Oklahoma Department of Transportation
200 NE 21st Street, Oklahoma City, OK 73105-3204

FHWA-OK-17-02

NATIONAL PERFORMANCE MANAGEMENT RESEARCH DATASET (NPMRDS) – SPEED VALIDATION FOR TRAFFIC PERFORMANCE MEASURES

Hazem H. Refai, Ph.D.
Naim Bitar, M.Sc.
Muhanad Shab Kaleia, M.Sc.
School of Electrical and Computer Engineering (ECE)

Gallogly College of Engineering

The University of Oklahoma

Norman, Oklahoma

October 2017

The Oklahoma Department of Transportation (ODOT) ensures that no person or groups of persons shall, on the grounds of race, color, sex, religion, national origin, age, disability, retaliation or genetic information, be excluded from participation in, be denied the benefits of, or be otherwise subjected to discrimination under any and all programs, services, or activities administered by ODOT, its recipients, sub-recipients, and contractors. To request an accommodation please contact the ADA Coordinator at 405-521-4140 or the Oklahoma Relay Service at 1-800-722-0353. If you have any ADA or Title VI questions email ODOT-ada-titlevi@odot.org.

NATIONAL PERFORMANCE MANAGEMENT RESEARCH DATASET (NPMRDS) – SPEED VALIDATION FOR TRAFFIC PERFORMANCE MEASURES

FINAL REPORT ~ FHWA-OK-17-02

ODOT SP&R ITEM NUMBER 2300 (16-01)

Submitted to:

Dawn R. Sullivan, P.E.
Director of Capital Programs
Oklahoma Department of Transportation

Submitted by:

Hazem H. Refai, Ph.D.
Naim Bitar, Graduate Student
Muhanad Shab Kaleia, Graduate Student
School of Electrical and Computer Engineering (ECE)
The University of Oklahoma

October 2017

TECHNICAL REPORT DOCUMENTATION PAGE

1. REPORT NO.	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.
FHWA-OK-17-02		

4. TITLE AND SUBTITLE	5. REPORT DATE
NATIONAL PERFORMANCE MANAGEMENT RESEARCH DATASET (NPMRDS) – SPEED VALIDATION FOR TRAFFIC PERFORMANCE MEASURES	Oct 2017
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT
Hazem H. Refai, PhD, Naim Bitar, Muhanad Shab Kaleia.	
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. WORK UNIT NO.
The University of Oklahoma 660 Parrington Oval, Norman, OK 73019.	
	11. CONTRACT OR GRANT NO. ODOT SPR Item Number 2300 (16-01)
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
Oklahoma Department of Transportation Office of Research and Implementation 200 N.E. 21st Street, Room G18 Oklahoma City, OK 73105	Final Report Oct 2015 - Oct 2017
	14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

This report presents research detailing the use of the first version of the National Performance Management Research Data Set (NPMRDS v.1) comprised of highway vehicle travel times used for computing performance measurements in the state of Oklahoma. Data extraction, preprocessing, and statistical analysis were performed on the dataset and a comprehensive study of dataset characteristics, influencing variables, outliers and anomalies was carried out. In addition, a study on filtering and removing speed data outliers across multiple road segments is developed, and a comparative analysis of raw baseline speed data and cleansed data is performed. A method for improved congestion detection is investigated and developed. Identification and a computational comparison analysis of travel time reliability performance metrics for both raw and cleansed datasets is shown. An outlier removal framework is formulated, and a cleansed and complete version of NPMRDS v.1 is generated. Finally, a validation analysis on the cleansed dataset is presented. In the end, research affirms that understanding domain specific characteristics is vital for filtering data outliers and anomalies of this dataset, which in turn is key for calculating accurate performance measurements. Thus, careful consideration for outlier removal must be taken into account when computing travel time reliability metrics using the NPMRDS.

17. KEY WORDS	18. DISTRIBUTION STATEMENT		
ITS, Traffic Analytics, Data Analytics, Machine Learning, Travel Time, NPMRDS.	No restrictions. This publication is available from the Office of Research and Implementation, Oklahoma DOT.		
19. SECURITY CLASSIF. (OF THIS REPORT)	20. SECURITY CLASSIF. (OF THIS PAGE)	21. NO. OF PAGES	22. PRICE
Unclassified	Unclassified	118	N/A

Form DOT F 1700.7 (08/72)

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

SYMBOL	WHEN YOU KNOW	MULTIPLY BY	TO FIND	SYMBOL
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
		NOTE: volumes greater than 1000 L shall be shown in m^3		
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
^oF	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	^oC
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

SYMBOL	WHEN YOU KNOW	MULTIPLY BY	TO FIND	SYMBOL
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
^oC	Celsius	1.8C+32	Fahrenheit	^oF
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380. (Revised March 2003)

Acknowledgements

Principal Investigator Dr. Refai and his research team recognize the Oklahoma Department of Transportation (ODOT) for providing funds to support the research activities of this project. Additionally, ODOT personnel are highly acknowledged for support in testing the developed system and engaging in insightful discussions throughout this project.

This page is blank.

Table of Contents

List of Tables

List of Figures

Executive Summary

Traffic congestion is customary in urban areas and is a main source for abated productivity (due to traffic delays) and increased imperil (due to the extended time in the automobile). Moreover, the effects of traffic congestion on society include an increase in fuel consumption, pollution, and vehicle wear. The economic effect is a major burden for citizens and states alike. One solution to alleviate this problem is to increase state roadway and highway capacity. Doing so, however, is cost prohibitive. A preferable alternative is to better manage current roadway assets using intelligent traffic management systems, which improve traffic flow and reduce road congestion. These systems, however, require improved traffic performance measurements that deliver accurate insight to roadway and traffic conditions.

Variables like segment travel time, speed, delay, reliability, and origin-to-destination trip time are measures frequently used to monitor traffic and improve traffic flow on state roadways. In 2014, ODOT was given access to the FHWA's first edition of the 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 also subsequently segregated into passenger vehicle travel time and freight travel time. Both travel time are calculated using Global Positioning System (GPS) locations transmitted from participating drivers traveling along interstate highways.

This report presents research detailing the use of the first version of the National Performance Management Research Data Set (NPMRDS v.1) dataset of highway vehicle travel times used for computing performance measurements in the state of Oklahoma. Data extraction, preprocessing, and statistical analysis were performed on the dataset. A comprehensive study of dataset characteristics, influencing variables, outliers and anomalies, and recommendations for improving accuracy and alleviating data anomalies are reported and presented. Furthermore, a study on filtering and removing speed data outliers across multiple road segments is developed, and a comparative analysis of raw baseline speed data and cleansed data is performed. Identification and computational comparison of travel time reliability performance measurements is provided. A method for improved congestion detection is investigated and developed. An outlier removal framework based on the analysis study is formulated, and finally a cleansed version of NPMRDS v.1 is generated and presented along with a validation analysis on the cleansed dataset which is shown.

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, budgeting 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.

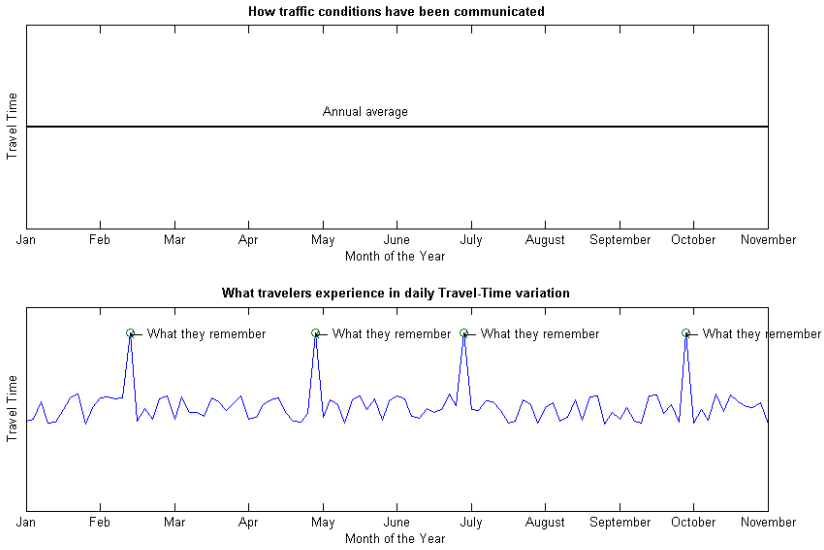


Figure 1 - Theoretical vs. perceived notion of congestion.

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 as "*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.

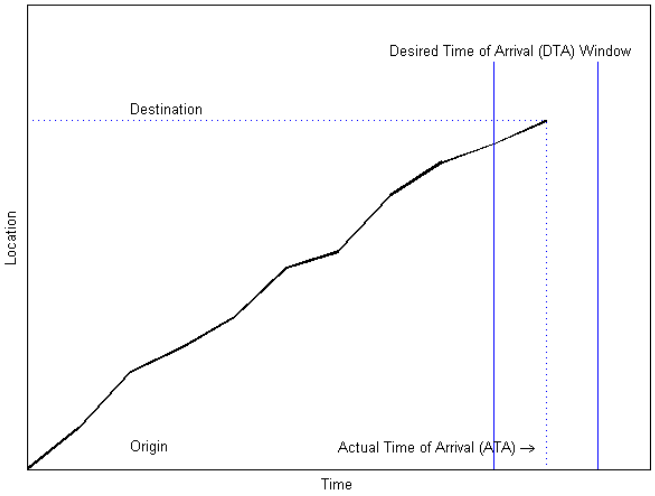


Figure 2 - Desired vs. actual times of arrival in defining travel time reliability.

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 an accident, can also impact traffic flow by distracting drivers that 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 such as elevated 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. Day-to-day variability in demand leads to higher traffic volume on some days more than 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 (due to unique 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 disjointed 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 many 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 as a key performance measure.

1.4. National Performance Management Research Data Set (NPMRDS)

The Federal Highway Administration (FHWA)—recognizing the importance of travel time reliability and its significance for quantifying the benefits of traffic management and roadway operations—offered state DOTs access to 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 was designed to complement existing state DOT's travel time measurements and reports. The relationship of the National Performance Management Research Data Set (NPMRDS), hereafter referred to as NPMRDS v.1 (to distinguish it from the second release procured recently in 2017 from a different provider, namely INRIX) and the Oklahoma Department of Transportation (ODOT) is the focus of this report and all work presented herein.

1.4.1. Overview of the NPMRDS

In 2013, the FHWA acquired a national dataset of average travel times: NPMRDS v.1. This information was intended for use in FHWA's 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) were encouraged to 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 collected 24 hours-a-day in 5-minute intervals report freight truck and passenger vehicle travel times, as well as combined vehicle travel time records.

NPMRDS v.1 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 generated by mobile phones, vehicle navigation systems, and portable navigation devices [8]. Freight data is embedded in fleet data-collection systems. Combined vehicle travel time data is a weighted average of passenger vehicle and freight travel times based on respective traffic volumes. Neither passenger nor freight volumes were 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 and vehicle type. Probe coordinates are based on GPS equipment (e.g., smartphones, navigation devices) located in vehicles. Recorded data is referenced to segments on a map. 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. Information is updated only when necessary changes are present. Table 1 details information

associated with the static NPMRDS file and provides a description of each entry.

A separate NPMRDS v.1 data file reported average travel times for roadways geo-referenced to each of the TMC location codes. Table 2 details the 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 would be performed on the dataset after it is collected by HERE, the dataset is known to contain unique characteristics that make traditional processing techniques routinely performed by DOT agencies ineffective, at best. 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 maintains 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). • **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 period to which the average speed applies (local time)
Travel_TIME_ALL_VEHICLES	Integer	44	Travel times calculated in seconds representing the time between segment length and the average speed on the segment. Average segment speed is determined from a combination of passenger and freight truck GPS probe speed observations.
Travel_TIME_PASSENGER_VEHICLES	Integer	76	Travel time 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 time calculated in seconds between the segment length and the average speed on the segment. Average segment speed is determined from only freight truck individual GPS probe speed observations.

1.4.2. Existing and related work using NPMRDS

Currently, DOTs, MPOs, and research institutions with some experience analyzing probe data and performing big data analytics are utilizing NPMRDS v.1 data in their performance measurements and reliability reports. Public documentation describing the NPMRDS v.1 dataset was first made available in November 2013, via a presentation given by the FHWA Office of Operations and Resource center, HERE, and The Volpe Center [7]. Soon afterward, research was reported by academic institutions and other organizations who were interested in investigating ways to utilize the dataset. One of the earliest presentations was made in February 2014 by the Wisconsin Traffic Operations and Safety Laboratory during the second quarterly NPMRDS webinar [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 v.1 and the I-95 Corridor Coalition's Vehicle Probe Project (VPP). Results indicated that direct comparisons between various 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 v.1 data gathered in-state 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 American Transportation Research Institute (ATRI) published work using NPMRDS v.1 data to compute congestion and the cost of delay incurred by the trucking industry [14]. Freight truck data from NPMRDS v.1 and data from ATRI's Freight Performance Measures database was used in the study. During the third quarterly NPMRDS v.1 webinar in May 2014, Iteris shared their work implementing performance measures for Utah DOT [15]. The presentation indicated that data imputation was the result of smoothing, although no filtering was applied to the dataset. A study comparing NPMRDS v.1 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 v.1 were accurate 95% of the time, and reliability measures were accurate only 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 v.1 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) publication [19]. Researchers discussed their methodology for processing NPMRDS v.1 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 v.1 data and Bluetooth traffic probe data from INRIX. Researchers recommended investigating NPMRDS v.1 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 v.1 data for analyzing road traffic congestion.

All related, published work relied on data imputation with no filtering or a process for outlier removal for the NPMRDS v.1 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 (STD) 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 v.1 dataset were the impetus for researchers to recommend data pre-processing by: 1) eliminating speeds below 2 mph, 2) resetting all speeds above the speed limit to the posted speed limit, and 3) 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 v.1 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 v.1 data is compared with probe data from INRIX. Results indicate NPMRDS v.1 data has a lower mean for travel time with a higher variance than data from INRIX. Finally, in 2016 the University of Oklahoma reported a detailed study on this dataset [27].

Several academic research communities have developed tools based on NPMRDS v.1 probe data. The University of Wisconsin developed a traffic tool for Wisconsin DOT that features an interactive map of the interstate system based on NPMRDS v.1 data [28]. A working prototype operations coordination mapping application, namely "The Interstate Mobility Performance Scanning Tool" (IMPST) [29], 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 v.1 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 [30]; INRIX, which provides roadway congestion information in real time and claims to report accurate real time traffic conditions; and Iteris [31], 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 v.1 data.

In spite of the aforementioned activity, one online investigation has proven that few DOT agencies are utilizing the NPMRDS v.1 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 v.1 dataset [32] (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, 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 v.1 data. In many cases, however, this finding was contrary to results presented herein for the current NPMRDS v.1 data. In fact, shorter segments exhibit an unknowingly problematic anomalous data, as will be shown in subsequent chapters of this report. 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 segments and those after that 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 this report

This work presents research detailing the use of the NPMRDS v.1 for computing performance measures in the state of Oklahoma. Data extraction, preprocessing, and statistical exploratory data analysis were performed on the NPMRDS v.1 dataset. Baseline historical raw calculations of road segment speed average (including outliers), variance, and STD across various time scales are shown. A comprehensive study of NPMRDS v.1 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 a comparative analysis of raw baseline NPMRDS v.1 speed data and cleansed data is presented. A method for improved congestion detection was also investigated and presented. Also, identification and computational comparison of free flow speed, 85th percentile, and travel time reliability performance measures computed using both raw and cleansed data is shown.

The main contributions of this work are summarized below.

- This work applies traffic data analytics and statistical analysis to the NPMRDS v.1 for developing 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 authors' 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 v.1 dataset. The process is highly beneficial for agencies and researchers interested in working with the NPMRDS v.1 dataset.

The balance of this report is organized in the following manner. The next chapter presents the framework and tools utilized for NPMRDS v.1 data acquisition and preprocessing. It also provides information for a necessary understanding of the unique characteristics of the NPMRDS v.1 data, with a focus on challenges associated with probe data. Chapter 3 presents a study on detecting anomalies/outliers in the dataset and develops models for alleviating said anomalies. 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 the final results. The chapter also includes an improved approach for detecting congested segments. Reliability performance measure computations follow in Chapter 5, wherein free flow, 85th 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 an inclusive framework process for filtering outliers for both types of vehicles, which caters to the NPMRDS v.1 domain, along with a validation study on the generated cleansed dataset. Finally, chapter 7 presents a developed web graphical user interface (GUI) for accessing and utilizing the cleansed and raw NPMRDS v.1.

Chapter 2: NPMRDS v.1 Acquisition, Characteristics and Processing

NPRMDS v.1 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 interstate highway in Oklahoma, namely Interstate 35 (hereafter, I-35). Furthermore, this chapter discusses limitations and challenges associated with utilizing NPMRDS v.1 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 v.1 data files from a shared FHWA repository accessible only by state DOT 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.

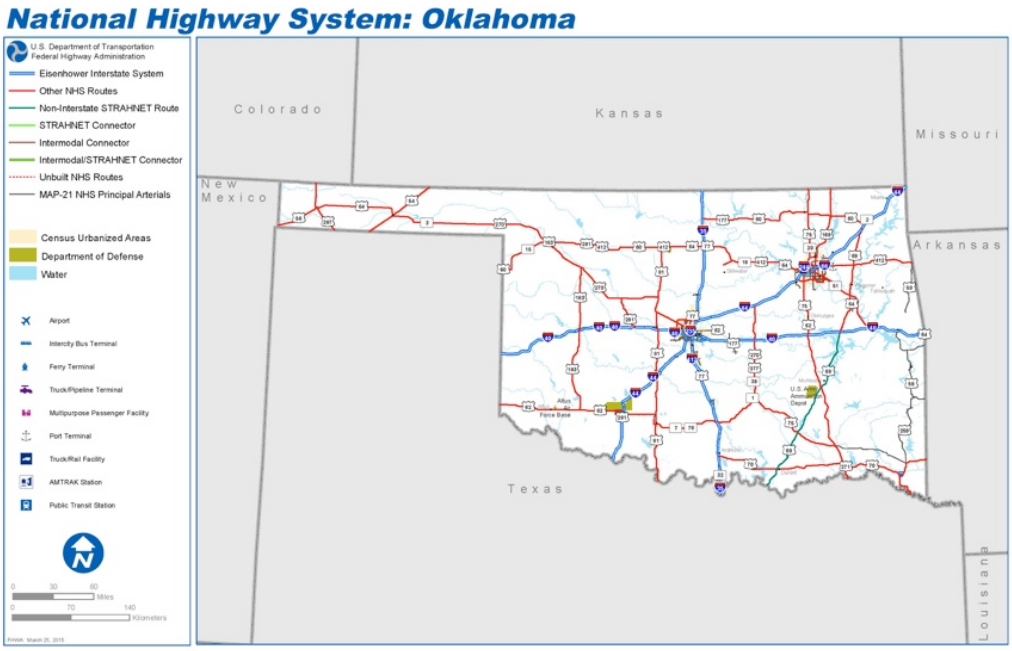


Figure 3 - NHS roadways in Oklahoma.

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 the NPMRDS v.1 static file for the 1st quarter of 2015, NHS roadways in Oklahoma are composed of 4,323 defined segments, each generating one epoch every five minutes. This 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 [33].

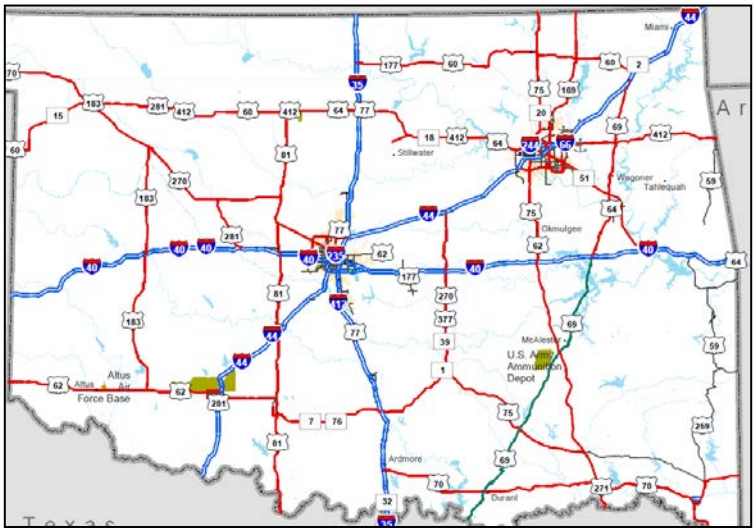


Figure 4 - NHS roadways in Oklahoma - magnified.

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 [34]. 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 v.1 data. See Figure 6.

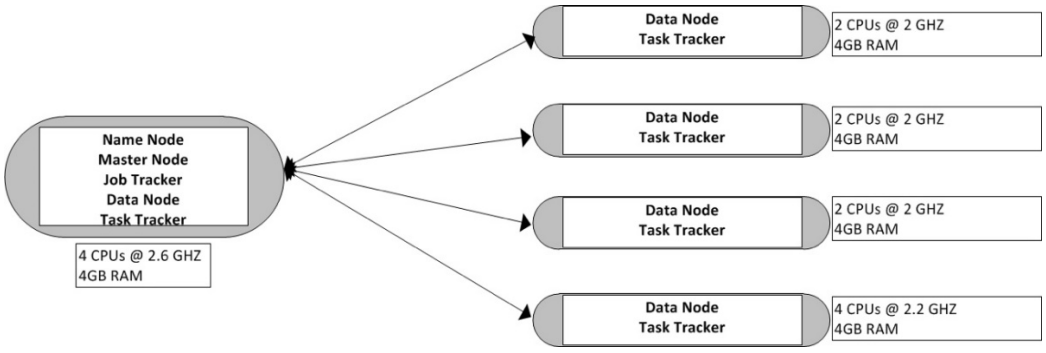


Figure 6 - Illustration of the 5 node Hadoop setup.

