

Application of Crowdsourced Data in Transportation Operations and Safety

**A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville**

**Nima Hoseinzadeh
December 2020**

Copyright © 2020 by Nima Hoseinzadeh
All rights reserved.

ACKNOWLEDGMENTS

I would like to express my deep gratitude to my advisor, Dr. Lee D. Han, for his contagious passion, support, and guidance during my graduate studies. More than an advisor during my Ph.D. program, he has been a role model and mentor. I am very fortunate to have the opportunity to work in his team, which taught me how to think critically and creatively. Without his encouragement and keen insights, this work would not have been possible. I thank Dr. Candace Brakewood, my co-advisor, for her contributions, encouragement, and valuable advice on my research. I also appreciate my statistics advisor, Dr. Russell Zaretzki, and my committee member, Dr. Anahita Khojandi, for serving as my committee members, and I thank them for their valuable suggestions. I am also thankful to Dr. Asad Khattak and Dr. Christopher Cherry for helping me improve my studies.

I will always be thankful to my colleagues and lab-mates: Yuandong Liu, Yangsong Gu, Harry Zhang, Zihua Zhang, Ramin Arvin, Iman Mahdinia, Amin Mohammadnazar, Mojdeh Azad, Ali Boggs, Mohsen Kamrani, Bumjoon Bae, Pankaj Dahal, Ruqing Huang, and many more. It was a great experience to work with them on various projects and share our experience and life. I am also grateful for all my friends in ISA-UTK, ITE, and WTS student chapters.

I especially thank my life partner, Ava Hedayatipour, for years of friendship and supports. We walked side by side in each step of our 7,500-mile journey, from Iran to the US, continuing to make our dreams come true. Thanks to my family, who supported me in all my decisions and motivated me to step up for my dreams. They sacrificed a lot to raise me in a free, encouraging, and confident environment called “home”. They always taught me of being a free and independent human, and now this is my personal worldview. I am extremely lucky to have supports from my wonderful friends (Hosseini, Saeed, Pegah, Parinaz, Golnaz, Shaghayegh, and many more) through this journey.

ABSTRACT

Crowdsourcing refers to the acquisition of data from users who contribute their information via smartphone, social media, or the internet. In transportation systems, crowdsourcing turns users into real-time sensors, providing data on traffic speed, travel time, mile traveled, incidents, roadway conditions, weather severity, irregularities in traffic patterns, and hazards. These data can be collected actively or passively in quantitative or qualitative forms. With the emergence of smartphones and navigation apps, crowdsourced data are gaining increased attention in transportation. Crowdsourced data have advantages over traditional fixed-location sensors and camera monitoring: low implementation costs, extended geographic coverage, high resolution, increased reliability, and the ability to perform proactive solutions. Transportation agencies can integrate crowdsourced data into their tools to manage the traffic and improve reliability and safety proactively. Also, crowdsourced data enables transportation researchers to propose innovative ideas and solutions not studied in the past.

This dissertation explores the applications of crowdsourced data in promoting transportation operations and safety. To this end, four studies are presented that integrate crowdsourced data in the transportation area. The first chapter evaluated and verified the quality of crowdsourced traffic speed data on surface streets in terms of accuracy and distribution. This study compared crowdsourced speed data collected from Waze to Bluetooth speed data. Based on the evaluated results, the crowdsourced data were used in the next two chapters. In the second chapter, a new methodology was proposed to assess traffic status and highway Level of Service to explore the application of crowdsourced data in transportation operations. This study exploited features from crowdsourced speed and travel time variation and incorporate them with crowdsourced user reports. The proposed methodology can be used in developing new tools for traffic status assessment on freeways with no need for fixed location sensors. Also, as the safety application, a spatiotemporal multisource lane-blocking incident detection system was provided in the third chapter. This study proposed a clustering algorithm called Weighted Spatiotemporal DBSCAN (WST-DBSCAN) that incorporates traffic speed pattern abnormalities with crowdsourced user reports to detect lane blocking incidents. The algorithm was evaluated by comparing it to the Tennessee Department of Transportation (TDOT) incident records. Lastly, the fourth chapter utilized crowdsourced vehicle mile traveled (VMT) and geographically weighted regression (GWR) models to explore the correlation between county-level spatial factors and pandemic-induced VMT changes (decline and recovery) during the COVID-19 outbreak.

Altogether, this dissertation provided a different framework to evaluate crowdsourced data and explore this data's capabilities in transportation applications. This dissertation showed that crowdsourced data are promising data sources in transportation analysis, operations, and safety.

TABLE OF CONTENTS

INTRODUCTION	1
Motivation	2
Data.....	3
Waze Travel Time/Speed Data.....	3
Waze Alerts	3
StreetLight	5
Contributions	5
Dissertation Structure.....	5
Chapter I: Crowdsourced Speed Data Quality	6
Chapter II: Assessment of LOS using Crowdsourced Data	6
Chapter III: Multisource Lane-blocking Incident Detection.....	6
Chapter IV: Geospatial Analysis of Mobility Amid COVID19.....	6
CHAPTER I Exploring Quality of Crowdsourced Speed Data on Surface Streets	8
Abstract	9
Introduction	10
Literature Review	11
Speed Collection Methods.....	11
Evaluating INRIX Data.....	11
Waze Data.....	12
Identifying the Gap.....	12
Methods	14
Step 0: Data Collection	14
Step 1: Waze Speed Validation	16
Step 2: Statistical Model of Factors Affecting Waze Data Quality.....	16
Step 3: Waze Data Quality Prediction.....	17
Case Study and Data Collection.....	18
Data Sets.....	18
Case Study Segments	20
Results	24
Step 1: Waze Speed Validation	24
Step 2: Statistical Model of Factors Affecting Waze Data Quality.....	33
Step 3: Waze Data Quality Prediction.....	37
Conclusions.....	37
CHAPTER II Assessment of Level of Service of Freeways Using Crowdsourced	
Data	40
Abstract.....	41
Introduction	41
Literature Review	43
Traffic Status and LOS Assessment Methods	43
Travel Time Reliability	43
Alternative LOS Methods.....	43

Waze Data.....	44
Gap in the Literature	44
Data.....	44
Waze Speed and Travel Time Data.....	44
Waze Crowdsourced Alert Data.....	47
Fixed Location Data.....	47
Study Area.....	47
Methodology.....	48
Step1: Data Collection	48
Step 2: Model Inputs.....	48
Step 3: Ground Truth LOS.....	50
Step 4: Machine Learning Methods	53
Results	54
Descriptive Statistics.....	54
LOS Classification Model Using Machine Learning	54
Sensitivity Analysis	58
Variable Importance.....	58
Conclusions.....	61
CHAPTER III Spatiotemporal Lane-Blocking Traffic Incidents Detection Using Multisource Crowdsourced Data	62
Abstract.....	63
Introduction	63
Literature Review	65
Existing Algorithms	65
Statistical Algorithms	65
Crowdsourcing in AID	65
Identifying the Gap.....	66
Data.....	66
Waze Speed and Travel Time Data.....	66
Waze Crowdsourced Alert Data.....	66
Study Area.....	67
Ground Truth (LocateIM data)	67
Methodology.....	67
Step1: Inter-quartile Range (IQR) Threshold Heatmap.....	67
Step 2: Detect IQR Potential Incidents	70
Step3: Match Waze Alert	70
Step 4: Incident Detection Using WST-DBSCAN.....	70
Performance Measurement	73
Results	74
Conclusion	77
CHAPTER IV Dynamics of Mobility Amid Covid-19 Pandemic – Geospatial Approach	78
Abstract.....	79
Introduction	79

Data.....	81
Study Area	81
StreetLight Data.....	81
Data validation	82
Dependent Variables	82
Explanatory Variables.....	84
Methodology.....	84
Variable Selection	84
Ordinary Least Squares (OLS)	87
Geographically Weighted Regression (GWR).....	87
Model Comparison Measures	88
Results	89
Descriptive Statistics.....	89
Modeling Estimation	89
Model Comparison.....	89
VMT Decline Phase Model	91
VMT Recovery Phase Model	94
Discussion	97
Conclusions.....	97
CONCLUSION	99
REFERENCES	103
VITA.....	113

LIST OF TABLES

Table 1. Summary of data sources used in this dissertation.....	4
Table 2. Summary of using probe vehicle speed data in the literature.....	13
Table 3. Summary of using Waze data in the literature.	13
Table 4. Descriptive statistics of length and AADT of different segments.	23
Table 5. Average number of hourly Bluetooth speed data update among all segments.	23
Table 6. Average percentage of historical speed data by Waze among all segments.	23
Table 7. Accuracy analysis of Waze speed data for different speed level.	27
Table 8. Definition of different quality clusters.	34
Table 9. Multinomial logistic regression model summary.....	36
Table 10. Different methods accuracy comparison.	38
Table 11. Summary of alternative LOS assessment methods in the literature....	45
Table 12. Summary of using Waze data in the literature.	46
Table 13. Model inputs equations.	51
Table 14. Label of different LOS (1).....	52
Table 15. Descriptive statistics of hourly data.	55
Table 16. Summary of classification methods with 3, 5, 10-fold cross validations.	60
Table 17. Lane blocking incident based on TDOT LocateIM.	68
Table 18. Data validation results.	83
Table 19. Explanatory variables of the study	86
Table 20. Descriptive Statistics of variables.....	90
Table 21. Comparison of OLS and GWR model (N=95 counties).	90
Table 22. Modeling results for VMT decline phase (N=95 counties).	92
Table 23. Modeling results for VMT recovery phase (N=95 counties).	95

LIST OF FIGURES

Figure 1. Dissertation Structure.	7
Figure 2. Framework of chapter I.	15
Figure 3. Data collection and case study. (a) Bluetooth data collection (map retrieved from www.sevierville.tn.gov/bluetoad), and (b) Waze data collection.	19
Figure 4. Case study location in Sevierville, TN.	21
Figure 5. Scatter plot: speed Difference (Bluetooth speed - Waze speed)	25
Figure 6. Boxplot of accuracy analysis results for 36 segments.	27
Figure 7. (a) Probability Density plot: Waze speed vs. Bluetooth speed (diagonal dashed line indicates perfect match).	29
Figure 8. Paired t-test result.	32
Figure 9. K-S test result.	32
Figure 10. K-means Algorithm results: Silhouette Method (Right) and Elbow Method (Left)	34
Figure 11. Framework of the proposed methodology.	49
Figure 12. Boxplot of number of alerts in each time period (hours).	56
Figure 13. Boxplot of different measures in each LOS category.	57
Figure 14. Different model accuracy and Kappa values.	59
Figure 15. Variable importance plot.	60
Figure 16. Study area on Interstate 40 (I-40) highway at westbound between mile marker 363 to 393.	68
Figure 17. Framework of chapter III.	69
Figure 18. Daily Threshold Heatmap.	71
Figure 19. WST-DBCAN implementation.	72
Figure 20. Comparing observed speed and IQR threshold.	75
Figure 21. Variation of DR and FAR for different IQR threshold constant and number of weeks of historical data	76
Figure 22. Comparison of proposed algorithm with other algorithms.	76
Figure 23. Boxplot of weekly VMT changes among 95 counties.	83
Figure 24. County-level VMT changes during decline and recovery phases (four major counties are labeled).	85
Figure 25. Decline Phase: GWR spatial distribution of estimated coefficients.	93
Figure 26. GWR spatial distribution of estimated coefficients: recovery phase.	96
Figure 27. Dissertation Key contributions and findings.	102

INTRODUCTION

Motivation

The Internet of Things (IoT) has emerged through the increased use of smartphones, Big data, and crowdsourced data as valuable tools for services and analysis. Statistics show that as of 2019, there are nearly 248.6 million smartphone users (with a penetration rate of 71.4%) in the United States [1]. Crowdsourcing refers to receiving data from a group of users who share information via smartphones, social media, or the internet. The application of crowdsourced data in transportation promotes researchers to a new era of proposing innovative ideas and solutions that were not studied in the past [2].

Road users may contribute crowdsourced information (e.g., speed, travel time, delay, incident, congestion) through mobile and navigation applications (e.g., Waze, Google Maps, Here, INRIX), connected and autonomous vehicles [3-8], social media (e.g., Facebook, Twitter), and other dedicated personal GPS devices. Transportation agencies and researchers gain from crowdsourced data by collecting and archiving them for further analyses. The data collected from these applications and devices are available in real-time or historical modes and lead to an opportunity to use it for different research directions.

Traditionally, transportation data are collected by fixed-location and infrastructure-mounted sensors such as loop detectors, remote traffic microwave sensors (RTMSs), magnetic sensors, laser sensors, video images, and License Plate Recognition (LPR) systems [9-11]. The costs associated with these collection methods usually include installation costs, power infrastructures, and maintenance fees [10, 12, 13]. Recently, new technologies such as location-based devices, cellular networks, Bluetooth sensors, and Wi-Fi to collect data has garnered more attention and applications [12, 14-18]. Crowdsourced data collected from floating cars and intensely active smartphone users have some advantages over fixed-location traffic [13]. Crowdsourcing is known as an emerging low-cost solution to improving safety and operations. Overall, crowdsourced data have the potential of following advantages for transportation analysis.

- Extended geographic coverage and expanded resolution
- Capability of real-time application
- Low/No implementation and maintenance costs
- Implementing reliable applications in operations and safety

This study aims to analyze and assess the potential of crowdsourced data in transportation applications in terms of mobility and safety. To this end, four different studies were designed. First, one study explores the quality of crowdsourced speed data. This study verifies data of the following studies with acceptable confidence. In the next chapter, the crowdsourced data (speed, travel time, and alerts) are used in a traffic status assessment algorithm. The third chapter links speed anomaly detection and crowdsourced alerts in an incident detection system. Finally, in the last chapter, the advantages and potentials of crowdsourced data in analyzing transportation mobility changes caused by due

to COVID19 will be discussed. This chapter will propose a county-based analysis to explore the impact of COVID19 on traffic.

Data

This dissertation will use different crowdsourced data sources that will be briefly introduced as follows. Waze crowdsourced data is the main data source of this study. Waze (<https://www.waze.com/about>) is a navigation app that shares data with agencies through a partnerships program called Waze for cities (WFC). This data source is used in first three chapter of this dissertation. Also, StreetLight (<https://www.streetlightdata.com/>) is a company that provides the estimate of daily vehicle miles traveled (VMT) and aggregates them at the county-level. It should be mentioned that other data sources such as roadside data sensors (RDS), TDOT incident logs (LocateIM), and Bluetooth speed data are used in different chapters to evaluate data and models. Table 1 provides data sources used in this dissertation,

Waze Travel Time/Speed Data

The Waze data estimate travel time and speed of specified links based on smartphone spatial coordinates generated by GPS sensors. The Waze App utilizes the speed and spatial coordinates information of the Waze users to provide roadway segments speed data. Waze provides a traffic view tool that allows traffic agencies to identify a list of road segments, referred to as a Watch-list. Authorized users can add any routes based on their data need or priority to this Watch-list. After that, travel time and the traffic speed data for the specified road segment is available and updated on a one-minute basis in real-time. Authorized users can use these data in real-time order or archive them in JavaScript Object Notation (JSON) format for further analysis or application. The archived JSON file of each time interval includes travel time, segment length, geospatial information of all predefined segments in that time interval. This data type will be used in Chapter I, Chapter II, and Chapter III.

Waze Alerts

User report data are another valuable crowdsourced data that Waze is providing to the partners, referring to as alerts. Waze alerts can be used in different applications and analysis such as incident detection, hot spot clustering, end of queue prediction. Waze users can report some predefined incidents while traveling in the Waze App. These alerts include accidents (major or minor), traffic jams (heavy, moderate, standstill, or light), hazards (severe weather, stopped cars, road potholes, etc.), construction, and closed roads.

Table 1. Summary of data sources used in this dissertation.

Data	Description	Source	Resolution	Coverage	Chapter
Traditional Data					
Roadside data sensors (RDS)	Speed and flow	roadside sensors	30 seconds	major urban areas highways	Chapter II and III
LocateIM	Incident logs	official incident records	-	major urban areas highways	Chapter III
Bluetooth Data	Speed and Travel time	Bluetooth MAC ID	1 minute	surface streets	Chapter I
Crowdsourced data					
Waze Speed	Crowdsourced Speed and travel time	Waze users' location and Speed	1 minute	statewide	Chapters I, II, III
Waze Alert	Crowdsourced user reports (Incidents, jams, accidents, severe weather, hazards)	Waze users report and location	1 minute	statewide	Chapters II and III
StreetLight Data	Vehicle Mile Travelled (VMT)	Travelers location	Daily	county and statewide	Chapter IV

The users also can verify the existing reports on the road. Waze shares all users' reports through the WFC program. Waze alerts crowdsourced data includes incident-unique ID, time, spatial coordinates, direction, reliability, and confidence of the reported alert. This study has access to Waze alerts in Tennessee State. Waze partners can employ real-time alerts or archive them in Extensible Markup Language (XML) format for their analysis. This data type will be used in Chapter II, and Chapter III.

StreetLight

Collecting and combining the users GPS data, StreetLight is a company that provides the estimate of daily vehicle miles traveled (VMT) and aggregates them at the county-level. StreetLight provided the average VMT for January as the reference under normal situations. Before this study, the StreetLight VMT dataset was used to examine health behaviors [19], policy intervention effects [20], air pollution [21, 22], and pandemic transmission [23]. In this dissertation, the StreetLight data will be used to explore the traffic changes during COVID19 outbreak. This data type will be used in Chapter IV.

Contributions

This dissertation provides novel frameworks and methodologies to evaluate crowdsourced data and explore the capabilities of these data in transportation applications. Overall, this dissertation (1) showed that crowdsourced data are promising data sources in transportation analysis, (2) incorporated crowdsourced data to propose a supplemental LOS assessment methodology in transportation operations, (3) proposed an incident detection algorithm using multisource crowdsourced data which improved detection rate and decreased false alarms, and (4) utilized crowdsourced data in exploring the mobility changes in unprecedented and unusual situations, such as COVID19 outbreak.

Dissertation Structure

Figure 1 presents the structure of this dissertation, including crowdsourced data used and four chapters. The first chapter explores the quality of crowdsourced speed data. The second chapter then discusses an operations application of crowdsourced data by proposing a methodology for LOS assessment. As a safety application, chapter three provides of crowdsourced data by proposing a multisource lane-blocking Incident detection algorithm. Finally, chapter four explores the mobility changes caused by COVID19 using geospatial analysis and crowdsourced VMT data. In the following, each chapter will be briefly discussed.

Chapter I: Crowdsourced Speed Data Quality

This chapter is designed to examine the quality of Waze speed data on real-world surface streets. Bluetooth speed data are used as the ground truth data. This chapter also aims to model data quality by exploring the contributing factors such as different times of day, days of the week, congestion levels, and speeds. This study showed the acceptable quality of Waze crowdsourced speed data and suggested that Waze speed data are promising data sources for using in further transportation analysis.

Chapter II: Assessment of LOS using Crowdsourced Data

This chapter proposes a new methodology that extracts features from crowdsourced data and speed/travel time deviation to assess LOS on freeways. This methodology can be used in developing new tools for LOS assessment and hourly traffic status on freeways with no need for fixed location traffic volume sensors. The proposed approach can be considered as a supplemental methodology for traditional HCM LOS calculation that relies on traffic density and flow.

Chapter III: Multisource Lane-blocking Incident Detection

This chapter aims to propose a spatiotemporal multisource lane-blocking incident detection framework with improved performance (a high detection rate and low false alarm rate). This framework incorporates traffic speed pattern with crowdsourced user reports in a proposed method called weighted spatiotemporal density-based clustering of applications with noise (WST-DBSCAN) to detect incidents. The proposed algorithm provided the highest detection rate and lowest false alarm in comparison to using a single dataset or using traditional methods. The proposed algorithm can help traffic agencies to implement real time incident detection with high network coverage.

Chapter IV: Geospatial Analysis of Mobility Amid COVID19

This chapter explores the correlation between county-level spatial factors and pandemic-induced vehicle mile traveled (VMT) changes (decline and recovery) during COVID-19 outbreak. It uses geographically weighted models to capture spatial heterogeneity in VMT changes and explore the association with county-level factors (commute to work, network, demographic, and economic).

Crowdsourced Data

 Waze Speed Speed and travel time Chapters I, II, III	 Waze Alerts Incidents, jams, accidents, severe weather, hazards Chapters II, III	 StreetLight VMT Vehicle miles traveled (VMT) Chapter IV
---	---	--

Dissertation Structure

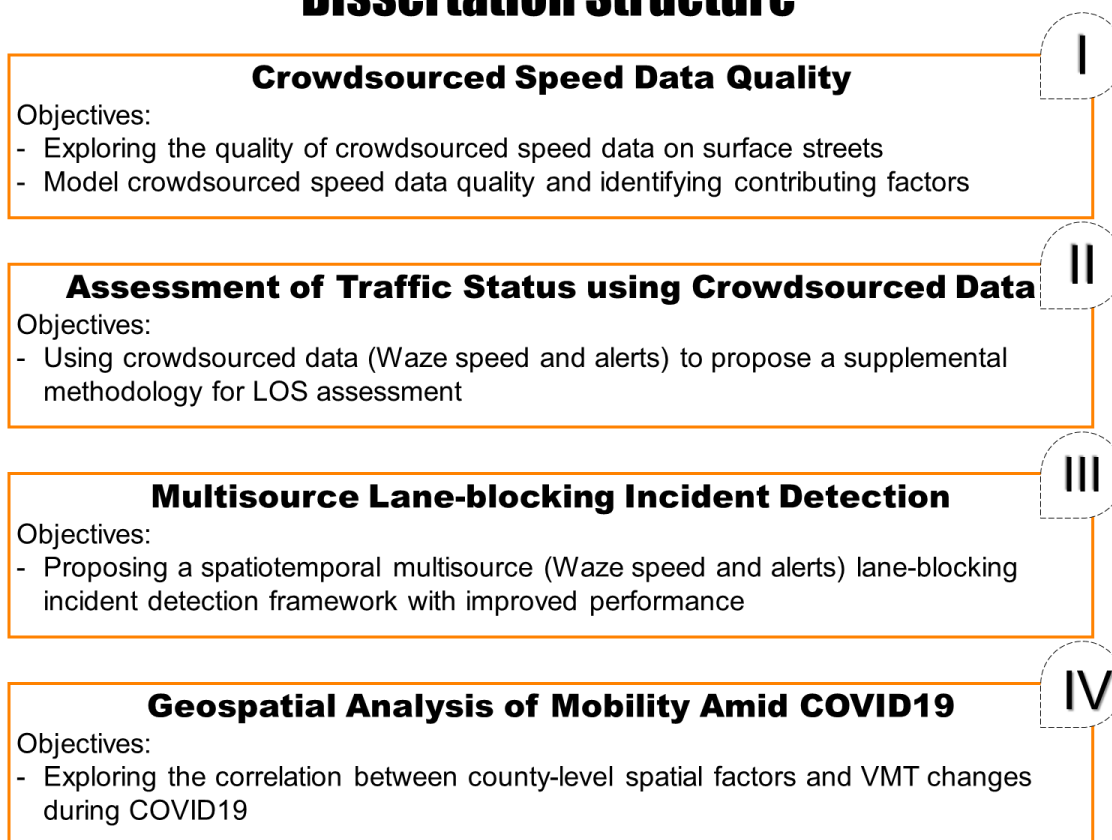


Figure 1. Dissertation Structure.

CHAPTER I
EXPLORING QUALITY OF CROWDSOURCED SPEED DATA ON
SURFACE STREETS

A version of this chapter was originally published by Nima Hoseinzadeh, Dr. Lee Han, and Dr. Brakewood:

Hoseinzadeh, N., Liu, Y., Han, L. D., Brakewood, C., & Mohammadnazar, A. (2020). Quality of location-based crowdsourced speed data on surface streets: A case study of Waze and Bluetooth speed data in Sevierville, TN. *Computers, Environment and Urban Systems*, 83, 101518.

Abstract

Obtaining accurate speed and travel time information is a challenge for researchers, geographers, and transportation agencies. In the past, traffic data were usually acquired and disseminated by government agencies through fixed-location sensors. High costs, infrastructure demands, and low coverage levels of these sensor devices require agencies and researchers to look beyond the traditional approaches. With the emergence of smartphones and navigation apps, location-based and crowdsourced Big Data are receiving increased attention. In this regard, big data collected from probe vehicles and road users can be used to provide speed and travel time information in different locations. Examining the quality of crowdsourced data is essential for researchers and agencies before using them. This study assessed the quality of Waze speed data from surface streets and conducted a case study in Sevierville, Tennessee. Typically, examining the quality of these data in surface streets and arterials is more challenging than freeways data. This research used Bluetooth speed data as the ground truth, which is independent of Waze data. In this study, three steps of methodology were used. In the first step, Waze speed data was compared to Bluetooth data in terms of accuracy, mean difference, and distribution similarity. In the second step, a k-means algorithm was used to categorize Waze data quality, and a multinomial logistics regression model was performed to explore the significant factors that impact data quality. Finally, in the third step, machine learning techniques were conducted to predict the data quality in different conditions. The result of the comparison showed a similar pattern and a slight difference between datasets, which verified the quality of Waze speed data. The statistical model indicates that that Waze speed data are more accurate in peak hours than in night hours. Also, the traffic speed, traffic volume, and segment length have a significant association on the accuracy of Waze data on surface streets. Finally, the result of machine learning prediction showed that a KNN method performed the highest prediction accuracy of 84.5% and 82.9% of the time for training and test datasets, respectively. Overall, the study results suggest that Waze speed data is a promising data source for surface streets.

Introduction

These days, nearly everyone has a smartphone. As of 2019, there are 248.6 million smartphone users (with a penetration rate of 71.4%) in the United States [1]. Through increased use of smartphones, the Internet of Things (IoT), Big data and crowdsourced data have emerged as valuable tools for location-based services and analysis. Crowdsourced data collected from smartphones and other location-based devices allow researchers to implement innovative ideas and propose smart cities and smart roadways solutions that were not feasible in the past [2].

Currently, it is possible to collect and archive fundamental information (e.g., speed, travel time, delay, incident, congestion) from road users through mobile location-based and navigation applications (Waze, Google Maps, Here, INRIX, etc.), connected and autonomous vehicles [3-8], social media (Facebook, Twitter, etc.), and other dedicated personal GPS devices. The high-resolution data collected from these applications and devices lead to an opportunity to collect big data and use it in different research purposes. These big data are available in real-time or historical modes and can be used in shortest path analysis, traffic signal control, queue detection, road users' decision models. Crowdsourced collected from floating cars and intensely active smartphones users have some advantages over fixed-location traffic sensors (e.g., Bluetooth and Wi-Fi speed sensors). For example, they are less expensive and are able to provide better coverage of the road network [13]. It should be noted that floating car data (FCD) is a term that refers to using data generated by vehicles in the traffic stream as a sample to estimate the spatial or temporal traffic data such as speed. Bluetooth and Wi-Fi speed data can be counted as an FCD [24].

Since leveraging crowdsourced data in smart city and smart transportation research gained more attention in the last few years, many cities and state departments of transportations (DOTs) have begun to cooperate with smartphone navigation apps such as Waze [8]. In this regard, Waze (<https://www.waze.com/about>) is a navigation app that employs probe vehicle data and crowdsourced road user reports to deliver information. Tennessee Department of Transportation (TDOT) has established partnerships with Waze through Waze for cities (WFC) to share data. Based on this agreement, Waze provides TDOT with incident reports, travel time, traffic speed data, and traffic jam levels on designated links. Waze is a crowdsourced data source that is provided to over one thousand cities and public sector organizations.

Employing crowdsourced brings up several questions for researchers about the quality of these data sources. Based on literature and the authors' experiences in different research projects, Waze speed data must be evaluated for further uses in future studies. Moreover, it is typically more challenging to examine probe vehicle data from arterials and surface streets than highways [25, 26]. The term "surface street" refers to an interrupted signalized roadway with

intersections. This term is typically used to differentiate and distinguish a road from freeways [27].

