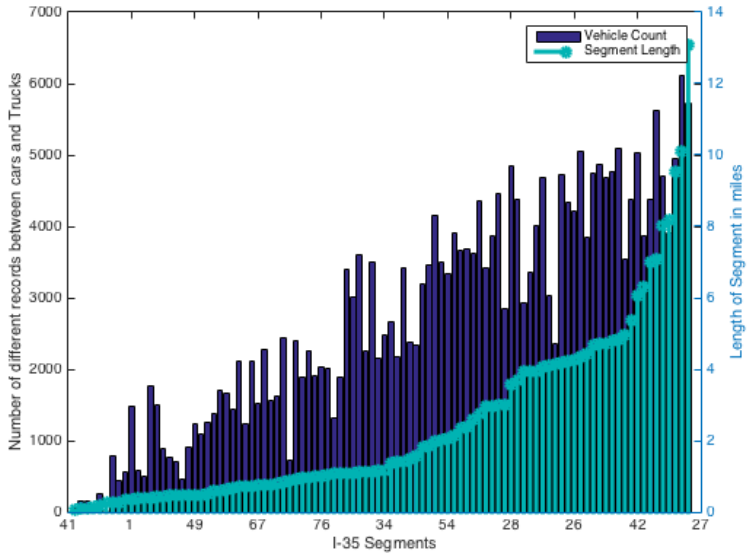


Figure 29 - Epoch record count for difference between car and truck matrices.

Figure 28 shows a plot of the maximum speed matrix subtracted from the NPMRDS dataset combined-all-vehicles matrix. Figure 29 shows a plot of the number of epochs when passenger car speeds were higher than truck speeds. Both figures nearly identical, indicate that the majority of slower speeds were caused by trucks slowing for vehicle-specific reasons rather than roadway conditions affecting all traffic. Figure 29 demonstrates that as segment length increases, the number of effected epochs averaged down from the maximum value increases, as well. This was confirmed when examining the percentage of down shifted epochs relative to the total number of epochs available per segment. See Figure 30 for a plot of this ratio. Results prove that the shorter the segment, the less epochs were averaged down. Nevertheless, as one would intuitively guess, an outlier would have a more profound effect due to the fact that fewer samples decreases the probability of correction when there is an outlier.

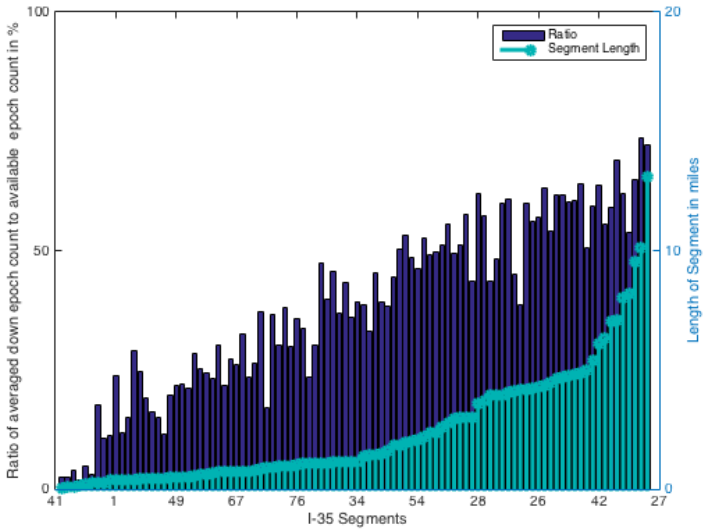


Figure 30 - Ratio of averaged epoch count to the total number of epochs available.

Figure 31 and Figure 32 show the mean and the standard deviation of the speed difference between the maximum and the combined vehicle speeds. Average difference for most segments is approximately 5 mph, and the standard deviation is approximately 2 to 3 mph. As segment length decreases, mean increases. Reported combined speeds in the NPMRDS dataset show on average a 5 mph reduction in speed compared to actual roadway speed as a result of slower freight trucks.

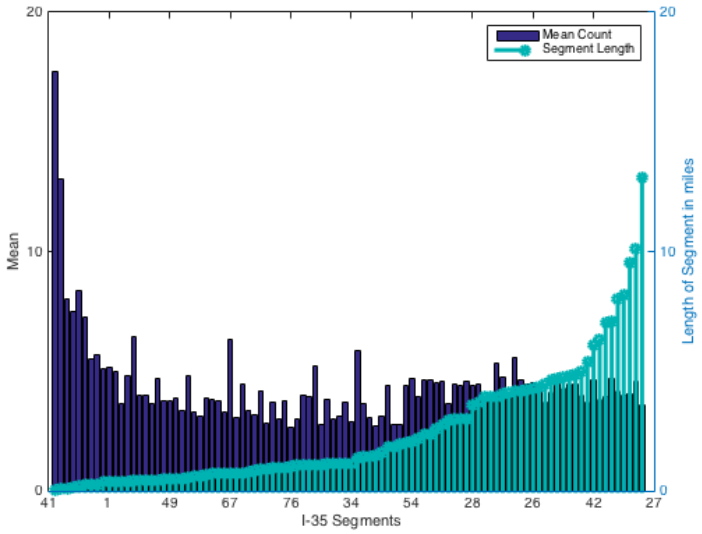


Figure 31 - Mean speed difference between max passenger and combined speeds.

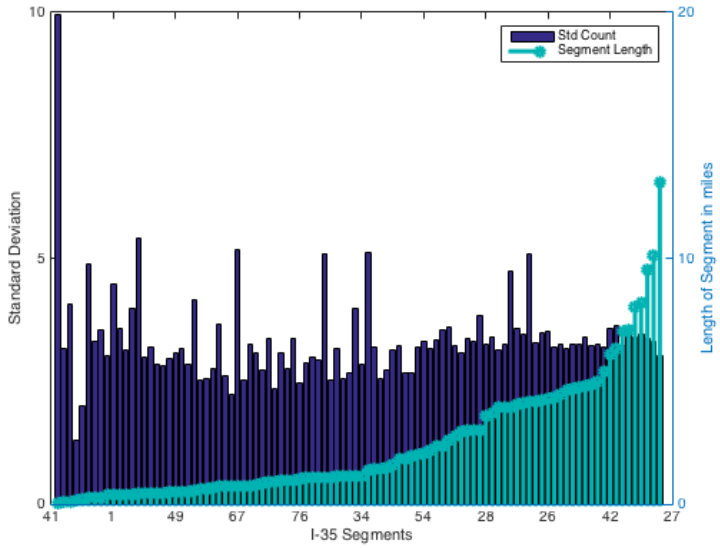


Figure 32 - Standard deviation of speed difference between max passenger and combined speeds.

3.2.3. Roadway geometry

When roadway geometry affected travel time, segments continually reported slow travel time when compared to speed limits. This phenomena builds on the assumption that slower travel times are a result of highway topography caused by the nature of the road itself, which consistently forces vehicles to slow down. Admittedly, roadway conditions might only affect larger truck speeds and not, passenger car speeds. In such a case, the power-weight ratio law would not consistently be cause for slowing down traffic. When slow trucks were identified based on passenger vehicles traveling at free flow speeds, changes were not made to the dataset. Instead, such cases were marked for post check in GIS. These cases are of interest to DOT agencies, as they show locations where segments could possibly undergo optimization for freight travel time.

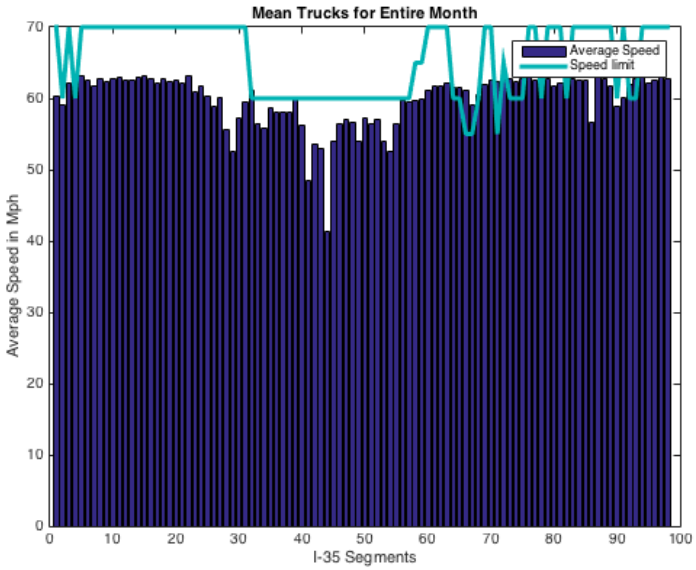


Figure 33 - Average epoch truck speed per segment for January 2015.

To investigate roadway segments, mean truck speeds were collectively monitored vis-a-vis speed limit during a one month time period. Figure 33 shows results for I-35 southbound. Average truck speed in January 2015 was somewhat below the speed limit. A plot of the highest mean day speed per segment is shown for trucks and passenger cars in Figure 34 and Figure 35, respectively.

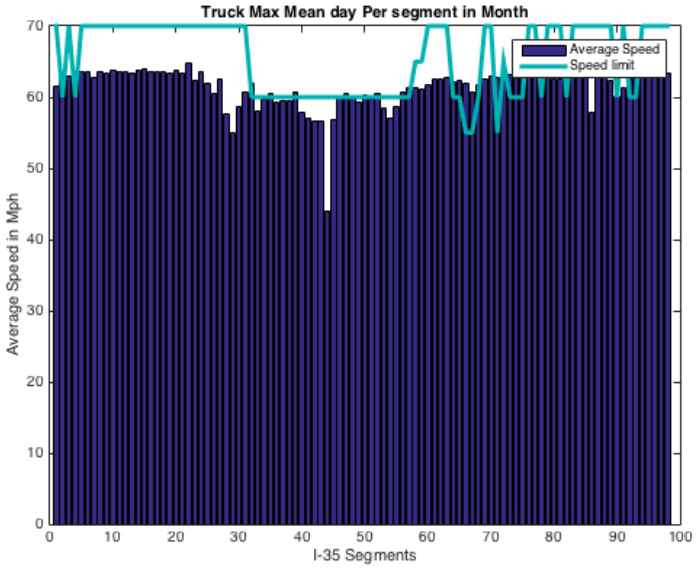


Figure 34 - Max day mean epoch truck speed for Januray 2015.

For most segments, average truck speed was recorded below the roadway speed limit. Also, some segments recorded average passenger car speed below the speed limit. Segment 44 in particular stands out for having speeds significantly below the speed limit throughout the month of January 2015. This result was consistent for both freight trucks and passenger cars.

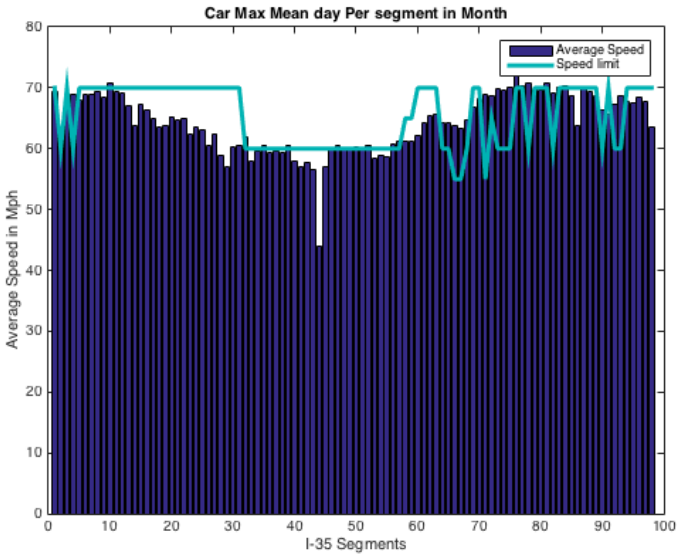


Figure 35 - Max day mean epoch car speed for January 2015.

Coordinates for segment 44 were extracted and are shown on the google map satellite image in Figure 36 and Figure 37.

Figure 36 - Segment 44 I-35 intersect with the centennial express way HW 235.

Segment 44 begins at the intersection of I-35 and Centennial Expressway Highway 235. The on-ramp is only one lane, which causes traffic slowdown for cars and trucks alike, as evidenced in the NPMRDS dataset.

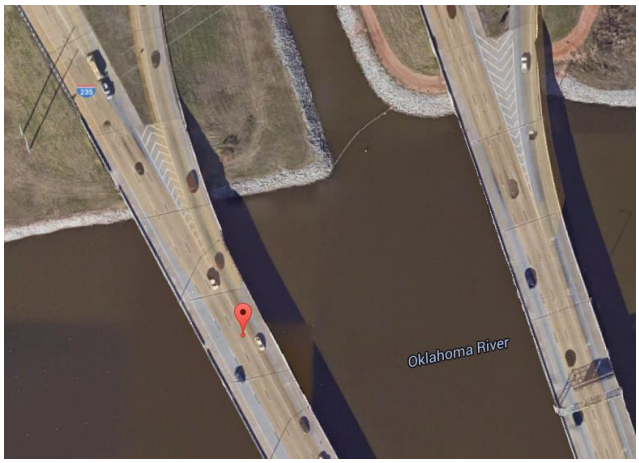


Figure 37 - Close view of segment 44 I-35 intersect with the centennial express way HW 235

3.2.4. GPS In-accuracy (non-NHS roadway data points).

Either faulty GPS units or insufficient positioning accuracy result in inclusion of data points that are not part of NHS roadways. As mentioned earlier, data records could actually belong to roadways adjacent to the NHS. When sample size is large, outlier effect is minimal. When the sample size is small, outlier effect is possibly measurable. Recall that detection relies on the assumption that there is a speed difference between NHS roadways and adjacent non-NHS roadways. Thus, any record mistakenly reported due to GPS inaccuracy would be different from lagging and leading epochs for any segment under study. Another indicator is when passenger car speeds are slower than truck speeds by one or more standard deviation in the same segment. By extracting all cases where trucks are faster than cars and removing all cases where cars are slower than trucks by

less than the maximum standard deviation (e.g.15 mph for I-35 southbound), all cases with noteworthy speed difference between cars and trucks can be identified. See Figure 38. Although such cases could be indicative of non-NHS roadways, the differences could be the results of a small sample size for passenger vehicles that reported outliers that were not representative of the average speed per segment. Threshold results were based on number of occurrences. Empirically, 20 occurrences were chosen, assuming the higher occurrence was indicative of GPS inaccuracies.

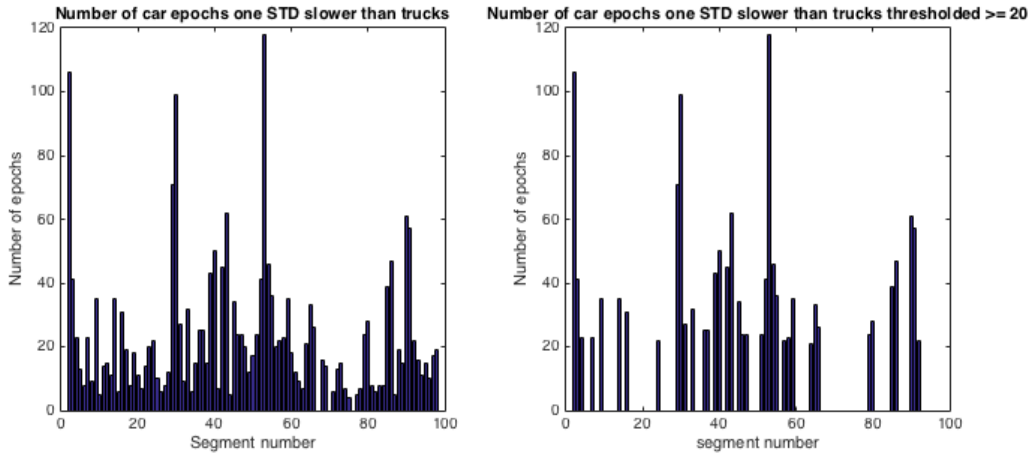


Figure 38 - (a) Cars one standard deviation less than trucks. (b) Threshold result for count >= 20.

Coordinates of a random sample of segments were extracted, and google maps were used for validation. In Figure 38, segment 53 is shown as the highest peak and was found to be adjacent to the I-35 southbound service road (See Figure 39). Similarly, segment 30, which proved to be the segment with the third highest error count, was found to be adjacent to the I-35 northbound service road (See Figure 40).

Figure 39 - Segment 53 adjacent to S I-35 service road.

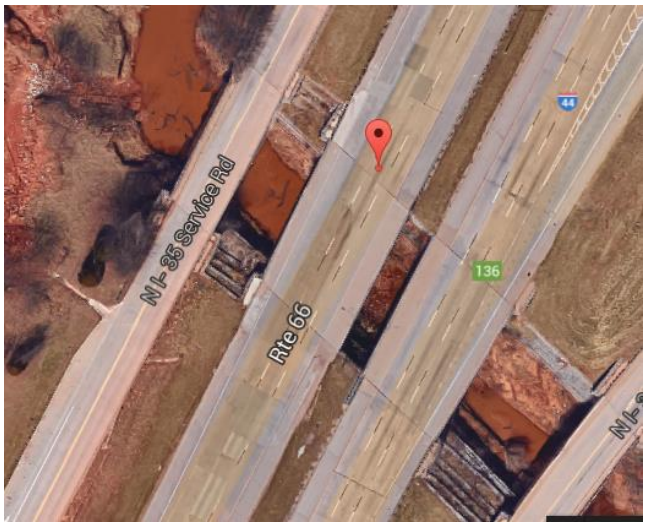


Figure 40 - Segment 30 adjacent to N-I35 service road.

To identify and remove outliers the following two procedures were performed:

1- A new output speed matrix was generated and consisted of the maximum speed record between both cars and trucks reported for each given epoch. The matrix alleviated non-NHS outliers when both car and truck speeds were available.

2- Building on the notion of congestion, as described earlier in this chapter, a mask filter was constructed to scan the entire database and to identify, then remove remaining outliers.

Figure 41 illustrates the mask used to scan the speed database. The mask filter identified three types of congestion: 1) New congestion evident in future epochs; 2) Present congestion evident in past epochs; and 3) Propagating congestion evident in adjacent segment epochs. Figure 42 illustrates a flow chart for the process used to remove outliers from the database. The process commences with thresholding a current segment epoch based on a modified congestion detection approach, which is described in Chapter 4. Once an epoch has been identified as likely congestion, all gray marked entries in the mask are thoroughly inspected likely congestion. If speed value of any grey entry is indicative of congestion, a flag is raised for the particular corresponding entry. If a check flag is detected at the end of the process, the current segment epoch is not altered. Given there is no flag, the current segment epoch is reset to the speed limit. A 20-minute detection range was chosen for the NPMRDS dataset, primarily because some missing epochs (i.e., epoch holes) were evident for consecutive records in particular segments in the dataset.

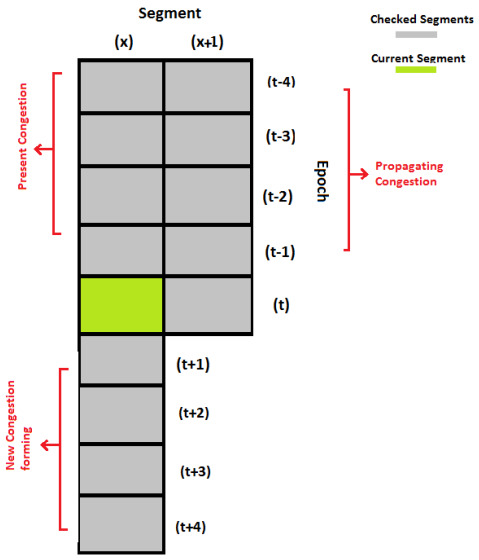


Figure 41 - Mask filter to scan for outliers.

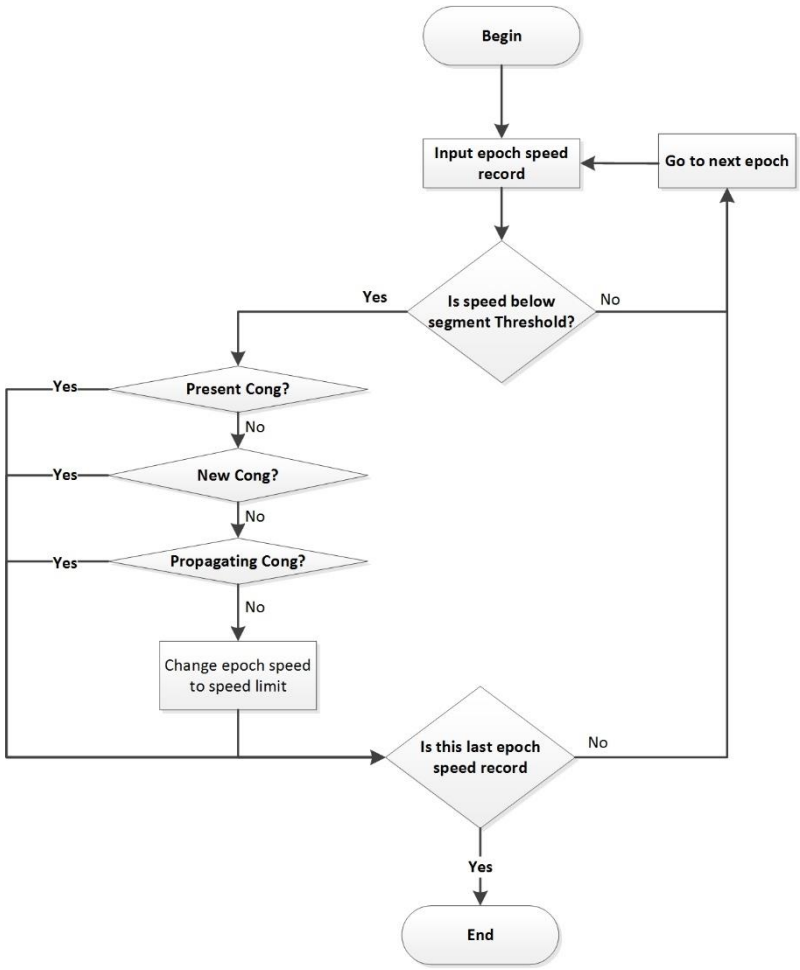


Figure 42 - Flow chart for scanning outliers using mask filter.

3.3. Cleansed dataset

After applying the aforementioned methods and processes, a cleansed dataset was generated. Table 9 shows a database example for segment 97 with outlier speed reported. Epoch 1818 speed of 34.6485 mph is considerably lower than previous, consecutive and adjacent recorded epoch speeds. As such, the value was considered an outlier, and was, accordingly, reset to the speed limit for the segment.

Table 9 - Database outlier for segment 97 in raw database.

	96	97	98
1805	62.7704	61.6913	61.8079
1806	62.3398	62.1969	59.1206
1807	64.3258	62.4529	NaN
1808	65.8407	62.7110	63.2453
1809	64.2122	77.8259	NaN
1810	64.3258	62.9712	NaN
1811	63.8736	64.0340	64.7511
1812	62.2330	65.9828	64.7511
1813	62.0206	62.9712	69.7320
1814	63.9861	62.4529	NaN
1815	65.3671	64.8549	63.2453
1816	63.8736	67.1507	64.7511
1817	58.1505	60.7042	NaN
1818	62.5544	34.6485	63.2453
1819	63.9861	64.0340	NaN
1820	64.2122	61.4415	61.8079
1821	64.2122	65.6972	61.8079
1822	64.3258	64.5789	66.3304
1823	63.4277	63.2335	60.4344

Figure 43 and Figure 44 illustrate a plot for segment 97 and segment 69 speed records in January 2015 composed of both raw speed data obtained from the travel time measurements without processing, as well as the cleansed dataset following anomaly and outlier detection and removal procedures.

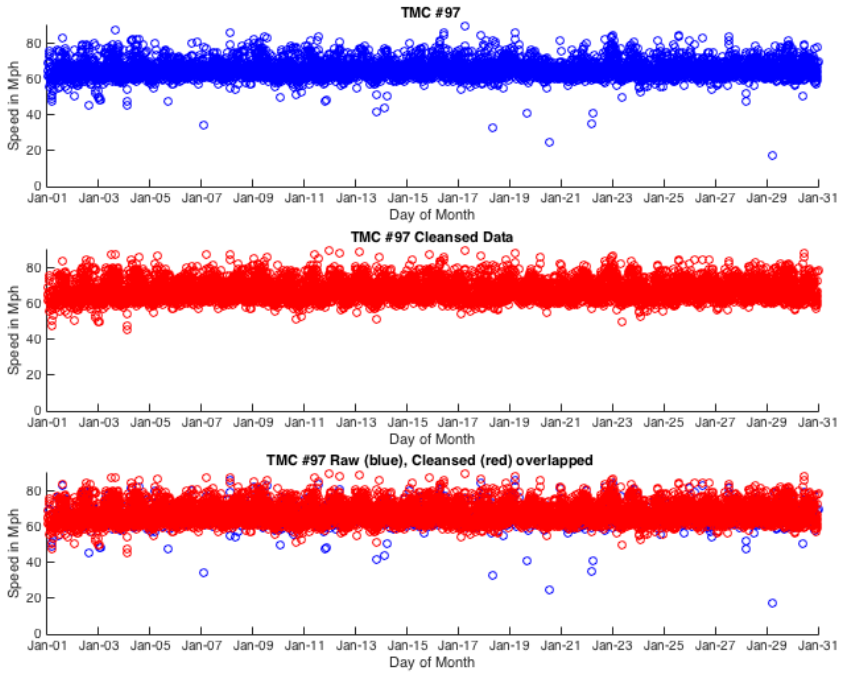


Figure 43 - Comparison for Segment 97 speed records, raw vs cleansed data for the month of January 2015.

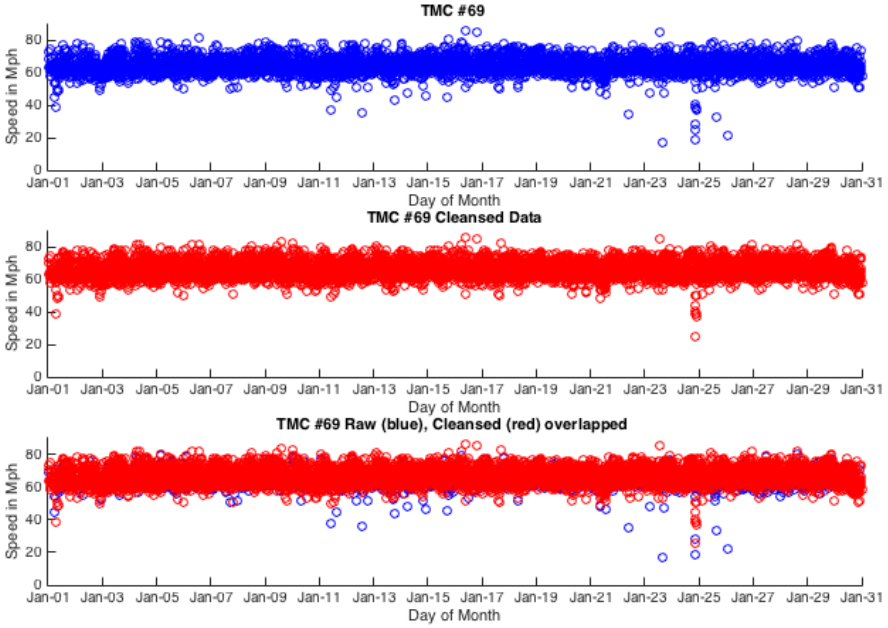


Figure 44 - Comparison for Segment 69 speed records, raw vs cleansed data for the month of January 2015.

Chapter 4: Dataset Exploration, Analysis and Congestion Detection

Classical applications of central tendency and variation—specifically means and standard deviations—are influenced by outliers. Appropriate measures discussed above were applied to alleviate the dataset of anomalous and outlier data points to obtain accurate aggregated measures of central tendency. In this section, comparative exploratory data analysis is performed for both the baseline raw dataset and the cleansed dataset as reported in the previous section. Limitations of standard statistical analysis for congestion detection are discussed, in particular the use of variance. This chapter also presents a robust method for detecting congestion by using the NPMRDS dataset to identify abnormal travel times on the roadway.

4.1. Statistical mean and variance

Utilizing travel time measurements in the NPMRDS, each segment extracted from the dataset was linked with its equivalent row of the geographical information system (GIS) static file provided by HERE. This fusion was then used to convert travel time to speed measurements using segment length. To determine speed limit per segment, ODOT provided a Google earth data file to facilitate manual-visual extraction of speed limits, as well as manual location coordinate matching for each segment. This task proved tedious and error prone. Nevertheless, as a preliminary tool for processing, the data served its purpose, noting that speed limit data has to be acquired with relatively higher accuracy for improved processing. Data linkage was done between extracted segments and the created speed limit file.

Figure 45 shows average speed of epochs for one month for all segments of I-35 southbound. Records were gathered for segments spanning from segment 1 at the Kansas

border to segment 98 at the Texas border. The top graph shows the raw dataset mean, and the lower graph shows the cleansed dataset mean after outliers were removed. Mean speed of the raw unprocessed dataset was 62.5475 mph across all segments. Cleansed dataset mean speed was 64.3716 mph across all segments. Average speed limit across all segments of I-35 southbound was 65.4082 mph.

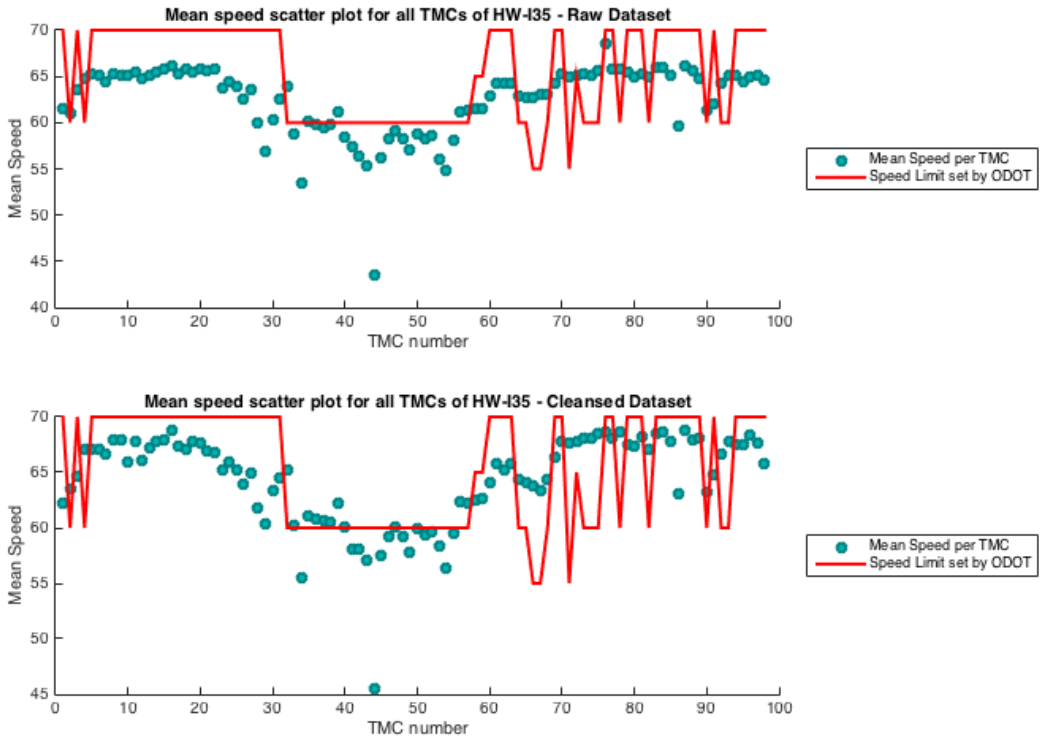


Figure 45 - Mean speed per segment vs. speed limit.

Raw data was utilized to calculate an average speed that was below the speed limit in nearly all segments, except those located in and around Oklahoma City. These are found in the center of the graph. Average speed correlated to speed limit in the cleansed dataset. Figure 46 shows speed variance per segment for all epochs during the month of January, 2015.