Processing using Hadoop begins with porting the travel time files from the PC storage 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 data manipulation and processing. In turn, users can quickly and efficiently extract any record according to a predefined criterion from the millions of available records. The following steps are done 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 [35]. 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 report 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 I-35 TMCs, 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
```

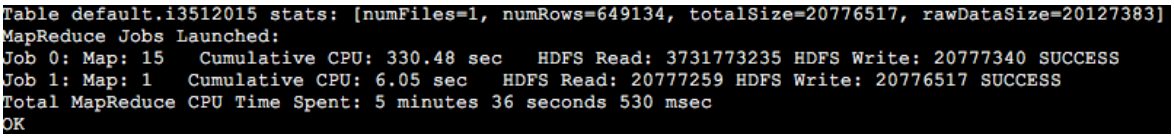


Figure 7 - Output of Hive.

Figure 7 shows Hadoop's 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 illustrates examples for I-35 southbound.

2.3. Dataset characteristics: challenges and limitations

As aforementioned, NPMRDS v.1 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 v.1 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 or 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 v.1 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 v.1 data requires database and scripting expertise [23].

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

NPMRDS v.1 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 45, segment (111N04920), per day for 31 days during the month of January 2015. Mean value of recorded epochs was 219.5806, and STD was 20.0678. Clearly, the number of epochs for the same segment fluctuate daily.

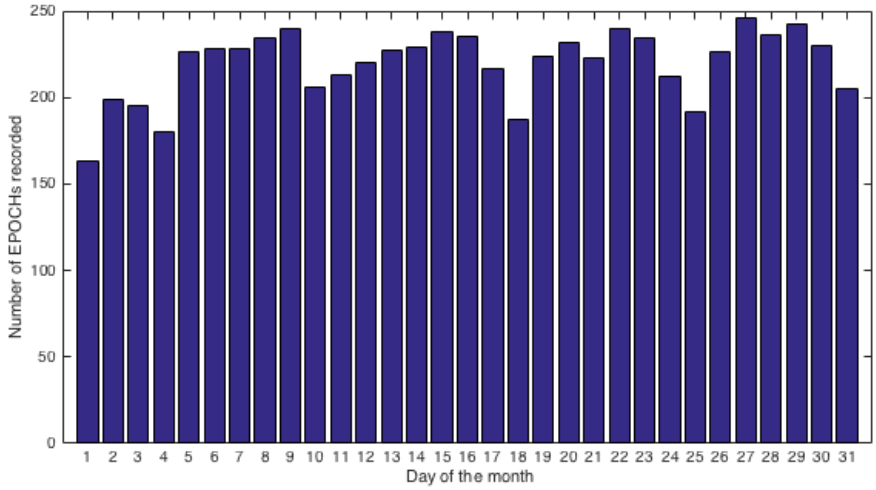


Figure 9 - Daily bar plot of epochs recorded for TMC 45 during January 2015.

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.

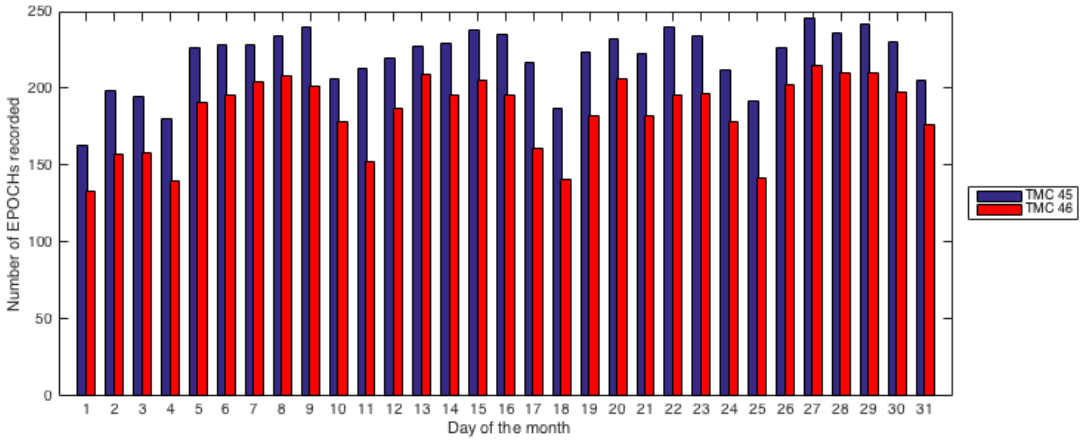


Figure 10 - Bar plot of epochs recorded for segments 45 and 46 during January 2015.

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 Epoch Percentage for Each Time Group 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 comparable results, as shown in Table 5 and

Table 6.

Table 5 - Probe Epochs Available per Time of 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 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 v.1 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, yet 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.

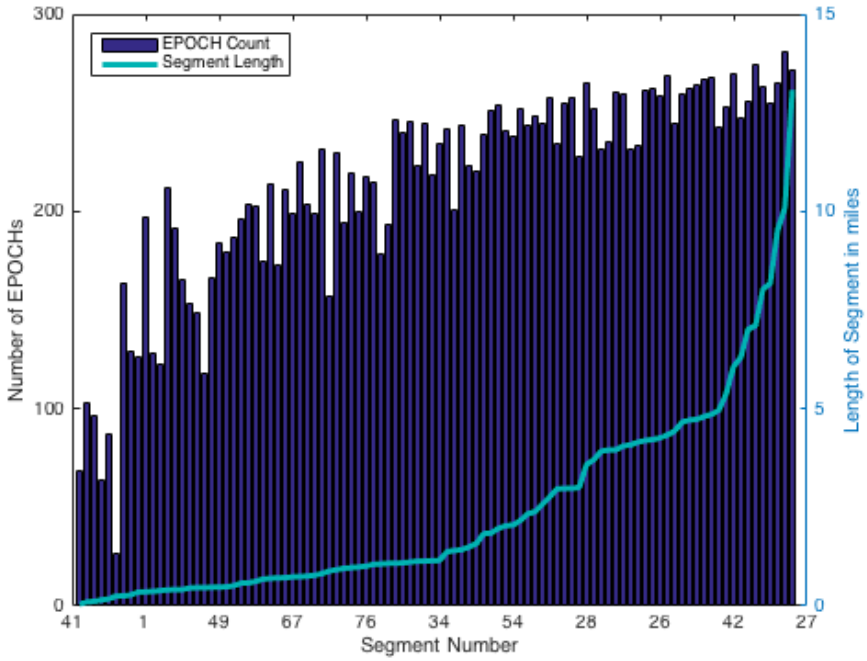


Figure 11 - Trend plot for number of epochs recorded versus length of segment.

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.

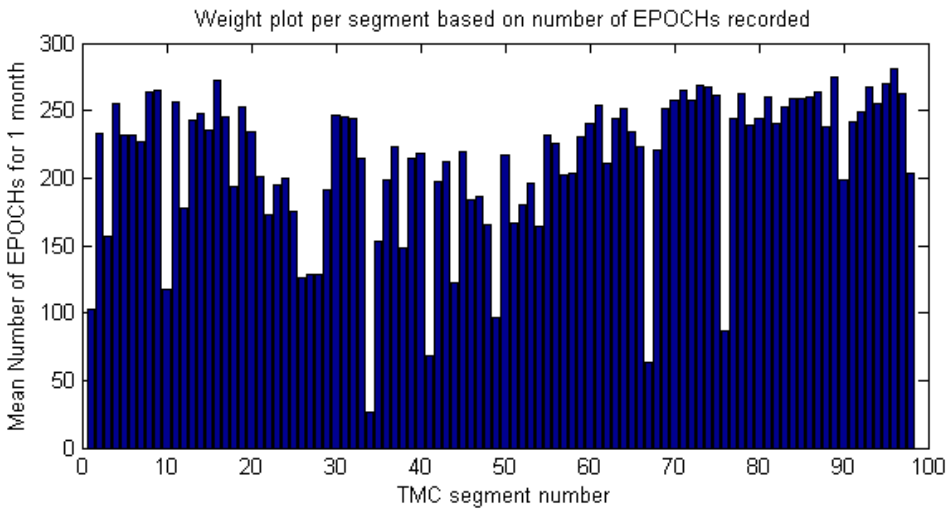


Figure 12 - Average number of epochs recorded per day reported per segment.

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 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 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.

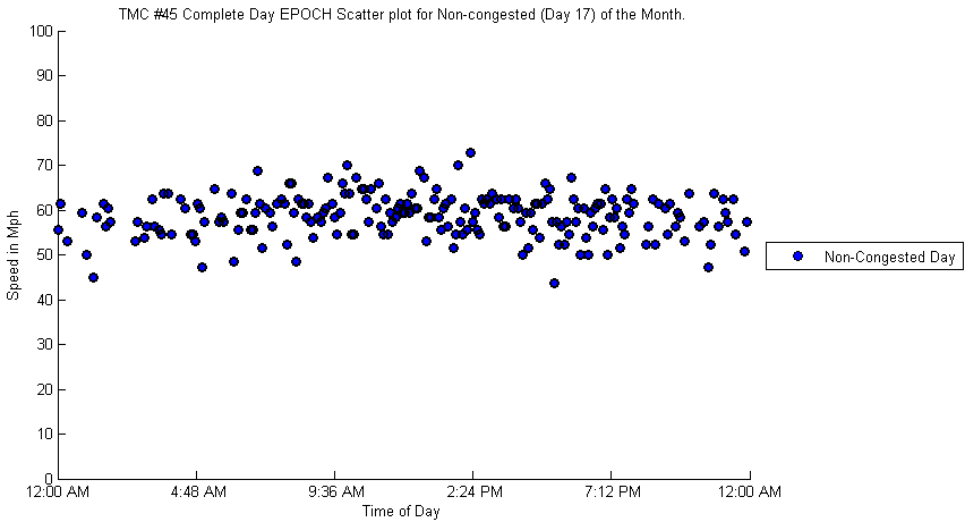


Figure 13 - TMC 45 complete day epoch scatter plot for non-congested day in January.

2.3.8. GPS in-accuracy:

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.

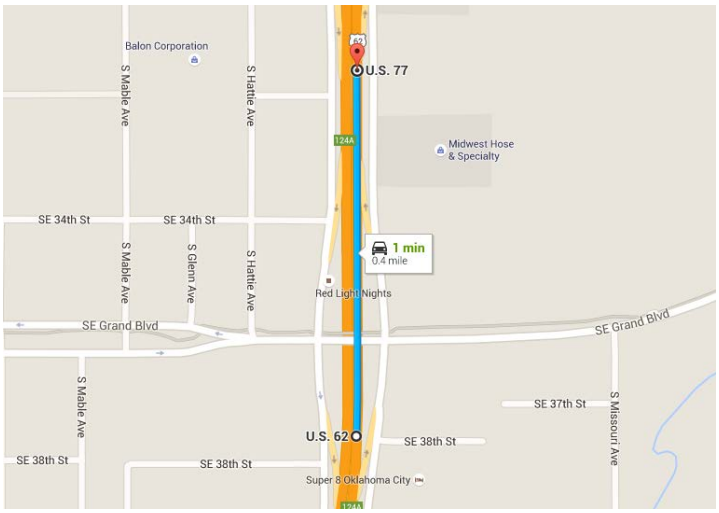


Figure 14 - Map view of TMC 47 crossroads with a major arterial.

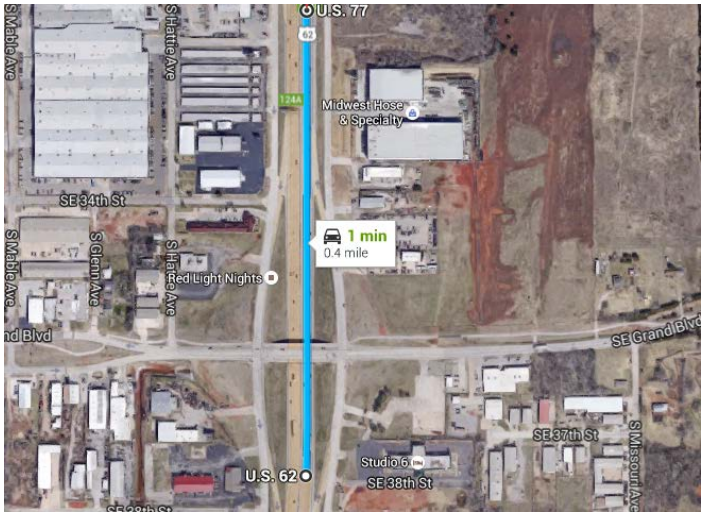


Figure 15 - Satellite view of TMC 47 crossing a major arterial.

Chapter 3: Anomaly and Outlier Study and Analysis

In the previous chapter, limitations and challenges inherit in the NPMRDS v.1 dataset were described and discussed. Despite the challenges, the NPMRDS v.1 dataset has important advantages that make it a valuable tool for crafting traffic performance measures. For example, because NPRMDS v.1 is a probe data set, travel times can be easily collected from different geographic regions. Compared to traditional fixed location detectors, NPRMDS v.1 data has higher granularity without the confines of location or forced infrastructural physical constraints. Moreover, NPRMDS v.1 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 v.1 data. A thorough understanding of the domain characteristic is necessary for accurate and effective statistical processing. Accordingly, the challenges 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 v.1) 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 v.1, 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 v.1 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.

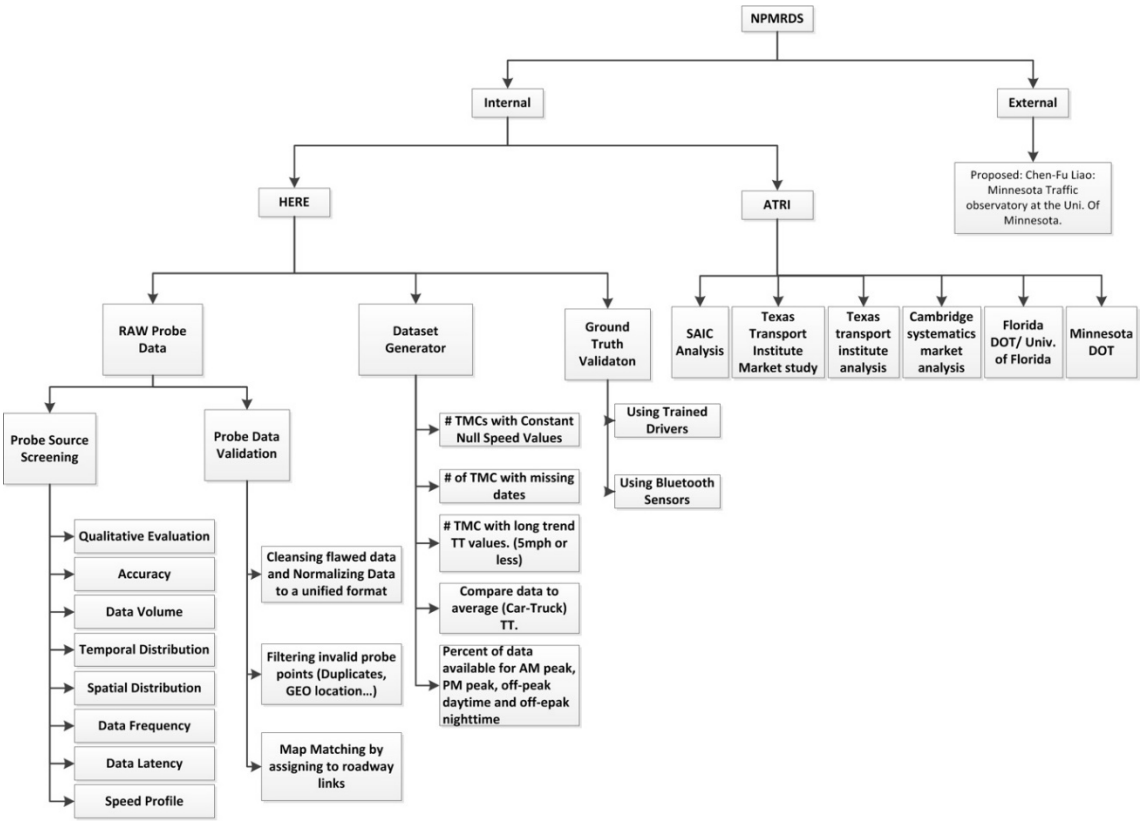


Figure 16 – NPMRDS v.1 data validation and quality assurance conducted by HERE.

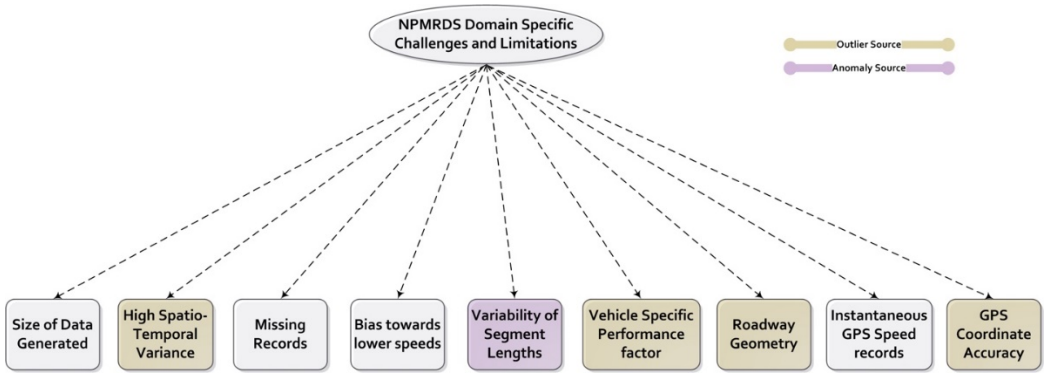


Figure 17 – Summary of limitations generating outliers and anomalies in the NPMRDS.

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 [36] and proved mathematically by T. Hill in 1995 [37]. Benford's law is applicable to occurrences of natural events [38]. 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 this 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 contrary, 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 [39]—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.

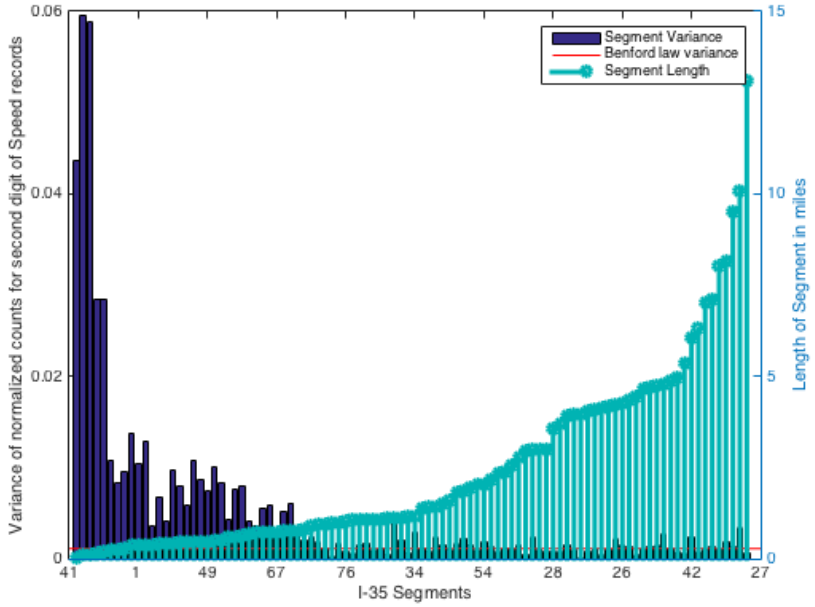


Figure 18 - Variance between percentages of digits vs. length of segment on I-35.

This fact gives insight that the NPMRDS v.1 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 v.1 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 v.1 data, it is clear that exactly 2 speeds were reported.

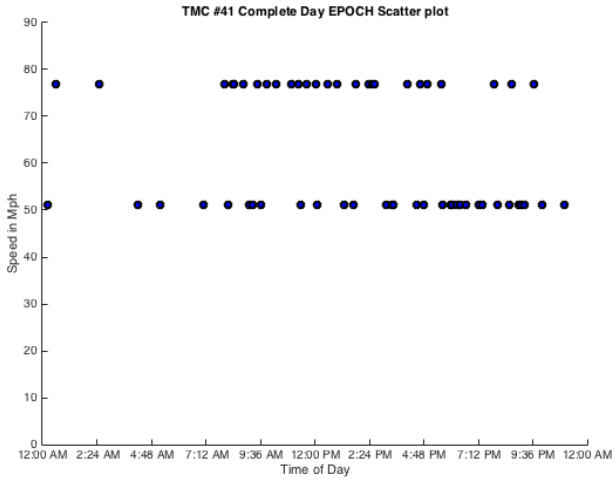


Figure 19 – Segment 41 daily epoch plot.

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∈{1,2,3,….}. 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.

$$\mathrm{V_2} - \mathrm{V_1} = E_r$$

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

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

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

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

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

where D is given in miles (M); $\mathrm{T}_{\mathrm{gran}}$ is the reported time granularity in seconds (s);$\mathrm{T}_{\mathrm{time}}$ 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.

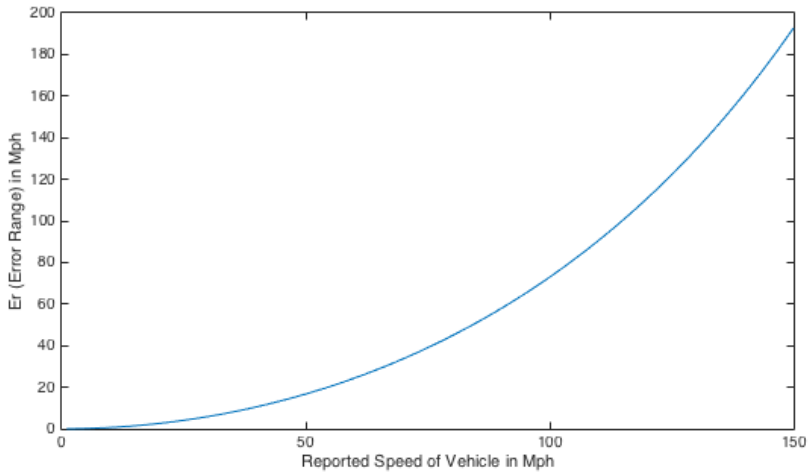


Figure 20 - Plot of vehicle speed vs. error range in mph for Segment 41.

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 examined for I-35 southbound. 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:

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

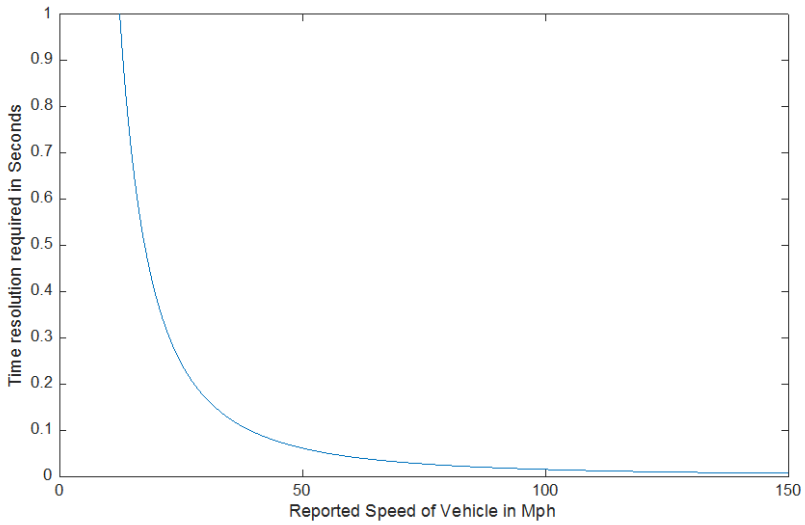


Figure 21 - Plot of vehicle speeds vs. time resolution for Segment 41.

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 to compute minimal segment length and 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 I-35 southbound, there are 50 segments shorter than this distance, meaning that 50 out of 98 segments are affected by this anomaly. Statistical analysis using NPMRDS v.1 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, as an example of a segment which is of length 1.373 miles—longer than the minimum distance calculated.

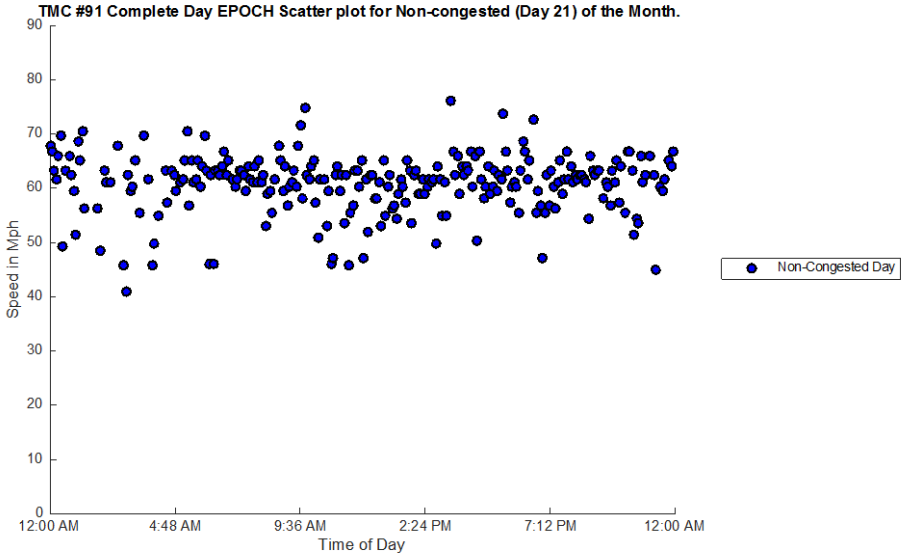


Figure 22 - Segment 91 reported speed scatter plot.

We observe the natural occurrence of randomness in speeds to be present in this segment. Moreover, for the purpose 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.

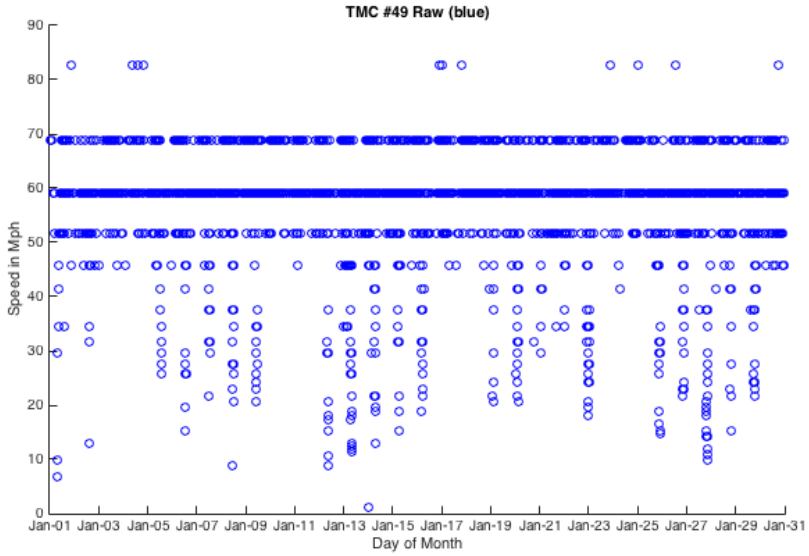