This study was designed to examine the quality of Waze speed data on real-world surface streets. Bluetooth speed data was used as the ground truth data. This study also aims to model data quality by exploring the contributing factors such as different times of day, days of the week, congestion levels, and speeds. Finally, this study also provides a method to predict the data quality in different conditions. The methods were applied to a case study in Sevierville, Tennessee, which is an area with a high level of tourist vehicles. The authors collected a month's worth of data from 36 signalized segments for the analysis that follows.

Literature Review

Crowdsourced data offers an opportunity to collect useful data from probe vehicles, bicycle, and pedestrians [28-32]. Because of the possibility of widespread use of crowdsourced data in the future, evaluation of quality, reliability, and validity of these data is critical to the future data collection and application [31]. This section, first, discusses speed data collection methods. Then, it provides a summary of probe-based speed data evaluation in literature and discusses different studies that evaluated INRIX data, which is similar to the data used in this chapter. Next, Waze data in the literature is discussed and finally, the research gaps are identified.

Speed Collection Methods

In the past, traffic speed and travel time data were collected by fixed-location and infrastructure-mounted sensors such as loop detectors, remote traffic microwave sensors (RTMSs), magnetic sensors, laser sensors, video images, and License Plate Recognition (LPR) systems [9-11]. The costs associated with these collection methods usually include installation cost, power infrastructures, and maintenance fees [10, 12, 13]. Recently, new technologies such as location-based devices, cellular networks, Bluetooth sensors, and Wi-Fi to collect data has garnered more attention [12, 14-18]. With the pervasive use of mobile phones and location-based devices, probe vehicle data has become one of the primary data sources in transportation. Private commercial vendors such as Waze, INRIX, TomTom, and HERE provide these types of data.

Evaluating INRIX Data

Table 2 summarizes studies that investigate the quality of probe vehicle speed data. It is shown that INRIX data have been used extensively as a data source [5, 6, 10, 33-36]. These studies used MAE [34, 37], MAPE [38, 39], distribution analysis [40], Speed Difference [5, 36], and t-test [39] to evaluate INRIX speed data. In addition, previous studies discussed different factors that influence data

quality [5, 6]. Congestion level is typically a significant factor in speed data quality [35, 41]. Traffic speed is discussed as another factor that has an association with speed data quality; the speed error was reported as being smaller at higher speed levels [42]. Time of day and segment lane also have been discussed as some of the factors that are prone to impact INRIX travel speed quality [5].

As shown in Table 2, most studies focused on highways and freeways [5, 34-38, 40]. It should be noted that few studies also select arterials for data evaluation [25, 39, 43]. Moreover, speed and travel time data errors from arterials are significantly higher than freeways [43]. Typically, it is more difficult to obtain the quality of crowdsourced data from arterials than freeways [25].

Different studies used various data sources as ground truth, e.g., Bluetooth data [6, 25, 36, 39, 40], and Loop detectors [35, 43]. Over the past few years, several studies have been conducted to evaluate floating cars and probe sourced vehicles such as Bluetooth and Wi-Fi speed data. These studies tried to verify the accuracy and reliability of the data by comparing it with fixed-location sensors [7, 34, 35, 42, 44-46]. As shown in Table 1, Bluetooth data has been widely used as ground truth in speed data evaluation. Bluetooth data collection technology has been used to estimate travel time and speed. Bluetooth speed and travel time data are known to be reliable and highly privacy-protected data sources [6]. Based on previous studies of Bluetooth data's reliability, it can be used as the ground-truth speed for this study [6, 15, 45, 46]

Waze Data

Table 3 provides a summary of prior studies that used or evaluated Waze data. Waze speed has not been widely evaluated in the literature. Waze provides speed and travel time data by collecting and analyzing the Waze application users' location. A more recent study evaluated Waze Incident report data reliability and coverage [8]. Also, the coverage of the Waze incident report [47] and Waze travel time computation [48] seems to be acceptable.

Identifying the Gap

As shown in Table 2 and Table 3, INRIX data were given the most attention in the prior literature, and Waze speed data has not been explored extensively. Moreover, most prior studies focus on highways or freeways, not arterials. This study addresses a gap in evaluating the quality of crowdsourced data (Waze speed data) on surface streets. Also, this study quantifies the impact of different factors on the quality of Waze data.

Table 2. Summary of using probe vehicle speed data in the literature.

Reference	Year	Type	Data	Ground Truth	Method
<i>Haghani et al. [34]</i>	2009	Freeway	INRIX	Bluetooth data	MAE
<i>Liu et al. [37]</i>	2012	Highway	INRIX	Pilot data	MAE
<i>Hamedi et al. [40]</i>	2013	Highway	INRIX	Bluetooth	Distribution Analysis
<i>Eleftheriadou et al. [38]</i>	2014	Freeway	INRIX	Floating Car	MAPE
<i>Kim and Coifman [35]</i>	2014	Freeway	INRIX	Loop detectors	Speed Error
<i>Zhang et al. [39]</i>	2015	Arterial	INRIX	Bluetooth	MAPE, t-test
<i>Tahmasseby [36]</i>	2015	Freeway	TomTom	Bluetooth	Speed Difference
<i>Ahsani et al. [5]</i>	2019	Freeway	INRIX	Wavetrax	Speed Bias

Table 3. Summary of using Waze data in the literature.

Reference	Year	Waze Data Used	Findings
Amin-Naseri et al. [8]	2018	Waze congestion and accident reports	Reasonable spatial and temporal accuracy of Waze data
Santos et al. [47]	2016	Waze accident report	Acceptable reliability of Waze accident report
Bahaweres et al. [48]	2017	Waze travel time	t-test showed the Waze travel time and ground truth are almost equal

Methods

Figure 2 presents the methodological framework of this study. First, the data should be collected and archived. In this part (Step 0), a data preprocessing and aggregation procedure were performed to make the data analyzable and comparable. Then, different methods were conducted to validate the quality of Waze speed data. The data analyses were performed in three main steps: Waze speed validation, Statistical Model of Factors Affecting Waze Data Quality, and Waze Data Quality Prediction. The following is a brief explanation of data processing and the different methods used in this study.

Step 0: Data Collection

Data Preprocessing

For each timestamp, Waze calculates the segment speed based on the observed probe vehicles. At the same time, the Bluetooth technology system detects and matches Bluetooth-enabled devices' Media Access Control (MAC) address in moving vehicles and estimates their segment speed. First, there should be a framework to collect and archive the datasets. After collecting and archiving data, a preprocessing procedure should be performed to (1) convert raw data to a useable format, (2) remove duplications, and (3) match Waze and Bluetooth data sources.

Data Aggregation

To enhance the comparability of the two time-series datasets, they need to be paired and set at the same times. Here, the question is how to group these data sources in fixed time intervals. If the data is aggregated in proper groups, each group can be assumed to be in the same traffic condition in its time interval. The archived Bluetooth speed data are collected in various time stamps. Also, this system updates the data when it observes and matches a new vehicle in origin and destination nodes. In this moving average system [9], as a new observation is detected, it will be replaced in the corresponding equations and the new speed will be reported. The Bluetooth data are collected in fixed travel distance in each segment. Thus, for Bluetooth data, the aggregated speed for each time interval group can be calculated as follows.

$$\bar{v}_i = \frac{n_i}{\sum_{k=1}^{n_i} \frac{1}{v_{ik}}} \quad (1)$$

Where,

$\bar{v}_i$ = aggregated speed for i^{th} time interval.

n_i = numbers of observation in i^{th} time interval.

v_{ik} = speed of k^{th} observation in i^{th} time interval.

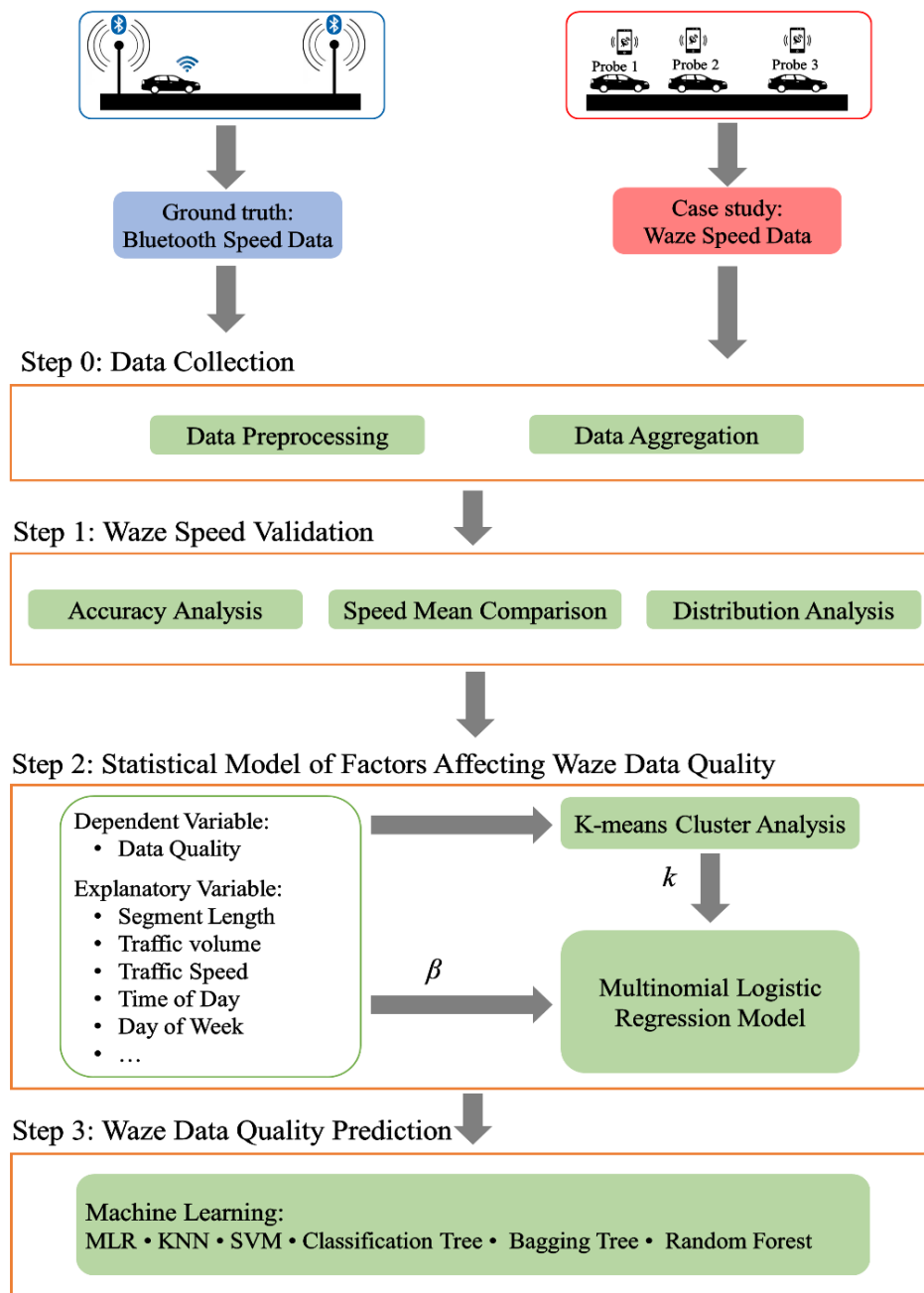


Figure 2. Framework of chapter I.

On the other hand, for Waze speed data, the time intervals are the fixed. Hence, the aggregated speed can be calculated as follows.

$$\bar{v}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} v_{ik} \quad (2)$$

In this study, based on previous studies [9], 5-minute time intervals were selected to aggregate the Waze and Bluetooth speed data.

Step 1: Waze Speed Validation

Step 1 is to validate the crowdsourced data (Waze speed data) compared to the ground-truth (Bluetooth speed) on surface streets. In terms of speed and travel time data quality assessment, prior studies used accuracy analysis such as Root Mean Squared (RMSE) [9, 49], Mean Average Error (MAE) [25, 34, 37], and Mean Average Percent Error (MAPE) [25, 38]. The acceptable error rate is typically 10 to 20 percent [38, 50, 51]. However, some studies have tried to conduct hypothesis testing such as t-test [7, 12, 15, 33, 48] and distribution analysis [40] to evaluate the quality of data. Other studies extract some features from speed and travel time to evaluate their data sources, including coefficient of variation, buffer index, 95th percentile, Level of Travel Time Reliability (LOTTR), standard deviation, data coverage, and data availability [5, 7, 40, 45, 52, 53].

As shown above, researchers can implement various time-series methods to validate and compare the speed data from one source (Waze data) by using an independent data as ground-truth data (Bluetooth data). In this chapter, pairwise data point analysis, speed mean difference, distribution comparison, and correlation analysis can be used in such a time-series analysis. Since both data sources are independent, the similar pattern and small value of deviation can verify the Waze data. RMSE, MAPE, MAE are well-known methods that can calculate the overall difference between the two data sources and evaluate the Waze data quality. T-test and KS-test are two helpful statistical methods that can perform further analysis of mean difference and distribution comparison, respectively. In the results section of this paper, these methods will be briefly discussed.

Step 2: Statistical Model of Factors Affecting Waze Data Quality

The second step will select factors such as time periods, days of the week, segment length, speed and congestion level to obtain a better understanding of the data quality. In the same vein, a statistical model is presented to analyze the contributing factors of data quality. Data quality is the dependent variable of the model and will be categorized in different levels of MAPE based on literature, which will also be confirmed by k-means analysis. The categories can represent different levels of Waze data quality from high to low. Subsequently, the model's goal is to explore the influence of a variety of factors on the quality of

crowdsourced data on surface streets. This model helps to estimate and quantify the significance of each variable. Multinomial Logistic Regression is a frequently used statistical method for analyzing categorical data analysis [54-57]. Suppose there are K categories of data quality and n explanatory variables. One category of data quality needs to be designated as the reference category, and P_k is the probability of association in k th category ($i=1:K-1$). Multinomial Logistic Regression calculates the probability of P_i compared to the probability of base category. This model requires to calculate $k-1$ equations to describe the relationship between explanatory variables and data quality index.

$$Q_k = \sum_i^n \beta_{i,k} X_i + \varepsilon_i \quad (3)$$

Where,

Q_k is the function of determining the impact of each variable on category k .

$\beta_{i,k}$ is the coefficient for k th category and i th variable.

X_i is i th explanatory variables.

ε_i is the error term.

The relationship between probability of being in the k th data quality (P_k) category and probability of reference category (P_r) can be written as

$$\ln\left(\frac{P_k}{P_r}\right) = Q_k = \beta_{i,k} X_i + \varepsilon_i \quad (4)$$

Subsequently, the probability of association in the k th category and probability of belonging to the reference category can be derived as following equations.

$$P_k = \frac{\exp(\beta_{i,k} X_i + \varepsilon_i)}{1 + \sum_{h=1}^{k-1} \exp(\beta_{i,h} X_i + \varepsilon_i)} \quad (5)$$

$$P_r = \frac{1}{1 + \sum_{h=1}^{k-1} \exp(\beta_{i,h} X_i + \varepsilon_i)} \quad (6)$$

Step 3: Waze Data Quality Prediction

After distinguishing the contributing factors that impact the data quality, it is essential to predict the data quality in different situations and circumstances.

Predicting data quality can be advantageous for researchers and data users (agencies and companies) for their crowdsourced data applications.

Furthermore, predicting and considering the data quality in every condition helps data users to prevent possible errors or compensate for the errors in their models or systems. For this purpose, several well-known machine learning approaches

are performed. K-Nearest Neighbors (KNN), Classification Tree, Bagging Tree, and Random Forest, and Support Vector Machine (SVM) are the main machine learning algorithms that have been used in this paper. These algorithms are selected based on their possible high rate of accuracy, fast processing time, and interpretability.

Case Study and Data Collection

In this part, first, some information about two main data sets of the study will be presented. Then, it will provide the spatial and temporal characteristics of this paper's case study.

Data Sets

The very first stage of using a new data source is to collect and archive them. To this end, primarily, the data users need to become familiar with the characteristics of the data. Then, they need to write a procedure and script to collect and archive the data. The following is presenting some information about Bluetooth and Waze speed data.

Bluetooth Data (Ground-Truth)

Bluetooth is a communication protocol that uses the MAC address of Bluetooth-enabled electronic devices (e.g., Smartphones, Bluetooth car stereo, Bluetooth car kit) to estimate travel time. To collect Bluetooth speed and travel time information of a segment, two Bluetooth sensors should be deployed at the start and end of the segment. If the start and end sensors match a MAC address, they can calculate the travel time and speed based on timestamps and segment length. The speed of each segment will be updated as a new pair of MAC addresses is observed [15]. In this study, the real-time Bluetooth data is provided by City of Sevierville. The city of Sevierville has invested in the Bluetooth traffic real-time monitoring system called "BlueTOAD". Figure 3 (a) shows the Bluetooth speed data collection procedure. Data is provided in Extensible Markup Language (XML) format. For each segment in three-seconds intervals, the real-time Bluetooth data will be retrieved and archived in a local data center. Here, the Bluetooth speed is analyzed and calculated by the data provider. As an end-user, this study only collected, archived, and used them. The Bluetooth data provider has its filtering process to remove outliers and present more reliable data. The authors of this study did not have any processing on removing outliers or calculating Bluetooth speed.

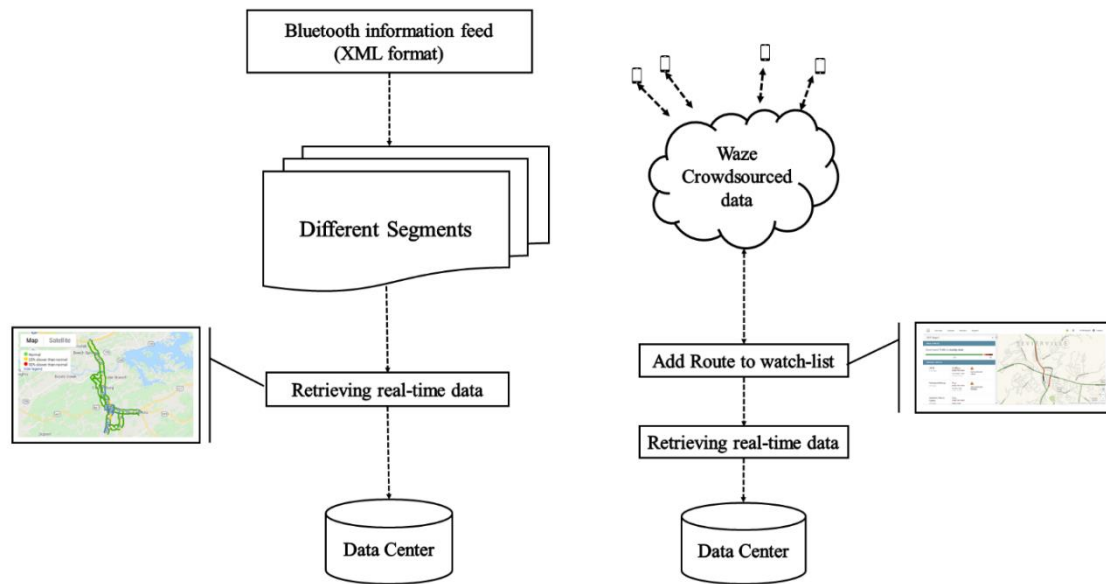


Figure 3. Data collection and case study. (a) Bluetooth data collection (map retrieved from www.seviervilletn.org/bluetoad), and (b) Waze data collection.

Waze Data

The Waze data is collected based on smartphone spatial coordinates generated by GPS sensors. The Waze App utilizes the speed and spatial coordinates information of the Waze users to provide roadway segments speed data. TDOT has established partnerships with Waze through Waze for Cities program to share data. Based on this agreement, Waze provides TDOT with travel time and traffic speed data as well as traffic jam levels on designated links, which is the primary dataset used in this study. Waze provides a traffic view tool that allows traffic agencies to identify a list of road segments, referred to as a Watch-list. Authorized users can add any routes based on their data need or priority to this Watch-list. After that, travel time and the traffic speed data for the specified road segment is available and updated on a one-minute basis in real-time. Waze investigates probe vehicles' space and speed data in the segment for each time interval. Then, it calculates and reports segment speed and travel time. If no probe vehicle is present, it reports historical speed and travel time. After archiving Waze data, it needs data preprocessing such as removing duplicate data [8, 58]. Figure 3 (b) displays the Waze data retrieval and archiving procedure: (1) adding routes to the watch-list, (2) extracting real-time data, (3) archiving data in the data management center.

Case Study Segments

Spatial and Temporal Characteristics

To assess the quality and verify Waze speed data and compare it to Bluetooth speed data on surface streets, a case study analysis was performed in Sevierville, Tennessee (Figure 3). The Sevierville city was chosen based on three significant reasons: (1) availability of Bluetooth data, (2) a large number of surface and signalized streets, and (3) tourist attraction of the area. The data in this study cover 36 surface street segments in Sevierville from January 1st to January 31st, 2019. Each segment in this study is located between two signalized intersections. The mean of Average Annual Daily Traffic (AADT) in these areas is 33,905 vehicles [59]. This area includes a big outlet mall and shopping center and is close to tourist attraction areas (Smoky Mountains National Park, Gatlinburg, Pigeon Forge, and Dollywood Park). These attractions and apparent high AADT levels make this place an advantageous case for this study. The city of Sevierville has invested in the Bluetooth traffic real-time monitoring system called "BlueTOAD". Meanwhile, the real-time traffic speed map of Sevierville is accessible via the City of Sevierville website (www.seviervilletn.org/bluetoad). In this study, the authors had access to real-time Bluetooth speed data in XML format and retrieved the data in 3 second time intervals. Figure 4 provides the location of bluetooth sensors and the segments in Sevierville.

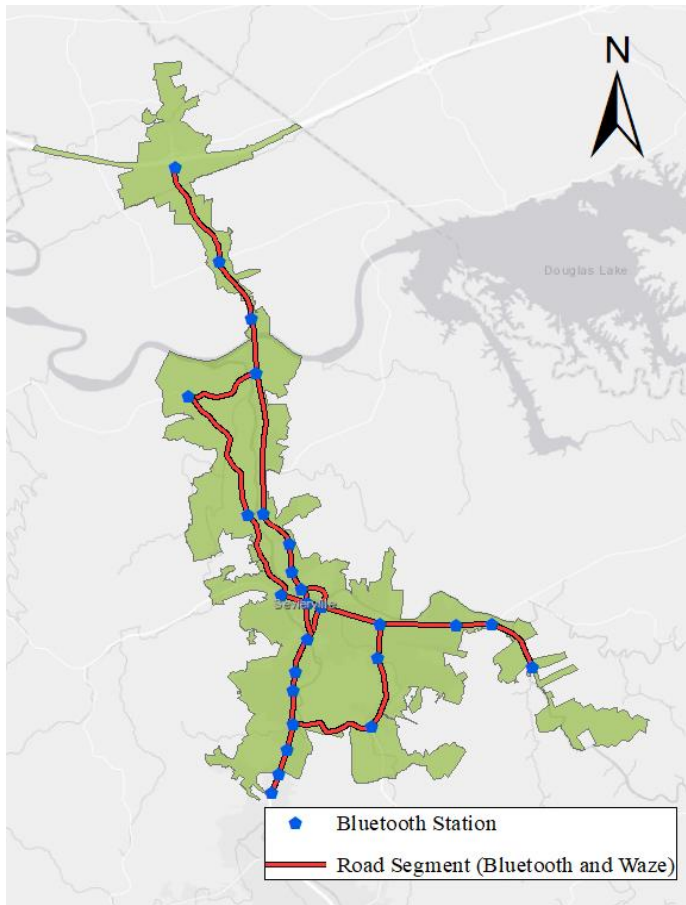


Figure 4. Case study location in Sevierville, TN.

At the same time, the exact segments were added to the Waze watch-list to collect the data. Data from 36 segments (18 segments in two directions) were available to compare with Bluetooth data. Table 4 provides the descriptive analysis of segments' length and AADT. The average segment length is 0.82 mile. However, the longest segment is in the Route 66 corridor (1.91 mile). In addition, based on Knoxville Regional Transportation Planning Organization (TPO), the AADT of case study segments is at least 20,000 vehicles per day [59]. As mentioned, this area is in a tourist attraction zone. Most vehicles in this area belong to tourists rather than commuters.

Data Availability in Different Traffic Flow

Before data were analyzed, it was important to compare traffic flow for different times of the day. First, each day was classified into four categories: AM Peak (7 AM to 10 AM), Mid-day (10 AM to 4 PM), PM Peak (4 PM to 7 PM), and Rest of Day (7 PM to 7 AM). As discussed, as the Bluetooth sensor system observes a new matched MAC address in a segment, it updates and reports the speed for that segment. Based on the matched data timestamp, the number of speed update in each segment was calculated. Table 5 compares the hourly average of the number of Bluetooth speed updates among all segments on weekdays and weekends. During the PM Peak and mid-day, there are, on average, 98 and 90-speed updates. However, this number declines to 27 at night. The same trend exists on weekends as well. From the observation, Route 66 holds the maximum amount, followed closely by I40, which has the highest AADT among all segments.

It is showing that for all segment the number of Bluetooth updates elevate in the daytime, especially PM peaks. The traffic count in each time of day can be compared based on the speed updates. This study used the number of Bluetooth speed updated in each time interval and established an explanatory variable called "Traffic Volume" index. In the statistical analysis section, the significance of this index in exploring data quality will be examined.

Like Bluetooth data, the vehicle count status can be estimated based on Waze data. As mentioned before, Waze uses historical speed data if it does not detect any probe vehicle in the segment. The percentage of time that Waze data used historical data was calculated for each segment. Table 6 compares the percentage of time that Waze reports historical speed. In both weekends and weekdays, the lowest percentage occurs during PM Peak, and the highest happens during night hours. During night hours, there are not many probe vehicles in the segment and the congestion level is low, so Waze uses more historical data.

Table 4. Descriptive statistics of length and AADT of different segments.

Characteristic	Mean	Median	S.D.	Min.	Max.
Length (mile)	0.82	0.76	0.46	0.22	1.91
AADT	33,905.75	29,782.5	9,166.587	20,865	50,561

Table 5. Average number of hourly Bluetooth speed data update among all segments.

Day Type	Time of Day	Mean	Median	S.D.	Max	Min
<i>Weekdays</i>	AM Peak	71.2	74.5	27.0	114.3	16.0
	Mid-day	90.8	94.3	24.8	134.8	34.2
	PM Peak	98.5	103.1	24.4	135.0	31.0
	Rest of day	27.3	25.8	9.5	44.3	5.2
<i>Weekends</i>	AM Peak	49.9	52.2	18.4	91.3	11.7
	Mid-day	83.3	85.6	22.6	124.6	35.8
	PM Peak	97.7	96.4	23.3	137.2	34.0
	Rest of day	40.1	40.9	13.2	62.2	8.4

Table 6. Average percentage of historical speed data by Waze among all segments.

Day Type	Time of Day	Mean	Median	S.D.	Max	Min
Weekdays	AM Peak	33.5	32.5	14.9	62.6	2.9
	Mid-day	13.6	9.9	9.9	42.7	0.6
	PM Peak	15.3	13.8	10.1	41.0	1.2
	Rest of day	59.7	59.2	13.5	82.4	34.9
Weekends	AM Peak	41.1	45.4	20.3	81.0	10.2
	Mid-day	29.0	24.1	21.0	82.9	2.9
	PM Peak	22.9	23.3	16.7	80.5	8.2
	Rest of day	82.9	82.4	6.7	93.5	68.9

Results

Step 1: Waze Speed Validation

Accuracy Analysis

A typical method for comparing two time-series is based on subtraction of pairwise data points. MAE, MAPE, and RMSE are some of the most common accuracy evaluation measures [9]. MAE and MAPE yield the absolute value and absolute percent of the difference between collected data and ground-truth data. RMSE captures the square root of the mean of squared differences between collected data and ground-truth data. The higher values of these indices show a higher deviation from the ground-truth data.

$$MAE = \frac{1}{N} \sum_{t=1}^N |V_{bl}^t - V_w^t| \quad (7)$$

$$MAPE = \frac{1}{N} \sum_{t=1}^N \frac{|V_{bl}^t - V_w^t|}{V_{bl}^t} \quad (8)$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^N (V_{bl}^t - V_w^t)^2}{N}} \quad (9)$$

Where V_{bl}^t and V_w^t are speed data obtained from Bluetooth and Waze at t^{th} time interval, respectively. Also, N denotes the number of observations.