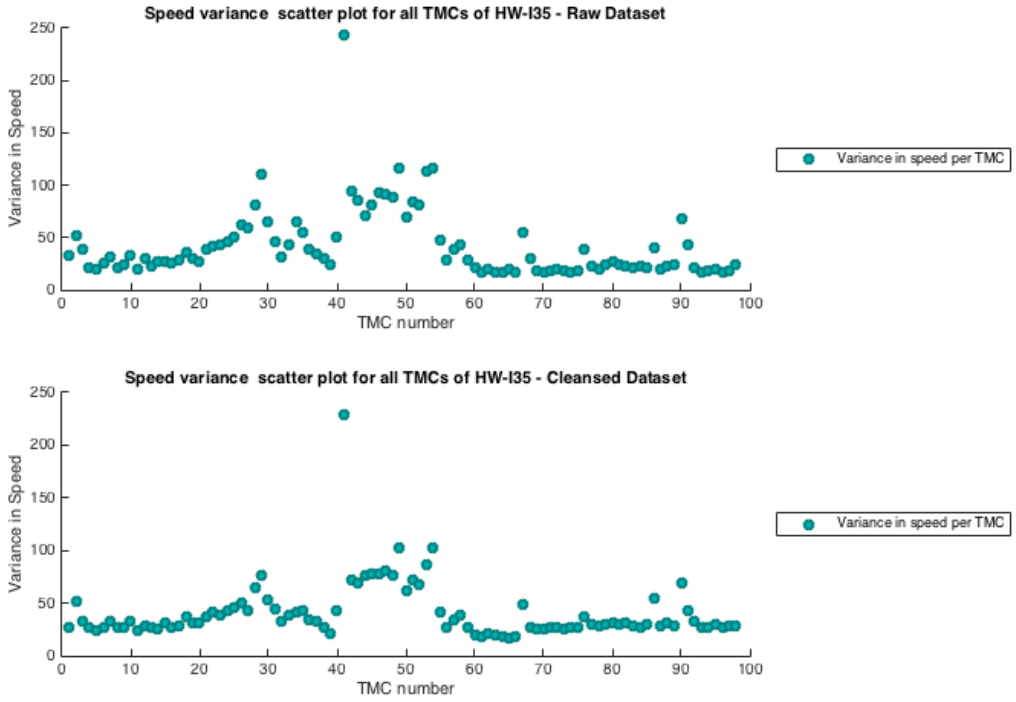


Figure 46 - Speed variance per segment for I-35.

Raw and cleansed graphs demonstrated that TMC stations had increased variance values, which could be indicative of abnormal, non-free flow traffic. Although variance in the cleansed dataset was slightly lower than variance in the raw dataset, the results were indicative of abnormal traffic speed (i.e., travel time fluctuations). [39] Suggested that a variance metric could be used to detect congested segments characterized with such abnormal traffic flow. Researchers concluded that travel time had little variance when estimated under non-congested conditions and high variance with increased value when estimated under congested conditions.

4.2. Epoch variance, segment weight and traffic correlation

As mentioned earlier, NPMRDS data is affected by several limitations and several challenges. One important factor is number of epochs generated per segment relative to the number of probes available at any location and at any specific point in time. Discontinuities in epoch availability can skew results and affect accuracy of computed travel time performance measures. Epoch availability is depicted in a 3D surface plot in Figure 47 which shows number of epochs per day for each segment of I-35. The plot shows a correlation of epoch numbers on most days of the month. Slight changes on weekends are visible as wave patterns for all segments throughout the month.

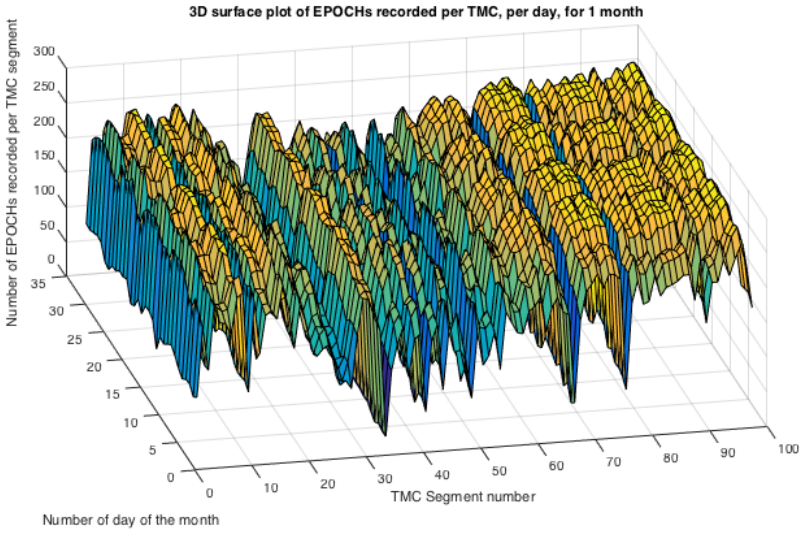


Figure 47 - 3D surface plot of epochs recorded per segment, per day, for January 2015.

Figure 48 shows an overlay epoch count plot for TMC segment per day during the month of January 2015. Each segment has to a large extent a repetitive pattern for nearly all segments.

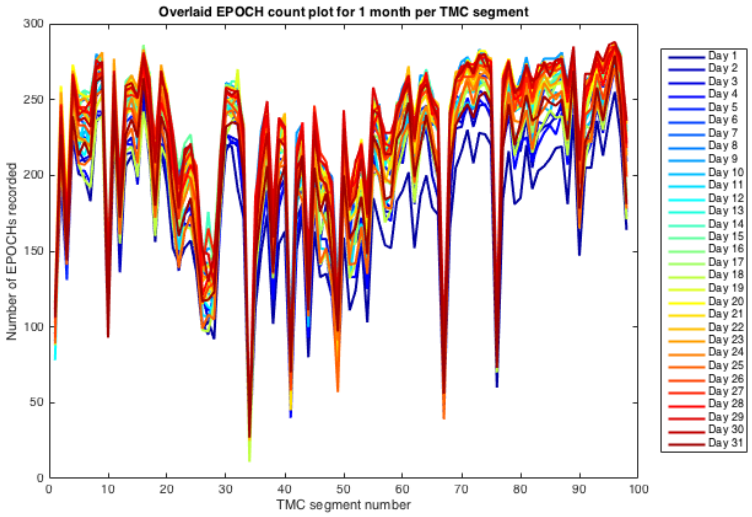


Figure 48 - Overlay epoch daily count for January 2015, per segment

Correlation between epoch counts can be validated numerically. Consider the correlation of two random variables A and B as a measure of their linear dependence. Given that each variable has N scalar observations, then the Pearson correlation coefficient can be applied as given in equation 4: [40]

$$\rho(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\overline{A_i - \mu_A}}{\sigma_A}\right)\left(\frac{\overline{B_i - \mu_B}}{\sigma_B}\right) \tag{4}$$

where μ_A, μ_B and σ_A, σ_B are the mean and standard deviation of A and B, respectively. Alternatively, this is also defined in terms of the covariance of A and B [40]:

$$\rho(A,B) = \frac{cov(A,B)}{\sigma_A \sigma_B}$$

$$R = \begin{pmatrix} \rho(A,A) & \rho(A,B) \\ \rho(B,A) & \rho(B,B) \end{pmatrix}$$

$$R = \begin{pmatrix} 1 & \rho(A,B) \\ \rho(B,A) & 1 \end{pmatrix}$$

The correlation coefficient matrix of two random variables is the matrix of correlation coefficients for each pairwise variable combination. Since A and B are always directly

correlated to themselves, diagonal entries are the value of 1. Figure 49 shows mean correlation coefficient results per segment.

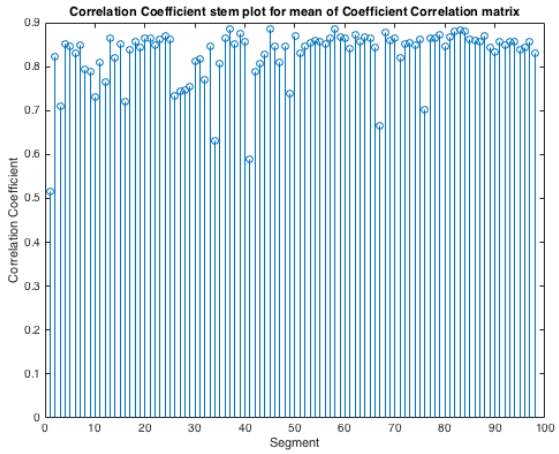


Figure 49 - Mean correlation coefficient per segment stem plot.

A box plot was used to generate the coefficient correlation matrix shown in Figure 50. The central mark of each box is the median; box edges are the 25th and 75th percentiles; whiskers extend to the most extreme data points not considered outliers; and outliers are plotted individually. The whiskers extend to a corresponding t +/–2.7σ, which should cover 99.3% of the data, assuming normal distribution. Correlation between epoch count patterns on I-35 is obvious for the majority of segments (i.e., there is a correlation in traffic flow across segments due to the fact that epochs are generated by probes). We note the following observations:

1- Most days, the effect of increasing or decreasing probe count spreads across the interstate from the Kansas border to the Texas. Assuming probe density is a fixed percentage of total traffic flow, traffic could be assumed to consist of mostly interstate transit vehicles.

2- Without prior knowledge of the type of highway being investigated. High correlation could be used as an indicator. i.e. Interstate or Non-Interstate roadways.

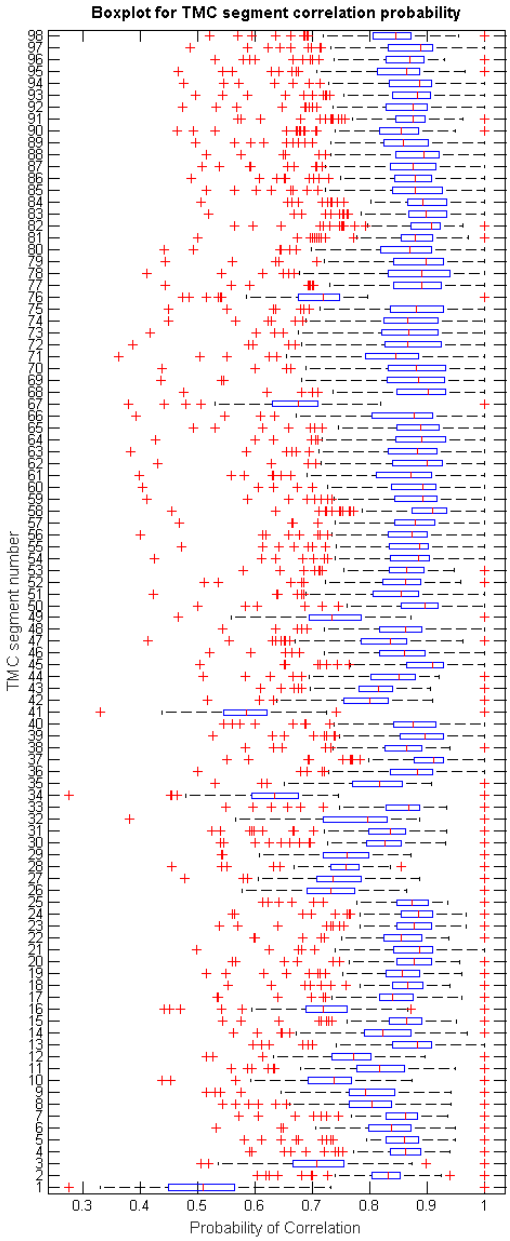


Figure 50 - Boxplot of correlation coefficient matrix.

Furthermore, each segment can be weighted based on average number of daily epochs over the course of the month— in this case January 2015. Figure 51 depicts the results of normalized weight per segment.

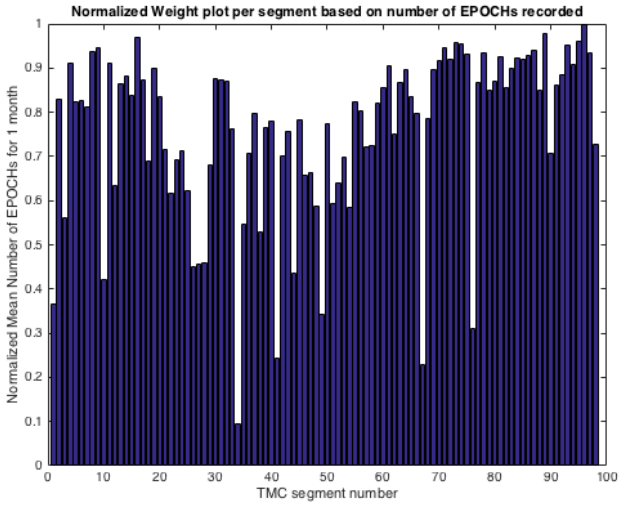


Figure 51 - Normalized epoch count weight plot

4.3. Congestion detection

Road traffic congestion has a negative environmental impact and causes significant loss to productivity to the economy. A beneficial use of the NPMRDS dataset is detecting congested roadway segments. By studying congestion and its correlation with various causes, a deeper understanding is gained about the impact each source has on traffic performance. Collective understanding of both the cause and the effect allow accurate inference and prediction for travel time and, more importantly, travel time reliability.

Literature shows two methods of congestion detection have been utilized. Statistical methods, and thresholding methods [39] [41]. The latter shows thresholds being defined in one of two ways. Either, using free-flow speed as a congestion threshold, or, establishing acceptable minimum speed for various types of facilities and operating environments. An example given is Washington DOT in [42] which defined a threshold

for congestion detection to be 75 % of the posted speed limit, resulting in a threshold for urban freeways with a speed limit of 60 mph to equal 45 mph. And for arterial streets with a posted speed limit of 40 mph to equal to 30 mph.

Assuming vehicles commuting under normal traffic conditions travel at free flow with speeds varying slightly above and below the mean. Given abnormal traffic conditions, speeds tend to vary to a greater extent. Determining statistical variance serves as a simple indicator of congestion [39].

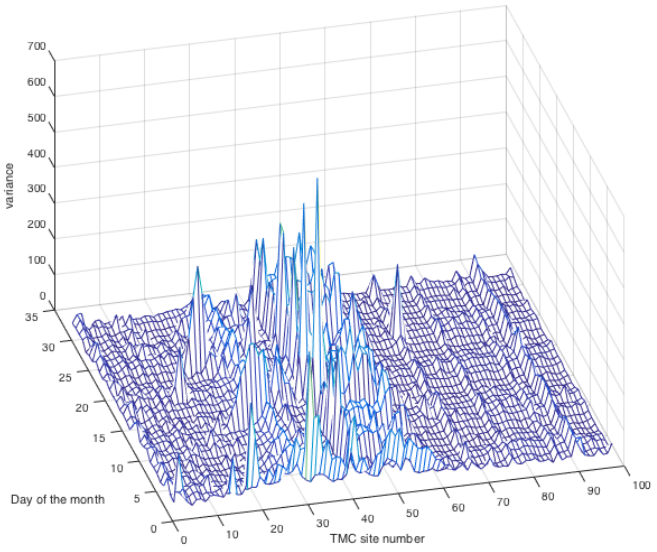


Figure 52 - Mesh plot for speed variance per segment, per day for I-35, Jan. 2015.

Figure 52 illustrates a mesh plot of speed variance per day per segment on I-35 southbound for January 2015. Figure 53 depicts a contour plot of speed variance where peaks of congestion can clearly be identified. Both figures show that commuters most often experience a variance in speed in and around segments 30 to 60 in the Oklahoma City area.

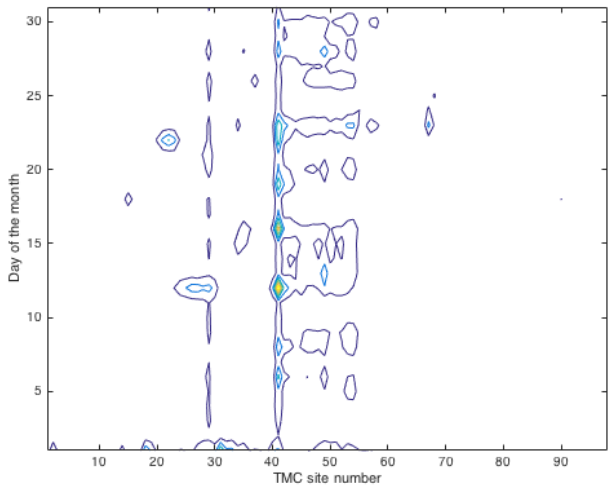


Figure 53 - Contour plot of speed variance per segment, per day for I-35 Jan. 2015.

Extracting the high variance segments and combining with the previously derived weights, a histogram plot shown in Figure 54 depicts congested segments and the number of congested days as well as segments in decreasing variance combined with the number of congestion days. Low reliability segments are marked based on these numbers, indicating the possibility of false congestion detection. In this work, a threshold of 55 epochs per day was chosen as the least number of epochs considered to provide an accurate daily measurement (i.e., any segment generating less 55/288 epochs on any given day was deemed a low reliability segment).

We observe 16 of 98 segments were congested on days that totaled half the month. The majority of the remaining segments experienced congestion on an average of only three days per month, indicating a significant drop in the number of congested days.

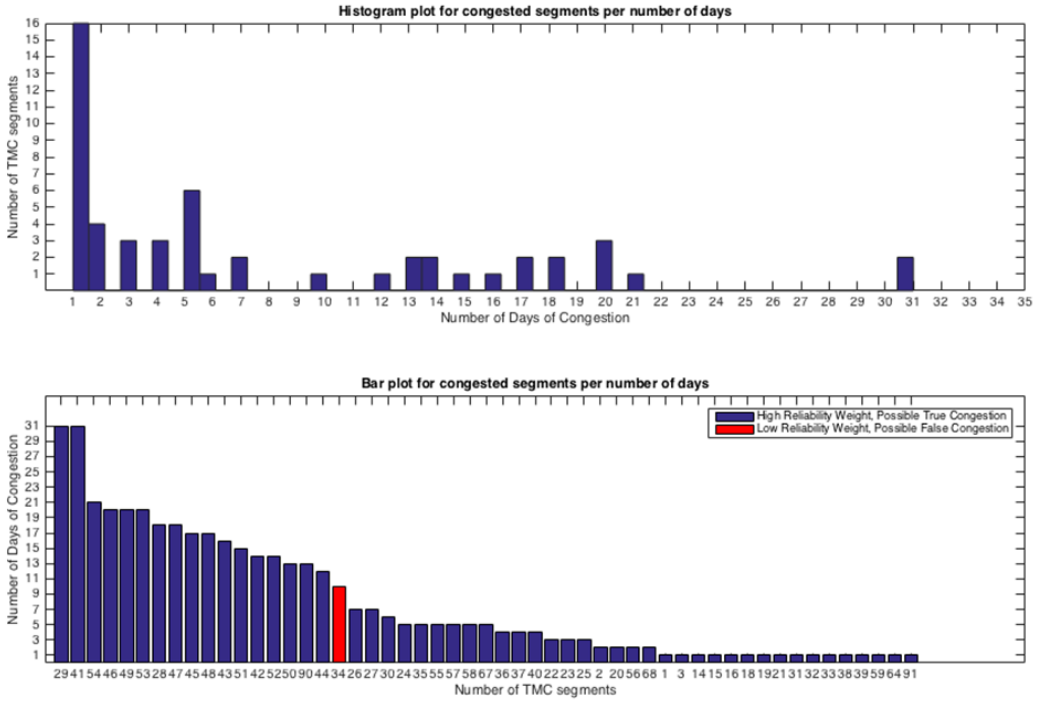


Figure 54 - Histogram and decreasingly sorted bar plots of congested segments on I-35.

It is noted that accuracy could be jeopardized when detecting congestion based on statistical variance. This drawback stems from reliance on false assumptions: 1) Congestion does not occur at all times; When congestion does occur for extended periods of time—equal to duration of analysis—variance measured does not accurately indicate congestion and 2) Variance is related to the number of samples obtained over time, meaning that when congested probes are measured over a short duration they are over masked by a higher number of normal samples. Thus, short bursts of congestion cannot be detected. Such an occurrence is evident in Figure 55, where congestion in segment 69 was not detected when merely considering variance in results. In fact, when examining the monthly plot of epochs for segment 69, undetected congestion occurred for a short period of time on January 25.

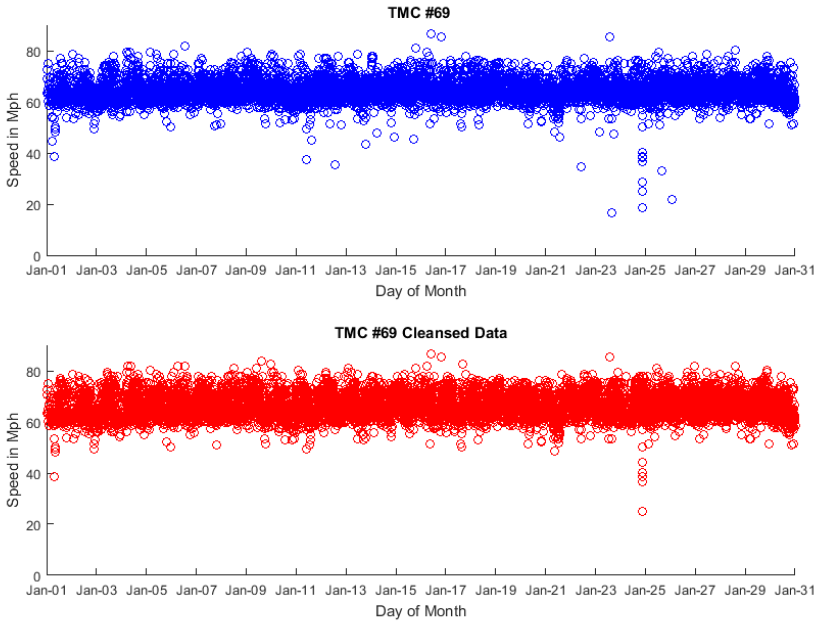


Figure 55 – Segment #69 shows congestion on both raw and cleansed datasets not detected using a standard variance test.

To remedy this problem, probability theory and decision theory independent of sample number daily congestion is proposed as a more robust approach for detecting congestion. Leveraging probability theory in combination with decision theory allows optimal decisions in situations involving uncertainty [43] [44].

4.3.1. Modified congestion detection approach

Assume all free flow traffic over segments can be modeled using a Gaussian distribution without loss of generality [45]. Figure 56 illustrates probability theory suggests that for a normal distribution, values less than one STD from the mean account for 68.27% of the set; two STD from the mean account for 94.45%; and three STD from the mean account for 99.73%. Figure 57 shows three examples of random segments collected on non-congested days and fitted to a normal distribution.

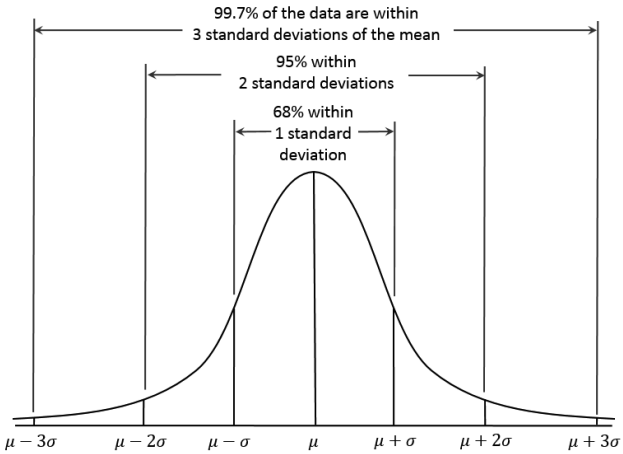
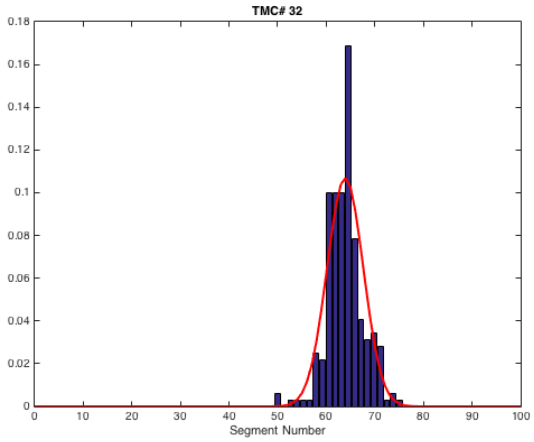
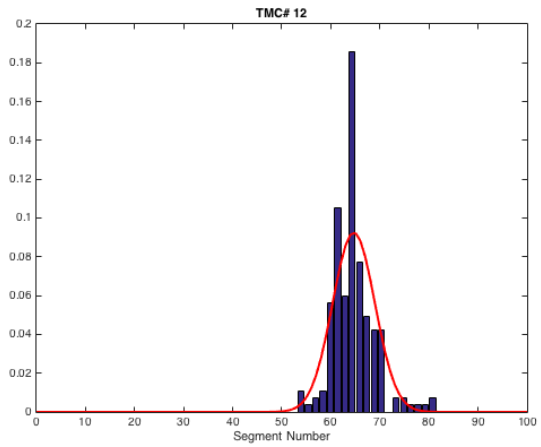


Figure 56 - Normal Gaussian distribution Model

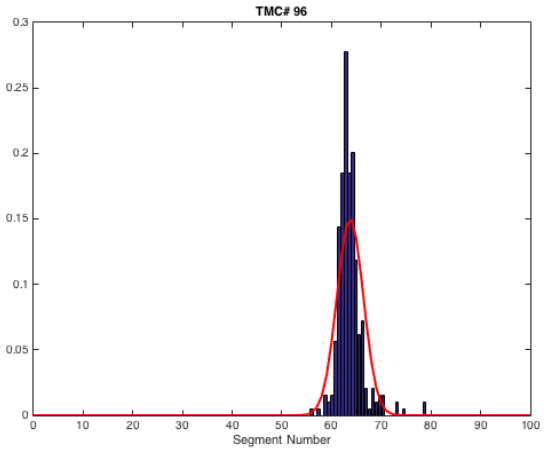
A decision threshold can be established by defining a specific threshold for each segment based on its free flow model at a chosen number below STD from the non-congested mean. Doing so aptly indicates congestion in each segment. The threshold chosen in this work was three STD from the non-congested mean, yielding a confidence of 99.7% approximate to free flow speed.



(a)



(b)



(c)

Figure 57 - Three random segments examples (a, b, c) depicting free flow Gaussian modeled segment speeds.

A database of STD free flow models was constructed, and thresholds per segment were set three STDs from the mean. On average four congested epoch counts occurred for most segments per non-congested days, as shown in Figure 61. Thus, a filter was applied for cases of five or fewer congested epochs during an entire day. Figure 58 shows the results for all segments per day on I-35 southbound during January 2015. Figure 59 and Figure 60 show results in contour and heat map plots. When comparing previous variance test results, it is clear that both results indicate the majority of congestion occurred in and around Oklahoma City in segments 30 through 60. The modified approach, on the other

hand, detected segments not previously discovered with the variance method (e.g., segment 69). Figure 62 illustrates a comparison of variance and threshold test results on segment 69 for detecting congestion.

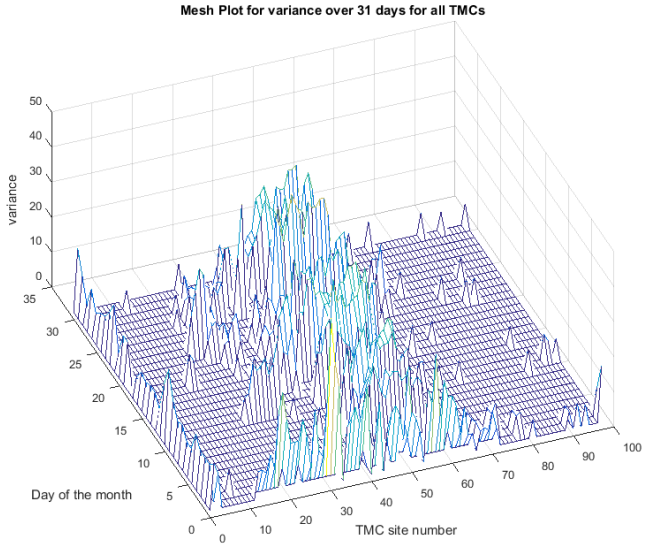


Figure 58 - Mesh plot for thresholded speed variance, per day for I-35, Jan. 2015.

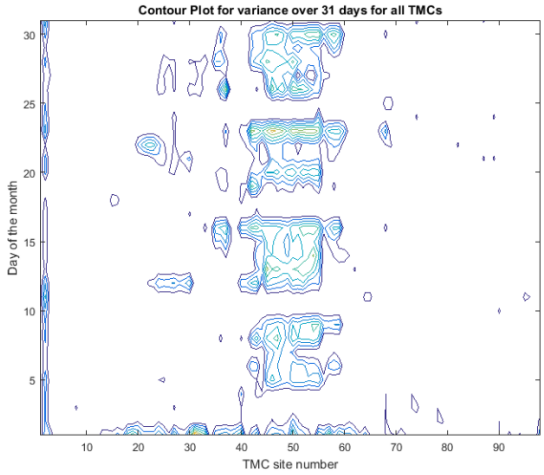


Figure 59 - Contour plot for thresholded speed variance per segment, per day for I-35, Jan. 2015.

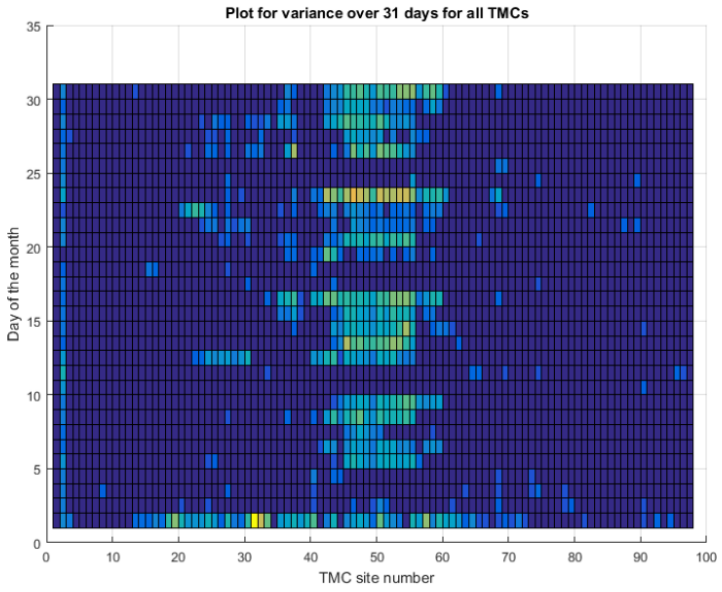


Figure 60 – Heat map for speed variance per segment, per day for I-35, Jan. 2015.

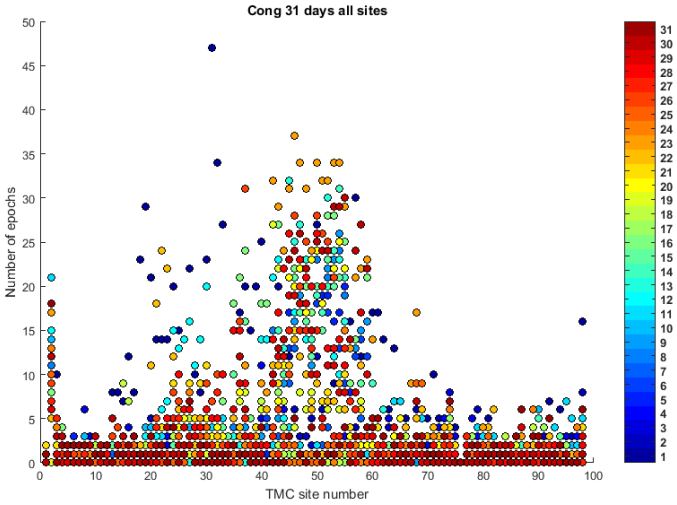


Figure 61 – Congested epoch count for January 2015 on I-35 southbound.

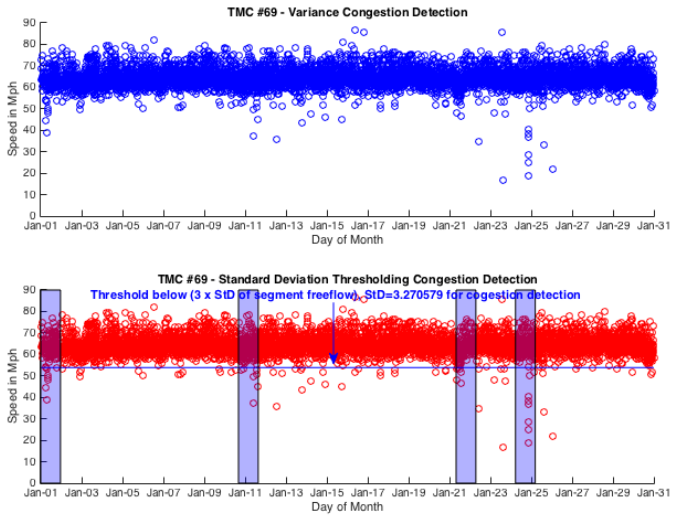


Figure 62 –Variance and threshold congestion detection comparison on segment 69

Further optimization of the congestion detection approach can be achieved by adjusting the filter for the number of epochs required for detection. The filter value establishes a tradeoff between false congestion due to dataset outliers and minimum duration required for the system to detect congestion.