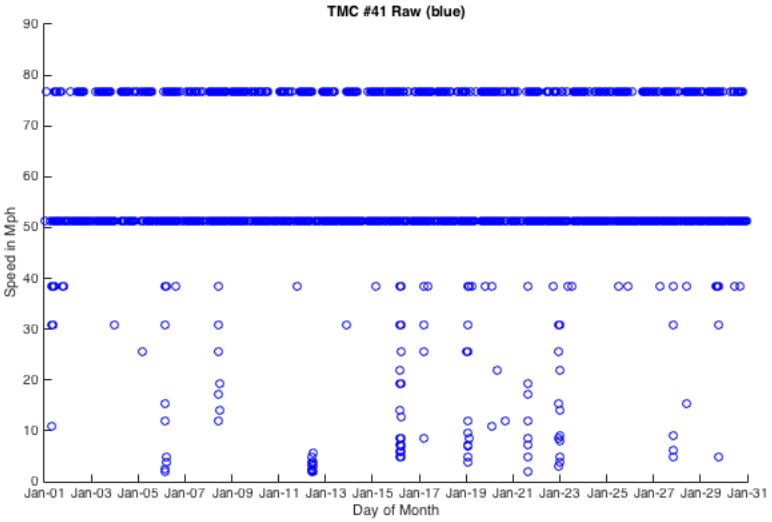


Figure 23 - TMC 49, January 2015 monthly speed plot of E_r at different speeds.

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 exist cases of extreme congestion where cars come to an almost complete stop.



.

Figure 24 - TMC 41, January 2015 monthly speed of E_r at different speeds.

In conclusion, the aforementioned study indicates that the FHWA should recommend to HERE changing time granularity of NPMRDS v.1 data reported according to Equation (2), which should alleviate inherent errors in the nationwide NPMRDS v.1.

3.2. Data Outliers:

Congestion on segmented roadways is a function of both time and space. In space, a shock wave like distribution of travel time 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 persists. 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. If 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 v.1dataset.

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 v.1 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 Wisconsin Madison in [23] pointed to this type of outlier and recommended scanning for observations that are several STDs 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 STDs above the mean. In the work presented in this report, average speed above 3 mean STDs from the speed limit is considered an outlier. This equates to approximately 90 mph on a roadway with a speed limit of 70 mph. Reported NPMRDS v.1 time/speed represent 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.

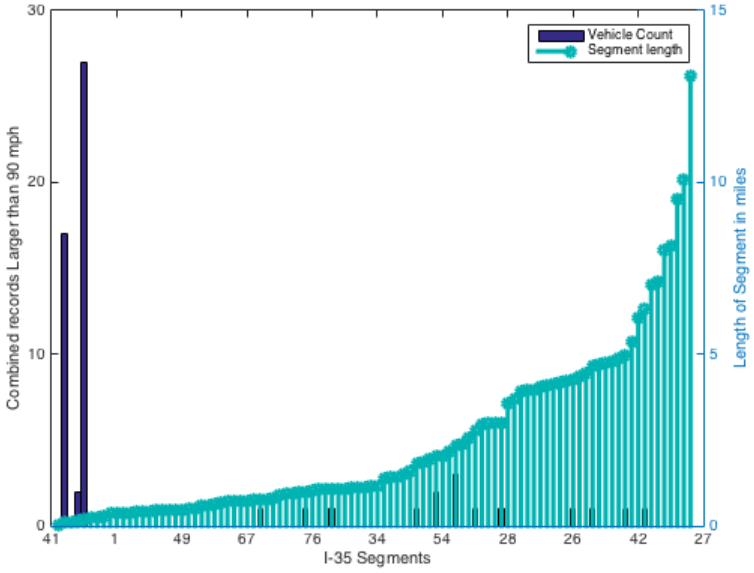


Figure 25 – Combined-vehicle count plot of epochs greater than 90 mph for I-35 S.

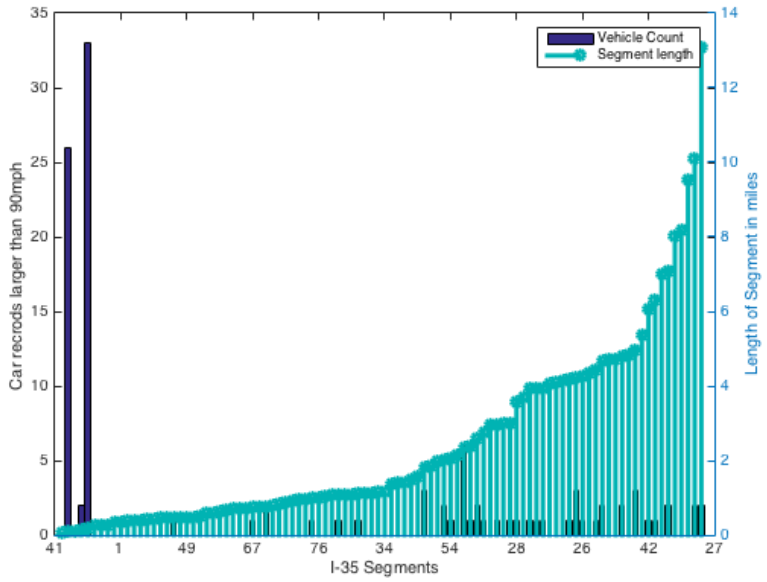


Figure 26 - Passenger vehicle count plot for epochs greater than 90 mph for I-35 S.

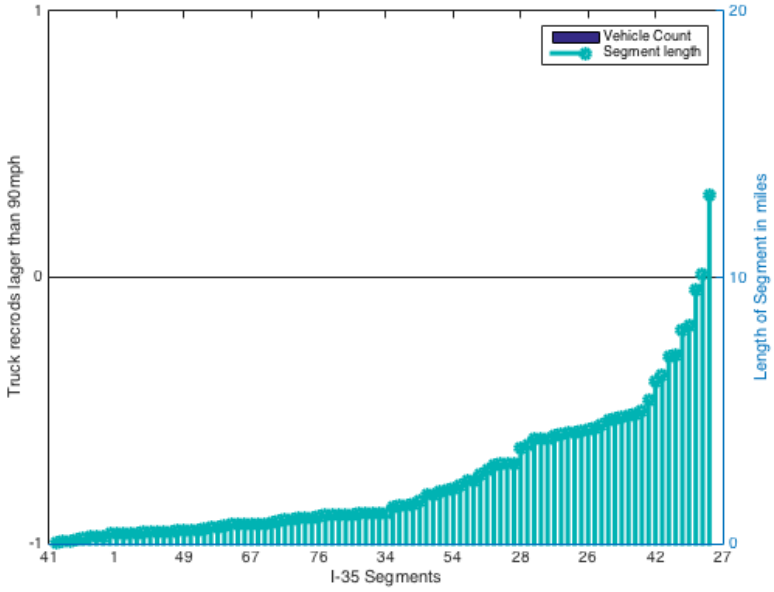


Figure 27 - Truck vehicle count plot for epochs greater than 90 mph for I-35 S.

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 instance 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 is 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 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 when congestion is detected in trucks recording slow speeds in correlation with passenger vehicles recording faster speeds (e.g.,15 mph higher than trucks), the faster speed characterized by the passenger vehicle represents a better approximation to the traffic flow condition of the road, while the slower truck speed characterizes 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 two and remove the outlier. Thus, detection is done by correlating speeds of freight and passenger vehicles for the same epoch and segment, and removal is done by replacing speed entries with the higher of the two speeds.

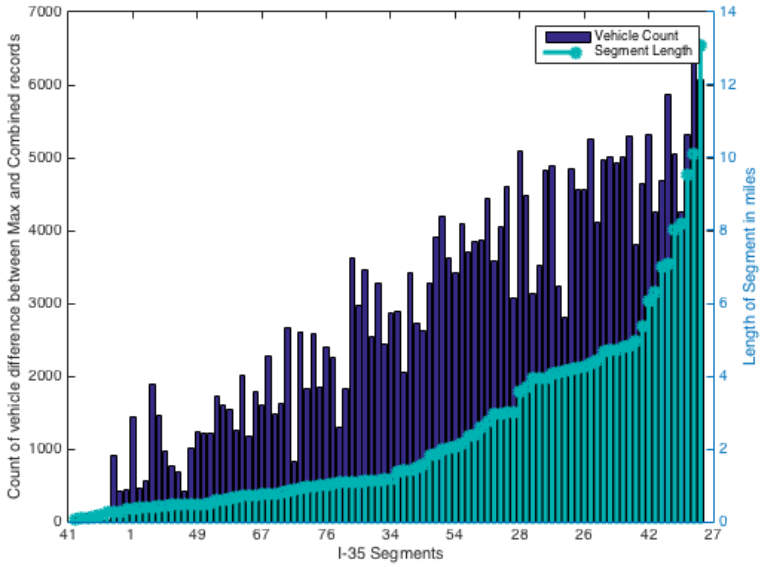


Figure 28 - Epoch count for difference of max (truck, car) to combined for I-35 S.

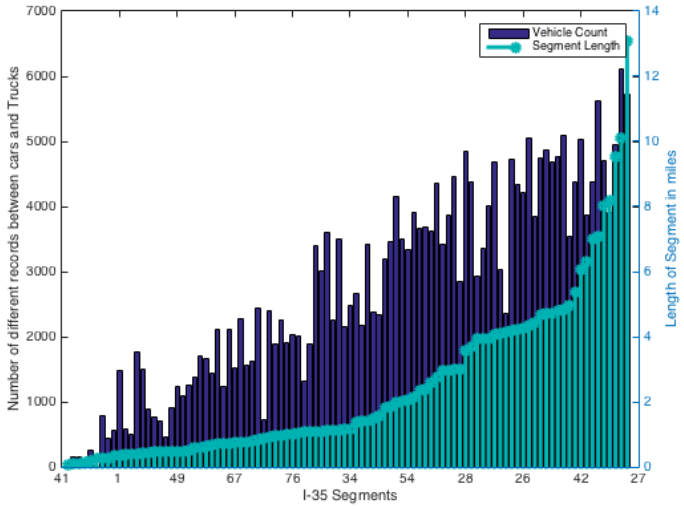


Figure 29 - Epoch record count for difference between car and truck matrices.

Figure 28 shows a plot of the maximum speed matrix generated, subtracted from the combined-all vehicles matrix. Figure 29 shows a plot of the number of epochs where passenger car speeds are 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, freight and passenger vehicle speed variation increased. This was confirmed when examining the percentage of epochs that

had the aforementioned speed difference (i.e., between trucks and cars) relative to the total number of epochs available per segment. See Figure 30 for a plot of this ratio.

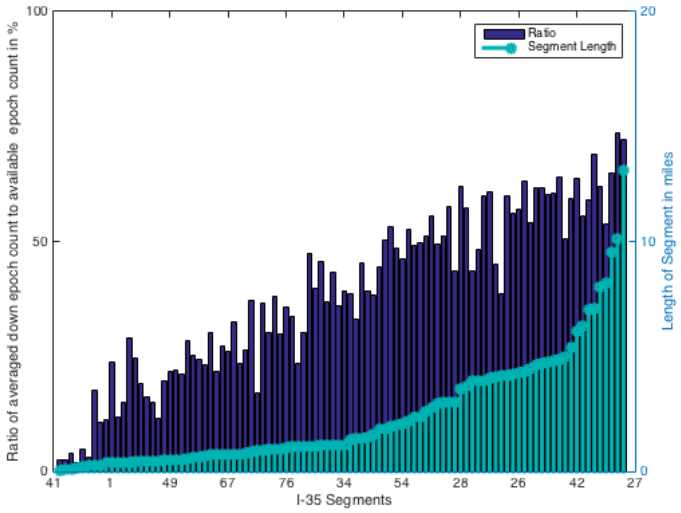


Figure 30 - Ratio of epoch count with difference in car-truck speed to total epochs.

Figure 31 and Figure 32 show the mean and the STD of the speed difference between the maximum and the combined vehicle speeds. Average difference for most segments is approximately 5 mph, and the STD is approximately 2 to 3 mph. As segment length decreases, mean increases. Reported combined speeds in the NPMRDS v.1 dataset show on average a 5 mph reduction in speed compared to slower freight trucks.

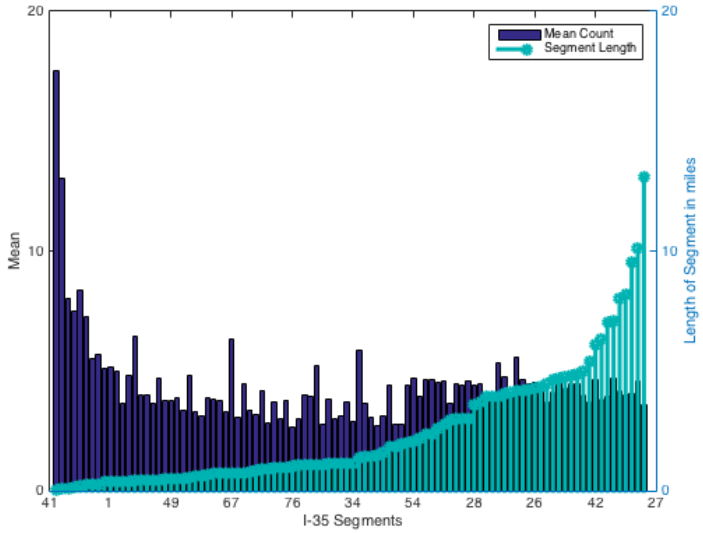


Figure 31 - Mean speed difference between max and combined speeds.

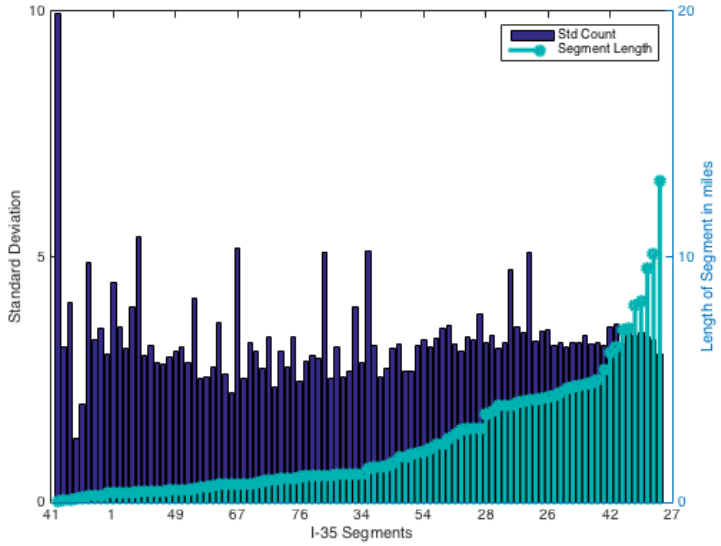


Figure 32 - Standard deviation of speed difference between max and combined speeds.

3.2.3. Roadway geometry

When roadway geometry is the underlining cause affecting traffic flow of vehicles on the road, then segment travel time reported should be affected at all times regardless of the record time or day. This phenomenon 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. However, roadway conditions in certain cases might only affect larger truck speeds and not passenger car speeds. As such, cases where slow traffic, and in particular freight speeds, were identified to be congested continually were marked for post check. No changes were done to the dataset for this type of outlier. However, it was thought that these cases would be of interest to DOT agencies, as they show locations where segments could possibly undergo optimization for freight travel time.

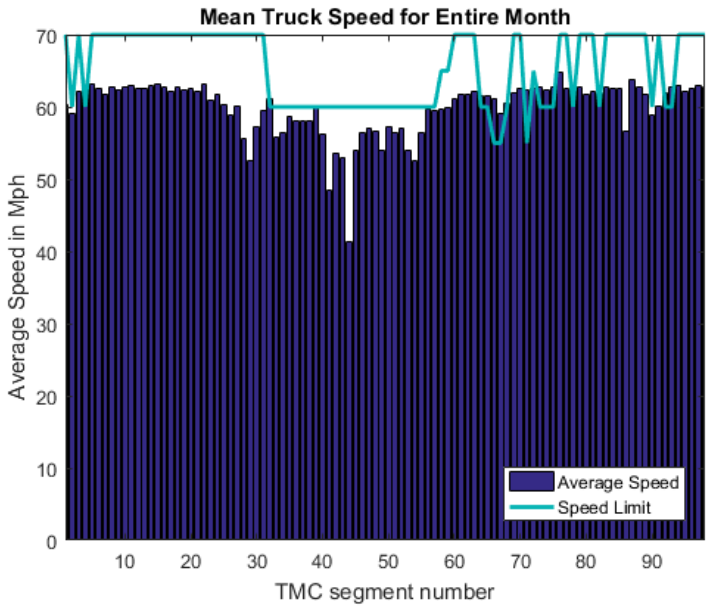


Figure 33 - Average epoch truck speed per segment for January 2015.

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

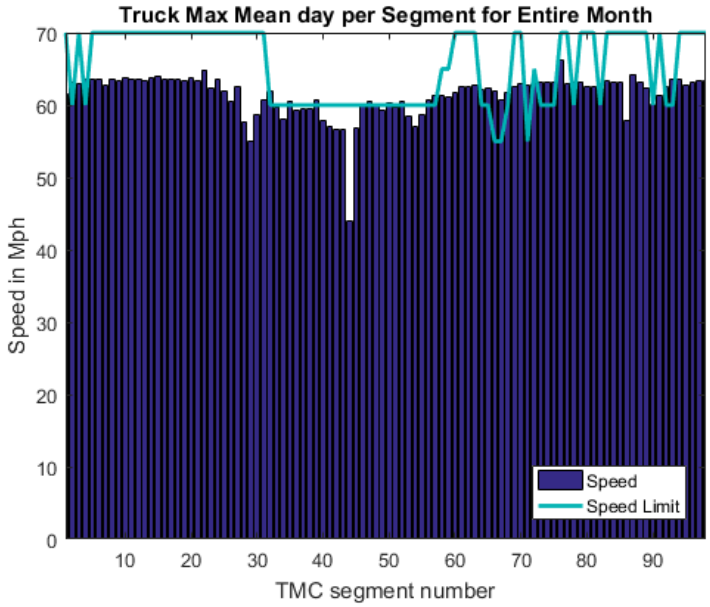


Figure 34 - Max day mean epoch truck speed for Januray 2015.

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.

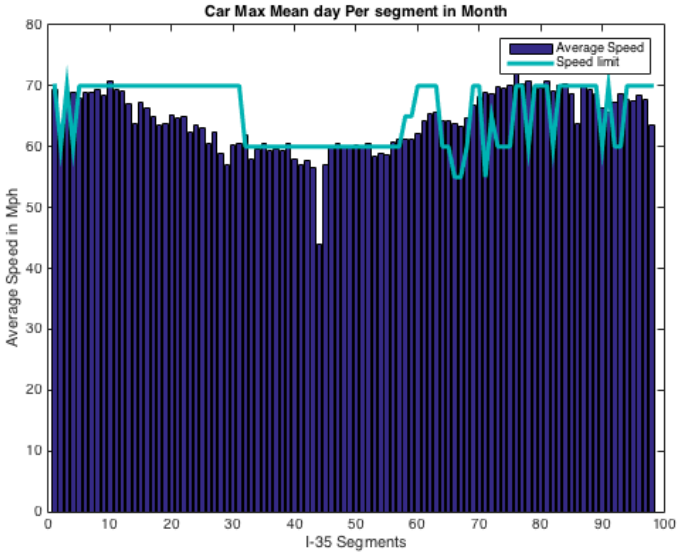


Figure 35 - Max day mean epoch car speed for January 2015.

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 Expressway HWY 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 v.1 dataset.

Figure 37 - View of segment 44 of I-35 intersect with the Centennial Expressway.

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

Either faulty GPS units or insufficient positioning accuracy could 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 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 STD 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 STD (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 result 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.

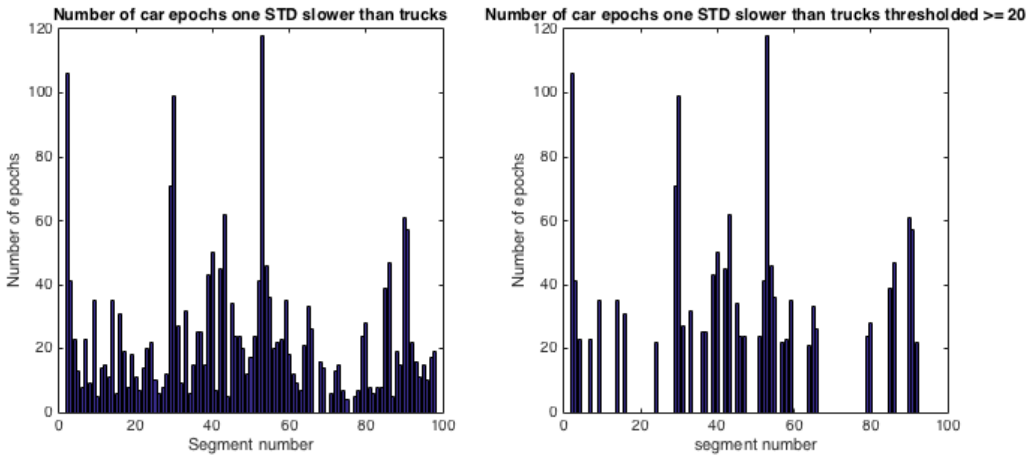


Figure 38 - (a) Cars one STD less than trucks. (b) Threshold result for count >= 20.

Coordinates of a random sample of segments were extracted, and Google Maps was 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 - I-35 S service road adjacent to segment 53.

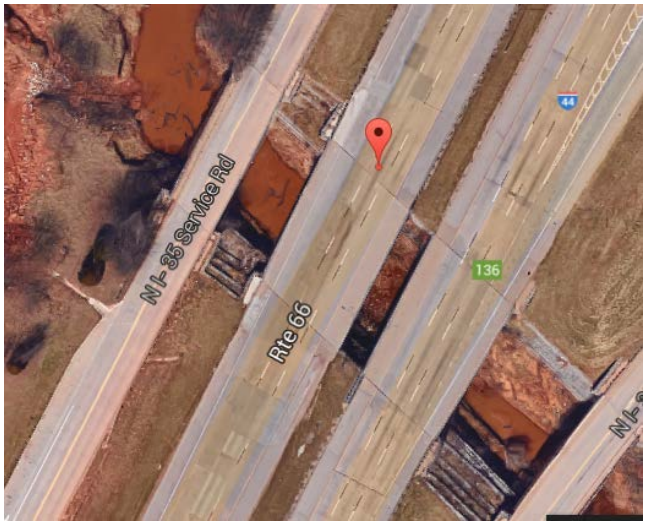
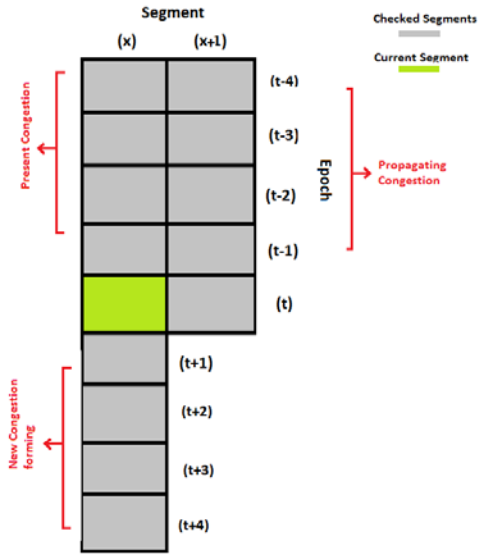


Figure 40 - Segment 30 adjacent to I-35 N service road.

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 detected for this case. 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 also inspected for 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 v.1 dataset, primarily because some missing epochs (i.e., epoch holes) were evident for consecutive records in particular segments in the dataset.



41

Figure 41 - Mask filter to scan for outliers.

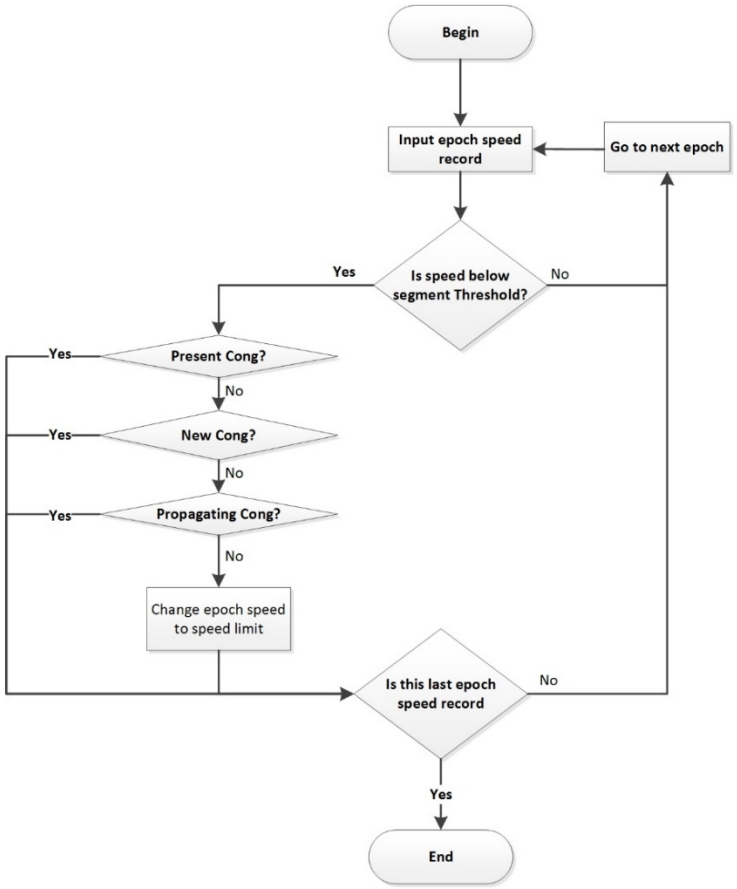


Figure 42 - Flow chart for scanning outliers using mask filter.

3.3. Cleansed dataset