In this section, the speed difference between ground truth data (Bluetooth speed) and Waze data was investigated. For this purpose, the speed difference (Bluetooth minus Waze speed) of one-month data were calculated for all segments. Then for each five-minute interval of both datasets (Waze speed and Bluetooth Speed), the average speed difference, and standard error of the mean value was calculated. Figure 5 shows the speed difference of the Bluetooth and Waze speed data over a month for different segments in each 5-minute aggregated speed data. It can be inferred that during peak hours, the speed difference is smaller than during night and early morning periods. In this figure, each dot represents the speed difference of a specific day for the corresponding time of day. Also, the red line and blue shadow represent the mean and standard error of each corresponding time intervals. It can be inferred that in the daytime and peak hours, the speed difference is, on average, closer to zero lines than in night times. This is an interesting issue that can be elaborated in the following analyses.

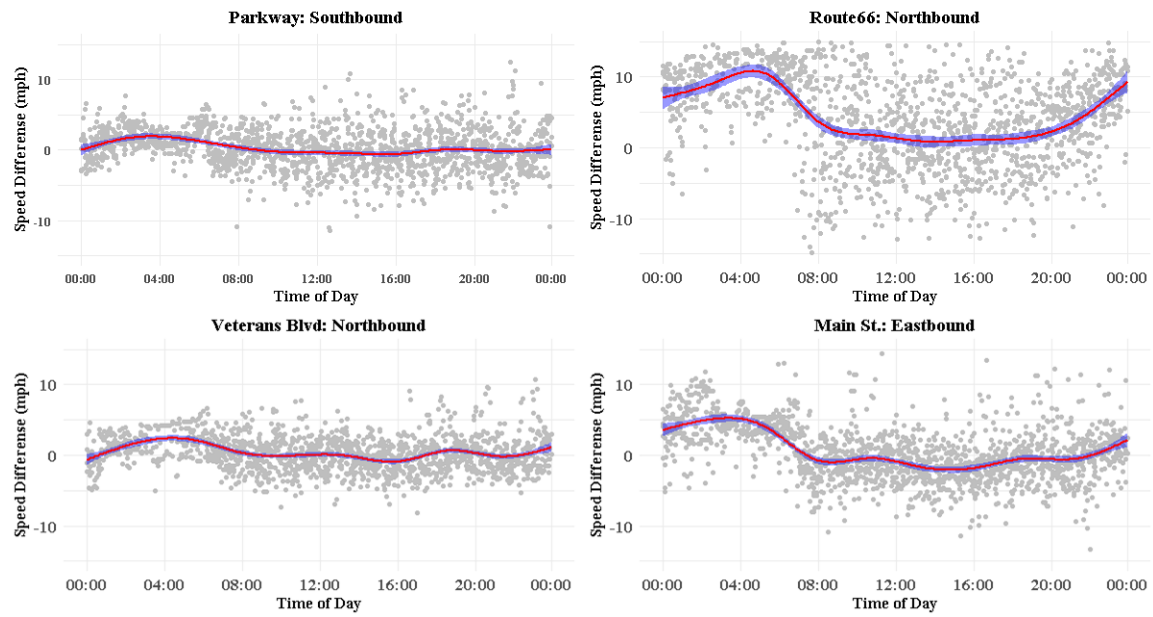


Figure 5. Scatter plot: speed Difference (Bluetooth speed - Waze speed)

To do the analysis, the proposed methods were applied to evaluate Waze speed data on surface streets. These methods compare Waze data with ground-truth (Bluetooth speed data) for accuracy analysis, mean difference, and distribution comparison. For this purpose, one month of Waze and Bluetooth data were collected in 36 segments. Afterward, to enhance the comparability of the paired data, both datasets were aggregated in a 5-minute interval basis.

Next, to perform the accuracy analysis, each day was first classified into four categories: AM Peak, Mid-day, PM Peak, and Rest of Day. Figure 6 presents the boxplot of the calculated value of RMSE, MAE, and MAPE for all 36 segments over a month in different periods. It shows that, on average, the Waze speed data is more accurate during peak hours and daytime than nonpeak hours and nighttime. That is an expected result because of low traffic flow during night hours. As discussed before, if the Waze system does not detect any probe vehicle in each time interval in the segment, it reports historical speed data, which causes more error and inaccuracy. It can be inferred that a higher level of vehicle counts would result in higher accuracy. In general, in most segments, the RMSE and MAE are less than 5 mph, and the average error is less than 10% of ground-truth data, which is an acceptable rate [38, 50, 51] in comparison to previous studies in similar speed and travel time verification.

In the following, the above analyses were repeated for different speed categories to clarify the sensitivity of the Waze accuracy to speed level. All data in various segments were divided into five speed-levels: (0-<20mph), (20-<30mph), (30-<40mph), (40-<50mph), and ($\geq$ 50mph). Table 7 shows the results for different speed levels. In the daytime (AM Peak, Mid-day, and PM Peak), the difference between Waze and ground-truth data slightly decreases as speed increases. In term of percentage of the difference, MAPE shows the same result. However, during nighttime (Rest of Day), the value of error is higher than daytime. Classifying speed shows that in higher speed categories Waze has a slightly higher rate of accuracy. Also, in daytime and peak hours, Waze speed data seem to be more accurate than nighttime and nonpeak hours. This trend is similar to what Hu et al. found in their study in Virginia [25].

Figure 7 (a) shows the comparison between the Waze and Bluetooth speed data over a month for four randomly selected segments using 5-minute aggregated speed data. The probability scatter plots of data from two data source are plotted for different time intervals of the day. The darker hexagons denote the higher density of data points, and the lighter color hexagons represent low-density data points. The diagonal dashed line ($y=x$) represents the perfect match between two datasets. The more probability on (or close) to this line, the higher the accuracy of Waze speed data. The color also shows the density of data points in the corresponding speed. As shown in the figures, for all segments in AM Peak, Middy, and PM Peak where more probe vehicles are present, the dark color surrounds the perfect match line. In contrast, during night hours and early morning, the higher densities are farther from perfect match line. This interpretation supports the results from the accuracy analysis.

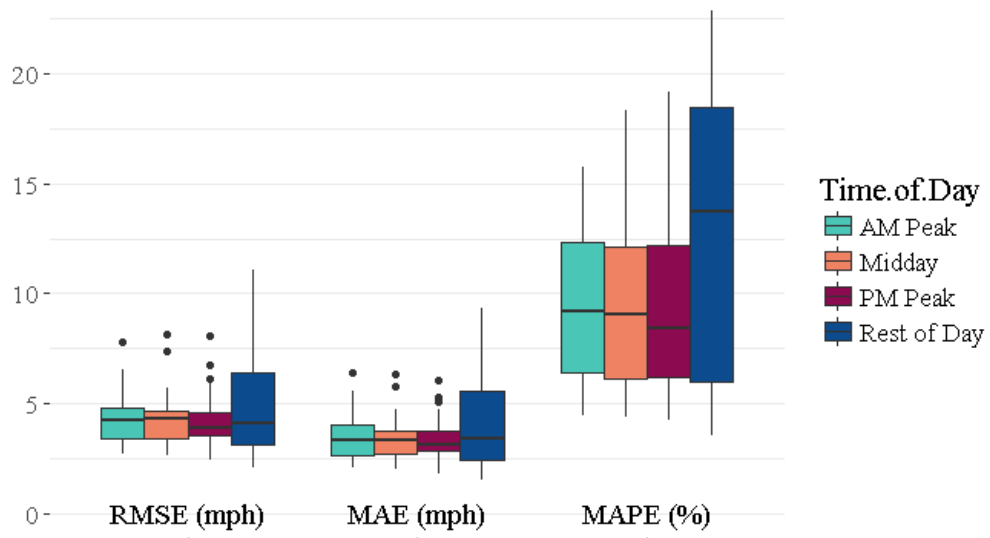


Figure 6. Boxplot of accuracy analysis results for 36 segments.

Table 7. Accuracy analysis of Waze speed data for different speed level.

Speed category (mph)	Day Time (AM Peak, Mid-day, and PM Peak)			Nighttime (Rest of Day)		
	RMES (mph)	MAE (mph)	MAPE (%)	RMES (mph)	MAE (mph)	MAPE (%)
0 - <20	3.3	3	13.2	4.1	3.4	21.8
20 - <30	4.5	3.6	14.1	4.3	3.7	14.4
30 - <40	3.8	3.1	8.8	5.7	5.3	14.7
40 - <50	4.7	4.2	8.6	8.0	7.5	17.1
≥50	3.9	3.7	7.1	11.5	11.1	21.2

Furthermore, in the figure, during night hours (Rest of day), the Waze speed in three out of four segments (Parkway, Veterans Blvd, and Main St.) exhibit a linear behavior. Again, this issue is caused by the small number of probe vehicles detected so it uses historical data that also have a small range of values (or sometimes constant value). However, on Route 66 the Waze data is less linear during night hours. Route 66 ends at Interstate 40, which has more probe vehicle during night hours than other segments. Also, Figure 7 (b) shows the comparison between the Waze and Bluetooth speed data for all segments together. The red line shows the nonlinear relationship between the two data sources. It can be inferred that during daytime, data sources are closer to the dashed line than night hours. Also, at higher speeds (over 45 mph), Waze and Bluetooth data are more similar than lower speed ranges (30 to 45mph).

Speed Mean Comparison

The paired t-test technique is an appropriate method for comparing two independent normally distributed time-series. This method first calculates the paired differences of two time-series datasets. The null hypothesis of this test indicates that the mean of differences (μ_{diff}) is equal zero.

$$\begin{cases} H_0: \mu_{diff} = 0 \\ H_1: \mu_{diff} \neq 0 \end{cases} \quad (10)$$

To test the null hypothesis, the t-statistics and difference confidence interval can be calculated as follow.

$$t - statistics = \frac{\mu_{diff}}{sd_{diff} / \sqrt{N}} \quad (11)$$

$$CI = \bar{\mu}_{diff} \pm t_{n-1, 1-\frac{\alpha}{2}} \frac{sd_{diff}}{\sqrt{N}} \quad (12)$$

Where

N = Number of observations

$\bar{\mu}_{diff}$ = Mean of differences between paired data.

sd_{diff} = Standard deviation of Mean of differences between paired data.

t = calculated t value

$t_{n-1, 1-\alpha/2}$ =t-value for N-1 degree of freedom and 100(1 – α) confidence.

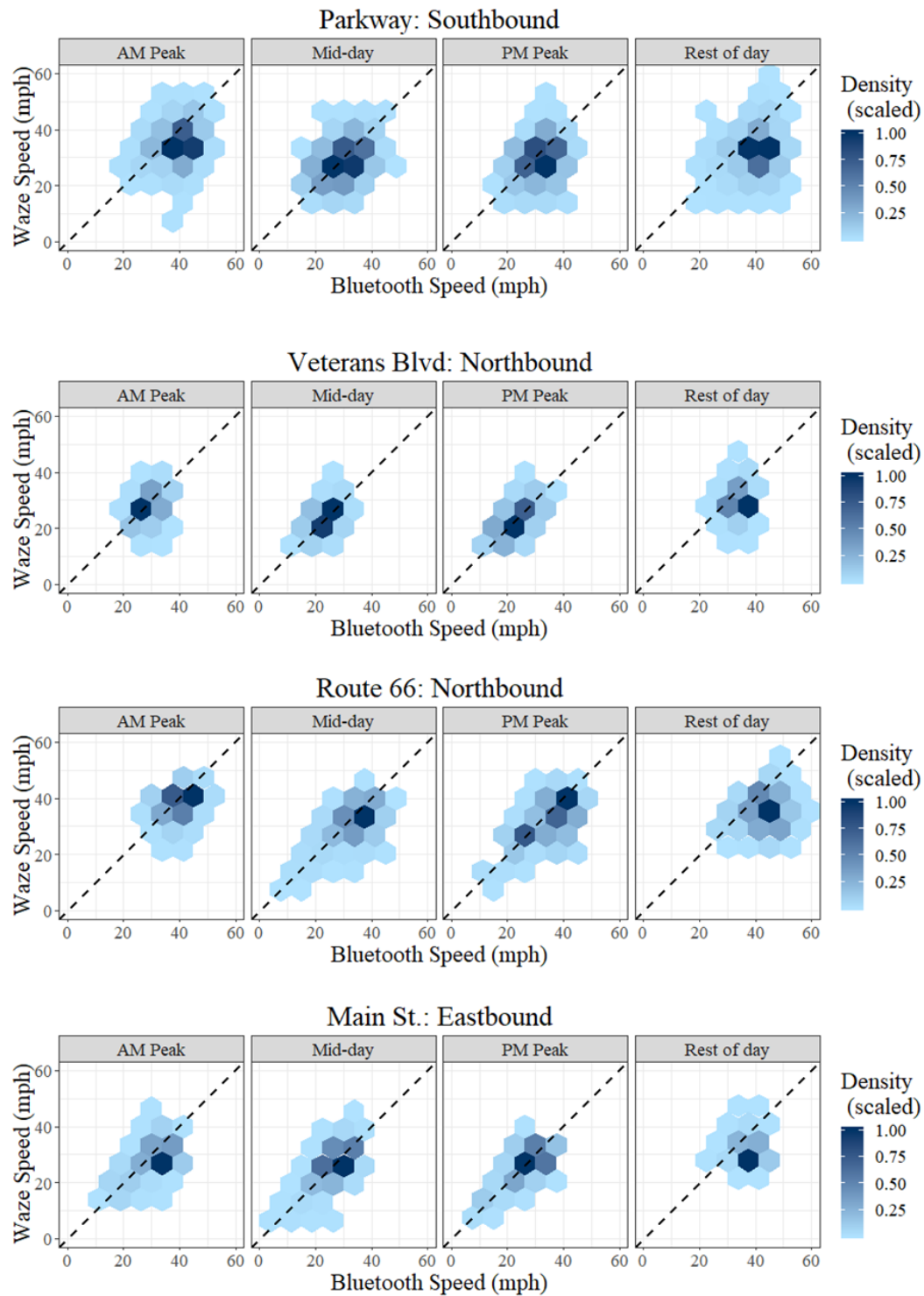


Figure 7. (a) Probability Density plot: Waze speed vs. Bluetooth speed (diagonal dashed line indicates perfect match).

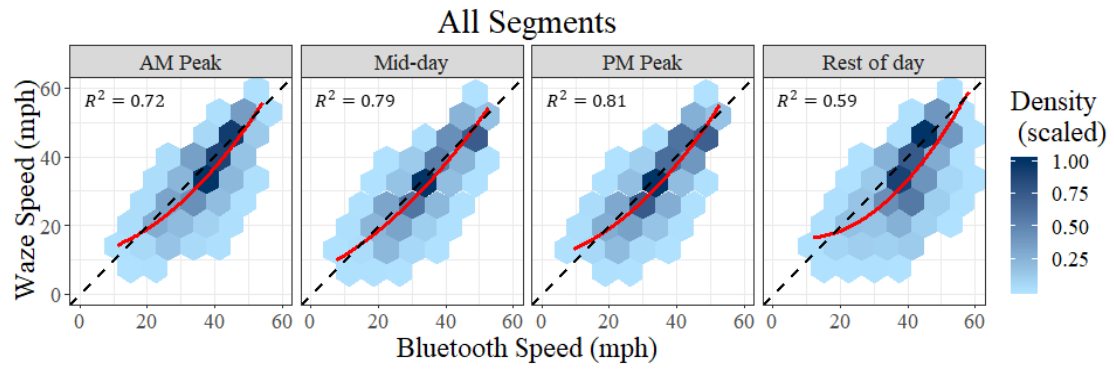


Figure 7. (b) Probability Density plot: Waze speed vs. Bluetooth speed all segments together (diagonal dashed line indicates perfect match).

Here, a paired t-test approach was tested on the null hypothesis, which indicates that the average difference between Waze and Bluetooth speed is equal to zero. The t-test was conducted on each hour and for all 36 segments over a month. For each different time of day and each segment, if the t-test includes zero value, it would be reported as “Passes t-test”. Figure 8 presents the percentage of segments that pass the t-test in each hourly time-interval. The results indicated that during daytime Waze and Bluetooth speed data are more similar than during night hours. Also, more segments in weekdays pass the t-test than in weekend. During daytime on weekdays, the percentage of passing t-test ranges between 60% to 80%. While on weekends, this range declined to 50% to 70%. Based on this result, it can be inferred that the mean value of two data sets for each time interval is significantly comparable to Bluetooth data. Like accuracy analysis, it is a reasonable interpretation that t-test validates the Waze speed data on surface streets. Also, when the number of observed probe vehicle increases (during daytimes), datasets are more similar and the trend of passing the t-test spiked. After 7 PM the trend dropped to 20%.

Distribution Analysis

The Kolmogorov-Smirnov (K-S) test is a common technique that measures the similarity between distributions of two datasets. This method tests the null hypothesis that two datasets have the same distribution. In this study, the K-S test can be used as follows:

$$\begin{cases} H_0: F_1 = F_2 \\ H_1: F_1 \neq F_2 \end{cases} \quad (13)$$

$$D_n = \max (F_1(t) - F_2(t)) \quad (14)$$

Where

F_1, F_2 = Cumulative Distribution Function (CDF) of two datasets.
 D_n = Kolmogorov-Smirnov Statistic.

Calculated D_n smaller than a critical value (a function of number of observations and confidence level) shows the data passes the K-S test. In contrast, the test would reject the null hypothesis if the D_n is larger than the predefined value [60]. Here, the Null hypothesis stated that the Waze data is distributed according to the Bluetooth data. Similar to the t-test, the K-S test was conducted for different segments in different times of day on an hourly basis. Then the percentage of segments that passed the K-S test were identified. Figure 9 summarizes the K-S test results for all segments. After 7 AM, the percentage of segments passing the K-S test climbed and this trend remains steady until 7 PM. During this interval, the K-S passing rate ranges from 65% to 80% on weekdays and from 55% to 75% on weekends as well. After this time, the trend falls to the range of 20%.

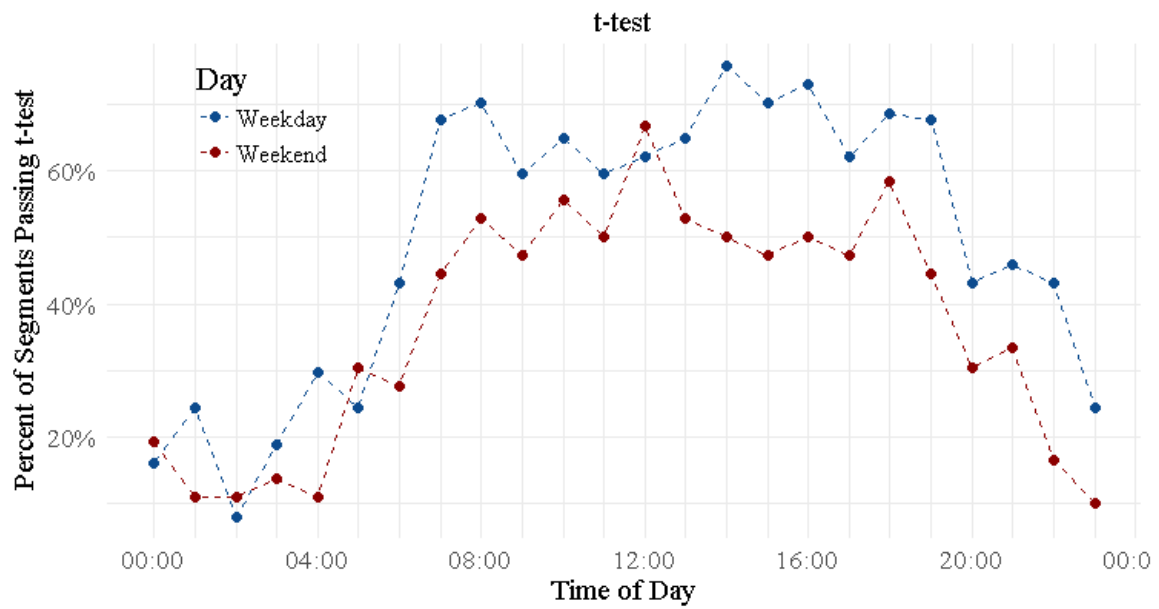


Figure 8. Paired t-test result.

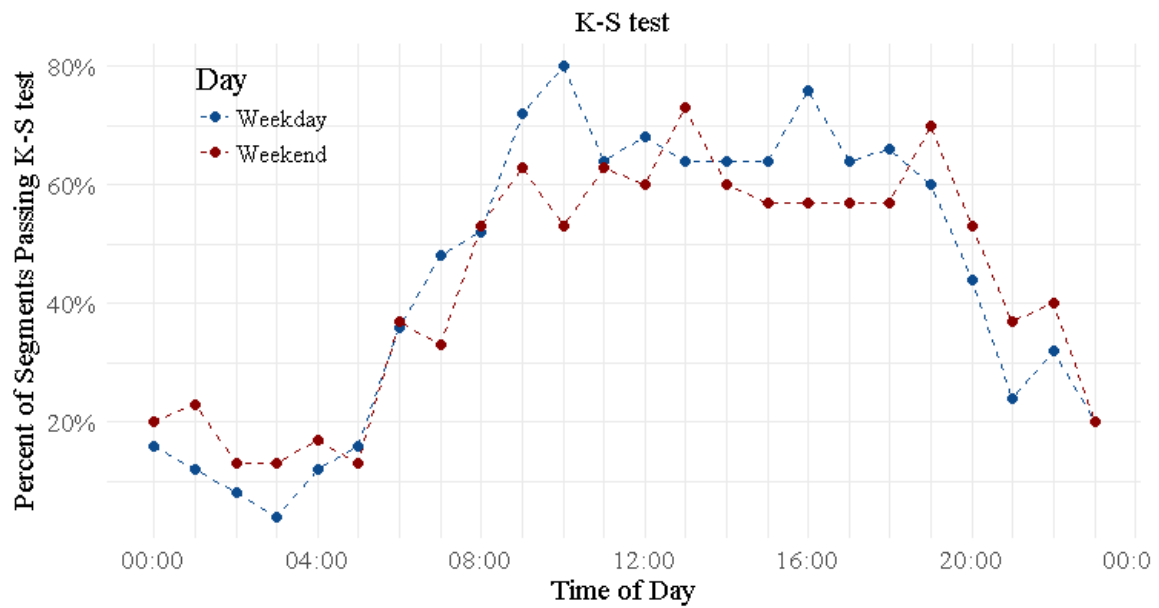


Figure 9. K-S test result.

It can be inferred that Waze data have a similar distribution to Bluetooth data when the number of detected probe vehicle increases. Overall, the result confirms that the Waze speed data show similar behavior to Bluetooth data. This verifies the Waze speed data as well.

Step 2: Statistical Model of Factors Affecting Waze Data Quality

Incorporating crowdsourced speed data into traffic operation and mobility analysis requires identifying different key factors that influence data quality. MAPE, as the “data quality index”, is defined as the percent of the difference between two data sources. Since speed data are collected by different equipment and different methods, there is always a difference between two independent data sources. Hence, a model is performed that presents insight into critical factors that influence Waze speed data quality to provide a further understanding of the quality of data, estimates the impact of each factor, and predicts data quality in different circumstances. A variety of possible contributing factors such as segment length, traffic volume indicator (number of Bluetooth updates), time of day, day of the week, and flow speed are used to model the characteristics of speed quality. Performing a statistical analysis helps to explore and quantify the significant factors of Waze speed data quality.

K-Means

First, speed data quality should be divided into different categories to model the contributing factors. For this purpose, a K-means algorithm was performed on a month’s worth of Waze data from 36 signalized segments to identify the number of possible clusters. K-means is an unsupervised algorithm that makes a reasonable inference from the contributing factors without referring to the labeled output. In this algorithm, k is referring to the number of clusters in the dataset. The Elbow Method and average Silhouette Method are two principal methods that help in identifying the optimal number of clusters in the datasets. The K-means algorithm was performed for different numbers of clusters. The Elbow Method choose the number of clusters that adding an extra cluster does not improve the total within-cluster sum of squares (WCSS). However, average silhouette method chooses the number of clusters with higher average silhouette width. Figure 10 presents the results of both methods for different number of clusters and the three clusters can were chosen as the optimal number of clusters.

Next, the speed quality number (MAPE) for each data observation was matched with the k-means clusters. The average speed MAPE for the three clusters showed 5.7%, 12.5%, and 22.9%. These numbers confirm the previous literature [38, 50, 51]. So, three categories were defined as high quality, medium quality, and low quality (Table 8).

Table 8. Definition of different quality clusters.

Cluster	High Quality	Medium Quality	Low quality
MAPE Range	0%-<10%	10%-<20%	20%>
Sample Percentage	52%	26%	22%

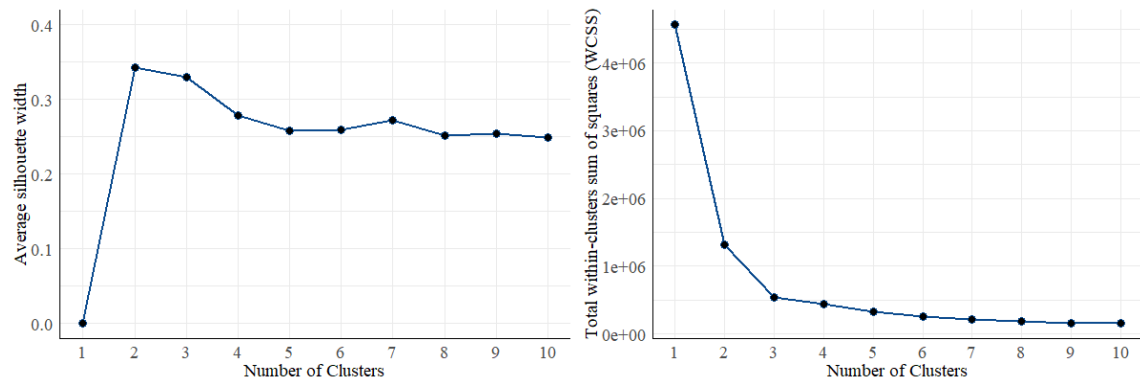


Figure 10. K-means Algorithm results: Silhouette Method (Right) and Elbow Method (Left)

Multinomial Logistic Regression Model

In this part, a statistical model should be selected to estimate the magnitude of the impact of each variable on the crowdsourced speed quality. First, a multinomial logistic regression model (AIC=32,905) and an ordinal logistic regression model (AIC=35,142) was performed. Akaike information criterion (AIC) was used to select the model. AIC is a well-known model selection method that evaluates the lack of fit of the model regarding the given data [61]. Based on AIC values, the multinomial logistic regression model was selected and reported in this study. In the multinomial logistic model, the order of the categories is not considered in the modeling. Also, this model provides a coefficient for each level of data quality. In this model, the “medium quality” category was selected as the reference category, and “high quality” and “low quality” categories are compared to this category. The model was statistically significant ($\chi^2 = 258.69, p - value < 0.001$) and explained 20% ($R^2 = 0.2$) of the variability of speed quality. The significant variables of the model are summarized in Table 9. Segment length ($p - value < 0.001$), traffic volume index ($p - value = 0.003$), and traffic speed ($p - value = 0.002$) have a significant impact on the quality of crowdsourced data. Based on the coefficient values, the higher values of segment length ($\beta = 0.001$), traffic volume ($\beta = 0.064$), and Bluetooth speed ($\beta = 0.041$) would increase the likelihood of having higher quality data. However, the coefficient for segment length ($\beta = -0.002$), traffic volume ($\beta = -0.045$), and Bluetooth speed ($\beta = -0.025$) are negative. This shows the small likelihood of Low-Quality category in higher segment length, traffic volume, and speed in comparison to High and Medium Quality. Moreover, time of day is another significant variable in the model. “AM Peak” and “Mid-day” times of days are prone to higher data quality than “PM Peak” and “Rest of day”. In general, daytime is more likely to have higher data quality than nighttime.