Table 10 offers a numeric comparison between results for raw and cleansed datasets. Figure 63 presents bar plots for both datasets. Congested segments are depicted in order according to decreasing number of congested days. Furthermore, each graph plots a histogram of the number of congested segments and number of congested days. As expected, the raw dataset generated a higher number of congested segments. Outliers present in the raw dataset cause a number of false detections. Three groups of congestion were identified: 1) Segments {12, 7, 6, 4, 91, 80, 84, 11, 85, 86, 9} detected only in the raw dataset (See the table to identify segments for this group), colored in red. Figure 64, Figure 65 and Figure 66 demonstrate outliers in the raw dataset caused false detection.

Table 10 - Result comparison between raw and cleansed dataset.

Segment number	# of congested days in Raw dataset	# of congested days in Cleansed dataset
2	30	30
29	28	4
53	28	18
30	27	7
55	27	21
42	26	12
27	24	14
46	23	22
49	23	22
52	23	20
40	22	10
43	22	21
47	22	21
48	22	21
50	22	22
51	22	18
54	21	18
68	20	9
45	19	19
44	18	14
33	17	4
59	17	9
3	16	2
25	16	8
58	16	13
65	16	3
89	16	2
37	15	12
26	14	4
35	13	7
36	13	10
57	13	11
91	13	0
23	11	5
24	11	8
56	11	7
90	11	4
32	10	4
39	10	1
60	10	4
98	10	4
31	9	3
61	9	4
4	8	0
74	8	5
6	7	0
17	7	1

Segment number	# of congested days in Raw dataset	# of congested days in Cleansed dataset
41	7	6
64	7	2
19	6	1
7	5	0
14	5	1
20	5	3
21	5	3
28	5	3
62	5	2
66	5	1
73	5	2
12	4	0
16	4	2
22	4	3
38	4	3
69	4	4
72	4	1
80	4	0
84	4	0
8	3	1
11	3	0
15	3	2
63	3	2
85	3	0
86	3	0
87	3	2
94	3	1
9	2	0
13	2	2
71	2	2
78	2	2
79	2	2
81	2	2
83	2	2
92	2	2
95	2	2
96	2	2
97	2	2
1	1	1
18	1	1
67	1	1
70	1	1
75	1	1
82	1	1
93	1	1

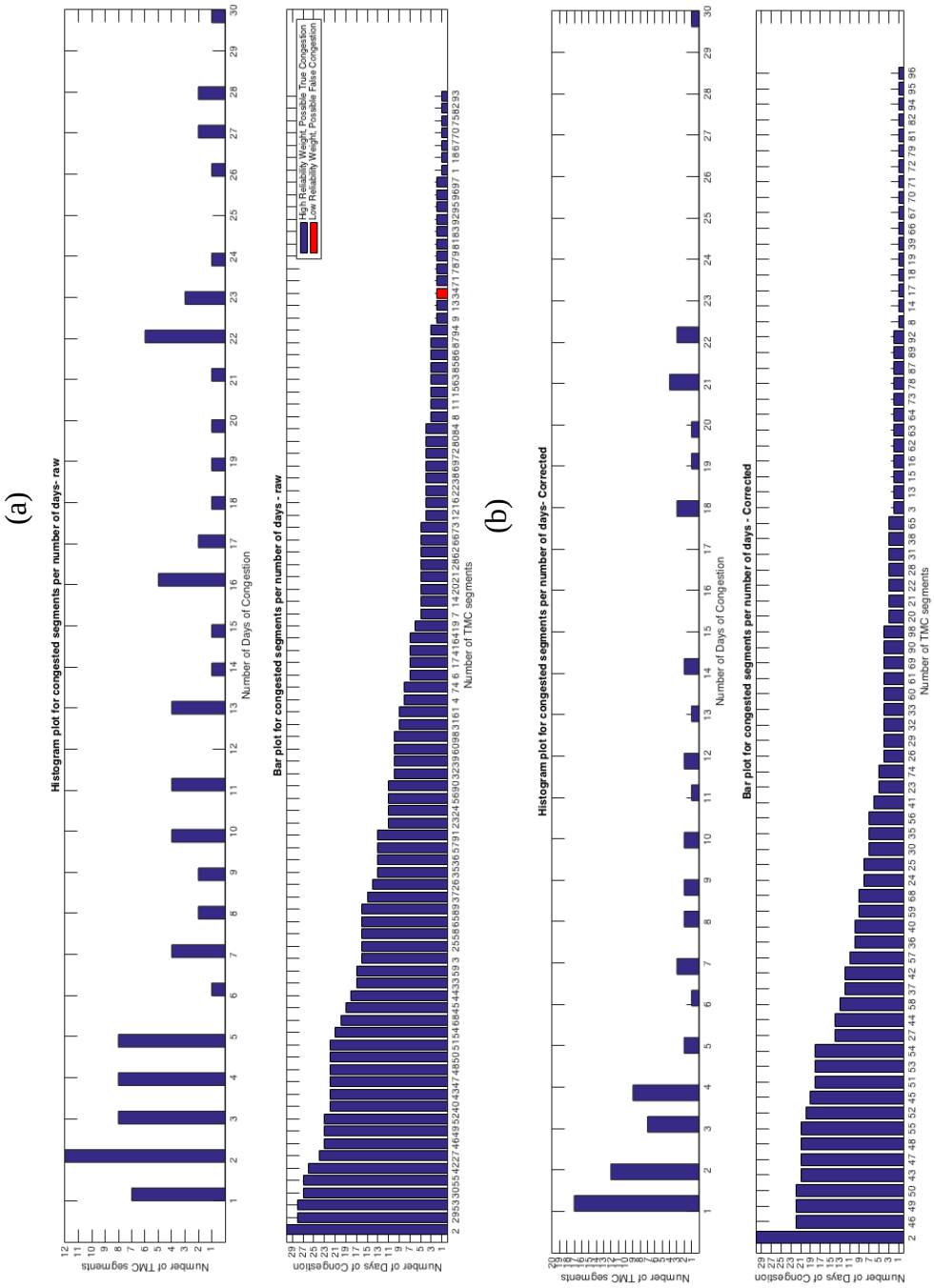


Figure 63 - Modified congestion detection results for raw (a) and cleansed dataset (b).

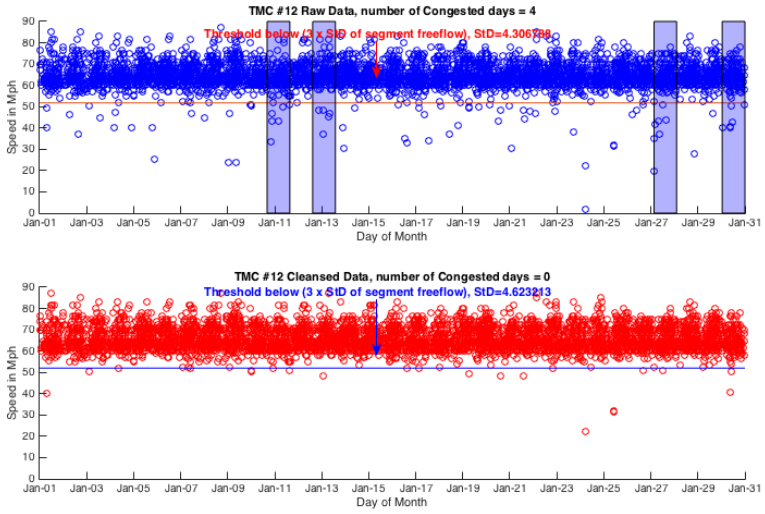


Figure 64 - Segment 12 congestion detection comparison for raw and cleansed datasets.

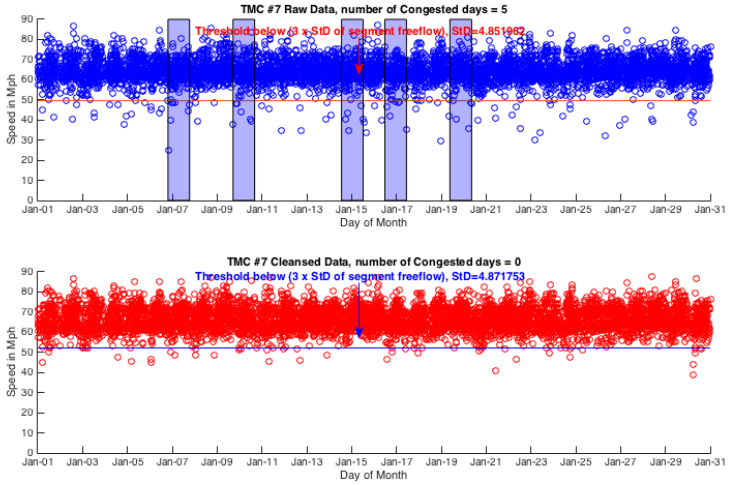


Figure 65 - Segment 7 congestion detection comparison for raw and cleansed datasets.

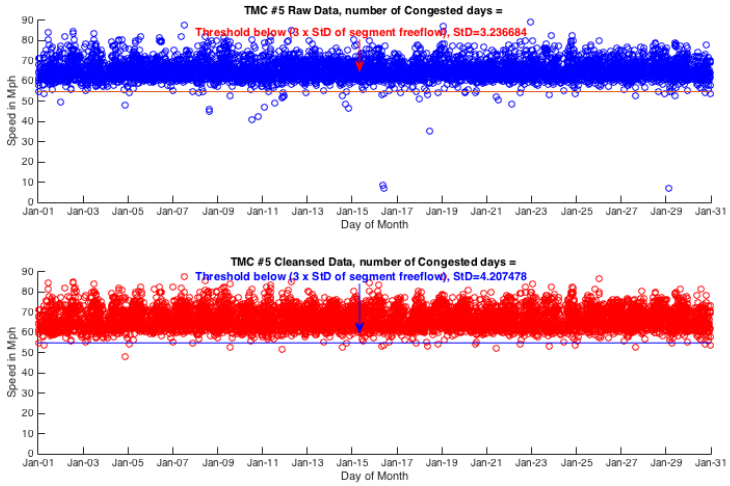


Figure 66 - Segment 6 congestion comparison for raw and cleansed datasets.

2) The second group contains segments detected in both datasets. Characterized by a large difference in the number of congested days, evident when comparing the two datasets (e.g., segments {24, 44, 33, 59, 3, 25, 65, 89, 26, 35, 23, 56, 90, 32, 39, 60, 98, 31, 61, 74, 17, 64, 19, 14, 62, 66, 73, 72, 29, 53, 30, 42, 27, 40, 68}). This group is colored in green. Three random examples are shown in Figure 67, Figure 68, and Figure 69.

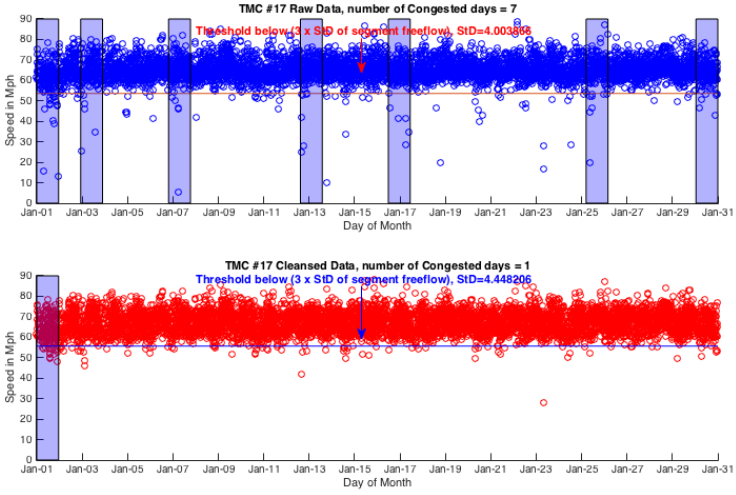


Figure 67 - Segment 17 congestion comparison for raw and cleansed datasets.

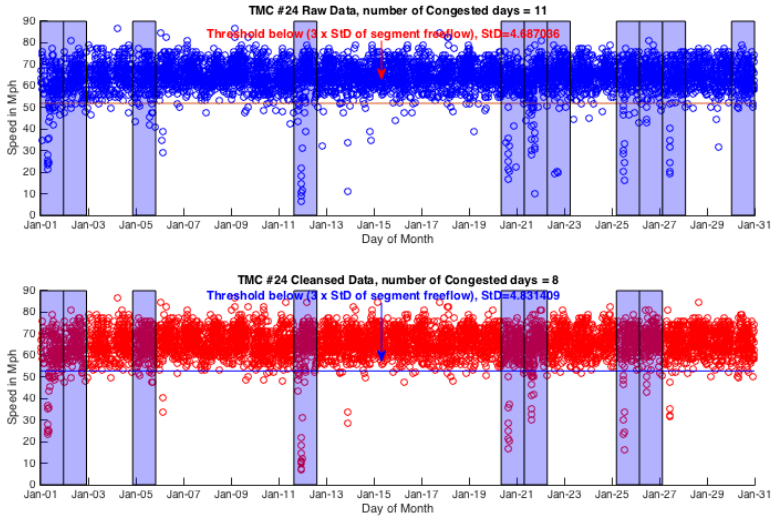


Figure 68 - Segment 24 congestion comparison for raw and cleansed datasets.

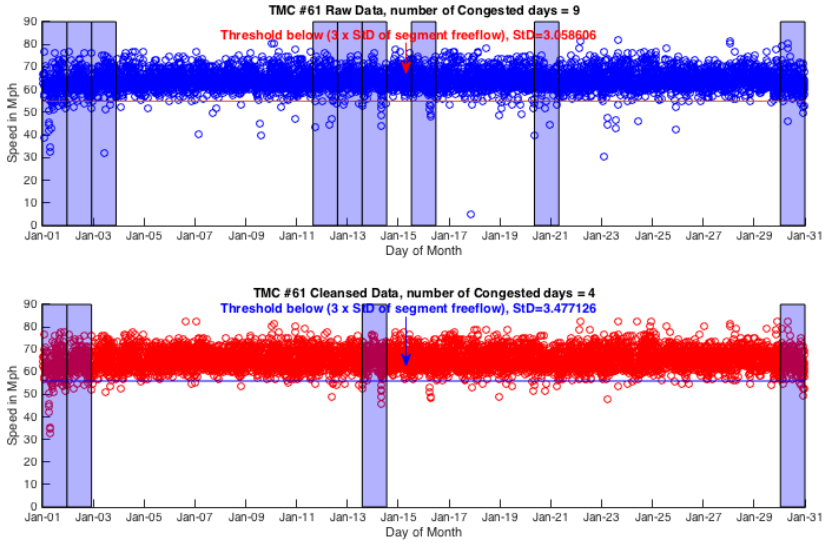


Figure 69 - Segment 61 congestion comparison for raw and cleansed datasets.

It is obvious that outliers were cause for false detection. 3) Includes segments detected in both datasets, characterized by the same or nearly the same number of congested days. This group is colored in white. Two examples of this group were randomly chosen and are depicted in Figure 70 and Figure 71. The cleansed dataset had no improvement over the raw dataset for this group.

As a result, for congestion detection, removal of outliers contributes to the reduction of false detections errors of congested segments and congested days for both variance and thresholding congestion detection methods alike.

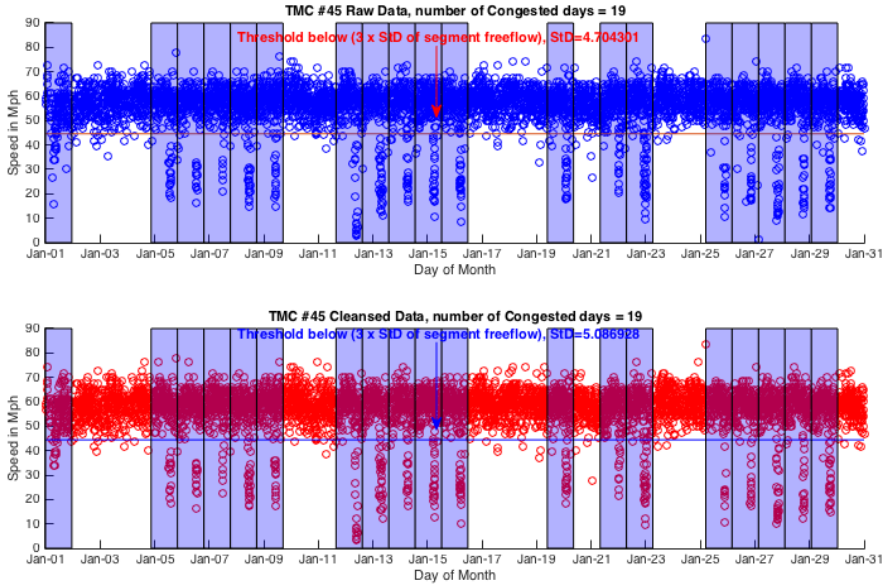


Figure 70 - Segment 45 congestion detection comparison for raw and cleansed datasets.

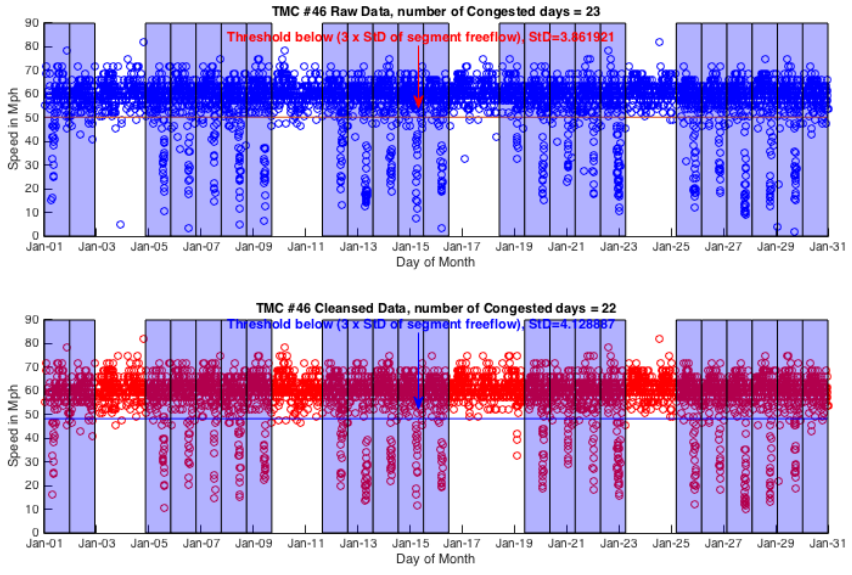


Figure 71 - Segment 46 congestion detection comparison for raw and cleansed datasets.

Chapter 5: Computing Performance Measures

Travel time, speed, and delay are closely related measures that convey the lag commuters experience and the time they expend in order to complete trips on a highway system. The purpose of computing traffic performance measures is to quantify the reliability of a traffic system. This chapter identifies and computes five basic travel time reliability measures which form the necessary building blocks for performance measurement of highway systems. Moreover, the study compares the results attained from these measurements using both the raw and the cleansed datasets demonstrating the effect outlier removal has on results attained.

5.1. Mean free-flow speed and travel time

Mean free-flow speed of a vehicle describes the average travel speed of a motorist driving in low volume traffic conditions in the absence of obstructions, traffic control devices, congestion, or other adverse conditions (e.g., bad weather) on the road [46]. The most typical, congestion-free workday flow for each segment was selected to determine free flow speed of each segment. Weekdays were first filtered from all days of the month, and then the highest mean, lowest variance day was identified. After the appropriate day was selected, standard deviation, variance, and mean measurements were recorded. Gaussian model fitting was performed.

Table 11 shows the segment-length weighted-average free flow speed, variance, and standard deviation of the datasets. The combined length, weighted-average speed limit for all segments was 67.007 mph. Both datasets showed mean free flow speed on I-35 southbound was very close to the weighted-average speed limit of the roadway. The raw

dataset had a slightly lower average speed than the cleansed dataset. Appendix A details the mean, variance, and standard deviation for each segment of I-35 southbound.

Table 11 – Free flow speed statistical measures for I-35 southbound.

	Cleansed Dataset	Raw Dataset
Mean:	67.13850 mph	64.31812 mph
Variance:	19.2384 mph	13.43946 mph
Std:	4.3590 mph	3.5999 mph

The maximum difference of the datasets relative to average free flow speed was 5.76332 mph for segment 96. Authors conclude, albeit minor, outlier removal has an impact on statistical analysis results for the NPMRDS. Figure 72 shows the difference of raw and cleansed mean free flow speed for all segments on I-35.

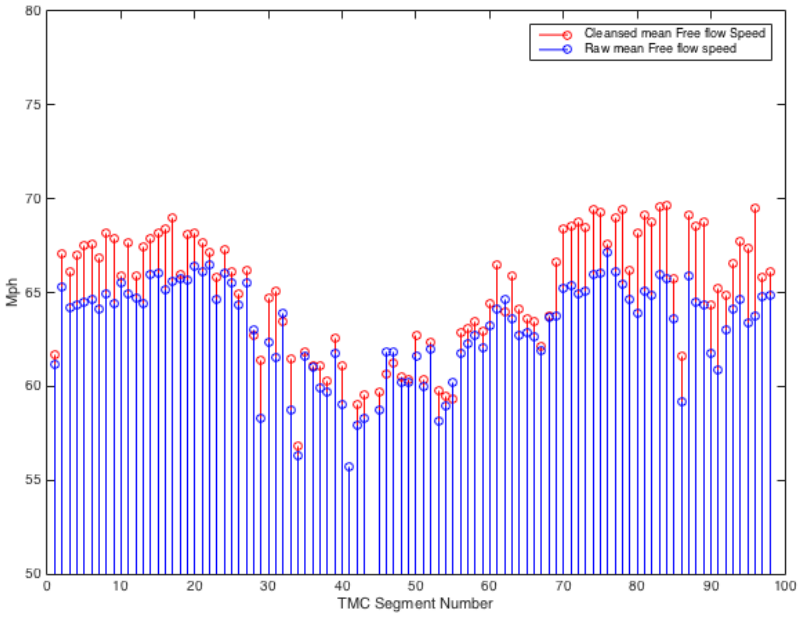


Figure 72 – Mean free flow speeds for all I-35 segments.

Mean travel time per segment was derived utilizing segment length obtained from the NPMRDS static file. Appendix A lists the mean travel time for each segment of I-35 southbound. The difference between the datasets for mean free flow travel of each

segment is small, yet notable. Measures for both datasets are shown in ascending length in Figure 73.

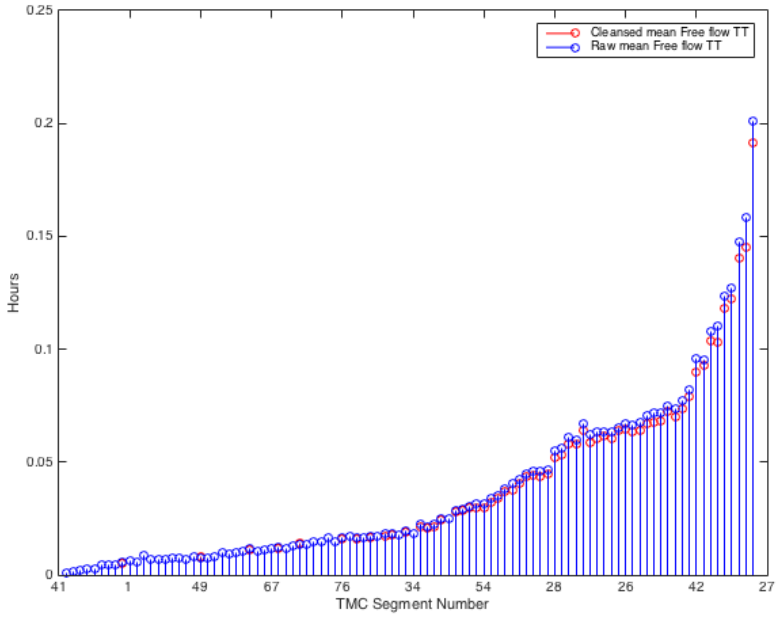


Figure 73 - Free flow travel time for I-35 southbound segments.

5.2. 85th percentile

Traffic engineers and transport planners typically use the 85th percentile speed as a key parameter. Standards like AS1742.4; traffic engineering text books; and federal reports [47], [48] define the 85th percentile speed as "*The speed at or below which 85% of all vehicles are observed to travel under free flowing conditions past a nominated point.*" [48]. The concept of the 85th percentile was first discovered in a comprehensive study entitled "*Accidents on main rural highways related to speed, driver, and vehicle*" conducted by David Solomon in the late 50s and early 60s. Findings were released in 1964 [49]. Figure 74 shows the Solomon curve, which is a graphical representation of collision rate of automobiles as a function of their speed compared to the average vehicle

speed on the same road. [49] The lowest collision rate conforms to the smallest variation from the average.

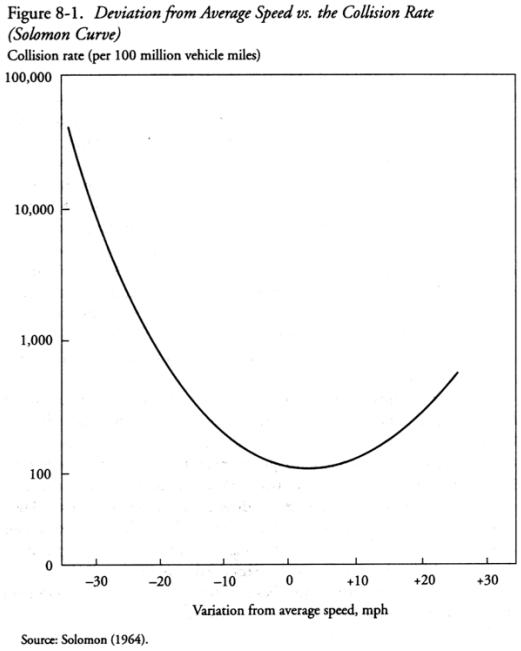


Figure 74 - Solomon Curve [49].

Several subsequent studies have been conducted, and each has reached similar conclusions. Thus, it is well documented that fewer and less severe collisions occur when speed limits are set near the 85th percentile. This practice is based on the premise that the majority of drivers are reasonable and prudent; want to avoid a crash; and desire to reach their destination in the shortest time possible. A speed at or below 85 percent of that which most people drive at any given location under good weather and visibility conditions is considered the maximum safe speed for that location.

Statistical techniques show that a normal probability distribution will occur when a random sample of traffic in free flow is measured [45]. Frequency distribution curves demonstrate that a certain percentage of drivers travel faster than conditions warrant. Likewise, a certain percentage of drivers travel at unreasonably slow speeds relative to

traffic trend. Most cumulative speed distribution curves "break" at approximately 15 percent and 85 percent of the total number of observations [45]. Consequently, motorists traveling in the lower 15 percent are considered to be traveling unreasonably slow, and those observed above 85 percent are assumed to be exceeding a safe and reasonable speed. Posting a speed below the 15 percent value would penalize a large percentage of reasonable drivers. The 85th percentile speed is considered a desirable characteristic of traffic for conforming to a speed limit that is considered safe and reasonable.

In this work, the 85th percentile segment value was found subsequent to detecting free flow values. Free flow Gaussian models leveraged Cumulative Distribution Functions (CDFs) to detect the 85th percentile. An example of this process is shown in Figure 75. Figure 76 shows segment 73 when using the cleansed dataset. 85th percentile speed was 72.7mph.

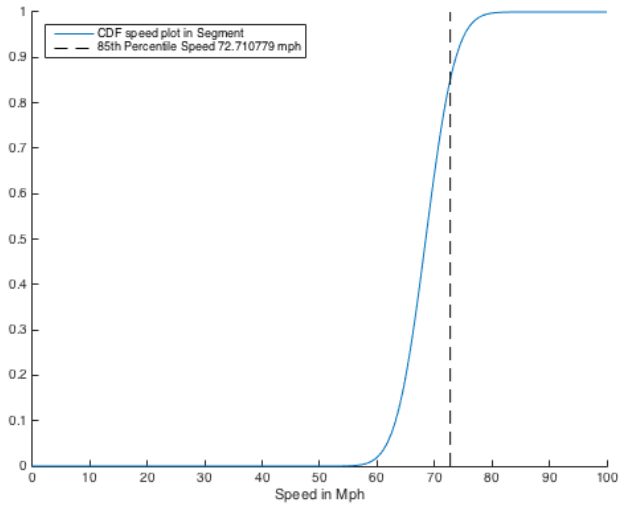


Figure 75 – Segment 73 CDF with 85th percentile speed. (Cleansed dataset).

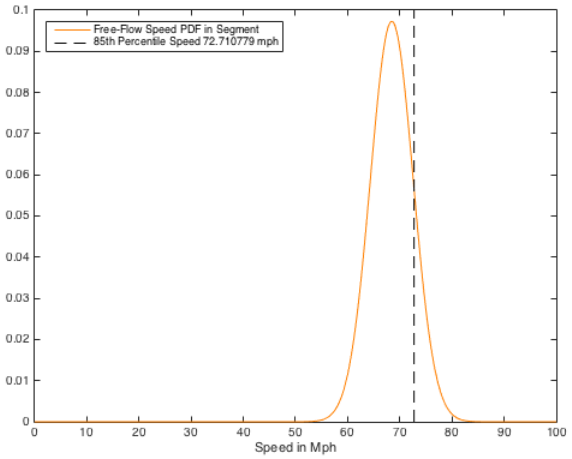


Figure 76 – Segment 73 PDF with 85th percentile speed (cleansed dataset).

The weighted mean 85th percentile for all segments of I-35 southbound were 68.0492 and 71.6563 mph for the raw and cleansed datasets, respectively. Appendix B illustrates the 85th percentile for each segment of I-35 southbound for both datasets. Figure 77 shows a stem plot depicting the 85th percentile of both datasets for all segments of I-35 southbound. A noticeable difference can be seen between 85th percentile results attained with and without the application of outlier removal measures.

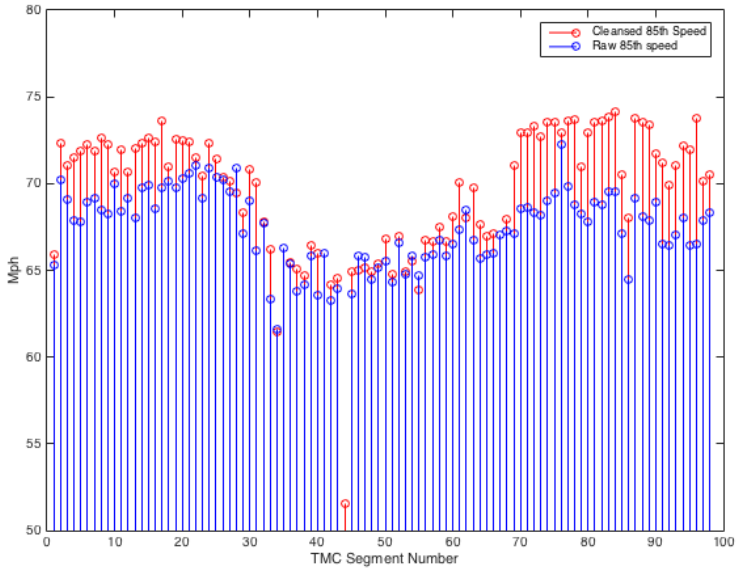


Figure 77 - I-35 85th percentile per segment.

5.3. Travel Time (TT) index,

Travel Time Index (TTI) compares peak period travel conditions to free-flow conditions. In other words, it is the ratio of measured travel time during average congestion to required travel time for the same trip at free-flow speeds. For example, a TTI of 1.3 indicates a 20-minute free-flow trip required 26 minutes [50].

$$TTI = \frac{TT_{MeanCongestion}}{TT_{FreeFlow}}$$

Appendix C reports TTI results per segment in each dataset.

The worst TTI value in the raw dataset was 5.1921 for segment 41, translating the 2.5690 second free-flow travel time to 13.3382 seconds. Segment 75 had the least congestion with a TTI of 1.031025, translating its 227.2721 second free-flow time to 234.3232 seconds. In general, free-flow travel time for I-35 southbound from state border to state border—distance of 236.06537 miles over all segments—was 3 hours and 18.76 minutes. Total TTI measured for all segments was 1.244, resulting in total travel time of 4 hours and 7.28 minutes.

For the cleansed dataset, the worst TTI was 5.0830 for segment 41, which is actually quite similar to the raw dataset. Free-flow travel time of 2.5690 translated to 12.97 seconds. Segment 65 had the best TTI of 1.0371, increasing its 63.98017 second free-flow to 66.35 seconds. Notably, both datasets indicated segment 41 had the worst TTI. However, each set indicated a different segment as having the best TTI, primarily because outlier points were removed in the cleansed dataset. See Figure 78 for a dataset comparison of outliers removed for segment 65.