After applying the aforementioned methods and processes, a cleansed dataset was generated. Figure 43 shows an example snapshot from the database 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.

	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 - Database Outlier for Segment 97 in Raw Database

Figure 44 and Figure 45 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.

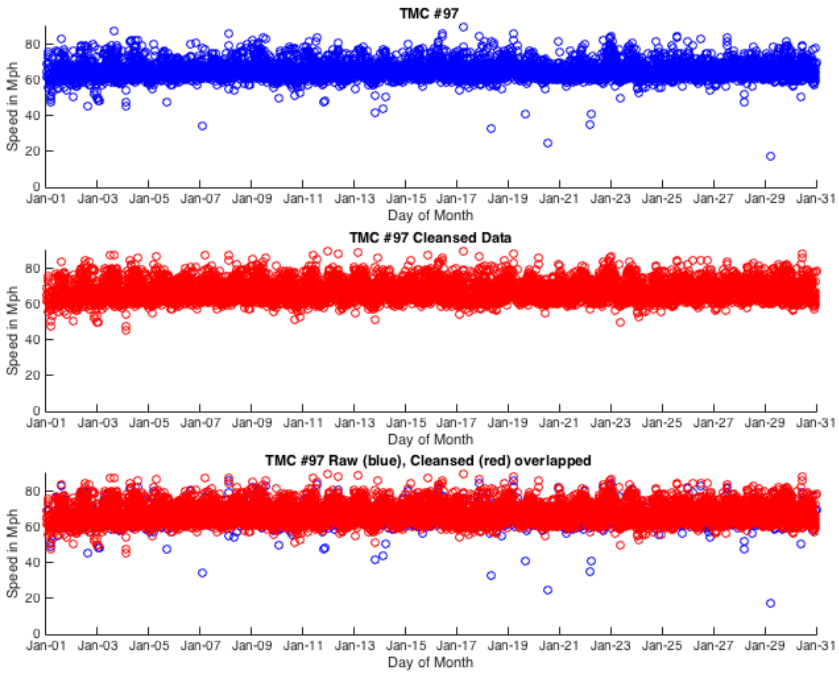


Figure 44 - Comparison for Segment 97 speed, raw vs cleansed, for January 2015.

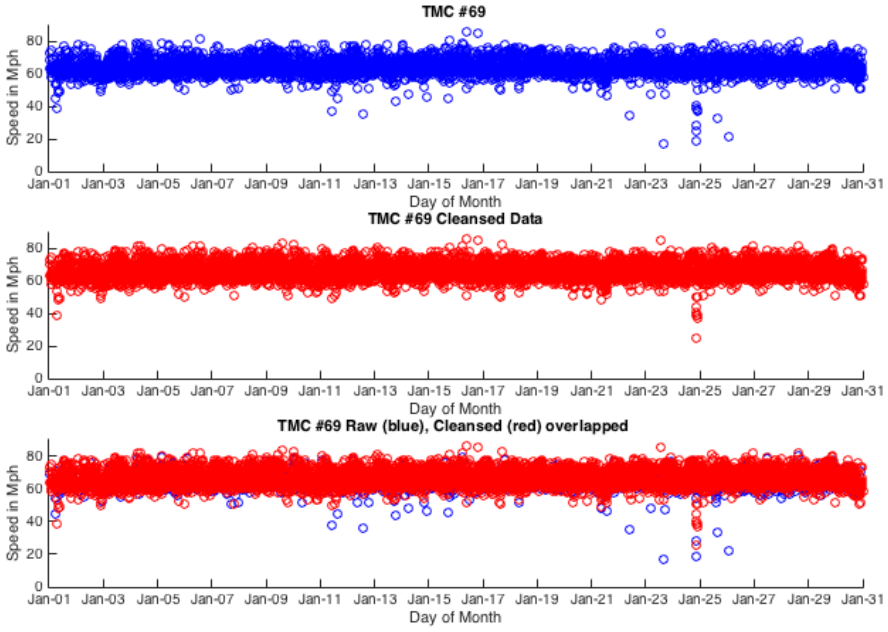


Figure 45 - Comparison for Segment 69 speed, raw vs cleansed, for January 2015.

Chapter 4: Dataset Exploration, Analysis and Congestion Detection

Classical applications of central tendency and variation—specifically means and STD—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 v.1 dataset to identify abnormal travel times on the roadway.

4.1. Statistical mean and variance

Utilizing travel time measurements in NPMRDS v.1, 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 46 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.

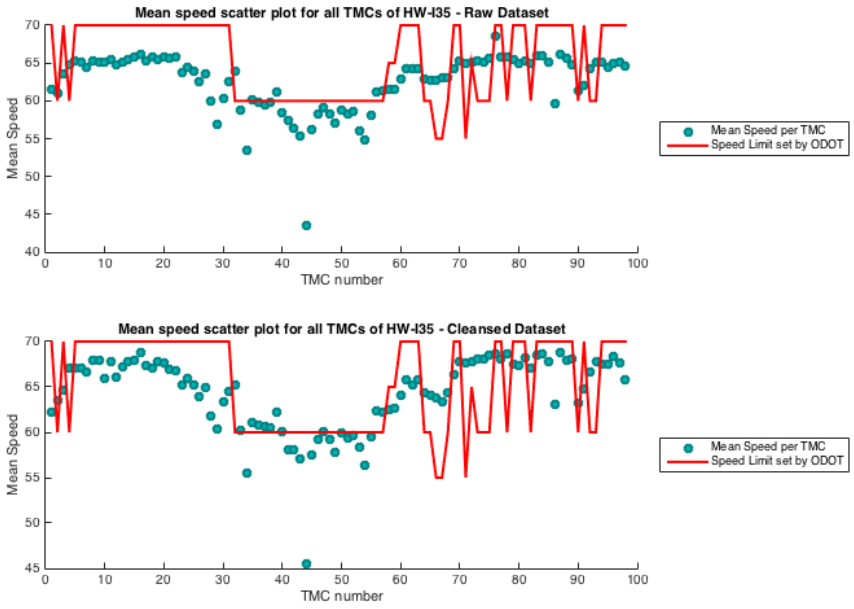


Figure 46 - Mean speed per segment vs. speed limit.

Raw data was utilized to calculate an average monthly speed across all epochs 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 47 shows speed variance per segment for all epochs during the month of January 2015.

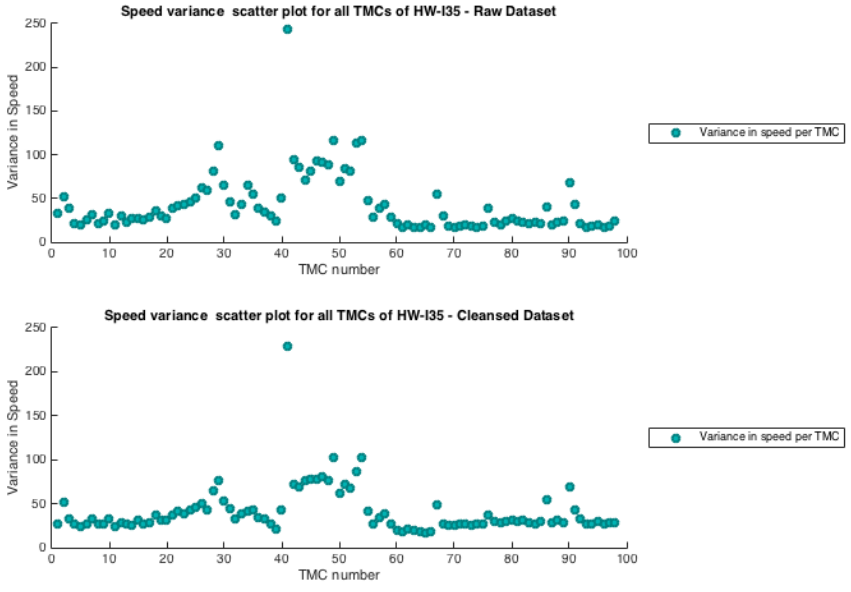


Figure 47 - Speed variance per segment for I-35.

Raw and cleansed graphs demonstrated that TMC stations had increased variance values, which could be indicative of many abnormal, non-free flow traffic cases occurring during the month. 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). [40] 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 v.1 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 48, which shows number of epochs per day for each segment of I-35 southbound for one month. 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.

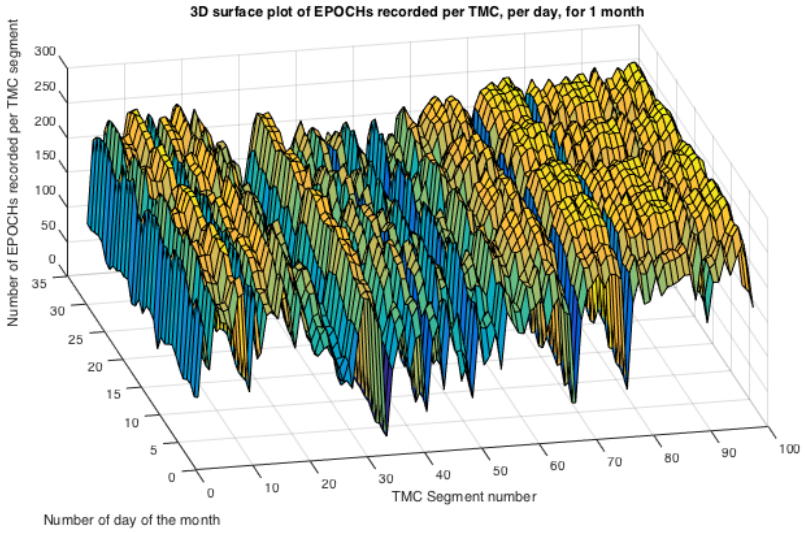


Figure 48 - 3D surface plot of epochs recorded per segment, per day, for January 2015.

Figure 49 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.

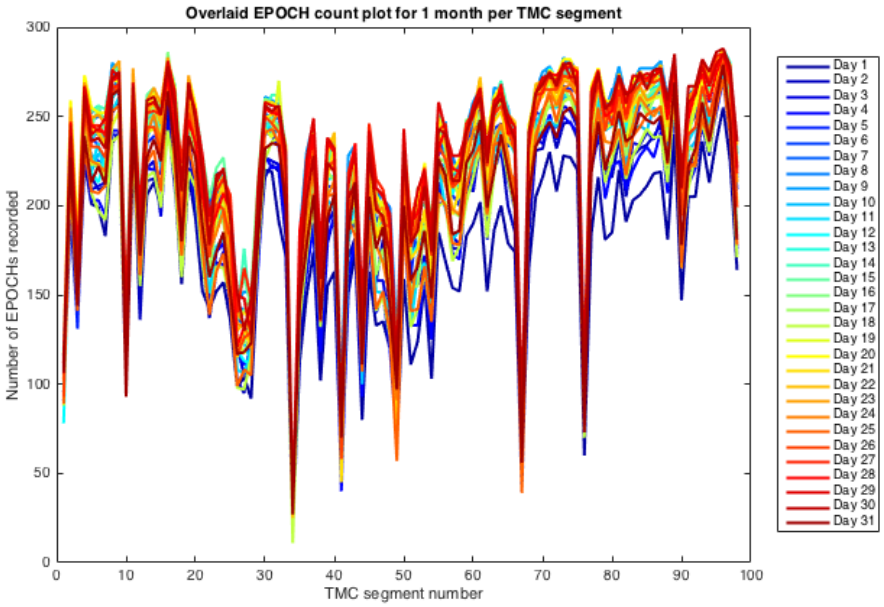


Figure 49 - Overlay epoch daily count for January 2015, per segment.

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 [41],

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

where μ_A, μ_B and σ_A, σ_B are the mean and STD of A and B, respectively. Alternatively, this is also defined in terms of the covariance of A and B [41]:

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

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

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

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 50 shows the mean correlation coefficient results per segment.

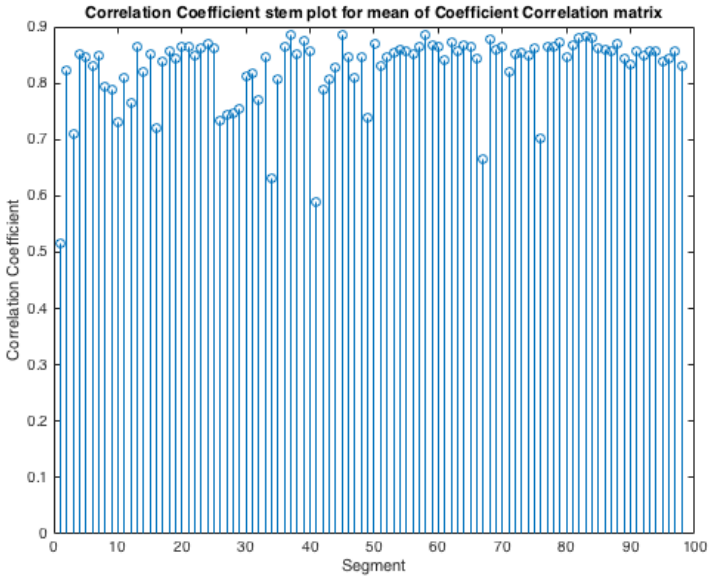


Figure 50 - Mean correlation coefficient per segment stem plot.

A box plot was used to generate the coefficient correlation matrix shown in Figure 51. 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 a large portion of 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).

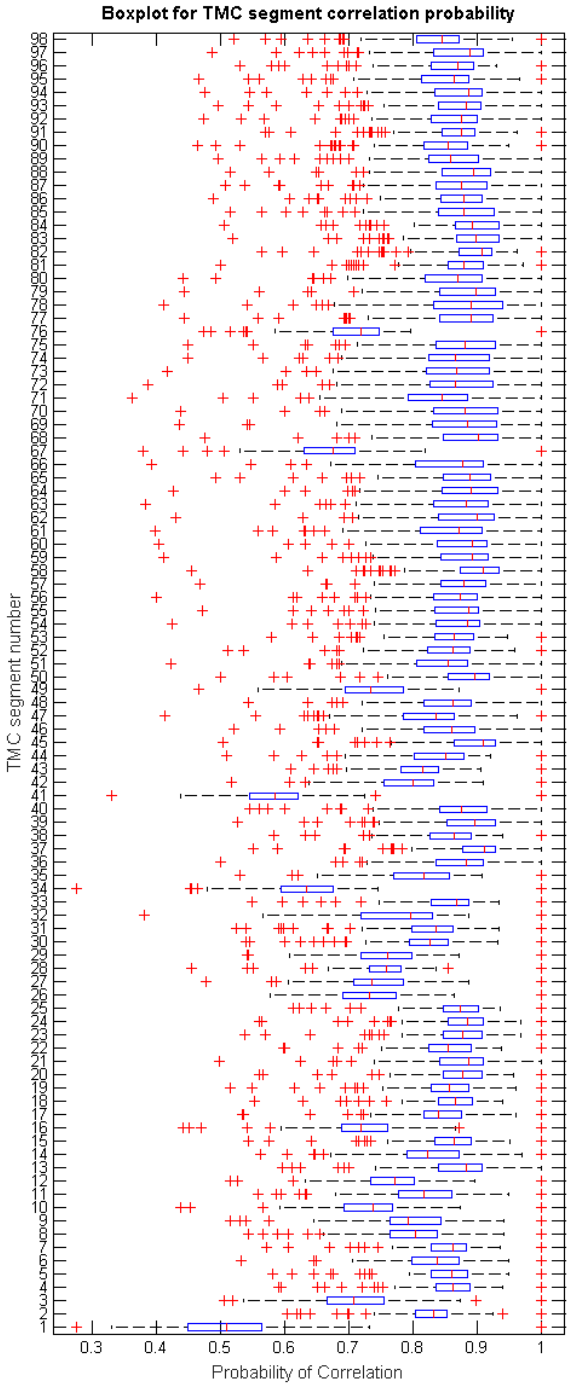


Figure 51 - Boxplot of correlation coefficient matrix.

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 52 depicts the results of normalized weight per segment.

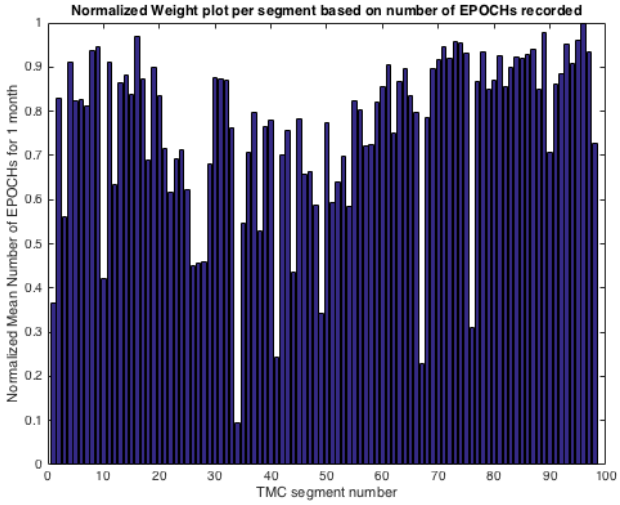


Figure 52 - Normalized epoch count weight plot.

4.3. Congestion detection

Road traffic congestion has a negative environmental impact and causes significant loss to productivity and to the economy. A beneficial use of the NPMRDS v.1 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 [40] [42]. 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 [43], 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, as well as for arterial streets with a posted speed limit of 40 mph to equal 30 mph.

Assuming vehicles commuting under normal traffic conditions travel at free flow with speeds varying slightly above and below the mean, and given abnormal traffic conditions, speeds tend to vary to a greater extent. Determining statistical variance serves as a simple indicator of congestion [40].

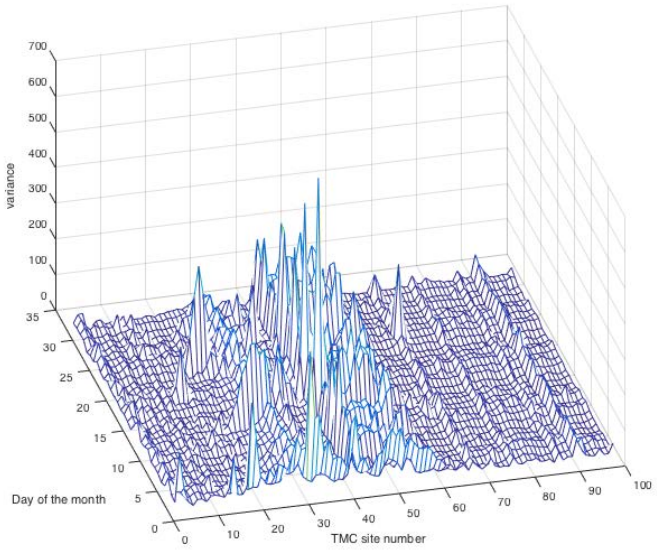


Figure 53 - Mesh plot for speed variance per segment, per day for I-35, January 2015.

Figure 53 illustrates a mesh plot of speed variance per day per segment on I-35 southbound for January 2015. Figure 54 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.

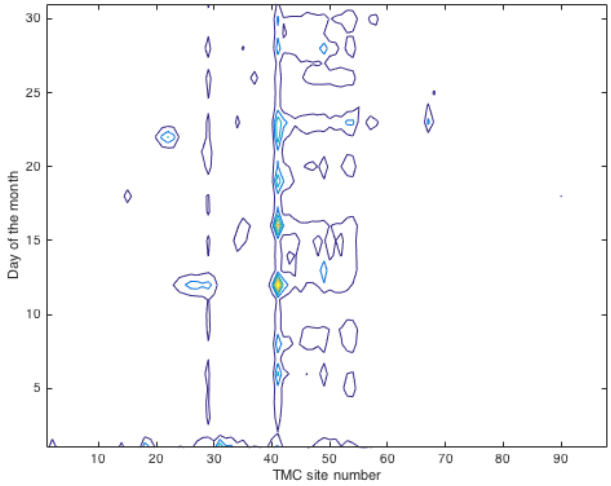


Figure 54 - Contour plot of speed variance per segment, per day, for I-35 January 2015.

Extracting the high variance segments and combining with the previously derived weights, a histogram plot shown in Figure 55 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.

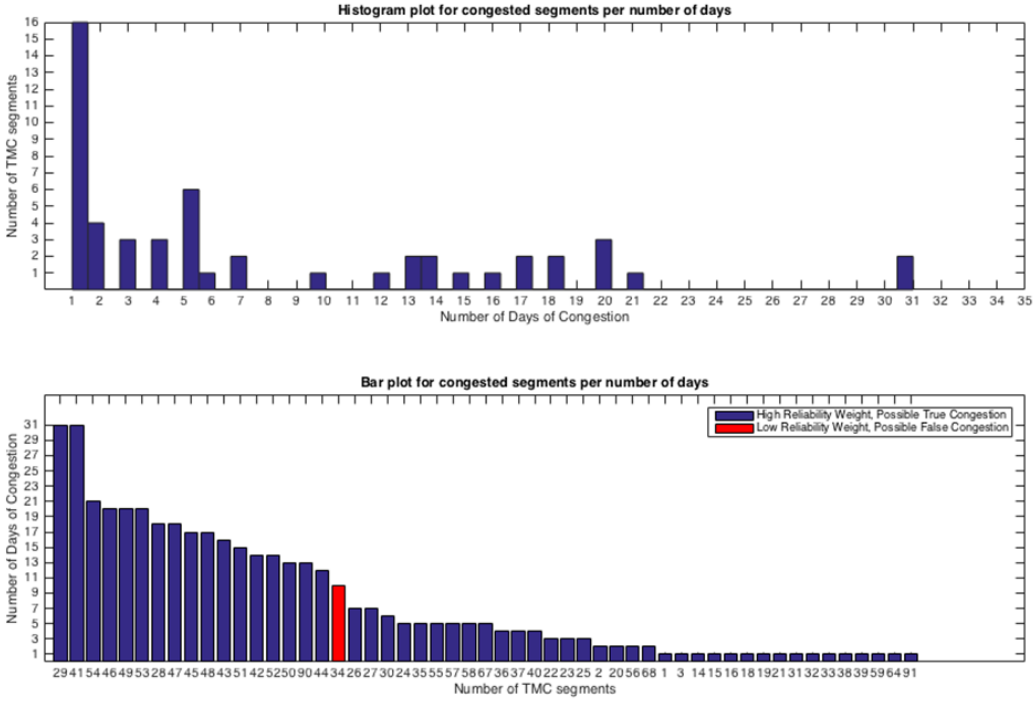


Figure 55 - Histogram and decreasingly sorted bar plots of congested segments on I-35.

It is noted that accuracy could be jeopardized when detecting congestion based on statistical variance. This drawback stems from reliance on false assumptions. First, congestion does not occur at all times; when it does occur for extended periods of time—equal to duration of analysis, variance measured does not accurately indicate congestion. Second, 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 56, 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.

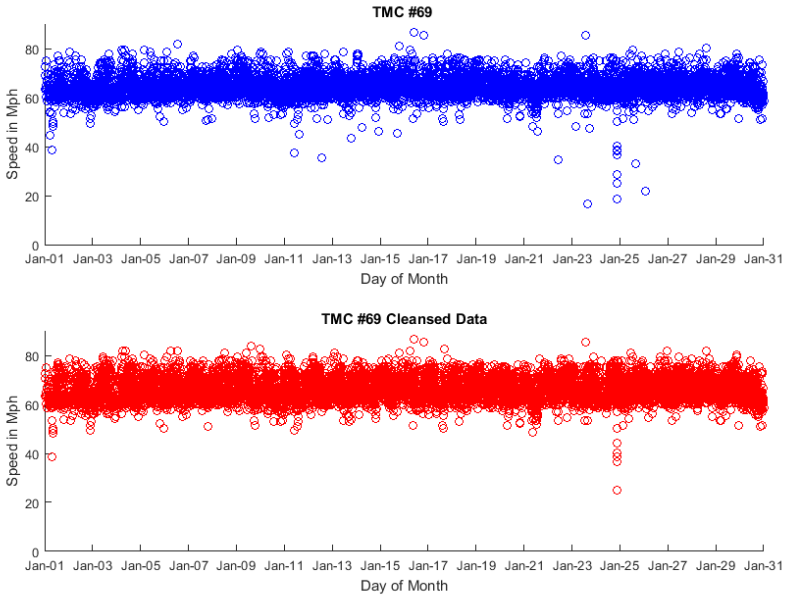


Figure 56 – Segment 69 congestion not detected using a standard variance test.

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 [44] [45].

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 [46]. Figure 57 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 58 shows three examples of random segments collected on non-congested days and fitted to a normal distribution.

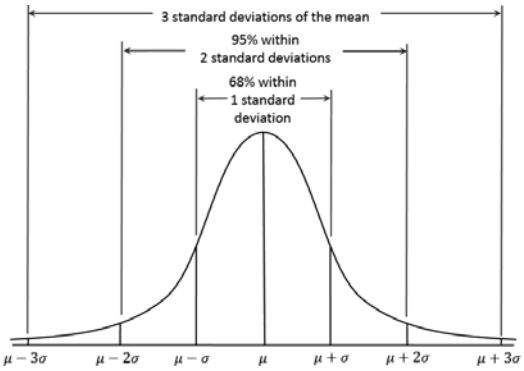
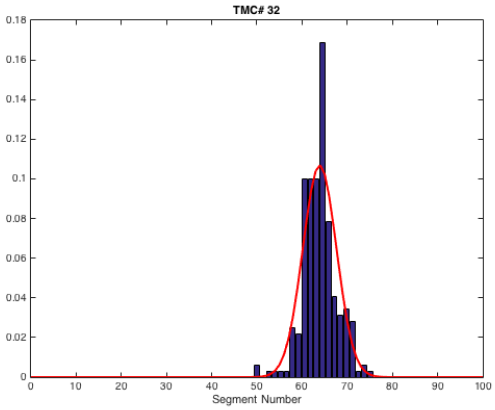
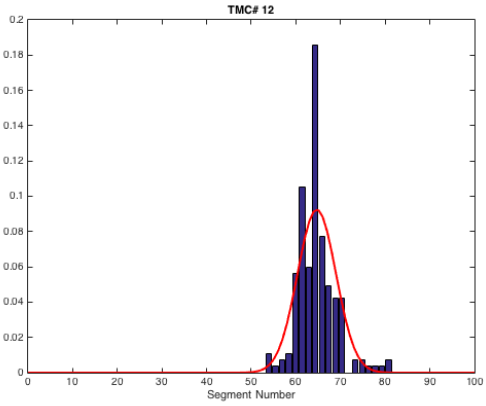


Figure 57 - Normal Gaussian distribution model.

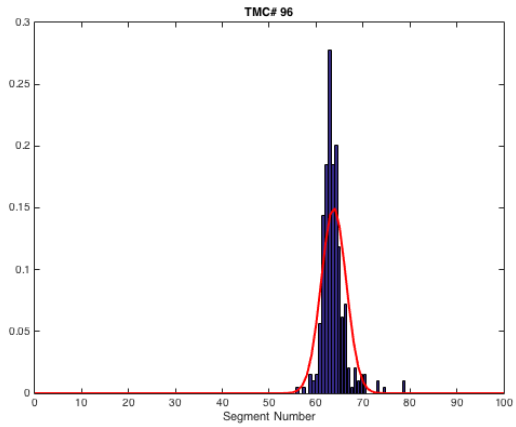
A decision threshold can be established by defining a specific threshold for each segment based on its free flow model at a chosen number below STD from the non-congested mean. Doing so aptly indicates congestion in each segment. The threshold chosen in this work was three STDs from the non-congested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



(a)



(b)



(c)

Figure 58 - Three random segments depicting free flow Gaussian modeled 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 62. Thus, a filter was applied for cases of five or fewer congested epochs during an entire day. Figure 59 shows the results for all segments per day on I-35 southbound during January 2015. Figure 60 and Figure 61 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 63 illustrates a comparison of variance and threshold test results on segment 69 for detecting congestion.

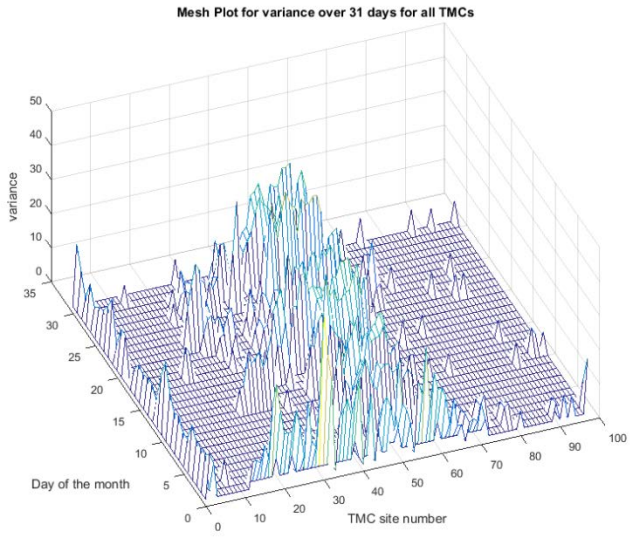


Figure 59 - Mesh plot for thresholded speed variance, per day for I-35 S, January 2015.

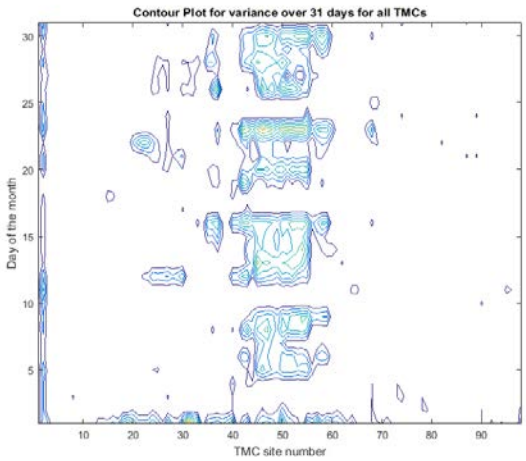