The traffic volume indicator, Bluetooth data updates, has a positive impact on data quality. Similarly, it can be inferred that the higher penetration rate of Waze users can increase the quality of the data. The model result confirms the observation in the previous sections of this paper. Also, the model indicated that longer segment length has a significant association with higher data quality, which is different from of Ahsani et. Al (2019) finding on INRIX data [5]. This might happen because of the different methods that Waze and INRIX calculate and aggregate speed data. The proposed model confirms the association of five main variables in estimating data quality. These five key variables can be used for further analysis and predicting data quality.

Table 9. Multinomial logistic regression model summary

Explanatory Variable	Categories	Quality category			
		High Quality		Low Quality	
		Coefficient	p-values	Coefficient	p-values
Constant		-2.553	0.025**	1.216	0.02 **
Segment Length		0.001	0.0001***	-0.002	0.0001***
Traffic Volume index		0.064	0.003***	-0.045	0.003***
Traffic speed		0.041	0.002***	-0.025	0.003***
Time of Day (Am Peak as base)	Mid-Day	0.174	0.047**	-0.269***	0.053
	PM-Peak	-0.126	0.05**	-0.407	0.049**
	Rest of Day	-0.248	0.05**	-0.253***	0.056
Type of Day (Weekends as base)	Workday	-0.250	0.045**	-0.089	0.052*
Model Goodness of fit					
	Adjusted R ²	0.20			
	Hosmer Lemeshow χ^2	258.69			
	df	16			
	p-value	< 000.1***			
Note:	*p<0.1; **p<0.05; p<0.01***				

Step 3: Waze Data Quality Prediction

Predicting the quality of crowdsourced data in different situations can help data users to have a better understanding of the data and be more confident in decisions made relying on the data. It also enables researchers to enhance the quality of data and prevent possible errors. For this purpose, a variety of machine learning techniques were employed to predict Waze data quality. The significant variables in the multinomial logistic regression model were used to train KNN, various Decision Trees (Classification Tree, Bagging Tree, and Random Forest), and SVM to classify and predict data quality.

The classification accuracy of different models is an index that can compare different techniques to choose the best model. The classification accuracy can be calculated as follow.

$$\text{Classification Accuracy} = \frac{\text{Number of correct prediction}}{\text{total number of prediction}} \times 100 \quad (15)$$

Next, the Waze data were aggregated in three different time intervals: 30-minute, 15-minute, and 5-minute. One-month worth of data (Bluetooth speed, segment length, volume indicator, time of day, day of the week, and data quality labels) for all segments were randomly divided into two data sets: a sample of 70% of data was used for training and fit the model, and a sample of 30% data was used for test the models.

Table 10 shows the accuracy percentage for training and testing data sets for all methods. Among all six methods, the KNN (with $k=11$) performs the best classification prediction accuracy for both training and test data for all time-interval aggregations. The 15-minute aggregation showed the highest accuracy. In the best scenario, KNN accurately predicted the category of the data quality 84.5% and 82.9% of time for training and test datasets, respectively. The tree models also confirm that segment length, traffic speed range, and volume indicator are the most important explanatory variables of the model. Overall, the KNN in 15-minute aggregation performed the best prediction.

Conclusions

Assessment of the quality of crowdsourced is vital before using them. This study examined and analyzed location-based crowdsourced speed data quality on surface streets. Waze speed data is a crowdsourced data that has not been evaluated in the literature. This study employed Bluetooth speed data as the ground truth, which is independent of Waze speed data. A case study was conducted in Sevierville, Tennessee, and one-month (January 2019) of speed data from 18 signalized segments (36 segments in two directions) were retrieved from Waze and Bluetooth systems. This study collected and archived real-time Waze and Bluetooth data by conducting different procedures and scripts.

Table 10. Different methods accuracy comparison.

Classifier	30-minute aggregation		15-minute aggregation		5-minute aggregation	
	Train	Test	Train	Test	Train	Test
Multinomial Logistic	68.9	69.7	72.9	72.7	68.8	68.7
KNN	81.7	79.5	84.5	82.9	78.9	76.9
Decision Tree	72.6	72.6	75.3	73.2	69.7	70.5
Bagging Tree	77.5	77.6	82.2	81.1	75.8	75.5
Random forest	78.0	77.8	80.1	81.2	75.2	74.7
SVM	75.6	74.5	78.6	77.7	72.9	71.5

After preprocessing (removing duplicated and outlier data), data were aggregated in 5-minute intervals. The aggregation process increases the comparability of the paired data. Then, the study method was conducted in three steps to evaluate the quality of Waze speed data. In Step 1, Waze data were compared with the ground truth. To this end, three techniques were used: (1) Accuracy analysis by calculating RMSE, MAE, and MAPE, (2) paired t-test to obtain the temporal comparison of the mean speed value in each segment, and (3) K-S test to analyze and compare the distribution of datasets in different time intervals and different segments. Since the two datasets are independent, the similar pattern and slight difference between datasets verified the quality Step 1 of the analysis:

- In almost all segments, the values of RMSE and MAE are less than 5mph during daytime (AM Peak, Midday, and PM Peak). Moreover, in these periods, the mean differences are less than 10 percent. However, during night times, the difference between Waze speed and Bluetooth gradually increased. Typically, Waze speed underestimates the speed during night hours.
- The Waze data tend to have a smaller percentage of error in higher speed ranges. The speed variability can affect the Waze speed accuracy.
- Paired t-test analysis shows that during peak hours two datasets are likely to have similar average speeds. However, during nighttime hours, fewer than 25 percent of segments passed the t-test. As for accuracy analysis and t-tests, two datasets show a more similar distribution during peak hours.

In the Step 2, data quality was divided in three categories (high quality, medium quality, low quality) based on MAPE value using k-means. Then, a multinomial regression model showed that segment length, traffic speed, traffic volume, time of day, and day of the week have significant associations with the data quality. For example, it can be inferred from the model that with a higher value of traffic volume and segment length, it is more likely to have a higher quality level.

Finally, in the Step 3, different machine learning techniques (Multinomial Logistics, KNN, SVM, Classification Tree, Bagging Tree, and Random forest) were applied to predict data quality based on contributing factors. These methods were applied for different data aggregation time-intervals (5-minute, 15-minutes, and 30-minutes). The result shows the KNN for 15-minute data aggregation resulted in predictions with the highest accuracy.

Overall, the case study results show the acceptable quality of Waze speed data on surface streets and explored the factors that impact on the quality. However, increasing the number of users can improve data quality and reduce errors. The crowdsourced speed data are promising data source for smart cities and transportation operations.

CHAPTER II
ASSESSMENT OF LEVEL OF SERVICE OF FREEWAYS USING
CROWDSOURCED DATA

A version of this chapter was originally presented by Nima Hoseinzadeh, Yangsong Gu, Dr. Lee Han, Dr. Brakewood, and Philip B. Freeze at Transportation Research Board (TRB) 2021 annual meeting

Abstract

In traffic operations, the aim of transportation agencies and researchers is typically to reduce congestion and improve safety. To attain these goals, agencies need continuous and accurate information about the traffic situation. Level-of-Service (LOS) is a beneficial index of traffic operations used to monitor freeways. The Highway Capacity Manual (HCM) provides analytical methods to assess LOS based on traffic density and highway characteristics. Generally, obtaining reliable density data on every road in large networks using traditional fixed location sensors and cameras is expensive and otherwise unrealistic. Traditional intelligent transportation systems (ITS) facilities are typically limited to major urban areas in different states. Crowdsourced data are an emerging, low-cost solution to potentially improve safety and operations. This study incorporates crowdsourced data provided by Waze to propose an algorithm for LOS assessment on an hourly basis. The proposed algorithm exploits various features from crowdsourced Waze user alerts and speed/travel time variation to perform LOS classification using machine learning models. Three categories of model inputs are introduced: basic statistical measures of speed, travel time reliability measures, and the number of hourly Waze alerts. Data collected from fixed location sensors were used to calculate ground truth LOS. The results reveal that using Waze crowdsourced alerts can improve LOS estimation accuracy by about 10% (accuracy=0.93, Kappa=0.83). The results of this research provide transportation agencies a LOS method based on crowdsourced data on different freeways segment (urban or rural) regardless of the availability of traditional fixed location sensors

Introduction

Intelligent transportation systems (ITS) are essential to assess the state of traffic. ITS traffic measurements can be used in different applications such as traffic operations, road work planning, assessing traffic queue, and congestion management. The United States Highway Capacity Manual (HCM) defines six Level-of-Service (LOS) to estimate traffic performance and state. The HCM provides analytical methods to assess LOS from traffic density and highway characteristics [62]. Traffic density, speed, and flow are key components of LOS assessment [62-64]. Department of Transportations (DOTs) and transportation agencies usually want real-time or historical hourly traffic status and LOS for different freeways segments.

Traditionally, traffic data (speed, travel time, flow, density) can be collected by a variety of fixed location sensors such as loop detectors, remote traffic microwave sensors (RTMS), magnetic sensors, laser sensors, video images, and License Plate Recognition (LPR) systems [9, 11]. Fixed location data collection methods are typically expensive and have limited network coverage. In recent years, data collection based on probe vehicles and floating cars has gained more attention. These use new technologies such as smartphones, cellular networks, Bluetooth sensors, Wi-Fi, and connected vehicles (CV) to provide traffic data [4, 16]. Meanwhile, these technologies provide a new opportunity to collect crowdsourced data.

Crowdsourcing refers to obtaining data from a group of users who contribute their information via smartphone, social media, or the internet. [2]. With increased use of smartphone applications, road users can share information (e.g., speed, travel time, delay, incident, hazards, hazards, severe weather, congestion) using navigation applications (e.g., Waze, Google Maps). Crowdsourced data are considered as a promising alternative to traditional data collection methods [58, 65-67]. Meanwhile, the advantages of crowdsourced data and probe vehicle over traditional fixed location data collection methods are: (I) expanded network coverage and resolution, (II) low/no implementation and maintenance costs, (III) improved real-time application, and (IV) ability to implement proactive applications [58, 67]. Since leveraging crowdsourced data, many cities governments and DOTs established partnerships with data providers such as Waze and INRIX [8, 58]. They utilized crowdsourced data in a variety of applications, such as performance measurement and incident detection. This chapter focuses on Waze. Waze is a navigation app that provides crowdsourced data such as speed, travel time, and road users reports (incidents, traffic jams, hazards) through the Waze for Cities (WFC) program. Tennessee Department of Transportation (TDOT) is a Waze partners that uses crowdsourced data. The acquisition of crowdsourced data brings up an opportunity to propose a new methodology to assess LOS based on this data's features and characteristics.

This chapter proposes a new methodology that exploits features from crowdsourced data and speed/travel time deviation to assess LOS on freeways. This methodology can be used in developing new tools for LOS assessment and hourly traffic status on freeways with no need for fixed location traffic volume sensors. The proposed approach can be considered as a supplemental methodology for traditional HCM LOS calculation that relies on traffic density and flow.

The remainder of this chapter is organized as follows. The next section summarizes related literature about LOS assessment, traffic situation prediction, and crowdsourced data in transportation. The methodology section presents the traditional LOS calculation and the proposed method of this paper. Also, some data mining approaches are discussed in this section. Then, the data used in this study are discussed, followed by results of implementing the methodology. Finally, the chapter is concluded, and area for future work are provided.

Literature Review

This part reviews the most relevant literature pertaining to this study in the following order. First, traffic status and LOS assessment methods are summarized. Then, travel time reliability and alternative LOS methods are discussed. The next parts will review Waze data used in previous studies and discuss the research gaps.

Traffic Status and LOS Assessment Methods

Most transportation agencies and DOTs focus on density or the volume to capacity (V/C) ratio to assess LOS [62, 68]. Traditionally, studies use one or a combination [69-71] of parameters such as speed [72], flow [70, 73], and density [74, 75] to explain traffic status and LOS. Previous studies used different data sources such as sensor data [76], probe vehicles [70, 77], camera images and videos [78], CVs [63, 79], and simulation [63, 73, 75, 79]. In terms of methodology, statistical modeling [75], Neural Networks [76, 77], Kalman Filters [77], Image Processing [80], and Machine Learning [70, 78] have been widely used.

Travel Time Reliability

Some more recent studies have captured useful information from speed and travel time variability and reliability to determine traffic status and performance. These studies used statistical measures such as average, standard deviation, percentiles, and range [52, 81, 82]. Additionally, the relation between speed deviation, travel time variability, planning time index (PTI) and buffer time index (BTI) with V/C ratio has been explored in prior literature [81, 83].

Alternative LOS Methods

SHRP 2 Reliability Project L08 discussed supplemental methods for LOS measurement. This project used a density-based definition of LOS to form the distribution of LOS and presented a distribution instead of a single value for LOS [84]. This study proposed an innovative approach for LOS based on travel time reliability perspectives. The travel speed range, the most restrictive condition, and travel time value were introduced in this project [84]. Travel time reliability and variability are measures of service quality [85].

Pulugurtha and Imran (2020) and Kodupuganti and Pulugurtha (2019) modeled the LOS of freeways and urban links using travel time variability indicators such as planning time index (PTI), buffer time index (BTI), average travel time, and the 95th percentile. They suggested a threshold travel time reliability to assess LOS [86, 87]. Singh et al. (2019) used Wi-Fi probe data to develop LOS thresholds based on travel time reliability and variability indices [68]. In a different approach, Altinatasi et al. (2016) used the average speed of

Floating Car Data (FCD) to quantify LOS [72]. Also, Khan, Dey, and Chowdhury (2017) used simulation and artificial intelligence to assess LOS based on different CV penetration rates [63]. Table 11 presents the most relevant studies that have proposed alternative methods for LOS assessment.

Waze Data

Crowdsourcing enables researchers to use data collected from road users, probe vehicles, bicycles, and pedestrians [31, 67]. This research uses Waze crowdsourced alerts and speed/travel time data as the primary crowdsourced data source. The company Waze analyzes the app users' location to provide speed and travel time data. Waze also provides different event reports (congestion, incidents, severe weather, road construction). Prior studies explored and verified Waze incident report and travel time data to assess reliability and coverage [8, 47, 48, 58, 88]. Some previous studies also used Waze alerts data for applications such as accident clustering [89] and improving dynamic traffic lights and mobility [90]. A more recent study also verified the quality of Waze speed data on surface streets [67]. Table 12 reviews prior studies using Waze data.

Gap in the Literature

As previously discussed, HCM density-based LOS were given the most attention in the prior literature. Moreover, some studies used travel time and speed variability for determining LOS. Crowdsourced data has not been used in to determine LOS. This study addresses a gap in integrating crowdsourced data (Waze incident report and speed data) for LOS assessment. This study's results can help agencies quantify LOS for different segments without installing new fixed location equipment.

Data

This part describes the primary datasets used in this study. This study uses Waze speed/ travel time and Waze alerts data for the LOS analysis. In the following, the Waze speed/travel time, Waze crowdsourced alerts, and fixed location data will be presented. Then, the study area will be introduced.

Waze Speed and Travel Time Data

Travel time and traffic speed for specific roadway segments are data sources that Waze shares with partners through the Waze for Cities (WFC) program. Waze obtains app users' kinematic information in specified segments to calculate and report speed and travel time. If no user is passing in that time interval or segment, it reports historical speed and travel time.

Table 11. Summary of alternative LOS assessment methods in the literature.

No.	Reference	Year	Data	Index Used	Method
1	<i>Chen et. al. [85]</i>	2003	Sensor data (speed, density)	Average travel time Standard deviation Percentiles of travel time	Travel time reliability Travel time cost
2	<i>Kittelson et al. [84]</i>	2008	Sensor data (speed, density, travel time)	Travel speed range Most restrictive condition Value of travel time	Travel time reliability threshold
3	<i>Altinatasi et al. [72].</i>	2016	Floating Car Data (speed)	Average speed	Speed threshold
4	<i>Khan, Dey & Chowdhury [63]</i>	2017	Simulation (speed, density)	Average speed CV penetration rate	Artificial intelligence
5	<i>Singh et al. [68]</i>	2019	Wi-Fi probe vehicle (speed, travel time)	Planning Time Index (PTI) Buffer Time Index (BTI) Travel Time Index (TTI)	Travel time reliability threshold Statistical regression
6	<i>Kodupuganti & Pulugurtha [87]</i>	2019	Travel time data provided by North Carolina DOT	Planning Time Index Buffer Time Index Average travel time 95th percentile travel time	Travel time reliability threshold Regression model
7	<i>Pulugurtha and Imran [86]</i>	2020	Simulation (travel time)	Planning Time Index Buffer Time Index	Travel time reliability threshold Statistical regression

Table 12. Summary of using Waze data in the literature.

No.	Reference	Year	Country	Waze Data Used	Findings and application
1	<i>Santos et al. [47]</i>	2016	Brazil	Waze accident alerts	Showed acceptable reliability of Waze accident report
2	<i>Bahaweres et al. [48]</i>	2017	Indonesia	Waze travel time	Performed t-test to show the Waze travel time and ground truth are almost equal
3	<i>Pack and Ivanov [58]</i>	2017	United States	Waze alerts	Explored the properties and possible benefits of Waze data
4	<i>Amin-Naseri et al. [8]</i>	2018	United States	Waze congestion and accident alerts	Showed reasonable spatial and temporal accuracy of Waze data
5	<i>Perez et al. [89]</i>	2018	Mexico	Waze alerts	Used Waze data for road accident clustering
6	<i>Raul Sanchez et al. [90]</i>	2019	Colombia	Waze jam and accident alerts	Used Waze data to improve dynamic traffic lights and urban mobility
7	<i>Turner et al. [88]</i>	2020	United States	Waze alerts	Found Waze data as a valuable safety data source especially for capturing unreported traffic incidents
8	<i>Hoseinzadeh et al. [67]</i>	2020	United States	Waze speed	Assessed the quality of Waze speed on surface streets

Waze implements a tool called “traffic view” (<https://www.waze.com/trafficview>) that allows transportation agencies and DOTs to specify a list of road segments. Authorized users can add links based on their priority or needs to the Watch-list. Subsequently, real-time travel time/speed data for the predefined road segments are available on a one-minute level. Authorized users can use these data in real-time or archive them in JavaScript Object Notation (JSON) format for further analysis. The archived JSON file for each time interval includes travel time, segment length, and geospatial information for all predefined segments in that time interval.

Waze Crowdsourced Alert Data

User report data are another valuable crowdsourced data that Waze provides to partners, which are referred as alerts. Waze alerts can be used in different analyses such as incident detection, hot spot clustering, and end of queue prediction. Waze users can report some predefined incident types in the Waze App while traveling. These alerts include accidents (major or minor), traffic jams (heavy, moderate, standstill, or light), hazards (severe weather, stopped cars, road potholes), construction, and closed roads. Users can also verify existing reports on the road. Waze shares all users' reports through the WFC program. Waze alerts data include an incident unique ID, time, spatial coordinates, direction, reliability, and confidence level of the reported alert. Waze partners can use real-time alerts or archive them in Extensible Markup Language (XML) format for their analysis. This study has access to Waze alerts for Tennessee State.

Fixed Location Data

Tennessee Department of Transportation (TDOT) uses Radio Data System (RDS) sensor data, which provides traffic information such as traffic count, speed, and occupancy in 30-sec time intervals. RDS stations are located on freeways close to four major cities, including Nashville, Memphis, Knoxville, and Chattanooga in Tennessee. This study uses traffic volume (flow) from RDS data and traffic speed from Waze to calculate density and LOS. The estimated LOS can be used as ground truth. This will be elaborated on the methodology section.

Study Area

To quantify hourly data and assess the LOS on freeways, a study area was designated in Knoxville, Tennessee. A segment on Interstate 40 (I-40) highway at westbound mile marker 385 was selected. The study segment length is about 1.5 miles. The speed limit for this segment is 65 mph. This location was selected based on two reasons: (1) variability of traffic and LOS during hours of the day, and (2) availability of roadway sensor data (flow) to calculate ground truth different LOS.

Methodology

The methodology used in this paper combines raw crowdsourced speed and user report data to obtain hourly LOS-based traffic status. This methodology uses the speed variation, travel time reliability, and user alerts in the selected segment to define measures of traffic conditions. Here, the Waze speed/travel time and crowdsourced alerts are used as the primary data source in the study, which will be elaborated on the following sections. Unlike some previous studies, this method does not depend solely on average speed [72] or density.

This section provides more details about the proposed algorithm of this study. As shown in the framework of the study (Figure 11), the different steps of the proposed method are listed below:

- **Step 1:** Data collection, which includes archiving Waze data and traditional fixed location sensor data, preprocessing and normalization.
- **Step 2:** Extract model inputs, which includes statistical measures, travel time performance measures, and crowdsourced Waze alerts.
- **Step 3:** Calculating ground truth LOS using fixed location sensors and labeling observations of Waze input data with the corresponding ground truth.
- **Step 4:** LOS assessment by performing different machine learning methods. This part includes feature selection, cross validation, and selecting the preferred method.

Step1: Data Collection

Waze continuously generates a massive amount of data. The first step in such a study is to archive Waze speed/travel time and Waze alert data. A python code was implemented to capture crowdsourced alert, speed, and travel time data for a 1-minute time intervals. Employing real-world raw data can always have some challenges, such as missing values or noise. In the next step, data was preprocessed by cleaning and removing possible errors. Possible missing values and outliers were removed/imputed. Next, RDS traffic volume data collected to calculate the hourly traffic flow and LOS ground truth, which will be elaborated in the Step 3.

Step 2: Model Inputs

As explained, previous studies have explored the variation of speed/ travel time to capture LOS. This study combines different speed and travel time variation indexes with crowdsourced data to assess LOS. Multiple indices were calculated as the inputs of the classification model. This paper divided these indicators into three categories as follows. Each index will be elaborated in the following paragraphs.

- Basic statistical measures
- Travel time performance measures
- Crowdsourced data

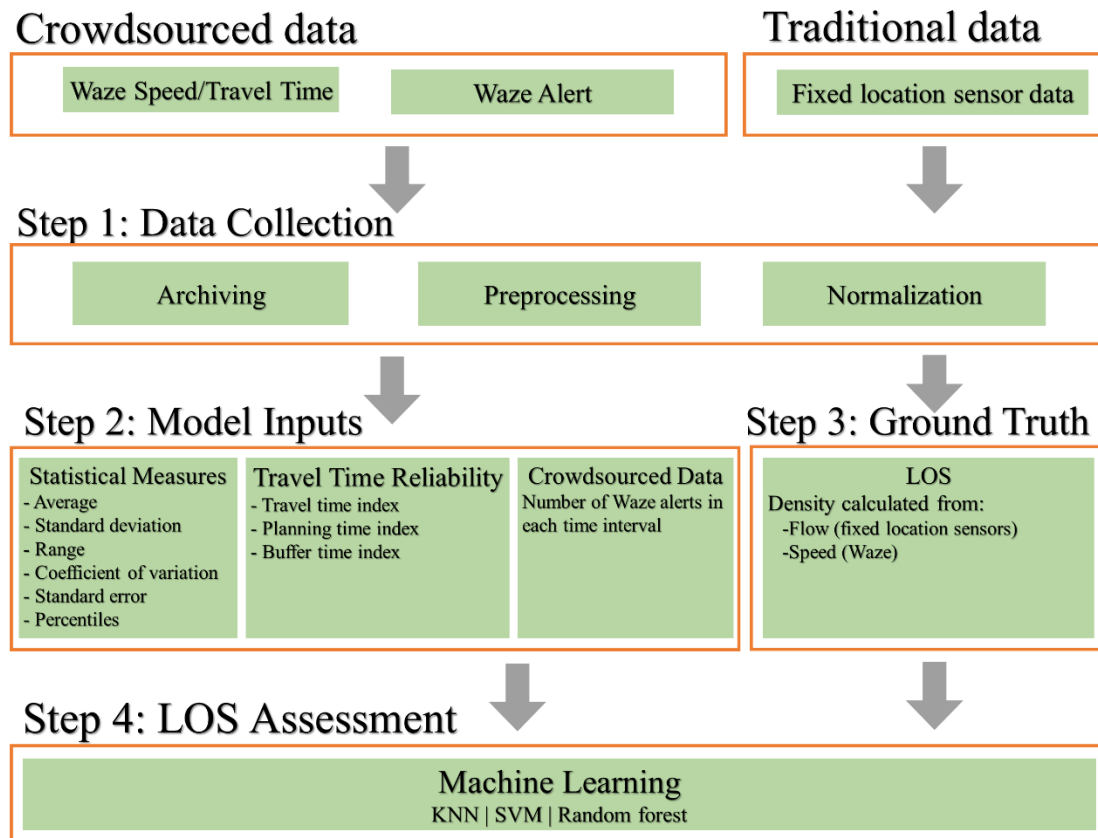


Figure 11. Framework of the proposed methodology

Basic statistical measures

Pertaining to speed variation, different statistical measures were considered. As discussed, speed variation has been considered in prior studies [63, 72, 84, 85]. All these measures were captured and measured during each period (in this study, hourly). Table 13 provides the different statistical measures used in this study.

Travel time performance measures

To analyze travel time variability for each time period, the following well-known travel time performances measures were also calculated. It should be noted that all the travel time reliability indexes were derived on a one-hour aggregation level. The travel time performance measures are as follows. Table 13 also presents the different travel time performance equations.

- **Travel Time Index (TTI)** captures the travel time variation by calculating the average travel time ratio to the free flow travel time in the segment. This index explores how travel time deviates from free flow travel time during the intended period, which is typically LOS A [68].
- **Buffer Time Index (BTI)** represents the amount of extra time that the traveler needs to be on time [68, 86, 87].
- **Planning Time Index (PTI)** calculates the ratio of the 95th percentile of travel time to the free flow travel time. A higher PTI value indicates a lower reliability and theoretically lower LOS [68, 86, 87].

Crowdsourced data

This study incorporates crowdsourced data along with speed and travel time variability. Here, the number of Waze user reports (alerts) in each period (one hour) for the study area was calculated. This number is used as an input to the final model of LOS assessment.

Step 3: Ground Truth LOS

LOS is a widely used performance measure of the quality of service on a road segment. The HCM identified six LOS categories for freeways and highways based on density and road characteristics. HCM employs traffic density as the primary measure of LOS on freeways segments [63, 72, 79, 91, 92]. Table 14 presents the density pertaining to each LOS [62]. In this study, traffic flow (from RDS sensors) and speed (from Waze) are used to calculate the hourly traffic density. The calculated density was used to obtain hourly LOS based on Table 14. The calculated LOS is used as the ground truth. The hourly input data are also labeled with ground truth values, which will be used in the LOS model in the next section.

Table 13. Model inputs equations.

Model Input	Measure	Equation	Eq. No.
Basic statistical measures of speed	Average Speed	$\bar{v} = \frac{\sum_1^n v_i}{n}$ <i>Where v_i is speed and n is number of observations in each time interval</i>	(16)
	Standard Deviation (SD)	$\sigma = \sqrt{\frac{\sum_1^n (v_i - \bar{v})^2}{n}}$	(17)
	Range	$Range(v_i) = \max(v_i) - \min(v_i)$ <i>where $i = 1, 2, \dots, n$</i>	(18)
	Coefficient of Variation (CoV)	$CoV = \frac{\sigma}{\bar{v}}$	(19)
	Standard Error (SE)	$SE = \frac{\sigma}{\sqrt{n}}$	(20)
	Percentiles (5 th , 25 th , 50 th , 75 th , 90 th)	$k^{th} \text{ percentile} = rank(\frac{k}{100}(n + 1))$ <i>Where $k = 5, 25, 50, 75, 90$</i>	(21)
	Interquartile Range (IQR)	$IQR = Q_3 - Q_1$ <i>Where Q_3 is 75th percentile and Q_1 is 25th percentile of v_i</i>	(22)
Travel time performance	Travel Time Index (TTI)	$TTI = \frac{TT_{Avg}}{TT_{free-flow}}$ <i>where TT_{Avg} is average travel time and $TT_{free-flow}$ is free flow travel time</i>	(23)
	Buffer Time Index (BTI)	$BTI = \frac{TT_{95th} - TT_{Avg}}{TT_{Avg}}$ <i>Where TT_{95th} is 95th percentile of travel time</i>	(24)
	Planning Time Index (PTI)	$PTI = \frac{TT_{95th}}{TT_{free-flow}}$	(25)
Crowdsourced data	Hourly Number of Alerts	$Count(Waze\ Alert_t^s)$ <i>Where s is the study segment and t is the time interval (hour of day)</i>	(26)

Table 14. Label of different LOS (1).