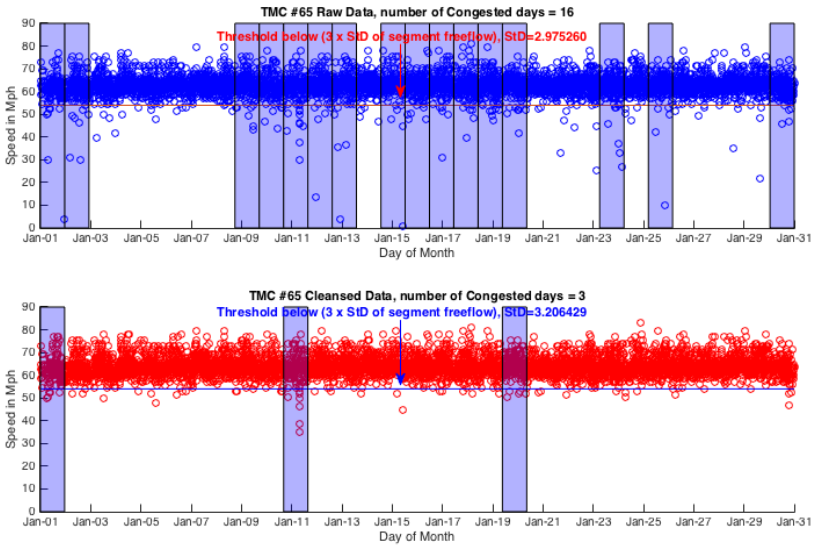


Figure 78 – Segment 65 comparison between cleansed and raw datasets.

For the cleansed dataset, free flow travel time for I-35 southbound from border to border was 3 hours and 11.7344 minutes. Total TTI in this dataset was 1.166, resulting in 3 hours and 43.685 minutes total travel time with congestion. Figure 79 illustrates results obtained using Google maps destination route information. Free flow travel time without congestion is estimated at 3 hours 13 minutes, which is very close to results from the cleansed dataset. Table 12 details a comparison of both datasets. Figure 80 illustrates TTI per segment for I-35 southbound for both datasets.

Table 12 – Free flow speed statistical measures for I-35 southbound.

	Cleansed Dataset	Raw Dataset	Google Maps
No-Congestion time	3 hours 11.7 mins	3 hours 18.7 mins	3 hours 13 mins
Normal Traffic time	3 hours 43.6 mins	4 hours 7.28 mins	3 hours 22 mins

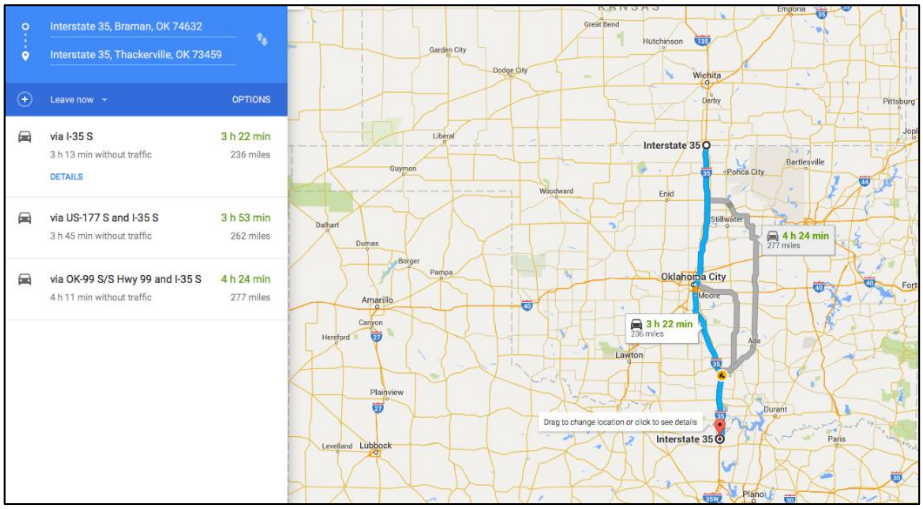


Figure 79 - Google maps route results for I-35 southbound. January 12, 2016.

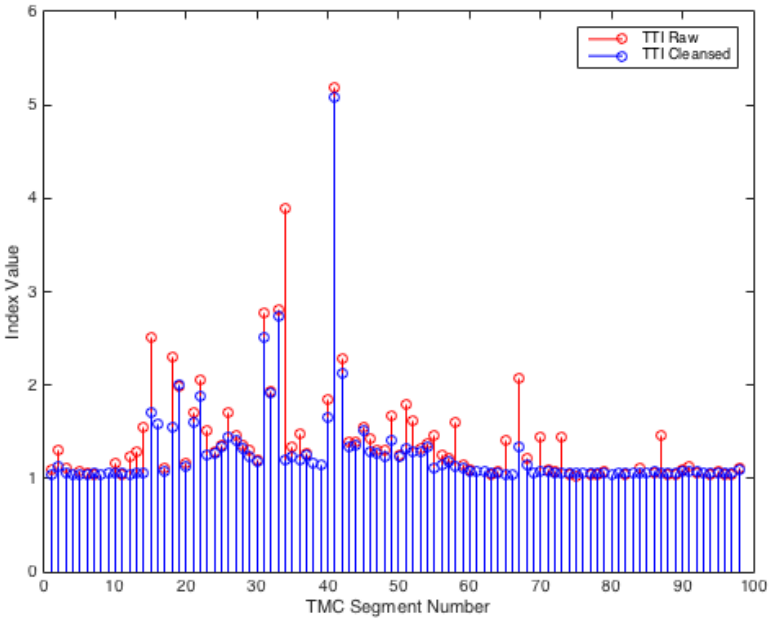


Figure 80 - Segment TTI comparison for raw and cleansed datasets.

5.4. Buffer Index (BI)

The Buffer Index (BI) represents the amount of time most travelers add to their average travel time when planning trips to account for any unexpected delay and ensure on-time arrival. BI is expressed as a percentage, and its value increases as reliability worsens. For example, a BI of 40% means that, given average travel time of 20-minutes,

a traveler should budget an additional 8 minutes to ensure on-time arrival most of the time (e.g., (20 minutes × 40% = 8 minutes buffer time). BI is computed as the difference between the 95th percentile travel time and average travel time, divided by the average travel time [51] and represents a near-worst case travel time.

Whether expressed as a percentage or in minutes, buffer time is the extra time a traveler should allow to arrive on-time for 95 percent of all trips. A simple analogy explains that a commuter who uses a 95 percent reliability indicator would be late only one weekday per month [51].

Figure 81 illustrates results per segment for I-35 southbound for both raw and cleansed datasets. Appendix C shows numerical results per segment per dataset.

$$BI = \frac{TT_{95\%} - TT_{MeanCongestion}}{TT_{MeanCongestion}}$$

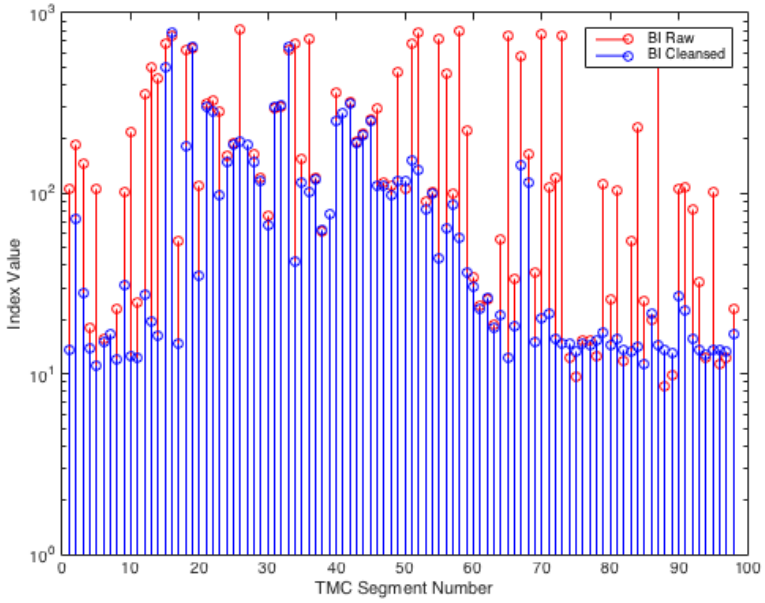


Figure 81 - BI for all segments I-35 raw and cleansed dataset.

5.5. Planning Time Index (PI)

Planning Time Index (PI) represents total travel time that should be planned when including adequate buffer time. PI differs from BI in that both typical delay and unexpected delay are included in the calculation. Thus, PI compares near-worst case travel time to light or free-flow traffic time. For example, given that PI is 1.60, total travel time for a 15-minute trip in light traffic should be 24 minutes (e.g., 15 minutes × 1.60 = 24 minutes). PI is useful for directly comparing the TTI measure of average congestion on similar numeric scales. PI is computed as the 95th percentile travel time divided by the free-flow travel time [51]. Figure 82 illustrates results per segment for I-35 southbound for both raw and cleansed datasets. Appendix C shows the numerical results per segment per dataset.

$$PI = \frac{TT_{95\%}}{TT_{FreeFlow}}$$

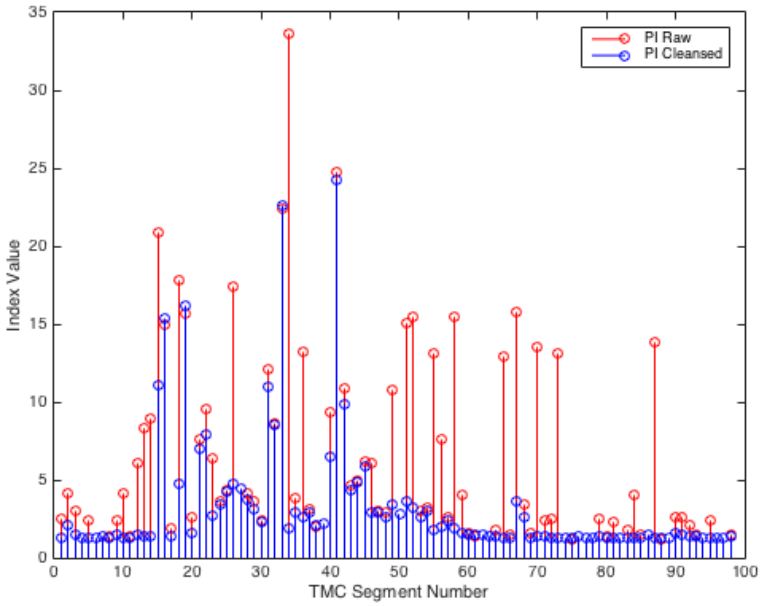


Figure 82 - PI for all I-35 segments, raw and cleansed datasets.

BI and PI statistics are significantly affected by outliers. Figure 82 shows a substantial difference between datasets for many segments. Figure 83 illustrates congestion comparison between datasets for segment 65. For the 15th of January a near 0 mph speed measurement was recorded in the raw dataset. Average travel time in the raw dataset for the 15th was 92.33 seconds. When the outlier was removed, average travel time for the cleansed dataset became 64.631seconds. Moreover, 85th percentile travel time was 78.088 seconds in the raw dataset and became 71.499 seconds in the cleansed dataset. Similarly, Figure 84 shows a near zero speed in the raw dataset for segment 34, which was removed in the cleansed dataset. A substantial effect is evident in the 95th percentile travel time of the raw dataset (See Figure 85).

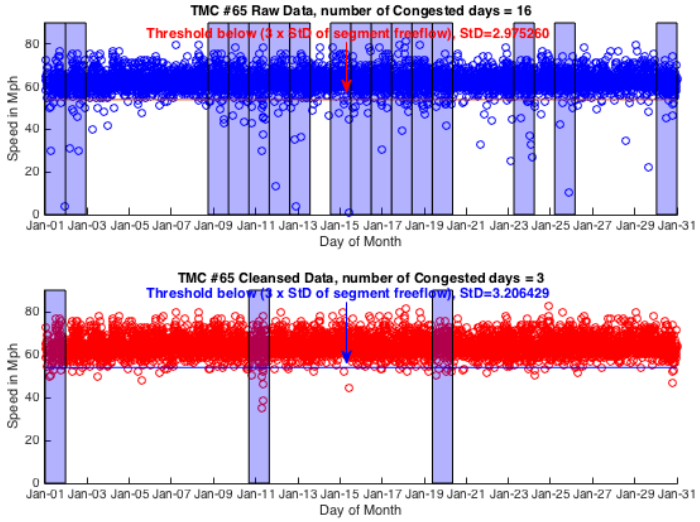


Figure 83 - Segment 65 Congestion comparison between raw and cleansed datasets.

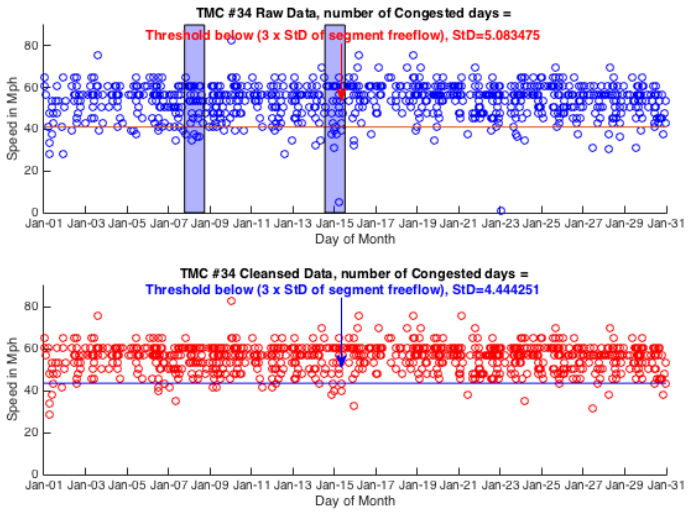
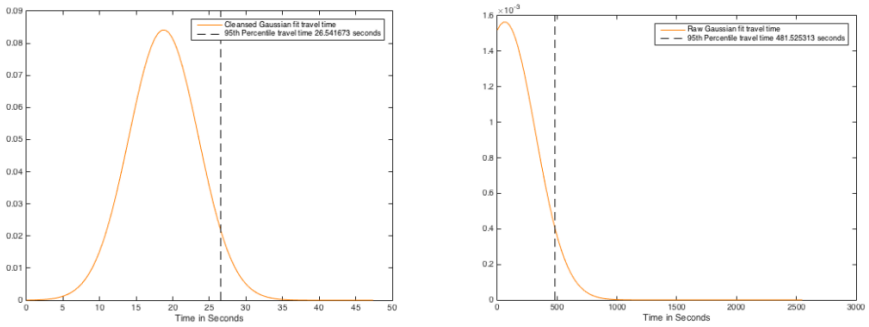


Figure 84 - TMC 34 January 2015 speed scatter plot.



(a) (b)

Figure 85 - 95th percentile travel time for (a) cleansed (b) raw dataset.

Chapter 6: Traffic Analytics

Chapter 6 builds on congestion analysis results detailed in previous chapters and introduces work focused on obtaining insight and extracting knowledge from traffic attributes, patterns, and characteristics evident in congested segments observed. First, inherent segment congestion groups and clusters are identified. Next, an optimum classifier is constructed to automatically classify congestion, given common characteristics found in all segments. Finally, a demonstration-of-concept is provided for identifying non-recurrent congestion using a Bayesian inference engine, which estimates likelihood models of non-recurrent congestion sources using traffic speed probe data measurements.

6.1. Congestion clustering

Results from previous congestion detection can be used to map the sole one-dimensional segment speed data—used as input criteria for subsequent learning stages—into a three-dimensional segment congestion data vector consisting of segment ID, number of congested days, total congestion duration, and number of consecutive congestion repetitions. A matrix of all data vectors can then be constructed. Total congestion duration is estimated as the number of congested epochs multiplied by the duration of an epoch (e.g., 5 minutes). Congestion Repetition (CR) represents congestion recorded on two or more consecutive days for any segment. A higher CR number is indicative of a congestion pattern in any given segment.

6.1.1. Cluster identification

Unsupervised exploratory data analysis was performed to gain insight from reported results for sections previously identified with traffic congestion. The goal was to understand the inherent group structure of congested segments and their common characteristics. Implementing the widely used hierarchical clustering approach provides visual assessment toward predicting the number of clusters intrinsic in the data. The advantage of this approach is that an initial estimate or assumption about the number of clusters is not necessary. Strategies for implementing hierarchical clustering are generally categorized in the following two ways [52].

- Agglomerative: This clustering is described as a "bottom up" approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This clustering is described as a "top down" approach where all observations in one cluster are split recursively as one moves down the hierarchy.

This work presented in this thesis used agglomerative clustering, as the method is faster than divisive clustering. Complexity of the former is $O(n^3)$ compared to $O(2^n)$ for the latter. Several methods exist for linkage criteria:

1- Single link clustering (i.e., nearest neighbor clustering) defines distance between two groups (G,H) as the distance between the two closest members of each group:

$$d_{SL}(G,H) = \min_{i \in G, i' \in H} d_{i,i'}$$

2- Average link clustering measures average distance between all pairs:

$$d_{avg}(G,H) = \frac{1}{n_G n_H} \sum_{i \in G} \sum_{i' \in H} d_{i,i}.$$

Where n_G and n_H are the number of elements in groups G and H. Because average link clustering requires averaging $d_{i,i'}$, any change to the measurement scale can change the result. In contrast, single linkage and complete linkage are invariant to monotonic transformation since they leave the relative ordering the same [52].

3- Complete link clustering (i.e., furthest neighbor clustering) defines the distance between two groups as the distance between the two most distant pairs:

$$d_{CL}(G,H) = \max_{i\in G, i'\in H} d_{i,i'}$$

Because single linkage requires proximity of only a single pair of objects among two groups to be considered close—regardless of the similarity of other members of the group, clusters can be formed that violate the compactness property (i.e., all observations within a group should be similar). Complete linkage, on the other hand, considers two groups close only if all observations in their union are relatively similar, which tends to produce clusters with small diameter (i.e., compact clusters). For this reason, complete linkage was chosen. Figure 86 shows the complete linkage agglomerative clustering performed on the data. Table 13 shows the percentage of total cluster per cut performed on the tree.

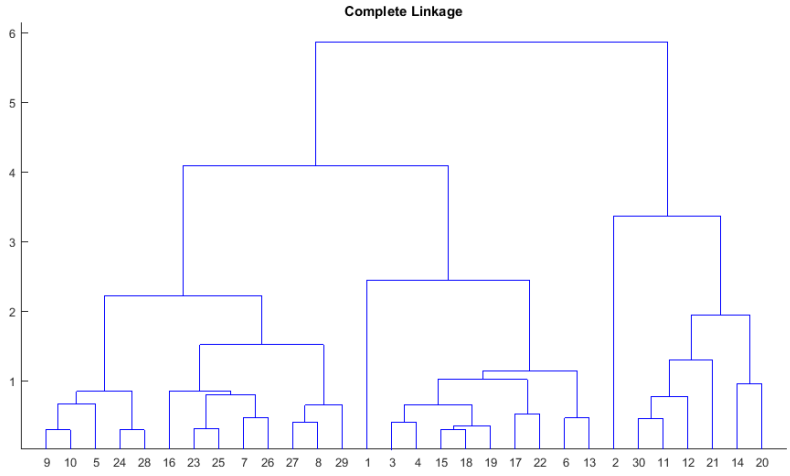


Figure 86 - Complete linkage agglomerative clustering

Table 13 - Clustering results per tree cut of complete linkage tree.

Cut Value	Cluster number	Count	Percent
5.869	1	14	14.29%
	2	84	85.71%
4.12	1	27	27.55%
	2	57	58.16%
	3	14	14.29%
3.362	1	1	1.02%
	2	13	13.27%
	3	27	27.55%
	4	57	58.16%
2.45	1	17	17.35%
	2	40	40.82%
	3	1	1.02%
	4	13	13.27%
	5	27	27.55%

From the dendrogram shown in Figure 86, it can be seen that clusters exist between segment data. Visible "gaps" in the lengths of the links in the dendrogram (representing the dissimilarity between merged groups) allow cluster patterns to be identified indicating the existence of 2 to 4 clusters The exact number of clusters yielding optimum results is unclear. Results of hierarchical clustering align with intuitive assumptions about congestion and the types of congestion patterns typical on roadways. Determining the exact number of clusters remains an open research problem for machine learning in clustering. Fortunately, numerical validation criteria are suitable for evaluating various values of number of clusters [53]. . Hierarchical clustering output provides an initial estimate of a suitable number of clusters available in the data. K-means utilizes this initial estimate as an optimization criterion and objective function to partition the data into K disjoint groups so that the within-group sum-of-squares is minimized. An advantage of an optimization-based method like K-means is that the method scales very well to large data sets [53]. The objective criterion of K-means is used to assess which K yields better results. The same distance metric (i.e., Euclidene distance) used in hierarchical clustering

was implemented in K-means. Thirty repitions for each K value were performed and best results were chosen as representative of a particular K value. Results of K-mean are affected by the initial location of the centroid. Table 14 shows percentage of data-per-cluster using each value of K. Table 15 shows total sum distance calculated for 30 runs for values of K=1, 2, and 3.

Table 14 - K-means clustering for K=2, 3 &4.

K-Value	Value	Count	Percent
K=2	1	81	82.65%
	2	17	17.35%
K=3	1	51	52.04%
	2	30	30.61%
	3	17	17.35%
K=4	1	17	17.35%
	2	32	32.65%
	3	28	28.57%
	4	21	21.43%

Table 15 - 30 Replicative run sum of distance results for K-means; K=2, 3 &4.

Run	K=2	K=3	K=4
Replicate 1, 2 iterations, total sum of distances	10510	2955.14.	1423.96.
Replicate 2, 2 iterations, total sum of distances	11026.2.	2955.14.	1423.96.
Replicate 3, 7 iterations, total sum of distances	11026.2.	2955.14.	1447.39.
Replicate 4, 10 iterations, total sum of distances	11026.2.	2955.14.	1447.39.
Replicate 5, 9 iterations, total sum of distances	11026.2.	2955.14.	1447.39.
Replicate 6, 2 iterations, total sum of distances	10510	2955.14.	1447.39.
Replicate 7, 8 iterations, total sum of distances	11026.2.	2955.14.	1423.96.
Replicate 8, 2 iterations, total sum of distances	10510	2955.14.	1447.39.
Replicate 9, 9 iterations, total sum of distances	11026.2.	2955.14.	1447.39.
Replicate 10, 2 iterations, total sum of distances	10510	2955.14.	1423.96.
Replicate 11, 2 iterations, total sum of distances	11026.2.	2955.14.	1447.39.
Replicate 12, 4 iterations, total sum of distances	11026.2.	2955.14.	1423.96.
Replicate 13, 3 iterations, total sum of distances	10510	2955.14.	1447.39.
Replicate 14, 9 iterations, total sum of distances	11026.2.	2955.14.	1447.39.
Replicate 15, 2 iterations, total sum of distances	10510	2955.14.	1447.39.
Replicate 16, 2 iterations, total sum of distances	10510	2955.14.	1447.39.
Replicate 17, 2 iterations, total sum of distances	10510	2955.14.	1447.39.
Replicate 18, 2 iterations, total sum of distances	10510	2955.14.	1423.96.
Replicate 19, 2 iterations, total sum of distances	10510	2955.14.	1447.39.
Replicate 20, 5 iterations, total sum of distances	11026.2.	2955.14.	1423.96.
Replicate 21, 9 iterations, total sum of distances	11026.2.	2955.14.	1423.96.
Replicate 22, 2 iterations, total sum of distances	10996.8.	2955.14.	1423.96.
Replicate 23, 10 iterations, total sum of distances	11026.2.	2955.14.	1447.39.
Replicate 24, 2 iterations, total sum of distances	10996.8.	2955.14.	1423.96.
Replicate 25, 2 iterations, total sum of distances	11026.2.	2955.14.	1447.39.
Replicate 26, 2 iterations, total sum of distances	10996.8.	2955.14.	1447.39.
Replicate 27, 2 iterations, total sum of distances	11026.2.	2955.14.	1423.96.
Replicate 28, 2 iterations, total sum of distances	10510	2955.14.	1447.39.
Replicate 29, 2 iterations, total sum of distances	10510	2955.14.	1447.39.
Replicate 30, 9 iterations, total sum of distances	11026.2.	2955.14.	1423.96.
Best total sum of distances	10510	2955.14	1423.96

To measure clustering accuracy and find the best K value for clustering, a silhouette plot was implemented. The silhouette plot is a special type of plot that uses output from clustering methods to display a measure of how close each data point is to observations in its own cluster, as compared to observations in other clusters [54]. Kaufman and Rousseeuw (1990) developed the silhouette width to measure the i_{th} observation:

$$sw_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

where a_i is the average distance to all other points in its cluster. b_i is found as follows. First, distance between the *i*-th point and all points in another cluster *c* are averaged, providing a measure of distance between the *i*-th point and cluster *c*. The minimum of these across all clusters is represented by b_i. Silhouette width ranges from -1 to 1. Given that an observation has a silhouette width close to 1, it is considered closer to observations in its own group rather than others, which is the objective of clustering. If an observation has a negative silhouette width, then it is not well clustered. Figure 87, Figure 88, and Figure 89 show the silhouette plot for K=2, 3 and 4.

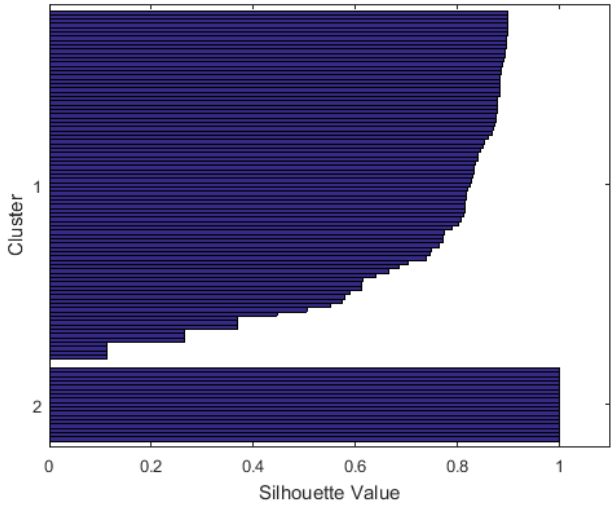


Figure 87 - Silhouette plot for K=2.

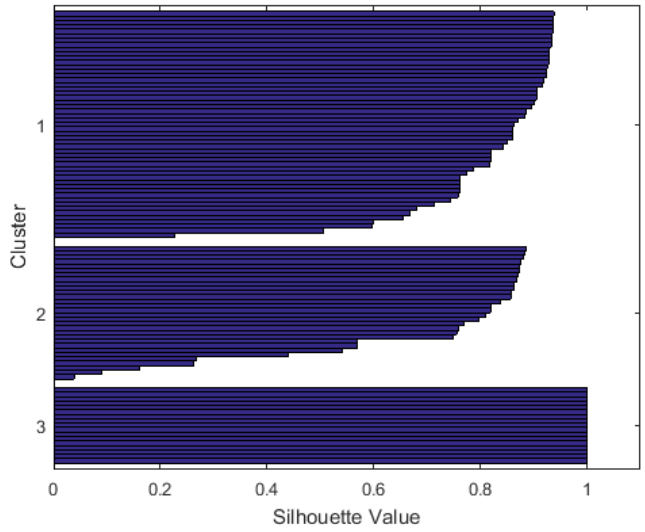


Figure 88 - Silhouette plot for K=3.

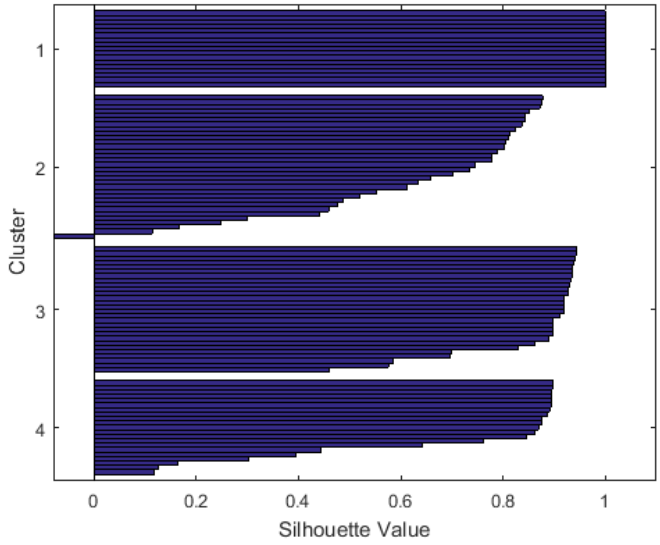


Figure 89 - Silhouette plot for K=4.

Figure 89 indicates that K=4 yields a negative silhouette width for particular segments in cluster 2, indicating that clustering accuracy deteriorated compared to lower values of K. Table 16 shows the average value of each silhouette plot. K=3 results are superior.

Table 16 - Average Silhouette plot value K=2, 3 &4.

	K=2	K=3	K=4
Average Silhouette Value:	0.7791	0.8128	0.7722

6.2. Congestion classification

Clustering analysis was performed to identify common characteristics and formulate labels for segments. Segments were subsequently assigned to individual group IDs based on K-means results with K=3. Table 17 shows clustering results of K-means for each segment with K=3, including the segment to which each cluster ID belongs. Group 2 is primarily centered on segments of I-35 southbound located near OKC. This group tends to have high values for congested days, congestion duration, and congestion runs (i.e., repeated pattern of congestion occurring over the entire month). Results of a scatter-matrix plot are illustrated in Figure 90. Congestion duration and congested days have a linear correlated relationship, indicating congestion duration is similar on separate days. Extracting sample segments and inspecting speed distributions provided further insight into the characteristics of each cluster.

Table 17 - K-means clustering results per segment for K=3.

Cluster Number	Segment Number														
Cluster 1: (51) elements	3	6	7	8	13	14	15	16	17	18	19	20	21	22	23
	25	26	28	31	32	34	38	41	60	61	62	63	64	65	66
	67	68	69	70	71	72	73	74	75	78	80	81	82	84	87
	89	91	94	95	96	98									
Cluster 2: (30 elements)	2	24	27	29	30	33	35	36	37	39	40	42	43	44	45
	46	47	48	49	50	51	52	53	54	55	56	57	58	59	90
Cluster 3: (17 elements)	1	4	5	9	10	11	12	76	77	79	83	85	86	88	92
	93	97													

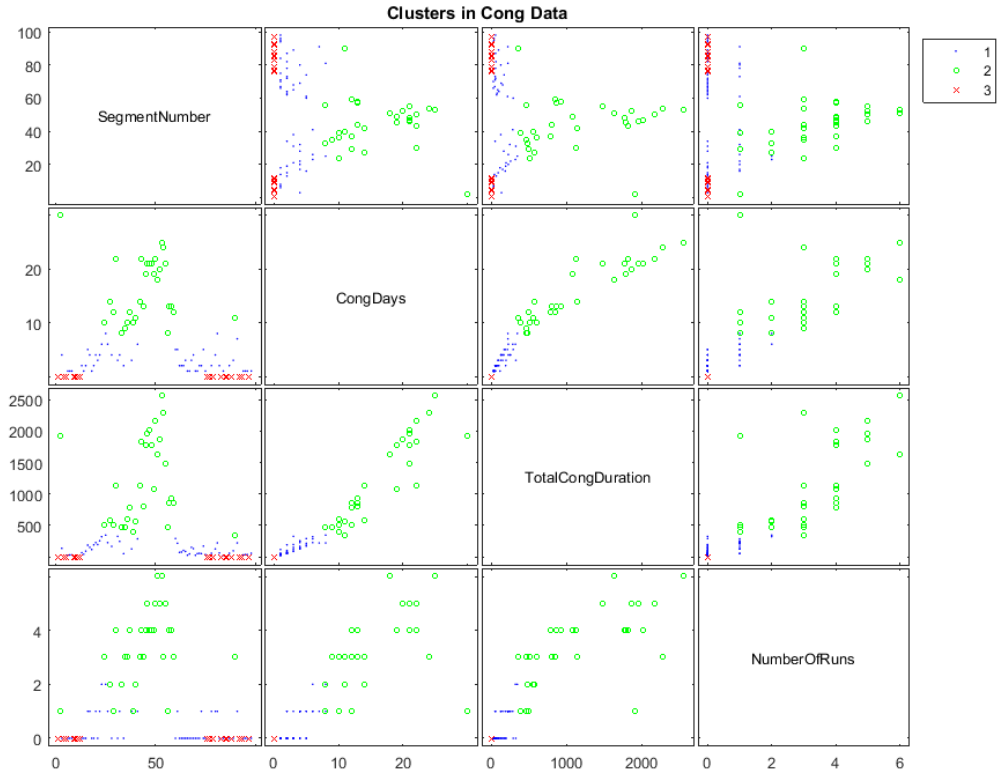


Figure 90 - Scatter matrix plot of input features assigned to clustering group IDs.

A checkerboard plot is constructed in Figure 91 for all segments in cluster 2. The majority of the segments in this cluster have high repetition on congested days. In particular, day 10, 11, 17, 18, 24, and 25 indicate reduced to nearly zero congestion compared with others days of the month. These particular days fall on weekends, suggesting these particular segments experience congestion on a repeated basis during weekdays. Several samples from Figure 91 were extracted for further illustration.