Figure 60 - Contour plot for thresholded speed variance, per day for I-35 S.

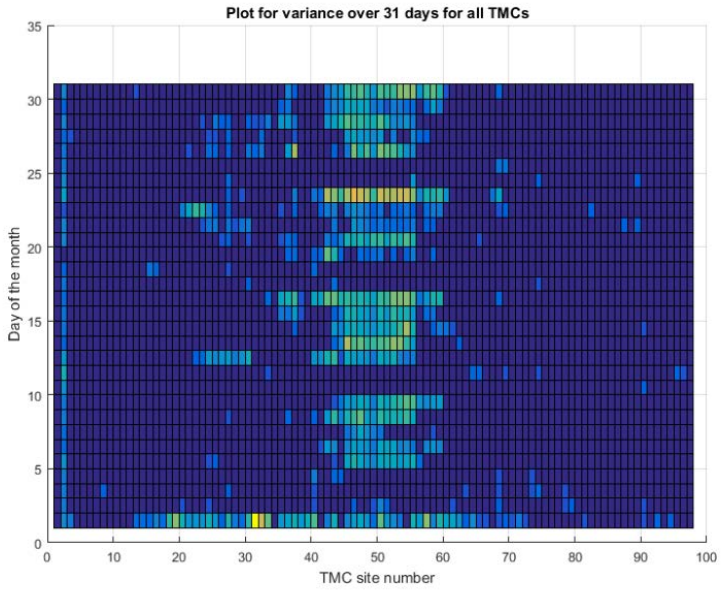


Figure 61 – Heat map for speed variance per segment, per day for I-35 S, January 2015.

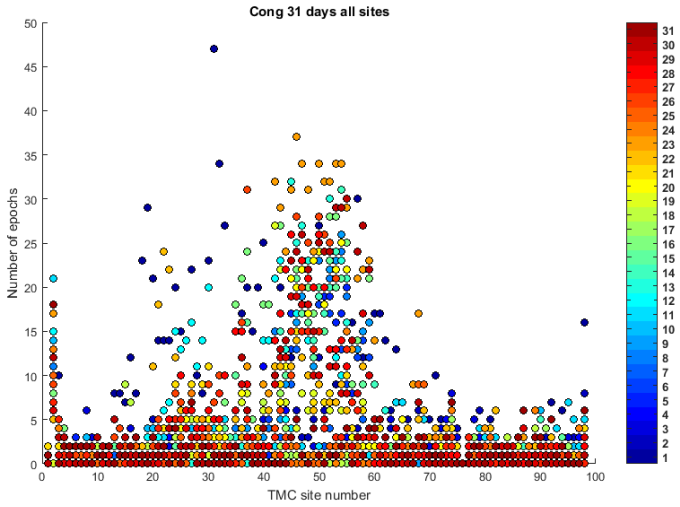


Figure 62 – Congested epoch count for January 2015 on I-35 S.

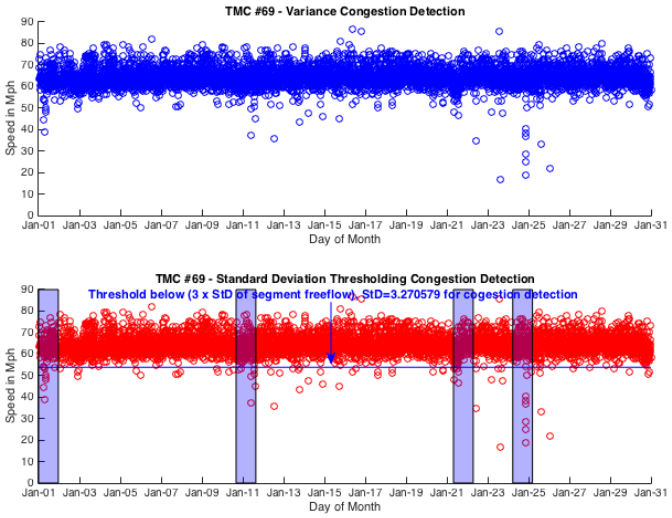


Figure 63 –Variance and threshold congestion detection comparison on Segment 69.

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 9 offers a numeric comparison between results for raw and cleansed datasets. Figure 64 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 65, Figure 66 and Figure 67 demonstrate outliers in the raw dataset caused false detection.

Table 9 - 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

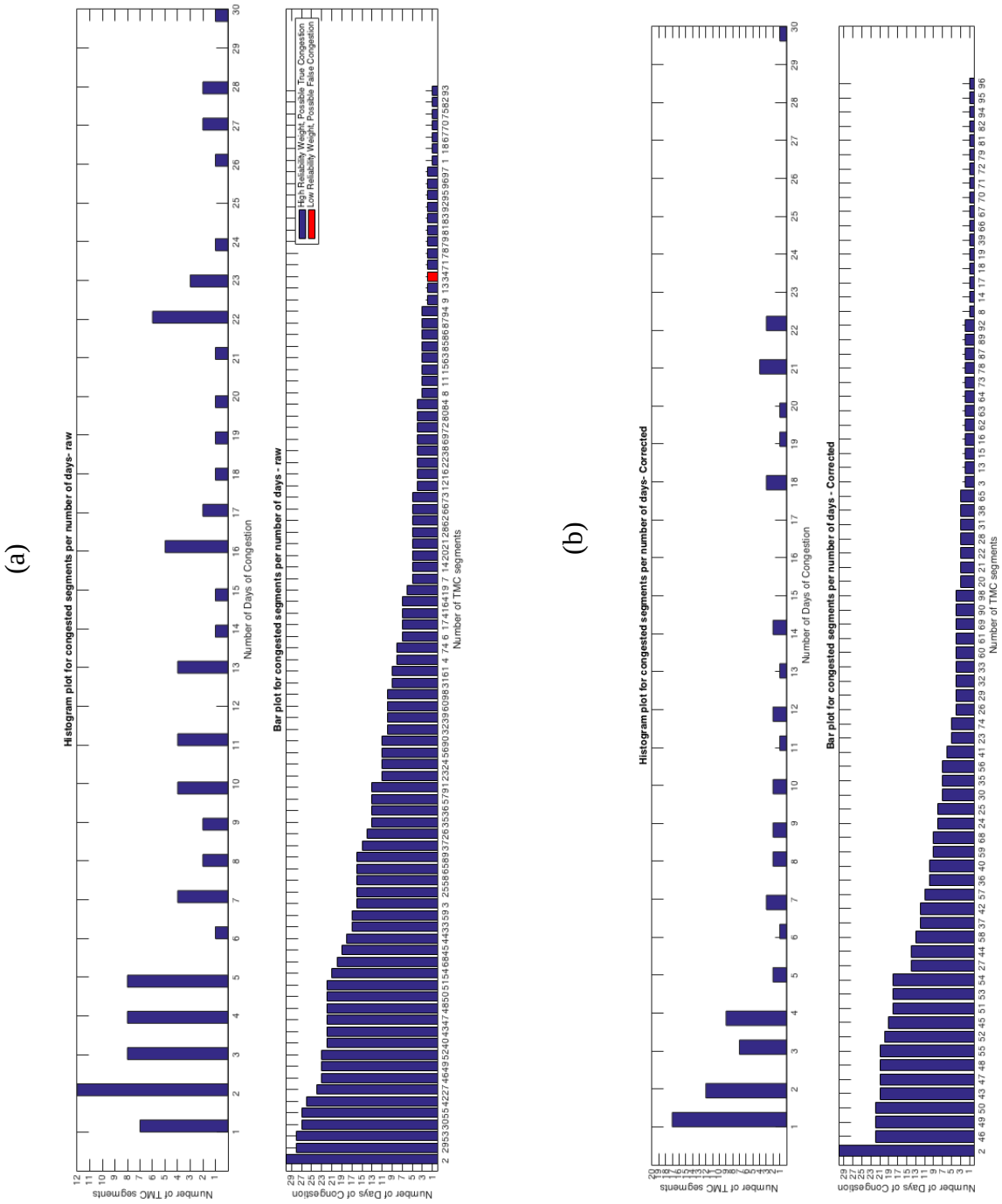


Figure 64 - Modified congestion detection results for raw (a) and cleansed dataset (b).

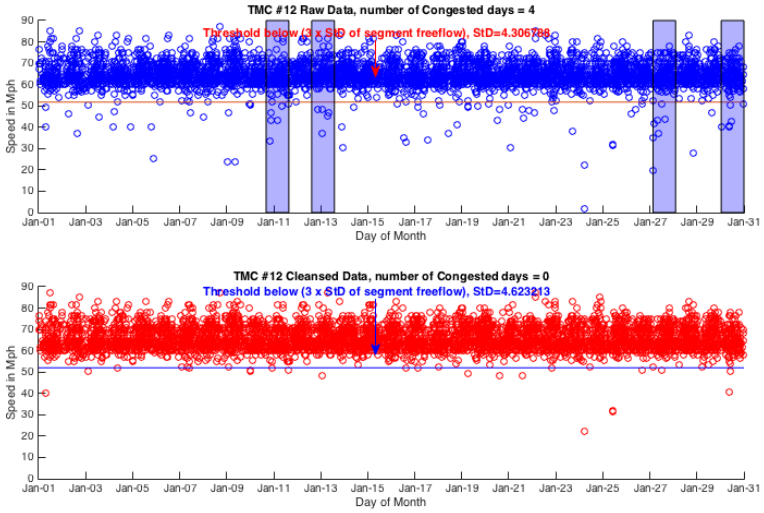


Figure 65 - Segment 12 congestion detection comparison for raw and cleansed datasets.

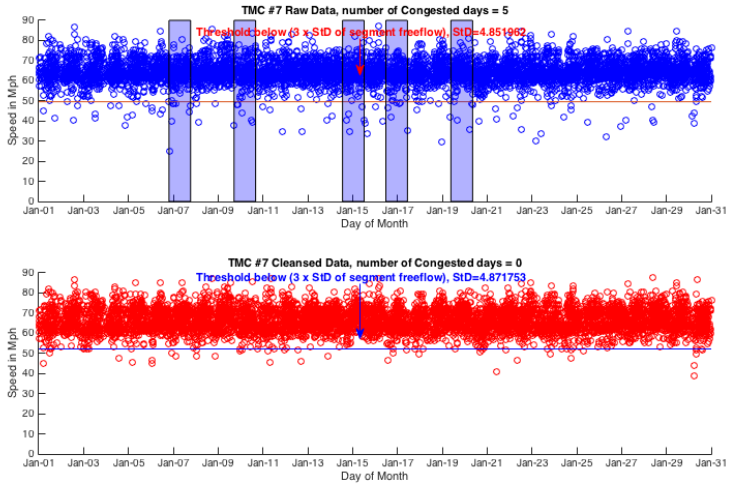


Figure 66 - Segment 7 congestion detection comparison for raw and cleansed datasets.

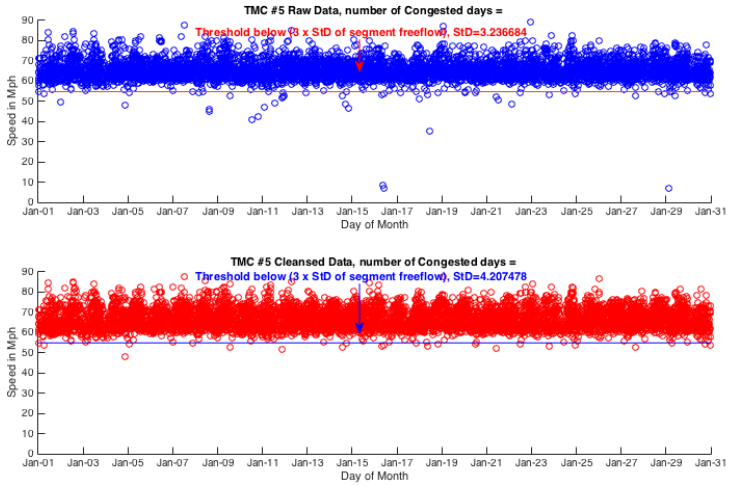


Figure 67 - Segment 6 congestion comparison for raw and cleansed datasets.

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 68, Figure 69, and Figure 70.

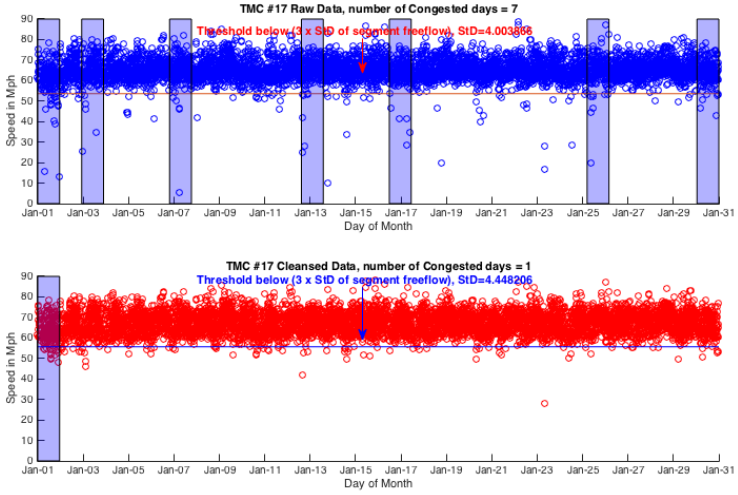


Figure 68 - Segment 17 congestion comparison for raw and cleansed datasets.

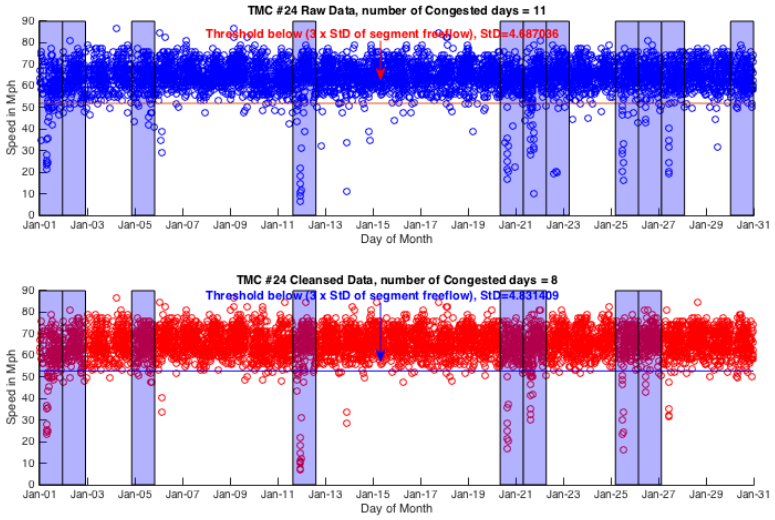


Figure 69 - Segment 24 congestion comparison for raw and cleansed datasets.

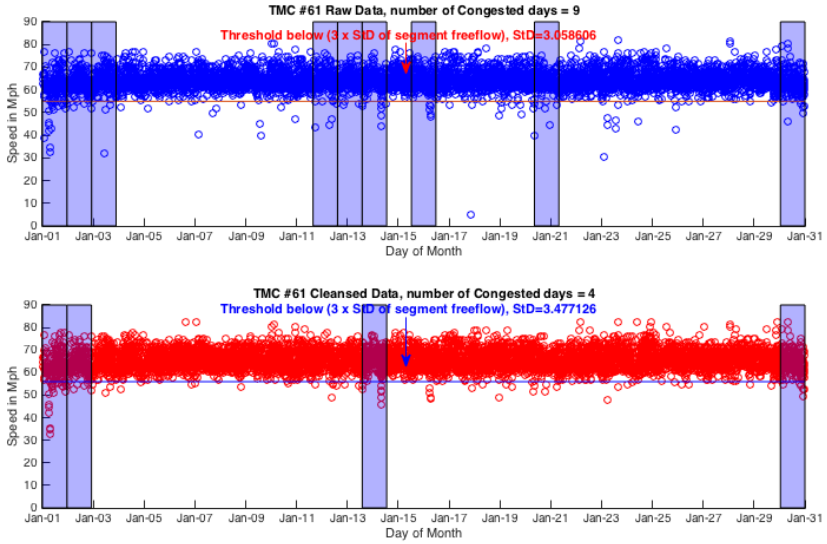


Figure 70 - Segment 61 congestion comparison for raw and cleansed datasets.

It is obvious that outliers were cause for false detection.

3) The third group 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 71 and Figure 72. 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.

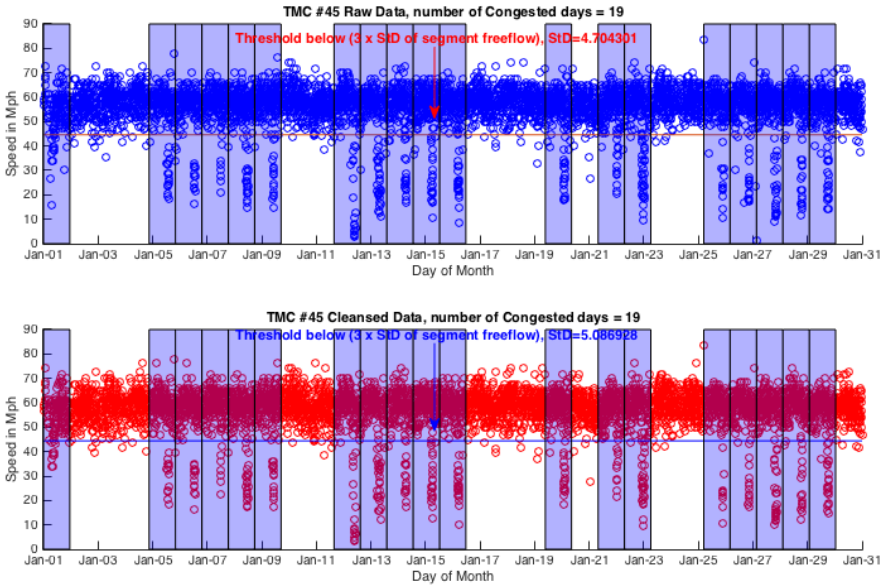


Figure 71 - Segment 45 congestion detection comparison for raw and cleansed datasets.

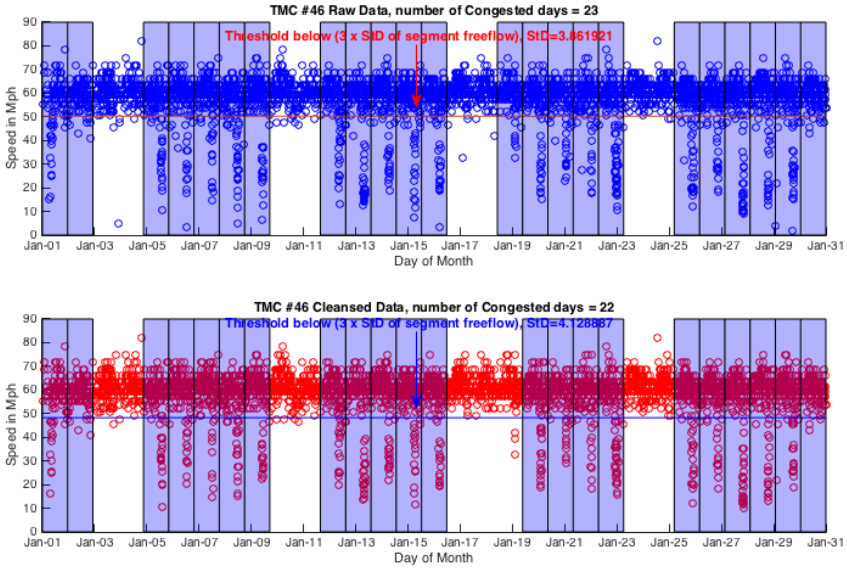


Figure 72 - Segment 46 congestion detection comparison for raw and cleansed datasets.

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 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 that 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 [47]. 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, STD, variance, and mean measurements were recorded. Gaussian model fitting was performed.

Table 10 shows the segment-length weighted-average free-flow speed, variance, and STD 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.

Table 10 – Free-flow Speed Statistical Measures for I-35 S

Measure	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 v.1. Figure 73 shows the difference of raw and cleansed mean free-flow speed for all segments on I-35.

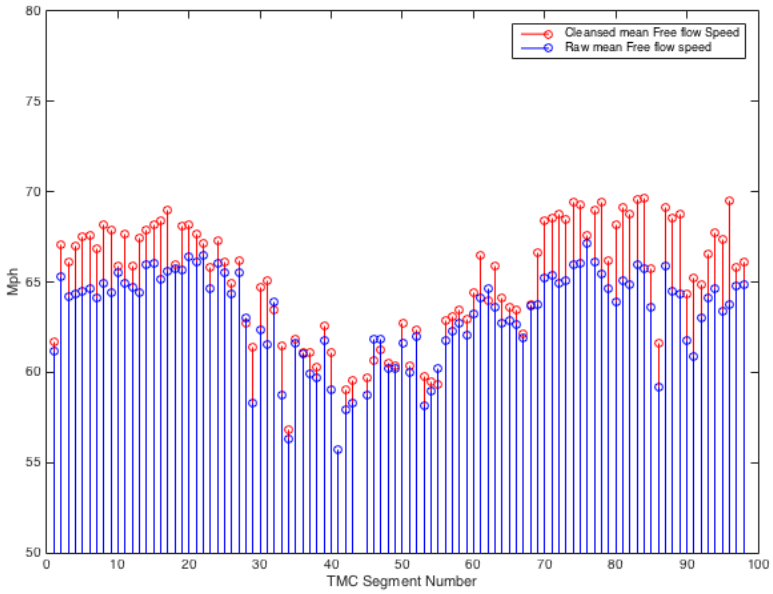


Figure 73 – Mean free-flow speeds for all I-35 segments.

Mean travel time per segment was derived utilizing segment length obtained from the NPMRDS v.1 static file. 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 segment length in Figure 74.

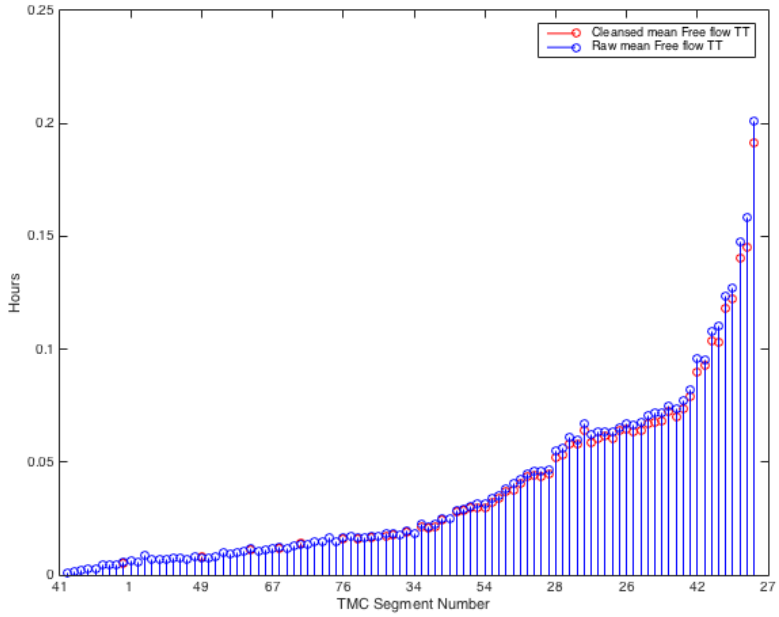


Figure 74 – Free-flow travel time for I-35 S segments.

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 [48], [49] 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."* [49]. 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 [50]. Figure 75 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. [50] The lowest collision rate conforms to the smallest variation from the average.

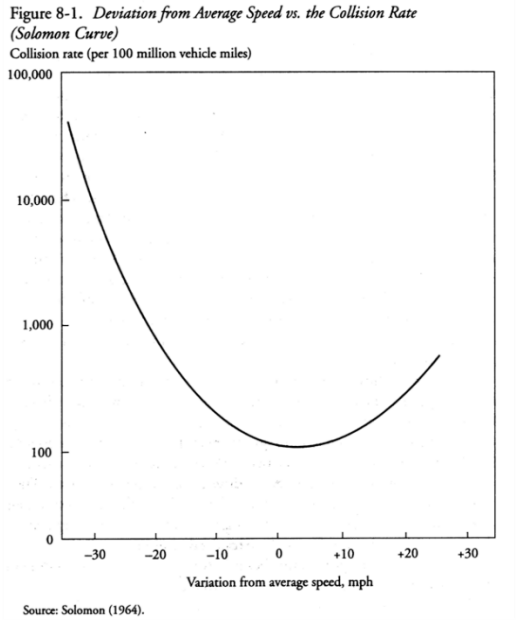


Figure 75 - Solomon Curve [50].

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 [46]. 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 [46]. 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 76. Figure 77 shows segment 73 when using the cleansed dataset. 85th percentile speed was 72.7mph.

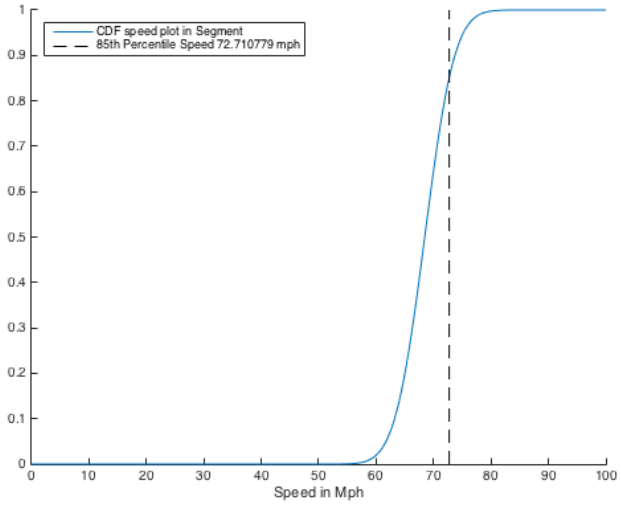


Figure 76 – Segment 73 CDF with 85th percentile speed (cleansed dataset).

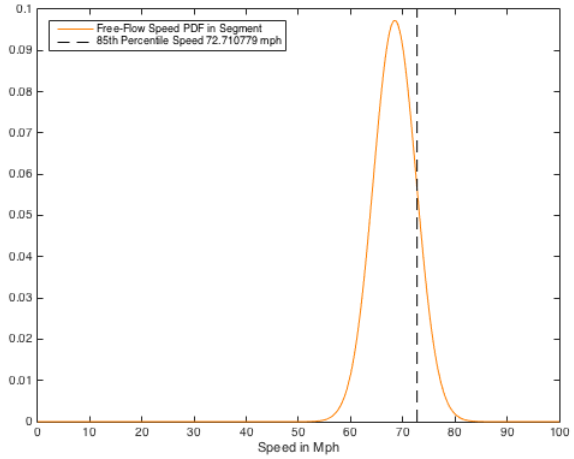


Figure 77 – Segment 73 PDF with 85th percentile speed (cleansed dataset).

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. Figure 78 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.

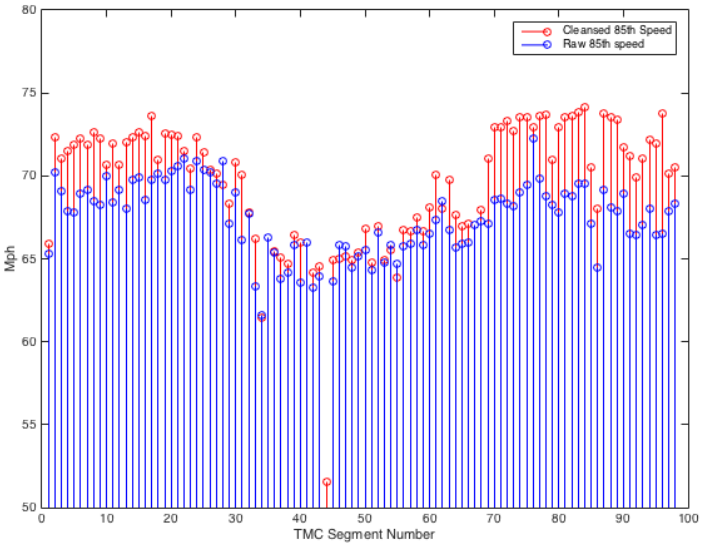


Figure 78 - I-35 85th percentile per segment.

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 [51].

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

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.

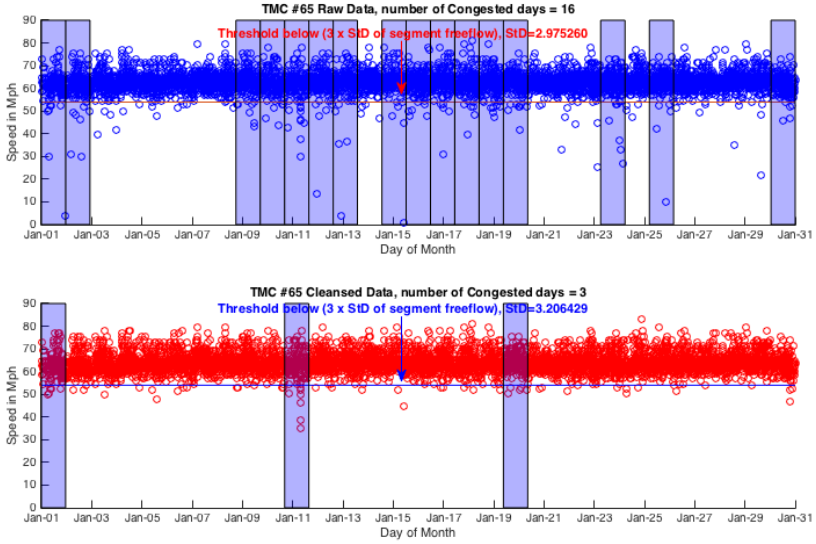


Figure 79 – Segment 65 comparison between cleansed and raw datasets.

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 80 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 11 details a comparison of both datasets. Figure 81 illustrates TTI per segment for I-35 southbound for both datasets.

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

Time	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

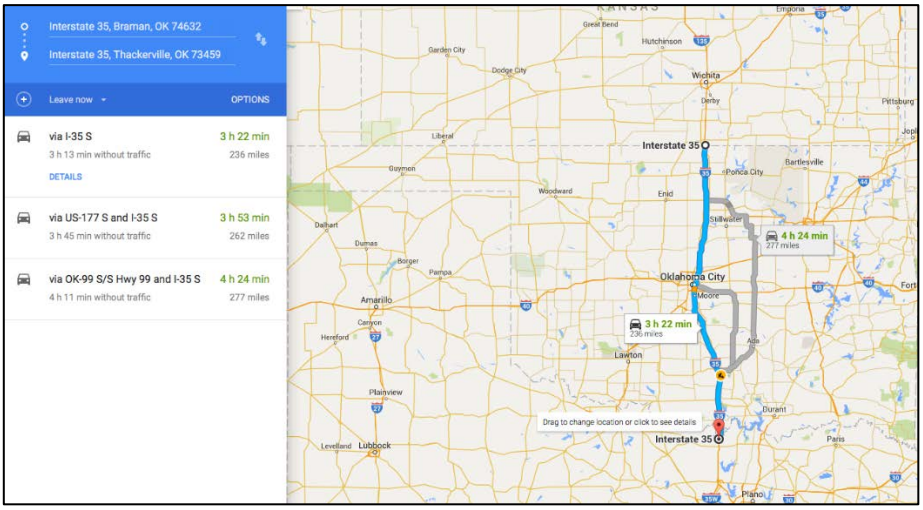


Figure 80 - Google Maps route results for I-35 S, January 12, 2016.

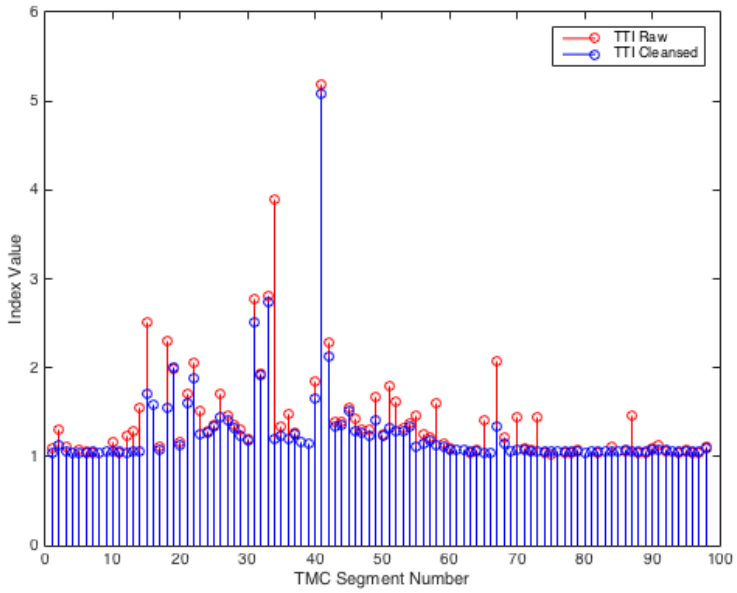


Figure 81 - Segment TTI comparison for raw and cleansed datasets.

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 [52]; the result 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 [52].

Figure 82 illustrates results per segment for I-35 southbound for both raw and cleansed datasets. Appendix C shows numerical results per segment per dataset.

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

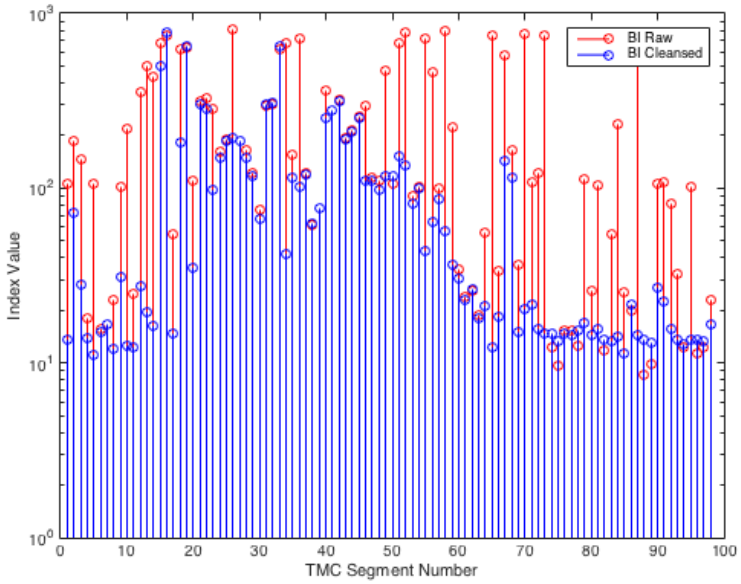


Figure 82 - BI for all segments I-35 raw and cleansed dataset.

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 [52]. Figure 83 illustrates results per segment for I-35 southbound for both raw and cleansed datasets. Appendix C shows the numerical results per segment per dataset.

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

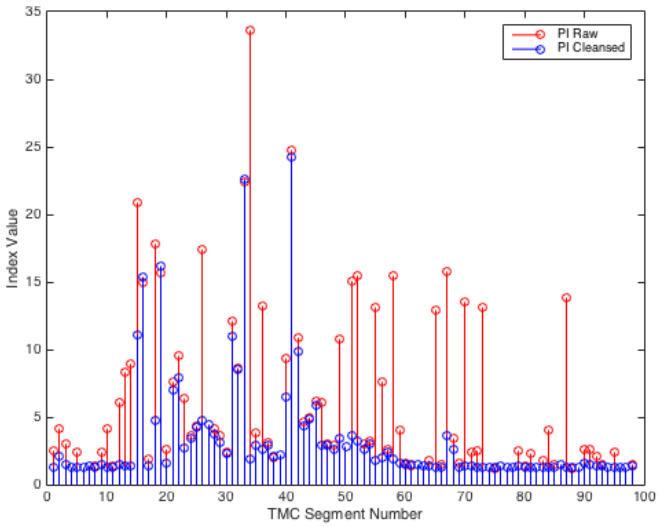


Figure 83 - PI for all I-35 segments, raw and cleansed datasets.

BI and PI statistics are significantly affected by outliers. Figure 83 shows a substantial difference between datasets for many segments. Figure 84 illustrates congestion comparison between datasets for segment 65. For the 15th of January 2015, 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 85 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 86).

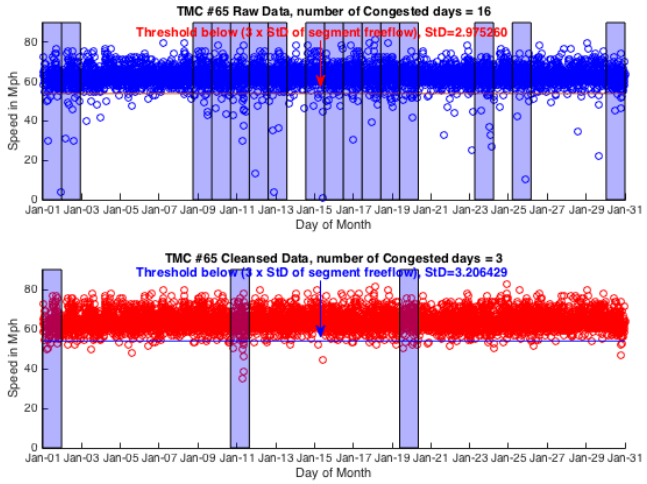


Figure 84 - Segment 65 congestion comparison between raw and cleansed datasets.

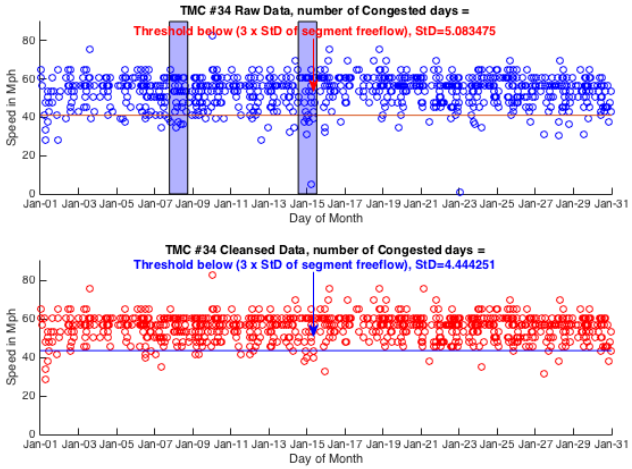
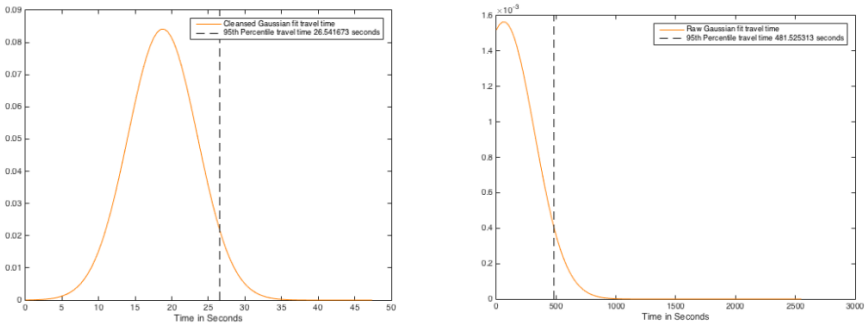


Figure 85 - TMC 34 January 2015 speed scatter plot.



(a) (b)

Figure 86 - 95th percentile travel time for (a) cleansed and (b) raw dataset.

Chapter 6: NPMRDS v.1 Cleansing and Validation Study