LOS	Density (vehicle/mile/lane)	Description
A	≤ 11	Free Flow
B	>11-18	Reasonably free flow
C	>18-26	Stable flow (acceptable delays)
D	>26-35	Speeds decline slightly with increasing flows
E	>35-45	Operation near or at capacity
F	>45	Forced or breakdown flow

Step 4: Machine Learning Methods

To accomplish the study objectives and estimate hourly LOS using crowdsourced data, machine learning classification methods were used. In this study, a variety of machine learning algorithms were tested. Among eight methods, the top three methods with the highest accuracy were selected and reported in this paper.

These are discussed below.

- **Random Forest (RF):** RF is an ensemble classification method that combines several random decision trees. In this method, all trees are built independently. Then, it classifies the data based on the majority of votes of all trees.
- **Support Vector Machines (SVM):** SVMs are well-known margin-based classification methods. For each class, the SVM algorithm finds the optimal support vector that provides the maximum distance to other classes. By calculating the optimal support vectors, the algorithm can identify the boundaries and classify the data.
- **K-Nearest Neighbor (KNN):** KNNs are non-parametric methods that are widely used for classification. All training data are considered in an n-dimensional feature space (n=number of input features) in this method. For each observation, the algorithm looks for the k (a predefined constant) nearest neighbors based on Euclidean distance. Then, it assigns the category based on the most frequent label of the neighbors.

Since this study implements different machine learning methods, they should be compared to find the preferred model. Classification accuracy and Cohen's kappa coefficient were used to choose the preferred model and features. Accuracy captures the ratio of correct classified predictions (LOS in this study) in comparison to ground truth. Kappa is another classification performance measure that calculates how close the classified instances are to the labeled ground truth. Kappa eliminates the correct predictions occurring by chance. Kappa is useful when the data are unbalanced due to the number of observations in each category. It should be noted that the higher accuracy and Kappa value, the better performance of the method. The accuracy and Kappa can be calculated using the following equations.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (27)$$

$$Kappa = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)} \quad (28)$$

Where, $\text{Pr}(a)$ is the ratio of correct classification or accuracy (equation 12), and $\text{Pr}(e)$ represents the probability of success due to chance.

Results

This section first provides the descriptive statistics for all input variables. Then, the machine learning model results are presented. It should be noted that R software (version 4.0.0) was used for all analyses and visualization presented in this section.

Descriptive Statistics

Table 15 presents the descriptive statistics for the hourly input data. All input measures have a range of values during different hours. For example, the hourly average speed has a range between 19.3-74.2 mph. Also, the number of Waze alerts has a range of 0-101 hourly alerts. This suggests that some of the measures need normalization to remove bias in the models. Therefore, some of the speed measures (average, maximum, minimum, and percentiles) were normalized to improve the dataset quality and prevent an imbalance bias of the dataset.

Figure 12 presents a boxplot of the number of hourly crowdsourced alerts for each time of day. It shows that typically during daytime and peak hours, there are a higher number of alerts than during night hours. Furthermore, Figure 13 shows the boxplot of the number of alerts, average speed, TTI, BTI, and PTI in each LOS category. This figure indicates that from LOS A to F, the range of the number of Waze alerts, TTI, BTI, and PTI for each LOS increase. On the other hand, the average speed decreases. Also, the range of measures in each LOS category is different. The results suggest that these features can be beneficial in describing the traffic status and LOS.

LOS Classification Model Using Machine Learning

To assess LOS, a variety of machine learning techniques were employed. Among eight tested methods, the highest accuracy methods are reported in this paper: SVM, RF, and KNN. The k-fold cross validation technique was used for all three techniques to remedy the overfitting problem, reduce the impact of unbalanced label frequencies, and maximize the use of data for both training and testing. In this technique, the datasets will be randomly divided into equal k-folds with approximately the same number of instances. One fold will be used as the validation set, and the remaining folds will be used for training. Each fold will be used once as the validation dataset. Then, the final accuracy and Kappa value will be the average of k validation results.

Here, three different models were estimated to elaborate on the impact of adding crowdsourced data in the LOS assessment accuracy. By comparing these models using different machine learning methods, the preferred model can be selected. The proposed models are described as following.

- **Model I** uses only travel time performance measures as the model inputs and shows how accurate travel time performance measures can determine LOS.

Table 15. Descriptive statistics of hourly data.

Model Input	Measure	Mean	Min.	Max.	Median	S.D.
<i>Basic statistical measures on speed</i>	Average Speed	62.5	19.3	74.2	68.8	13.5
	Speed standard deviation	4.1	0.0	24.5	2.5	3.9
	Minimum speed	54.9	11.2	73.6	65.5	16.0
	Maximum speed	69.0	24.5	90.7	73.6	12.5
	Range of speed	14.1	0.0	56.6	9.0	11.3
	CoV of speed	0.0	0.0	0.5	0.0	0.1
	SE of speed	0.5	0.0	3.1	0.3	0.5
	25 th percentile	60.2	16.1	71.1	68.2	15.2
	50 th percentile	62.7	18.1	73.6	68.8	14.2
	75 th percentile	65.3	22.7	75.2	69.3	13.1
	90 th percentile	66.9	24.5	78.6	73.6	12.6
	IQR	5.1	0.0	51.9	3.8	7.7
<i>Travel time performance</i>	TTI	1.2	1.0	3.9	1.0	0.5
	BTI	0.2	-0.4	3.4	0.0	0.4
	PTI	1.4	1.0	5.8	1.1	0.7
<i>Crowdsourced data</i>	Number of Waze Alerts	9.0	0.0	101.0	4.0	20.0

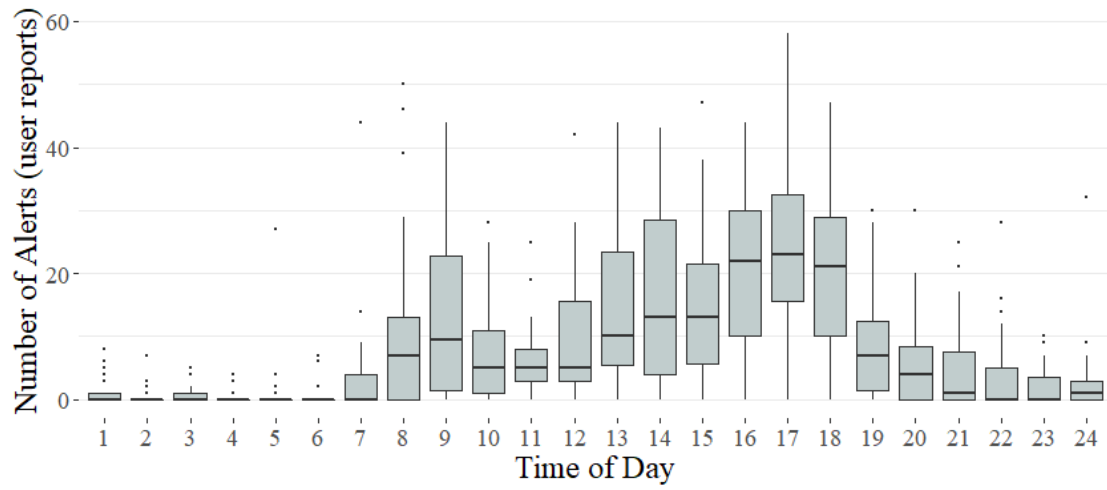


Figure 12. Boxplot of number of alerts in each time period (hours).

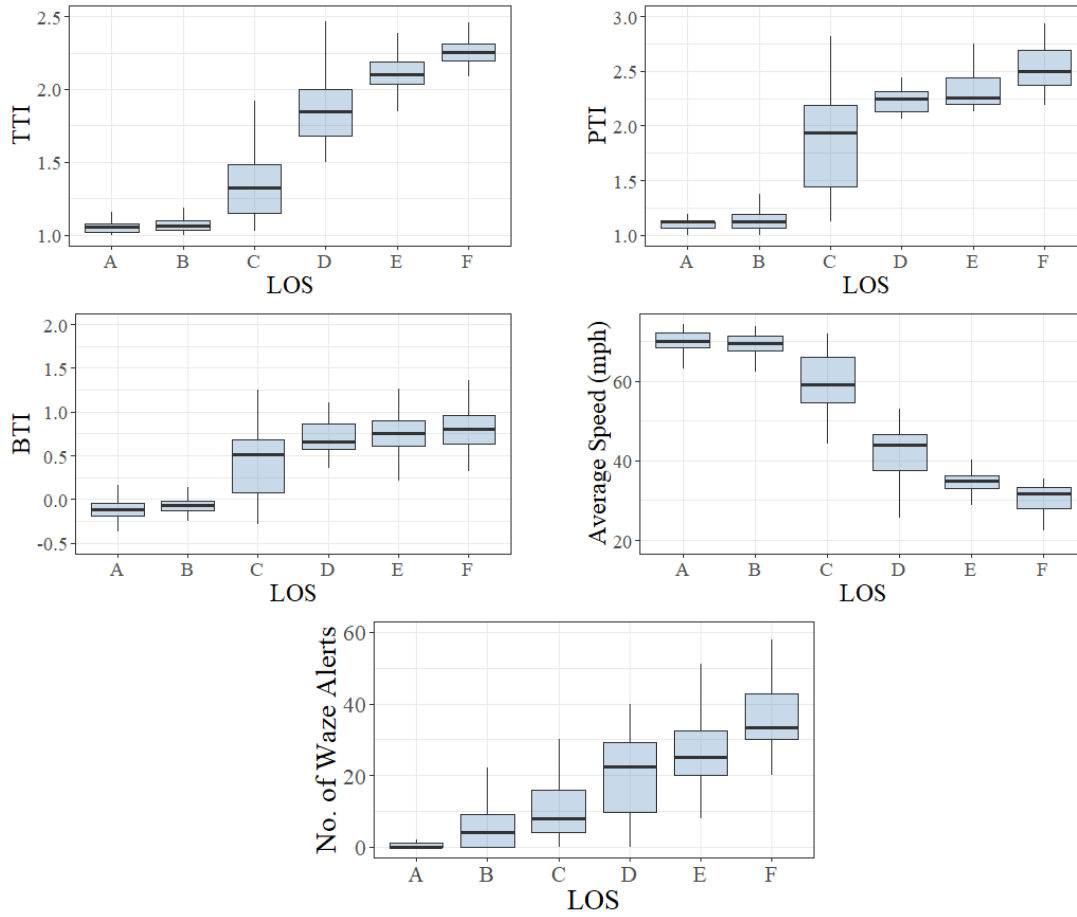


Figure 13. Boxplot of different measures in each LOS category.

- **Model II** uses travel time performance and statistical measures as the.
- **Model III** incorporates crowdsourced Waze alerts and uses all three types of inputs.

Next, 10-fold cross validation was performed for the three models. Figure 14 displays the result of cross validation for the different machine learning techniques for each model. It can be inferred that for most methods, adding statistical measure (Model II) improves the accuracy and Kappa in comparison to Model I. Also, adding crowdsourced data to the input measures (Model III) increased the performance of all methods. Based on this step, it can be concluded that crowdsourced data improved the LOS classification performance. All three types of input measures were used in the final model.

In order to assess the sensitivity to the number of folds in cross validation, the values of k (3,5,10) were selected. Table 16 compares selected classification methods with 3, 5, and 10 cross validation folds.

The result of this study shows that machine learning techniques can determine LOS. The RF accuracy of method with 3, 5, and 10 cross validation was 0.91, 0.93, and 0.92, respectively. Also, all Kappa values for RF was above 0.8, which is acceptable. Among the selected machine learning techniques, the RF performed the best result using model III inputs.

Sensitivity Analysis

The sensitivity of the preferred model (Model III using RF) to different hours of the day was investigated. In the RF model, the corresponding accuracy of each hour of the day was calculated. The accuracy values have a range between 0.92 to 0.94 during 24 hours of the day. The result highlights that the LOS estimation model is not dependent on the time of day. This method can be used for both peak hours and non-peak hours to estimate traffic state and LOS. Additionally, the classification accuracy for each ground truth LOS was calculated based on the confusion matrix. Similar to hourly analysis, each LOS category's accuracy did not deviate from the total accuracy (0.93). This suggests that the proposed method results are not biased due to the frequency of each LOS category.

Variable Importance

This study suggests that crowdsourced data can improve LOS classification accuracy. Accordingly, variable importance analysis was performed based on the preferred RF model (Figure 15). The mean decrease in Gini index was computed from the RF model. A higher value of this index indicates higher variable importance. Average speed, number of crowdsourced alerts, TTI, BTI, minimum speed, PTI, and standard deviation (SD) of speed are the most important variables in determining the LOS, respectively. This result is consistent with previous studies in the literature that employ speed and travel time reliability measures in determining LOS thresholds. However, the number of crowdsourced alerts seems to impact LOS prediction accuracy significantly more than TTI, BTI, and TTI.

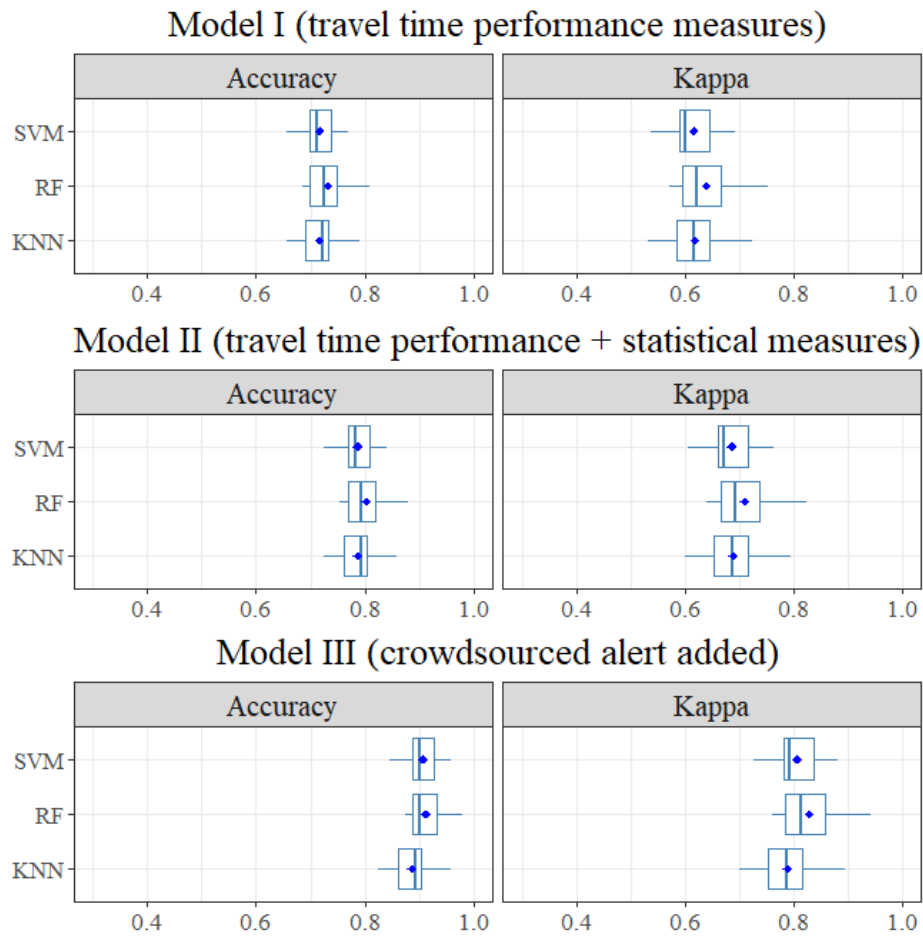


Figure 14. Different model accuracy and Kappa values.

Table 16. Summary of classification methods with 3, 5, 10-fold cross validations.

Classifier	3-fold cross validation		5-fold cross validation		10-fold cross validation	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
SVM	0.91	0.81	0.91	0.81	0.90	0.79
RF	0.91	0.82	0.93	0.83	0.92	0.83
KNN	0.88	0.77	0.89	0.79	0.88	0.76

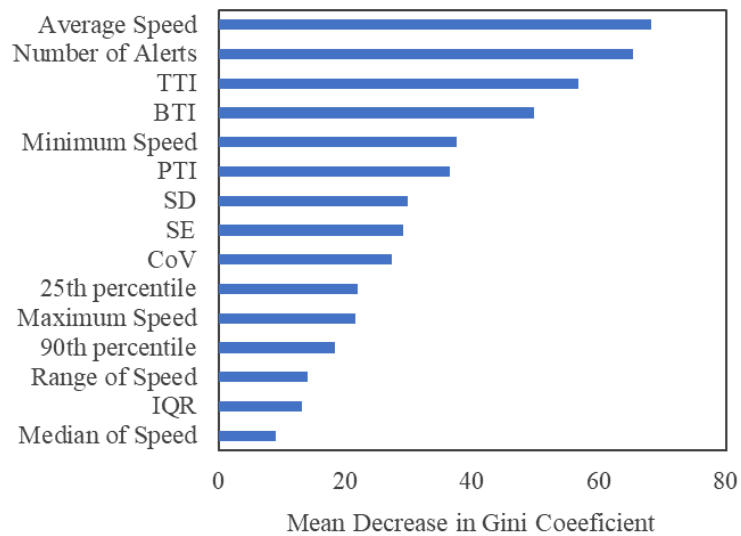


Figure 15. Variable importance plot.

Conclusions

Crowdsourced data availability is increasing rapidly, and machine learning offers the opportunity to analyze it. This study proposed a new methodology to incorporate crowdsourced data in LOS assessment. The method was applied to a 1.5-mile segment of freeway on I-40 in Knoxville, Tennessee. Crowdsourced data from Waze were collected, and three categories of input measures (basic statistical measures, travel time reliability, and Waze crowdsourced alerts) were calculated. The methodology performed machine learning techniques to classify LOS on an hourly basis. Also, data collected from fixed location RDS sensors were used to calculate the traffic density and estimate LOS ground truth using HCM density thresholds.

The results of this chapter highlight that crowdsourced data and machine learning techniques can be used to estimate LOS. The results revealed that using crowdsourced alerts as an input can significantly improve model accuracy (about 10%). Also, the RF method shows the highest performance among other classification methods (accuracy=0.93 and Kappa=0.83). In this method, the LOS estimation accuracy value is relatively consistent among different times of days and LOS categories. Sensitivity analysis confirms that the accuracy of this methodology does not deviate in traffic peak-hours or non-peak hours. The results also suggest that average speed, number of alerts, TTI, and BTI are the most important variables in determining LOS.

This method helps to explore traffic status on freeways without relying on fixed location sensors, time of day, or day of the week. The proposed method has the potential to be applied to different freeways segments to assess LOS. This method does not need fixed location sensors, potentially resulting in less implementation and maintenance costs. Transportation agencies and DOTs can utilize this method for traffic operations purposes. This method can also analyze freeways traffic in locations outside of urban areas with no fixed location sensor. It benefits from crowdsourced data and can be applied for different time periods such as hourly, daily, and traffic peak hours.

The proposed method used crowdsourced data, travel time variability, and speed statistics measures to estimate LOS. The travel time reliability and speed statistics measures captured the temporal variability of speed and travel time. In future research, spatial variation can be used as an input variable for LOS assessment. The speed deviation from upstream and downstream segments can also be addressed in LOS estimation. Additionally, the reliability of the proposed method during severe events such as the COVID-19 outbreak can be evaluated in future research. Finally, this chapter used the Waze alert counts regardless of the event type (jam, accident, hazards). In the future, the impact of each type of event on traffic status and LOS should be evaluated.

CHAPTER III
SPATIOTEMPORAL LANE-BLOCKING TRAFFIC INCIDENTS
DETECTION USING MULTISOURCE CROWDSOURCED DATA

Abstract

Automatic traffic detection (AID) is important in traffic management, which can reduce non-recurrent congestion caused by incidents, decrease traffic delay, and decrease the probability of secondary and subsequent crashes. This chapter proposed a multisource traffic lane blocking incident detection algorithm. Crowdsourced data are known to be the emerging, low-cost solution to potentially improve safety and operations. This study incorporates crowdsourced data (speed and user alerts) provided by Waze to propose an AID algorithm. To implement this algorithm, first, an anomaly detection algorithm called inter-quartile range (IQR) is used to set speed threshold based on historical data. This threshold identifies the abnormality in the speed pattern as potential incident. Then, the algorithm matches the Waze user accident alerts with IQR potential incidents in a proposed clustering algorithm called weighted spatiotemporal density-based clustering of applications with noise (WST-DBSCAN) to identify incidents. The proposed algorithm set a predefined weight to Waze accident alerts. To capture the performance of proposed method, a 30-miles segment on Interstate 40 (I-40) highway at westbound between mile marker 363 to 393 was selected, and six month of speed (1-min resolution) and Waze accident alerts were selected. The results revealed that the proposed algorithm can improve the incident detection performance significantly. Also, the idea of using multisource data in AID increases the detection performance rather than using a single data source. The results of this research provide transportation management centers an AID method based on crowdsourced data on different freeways segment (urban or rural) with high detection rate and low false alarm rate, regardless of the availability of traditional fixed location sensors.

Introduction

With the growth in the number of vehicles, traffic incidents are increasing. Meanwhile, uninterrupted roads (highways and freeways) are more prone to have incidents due to higher speeds in those road segments. The traffic incidents affect the transportation systems in both safety and operation performance sides. Automatic incident detection (AID) has been developed as an essential tool for traffic operation performance, incident management, and safety. Studies show that the nationwide cost contributed by congestion in 2014 was estimated to be \$160 billion, or \$960 per commuter, for the top 471 urban areas of the United States. This study also estimated 6.9 billion hours of wasted time and 3.1 billion gallons of wasted fuel due to congestion [93].

Over the past decade, different studies have developed data-driven algorithm and statistical models of AID. The methods incorporate various

approaches such as statistical models, time series analysis, anomaly detection, clustering algorithms, pattern recognition, artificial intelligence, and data mining. Most of AID algorithms analyze historical data to detect the traffic patterns (speed, flow, density) for different time/location. If the real-time observed data do not follow the learned patterns, these algorithms detect incidents.

Freeways congestion is typically classified in two categories: recurrent and non-recurrent congestion. Recurrent congestion refers to the traffic pattern that occurs in a daily/weekly basis such as morning and afternoon peak. However, non-recurrent congestion happens when there is an incident such as abandoned vehicle or vehicle accident. Identifying these two main patterns is a major challenge for AID developers which can improve false alarm.

Fixed-location sensors are the main method of collecting traffic data such as speed, flow and density that are used in AID. License Plate Recognition (LPR) systems, remote traffic microwave sensors (RTMSs), magnetic and laser sensors, loop detectors and video recordings are examples of these fixed location sensors [9-11]. Increase in smartphone use has led to emergence of modern data collection. The newer methods of data collection are based on probe-vehicles and floating cars (smartphones, cellular networks, Bluetooth sensors, Wi-Fi, and connected vehicles) [4, 6, 12, 14, 16-18]. Deploying smartphones by users to contribute their information via smartphone, social media, or the internet is called crowdsourced data. Crowdsourced data and probe-vehicle as novel data source for transportation analysis are advantageous over fixed-location data collection methods are since they are (I) lower in cost, (II) higher in network coverage, and (III) lower in maintenance fees [13, 79]. These new tools emerged in the last few years has enabled researchers proposal of innovative ideas and solutions [2]. Acquisition of crowdsourced data is possible through partnership between the Department of Transportations (DOTs) in many cities such as Waze and INRIX [8]. In this partnership, information such as speed, travel time, delay, incident, hazards, hazards, severe weather, congestion is shared by the data provider. Tennessee Department of Transportation (TDOT) has partnered with Waze navigation app through Waze for Cities (WFC) program to provide crowdsourced data (e.g. Speed, travel time, Incidents, traffic jams, hazards).

This study chapter aims to propose a spatiotemporal multisource lane-blocking incident detection framework with improved performance (a high detection rate and low false alarm rate). This chapter will clarify the advantages of using crowdsourced data in AID applications. This framework incorporates traffic speed pattern with crowdsourced user reports. To this end, a real-time anomaly detection methodology will be used to detect potential incidents based on speed. By using the real-time and user reports, a method called weighted spatiotemporal density-based clustering of applications with noise (WST-DBSCAN) would be proposed to detect incidents.

Literature Review

This part reviews relevant literatures pertaining to this chapter in the following order. First, existing AID algorithm are summarized. Then, statistical algorithms are discussed. The next parts will review crowdsourcing in AID and discuss the research gaps.

Existing Algorithms

Most of existing and traditional AID algorithms are based on fixed-location sensors data and probe vehicles in more recent decade [94]. Comparing algorithms are widely used in AID methodologies. Meanwhile, some algorithm compares the current traffic parameter values with historical observed value with similar condition such as time of day and day of the week. On the other hand, some studies focused on comparing present traffic parameters with immediate previous time window [95-97]. Overall, different AID algorithms have been proposed using pattern recognition [98, 99], traffic model [100], artificial intelligence [101, 102], data mining [103], statistical models [96, 97], and clustering [104].

Statistical Algorithms

Different statistical algorithms have been performed to detect incidents in freeways. Outlier detection method, such as standard normal deviate (SND), is known as one of the popular statistical methods of AID. SND uses historical data (mean and standard deviation) to define a threshold. If the observed parameter passes the defined threshold, the AID will trig an incident alarm [105]. More recent studies modified the outlier detection method to reduce false alarm and increase algorithm performance [96, 97].

Crowdsourcing in AID

Recently with the growth of internet applications, the newly emerged data sources, social media, has gained more attention. Studies presented the capabilities of social media data in transportation applications [106-108]. AID is one of the applications of social media data in transportation. Zhang et al. employed deep learning and accident-related tweets to detect traffic accident. They showed using tweets improved the accuracy of algorithm [108]. Gu et al. also used classification and text mining to detect incidents [109]. Zhang et al. also used clustering algorithm to improve incident detection accuracy [110]. Salas et al. also used Natural Language Processing (NLP) techniques with a Support Vector Machine algorithm (SVM) for AID [111].

Some studies used Waze data in the incident detection algorithm [112]. Waze provides speed and travel time data by collecting and analyzing the Waze application users' location. A more recent study evaluated Waze Incident report data reliability and coverage [8]. Also, the coverage of the Waze incident report [47] and Waze speed/travel time computation [48, 67] seems to be acceptable.

Identifying the Gap

As discussed, Crowdsourced data has not been employed extensively in AID studies. This study addresses a gap in AID by incorporating statistical outlier detection method with crowdsourced user reports (Waze alerts). This study proposed a new methodology called WST-DBSCAN.

Data

This part describes the crowdsourced data set in this study. This study uses Waze speed and Waze alerts data for the AID purpose. In the following, the Waze speed/travel time and Waze crowdsourced alerts will be presented. Then, the study area will be introduced.

Waze Speed and Travel Time Data

the Waze for Cities (WFC) program collects users' kinematic information to calculate and report speed and travel time used in the proposed algorithm. In the event of no user passing through the segment in the data collecting time interval, a historical speed and travel time is reported. This data is available in specific road segments through a tool called “traffic view”

(<https://www.waze.com/trafficview>), implemented by Waze. Using this tool, authorized users like transportation agencies and DOTs can specify a list of road segments and add links based on their priority or needs to the Watch-list.

The real-time travel time/speed data for the predefined road segments are available on a one-minute level. Apart from Authorized users using the real-time data using the traffic view, this data can be archived in JavaScript Object Notation (JSON) format for further analysis. In the archived JSON file, travel time, segment length, and geospatial information for all predefined segments is stored for each time interval.