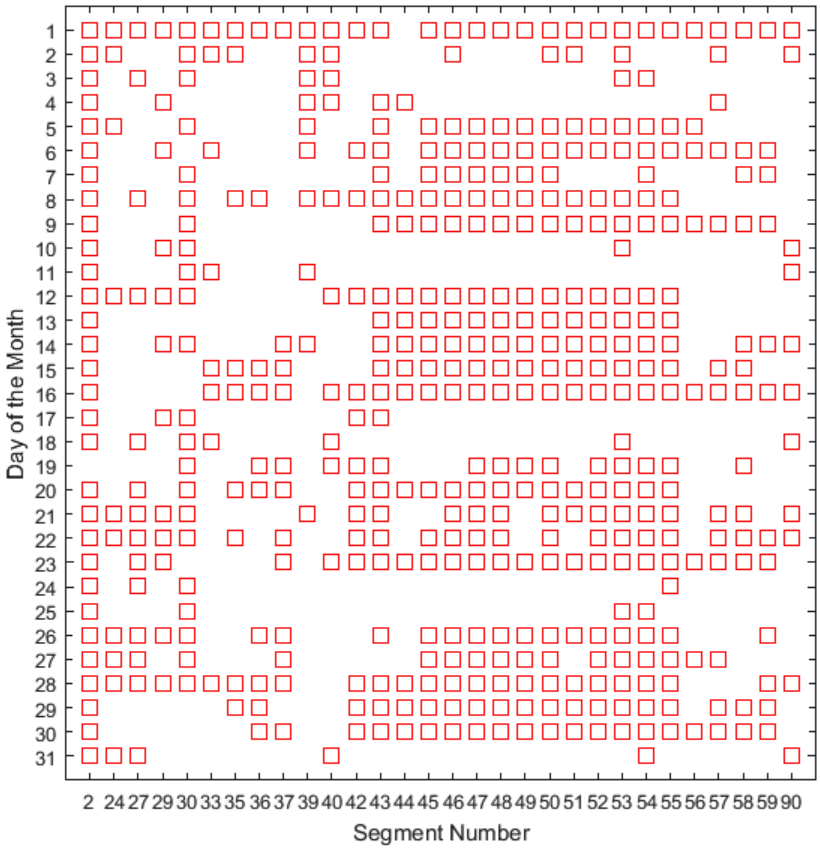


Figure 91 – Checkerboard plot of segments in cluster 2.

Figure 92, Figure 93, and Figure 94 show a scatter plot of speeds for segments 46, 47, and 48 consecutively. Congestion is repetitive to the same extent at the end of the day for weekdays throughout the month. This validates the notion that these segments are experiencing recurrent congestion, as described in Chapter 1. Figure 95, Figure 96, and Figure 97 show scatter plots for segments 27, 29, and 40. These segments also indicate recurrent congestion, but to a lesser extent than the previous example. These clusters share recurrent congestion (Recurrent_Cong) as a common characteristic. This particular cluster is important to ODOT agencies because it indicates segments that have an imbalance between high demand and low capacity during peak travel hours.

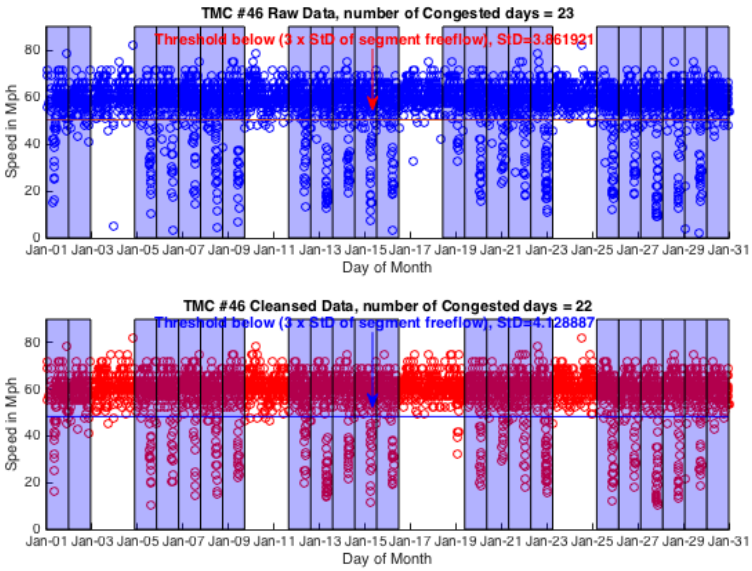


Figure 92 – Scatter plot of speeds segment 46 on I-35 southbound.

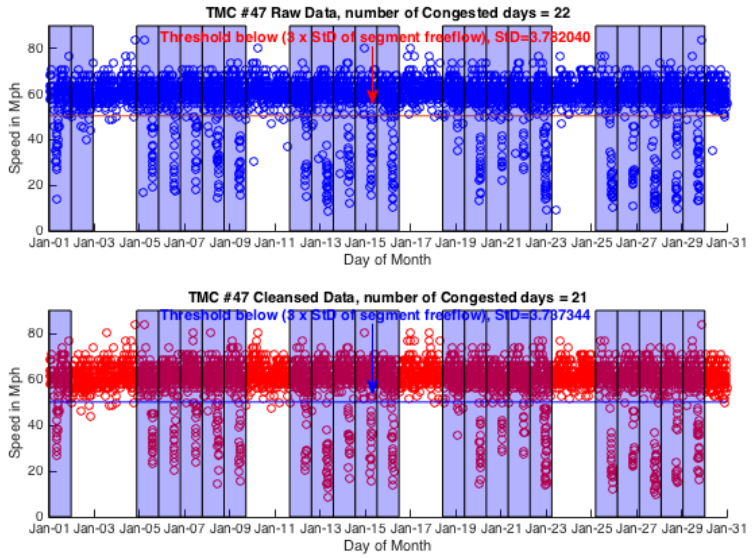


Figure 93 - Scatter plot of speeds segment 47 on I-35 southbound.

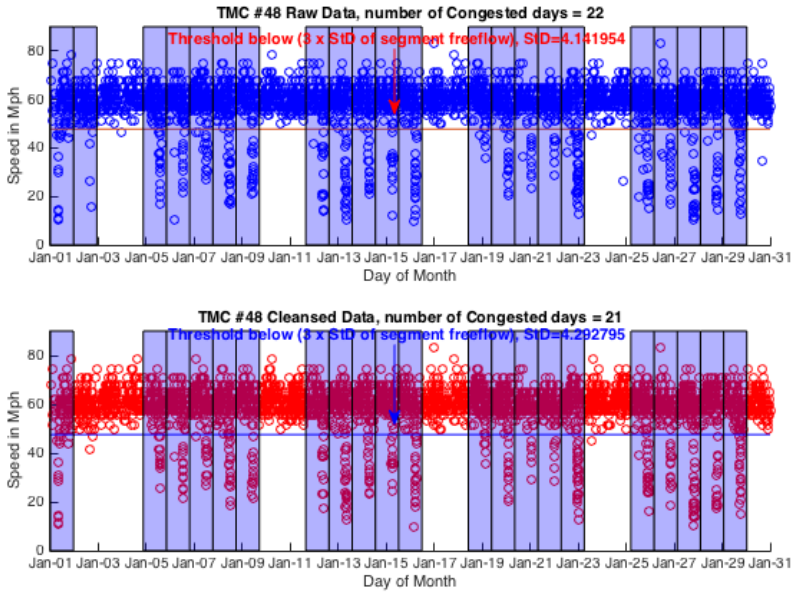


Figure 94 - Scatter plot of speeds segment 48 on I-35 southbound.

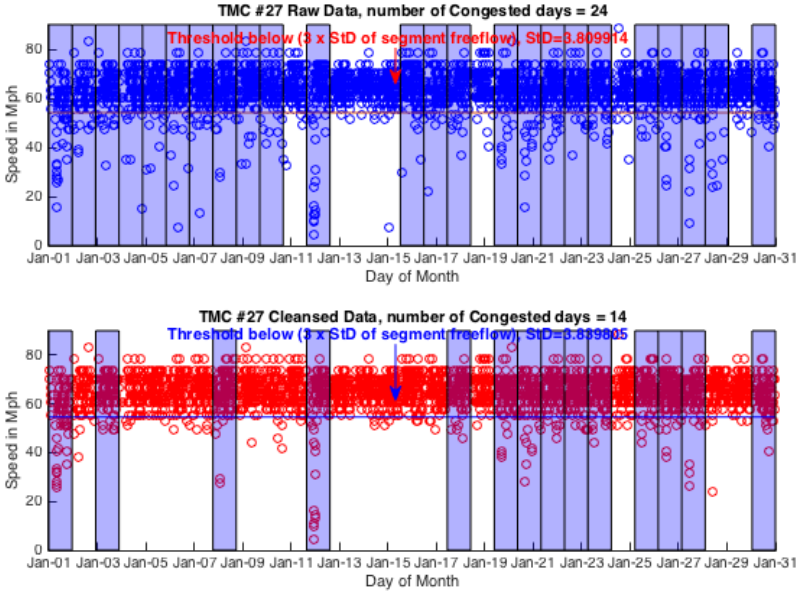


Figure 95 - Scatter plot of speeds segment 27 on I-35 southbound.

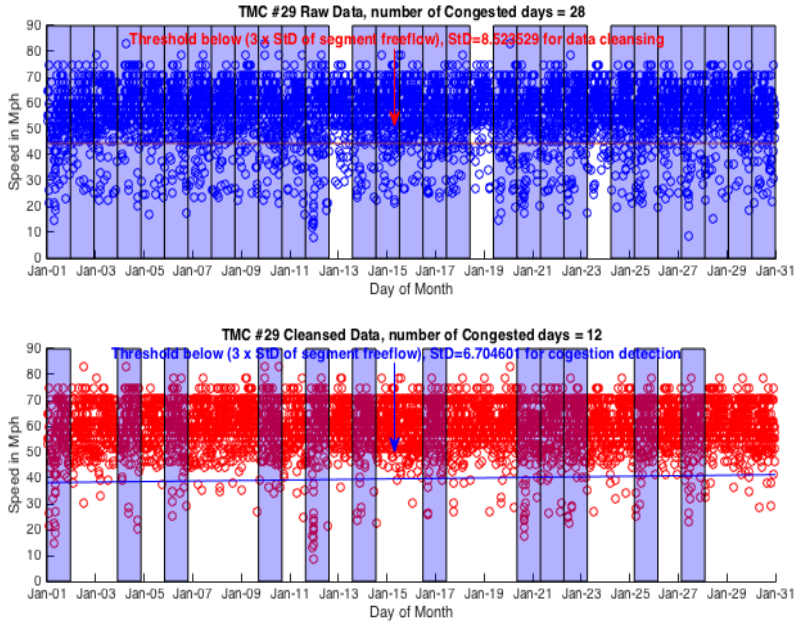


Figure 96 - Scatter plot of speeds segment 29 on I-35 southbound.

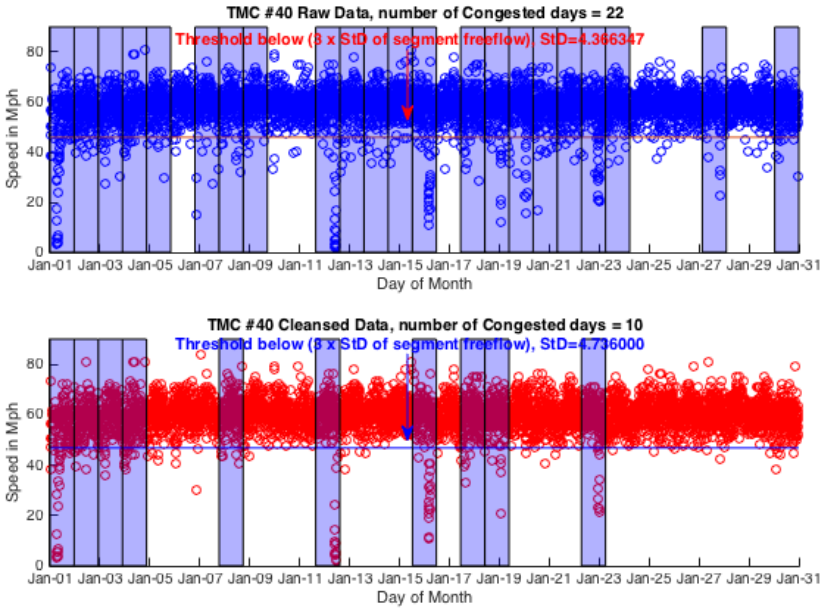


Figure 97 - Scatter plot of speeds segment 40 on I-35 southbound.

.

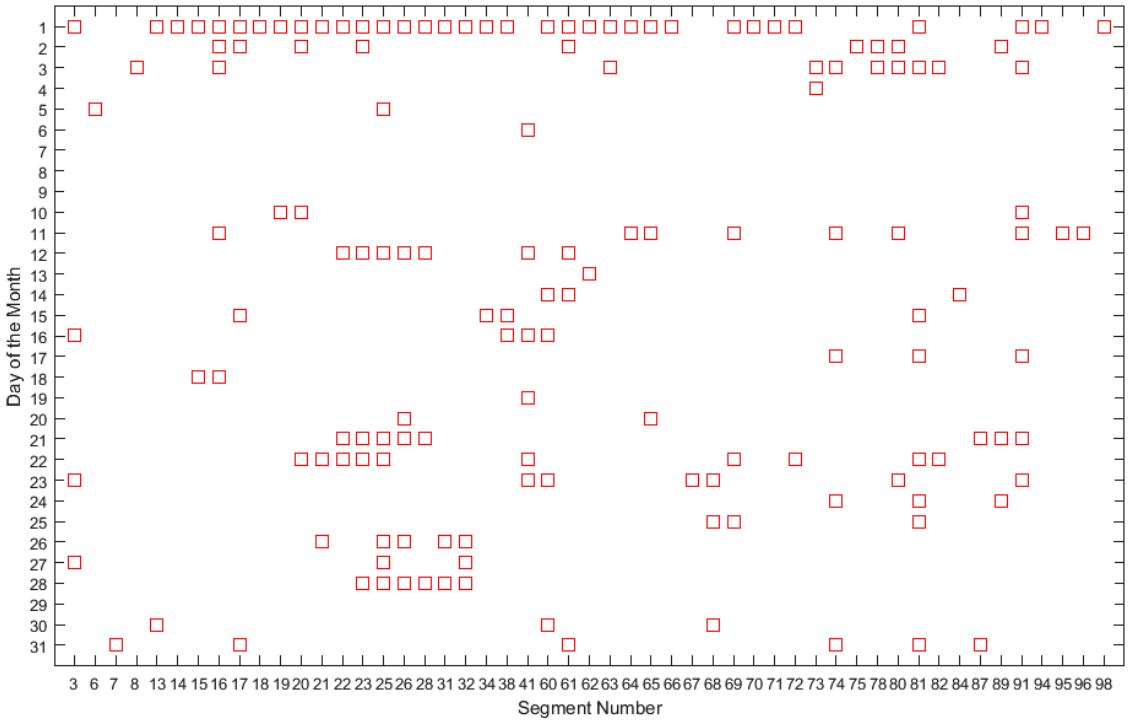


Figure 98 - Checkerboard plot of segments in cluster 1.

Figure 98 depicts a checkerboard plot for segments in cluster 1. Congestion is sparse and is distributed randomly in space and time. Figure 99, Figure 100, and Figure 101 illustrate segments 15, 65, and 23 randomly selected from among the set of segments in cluster 1. It is evident in the cleansed dataset that sparse, non-recurrent congestion occurs in a non-periodic manner. This could be indicative that non-recurrent congestion (Non-Recurrent_Cong) is caused by non-recurring external conditions, such as weather, traffic incidents, or other factors discussed in Chapter 1. Figure 102, Figure 103, and Figure 104 show the checkerboard plot for cluster 3, as well as two randomly selected segment samples—10 and 76. The cleansed dataset indicates there is no congestion (No-Cong) in these segments throughout the month.

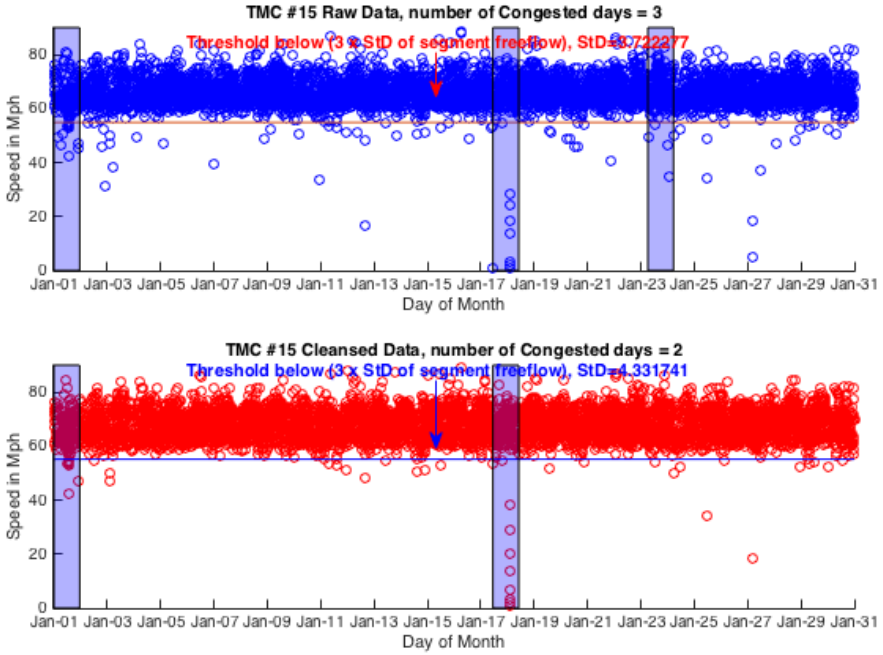


Figure 99 - Scatter plot of speeds segment 15 on I-35 southbound.

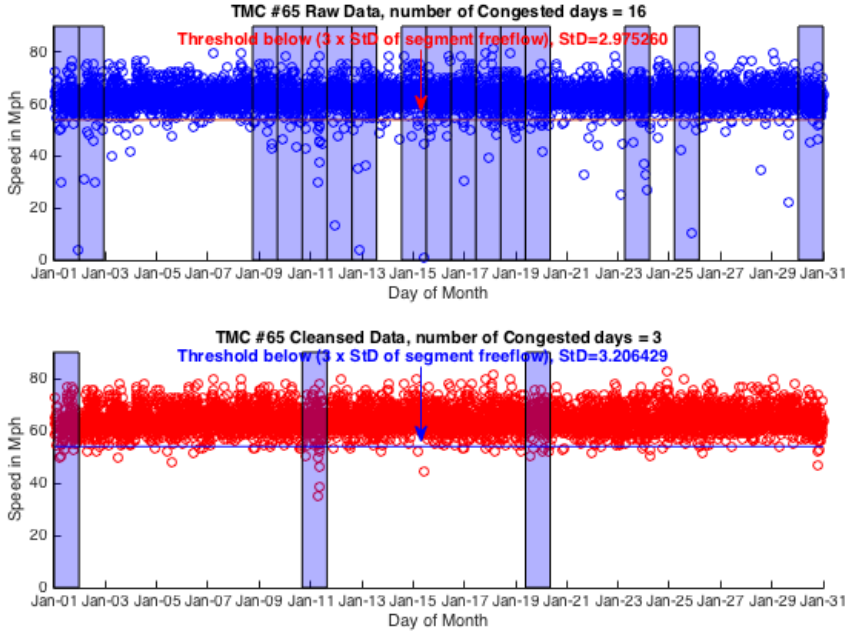


Figure 100 - Scatter plot of speeds segment 65 on I-35 southbound.

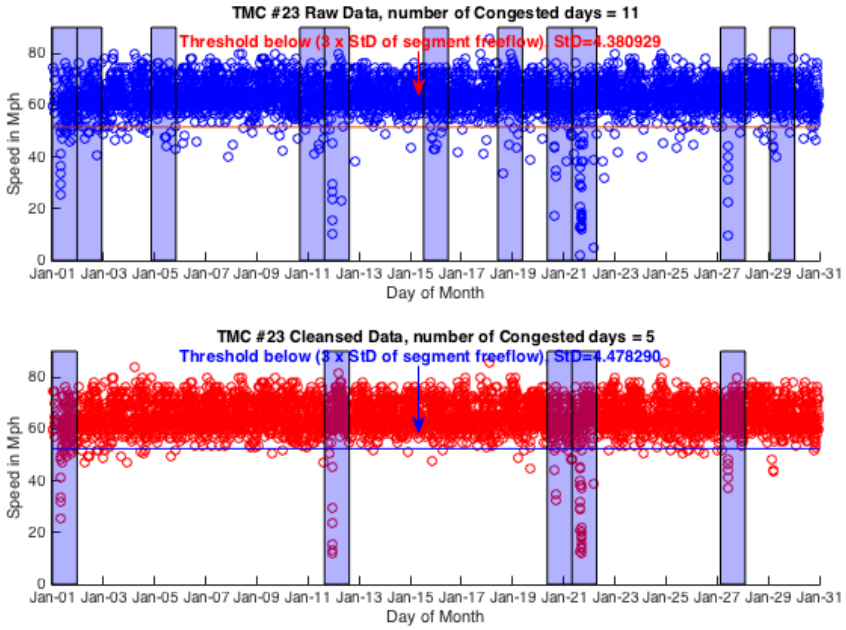


Figure 101 - Scatter plot of speeds segment 23 on I-35 southbound

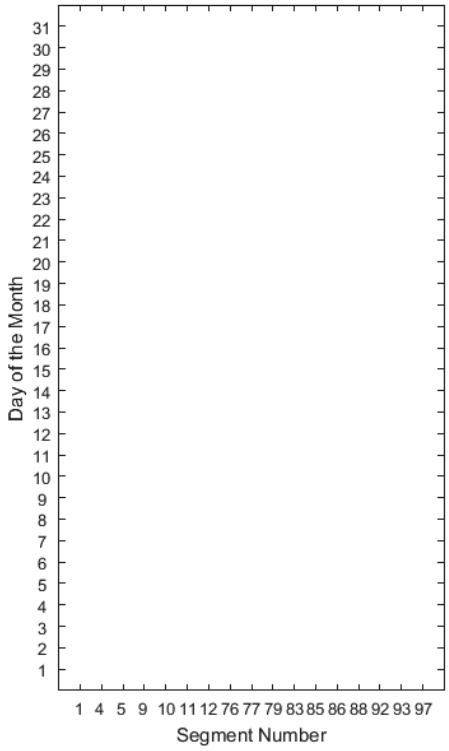


Figure 102- Checkerboard plot of segments in cluster 1.

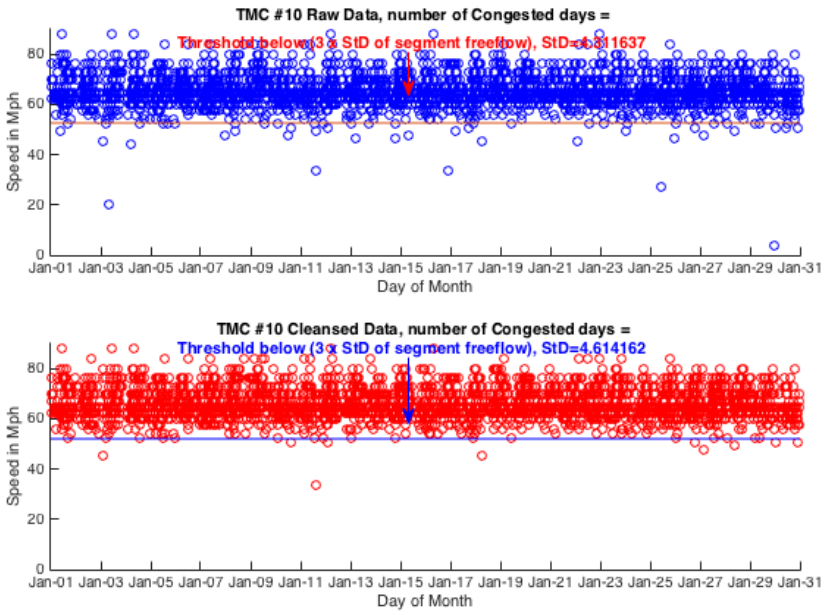


Figure 103 - Scatter plot of speeds segment 23 on I-35 southbound.

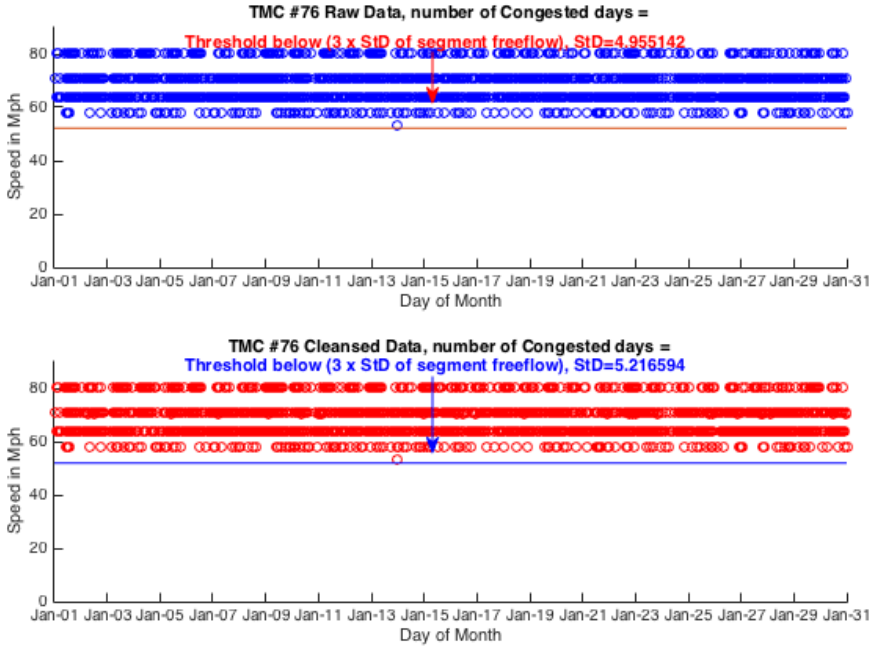


Figure 104 - Scatter plot of speeds segment 23 on I-35 southbound.

6.2.1. Constructing the classifier

Unsupervised learning allows separate groups of segments to be identified as distinguished characteristic clusters, as described in the previous section, namely recurrent congestion, non-recurrent congestion, and no congestion clusters. A classifier for online use was trained by labeling the data accordingly. Supervised learning enables classification of data with a known number of classes. Hence, a classifier was built based on observations of known true class labels (i.e., training the classifier). Classifier performance was assessed using a 10-fold cross validation operation, confusion matrices, and Receiver Operating Characteristic (ROC) measures. Response variable for classification is the class label, and predictor variables are the three aforementioned congestion data features. Results of several classifiers were compared for accuracy, speed, and interpretability of the algorithms under test.

6.2.1.1. Naïve Bayes:

Bayesian methods are highly desirable, as they avoid overfitting [55] by making early assumptions about the likely distribution of the answer. In Bayes decision theory, the classification problem is specified in terms of probabilities. Consider the following notations. Class membership w_m, with $m = 1\ldots M$; features (or variables) are a *p*-dimensional observation vector; and posterior probability of an observation belonging to the *m*-th class is $P(w_m|x_i); m = 1, \ldots, M$.

Bayes' Theorem decomposes the posterior probability as:

$$P(w_m|x_i) = \frac{P(w_m)P(x_i|w_m)}{P(x_i)}$$

Where

$$P(x_i) = \sum_m P(w_m)P(x_i|w_m)$$

Probability $P(w_m)$ is called the prior probability, and $P(x_i|w_m)$ is called the likelihood or class-conditional probability. Prior probability represents the likelihood that an observation is placed into a class without knowledge about the observation, such as measured features discussed earlier. The class-conditional probability is the probability of observing a feature vector x_i given it is in class w_m.

Prior probability can be inferred from prior domain knowledge, estimated from the observed data, or assumed equal across classes. Given these are estimates from the data, prior probabilities are considered the relative frequency of observations in each class.

$$\hat{P}(w_m) = \frac{n_m}{n}$$

where n_m is the number of observations in the m-th class. Accordingly, a naïve Bayes classifier can be constructed to estimate class-conditional probabilities. The approach assumes that individual features are independent, given the class. Therefore, the probability density function for the within-class conditional probability is written as:

$$P(x|w_m) = P(x_1|w_m) \times \ldots \times P(x_p|w_m)$$

In other words, when using data within a class, univariate density for each feature or dimension is first estimated and then multiplied together to obtain joint density. Results from implementing a Naïve Bayes classifier are shown in Figure 105, Table 18, and Table 19.

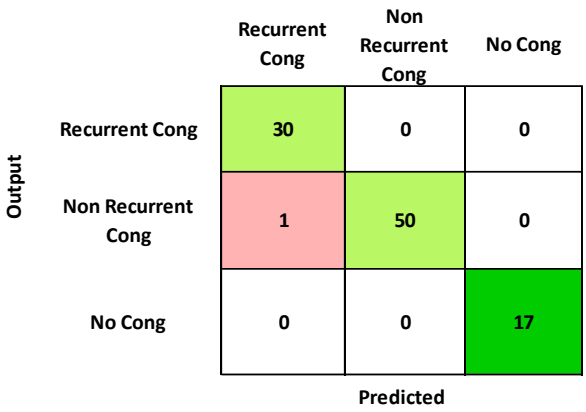


Figure 105 - Confusion matrix for Naïve Bayes classifier.

Table 18 - Detailed accuracy by class for Naïve Bayes classifier.

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1	0.015	0.968	1	0.984	0.976	0.999	0.998	Recurrent_Cong
0.98	0	1	0.98	0.99	0.98	0.993	0.995	Non_Recurrent_Con
1	0	1	1	1	1	1	1	No_Cong
0.99	0.005	0.99	0.99	0.99	0.982	0.996	0.997	**Weighted Avg.**

Table 19 - Classification results of Naive Bayes classifier.

	Naive Bayes Classifier
Time taken to build model	0
Correctly Classified Instances	98.98%
Incorrectly Classified Instances	1.02%
Kappa statistic	0.9832

6.2.1.2. K-Nearest Neighbor (K-NN)

In k-NN classification, output is class membership. Any given object is classified by a majority vote of its neighbors with the object being assigned to the class most common among its k-nearest neighbors [56]. K-NN has strong consistency results: as the amount of data approaches infinity, the algorithm is guaranteed to yield an error rate no worse than twice the Bayes error rate (i.e., the minimum achievable error rate given the

distribution of the data) [56]. Results of implementing K-NN are shown in Figure 106, Table 20, and Table 21.

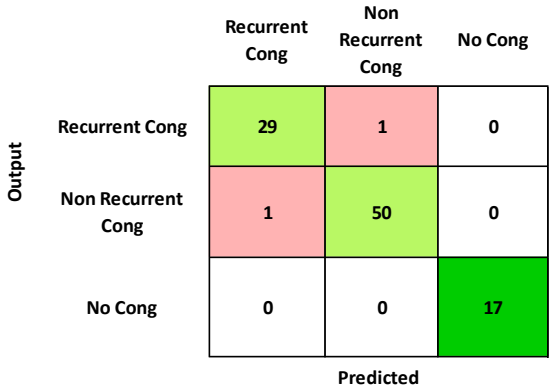


Figure 106 - Confusion matrix for K-NN classifier.

Table 20 - Detailed accuracy by class for K-NN classifier.

Detailed Accuracy By Class								
TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Class**
0.967	0.015	0.967	0.967	0.967	0.952	0.976	0.997	Recurrent_Cong
0.98	0.021	0.98	0.98	0.98	0.959	0.98	0.998	Non_Recurrent_Con
1	0	1	1	1	1	1	1	No_Cong
0.98	0.016	0.98	0.98	0.98	0.964	0.982	0.998	**Weighted Avg.**

Table 21 - Classification results of K-NN classifier.

	K-NN Classifier
Time taken to build model	0
Correctly Classified Instances	97.96%
Incorrectly Classified Instances	2.04%
Kappa statistic	0.9663

6.2.1.3. Decision Trees:

Decision trees are graphs that utilize a branching method to illustrate every possible outcome of a decision. Amongst other data mining methods, decision trees have various advantages [57]:

1- Simple to understand and interpret.

2- Require little data preparation**.** **(**Other techniques often require data normalization, wherein dummy variables must be created and blank values must be removed).

3- Manage both numerical and categorical data.

4- Are robust and perform well even if assumptions are somewhat violated by the true model from which data were generated.

Results of implementing K-NN are shown in Figure 107, Table 22, and Table 23.