Building on the insights obtained from the study detailed in the previous chapters. This chapter presents the work developed to cleanse the entire NPMRDS v.1 dataset per category of vehicles (i.e., cleansed freight vehicles, cleansed passenger vehicles). The work adapts the previous study findings of outlier and anomaly methods and enhances them to formulate an outlier algorithm for each vehicle category, then process the entire NPRMDS v.1 raw dataset. Furthermore, the chapter presents a validation study to statistically analyze the effects of outlier removal on the raw dataset. In the end, the outcome is a cleansed dataset analogous to the raw one that can be used for analysis and performance measurement computation.

6.1. NPMRDS Dataset Cleansing

Figure 87 illustrates the outlier removal algorithm flow chart depicting the steps taken to process data outliers in freight vehicles. A replica process was conducted to extract outliers of passenger vehicle data with a slight change in the subroutine, as explained below. Outliers generated by the following factors were addressed in this process:

1. High spatial-temporal probe and record data variability outliers.
2. Vehicle-specific performance outliers affecting freight truck speed on roadways.
3. Roadway geometry outliers.
4. GPS coordinate in-accuracy outliers.

A detailed description of each type of these outliers was previously presented in Chapters 2 and 3. As can be seen in Figure 87, the data files obtained from HERE are first transformed into three separate datasets: two pertaining to each separate category of vehicles (e.g., freight and passenger vehicles), and one as an average of both datasets termed the “all-vehicles” dataset. Outlier processing begins by removing extreme values over 95mph (this was a slight 5-mph increase above the threshold defined in Chapter 3). Once extreme value outliers were removed, the freight dataset was split into two new datasets according to the directionality of the segments (e.g., northbound–southbound or eastbound–westbound). Once the split was completed, segments were ordered according to directionality (e.g., increasing mile marker [positive] and decreasing mile marker [negative]). This ordering stage was crucial for outlier removal, as it allowed implementing the scanning outlier procedure previously discussed in Chapter 3. After the completion of the ordering stage, a subroutine is triggered that extracts each highway in every directional dataset and performs a revised outlier mask-scanning process, as explained below.

The outlier mask-scanning process, depicted in Figure 87 as a subroutine colored in gray, scans each epoch of each segment in the highway. If the epoch is deemed to be

congested (recall from Chapter 3 that congestion is defined to be a speed below 3 STDs from the mean free-flow speed of the segment), then a check process commences. The check process scans present congestion (as described by the mask shown in Figure 41 in Chapter 3), future congestion, and propagating congestion. If any of the checked epochs report congestion, then the congested epoch is left as is.

If, however, all of the checked epochs are not deemed to be congested, then a final check is done on the other category dataset (i.e., in the freight dataset case, that being the passenger vehicle dataset). The investigated epoch of both datasets is compared, and if the reported passenger speed epoch was found to be 15 mph or more above the freight dataset epoch, then the freight epoch entry is removed. Processing continues until all epochs of the dataset are processed.

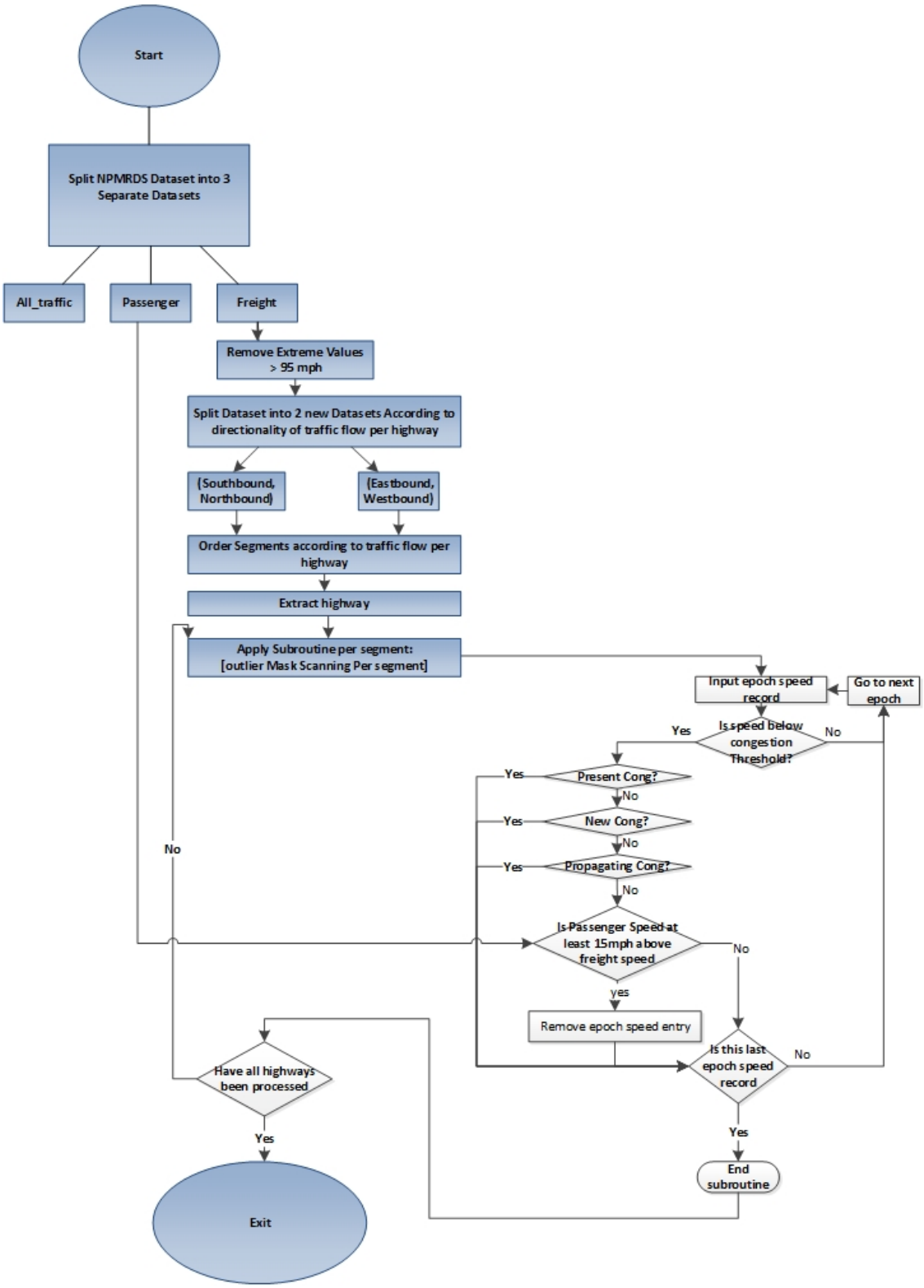


Figure 87 – Outlier removal flowchart for freight truck vehicles.

6.2. Sequencing Process

The sequencing process studied below was examined for Oklahoma NHS segments. The process output of sequenced segments was checked manually for 2015. As aforementioned, a “static file” and a “shapefile” provided by HERE contain all segment information for the NHS roadways in the NPMRDS v.1. The developed sequencing algorithm relies on these files to order and sequence segments of highways. Figure 88 depicts the high-level flow chart of the steps performed in the algorithm. First a matrix of distances between all segments, depending on the coordinate locations, is calculated. Then, after finding start and end segments of each highway, adjacent segments are picked up one after the other until the end segment of the highway is reached.

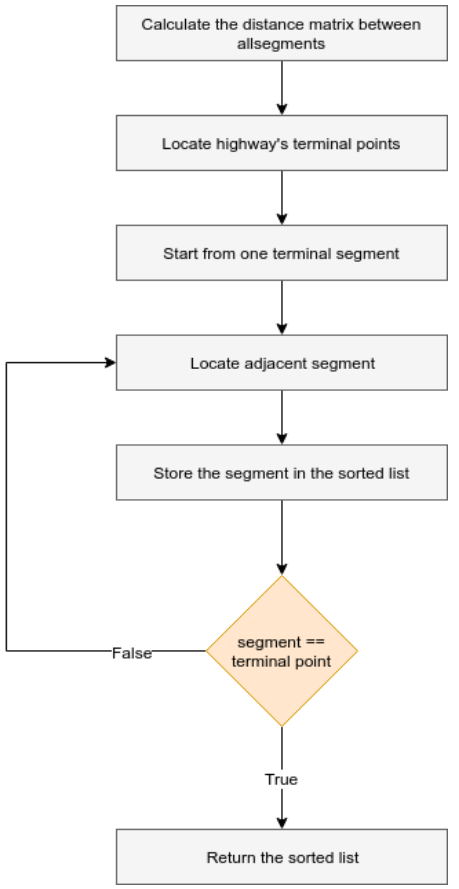


Figure 88 - Flowchart of sequencing algorithm.

Calculating the distance matrix:

To get accurate estimates of distance based on GPS coordinates, the "Great Circle Distance" method was used, where distance is calculated between two points on a sphere. This method gives more accurate estimates of distance than Euclidean distance on planer coordinate systems. We employed the data available in the "Shapefile" (.shp) using a geographic information system. In the '.shp' file, each segment is represented by a link. Using QGIS we converted each link into a set of points, and then represented each point by its GPS coordinates. Resulting points then are merged with the original static file to generate an extended shape file. For example, the resolution of I-35 increased from 198 points into 7482 points.

The great-circle distance method is represented in Equations 5 and 6.

Let $\phi_1, \lambda_1 \; and \; \phi_2 \; \lambda_2$ be the geographical latitude and longitude in radians of two points 1 and 2, and $\Delta\phi, \Delta\lambda$ be their absolute differences; then $\Delta\,\sigma$ the central angle between them, is given by the spherical law of cosines:

$$\Delta\,\sigma = arccos(sin\phi_1 \cdot sin\phi_2 + cos\phi_1 \cdot cos\phi_2 \cdot \cos(\Delta\lambda)) \quad (5)$$

The distance d (i.e., arc length) for a sphere of radius r and $\Delta\,\sigma$:

$$d = r \cdot \Delta\,\sigma \quad (6)$$

Once distance between all segments is calculated, the result is stored in a 2-D matrix called the distance matrix.

Locate highway's terminal points:

A simple algorithm was developed for Oklahoma segments to return the terminal points based on distance matrix algorithm along with Google Maps services.

Algorithm steps:

1. Calculate the distance matrix between all points.
2. Locate two points with the largest distance.
3. Consider them as terminal points of the road.
4. Starting from one terminal point, find all adjacent points.
5. Starting from the other terminal point, find all adjacent points.
6. Compare the two lists.
7. If they are identical, then the founded sequence list is correct.
8. If they are different then:
    1. Apply Google Map distance service to calculate the distance between rounded terminal points among the two lists originating in Steps 4 and 5.
    2. Compare distance results from Google Map.
    3. Consider the list with the largest distance value as the correctly ordered list.

The following case scenario, illustrated in Figure 89, further explains the solution. By applying the first three steps from the algorithm, we find that the terminal points—

according to the distance measurements between all points—are points 5 and 2. By proceeding with the algorithm in Steps 4 and 5, we order all points starting from the first terminal point, resulting in the first ordered list. We then perform the same steps starting from the second terminal point, resulting in the second ordered list.