Waze Crowdsourced Alert Data

Another valuable crowdsourced data is user report data provided by Waze that is referred to as alerts. These alerts reported by users include accidents (major or minor), traffic jams (heavy, moderate, standstill, or light), hazards (severe weather, stopped cars, road potholes), construction, and closed roads that can be used in different research of analysis such as incident detection, hot spot clustering, and end of queue prediction. Waze app users can report predefined incident types or verify existing reports on the road in real time while travelling, all users' reports are then shared through the WFC program. Each alert is differentiated from another one by its unique ID, time, spatial coordinates, direction, reliability, and confidence level of the reported alert. This study has access to real-time Waze alerts for Tennessee State that is available to Waze partners, archived in Extensible Markup Language (XML) format.

Study Area

To perform the proposed incident detection algorithm, a study area was designated in Knoxville, Tennessee. A 30-miles segment on Interstate 40 (I-40) highway at westbound between mile marker 363 to 393 was selected (Figure 16). This segment includes more than 47 links (each point on the map represent median of a link in the study segment) The Waze and LocatIM data used in this study covers six months of data from July 1, 2019 to December 31, 2019 in the selected segment.

Ground Truth (LocatIM data)

TDOT provide traffic events records on interstate highways, including location, time, number of blocked lanes, and incident duration. This data is registered and validated by TDOT operators. This data is used as ground truth to evaluate proposed method accuracy. Table 17 provides the monthly and total lane blocking incident occurred in study segment based on LocatIM registered data. In total 203 lane blocking has been observed and registered by TDOT operators. Multivehicle crash is the most frequent lane blocking incident type.

Methodology

This study incorporates crowdsourced data and traffic speed pattern to detect an incident. This algorithm includes four main steps discussed below. This section provides more details about the proposed algorithm of this study. As shown in the framework of the study (Figure 17), the different steps of the proposed method are:

- **Step 1:** Extract heatmap based on Inter-quartile range (IQR) method for each day based on last four weeks speed data.
- **Step 2:** Detect potential incident based on IQR heatmap and traffic speed.
- **Step 3:** Match Waze Alerts (Accidents) with corresponding mile marker and time.
- **Step 4:** Employ WSD-DBSCAN method for incident detection.

Step1: Inter-quartile Range (IQR) Threshold Heatmap

This study uses outlier detection to evaluate traffic speed pattern and detect potential incidents. Inter-quartile range (IQR) is a robust outlier detection method in statistics. Here, we adapt a modified version of IQR from Chakraborty et al.[96]. For each day, last four weeks' corresponding day of the week should be extracted as reference speed profile to detect anomalies. Then, each day should be divided in 15-minute intervals. The speed threshold for each time interval can be calculated as follows.

$$\tau_s^{t,d} = \text{Min}[45, (M_s^{t,d} - c_{IQR} \times IQR_s^{t,d})] \quad (29)$$

$$IQR = 75^{th} \text{ percentile pf speed} - 25^{th} \text{ percentile of speed} \quad (30)$$



Figure 16. Study area on Interstate 40 (I-40) highway at westbound between mile marker 363 to 393.

Table 17. Lane blocking incident based on TDOT LocateIM.

Incident Type	Month						Total of six month
	Jul	Aug	Sep	Oct	Nov	Dec	
Abandoned Vehicle	0	1	0	3	0	1	5
Debris	5	6	1	0	0	1	13
Disabled Vehicle	4	6	8	5	7	6	36
Multivehicle Crash	23	14	16	23	31	21	128
Single Vehicle Crash	6	2	2	2	5	1	18
Vehicle Fire	0	1	0	0	1	1	3
Total Incidents	38	30	27	33	44	31	203

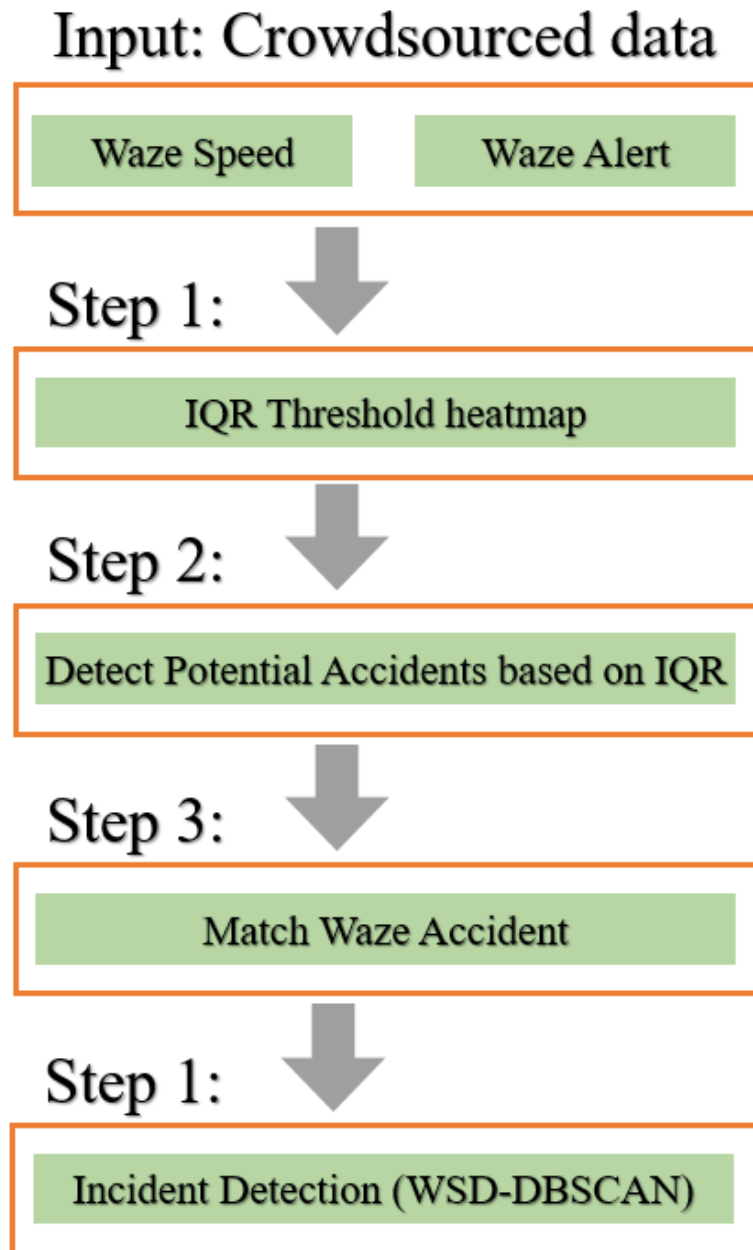


Figure 17. Framework of chapter III.

Where,

t : 15 – minute time window

d : day of week

s : Segment

M : Median of speed

c_{IQR} : constant value

To reduce the false alarms due to fluctuation during non-peak hours, the threshold is calculated by minimum of $(M_s^{t,d} - c_{IQR} \times IQR_s^{t,d})$ and 45 mph. In this step for each time of day and each segment a threshold will be assigned. The calculated IQR thresholds can be used for plotting a daily heatmap. As an example, for a weekend and a weekday the IQR threshold heatmap are presented in Figure 18. For a weekend, when heavy traffic does not occur, the threshold for almost all day for all segments are 45 mph. However, for a typical weekday, in some segments the peak hour thresholds are less than 45.

Step 2: Detect IQR Potential Incidents

This step compares travel speed with the calculated thresholds. For each time of day and location if the speed is less than corresponding threshold in IQR heatmap, the algorithm will add a potential incident for that mile marker and time of day. As an example, using the above heatmap, if the speed in a mile marker 385 at time 4:00 PM is less than 25, the algorithm considers this time and location as a potential incident.

Step3: Match Waze Alert

In the next step, the algorithm matches the Waze incidents alerts to the corresponding time and mile marker location.

Step 4: Incident Detection Using WST-DBSCAN

Spatiotemporal DBSCAN (ST-DBSCAN) was proposed to extend the dimensions of conventional DBSCAN using temporal dimension [113]. This study uses two different data sources as input of data clustering: (1) IQR potential incidents and (2) Waze accident Alerts. Both inputs include spatial and temporal dimensions. To detect the incident, the proposed algorithm finds the clusters in the input data. When a cluster is detected, the algorithm notifies a lane-blocking incident.

This study uses Waze accident alert to reduce the false detection. Since two inputs are coming from two different types of data, this study proposes a Weighted Spatiotemporal DBSCAN (WST-DBSCAN) method to cluster IQR alerts and Waze incident alerts. This algorithm adds a weight to the Waze alert. The WST-DBSCAN algorithm is defined as Figure 19.

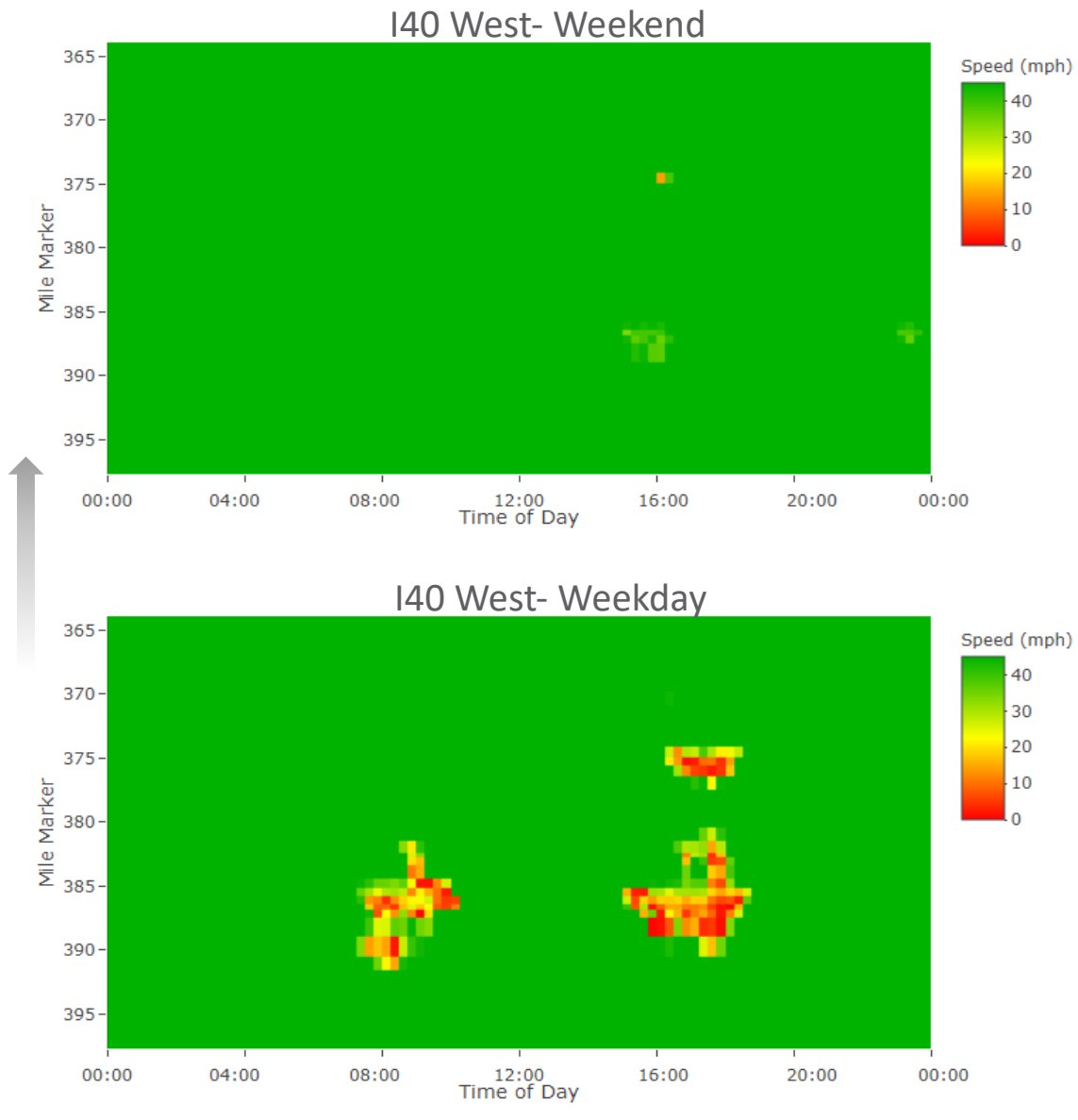


Figure 18. Daily Threshold Heatmap.

```

WST-DBSCAN( $D, \varepsilon_t, \varepsilon_s, \omega_i, minPts$ )
   $C \leftarrow 0$ 
  for each data point  $d$  in dataset  $D$  do
    If  $d$  is visited, then
      Continue to next  $d$ 
    end
    else
      mark  $d$  as visited
       $\varepsilon_s$ -neighbors = SpatialNeighbors( $D, d, \varepsilon_s$ )
       $\varepsilon_t$ -neighbors = TemporalNeighbors( $D, d, \varepsilon_s$ )
       $N = \varepsilon_s$ -neighbors  $\cap$   $\varepsilon_t$ -neighbors
       $n = \sum_{i=1}^{sizeof(N)} (d_i \in N) \times \omega_i$ 
      If  $n < minPts$ :
         $mark(d) \leftarrow Noise$ 
      end
      else
         $N \leftarrow N \cup d$ 
        Split  $N$  in  $Labeled\_N$  and  $Unlabeled\_N$ 
         $N_{clusters} \leftarrow unique(Labeled\_N)$ 
        If  $N_{clusters} \equiv 0$ 
           $C \leftarrow C + 1$ 
           $label(N) \leftarrow C$ 
        endif
        if  $N_{clusters} \equiv 1$ 
           $label(Unlabeled\_N) \leftarrow C$ 
        end
        if  $N_{clusters} > 1$ 
          Merge  $C$  with every cluster point in  $N$  belongs
        end
      end
    end
  end
end WST-DBSCAN

```

Figure 19. WST-DBCAN implementation.

In WST-DBSCAN four main parameters are used: ε_t , ε_s , ω_i , $minPts$. Here, ε_t and ε_s specify the temporal and spatial distance thresholds. Two neighbor points can be considered in a cluster only if their distance is within ε_t and ε_s . ω_i is the Waze accident alert weight in the algorithm. Using this weight parameter, the Waze accidents will be considered ω_i times of IQR potential incident in the algorithm. $minPts$ also specifies the minimum number of neighbor points needed for each point to be considered in the core of a cluster. Any point that is not cored in a cluster will be assigned as noise.

As the WST-DBSCAN identifies a cluster based on IQR potential incidents and Waze accident alerts, the proposed algorithm will report a lane-blocking algorithm. Going forward in time dimension, the WST-DBSCAN will continue and add more point to the identified cluster. In each time interval, this algorithm can identify the end of queue caused by incident. This help traffic management centers (TMC) to manage the incident.

Performance Measurement

To evaluate the proposed incident detection algorithm, the case study results (detected incidents) are compared to LocateIM data. Detection Rate (DR), False Alarm Rate (FAR), Mean Time to Detection (MTTD) are three measures used to qualify the performance of the algorithm based on previous studies [96].

- **Detection Rate (DR):** captures the ratio of number of true incidents detected to the total number of incidents occurred.

$$DR = \frac{\text{Number of true detection}}{\text{Total number of incidents}} \times 100 \quad (31)$$

- **False Alarm Rate (FAR):** calculates the ratio of total false alarm reported to total number of algorithm application [8].

$$FAR = \frac{\text{Number of flase alarm cases}}{\text{Total number of algorithm application cases}} \times 100 \quad (32)$$

- **Mean Time to Detection (MTTD):** considers the latency of incident detection algorithm.

$$MTTD = \frac{1}{n} \sum_{i=1}^n (T_{detection} - T_{incident}) \quad (33)$$

Here, n represent the number of detected incidents. Also, $T_{detection}$ and $T_{incident}$ are the time when the incident is detected and incident start time, respectively.

Results

As mentioned earlier, a 30-mile segment was selected to evaluate the proposed algorithm. First step of the proposed algorithm was determination of speed threshold heatmap for each day based on historical data. Next the observed speed of each segment was compared to IQR speed threshold for that segment based on historical data. Each time the observed speed was smaller than IQR threshold, the algorithm will assign this time and location as a potential lane blocking incident. As an example, for two different days in mile marker 386 the observed speed and threshold speed are compared (Figure 20). In case (a) at 5:30 PM the speed is smaller than threshold. The same thing happened in case (b) at 12:00 to 2:00 PM. IQR potential incidents in both cases were used as input of WST-DBSCAN, but case (b) was identified as noise, not a real incident.

Before deploying the algorithm, we need to select the c_{IQR} value in IQD algorithm. To determine the c_{IQR} , the proposed algorithm was applied on data with pre-selected $\varepsilon_t, \varepsilon_s, \omega_i, minPts$ and different c_{IQR} values. Figure 21 shows the variation of DR and FAR for different IQR threshold constant and number of weeks of historical data. The results showed that $c_{IQR} = 2$ and 4 weeks of historical data is the optimum numbers with the highest DR and lowest FAR. Finally, to identify the improvements, the proposed algorithm was compared to following algorithm.

- ST-DBSCAN using only IQR potential incidents from Waze Speed data
- ST-DBSCAN using only IQR potential incidents from RDS Speed
- Standard normal deviates (SND) algorithm using Waze Speed Data [96]
- Maximum absolute deviation (MAD) using Waze Speed Data [96]
- Modified IQR algorithm using Waze Speed Data [96]
- ST-DBSCAN using Waze Alarm (Accident and Jam Report)

Figure 22 presents the comparison of the proposed method with other algorithms mentioned above using six months of data and 4 weeks of last corresponding historical speed data. Here, the x-y dimensions represent the FAR and DR values, and the label of each point represents the value of MTTD. The proposed method has the highest DR (94%) and lowest FAR (0.19%) and MTTD (4.8 min) among other algorithms. It should be noted that incorporating speed (IQR potential incidents) with crowdsourced data increased the performance of algorithm significantly rather than using a single data source as input of ST-DBSCAN. Also, it can be shown that using speed RDS sensor data instead of Waze speed have similar performance in the algorithm. Since the proposed algorithm consider spatiotemporal properties in incident detection, the MTTD is lower than IQR, SND, and MAD.

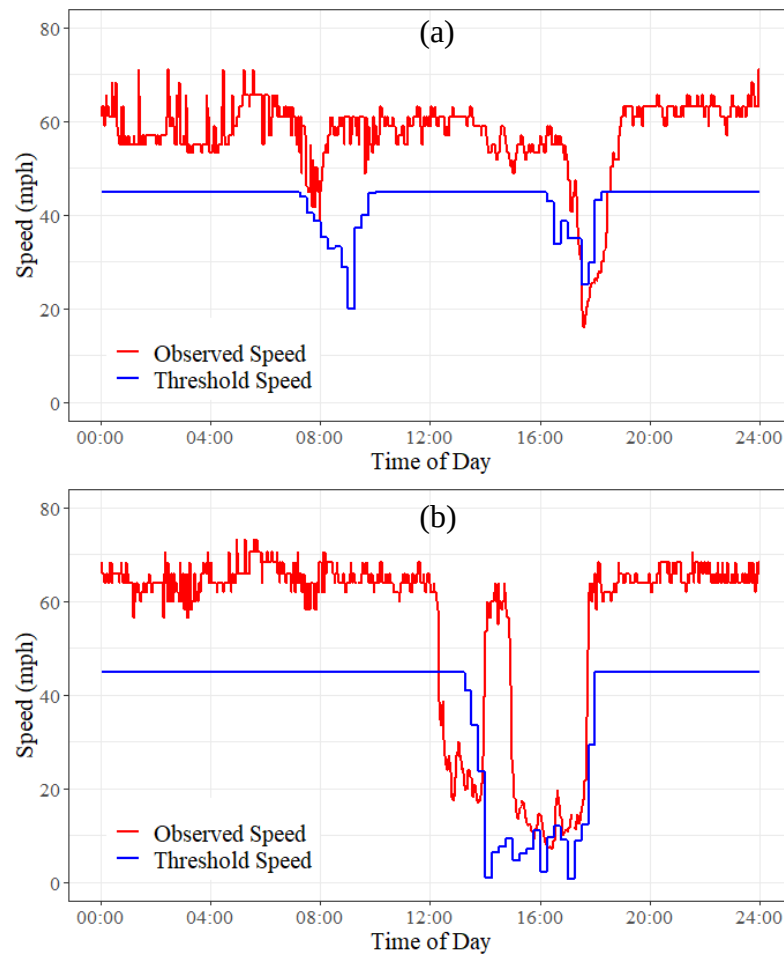


Figure 20. Comparing observed speed and IQR threshold.

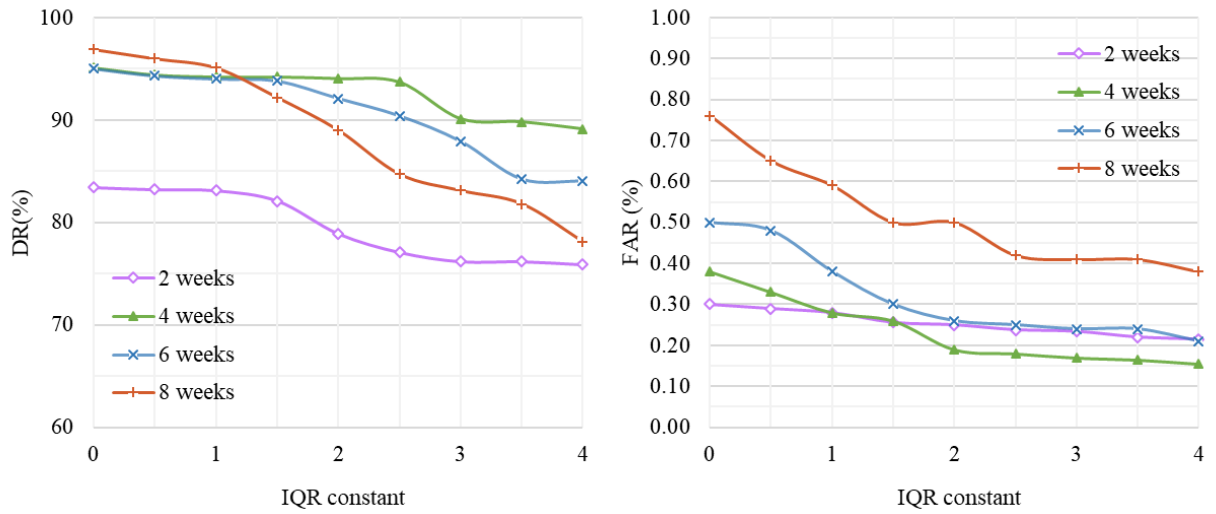


Figure 21. Variation of DR and FAR for different IQR threshold constant and number of weeks of historical data

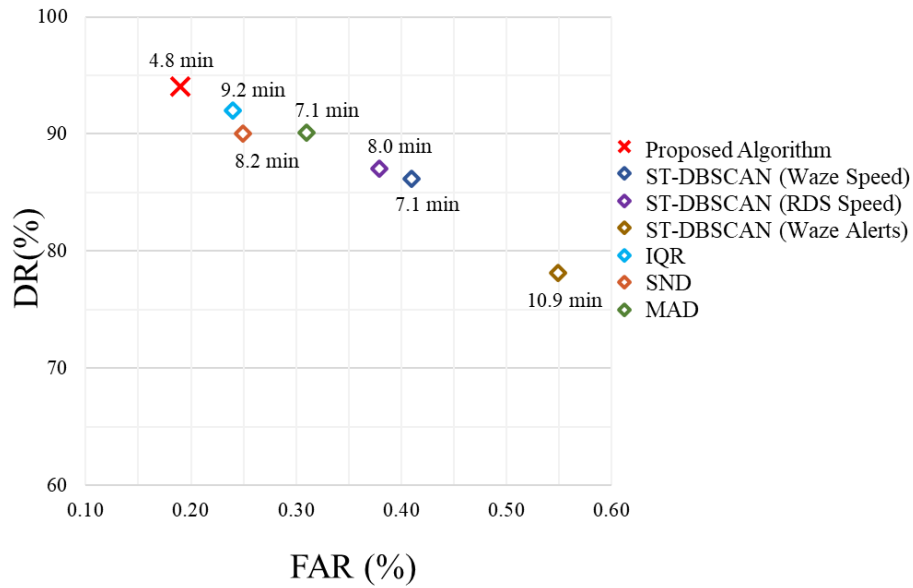


Figure 22. Comparison of proposed algorithm with other algorithms.

Conclusion

In the era of smartphones, crowdsourced data is a newly available tool with growing applications, and Machine learning makes analyzing of crowdsourced data easier. This study proposed a new methodology to incorporate crowdsourced data in incident detection. Automatic traffic incident detection play an important role in traffic management, reduction of congestion caused by incidents, and avoiding secondary crashes. This chapter proposed an incident algorithm that incorporates multisource crowdsourced data to identify traffic lane-blocking incident. The proposed algorithm uses two main data sources: (1) traffic speed to calculate abnormality in traffic pattern known as potential traffic incidents, (2) Waze user accident alert. A weighted spatiotemporal DBSCAN is then proposed to cluster the two data sources point as incidents.

The algorithm was applied to a 30-mile segment (including 47 links) of freeway on I-40 West in Knoxville, Tennessee. Six months (July 1, 2019 to January 31, 2019) of speed and incident crowdsourced data from Waze were collected. TDOT incident records (LocateIM data) was used as ground truth of the study. The records showed 203 lane blocking incident cases.

The proposed algorithm was then compared with single data source (speed or crowdsourced data) implementation of the algorithm. The algorithm was also compared with some known algorithms in the literature (SND, MAD, and IQR). Three performance measurements (DR, FAR, and MTTD) were selected regarding the comparison. The results of this study highlights that proposed algorithm revealed the highest performance among other algorithms in all three performance measurements. The results showed that using multisource (speed pattern and crowdsourced accident alerts) as an input can significantly improve detection performance.

CHAPTER IV
DYNAMICS OF MOBILITY AMID COVID-19 PANDEMIC –
GEOSPATIAL APPROACH

A version of this chapter was originally presented by Nima Hoseinzadeh, Yangsong Gu, Hairuolong Zhang, Dr. Lee Han, Dr. Hyun Kim, and Philip B. Freeze at Transportation Research Board (TRB) 2021 annual meeting and accepted for publication at Transportation Research Record (TRR)

Abstract

The year 2020 has marked the spread of a global pandemic, COVID-19, challenging many aspects of our daily lives. Different organizations have been involved in controlling this outbreak. The social distancing intervention is deemed to be the most effective policy in reducing face-to-face contacts and slowing down the rate of infections. The stay-at-home and shelter-in-place orders have been conducted in different states and cities, affecting daily traffic patterns. Social distancing interventions and fear of the disease result in a traffic decline in cities and counties. However, after stay-at-home orders end and some public places reopen, traffic gradually started reverting to pre-pandemic levels. It can be shown that counties have diverse patterns in the decline and recovery phases. This chapter analyzes county-level mobility change after the pandemic, explores the contributing factors, and identify possible spatial heterogeneity. To this end, 95 counties in Tennessee have been selected as the study area to perform geographically weighted regressions (GWR) models. The results show that density on non-freeway roads, median household income, percent of unemployment, population density, percent of people over age 65, percent of people under age 18, percent of work from home, and mean time to work are significantly correlated with vehicle mile traveled (VMT) change magnitude in both decline and recovery phases. Also, the GWR estimation captures the spatial heterogeneity and local variation in coefficients among counties. Finally, the results imply that recovery phase could be predicted depending on the identified spatial attributes

Introduction

The world faced a significant challenge in 2020 from the coronavirus disease 2019 (COVID-19) pandemic [114]. COVID-19 was first detected in December 2019 in Wuhan, China. Human to human transmission of this virus was confirmed in January 2020 [115]. By November 15, 2020, COVID-19 resulted in over 54.8 million confirmed cases and over 660,000 death globally, increasing daily. Among all 213 countries and territories with confirmed cases, the United States reported the highest number of confirmed cases (over 11.3 million) and deaths (over 251,800) caused by this pandemic by November 15, 2020 [116].