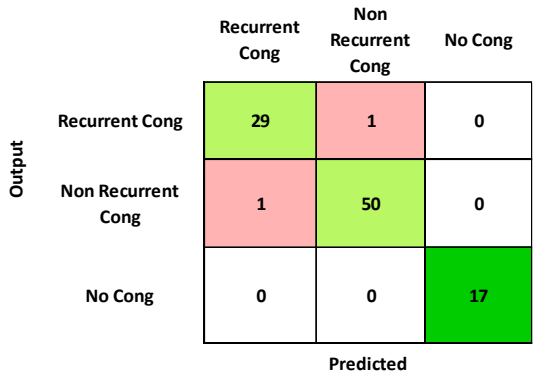


Figure 107 - Confusion matrix for simple decision tree classifier.

Table 22 - Detailed accuracy by class for simple decision tree classifier.

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.967	0.015	0.967	0.967	0.967	0.952	0.976	0.945	Recurrent_Cong
0.98	0.021	0.98	0.98	0.98	0.959	0.98	0.971	Non_Recurrent_Con
1	0	1	1	1	1	1	1	No_Cong
0.98	0.016	0.98	0.98	0.98	0.964	0.982	0.968	**Weighted Avg.**

Table 23 - Classification results of simple decision tree classifier.

	Decision Tree
Time taken to build model	0
Correctly Classified Instances	97.96%
Incorrectly Classified Instances	2.04%
Kappa statistic	0.9663

6.2.1.4. Support Vector Machine (SVM)

In support vector machines (SVMs), a data point is viewed as a p-dimensional vector. SVMs determine the boundary that separates classes with a (p-1) - dimensional hyperplane by as wide a margin as possible. Given that two classes cannot be clearly separated, algorithms determine the best boundary. Such a hyperplane is recognized as the maximum-margin hyperplane, and its classifier is the maximum margin classifier [58]. In addition to performing linear classification, SVMs can efficiently perform non-linear classification using a non-linear kernel. A linear and radial SVM kernel were applied, and results are presented in Figure 108, Table 24, Table 25, and Table 26.

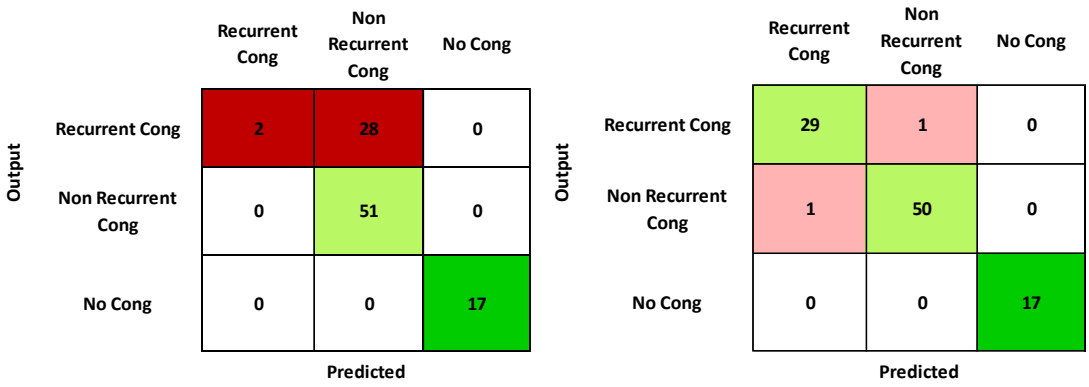


Figure 108 - Confusion matrix for SVM classifier. (a) Radial kernel, (b) Linear kernel.

Table 24 - Detailed accuracy by class for SVM radial kernel.

Detailed Accuracy By Class – Radial kernel

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.067	0	1	0.067	0.125	0.217	0.533	0.352	Recurrent_Cong
1	0.596	0.646	1	0.785	0.511	0.702	0.646	Non_Recurrent_Con
1	0	1	1	1	1	1	1	No_Cong
0.714	0.31	0.816	0.714	0.62	0.506	0.702	0.617	**Weighted Avg.**

Table 25 - Detailed accuracy by class for SVM linear kernel.

Detailed Accuracy By Class – Linear Kernel

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.967	0.015	0.967	0.967	0.967	0.952	0.976	0.945	Recurrent_Cong
0.98	0.021	0.98	0.98	0.98	0.959	0.98	0.971	Non_Recurrent_Con
1	0	1	1	1	1	1	1	No_Cong
0.98	0.016	0.98	0.98	0.98	0.964	0.982	0.968	**Weighted Avg.**

Table 26 - Classification results of SVM classifier.

	SVM (Radial basis)	SVM (Linear basis)
Time taken to build model	0.1	0.03
Correctly Classified Instances	71.43%	97.96%
Incorrectly Classified Instances	28.57%	2.04%
Kappa statistic	0.4749	0.9663

In the end a simple decision tree classifier was chosen based on Occam's razor. Factors for this choice include minimal execution time for the 2-diminesional classifier and the high interpretability. The classifier is illustrated in Figure 109. Clearly, two predictors suffice to classify segments with a 98% accuracy.

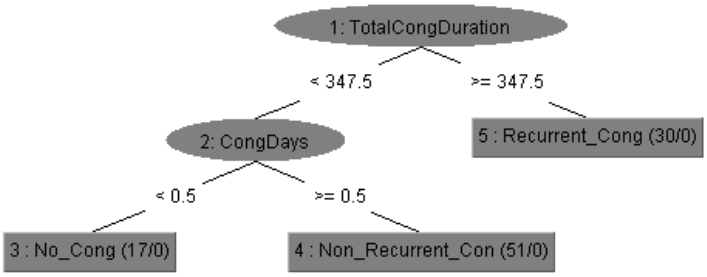


Figure 109 – Simple decision tree classifier.

6.3. Congestion Identification

Three distinct categories of congestion, namely recurrent, non-recurrent and no congestion, were identified. Re-current congestion is largely caused by lack of roadway capacity to support demand or load delivered onto it. Moreover, this lack is more apparent (i.e., more so than its non-recurrent counterpart) given a repeated pattern-like characteristic. Identifying recurrent congestion is important for DOTs and passengers

alike. DOTs recognize re-current congestion as segments and sections of roadway that require improvement, conditioning, and optimization to support present and future demand. Passengers gain an understanding that recurrent congestion should be considered to adequately plan travel time and make route decisions to circumvent delays caused during peak operating hours. Non-recurrent congestion, on the other hand, is sparse and disjoint in nature, thus difficult to identify. Causes of non-recurrent congestion are numerous, and identifying the source of each non-recurrent cause in a timely manner remains an open research problem. Understanding non-recurrent congestion is vital to alleviate its negative effect on traffic performance. Furthermore, obtaining insight on the effect and impact of various sources of non-recurrent congestion allows adequate trip planning, buffer time, and necessary resources to enhance travel time and improve traffic performance in an efficient, holistic manner. The next section presents a Bayesian probability approach to identify underlining operating conditions that cause non-recurrent congestion on a roadway.

6.3.1. Bayesian probability and Bayesian methods

Bayesian probability, in contrast to frequentist probability that interprets probability as the long run frequency or propensity of some phenomenon, is a quantity assigned to represent a state of knowledge or a state of belief [59]. Bayesian probability expresses a subjective degree of belief that rationally changes over time accounting for new evidence. A comical depiction of both [60] is shown in Figure 110.

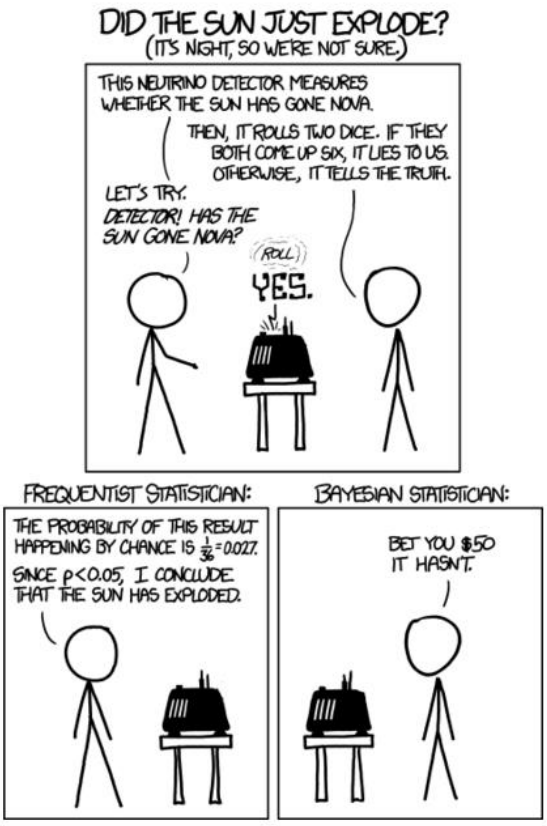


Figure 110 - Frequentists vs. Bayesians

Bayesian inference relies on Bayesian probability as a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. As aforementioned, Bayesian inference derives the posterior probability as a consequence of two antecedents, a prior probability, and a "likelihood function" derived from a statistical model for the observed data. Bayesian inference computes the posterior probability according to Bayes' theorem:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

where $|$ denotes a conditional probability; H is a hypothesis whose probability may be affected by data; evidence E corresponds to new data not used in computing the prior probability; and $P(H)$ is the prior probability indicating a previous estimate of the

probability that a hypothesis is true before gaining current evidence. Thus, $P(H|E)$ is the posterior probability that tells us what we want to know: the probability of a hypothesis given the observed evidence. $P(E|H)$ is the probability of observing E given H, and $P(E)$ is the marginal likelihood.

The primary advantage of using Bayesian methods is that they incorporate probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, Bayesian methods offer several advantages for data analysis [61]. First, because they encode interdependencies between variables, they can manage instances in which data is missing. Second, the methods have the ability to represent causal relationships. Therefore, they can predict consequences of an event or action. Lastly, because the methods have both causal and probabilistic relationships, they can model problems given a need to combine prior knowledge with current data.

Bayesian methods have previously been incorporated in travel time prediction studies [62] [63]. Bayesian Network (BN) models have been used in accident severity analysis [64] [65], non-recurrent incident detection [66], and other traffic studies. Recently, real-time crash prediction using BN has been investigated [67] [68] [69]. Non-recurrent weather conditions have also been extensively studied and modeled by researchers. Studies concerning the effect of weather on traffic flow [70] [71] [72] and the impact of different weather conditions with the temporal and spatial variations of traffic have been reported [73] [74]. A survey of numerous weather characteristics and their effect on traffic can be found in [75]; weather forecasting and prediction using BN have also been heavily investigated [76], [77], [78] [79].

However, the majority of previous studies have adopted speed data collected from AVI sensors or loop detectors for modeling and analysis. Volume/occupancy/speed parameters were used to predict the likelihood of incidents. These measures are invalid for roads on which traffic conditions are estimated using only speed data extracted from sampled floating cars or smart phones (i.e., probe traffic). Furthermore, probe-based data permits collecting information in remote and non-urbanized locations where conventional data acquisition instruments and data collection stations are not available. Because current and historical traffic conditions could be factors used to predict future traffic conditions, it is vital to identify re-current congestion sources in locations that do not have adequate data acquisition sources. For example, weather data is not available at all roadway locations: weather stations are densely located in and around metropolitan areas and large cities, but few are located on stretches of highways connecting cities. Notably, because Bayesian forecasting revises the state of a priori knowledge with a posterior distribution per condition given real-time measurements of TT, a Bayesian system can promptly respond to real time changes in traffic pattern [80].

6.3.2. Identification using Bayesian probability

Various non-recurrent conditions characterize the manner in which vehicle speed is affected on road segments and routes. These conditions correspond to a variety of characteristic models, the impact of which are clearly visible and identifiable on the baseline distribution. Thus, distinguishable statistical models can be used to reveal assorted information for each condition. Combining distribution models with Bayesian probability, an approach can be determined to identify the underlining condition occurring in both offline and real-time speed analysis. This thesis proposes a Bayesian

engine for congestion identification. The engine utilizes statistical models derived from observed data records per condition, and then estimates a posterior credibility for each hypothesis. Figure 111 illustrates this concept for identifying three situations: incident (e.g., cash and collision), weather (e.g., snow) and free-flow traffic. Histogram distributions of speeds are used to create distribution density models from travel time data.

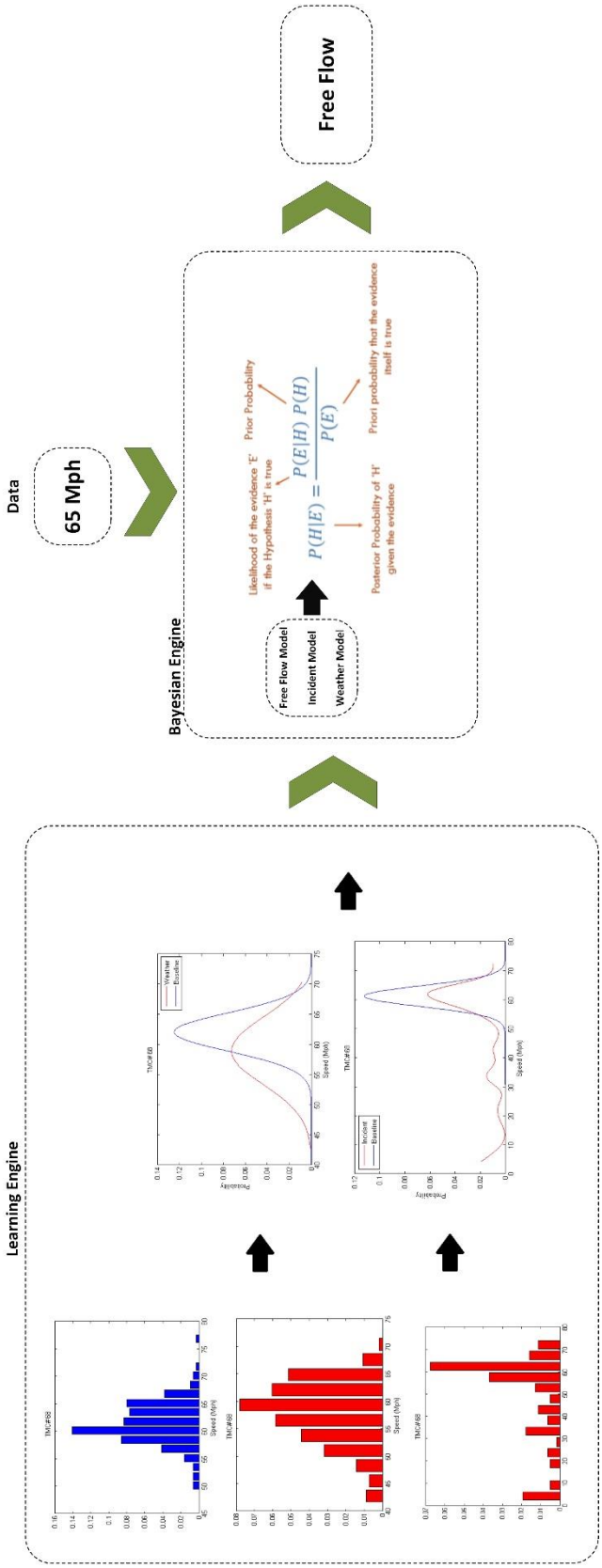


Figure 111 – Bayesian inference engine concept illustration

6.3.3. External Data Sources

A major challenge to implementing the proposed concept is obtaining accurate and reliable information of non-recurrent congestion sources collocated with each segment data for NHS highways. Although extensive weather data is available in major metropolitan areas, this type of information is not collected on segments located near border crossings, un-populated rural areas, or large stretches of highway. Furthermore, categorical historical data (e.g., snow, hail, fog, visibility, thunderstorm) is neccassary for identification. The amount of this data type is rarely stored by weather data centers. Instead, temperature levels, perceipitation levels, wind direction, wind speed, and other numerical weather indicators are typically captured and retained.

Historical categorical and numerical weather data for I-35 southbound was obtained online from www.wunderground.com. Only 10 sensors are used to report data for the entire 236 mile stretch of roadway across the state. Accordinlgy, there are concerns about data accuracy for segments located a signficant distance from weather sensor locations. Furthermore, segments around OKC experiencing recurrent congestion were excluded from analysis. This further reduced the number of sensors available for use in the analysis. Figure 112 illustrates weather senser locations. Incident data was obtained online from www.navibug.com, a website that relies on crowdsourced information collected from users in realtime, as well as aggregated online information captured from local news agency reports. Figure 113 illustrates a snapshot of the online archive and news reports.

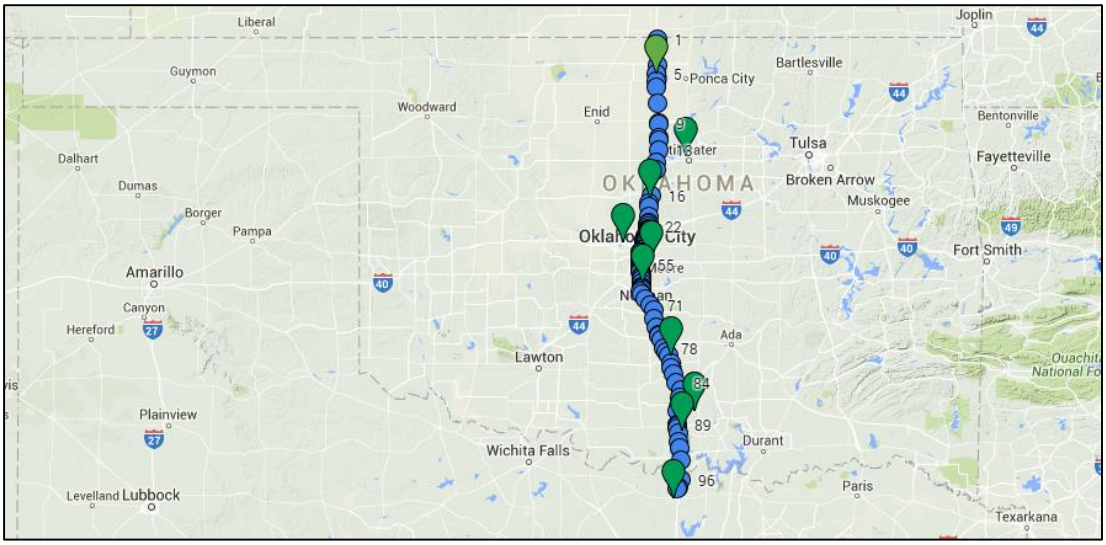


Figure 112 – I-35 southbound segments and weather station locations.

A public government database of historical incidents in not available online. Although discussions are currently underway with ODOT to provide access to collected and stored highway incident information from public safety agencies, the information was not available at the time this thesis was prepared.

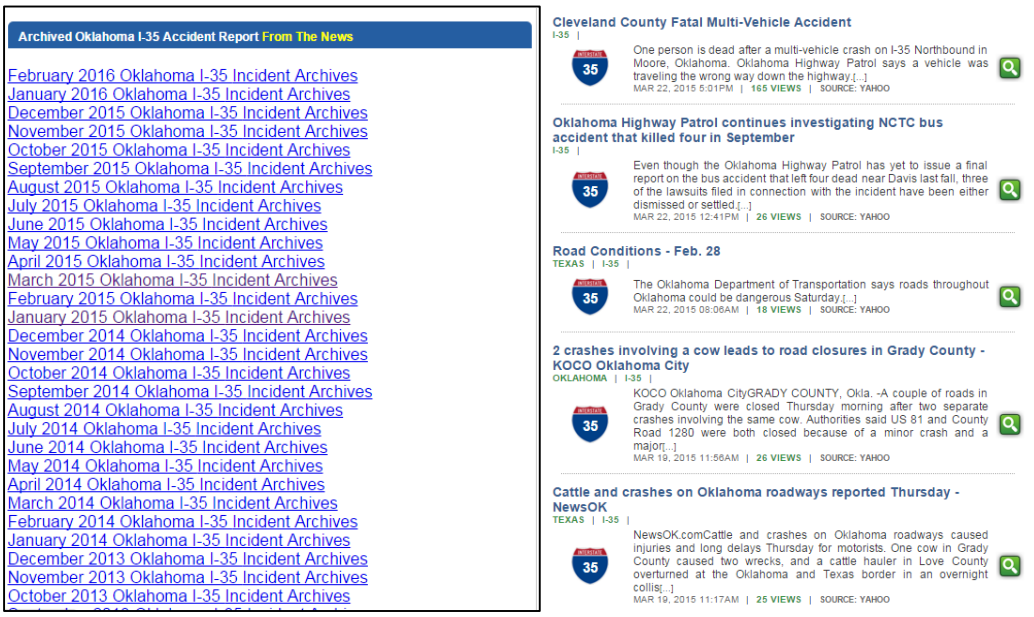


Figure 113 – Incident data online arhive found at www.navibug.com.

As a result, the set of data samples collected for weather and incident data is not adequate for accurate results. Instead, a demonstration of concept is shown in the remainder of this work. Future work will include a through validation of the approach detailed below for a larger sample size when data is available.

6.3.4. Implementation of Bayesian congestion identification

Models pertaining to three distinct conditons, namely free flow, incident, and weather (snow) were constructed. A particular segment was chosen to derrive distribution models, which were subsequently used to evaluate the proposed Bayesian identification approach on additional segments for all three non-recurrent conditions cases. Segment 64, west of Norman on highway I-35 (See Figure 114) is located proximate to a dedicated weather sensor gathering accurate historical weather data.

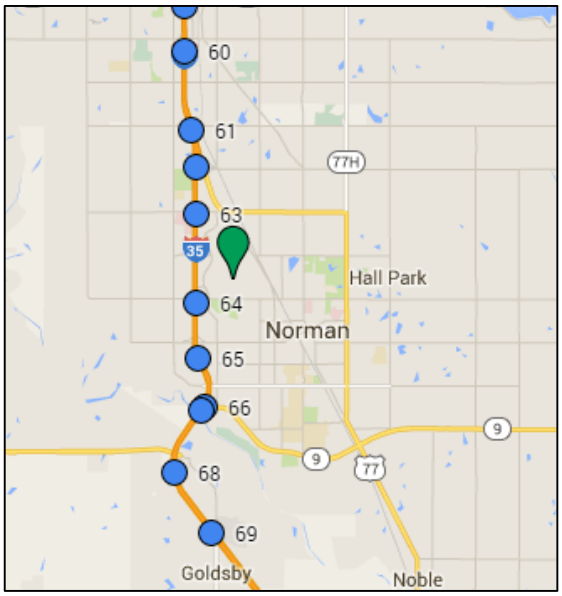


Figure 114 - Segment 64 on I-35.

A historic crash (incident) event occurred on segment 64 on March 13, 2015. Time and location were confirmed by local news agencies' online articles. A snow fall event

occurring on March 4 was also confirmed in such a manner. No congestion (i.e., free flow traffic) was observed on March 2. Figure 109 illustrates the monthly epoch plot for segment 64 during March 2015.

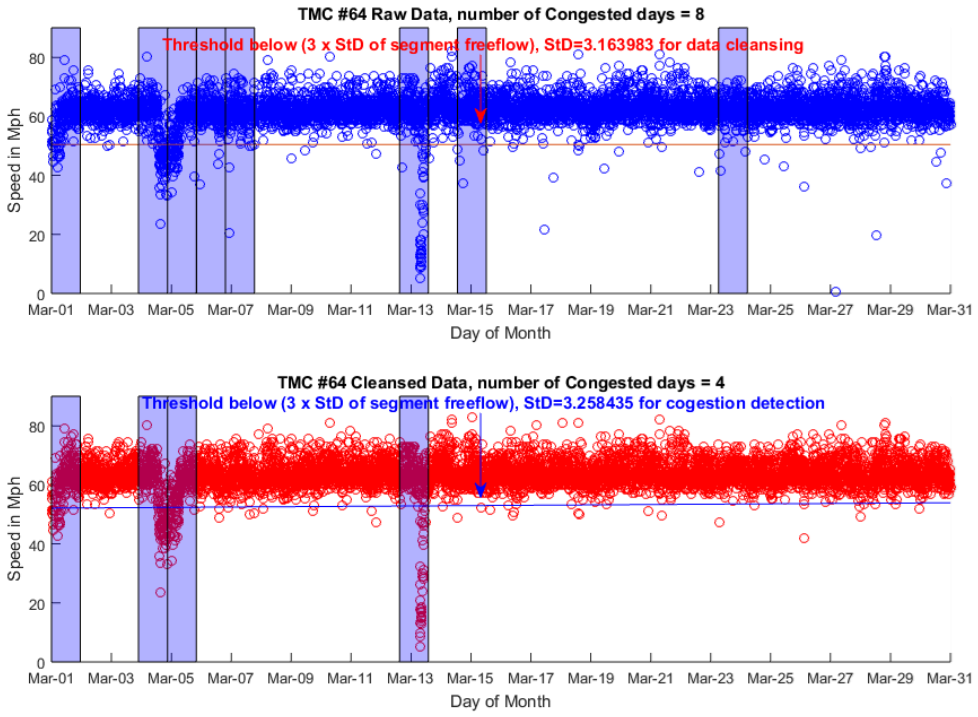


Figure 115 - Epoch speed plot for segment 64 during the month of March.

Free-flow travel and snow travel cases were characterized by mean and standard deviation modeled according to a gaussian distribution model. See Table 27. Incident event modeling was performed using a non-parametric Kernal density estimater to generate probability density function (PDF). The formula for the model is given by a smoothing spline, 3rd degree piecewise polynomial. The resulting formula has 70 parameter coeffificents, shown in Appendix D. Figures 114, 115, and 116 demonstrate fitted models per case. Although fitting of snow showed less goodness-of-fit than normalfree-flow traffic with regard to normal distribution, Bayesian inference results exhibited robustness in decison making and correctly identifying cases, as evident in

subsequent results. Results are indicative of the suitability of Bayesian inference for solving problems when accurate, closed form models are not possible.

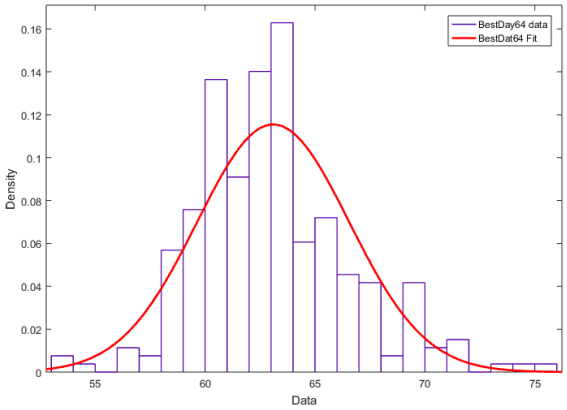


Figure 116 – Free Flow model fit.

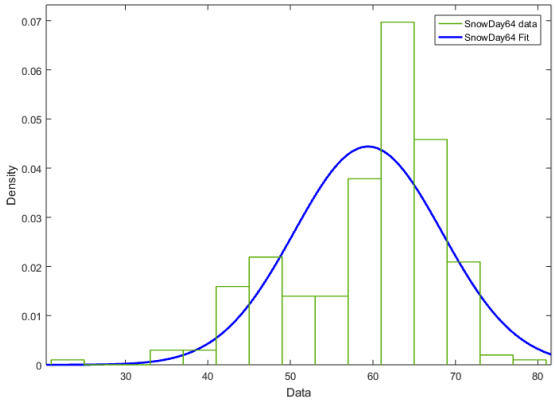


Figure 117 – Snow (weather) model fit.

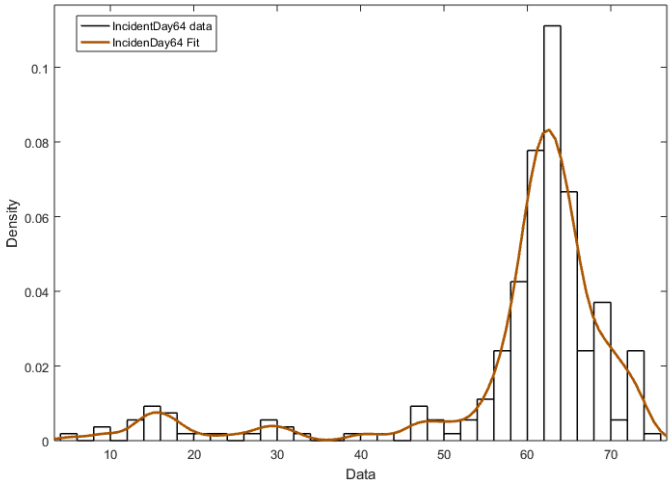


Figure 118 – Incedent model fit

Table 27 – Free flow- snow distribution model parameters

	Free Flow	Weather
Mean	67.9814	59.412
Standard Deviation	4.88	8.9821

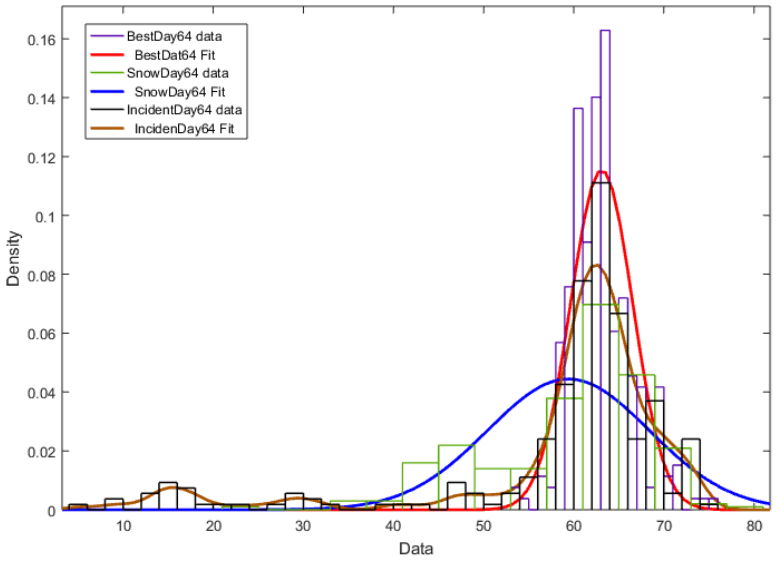


Figure 119 - Distribution fitting for 3 distinct events overlaid.

Figure 117 illustrates probability for the three categories overlaid the model. High overlap occurs when value of free flow mean speed is near the speed limit.

Figure 120 depicts a probablity plot showing various probability values for each model relative to various speed measurements on the highway. Three distinct regions are visible. Lower speeds of 0 to 30 mph result in higher probability of incident occurrance. As speeds increase to between 30 and 60 mph, the snow model tends to dominate with higher probability values over-all. For travel near the speed-limit, the free flow model dominates in probability values, in spite of overlap among distribution models in this region.

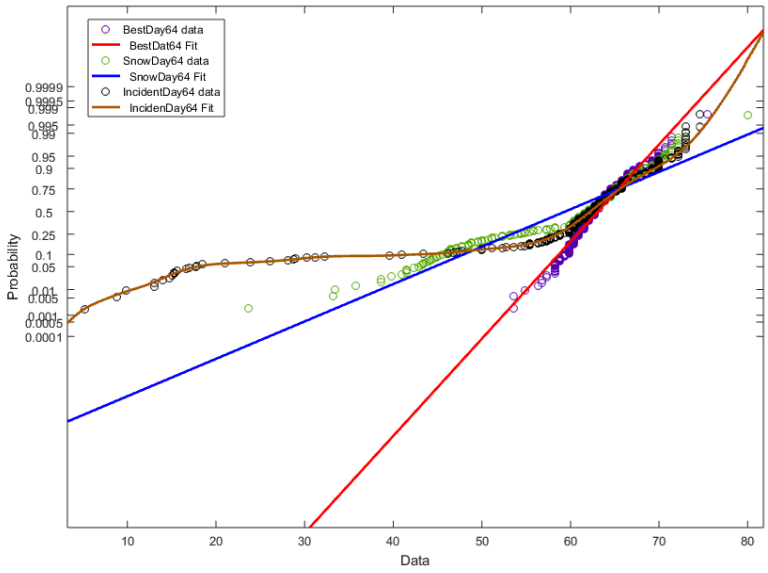


Figure 120 - Probability plot for 3 distinct distribution models

6.3.4.1. Bayesian updating

Implementing Bayes theorem in a time series input requires updating the prior. Each of the aforementioned cases occurs over individual time intervals. Snow, for instance, accumulates with time, and the effect on roadways becomes apparent after several hours of continuous snowfall. By contrast, effects of a road incident occur almost instantaneously. Thus, one can intuitively suggest that updating prior probabilities is related to the duration of the event and the time required for its effects to manifest. As a result, posterior probability was averaged over the course of an hour and a half for incident data and over five hours for the snow event. Free flow update time was chosen to match the shortest length of time for all cases. Values were chosen based on the duration each event modeled for one day. Prior update time remains an optimization research problem that requires a larger sample size to be studied. Furthermore, there is a tradeoff between the system's ability to instantly detect an event (i.e., response time) and the stability and accuracy of the system. Decreasing update time results in near

instantaneous updating of the prior, which causes fast inference decisions. False detection is expected to occur when small values are used, particularly in cases where speed measurements caused by outliers and anomalies were present in the data, or, in cases when there is a high variance between consecutive data samples. On the other hand, increasing prior update time could result in the system's inability to detect extremely short incident occurrences for durations of 15 or fewer minutes. After taking into account the aforementioned details, the Bayesian inference engine was coded using Matlab. Inputs of actual speed measurements obtained from the NPRMDS dataset, which simulated real-time measurements, were fed to the system. System output was a prediction of the type of condition (event) causing the input speed measurements given. Results per case are offered below.