2	1	3	4	5

The first ordered list

5	4	3	2	1

The second ordered list

1	2	3	4	5

The second ordered list (reversed)

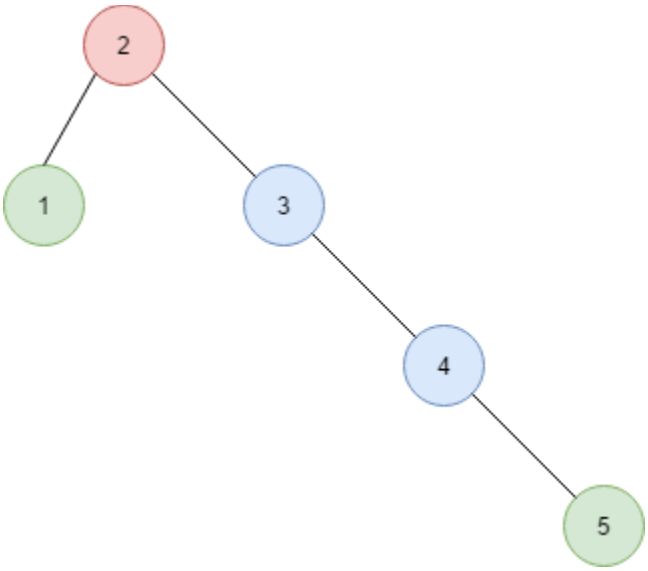


Figure 89 - Sequencing example.

Reversing the second list and comparing with the first list, we find a difference between the two; accordingly, an error alert is raised. The Google Maps step is triggered to resolve the problem and determine the correct sequence. By applying Google Maps service to calculate the distance on the road between points (1, 5) and points (2, 5), we can easily see the distance between points (1, 5) is bigger than the distance between points (2, 5), which leads us to consider points (1 and 5) as the correct terminal points and return the second list as the correct order.

The aforementioned developed method could cause errors in cases where the longest distance between two segments does not represent the start and end nodes of the road, also in cases where both lists returned in Steps 4 and 5 did not have the correct start and end segments or ordering of the highway. This is also true for cases where changes in

highway topology cause the closest distance segment according to our distance metric (which is based on coordinates) to not be the next contiguous segment on the highway path. As such, manual sequencing was done for segments of the 2015 NPMRDS v.1 data, and the results of the developed algorithm were cross compared with the manual results of that year. It was found that the accuracy was 100% for Oklahoma Interstate highways and Oklahoma US highways; and was 96.36% for Oklahoma Intrastate highways.

6.3. Cleansed Dataset Validation

The validation study was performed on the freight cleansed dataset. Prior to statistical analysis, data was preprocessed and re-arranging, as shown in Figure 90. The following gives a description of each column.

- **raw_speed**: freight speed taken from the raw dataset (before cleansing)
- **raw_speed_passenger**: passenger vehicles speed, taken from the raw dataset
- **cleansed_speed**: freight vehicle speed, taken from the cleansed dataset
- **removed**: a logical indicator, assuming values were removed (considered as outliers)
- **distance**: segment length, taken from the static file
- **distance_category**: segments categorized based on length by a 0.5 mile-length interval
- **hour**: hour of the day (EPOCH*5/60)
- **speed_difference**: variance between passenger vehicle speed (if available) and cleansed freight vehicle speed (Vp – Vf)
- **raw_speed_difference**: variance between passenger vehicles speed (if available) and the raw freight vehicle speed (Vp – Vf)
- **speed_category**: speed difference categorized as multiple of ten, based on speed_difference field

	DATE	EPOCH	TMC	raw_speed	raw_speed_passenger	cleansed_speed	removed	distance	distance_category	hour	speed_difference	raw_speed_difference	speed_category
17487	01302015	206	111P05483	65.09047	67.88006	65.09047	0	1.31989	1.5	17	2.7895914	2.7895914	0
17488	01302015	207	111P05483	63.35472	NA	63.35472	0	1.31989	1.5	17	NA	NA	NA
17489	01302015	208	111P05483	61.70914	NA	61.70914	0	1.31989	1.5	17	NA	NA	NA
17490	01302015	209	111P05483	57.94639	NA	57.94639	0	1.31989	1.5	17	NA	NA	NA
17491	01302015	210	111P05483	60.91800	73.10160	60.91800	0	1.31989	1.5	17	12.1836000	12.1836000	10
17492	01302015	211	111P05483	62.52111	NA	62.52111	0	1.31989	1.5	17	NA	NA	NA
17493	01302015	212	111P05483	65.09047	NA	65.09047	0	1.31989	1.5	17	NA	NA	NA
17494	01302015	213	111P05483	60.14689	NA	60.14689	0	1.31989	1.5	17	NA	NA	NA
17495	01302015	214	111P05483	61.70914	64.21086	61.70914	0	1.31989	1.5	17	2.5017220	2.5017220	0
17496	01302015	215	111P05483	57.24824	69.87653	57.24824	0	1.31989	1.5	17	12.6282884	12.6282884	10
17497	01302015	216	111P05483	60.14689	59.39505	60.14689	0	1.31989	1.5	18	-0.7518361	-0.7518361	0
17498	01302015	217	111P05483	60.14689	NA	60.14689	0	1.31989	1.5	18	NA	NA	NA
17499	01302015	218	111P05483	61.70914	66.92400	61.70914	0	1.31989	1.5	18	5.2148571	5.2148571	10
17500	01302015	219	111P05483	61.70914	73.10160	61.70914	0	1.31989	1.5	18	11.3924571	11.3924571	10
17501	01302015	220	111P05483	56.56671	74.24381	56.56671	0	1.31989	1.5	18	17.6770982	17.6770982	20
17502	01302015	221	111P05483	62.52111	NA	62.52111	0	1.31989	1.5	18	NA	NA	NA
17503	01302015	222	111P05483	57.24824	NA	57.24824	0	1.31989	1.5	18	NA	NA	NA
17504	01302015	223	111P05483	62.52111	NA	62.52111	0	1.31989	1.5	18	NA	NA	NA
17505	01302015	224	111P05483	58.66178	70.91946	58.66178	0	1.31989	1.5	18	12.2576849	12.2576849	10
17506	01302015	225	111P05483	69.87653	NA	69.87653	0	1.31989	1.5	18	NA	NA	NA
17507	01302015	226	111P05483	NA	NA	NA	NA	1.31989	1.5	18	NA	NA	NA
17508	01302015	227	111P05483	64.21086	64.21086	64.21086	0	1.31989	1.5	18	0.0000000	0.0000000	0
17509	01302015	228	111P05483	60.91800	NA	60.91800	0	1.31989	1.5	19	NA	NA	NA

Figure 90 Rearranged dataset representation.

I-35 and US-69 were examined in this study. First, distribution of speed difference between passenger vehicles and freight vehicles was explored and compared with the speed difference of the raw dataset. By fitting the data into a distribution, we found the mean and the STD for both cases. Figure 91 and Figure 92 represent the histogram of the speed difference between passengers and freight in both cases.

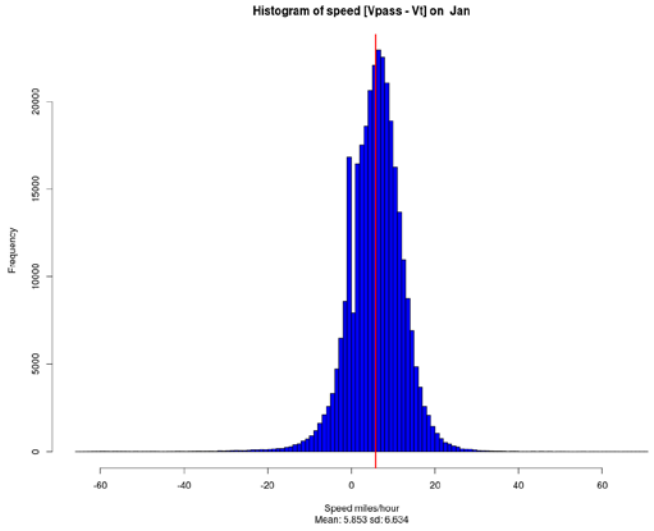


Figure 91 - Speed difference with outlier removal b/w passenger and freight speed.

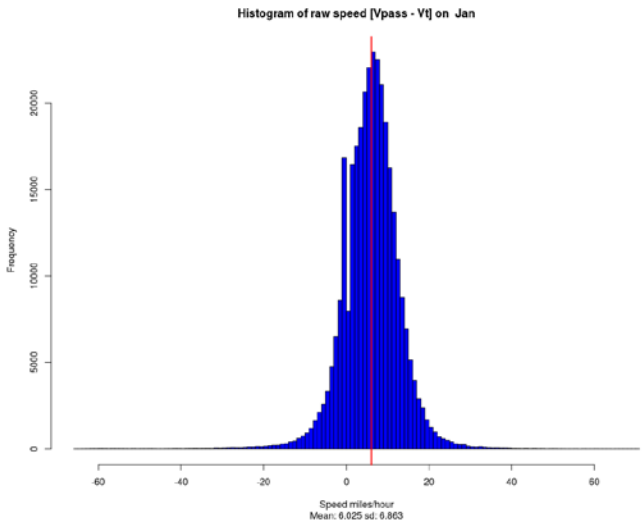


Figure 92 - Speed difference without outlier removal b/w passenger and freight speeds.

As we can see by comparing the raw and cleansed histogram, the mean and STD become less after cleansing the data, which make sense by removing the outlier values.

Exploring Outliers:

From the previous section, we saw the distribution of the speed difference before and after cleansing the data, and how the cleansing does not affect the distribution. We now focus on the removed values, to explore the behavior of the cleansing algorithm. Figure 93, shows a histogram of removed values categorized based on speed difference between passenger and freight vehicles.

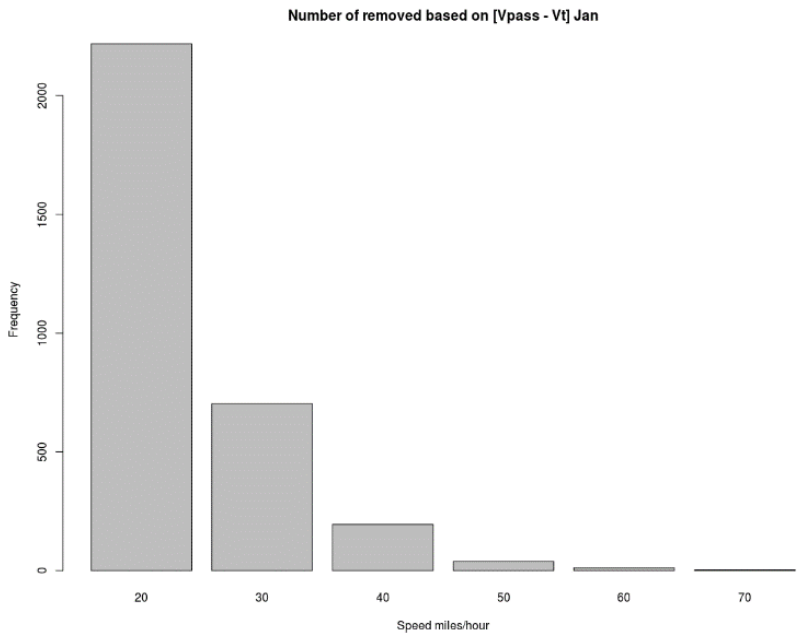


Figure 93 - Histogram of removed values based on speed difference.

Most removed values are in the category of 20 mile/hours, which matches with the threshold in the cleansing algorithm and the speed difference distribution we demonstrated in the previous section.
Figure 94 and Figure 95 show the number of removed values in I-35 data between January and Jun 2015. Most removed values per hour are between 8 and 14 values, which is considered normal.

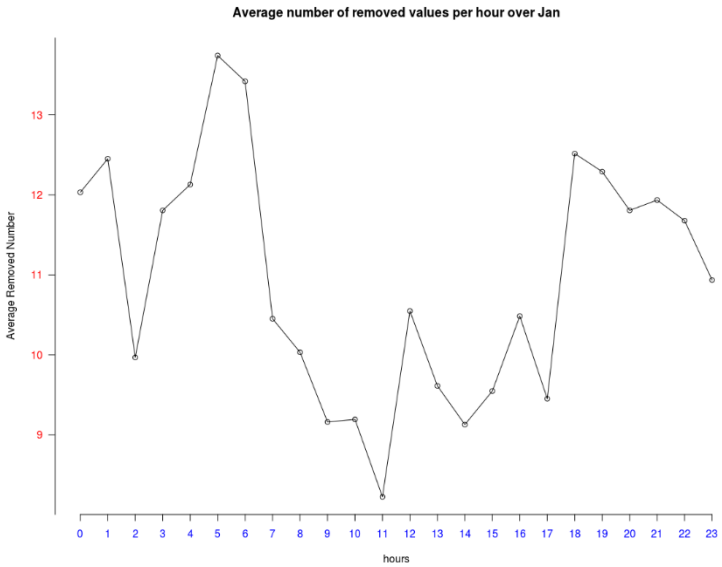


Figure 94 - Average number of removed values per hour during January.

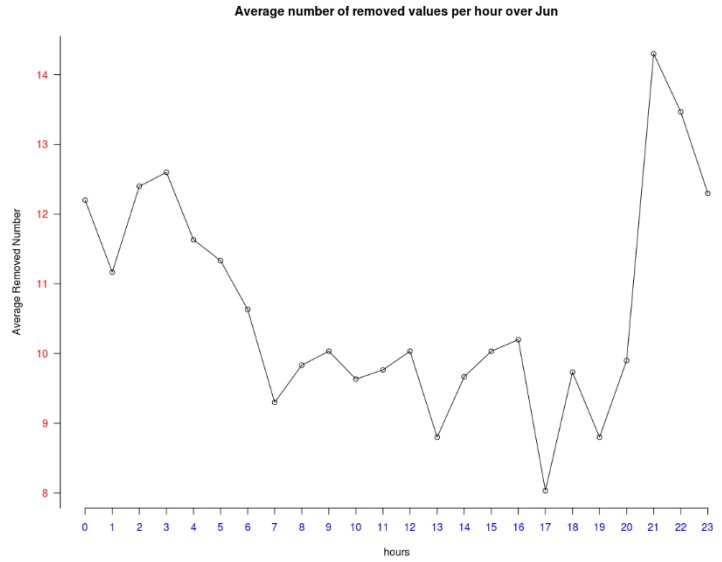


Figure 95 - Average number of removed values per hour during June.

Depending on “distance_category” created in the rearranging phase, we explore outliers values based on segment length. Figure 96 shows percentage of removed values during January. Most segments lengths experience nearly the same outlier removal percentage. A slight increase occurs at segment lengths of 8 to 9 miles, when compared to the others.