Centers for Disease Control and Prevention (CDC) announced the first case of COVID-19 in the United States on January 22, 2020 [117]. Because of the absence of a vaccine and effective pharmacological treatments, social

distancing became the primary policy of states and local governments in the United States to mitigate the outbreak. The CDC recommended these behaviors for social distancing: stay six feet away from others, do not gather in groups, and avoid crowded public places [19]. Overall, the social distancing policy includes travel restrictions, border closures, limited public events, limited group gatherings, encouraging work from home, and closure of schools and universities. The social distancing policy includes stay-at-home and shelter-in-place orders.

Studies show that a variety of public health policies and interventions can change people's behavior [118, 119]. Previous studies show that the spread of the pandemic and contagious disease is spatially distributed. Understanding the spatial pattern of a pandemic can help manage the pandemic [120-122]. Meanwhile, different recent studies have explored the changes of mobility caused by COVID-19 and the effectiveness of social distancing policies. Some recent studies in China evaluated the association between COVID-19 growth and mobility. They show that social distancing drastically decreases its growth rate [123, 124]. Some other studies show that travel restriction policies delay the disease spread [125, 126]. In this regard, a study of 14 different countries show that a significant reduction in mobility is required to reduce the growth of such an epidemic disease [127].

In terms of traffic and mobility reduction, a study on major metropolitan areas in the U.S. identifies a 50% reduction in number of people's commute to work after the outbreak and social distancing implementation [128]. Similarly, a study in Italy showed a decrease in traffic after its national lockdown from the outbreak [129]. Declaring a state of emergency also caused around 10% reduction in time spent away from residence [130]. As counter studies, some initial analyses and interpretations indicate that changes people have made to their mobility during the pandemic are not directly associated with governmental orders [131, 132]. Some studies also show that stay-at-home orders are insufficient to understand people's movement reduction [20, 133].

Obviously, the COVID-19 outbreak has influenced the transportation system and population mobility. Social distancing interventions and fear of the disease result in a traffic decline in cities and counties. During social distancing, mobility and traffic indices have declined compared to before implementing the orders. However, after ending stay-at-home orders and reopening restaurants and public places, traffic is gradually recovering, though the degree varies due to different policy direction by states. Notice that recent studies have used crowdsourced data such as Waze, Google, Apple, and Streetlight reports to quantify mobility changes as they are recognized as readily available real-time data resources. As discussed, most previous studies evaluated the association between mobility and the COVID-19 outbreak. Some studies quantify the impact of social distancing policies on people's mobility. Few studies, however, explore the mobility changes on the county-level [19] or spatial distribution of COVID-19 [134] to identify crucial spatial factors which link to the dynamics of mobility

changes. Although the traffic has been declined due to COVID-19 or social distancing order, other factors might also impact the magnitude of mobility change. There are gaps in the literature on surveying the contributing factors to traffic change at the county-level. Also, the difference between traffic patterns in decline and recovery phases has not been observed.

This chapter's contribution is to address these gaps in exploring the correlation between county-level spatial factors and pandemic-induced vehicle mile traveled (VMT) changes (decline and recovery). This study is timely and original by responding to a present pandemic and captures dynamics of mobility caused by pandemic and public interventions. To this end, we conduct a case study in Tennessee's 95 counties with three aims. First, we intend to quantify and classify VMT changes in different counties for both decline and recovery phases. Secondly, this study displays the spatial distribution of VMT change and county-level factors across the study area. Finally, it uses geographically weighted models to capture spatial heterogeneity in VMT changes and explore the association with county-level factors (commute to work, network, demographic, and economic). This study also addresses the difference between VMT decline and recovery. The result of this study can help states and cities to identify the counties and regions that are more prone to VMT changes arising from the pandemic and social distancing orders.

The remainder of this chapter is organized as follows. The next section presents study area, data used in this study, and data validation. Also, the dependent and explanatory variables will be introduced. The methodology section presents variable selection, repression models, and model comparison measures. Then, the results and discussion of the study are provided, followed by conclusions.

Data

Study Area

This study surveys all 95 counties in Tennessee to examine the mobility changes caused by COVID-19 and explores the contributing factors. The study has been conducted on mobility data from March 1 to June 30, 2020 which covers the range of two phases (mobility decline/recovery) in Tennessee data sets selected for the analysis and modeling are as follows.

StreetLight Data

Mobility data were made available through a request to the StreetLight website [135]. Combining the GPS movement data collected by the Cuebiq company, StreetLight employs its proprietary algorithm to produce the estimate of daily vehicle miles traveled (VMT) and aggregates them at the county-level. This study uses VMT as the mobility indicator to capture changes caused by pandemic. The

free public data starts on March 1, defined as the start of this study. StreetLight provided the average VMT for January as the reference under normal situations. Before this study, the StreetLight VMT dataset was widely used as the social distancing or movement indicator to examine the association with conventional health behaviors [19], policy intervention effects [20], air pollution [21, 22], and pandemic transmission [23].

Data validation

Using VMT as the mobility indicator has its advantages because it tracks the traveled distance and volume simultaneously. However, compared to the traditional VMT detected by vehicle miles and volume, the StreetLight VMT data still have many uncertainties in GPS data collection, such as device penetration, coverage, travel mode, movement context, and transformation process when converting GPS data to VMT numbers. Therefore, it entails a validation process to the data before conducting the rest of analysis. For this purpose, this study uses the detector data collected from the Radar Detection System (RDS) on interstates and provided by the Tennessee Department of Transportation (TDOT). The RDS records the interstate 30-seconds traffic information in four major Tennessee cities: Nashville, Memphis, Knoxville, and Chattanooga. The corresponding four counties are Davidson, Shelby, Knox, and Hamilton, respectively. Considering VMT and RDS data are collected at different geo-levels, this study assumes that those cities are located centrally within the county. To eliminate the two dataset units, we normalized the data using the average for the first two weeks in March.

To examine the correlation between daily Streetlight VMT change and RDS volume changes from March 1 to June 26, we eliminated the effect of units from two datasets through a normalization process using the average of the first two weeks in March. Table 18 shows the relative high Pearson correlation coefficients (greater than 0.89) of the four counties, confirming all correlations are statistically significant with $p\text{-value} < 0.0001$. The high degree of mobility change from these two measures suggests that the algorithm applied to the StreetLight VMT data does not warp the mobility much from the real situation. thus, the VMT data are reliable to be used in the further in-depth analysis at county level in Tennessee.

Dependent Variables

On March 5, 2020, the first confirmed case of coronavirus was announced in Tennessee. On March 12, 2020, Tennessee entered the emergency state, and many activities were canceled. Since March 13, a significant drop in daily VMT has been observed. As summarized in the boxplots of weekly VMT variation in 95 counties of Tennessee, Figure 23 highlighted a significant drop in daily VMT has been observed statewide since March 13. Here, the last two weeks of January 2020 were selected as the baseline. It is evident that, from the third week of March 2020, a significant VMT decline has occurred in almost all counties. The VMT decline phase has been continued for the next two weeks. In this study, these three weeks are called the Decline Phase.

Table 18. Data validation results.

	Pearson correlation coefficient	P-value
Nashville - Davidson	0.91	<0.0001
Memphis - Shelby	0.90	<0.0001
Knoxville - Knox	0.89	<0.0001
Chattanooga - Hamilton	0.91	<0.0001

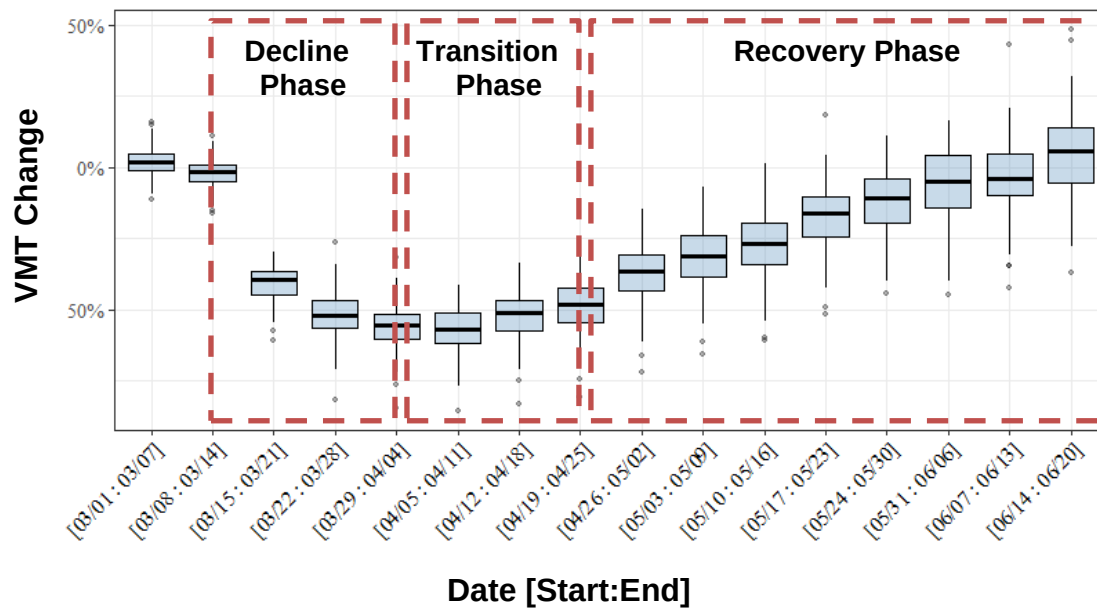


Figure 23. Boxplot of weekly VMT changes among 95 counties.

Next, for about three weeks, there was a slightly steady state in VMT level, called the Transition Phase. However, from April 27, with the restaurant reopening policy in Tennessee, the VMT trend began increasing, and this is called the Recovery Phase. Based on the boxplot, the range of changes in the recovery phase is broader than the decline phase. It can be inferred that various counties have diverse patterns in decline and recovery.

Figure 24 displays the VMT decreases during the decline phase and VMT increase in the recovery phase. VMT changes vary from county to county. Also, VMT changes are higher in both phases for the major cities in this study. Hence, this study considers the county-level magnitude of VMT decrease in the decline phase and the average weekly VMT increase during the recovery phase as the response variables. In this regard, two different models will be conducted for decline and recovery phase.

Explanatory Variables

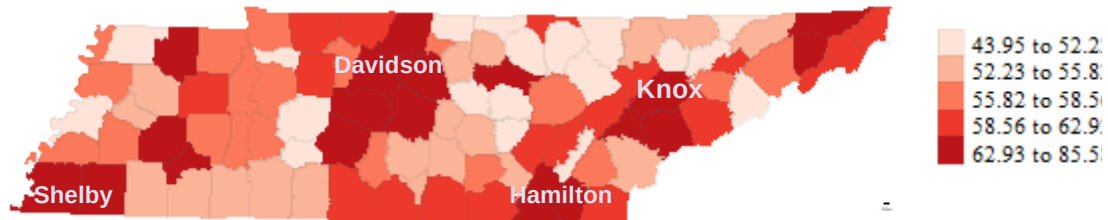
This study gathered four categories of county-level spatial factors as explanatory variables. This study tried to estimate the impact of each variable on VMT changes in decline and recovery phases. The explanatory variables are gathered from National datasets (Table 19).

Methodology

Variable Selection

Choosing the features of the variable selection is a technique to reduce the number of variables and the computational complexity. This study uses several variables, which can have some correlation. Here, variable selection determines the smallest set of variables to explain the dependent variable with high accuracy. The variable selection techniques decrease overfitting, increase accuracy, improve model performance, eliminate collinearity, and prevent possible correlations between variables. In other words, subset selection removes the variable that does not add much value in model estimation. Typically, t-value, Akaike Information Criterion (AIC), and R^2 are used to select the best subset. This study uses a well-known subset selection method called stepwise regression method. This method is a combination of the forward selection and backward elimination methods. Stepwise regression starts with no candidate variable. At each step, it adds the variable that increases the R^2 the most, but at the end of each step, it tests whether eliminating each candidate variable can improve the model. This method stops adding variables when none of the remainders improve the model.

Decline phase change (%)



Recovery phase weekly change (%)

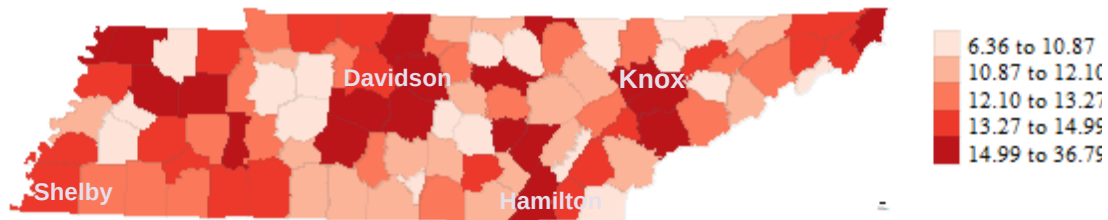


Figure 24. County-level VMT changes during decline and recovery phases (four major counties are labeled).

Table 19. Explanatory variables of the study

Category	Variable	Data Source
Commute to work	Percent of driving alone	American Community Survey (ACS), 2014-2018. (https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/)
	Percent of taking carpool	
	Percent of taking public transportation	
	Percent of walking to work	
	Percent of working from home	
	Mean time to work	
Network	Freeway density	Highway Performance Monitoring System (HPMS), 2017. (https://www.fhwa.dot.gov/policyinformation/hpms/shapefiles_2017.cfm)
	Non-Freeway density	
	Freeway AADT per mile	
	Non-Freeway AADT per mile	
	Freeway length	
	Non-Freeway length	
Demographic	Population density	County Health Rankings & Roadmaps (CHRR), 2020. (https://www.fhwa.dot.gov/policyinformation/hpms/shapefiles_2017.cfm)
	Percent of Population under 26 years	
	Percent of population over 65 years	
	Percent of minorities	
Economic	Median household income	U.S. Bureau of Economic Analysis (BEA), 2018 (https://www.bea.gov/news/2019/local-area-gross-domestic-product-2018)
	Percent of household with severe cost burden	
	Unemployed population	
	Labor force	
	Gross domestic product (GDP)	

Here, Based this method, the following variables were selected as explanatory variables of the model: freeway density, non-freeway density, median household income, percent of unemployment, population density, percent of people over 65, percent of people under 18, percent of people working from home, and mean time to work. From each explanatory variable category, at least two variables were selected.

Ordinary Least Squares (OLS)

OLS estimates the dependent variable based on a set of variables using the linear relation. OLS can be specified as the following equation.

$$y_i = \beta_0 + \sum_{k=1}^m \beta_k X_{ik} + \varepsilon_i \quad (34)$$

Here, y_i , denotes the dependent variable, and X_{ik} denotes k th independent variables at i th observation. Also, β_k represents the estimated coefficients, and ε_i is normal random error.

Geographically Weighted Regression (GWR)

The GWR model is typically used to capture the variation of different variables in estimating the dependent variable across the spatial dimension. The basic GWR model can be constructed as below [136].

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^m \beta_k(u_i, v_i) X_{ik} + \varepsilon_i \quad (35)$$

Where, $\beta_k(u_i, v_i)$ denotes the coordinates of location i . The $\beta(u_i, v_i)$ can be estimated as following equation. Also, m is number of explanatory variables.

$$\hat{\beta}(u_i, v_i) = [X^T W(u_i, v_i) X]^{-1} X^T W(u_i, v_i) Y \quad (36)$$

Where, $\hat{\beta}(u_i, v_i)$ represents the vector of the estimated coefficient at location i , also, X and Y are the matrix of independent variables and vector of the dependent variable, respectively. $W(u_i, v_i)$ is also an $n \times n$ (n is number of observations) spatial weighted matrix.

$$W(u_i, v_i) = \begin{bmatrix} W_{i1} & 0 & \dots & 0 \\ 0 & W_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & W_{in} \end{bmatrix} \quad (37)$$

Also, spatial weighted matrix values can be estimated using adaptive Gaussian function as the independent variables are normally distributed as pre-tests.

$$W_{ij} = \exp\left(-0.5\left(\frac{d_{ij}}{B}\right)^2\right) \quad (38)$$

Where, d_{ij} is the Euclidean distance between location i and j , B is also the distance bandwidth. Also, n is the number of observations. The optimal model was prescribed which produced the lowest AIC value among a set of bandwidth options. Here, we used adaptive Gaussian kernel, because adaptive kernel performed better than fixed kernel based on their AIC score.

Model Comparison Measures

In order to evaluate the performance of OLS and GWR methods, different statistical measures are selected. R^2 and VIF are important frequently used factors in the literature to determine the goodness of fit of a model. A model with high accuracy is preferred and selected if it shows a higher R^2 and lower VIF . Other indicators of goodness of fit and accuracy of model used in this study are presents as follows.

- **Akaike Information Criterion (AIC):** AIC compares models' performance. The model with the lower AIC shows better goodness of fit. AIC can be calculated as the following equation.

$$AIC = D + 2m \quad (39)$$

Where, D represents the model deviance, and m is the number of explanatory variables.

- **Mean Absolute Error (MAE):** MAE measure shows the deviation of the fitted model from observed values. The smaller MAE indicated better model estimation. It can be written as:

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (40)$$

y_i , and $\hat{y}_i$ denote observed dependent value and predicted value, respectively.

- **Root Mean Squared Error (RMSE):** This indicator also captures the distance between observation and fitted model estimation as follows.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (41)$$

Results

Descriptive Statistics

As mentioned earlier, nine variables were selected using stepwise regression method. Table 20 provides the descriptive statistics of the dependent and explanatory variables of this study. Here, mean, standard deviation, minimum, maximum, and median of each variable are provided for 95 counties. It can be seen that different variables vary in different counties.

Modeling Estimation

According to the discussed methodology section, the OLS and GWR models were developed to explore the impact of explanatory variables on VMT changes to assess the model's performance as global and local statistics, respectively. Before performing the models, some of the variables should be transformed to enhance the goodness of fit of the model. This study used Logarithmic Transformations to handle the highly skewed variables. Here, median household income and population density were treated with log-transform to conform the condition of normal distribution for both OLS and GWR. In the following, the model comparison is provided to clarify more insights about each model's performance. Then, estimated models are provided.

Model Comparison

As discussed, the OLS and GWR models were conducted for both decline and recovery phase. To compare the models, the *VIF*, *R*², *AIC*, *MAE*, and *RMSE* were calculated for OLS and GWR models. Table 21 compares the result of OLS and GWR for the decline and recovery phases. Based on the results, the *R*² of GWR in the decline and recovery phase is significantly higher than the OLS model. Also, for both models, GWR performs better *AIC*, *MAE*, and *RMSE* than the OLS model. The result of the model comparison suggests that using GWR improves the model. Hence, it can be concluded that the GWR model is suitable for fitting decline and recovery phases. *VIF* values (average=2.92, max=6.52) for all variables in both models were tested and all of them were <7.5, which indicate the constructed models does not have any redundant variable and is not biased with multicollinearity.

Table 20. Descriptive Statistics of variables.

Variable	Mean	S.D.	Min	Max	Median
Dependent Variables					
Average VMT Change (Decline Phase)	58.33	7.56	43.95	85.58	57.33
Weekly VMT Change (Recovery phase)	13.13	3.71	6.36	36.79	12.83
Explanatory Variables					
Freeway density	0.04	0.07	0.00	0.43	0.01
Non-freeway density	0.39	0.21	0.13	1.52	0.33
Median household income	47251	10896	30691	115930	44684
Percent of unemployment	1.71	0.22	1.28	2.39	1.67
Population density	146.9	216.4	19.1	1310.1	71.5
Percent of over 65	19.16	3.39	9.23	30.68	19.63
Percent of under 18	21.20	2.30	14.17	27.22	21.04
Percent of work from home	3.83	1.77	1.20	10.90	3.50
Mean time to work	26.67	4.03	18.70	36.20	26.50

Table 21. Comparison of OLS and GWR model (N=95 counties).

Model	Goodness of Fit	OLS Model	GWR Model
Decline phase model	R^2	0.76	0.86
	<i>Adjusted- R^2</i>	0.73	0.84
	<i>AIC</i>	528.6	513.4
	<i>MAE</i>	3.631	3.09
	<i>RMSE</i>	3.58	3.28
Recovery phase model	R^2	0.71	0.89
	<i>Adjusted- R^2</i>	0.68	0.88
	<i>AIC</i>	477.3	469.1
	<i>MAE</i>	2.05	1.93
	<i>RMSE</i>	2.72	2.01

VMT Decline Phase Model

First, we will consider the results from OLS, which are given in Table 22. Note that almost every explanatory variable is statistically significant except for Freeway Density. Moreover, most variables have expected signs. For instance, median household income positively affects average VMT decline because people with higher incomes tend to follow the stay-at-home order more strictly. They are more likely to withstand the economic pressure brought by COVID-19. It is strange to see a negative coefficient for percent of work from home in the decline model. This happens because the work from the home index was collected before the pandemic. Another counterintuitive finding is that those coefficients of percent of people over age 65 and under 18 have negative values. In terms of the explanatory power, the *Adjusted-R²* value for the OLS is 0.73 (see Table 21), indicating a good fit of the model. However, what the OLS model failed to indicate is the spatial variation across the whole study region. Therefore, a GWR model must be considered. Comparing GWR results with the OLS model helps interpret the extent to which the spatial variation lies within our data. According to Table 21, the *Adjusted-R²* for the GWR model improved by 0.11 compared to OLS, indicating better goodness of fit for the GWR model to explain the variation within the data. As shown in Figure 25, there is evident spatial variation in estimated coefficients distributed for every explanatory variable. It should be noted that the spatial variation of freeway density, household median income, and percent of people over age 65 explain the relationship of VMT decline phase clearly. First, the distribution map for the median household income coefficient elaborates the nature of the variation of traffic decline phase. The coefficients are all positive across the whole state. Particularly in West Tennessee, areas with relatively lower median household income, there is a stronger positive relationship than anywhere else. The result implies that there seems to be a “diminishing marginal utility” in the relationship between median household income the drop of average VMT change during decline phase. With the increase of median household income, the positive effects become less intense. The reason for this distribution might require further investigation. However, it is clear that a socio-economic condition of household is an explanatory variable to understand the outcome of declined VMT at county level in Tennessee. Additionally, what should have been interesting is that the coefficients for freeway density take both positive and negative values, regardless of the statistical significance problem.

Table 22. Modeling results for VMT decline phase (N=95 counties).

variable	OLS		GWR				
	Coef.	p-value ¹	Min	1st Q	Med	3rd Q	Max
(Intercept)	-150.4	0.00***	-181	-159	-120	-114	-92
Freeway density	4.2	0.68	-6.5	-1.0	1.9	4.2	7.7
Non-freeway density	8.4	0.04**	2.2	2.7	4.5	6.5	7.1
Median household income	22.0	0.00***	14.8	17.2	18.8	21.7	24.2
Percent of unemployment	3.0	0.00***	1.3	1.8	2.8	3.1	3.8
Population Density	-2.7	0.06*	-0.4	-0.3	-0.3	-0.2	-0.2
Percent of over 65	-0.3	0.06*	-0.4	-0.3	-0.1	-0.1	0.0
Percent of under 18	-1.0	0.00***	-1.1	-1.0	-1.0	-0.9	-0.8
Percent of work from home	-0.8	0.00***	-1.6	-1.2	-1.1	-0.6	-0.4
Mean time to work	-0.6	0.00***	-0.7	-0.6	-0.6	-0.4	-0.2

¹ Significance at *** 1%, ** 5%, and * 10%.

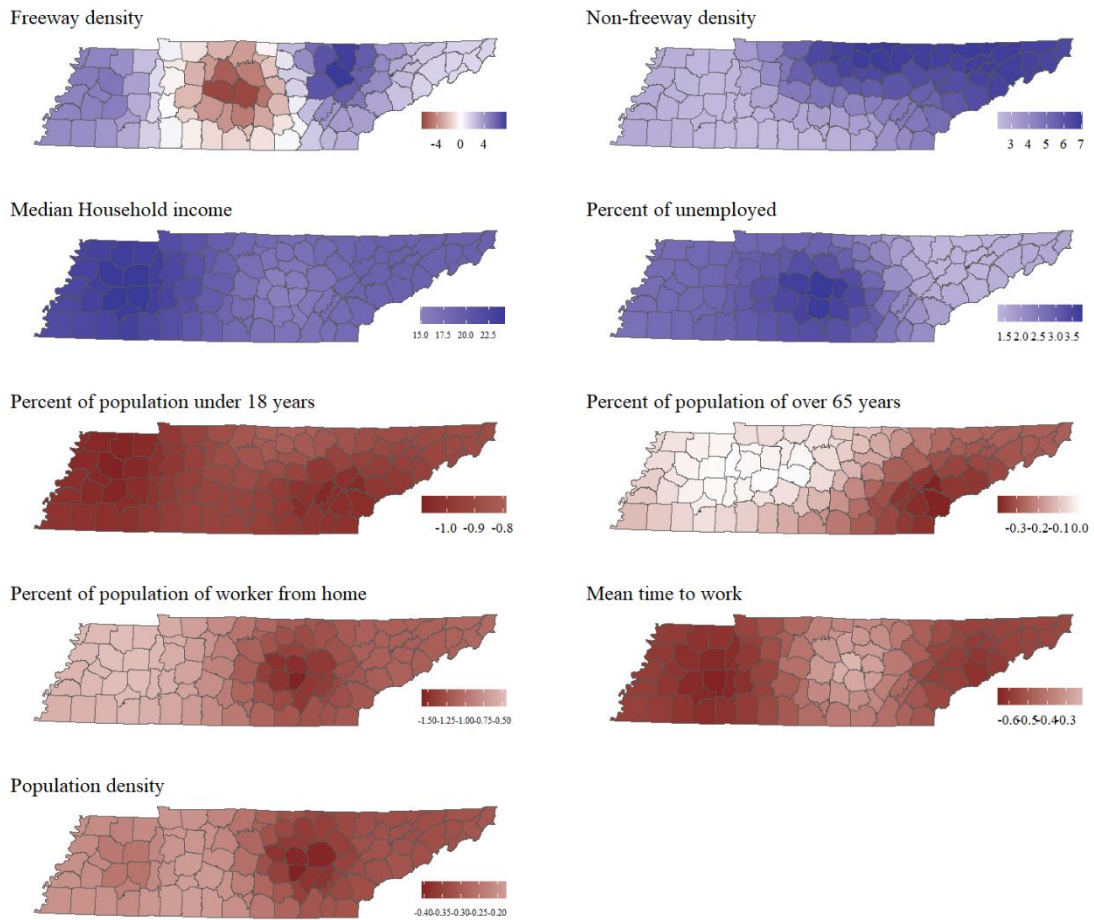


Figure 25. Decline Phase: GWR spatial distribution of estimated coefficients.

VMT Recovery Phase Model

As shown in Table 23, the output of OLS indicates that almost all explanatory variables have a statistically significant effect on weekly VMT change during the recovery phase except for non-freeway density. The signs of coefficients are consistent with expectations except for Freeway density. What is interesting is that the coefficient of freeway density implies a negative relationship with weekly VMT change. One should note that this relationship in the decline model is positive despite not being statistically significant. This negative value indicates that counties with higher freeway density tend to have a slower recovery. By comparison, the negative coefficient for the percent of work from home makes sense in the recovery model because more people working at home imply a slow recovery speed, thus resulting in a smaller average weekly VMT increase. Another interesting finding is that the coefficient of population density is negative in both the decline and recovery models. This might be because areas of high population density are more resistant to changes, no matter how they decrease or increase. Regarding the model's performance, the R^2 value for OLS is 0.68, not indicating a good fit. In a sense, the geospatial variation is more evident during the recovery phase since the global OLS fails to capture enough variation underlying the data. There must be some spatial non-stationarity in these relationships.

The *Adjusted- R^2* value for the GWR model is 0.88. This dramatic increase of 0.2 is due to GWR's capability of capturing the local spatial variation across the data. Figure 26 illustrates the spatial variation in all coefficients during the recovery phase. For the GWR model for the recovery phase, it is particularly interesting to see the coefficients for several independent variables take both positive and negative values across the study area. For instance, in west Tennessee and areas surrounding Overton county, there is a strong positive relationship between freeway density and weekly VMT increase so that high freeway density implies significant VMT change.