6.3.4.2. Incident

Figure 121 demonstrates a traffic accident at approximately 3 p.m. on I-35 southbound over segment 15. Figure 122 shows a snap shot of the Bayesian inference engine GUI final output for a day of monitoring. The top subplot illustrates speed records arriving in real time. The bottom subplot illustrates the probability of the Bayesian engine pertaining to each of the three defined states; free-flow, incident and weather (in this case, snow). The right subplot illustrates system output. For this implementation, a threshold of 40% confidence was required for decision-making. The threshold is flexible and can be modified, as necessary. Figure 122 indicates incident detection between 4:23 and 5:20 PM. Figure 123 and Figure 124 illustrate an incident on segment 78 on March 13, 2015. System output indicates the incident was detected between 2:38 and 5:18 PM. Detection and incident time reported by news agencies was highly correlated, primarily because the

effect of an incident is profound and nearly instantaneous on traffic flow. As a result, Bayesian inference will allocate increased credibility and confidence to its probability.

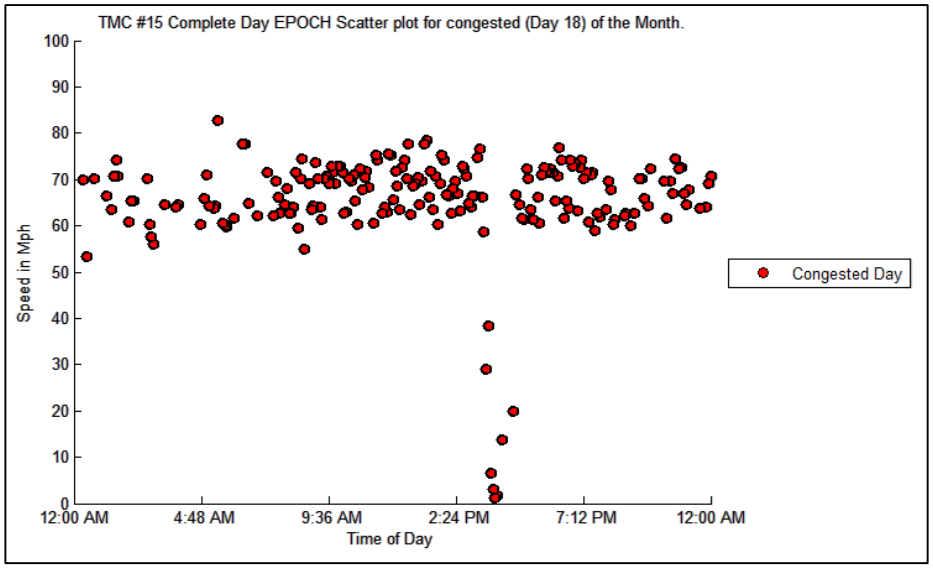


Figure 121 - Scatter plot of incident data which occurred at segment 15 during the 18th of January 2015

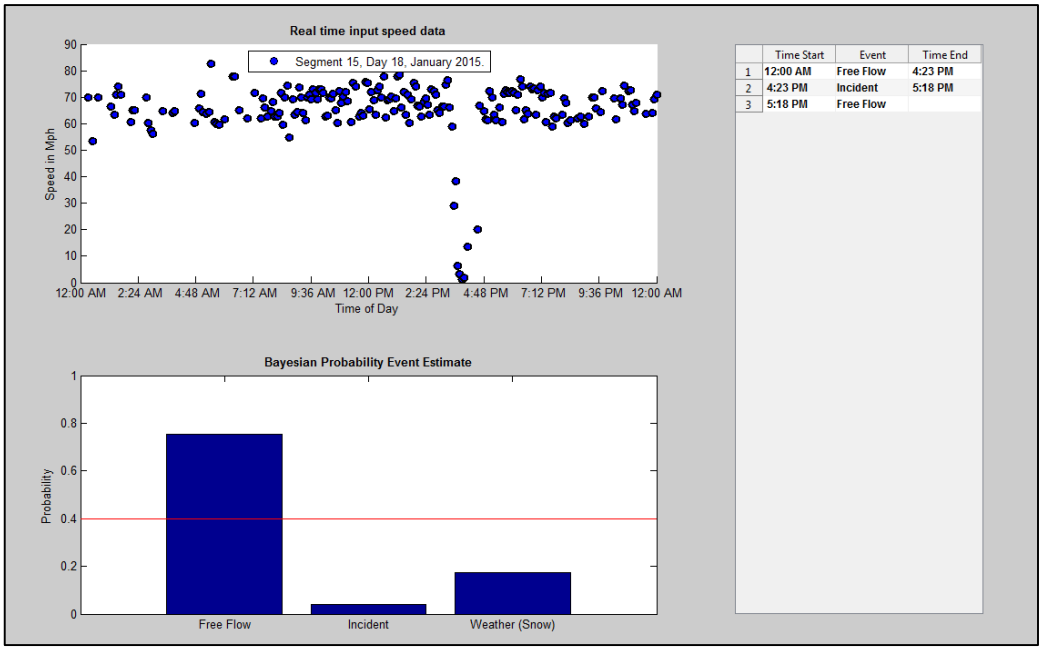


Figure 122 – System output of incident which occurred at segment 15 during the 18th of January 2015.

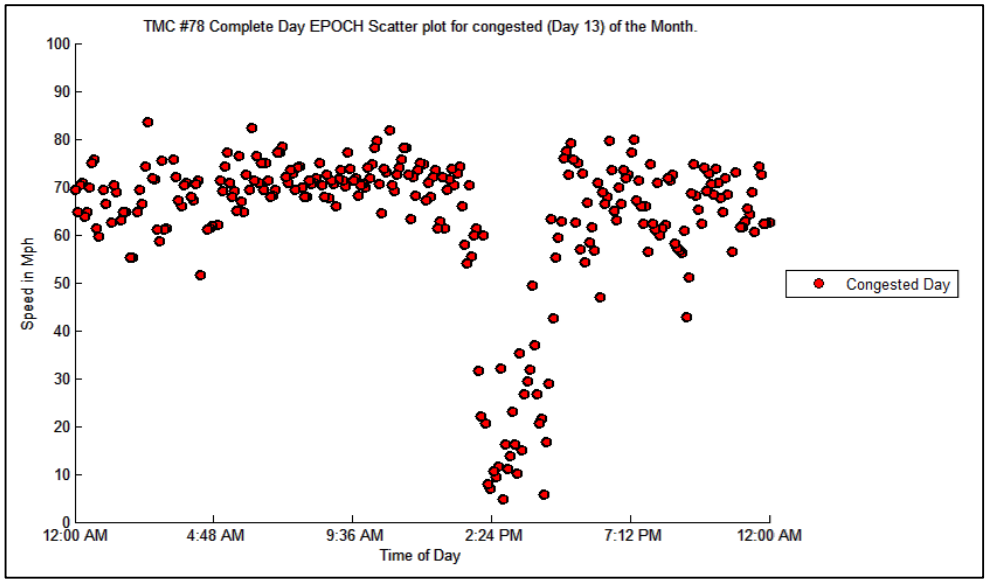


Figure 123 – Scatter plot of incident data which occurred at segment 78 on March 13 2015.

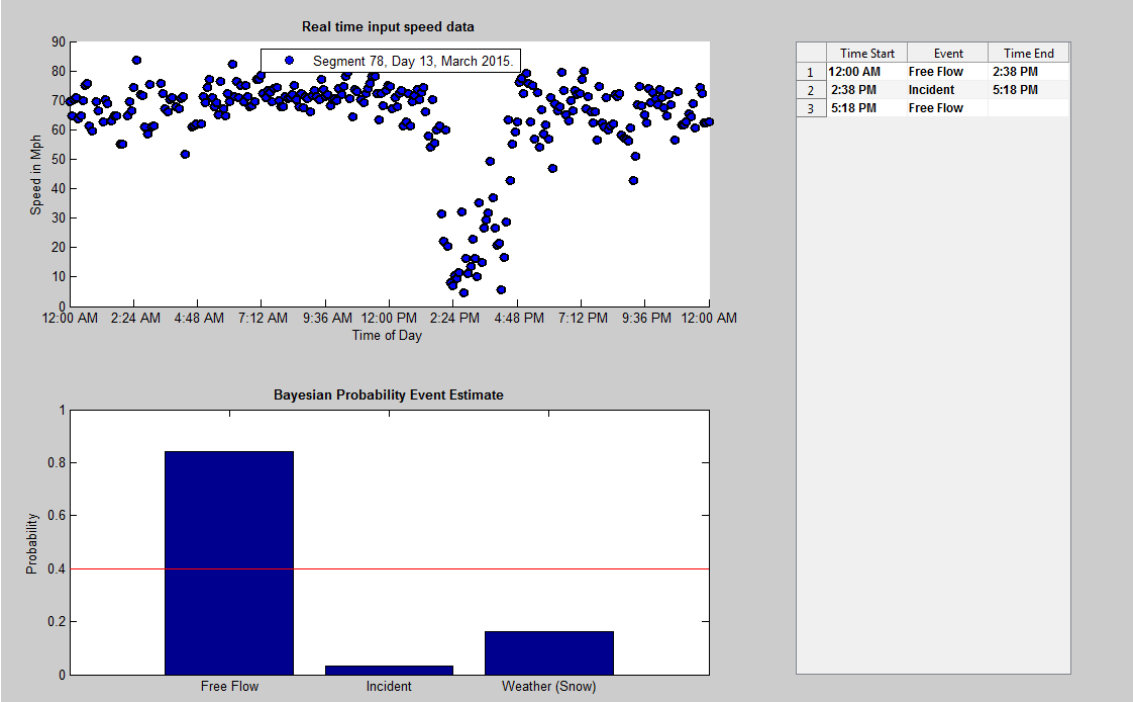


Figure 124 – System output of incident which occurred at segment 78 on March 13 2015.

6.3.4.3. Snow (Weather)

Figure 125 illustrates a scatter plot segment 80 on a snowy March 4, 2015. Notably, rain also occurred between 5 and 9 AM that morning, followed by light snow and then heavier snow for nearly the entire second half of the day. Figure 126 illustrates that the system was capable of detecting the snow event, with no false detection during rainfall, primarily because rain results in a larger spread of speeds quite different from snow. Detection time of snow was much later than the instantaneous detection of incidents. Reported output was approximately 7:44 PM, whereas snow was reported to have started several hours earlier, and accumulation increased gradually over many hours. Credibility allocation to its probability is similarly affected by the duration resulting in a delayed response. Although this can be affected by changing the prior update time, nevertheless, accuracy will be affected, as will an increase in the rate of false detection. When incidents occur, real time response is critical. However, this might not be needed for weather events, such as rain and snow. Accordingly, a less stringent response time can be tolerated for weather as a tradeoff between improved accuracy and error due to an increase of false detection. Figure 127 and Figure 128 illustrate a second weather event that occurred March 4, 2015 on segment 90. Delayed identification between time reported and time predicted in the output of the Bayesian inference engine was two hours.

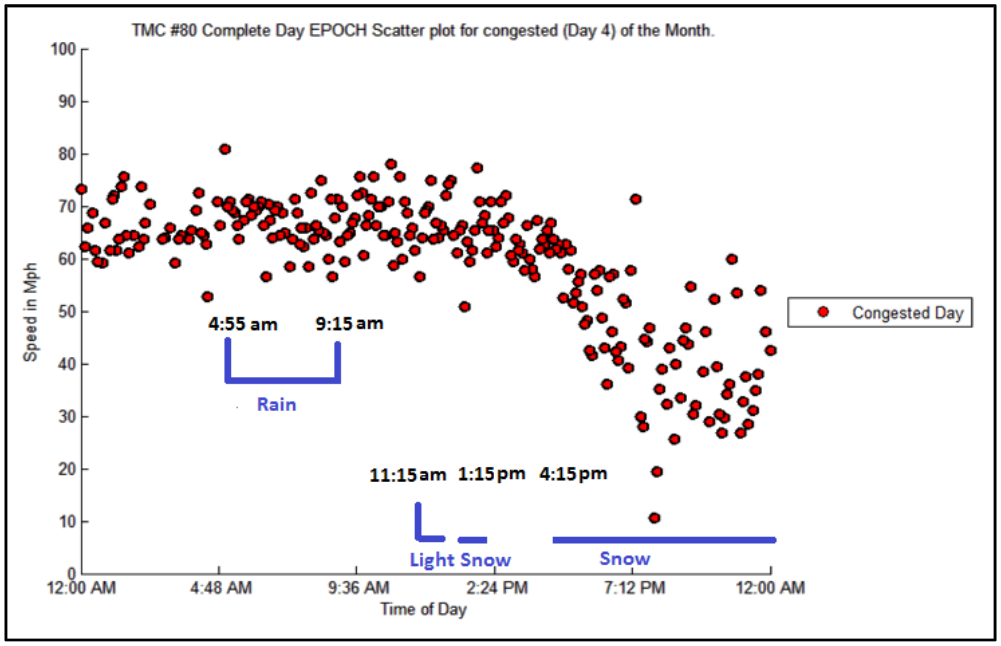


Figure 125 - Scatter plot of snow data which occurred at segment 80 on March 4 2015

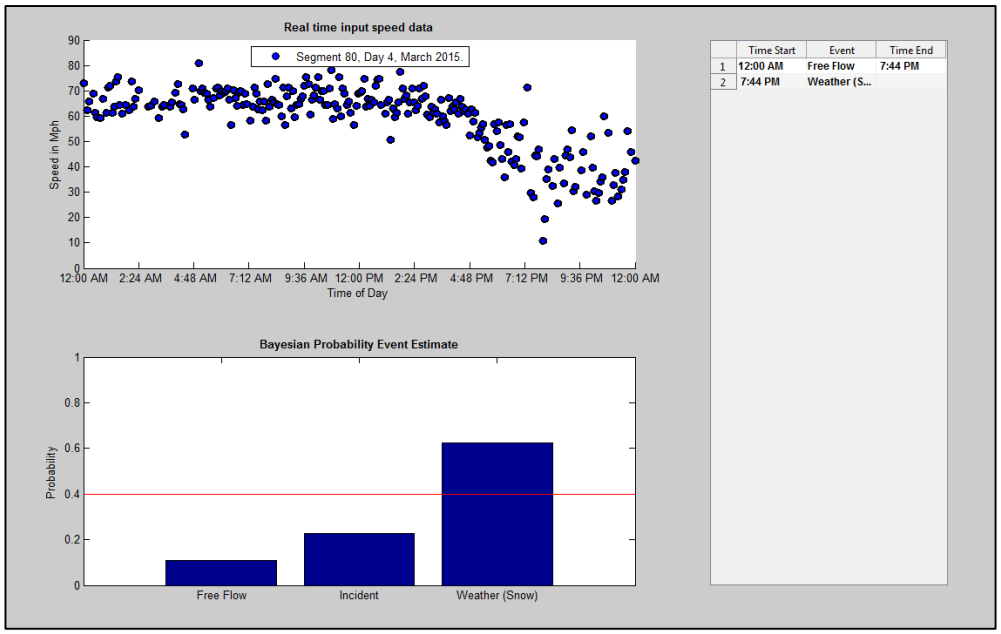


Figure 126 – Snow (weather) congestion on segment 80 during the 4th of March 2015.

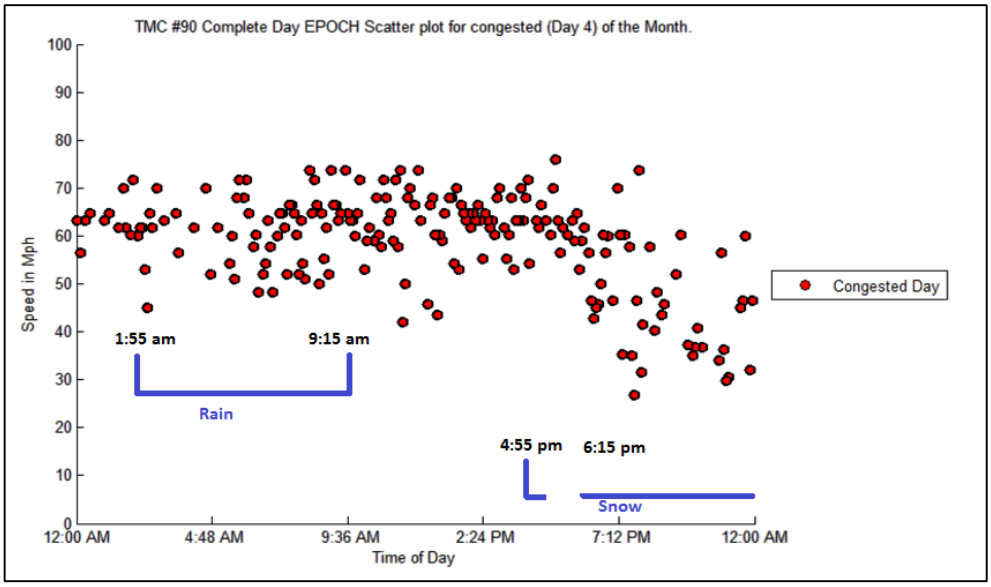


Figure 127 - Scatter plot of snow data which occurred at segment 90 on March 4 2015

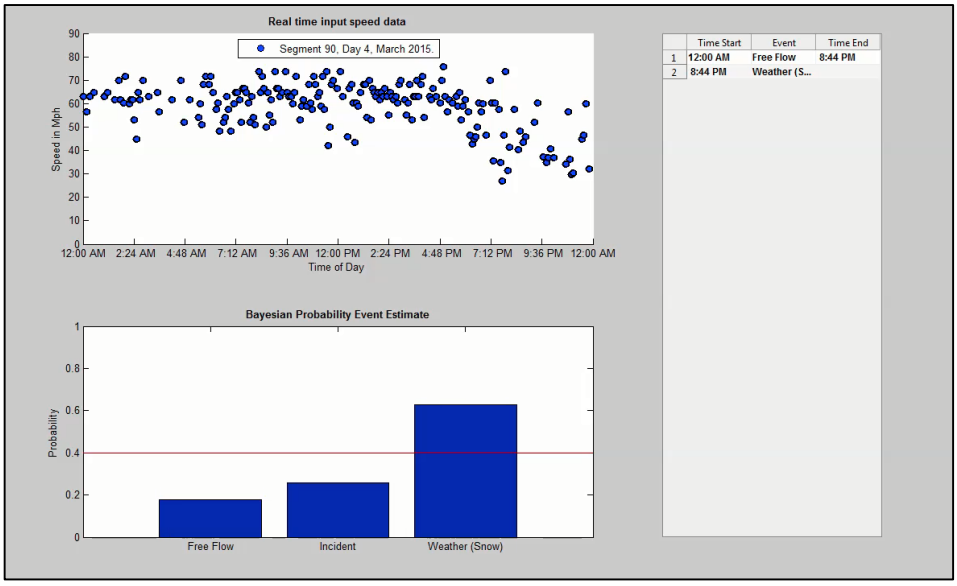


Figure 128 - Snow (weather) congestion on segment 90 during the 4th of March 2015.

6.3.4.4. Free Flow

Figure 129 depicts a case of free flow where the system was able to identify traffic conditions with a high degree of accuracy, primarily because probability for the range of

speeds were dominate in the Gaussian model previously shown when compared to that which fit both weather and incident models.

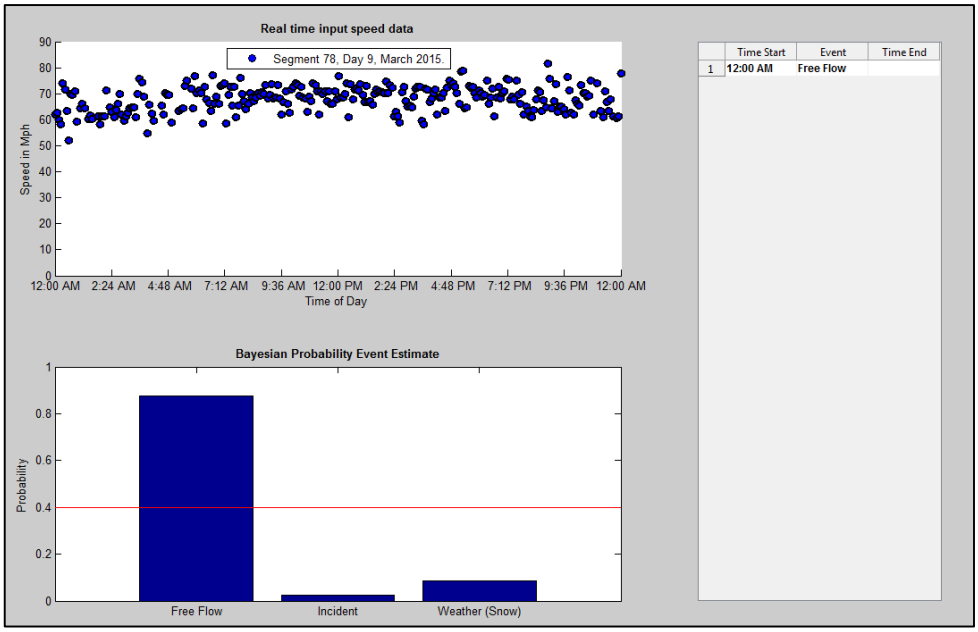


Figure 129 - Free flow occurring all day on March 9, 2015.

Overall, 12 cases—four incident, four weather (snow), and four free flow—were tested and identified. Results inclusively demonstrated accurate identification when using the proposed method. The proposed approach shows promising results and could be integrated with real-time incident detection technologies. Validating the accuracy of the proposed approach requires a larger sample size that researchers will obtain in future work. Furthermore, models must be extended to account for additional sources of congestion (e.g., weather events such as rain, fog, or hail and non-weather, non–recurrent causes such as work zones). Finally, optimizing prior update time for each case remains a research problem that affects identification time, which is critical for reducing identification time response without decreasing accuracy and increasing false alarm rate.

Chapter 7: Conclusion and Future Work.

Future ITS systems are expected to handle and resolve the arduous challenges of maintaining and improving roadway performance, facing today's transportation engineers and agencies alike. This is achieved through systems incorporating intelligence, coupled with the ability to ingest highly heterogeneous data in real-time, to perform various types of inferences i.e., (analysis, diagnosis, exploration and predictions) that allow insight and knowledge to be extracted and optimal solutions to be employed.

This thesis presented research detailing the use of one of the nation's largest datasets of roadway travel times; the NPMRDS. A comprehensive study of dataset characteristics, including influencing variables that affect data measurements have been presented. Research affirms that understanding domain specific characteristics is vital for filtering data outliers and anomalies, and is key for accurate statistical analysis to be performed. Moreover, a process for identifying anomalies using Benford's law was developed and models validating speed accuracy, computing optimum system time granularity, and computing minimum segment length for a desired CI were formulated. Models serve as tools for validating, designing and understanding the characteristics of travel time measurement systems. Furthermore, recommendations for improving accuracy and alleviating data anomalies in the NPMRDS were reported. Research affirms careful consideration of system capture time granularity and segment length has to be taken into account as the interaction between the two, coupled with the speed of vehicles on the road, could result in anomalous data being generated. Statistical analysis confirms that while summary statistics of data averaged over the course of a month is not highly effected by outliers, granular time periods are. Mean and variance statistics exhibited a

difference of around 3-5 mph when summarization was done over a period of one day. For congestion detection, removal of outliers contributed to the reduction of false alarm rate errors for congestion of segments and congested days for both variance and thresholding detection methods alike. More importantly, the effect of outliers was found to be severe on travel time reliability measures such as travel time index, buffer time index and planning time index. Thus, careful consideration for outlier removal has to be taken when computing these measurements. Finally, a novel approach for identifying non-recurrent congestion sources using Bayesian inference of speed data was developed and introduced. Results inclusively demonstrated accurate identification when using the proposed method. The proposed approach shows promising results and could be integrated with real-time incident detection technologies. Future work includes validating the accuracy of the proposed approach on a larger sample size, and extending the work to include models that account for additional sources of congestion (e.g., weather events such as rain, fog, or hail and non-weather, non–recurrent causes such as work zones). Optimizing prior update time for each case remains a research problem critical for reducing identification time response without decreasing accuracy and increasing false alarm rate.

References

[1] FHWA Office of Operations, "Travel Time Reliability: Making It There On Time, All The Time," Prepared by Texas Transporation Institute with Cambridge Systems, Inc., 1 January 2006. [Online]. Available: http://ops.fhwa.dot.gov/publications/tt_reliability/ . [Accessed 10 September 2015].

[2] L. a. X. C. Elefteriadou, "Review of Definitions of Travel Time Reliability," in *86th Annual Meeting of the Transportation Research Board*, Washington, D.C., 2007.

[3] C. Ebeling, Introduction to Reliability and Maintainability Engineering, McGraw- Hill Companies Inc., 1997.

[4] G. F. List, B. Williams, N. Rouphail, R. Hranac, T. Barkley, E. Mai, A. Ciccarelli, L. Rodegerdts, A. F. Karr, X. Zhou, J. Wojtowicz, J. Schofer, and A. Khattak, "Guide to Establishing Monitoring Programs for Travel Time Reliability: SHRP 2 Report S2-LO2-RR-2," FHWA, Washington, D.C., 2014.

[5] FHWA Office of Operations, "National performance management research data set (NPMRDS) information," FHWA, 23 June 2015. [Online]. Available: http://www.ops.fhwa.dot.gov/perf_measurement/. [Accessed 9 September 2015].

[6] FHWA Office of Operations, "2013 urban congestion trends," FHWA, 23 April 2015. [Online]. Available: http://www.ops.fhwa.dot.gov/publications/fhwahop15005/index.htm. [Accessed 9 September 2015].

[7] FHWA Office of Operations and Resource Center, "Introduction to the national performance management research data set (NPMRDS)," HERE and the Volpe Center, 1 August 2013. [Online]. Available: http://connectdot.connectsolutions.com/p42seglc752/. [Accessed 1 September 2015].

[8] FHWA Office of Operations, "National performance management research data set (NPMRDS) information, technical frequently asked questions," FHWA, 28 January 2014. [Online]. Available: http://www.ops.fhwa.dot.gov/freight/freight_analysis/perform_meas/vpds/npmrdsfaqs.htm. [Accessed 9 September 2015].

[9] Rajat Rajbhandari, "Exploring the applicability of commercially available speed and travel time data around border crossings. Final Report 186051- 00001," Center for International Intelligent Transportation Research, Texas Transportation Institute. The Texas A&M University, Texas, 2012.

[10] FHWA Office of Operations and Resource Center, "Second Quarterly NPMRDS Webinar," HERE and the Volpe Center, 1 February 2014. [Online]. Available: https://connectdot.connectsolutions.com/p36vxid1rr5/. [Accessed 1 August 2015].

[11] Rafferty, P., and C. Hankley, "National Performance Management Research Data Set (NPMRDS)," Wisconsin Traffic Operations and Safety Laboratory, 12 February 2014. [Online]. Available: http://www.topslab.wisc.edu/its/topms/tops_npmrds_20140212.pdf. [Accessed 1 August 2015].

[12] Rafferty, P., and C. Hankley, "NPMRDS Travel Time Reliability - Travel time reliability in the Mid America Freight Coalition Regions," TOPS Lab, 1 January 2014. [Online]. Available: http://www.arcgis.com/home/item.html?id=7089b0b5870e4505a2f9f175c157563c. [Accessed 1 August 2015].

[13] Liao, C, "Using Truck GPS Data for Freight Performance Analysis in the Twin Cities Metro Area," Research Services and Library, Office of Transportation System Management. Minnesota Department of Transportation, Minnesota , 2014.

[14] Pierce, D., and D. Murray., "Cost of Congestion to the Trucking Industry.," American Transportation Research Institute, 2014.

[15] HERE and the Volpe Center, "Third Quarterly NPMRDS Webinar," FHWA Office of Operations and Resource Center, 1 May 2014. [Online]. Available: https://connectdot.connectsolutions.com/p1ubotswuel/. [Accessed 1 August 2015].

[16] Kaushik K., E. Sharifi, S. E. Young, and B. Baghaei, "Comparison of National Performance Management Research Data Set (NPMRDS) with Bluetooth Traffic Monitoring (BTM) Data and I-95 Corridor Coalition Vehicle Probe Project (VPP) Data," in *Presented at the 31st ITS World Congress*, Detroit, September 2014.

[17] Sepideh Eshragh, Kaveh Farokhi Sadabadi, Kartik Kaushik and Reuben M. Juster, "Truck and Auto Performance Measurement Using Probe-Based Speed Data: Case Study I-95 Corridor in Maryland," in *94th Annual Meeting of Transportation Research Board*, Washington D.C., January 11-15, 2015.

[18] Kaveh Farokhi Sadabad, Thomas H. Jacobs, Sevgi Erdoga, Fredrick W. Ducca and Lei Zhang, "Value of Travel Time Reliability in Transportation Decision Making: Proof of Concept—Maryland," TRB Publications, February 2015.

[19] Kartik Kaushik, Elham Sharifi, and Stanley Ernest Young, "Computing Performance Measures with National Performance Management Research Data Set," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2529, p. 10–26, 2015.

[20] Filmon Habtemichael, Rajesh Paleti and Mecit Cetin, "Performance Measures for Freight & General Traffic: Investigating Similarities and Differences Using Alternate Data Sources," Virginia Center for Transportation Innovation and Research, Virginia, 2015.

[21] John Wisdom, "Using Travel Time Data for Analyzing Congestion," NCG Conference, 26 February 2015. [Online].

Available:http://ncgisconference.com/presentations/pdf/306B_1-3_Wisdom.pdf. [Accessed August 2015].

[22] CDM Smith, "Travel Time Based Oklahoma Congestion Analysis: Pilot Study," Prepared for the Oklahoma Department of Transportation, 2014. [Online]. Available: http://www.okladot.state.ok.us/p-r-div/lrp_2015_2040/2040_LRTP_TM_Travel_Time.pdf. [Accessed 11 January 2015].

[23] Peter Rafferty and Chip Hankley, "Crafting measures from the national performance management research data set," in *22nd ITS World Congress*, Bordeaux, 2015.

[24] Mark E. Hallenbeck, Ed McCormack and Saravanya Sankarakumaraswamy, "Developing A System for Computing and Reporting MAP-21 and Other Freight Performance Measures," The State of Washington, Department of Transportation, Washington, 2015.

[25] Sabya Mishra, Mihalis Golias, Maxim Dulebnets, and Mania Flaskou, "A Guidebook for Freight Transportation Planning Using Truck GPS Data," Wisconsin Department of Transportation, Madison, 2016.

[26] "Travel Time Reliability Reference Manual," Upper Midwest Reliability Resource Center, [Online]. Available:http://en.wikibooks.org/wiki/Travel_Time_Reliability_Reference_Manual. [Accessed 1 August 2015].

[27] "http://www.glrtoc.org/operations/performance/tmc-map/," [Online]. [Accessed August 2015].

[28] "http://www.glrtoc.org/operations/performance/scanner/," [Online]. [Accessed August 2015].

[29] "https://company.here.com/enterprise/location-content/here-traffic/," [Online]. [Accessed August 2015].

[30] " http://www.iteris.com/products/services," [Online]. [Accessed August 2015].

[31] " http://www.ops.fhwa.dot.gov/perf_measurement/ucr/," [Online]. [Accessed August 2015].

[32] " http://www.fhwa.dot.gov/planning/national_highway_system/nhs_maps/," [Online]. [Accessed August 2015].

[33] "https://hadoop.apache.org/," [Online]. [Accessed August 2015].

[34] " https://hive.apache.org/," [Online]. [Accessed September 2015].

[35] Frank Benford, "The Law of Anomalous Numbers," *Proceedings of the American Philosophical Society,* vol. 78, no. 4, pp. 551-572, 31 Mar 1938.

[36] Hill, Theodore P., "A Statistical Derivation of the Significant-Digit Law.," *Statist. Sci.*, vol. 10, no. 5, pp. 354-363, 1995.

[37] Z. Jasak, L. Banjanovic'-Mchmcdovic, "Detecting Anomalies by Benford's Law," in *Signal Processing and Information Technology, 2008. ISSPIT 2008. IEEE International Symposium on*, 16-19 Dec. 2008..

[38] Cindy Durtschi, WIlliam Hillison and Carl Pacini, "The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data," *Journal of Forensic Accounting,* vol. 5, pp. 17-34, 2004.