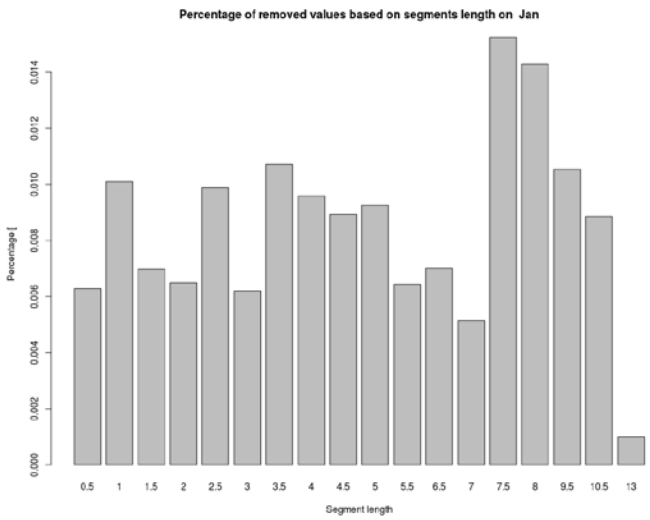


Figure 96 - Percentage of removed values based on segment length during January.

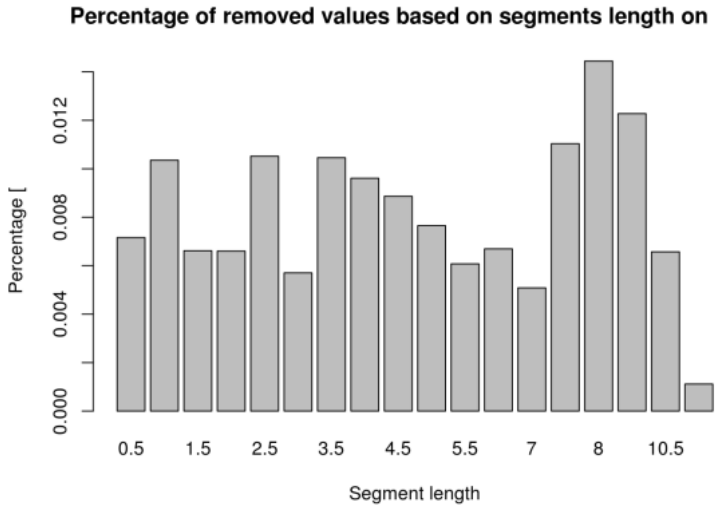


Figure 97 - Percentage of removed outliers for I-35 during February.

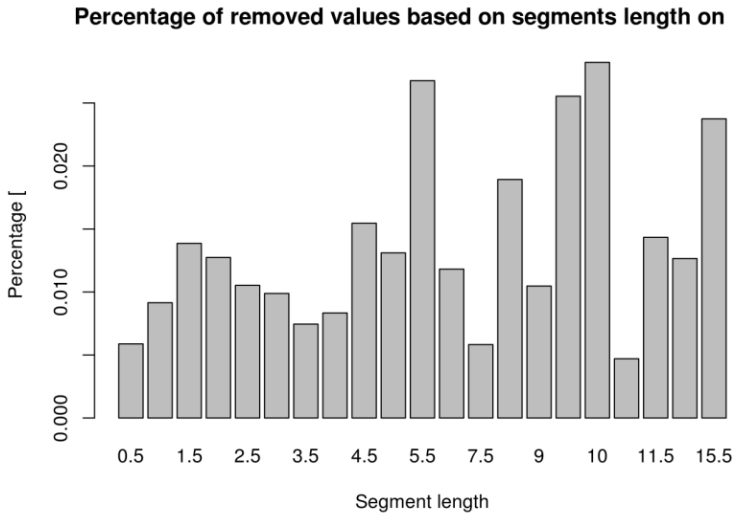


Figure 98 - Percentage of outliers removed for US-60 during July.

Table 1Table 12 and Table 13 represent statistical information about removed outliers for US-69 during January 2015.

Table 12 - Statistics for Removed Values for US-69 During January

Measure	Value
Total number of epochs	473184
Number of removed epochs	6135
Total number of missing epochs	141398
Percentage of removed values	1.85%

Table 13 - Statistics for Removed Values for I-35 During January

Measure	Value
Total number of epochs	892800
Number of removed epochs	8139
Total number of missing epochs	343751
Percentage of removed epochs	1.48%

The percentage of removed values is calculated as follows:

percentage = 100 * removed.values / raw.values.number,

where “raw.value.number” represents available epochs (total number of epochs – total number of missing values).

6.4. Challenges in the Outlier Removal Process

In the end, researchers note a general list of problematic points that affected outlier removal and the cleansing process of the NPMRDS v.1. The following points summarize the main challenges that should be addressed in future updates to the NPMRDS v.1.

1- **Unexplainable (illogical) change of directionally**: Cases of NHS roadways were found to include illogical directionality designations at the start or end segments.
2- **Multiple zero-mile markers:** Cases of roadways were found to have multiple zero-mile markers where flow direction would change from the positive (increasing mileage) to the negative (decreasing mileage) direction numerous times. This made determining the start and end of a roadway very difficult.
3- **Disconnections in roadways**: Cases of roadways were found to have large gaps between segments, making automatic sequencing of segments more difficult.
4- **Traffic lights and stop signs**: Traffic lights and stop signs disrupt traffic flow that could cause problems when generating outliers in the dataset.
5- **Quarterly static file changes**: Quarterly changes were made in the HERE static file (new segments added; old segments removed; segment lengths modified), causing inconsistencies in the database and additional difficulty in processing procedures for combined annual data.
6- **Errors in geographical directionality**: Cases of roadways were assigned wrong directions for their segments (i.e., [northbound, southbound] highway labeled as [eastbound, westbound] or vice versa), rendering any algorithm that relies on static file information prone to error when processing decisions based on highway directionality are made.

Chapter 7: CONCLUSION

Future ITS systems are expected to manage and resolve the arduous challenges of maintaining and improving roadway performance faced by today's transportation engineers and agencies. This can be achieved through systems incorporating intelligence, coupled with the ability to ingest highly heterogeneous data in real-time for performing 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 report presented research detailing the use of one of the nation's largest datasets of roadway travel times—the NPMRDS v.1. 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, which is key for performing accurate statistical analysis. 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 v.1 were reported. Research affirms careful consideration of system capture time granularity and segment length must be considered, as the interaction between the two—coupled with the speed of vehicles on the road—could result in generating anomalous data. 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. Finally, for congestion detection, removal of outliers contributed to the reduction of false alarm rate errors for segment congestion and congested days for both variance and thresholding detection methods alike. More importantly, the effect of outliers was severe on travel time reliability measures, such as travel time index, buffer time index, and planning time index. Thus, careful consideration for outlier removal must be taken into account when computing these measurements.

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] Naim Bitar, "Big Data Analytics in Transportation Networks Using the NPMRDS," 2016.

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

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

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

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

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

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

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

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

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

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

[38] 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..

[39] 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.

[40] 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.

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

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

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

[44] N. Bitar and H. H. Refai,, "A Probabilistic Approach to Improve the Accuracy of Axle-Based Automatic Vehicle Classifiers," *in IEEE Transactions on Intelligent Transportation Systems,* vol. 18, no. 3, pp. 537-544, March 2017.

[45] 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.

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

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

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

[49] 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.

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

[51] 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].

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

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

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

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

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

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

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

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

[60] 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.

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

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

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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 .

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] Antonio S. Cofıno, 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.

[78] 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.

[79] 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.

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

[81] 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.