[39] J. W. C. van Lint and H. J. van Zuylen, "Monitoring and Predicting Freeway Travel Time Reliability Using Width and Skew of Day-to-Day Travel Time Distribution," *Transportation Research Record: Journal of the Transportation Research Board,,* vol. 1917, pp. 54-63, 2005.

[40] R. Fisher, Statistical Methods for Research Workers, 13th Ed. Hafner, 1958.

[41] John C. Falcocchio and Herbert S. Levinson, Road Traffic Congestion: A concise Guide, vol. 7, New York: Springer, 2015.

[42] Paula J, Hammond, "The 2011 Congestion Report," Washington State Department of Transporation, Washington, 2011.

[43] Naim Bitar, Hazem Refai, "A Probabilistic Approach to Improve the Accuracy of Axle-based Automatic Vehicle Classifiers," *in Intelligent Transportation Systems, IEEE Transactions on*, 2016, To be Published..

[44] H Refai, N Bitar, J Schettler, O Al Kalaa, "The Study of Vehicle Classification Equipment with Solutions to Improve Accuracy in Oklahoma," (No. FHWA-OK-14-17), 2014.

[45] Carol T. Rawson, P.E, "Procedures for Establishing Speed Zones," Texas Department of Transportation.

[46] National Research Council, "Highway Capacity Manual - HCM2000," Transportation Research Board, 2000.

[47] National Cooperative Highway Research Program, "Cost-Effective Performance Measures for Travel Time Delay, Variation and Reliability," Transporation Research Board, 2008.

[48] Eric T. Donnell, Ph.D., P.E; Scott C. Hines, Kevin M. Mahoney, D. Eng., P.E., Richard J. Porter, Ph.D., Hugh McGee, Ph.D., P.E., "Speed Concepts: Information Guide," U.S Department of Transporation, Federal Highway Administration , FHWA-SA-10-001, 2009.

[49] David Harris Solomon, "Accidents on main rural highways related to speed, driver, and vehicle," United States. Bureau of Public Roads, 1964.

[50] Urban Congestion Report, "The Urban Congestion Report (UCR): Documentation and Definition," Office of Operations, FHWA, 22 September 2015. [Online]. Available: http://www.ops.fhwa.dot.gov/perf_measurement/ucr/documentation.htm. [Accessed 16 3 2016].

[51] Tim Lomax, David Schrank, Shawn Turner and Richard Margiotta, "Selecting Travel Reliability Measures," Texas Transportation Institute, Cambridge Systematics, Inc., May 2003.

[52] Rokach, Lior, and Oded Maimon, "Clustering methods.," Data mining and knowledge discovery handbook, Springer US, 2005. 321-352.

[53] Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, 2012.

[54] Kaufman, L. and P. J. Rousseeuw., An Introduction to Finding Groups in Data, New York: John Wiley & Sons., 1990.

[55] Ralf Herbrich, Thore Graepel and Colin Campbell, "Bayes Point Machines," *Journal of Machine Learning Research,* vol. 1, p. 245–279, 2001.

[56] T. Cover and P. Hart, "Nearest neighbor pattern classification,," *in IEEE Transactions on Information Theory,* vol. 13, no. 1, pp. 21-27, January 1967.

[57] Haim Dahan, Shahar Cohen, Lior Rokach and Oded Maimon, Proactive Data Mining with Decision Trees, New York: Springer, 2014, p21-22.

[58] Christopher J.C Burges, "A Tutorial on Support Vector Machines for Pattern," *Data Mining and Knowledge Discovery,* vol. 2, pp. 121-167, 1998.

[59] Jaynes, E.T, Bayesian Methods: General Background, n Maximum-Entropy and Bayesian Methods in Applied Statistics, by J. H. Justice (ed.). Cambridge: Cambridge Univ. Press, 1986.

[60] https://xkcd.com/1132/, [Online]. [Accessed 25 February 2015].

[61] D. Heckerman, "A Tutorial on Learning With Bayesian Networks," Microsoft Research, Technical Report MSR-TR-95-06, March 1995.

[62] C. van Hinsbergen and J. van Lint, "Bayesian combination of travel time prediction models," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 2064, pp. 73-80, 2008.

[63] Xiang Fei, Chung-Cheng Lu, Ke Liu, "A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction," *Transportation Research Part C: Emerging Technologies,* vol. 19, no. 6, pp. 1306-1318, December 2011.

[64] Juan de Oña1, Randa Oqab Mujalli1, Francisco J. Calvo, "Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks," *Accident Analysis & Prevention,* vol. 43, no. 1, pp. 402-411, January 2011.

[65] Rongjie Yu, Mohamed Abdel-Aty, " Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes," *Accident Analysis & Prevention,* vol. 58, pp. 97-105, September 2013.

[66] Kun Zhang, Michael A.P. Taylor, "Effective arterial road incident detection: A Bayesian network based algorithm," *Transportation Research Part C: Emerging Technologies,* vol. 14, no. 6, pp. 403-417, December 2006.

[67] Mohamed M. Ahmed, Mohamed Abdel-Aty, and Rongjie Yu, "Bayesian Updating Approach for Real-Time Safety Evaluation with Automatic Vehicle Identification Data," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 2280, pp. 60-67, 2012.

[68] Moinul Hossain, Yasunori Muromachi, "A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways," *Accident Analysis & Prevention,* vol. 45, pp. 373-381, March 2012.

[69] Moinul Hossain, Yasunori Muromachi, "A real-time crash prediction model for the ramp vicinities of urban expressways," *IATSS Research,* vol. 37, no. 1, pp. 68-79, July 2013.

[70] Hesham Rakha, Mohamadreza Farzaneh, Mazen Arafeh, and Emily Sterzin, "Inclement Weather Impacts on Freeway Traffic Stream Behavior," *Journal of the Transportation Research Board,* vol. 2071, pp. 8-18, 29 January 2008 .

[71] Mario Cools, Elke Moons, and Geert Wets, "Assessing the Impact of Weather on Traffic Intensity," *Weather, Climate, and Society. American Meteorological Society,* vol. 2, pp. 60-68, 2010.

[72] Sandeep Datla, Prasanta Sahu, Hyuk-Jae Roh, Satish Sharma, "A Comprehensive Analysis of the Association of Highway Traffic with Winter Weather Conditions," *Procedia - Social and Behavioral Sciences,* vol. 104, pp. 497-506, 2 December 2013.

[73] Sandeep Datla, Satish Sharma, "Impact of cold and snow on temporal and spatial variations of highway traffic volumes," *Journal of Transport Geography,* vol. 16, no. 5, pp. 358-372, September 2008.

[74] Roh, Hyuk-Jae and Sharma, Satish and Sahu, Prasanta K. and Datla, Sandeep, "Analysis and modeling of highway truck traffic volume variations during severe winter weather conditions in Canada," *Journal of Modern Transportation,* vol. 23, no. 3, pp. 228-239, 2015.

[75] Athanasios Theofilatos, George Yannis, "A review of the effect of traffic and weather characteristics on road safety," *Accident Analysis & Prevention,* vol. 72, pp. 244-256, November 2014.

[76] Antonio S. Cofino, Rafael Cano, Carmen Sordo and Jose M. Gutierrez, "Bayesian Networks for Probabilistic Weather Prediction," in *in ECAI 2002. Proceedings of the 15th European Conference on Artificial Intelligence*, 2002.

[77] Michael J. Erickson, Brian A. Colle, Joseph J. Charney, "Impact of Bias-Correction Type and Conditional Training on Bayesian Model Averaging over the Northeast United States.," *Weather and Forecasting,* pp. 1449-1469, December 2012.

[78] R. Marty, V. Fortin, H. Kuswanto, A.-C. Favre, E. Parent, "Combining the Bayesian processor of output with Bayesian model averaging for reliable ensemble forecasting," *Journal of the Royal Statistical Society: Series C (Applied Statistics)* , vol. 64, no. 1, pp. 75-92, 2014.

[79] Joint probabilistic forecasting of wind speed and temperature using Bayesian model averaging, *Environmetrics,* vol. 26, no. 2, pp. 120-132, March 2015.

[80] Xiang Fei, Chung-Cheng Lu, Ke Liu, "A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction," *Transportation Research Part C: Emerging Technologies,* vol. 19, no. 6, pp. 1306-1318, December 2011.

Appendix A- Segment Free Flow Statistics

Mean, Variance and Standard Deviation Free Flow Statistics for I-35 southbound segments.

	Cleansed Dataset		Raw Dataset	
Segment number	Mean Free-Flow Speed (Mph)	Mean Free-Flow TT (Hour)	Mean Free-Flow Speed (Mph)	Mean Free-Flow TT (Hour)
1	61.66524054	0.001649227	61.17955485	0.00166232
2	67.08731889	0.061870262	65.33022804	0.063534295
3	66.06997434	0.013342369	64.17450277	0.013736452
4	66.9601723	0.122051359	64.32567135	0.127050054
5	67.49851698	0.058101128	64.4582114	0.060841589
6	67.55270124	0.060449544	64.62530094	0.063187791
7	66.82871201	0.044904352	64.11301232	0.046806411
8	68.154124	0.11782075	64.9426361	0.123647121
9	67.89004955	0.14025207	64.4348242	0.147772887
10	65.85834018	0.007057117	65.53922647	0.007091478
11	67.67066542	0.103639442	64.958814	0.107966103
12	65.90360087	0.016144793	64.67223829	0.016452191
13	67.44063037	0.073530896	64.39276274	0.077011294
14	67.910604	0.092766809	65.97099013	0.095494247
15	68.13926467	0.057985657	66.0656925	0.059805625
16	68.42562839	0.191117573	65.13803044	0.200763516
17	69.0061864	0.064235255	65.60588673	0.067564516
18	65.97643487	0.016374938	65.76510832	0.016427556
19	68.06153442	0.07891741	65.682188	0.081776204
20	68.13643832	0.043599285	66.36694617	0.04476174
21	67.64449233	0.020750987	66.11406622	0.021231337
22	67.15088743	0.010612369	66.44087892	0.010725776
23	65.79103068	0.01447173	64.61205067	0.014735796
24	67.29946605	0.014680206	66.00415522	0.0149683
25	66.12270967	0.010302814	65.49048824	0.010402274
26	64.94883504	0.005397633	64.35887496	0.005447112
27	66.15905419	0.005592885	65.55209539	0.005644671
28	62.74752951	0.004446709	62.99768748	0.004429051
29	61.37458064	0.006754588	58.29798668	0.007111052
30	64.72207619	0.016730149	62.31939187	0.01737517
31	65.07052196	0.017121578	61.55176473	0.018100375
32	63.47771023	0.01790471	63.88912054	0.017789414
33	61.45625686	0.011404046	58.69982282	0.011939559
34	56.80647264	0.004432593	56.29981718	0.004472483
35	61.81037912	0.007400861	61.64593979	0.007420602
36	61.09200709	0.012729488	61.03556378	0.01274126
37	61.12660238	0.024466925	59.91950382	0.02495982

	Cleansed Dataset		Raw Dataset	
Segment number	Mean Free-Flow Speed (Mph)	Mean Free-Flow TT (Hour)	Mean Free-Flow Speed (Mph)	Mean Free-Flow TT (Hour)
38	60.25622985	0.007701776	59.7181227	0.007771175
39	62.59966293	0.016836832	61.78030735	0.017060129
40	61.06040563	0.018758146	59.06146461	0.019393017
41	55.73322581	0.000765432	55.73322581	0.000765432
42	59.0565178	0.006087389	57.90649905	0.006208284
43	59.50885485	0.006816465	58.31667499	0.006955815
44	45.40878513	0.008546364	43.56719285	0.00890762
45	59.67275441	0.016289176	58.73563591	0.016549067
46	60.68311543	0.007879457	61.84720832	0.007731149
47	61.27757297	0.008370599	61.83553852	0.008295068
48	60.48457311	0.006857616	60.20945043	0.006888952
49	60.35539545	0.001900079	60.21293582	0.001904574
50	62.67822894	0.016087723	61.64102904	0.016358423
51	60.33337211	0.007845244	59.98279906	0.007891096
52	62.3709272	0.007737419	61.97727192	0.007786564
53	59.7644505	0.009690376	58.14718092	0.009959898
54	59.44307705	0.004279556	58.97403454	0.004313593
55	59.29862281	0.013847202	60.22639696	0.013633889
56	62.82815139	0.01183721	61.73034293	0.012047722
57	63.06477251	0.00989887	62.25236524	0.010028053
58	63.45911062	0.009188909	62.73519521	0.009294942
59	62.91178717	0.01462683	62.02384027	0.014836231
60	64.425238	0.016896329	63.20654916	0.017222108
61	66.45059617	0.029548569	64.13896686	0.030613527
62	63.95021684	0.011166811	64.63241595	0.011048945
63	65.8510601	0.021628809	63.5839177	0.022400004
64	64.0931081	0.028794048	62.69267084	0.029437253
65	63.61809243	0.018103498	62.84598588	0.018325912
66	63.43268039	0.017900237	62.67005092	0.018118064
67	62.10019048	0.00244186	61.89103186	0.002450113
68	63.71379311	0.024715527	63.65203372	0.024739508
69	66.63357803	0.032441302	63.71224071	0.033928802
70	68.4063688	0.040409249	65.19442284	0.042400099
71	68.52314306	0.052271099	65.34501489	0.054813363
72	68.7496849	0.043377217	64.89445626	0.045954157
73	68.46112402	0.063265686	65.09740065	0.066534761
74	69.429769	0.069925193	65.98686188	0.073573585
75	69.29693449	0.060533847	66.05350265	0.063506246
76	67.54711882	0.002634013	67.10481199	0.002651375
77	68.9493956	0.033883836	66.12933665	0.035328798
78	69.40065146	0.06788236	65.43606963	0.071995155
79	66.21493772	0.027654334	64.60356577	0.028344101
80	68.13304135	0.037849771	63.89422431	0.040360769

	Cleansed Dataset		Raw Dataset	
Segment number	Mean Free-Flow Speed (Mph)	Mean Free-Flow TT (Hour)	Mean Free-Flow Speed (Mph)	Mean Free-Flow TT (Hour)
81	69.11564707	0.058634624	65.054351	0.062295141
82	68.74351664	0.029569043	64.86759516	0.031335831
83	69.54669978	0.053320143	65.94259672	0.056234364
84	69.6324421	0.066895399	65.74340628	0.070852581
85	65.71455304	0.064823236	63.63465181	0.066941986
86	61.57738067	0.064198249	59.16915332	0.066811164
87	69.1120471	0.068426421	65.86345312	0.071801428
88	68.53741509	0.029945104	64.46393814	0.031837335
89	68.79540773	0.103214738	64.36735015	0.110315245
90	64.33101384	0.01145917	61.76074717	0.01193606
91	65.20729204	0.021055161	60.84745779	0.022563802
92	64.8212189	0.036800604	63.00393087	0.037862082
93	66.54106319	0.07215304	64.10544052	0.074894423
94	67.75966665	0.043913292	64.66671716	0.046013624
95	67.36800295	0.090025231	63.36008548	0.095719883
96	69.49386424	0.145272969	63.73054155	0.158410391
97	65.77793942	0.0640879	64.74581694	0.065109534
98	66.14158063	0.011421408	64.84266484	0.0116502

Appendix B – Segment 85th percentile

85th Percentile Statistics for I-35 southbound segments.

Segment	Cleansed Dataset 85th Speed (Mph)	Raw Dataset 85th Speed (Mph)
1	65.86968377	65.26930547
2	72.34870957	70.2403965
3	71.07483392	69.05935077
4	71.51742405	67.84211459
5	71.8592873	67.81281833
6	72.23894079	68.9090222
7	71.87795974	69.14174774
8	72.61346272	68.49635651
9	72.24792589	68.2267387
10	70.64061208	70.00795137
11	71.9499672	68.3872716
12	70.69525299	69.13593709
13	72.01360687	68.03164656
14	72.33030844	69.73560137
15	72.62882577	69.92358513
16	72.42371343	68.56410637
17	73.61645542	69.75562667
18	70.97247156	70.15615174
19	72.547095	69.74972286
20	72.44961823	70.29799058
21	72.39799505	70.60129169
22	71.51054602	71.04467856
23	70.43248047	69.15259184
24	72.30689983	70.86195543
25	71.42540555	70.39021558
26	70.3797053	70.20233337
27	70.13875642	69.50081753
28	69.48681331	70.86592249
29	68.32345341	67.13205625
30	70.85124389	68.99065538
31	70.06500745	66.11549515
32	67.79082085	67.75494006
33	66.20243128	63.36386536
34	61.41264265	61.56850081
35	66.29032548	66.30565407
36	65.47507684	65.38079656
37	65.07777281	63.7712409
38	64.67946454	64.20319899
39	66.44957663	65.80001587
40	65.96895429	63.5868929
41	65.95047923	65.95047923
42	64.18180167	63.27311988
43	64.52024831	63.96446706
44	51.53899361	48.48420843
45	64.94501663	63.61133081
46	64.96243211	65.84983236

Segment	Cleansed Dataset 85th Speed (Mph)	Raw Dataset 85th Speed (Mph)
47	65.15108078	65.75537094
48	64.93376969	64.50230935
49	65.40016529	65.14536937
50	66.82967019	65.55977195
51	64.8053382	64.34480645
52	66.93830859	66.57387229
53	64.93880868	64.76044573
54	65.50450725	65.84957454
55	63.88279413	64.70832981
56	66.77113937	65.74112107
57	66.64629936	65.88827599
58	67.48900609	66.73407088
59	66.62485146	65.8124439
60	68.13472954	66.51034781
61	70.05440568	67.3090086
62	68.00987464	68.43905194
63	69.73594782	66.73945257
64	67.61302466	65.70823979
65	66.94134299	65.9296451
66	67.09231802	65.96599142
67	67.01880376	67.00770684
68	67.97744209	67.30110818
69	71.06774055	67.10197826
70	72.91213296	68.54952977
71	72.91105445	68.63542442
72	73.34535042	68.32470443
73	72.71077907	68.20130845
74	73.55524823	69.01425245
75	73.56020382	69.42101829
76	72.9537708	72.24048641
77	73.57812093	69.85396119
78	73.71980207	68.81547404
79	70.96546918	68.22892562
80	72.90797316	67.81317122
81	73.56111548	68.91104048
82	73.58976715	68.78220388
83	73.84708072	69.50912971
84	74.1377454	69.54628597
85	70.50337691	67.11530434
86	68.05492424	64.46671854
87	73.76753277	69.17747246
88	73.53606918	68.06676664
89	73.37918982	67.90756958
90	71.74779418	68.89733642
91	71.16520626	66.4881897
92	69.90900194	66.43048627
93	71.05810536	67.02308329
94	72.1841866	68.00657325
95	71.95376168	66.45030258
96	73.78785994	66.47864967
97	70.17351327	67.89593266
98	70.52800067	68.34023476

Appendix C – Segment Reliability Index Results

TTI, BI and PI for all segments of I-35 Southbound.

	Cleansed Dataset			Raw Dataset		
Segment	TTI	BI	PI	TTI	BI	PI
1	1.139358683	0.188340803	1.294207541	1.193652943	1.27053324	2.465210301
2	1.222937137	0.941895726	2.10137121	1.427348576	2.754765009	4.085241935
3	1.153085067	0.360080111	1.473022591	1.203651387	1.738778053	2.968346906
4	1.124710015	0.191754465	1.280487944	1.120122617	0.219216446	1.320511257
5	1.120834429	0.159637837	1.245972986	1.143874034	1.210802815	2.360346944
6	1.140245516	0.198632403	1.311072422	1.142912217	0.216958096	1.323162172
7	1.141222065	0.222989588	1.329521984	1.139397879	0.217462081	1.329147935
8	1.12343795	0.168482612	1.257839888	1.104345758	0.275289252	1.355667708
9	1.126319642	0.373102925	1.471924708	1.169809535	1.134154341	2.35274804
10	1.172722807	0.19627674	1.320232636	1.296814926	2.70698406	4.103849897
11	1.142093846	0.180189595	1.281680677	1.13273872	0.301868869	1.412799746
12	1.143672207	0.329676582	1.46005814	1.337500522	4.657759301	6.099516647
13	1.157935496	0.276064024	1.384757465	1.379306933	6.715918795	8.282164057
14	1.141267154	0.225433244	1.327264358	1.664800381	7.267064619	8.905211485
15	1.840234637	9.265768936	11.07844428	2.700214632	18.336497	20.87138286
16	1.752599548	12.79834397	15.33813579	1.7660273	12.43703469	14.92670002
17	1.160749771	0.242982316	1.333108448	1.19180989	0.722523512	1.839530407
18	1.677838574	3.407957994	4.759774575	2.472111212	15.59523284	17.7894932
19	2.153678192	14.05546207	16.18823809	2.122090528	13.60339883	15.61418527
20	1.205808497	0.514819659	1.625025979	1.237915628	1.427382031	2.592317735
21	1.734061404	5.500993987	7.012370653	1.827165344	6.073255732	7.602530178
22	2.056942637	6.276760229	7.945306848	2.243342029	7.712967218	9.516187269
23	1.344045684	1.462269432	2.66249112	1.651687074	4.833293972	6.352082891
24	1.387157924	2.153293511	3.454471196	1.407223123	2.334434471	3.66054684
25	1.470019048	2.862201032	4.219381896	1.495146028	2.953697564	4.358943684
26	1.60589859	3.23553081	4.713988403	1.921280864	14.45549005	17.36993648
27	1.544828298	3.043572581	4.44452547	1.555463667	3.145825024	4.425164114
28	1.495097193	2.283130754	3.724554133	1.577761185	2.568157187	4.162415898
29	1.445354341	1.65753905	3.128278513	1.662566671	1.869196661	3.670879589
30	1.354829678	0.974493509	2.265072601	1.367794435	1.109488546	2.392279078
31	2.723404505	9.094061193	10.9691557	3.060216745	9.998289379	12.10933622
32	2.084434263	6.818953214	8.514703862	2.121287707	6.81499158	8.579285172
33	3.018146871	19.53179571	22.573061	3.097996761	19.31342399	22.36827013
34	1.3573433	0.700604423	1.921578726	4.346577803	29.09533205	33.60541473
35	1.361769551	1.648229253	2.926557706	1.510090033	2.418088655	3.855575201
36	1.303305532	1.40343139	2.625927448	1.604872138	11.20172785	13.18373126
37	1.344887347	1.720818354	2.94539697	1.383637732	1.836224967	3.078439488
38	1.260693504	0.87977508	2.045740527	1.256397681	0.870989028	2.035013754
39	1.235663612	1.009090635	2.175487628	1.228114708	1.003363445	2.163804785
40	1.834659123	4.836649717	6.461981309	2.01474297	7.556842103	9.316071413
41	6.383835663	18.29057589	24.22707262	6.502588669	18.75599413	24.74258506
42	2.377341883	7.840921015	9.883149352	2.576347918	8.692818329	10.91321803
43	1.490288579	2.895528653	4.310731575	1.587799681	3.048824212	4.634719791

	Cleansed Dataset			Raw Dataset		
Segment	TTI	BI	PI	TTI	BI	PI
44	1.554221055	3.233273734	4.840271075	1.568954307	3.349843485	4.930734977
45	1.680487006	4.312954469	5.907326532	1.72259753	4.566832277	6.177566175
46	1.3806581	1.696889469	2.908122556	1.53151194	4.63983094	6.044781843
47	1.369414302	1.673445589	2.877224308	1.40721149	1.814328755	3.027387988
48	1.340352934	1.439590025	2.639363782	1.409165626	1.728643306	2.95839673
49	1.565238101	2.057697338	3.389074732	1.89107914	8.532447449	10.7984543
50	1.319834482	1.662970211	2.853698571	1.341738961	1.590040599	2.769771181
51	1.432216271	2.339567774	3.627332653	1.940858089	12.91616125	15.08884015
52	1.388083566	1.991526457	3.242840761	1.751621407	13.31817986	15.42241153
53	1.456327425	1.335452413	2.642270932	1.569327555	1.520655049	2.992955511
54	1.502247354	1.663985903	2.999614646	1.597896905	1.756132907	3.218609832
55	1.225201122	0.606309516	1.763027071	1.613776049	10.92638658	13.11384491
56	1.239101664	0.898311653	2.037578816	1.359185537	5.979001351	7.576369335
57	1.269464597	1.201669518	2.36326166	1.312793627	1.442757414	2.624613384
58	1.221831928	0.771622835	1.917904261	1.729275389	13.34445613	15.46752921
59	1.19752333	0.512988063	1.630297786	1.252387553	2.736752202	4.07072471
60	1.153226659	0.41317846	1.504772367	1.163530702	0.461060239	1.560606498
61	1.169532107	0.322714976	1.435449203	1.133520431	0.325202587	1.402314761
62	1.166218186	0.363690724	1.474871711	1.158255404	0.353909401	1.45687454
63	1.135882458	0.248156803	1.339426658	1.119682757	0.244695266	1.330078488
64	1.124861684	0.282218349	1.364845757	1.145159703	0.691632959	1.787309691
65	1.10720321	0.165636369	1.24433707	1.525315759	10.90557695	12.89991789
66	1.105971691	0.237196507	1.309574387	1.101057086	0.393732605	1.469768972
67	1.497087845	2.283250172	3.651398777	2.331783702	13.07295849	15.78183327
68	1.228635111	1.458639229	2.646789248	1.302111904	2.198884764	3.44615586
69	1.14082778	0.211774335	1.31181476	1.128722425	0.442672541	1.539911424
70	1.150543176	0.287894727	1.383020549	1.564754634	11.49931186	13.49306005
71	1.154836053	0.300644789	1.404252954	1.169807843	1.274963232	2.435236134
72	1.144347448	0.221663424	1.322648879	1.14210633	1.37469669	2.532394874
73	1.137562999	0.211600995	1.304318838	1.543724645	11.31159682	13.10421826
74	1.127767063	0.214913989	1.293547854	1.115563143	0.168153406	1.25347729
75	1.138129985	0.19307852	1.288824255	1.103802005	0.130234349	1.210014148
76	1.164527802	0.212074283	1.334949722	1.184697431	0.219458251	1.366647137
77	1.14808477	0.203262907	1.311990829	1.122522105	0.194461438	1.294943027
78	1.140720216	0.212000863	1.315708711	1.11884071	0.170841496	1.258188731
79	1.149633566	0.233693024	1.346012129	1.166158418	1.286495973	2.473064387
80	1.144213338	0.20105088	1.310738558	1.120976925	0.30648243	1.41134385
81	1.136600897	0.224174459	1.31526414	1.151637159	1.159741919	2.343652182
82	1.143381357	0.206331603	1.298885071	1.113005577	0.161689444	1.243426238
83	1.136906711	0.199708729	1.288108936	1.14394732	0.637786935	1.769723966
84	1.136083581	0.201969153	1.295166596	1.203746658	2.707028078	4.017876787
85	1.144912699	0.175689167	1.274874917	1.146730584	0.321980225	1.439576422
86	1.217744698	0.303403199	1.479365107	1.219496263	0.279338715	1.463138843
87	1.145499863	0.206760562	1.310065706	1.578880807	11.85111423	13.82925512
88	1.156930067	0.203362402	1.313361844	1.113756326	0.123428371	1.209068382
89	1.161411268	0.206620157	1.313597356	1.163995625	0.149257044	1.277353425
90	1.264036889	0.374529244	1.605376999	1.289182065	1.242498653	2.64103447
91	1.195715515	0.316335337	1.464898244	1.262923297	1.321266425	2.617355012
92	1.176995036	0.235954497	1.361438276	1.142253641	0.905741841	2.065587475
93	1.17170211	0.208102838	1.331688284	1.145701256	0.392377517	1.513475419

	Cleansed Dataset			Raw Dataset		
Segment	**TTI**	**BI**	**PI**	**TTI**	**BI**	**PI**
94	1.154143718	0.200040035	1.303046933	1.123548189	0.175493617	1.261481225
95	1.144131744	0.205971177	1.299069287	1.159710329	1.181900825	2.350134964
96	1.154068261	0.207461284	1.309302846	1.130873913	0.163746406	1.259901792
97	1.162582542	0.206692858	1.315841437	1.121703382	0.171662203	1.259749452
98	1.19813794	0.266793986	1.397365338	1.221204527	0.371941573	1.499604501

Appendix D – Incident Model Coefficients

Coefficient Number	θ1	θ2	θ3	θ4
1	0	0	0.0002	0.001
2	0	0	0.0002	0.0018
3	0	0.0001	0.0003	0.002
4	-0.002	0.0003	0.0017	0.0048
5	-0.0001	0	0.0017	0.0049
6	-0.0001	-0.0003	0.0015	0.0064
7	0	-0.0005	0.0008	0.0072
8	0	-0.0005	0.0005	0.0075
9	0	-0.0005	0.0003	0.0076
10	0	-0.0005	0.0003	0.0076
11	0	-0.0005	0	0.0076
12	0.0001	-0.0003	-0.0009	0.007
13	0.0001	-0.0003	-0.0012	0.0067
14	0	-0.0002	-0.0015	0.0058
15	0.0001	-0.0002	-0.0015	0.0055
16	0	0.0001	-0.0015	0.0046
17	0	0.0003	-0.0006	0.0018
18	0	0	0.0002	0.0015
19	0	0.0002	0.0005	0.0022
20	-0.0001	-0.0001	0.0006	0.0035
21	0	-0.0003	0.0003	0.0038
22	0	-0.0003	0.0002	0.0039
23	0	-0.0002	-0.0004	0.0038
24	0.0001	-0.0002	-0.0008	0.0032
25	0	0.0001	-0.0008	0.0023
26	0	0	0.0003	0.0015
27	0.0001	-0.0002	0	0.0018
28	0	0.0003	0.0003	0.0018
29	-0.0002	0	0.0011	0.0041
30	0.0001	-0.0004	0.0008	0.0047
31	0	-0.0003	0.0006	0.0049
32	0.0001	-0.0003	0.0004	0.0051
33	0	-0.0002	0.0002	0.0052
34	0	0.0001	-0.0001	0.0051
35	0.0001	0.0001	0.0001	0.0052
36	0.0001	0.0005	0.0009	0.0057
37	0	0.0006	0.0013	0.0061
38	0	0.0007	0.0024	0.0077
39	0	0.0007	0.0036	0.0103
40	0.0001	0.0008	0.0043	0.0121
41	0.0001	0.0009	0.0051	0.0143
42	0.0002	0.0012	0.007	0.0199
43	0.0001	0.0014	0.0083	0.0235
44	0.0001	0.0016	0.0098	0.0279
45	-0.0001	0.0017	0.0115	0.0332
46	-0.0004	0.0015	0.0131	0.0395
47	-0.0006	0.001	0.0144	0.0466
48	-0.0007	0	0.015	0.0544
49	-0.0007	-0.0011	0.0144	0.0623

Coefficient Number	θ1	θ2	θ3	θ4
50	-0.0007	-0.0014	0.0141	0.0644
51	-0.0005	-0.0022	0.0126	0.0697
52	-0.0002	-0.003	0.0098	0.0759
53	0	-0.0033	0.0062	0.0805
54	0.0001	-0.0034	0.0024	0.083
55	0.0002	-0.0032	-0.0015	0.0832
56	0.0003	-0.0029	-0.0051	0.0813
57	0.0004	-0.0025	-0.0084	0.0771
58	0.0005	-0.0018	-0.0111	0.071
59	0.0006	-0.0009	-0.0127	0.0635
60	0.0006	0.0003	-0.0131	0.055
61	0.0004	0.0014	-0.012	0.0467
62	0	0.0021	-0.0096	0.0394
63	-0.0004	0.0021	-0.0067	0.0339
64	-0.0004	0.0012	-0.0044	0.0301
65	-0.0002	0.0004	-0.0032	0.0274
66	0	0	-0.003	0.0252
67	0	0	-0.003	0.023
68	-0.0002	0	-0.003	0.0207
69	0	-0.0005	-0.0034	0.0182
70	0.0001	-0.0005	-0.0042	0.0152

Glossary

ATA : Actual Time of Arrival

ATRI : American Transportation Research Institute

CATT : Center for Advanced Transportation Technology

CDF : Cumulative Distribution Function

DOT : Department Of Transportation

DTA : Desired Time of Arrival

FHWA : Federal Highway Administration

FPM : Freight Performance Measures

GIS : Geographic Information System

GLRTOC : Great Lakes Regional Transportation Operations Coalition

HDFS : Hadoop File System

I-35: Interstate 35

IMPST : Interstate Mobility Performance Scanning Tool

MAP-21 : Moving Ahead for Progress in the 21st Century Act

MPO : Metropolitan Planning Organization

ODOT : Oklahoma Department Of Transportation

RITIS : Regional Integrated Transportation Information System

SQL : Structured Query Language

STD : Standard Deviation

TMC : Traffic Message Channel

UCR : Urban Congestion Report

VPP : Vehicle Probe Project