In contrast, in the Nashville area, the relationship is negative. One reason could be that the Nashville area has high freeway density and heavy traffic, which causes slow recovery speed for traffic in the greater Nashville area. It is hard to conclude specific reasons now, but it would be helpful to explore more data points. The coefficient for percent of the population over age 65 also distributes with typical geographic features. In the Nashville area, where the population over 65 is at a low level, the sign of the coefficient is negative, while the relationship is positive in most other areas. It seems that the highly critical COVID-19 areas including Nashville are sensitively responding to the recovery phase. There is also a similar pattern for the distribution of coefficient of population density. In areas with the lowest population density, the relationship tends to be mostly negative. It is hard to explain because the sign is not consistent with our intuition. There are certainly more stories behind the coefficients' geographic distribution, which requires new measurement and more experiments to give a reasonable explanation.

Table 23. Modeling results for VMT recovery phase (N=95 counties).

variable	OLS		GWR				
	Coef.	p-value ¹	Min	1st Q	Med	3rd Q	Max
(Intercept)	-71.04	0.00***	-276	-216	-126	-78	53
Freeway density	-15.48	0.05*	-24.5	-5.1	1.7	15.8	39.6
Non-freeway density	4.45	0.15	-2.1	2.1	4.0	7.3	13.1
Median household income	9.61	0.00***	0.0	12.7	19.7	26.9	35.9
Percent of unemployment	1.31	0.08*	0.1	2.0	2.8	4.1	6.2
Population Density	-1.90	0.08*	-5.8	-0.9	-0.5	-0.3	2.3
Percent of over 65	-0.32	0.01***	-0.9	-0.3	-0.1	0.1	0.5
Percent of under 18	-0.48	0.01**	-1.7	-1.3	-0.9	-0.7	0.1
Percent of work from home	-0.82	0.00***	-2.2	-1.8	1.1	-0.5	-0.2
Mean time to work	-0.27	0.00***	-1.0	-0.8	-0.7	-0.3	0.2

¹Significance at *** 1%, ** 5%, and * 10%.

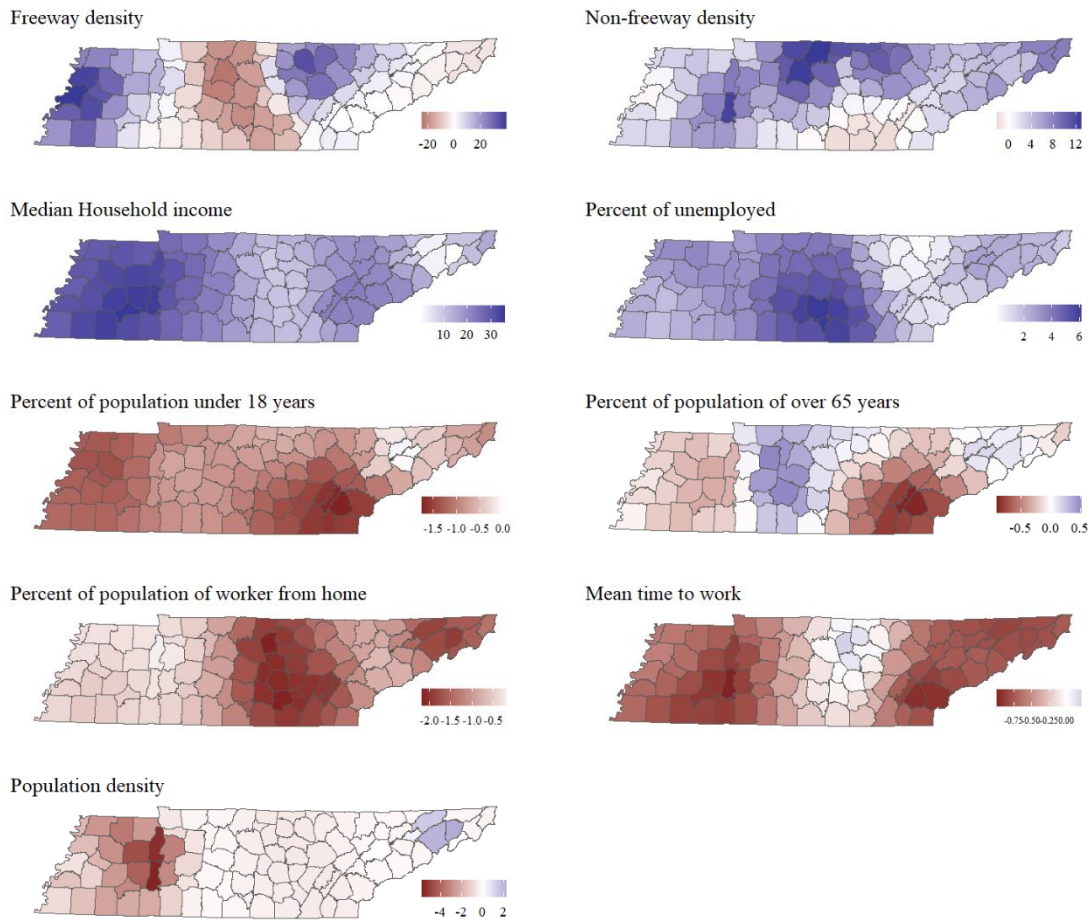


Figure 26. GWR spatial distribution of estimated coefficients: recovery phase.

Discussion

Unlike the decline phase, VMT recovery depends more on local attributes since it is a natural process without intervention. The rapid decline is due in part to the administrative order and policy change. Hence, the recovery phase's GWR model acquires a high R^2 value because the model captures much of the spatial variation.

It is not challenging to fit a geographically weighted regression model but explaining the variation of the output remains a challenge given limited information. Usually, more external factors and information are needed to give a reasonable interpretation of the model. Some may argue that the spatial variation may be caused by factors that are not accounted for in our models. The variation could be removed by the addition of further explanatory variables [136]. However, increasing the model complexity may cause other problems like multicollinearity, lack of geographic knowledge, and the robustness of the OLS or GWR. Besides, the GWR model performs acceptably on visualizing the spatial heterogeneity with a limited number of explanatory variables.

Conclusions

This study explored the impact of county-level factors (commute to work, network, demographic, economic) on mobility changes caused by the COVID-19 pandemic from the date with the first mandating order of stay-at home and reopening dates, which entailed the phases of declining and recovery of traffic. The goal of this study was to create a framework to display the mobility changes after the pandemic and model the spatial association of county-level factors and mobility changes. For this purpose, all Tennessee 95 counties were selected as the study area. Also, the mobility data were collected from the StreetLight website, and the VMT changes were calculated using the normal situation as the baseline for comparison. The mobility changes (dependent variables) have been classified in two phases: VMT decline after social distancing policy intervention, and VMT recovery after reopening some public places.

OLS and GWR were developed to quantify the impact of different county-level factors on mobility changes from the methodological viewpoint. Note that the individual models were conducted for the decline and recovery phases. The GWR model allows us to capture and identify the unobserved spatial heterogeneity in the data and variables. It should also be emphasized that the GWR model performed better fit in terms of AIC and R^2 for both recovery and decline phases.

In both decline and recovery models, non-freeway density, median household income, percent of unemployment, population density, percent of people over 65, percent of people under 18, percent of people who work from home, and mean time to work are significantly correlated with VMT change magnitude. Based on the modeling results, higher values of recovery models,

non-freeway density, median household income, and percent of unemployment of counties increase the VMT change. On the other hand, higher population density, percent of people over 65, percent of people under 18, percent of people who work from home, and mean time to work values decrease VMT changes. Also, the GWR model results indicate there is a spatial variation among estimated coefficients in different counties.

The results highlight the significance of spatial variation in analyzing VMT changes in events like the COVID-19 outbreak. It shows that different counties have different behaviors in confronting enacted policies. Stay-at-home policies are not sufficient to describe and quantify county mobility changes. The impact of different variables is spatially distributed among counties. On the other hand, the findings clarify that people do not immediately revert to pre-pandemic daily commutes and routines after ending stay-at-home orders. Thus, VMT recovery depends more on local and spatial attributes than the decline phase because VMT recovery is a natural process without direct intervention and policies.

CONCLUSION

This study compiled a series of studies on the application of crowdsourced data in transportation operation and safety. These studies were conducted to verify crowdsourced speed data quality, assess traffic status and LOS using crowdsourced data, propose a multisource lane blocking incident detection, and capture mobility dynamics amid the COVID-19 pandemic. Figure 27 provides Dissertation Key contributions and findings each chapter of this dissertation.

First, a study was designed to examine and evaluate Waze speed data quality on surface streets. Here, Bluetooth speed data were used as the ground truth data. This study explored the quality of Waze speed data in terms of accuracy, mean comparison, and distribution. This study also performed a multinomial logistic regression model on data quality to identify the contributing factors such as different times of day, days of the week, congestion levels, and speeds. The results of this study showed the acceptable quality of Waze crowdsourced speed data. Also, increasing the number of users can improve data quality and reduce errors. The crowdsourced speed data are promising data source for using in transportation research and development.

Second, a new methodology was proposed that incorporates crowdsourced data collected from Waze user reports with features extracted from speed/travel time variation to assess LOS on freeways. This method used the machine learning classification to assess six LOS categories. Multiple indices were calculated as the inputs of the classification model. This paper divided these indicators into three categories: (1) basic statistical measures, (2) travel time performance measures, and (3) crowdsourced data (Waze user reports). This study showed that using Waze user reports increased (10percent) the classification performance to 93 percent. The proposed approach showed the value of using crowdsourced data in transportation operations. It can be considered a supplemental methodology for traditional HCM LOS calculation that relies on traffic density and flow.

Third, a spatiotemporal multisource lane-blocking incident detection framework with improved performance (a high detection rate and low false alarm rate) was performed. This framework incorporates traffic speed pattern abnormality with crowdsourced user reports in a proposed clustering method called weighted spatiotemporal density-based clustering of applications with noise (WST-DBSCAN) to detect incidents. The results of implementing the algorithm showed that using multisource crowdsourced data provided the best detection performance in comparison to using single dataset, with 93 Percent, 0.19 Percent and 4.8 minute in DR, FAR, and MTTD, respectively.

Finally, the last study benefits from crowdsourced data in an unusual event such as the COVID19 pandemic. This chapter explored the correlation between county-level spatial factors and pandemic-induced vehicle mile traveled (VMT) changes (decline and recovery) during the COVID19 outbreak. This study performed geographically weighted regression (GWR) models to capture spatial heterogeneity in VMT changes and explore the association with county-level factors (commute to work, network, demographic, and economic). The results

showed different counties have different behaviors in confronting enacted policies. The GWR model results indicate there is a spatial variation among estimated coefficients in different counties. Also, VMT recovery depends on local attributes more than decline phase.

Dissertation Key contributions and findings

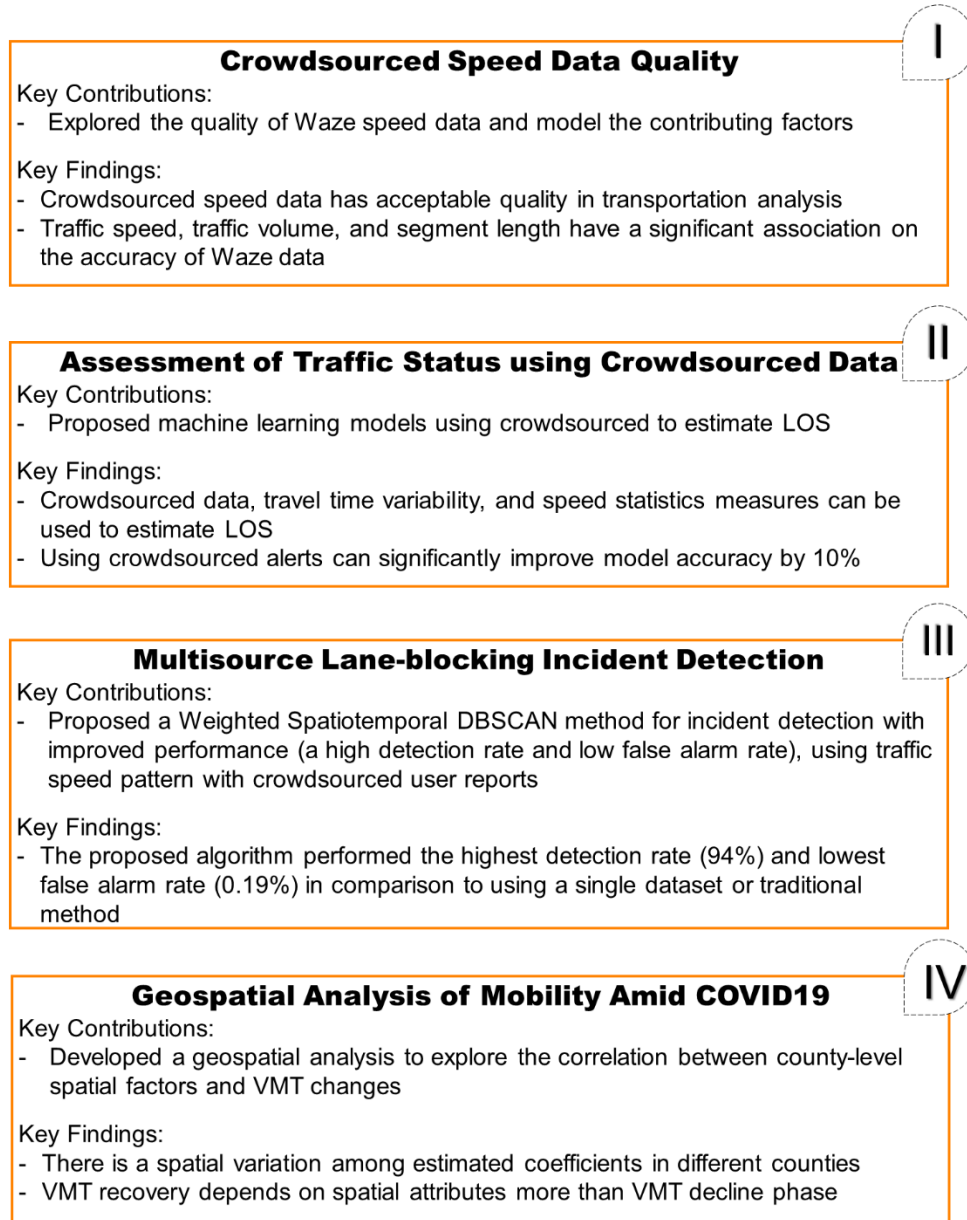


Figure 27. Dissertation Key contributions and findings.

REFERENCES

- [1] Statista. (2019, July 6). *Number of smartphone users in the United States from 2010 to 2023 (in millions)**. Available: <https://www.statista.com/statistics/201182/forecast-of-smartphone-users-in-the-us/>
- [2] G. Chatzimilioudis, A. Konstantinidis, C. Laoudias, and D. Zeinalipour-Yazti, "Crowdsourcing with smartphones," *IEEE Internet Computing*, vol. 16, no. 5, pp. 36-44, 2012.
- [3] M. Azad, N. Hoseinzadeh, C. Brakewood, C. R. Cherry, and L. D. Han, "A Literature Review on Fully Autonomous Buses," presented at the Transportation Research Board 98th Annual Meeting, Washington DC, Unites States, 2019.
- [4] N. Hoseinzadeh, R. Arvin, A. J. Khattak, and L. D. Han, "Integrating safety and mobility for pathfinding using big data generated by connected vehicles," *Journal of Intelligent Transportation Systems*, pp. 1-17, 2020.
- [5] V. Ahsani, M. Amin-Naseri, S. Knickerbocker, and A. Sharma, "Quantitative analysis of probe data characteristics: Coverage, speed bias and congestion detection precision," *Journal of Intelligent Transportation Systems*, vol. 23, no. 2, pp. 103-119, 2019.
- [6] A. Haghani, X. Zhang, and M. Hamedi, "Validation and augmentation of Inrix arterial travel time data using independent sources," Maryland. State Highway Administration 2015.
- [7] Y. Aliari and A. Haghani, "Bluetooth sensor data and ground truth testing of reported travel times," *Transportation Research Record*, vol. 2308, no. 1, pp. 167-172, 2012.
- [8] M. Amin-Naseri, P. Chakraborty, A. Sharma, S. B. Gilbert, and M. Hong, "Evaluating the reliability, coverage, and added value of crowdsourced traffic incident reports from Waze," *Transportation Research Record*, vol. 2672, no. 43, pp. 34-43, 2018.
- [9] S. R. Hargrove, H. Lim, L. D. Han, and P. B. Freeze, "Empirical evaluation of the accuracy of technologies for measuring average speed in real time," *Transportation Research Record*, vol. 2594, no. 1, pp. 73-82, 2016.
- [10] A. Sharma, V. Ahsani, and S. Rawat, "Evaluation of opportunities and challenges of using INRIX data for real-time performance monitoring and historical trend assessment," 2017.
- [11] S. E. Underwood, "A review and classification of sensors for intelligent vehicle-highway systems," 1990.
- [12] F. Ding, X. Chen, S. He, G. Shou, Z. Zhang, and Y. Zhou, "Evaluation of a Wi-Fi signal based system for freeway traffic states monitoring: An exploratory field test," *Sensors*, vol. 19, no. 2, p. 409, 2019.
- [13] G. A. Gravvanis, A. I. Salamanis, and C. K. Filelis-Papadopoulos, "Advanced Parametric Methods for Short-Term Traffic Forecasting in the

- Era of Big Data," in *Machine Learning Paradigms*: Springer, 2019, pp. 199-231.
- [14] C. Bachmann, M. J. Roorda, B. Abdulhai, and B. Moshiri, "Fusing a bluetooth traffic monitoring system with loop detector data for improved freeway traffic speed estimation," *Journal of Intelligent Transportation Systems*, vol. 17, no. 2, pp. 152-164, 2013.
 - [15] A. Haghani, M. Hamed, K. F. Sadabadi, S. Young, and P. Tarnoff, "Data collection of freeway travel time ground truth with bluetooth sensors," *Transportation Research Record*, vol. 2160, no. 1, pp. 60-68, 2010.
 - [16] J. C. Herrera, D. B. Work, R. Herring, X. J. Ban, Q. Jacobson, and A. M. Bayen, "Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 4, pp. 568-583, 2010.
 - [17] R. Omrani, P. Izadpanah, G. Nikolic, B. Hellinga, A. Hadayeghi, and H. Abdelgawad, "Evaluation of wide-area traffic monitoring technologies for travel time studies," *Transportation Research Record*, vol. 2380, no. 1, pp. 108-119, 2013.
 - [18] J. Zhang, S. He, W. Wang, and F. Zhan, "Accuracy analysis of freeway traffic speed estimation based on the integration of cellular probe system and loop detectors," *Journal of Intelligent Transportation Systems*, vol. 19, no. 4, pp. 411-426, 2015.
 - [19] K. J. Bourassa, D. A. Sbarra, A. Caspi, and T. E. Moffitt, "Social Distancing as a Health Behavior: County-Level Movement in the United States During the COVID-19 Pandemic Is Associated with Conventional Health Behaviors," *Annals of Behavioral Medicine*, 2020.
 - [20] K. J. Bourassa, "State-level Stay-at-home Orders and Objectively Measured Movement in the United States During the COVID-19 Pandemic," 2020.
 - [21] C. Jia, X. Fu, D. Bartelli, and L. J. A. Smith, "Insignificant Impact of the "Stay-At-Home" Order on Ambient Air Quality in the Memphis Metropolitan Area, USA," vol. 11, no. 6, p. 630, 2020.
 - [22] S. Cicala, S. P. Holland, E. T. Mansur, N. Z. Muller, and A. J. Yates, "Expected Health Effects of Reduced Air Pollution from COVID-19 Social Distancing," National Bureau of Economic Research 0898-2937, 2020.
 - [23] L. S. Robertson, "Predicting Growth of COVID-19 Confirmed Cases in Each US County with a Population of 50,000 or More, Revised July 8, 2020."
 - [24] R.-P. Schäfer, K.-U. Thiessenhusen, and P. Wagner, "A traffic information system by means of real-time floating-car data," in *ITS world congress*, 2002, vol. 2002.
 - [25] J. Hu, M. D. Fontaine, and J. Ma, "Quality of private sector travel-time data on arterials," *Journal of Transportation Engineering*, vol. 142, no. 4, p. 04016010, 2016.

- [26] C. Systematics, "Travel time data collection white paper," *Florida Department of Transportation, District IV*, 2012.
- [27] Z. Z. Tian, K. Balke, R. Engelbrecht, and L. Rilett, "Integrated control strategies for surface street and freeway systems," *Transportation Research Record*, vol. 1811, no. 1, pp. 92-99, 2002.
- [28] G. Barbier, R. Zafarani, H. Gao, G. Fung, and H. Liu, "Maximizing benefits from crowdsourced data," *Computational and Mathematical Organization Theory*, vol. 18, no. 3, pp. 257-279, 2012.
- [29] A. Feizi, S. Joo, V. Kwigizile, and J.-S. Oh, "Analysis of Bicycle Environment Using Instrumented Probe Bicycle," *Journal of Transport & Health*, vol. 9, p. S32, 2018.
- [30] A. Feizi, J.-S. Oh, V. Kwigizile, and S. Joo, "Cycling environment analysis by bicyclists' skill levels using instrumented probe bicycle (IPB)," *International Journal of Sustainable Transportation*, pp. 1-11, 2019.
- [31] A. Misra, A. Gooze, K. Watkins, M. Asad, and C. A. Le Dantec, "Crowdsourcing and its application to transportation data collection and management," *Transportation Research Record*, vol. 2414, no. 1, pp. 1-8, 2014.
- [32] M. S. Sharifi, Z. Song, H. N. Esfahani, and K. Christensen, "Exploring heterogeneous pedestrian stream characteristics at walking facilities with different angle intersections," *Physica A: Statistical Mechanics and its Applications*, vol. 540, p. 123112, 2020.
- [33] F. Alhajri, "A comparative assessment of crowded source travel time estimates: A case study of Bluetooth vs INRIX on a suburban arterial," 2014.
- [34] A. Haghani, M. Hamed, and K. F. Sadabadi, "I-95 Corridor coalition vehicle probe project: Validation of INRIX data," *I-95 Corridor Coalition*, vol. 9, 2009.
- [35] S. Kim and B. Coifman, "Comparing INRIX speed data against concurrent loop detector stations over several months," *Transportation Research Part C: Emerging Technologies*, vol. 49, pp. 59-72, 2014.
- [36] S. Tahmasseby, "Traffic data: Bluetooth sensors vs. crowdsourcing—A comparative study to calculate travel time reliability in calgary, alberta, canada," *Journal of Traffic and Transportation Engineering*, vol. 3, pp. 63-79, 2015.
- [37] X. Liu, S. Chien, and K. Kim, "Evaluation of floating car technologies for travel time estimation," *Journal of Modern Transportation*, vol. 20, no. 1, pp. 49-56, 2012.
- [38] L. Elefteriadou, A. Kondyli, and B. St George, "Comparison of methods for measuring travel time at Florida freeways and arterials," *Florida. Dept. of Transportation* 2014.
- [39] X. Zhang, M. Hamed, and A. Haghani, "Arterial travel time validation and augmentation with two independent data sources," *Transportation Research Record*, vol. 2526, no. 1, pp. 79-89, 2015.

- [40] M. Hamed, A. Haghani, and Y. Zhang, "Travel Time Reliability Measurement Sensitivity to Data Source Selection," *IRF*, p. 13, 2013.
- [41] INRIX, "I-95 corridor coalition: vehicle probe project data validation summary.," 2015, Available: <https://inrix.com/wp-content/uploads/2016/11/INRIX-I-95-VPP-Data-Summary-Validation-1.pdf>.
- [42] Y. O. Adu-Gyamfi, A. Sharma, S. Knickerbocker, N. Hawkins, and M. J. T. R. R. Jackson, "Framework for evaluating the reliability of wide-area probe data," vol. 2643, no. 1, pp. 93-104, 2017.
- [43] S. M. Turner and P. Koeneman, "Validating the National Performance Management Research Data Set (NPMRDS) for South Dakota," South Dakota. Dept. of Transportation. Office of Research 2018.
- [44] W. Feng, A. Y. Bigazzi, S. Kothuri, and R. L. Bertini, "Freeway sensor spacing and probe vehicle penetration: Impacts on travel time prediction and estimation accuracy," *Transportation research record*, vol. 2178, no. 1, pp. 67-78, 2010.
- [45] B. N. Araghi, J. Hammershøj Olesen, R. Krishnan, L. Tørholm Christensen, and H. Lahrmann, "Reliability of bluetooth technology for travel time estimation," *Journal of Intelligent Transportation Systems*, vol. 19, no. 3, pp. 240-255, 2015.
- [46] J. J. V. Díaz, A. B. R. González, and M. R. Wilby, "Bluetooth traffic monitoring systems for travel time estimation on freeways," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 123-132, 2015.
- [47] S. R. dos Santos, C. A. Davis Jr, and R. Smarzaro, "Integration of data sources on traffic accidents," in *GeoInfo*, 2016, pp. 192-203.
- [48] R. B. Bahaweres and M. R. Akbar, "Analysis of travel time computation accuracy from Crowdsourced data of hospitality application in South of Tangerang City with estimated travel time method," in *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, 2017, pp. 1-5: IEEE.
- [49] L. Lu, J. Wang, Z. He, and C.-Y. Chan, "Real-time estimation of freeway travel time with recurrent congestion based on sparse detector data," *IEEE Intelligent Transport Systems*, vol. 12, no. 1, pp. 2-11, 2017.
- [50] C. Jia, Q. Li, S. Oppong, D. Ni, J. Collura, and P. W. Shuldiner, "Evaluation of alternative technologies to estimate travel time on rural interstates," 2013.
- [51] A. Kondyli, B. St. George, and L. Elefteriadou, "Comparison of travel time measurement methods along freeway and arterial facilities," *Transportation Letters*, vol. 10, no. 4, pp. 215-228, 2018.
- [52] D. Schrank, B. Eisele, and T. Lomax, "TTI's 2012 urban mobility report," Texas A&M Transportation Institute. The Texas A&M University System 2012, vol. 4.

- [53] H. Rakha, I. El-Shawarby, and M. Arafeh, "Trip travel-time reliability: issues and proposed solutions," *Journal of Intelligent Transportation Systems*, vol. 14, no. 4, pp. 232-250, 2010.
- [54] D. Bohning, "Multinomial Logistic-Regression Algorithm," (in English), *Annals of the Institute of Statistical Mathematics*, vol. 44, no. 1, pp. 197-200, Mar 1992.
- [55] H. Nasr Esfahani, R. Arvin, Z. Song, and N. Sze, "Prevalence of cell phone use while driving and its impact on driving performance, focusing on near-crash risk: A survey study in Tehran," *Journal of Transportation Safety Security*, pp. 1-21, 2019.
- [56] G. Azimi, H. Asgari, A. Rahimi, and X. Jin, "Investigation of heterogeneity in severity analysis for large truck crashes," 2019, vol. No. 19-02748.
- [57] G. Azimi, A. Rahimi, H. Asgari, and X. Jin, "Severity analysis for large truck rollover crashes using a random parameter ordered logit model," *Accident Analysis & Prevention*, vol. 135, 2020.
- [58] M. Pack and N. Ivanov, "Are You Gonna Go My WAZE?," *Institute of Transportation Engineers. ITE Journal*, vol. 87, no. 2, p. 28, 2017.
- [59] TPO. (2019). *Knoxville Region Traffic Count Program*. Available: <https://knoxtrans.org/traffic-counts>
- [60] M. J. Slakter, "A comparison of the Pearson chi-square and Kolmogorov goodness-of-fit tests with respect to validity," *Journal of the American Statistical Association*, vol. 60, no. 311, pp. 854-858, 1965.
- [61] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected papers of hirotugu akaike*: Springer, 1998, pp. 199-213.
- [62] T. R. Board, *Highway Capacity Manual 6th Edition: A Guide for Multimodal Mobility Analysis*. Washington, DC: The National Academies Press, 2016, p. 6.
- [63] S. M. Khan, K. C. Dey, and M. Chowdhury, "Real-time traffic state estimation with connected vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1687-1699, 2017.
- [64] S. Hernandez, A. Tok, and S. G. Ritchie, "Density Estimation using Inductive Loop Signature based Vehicle Re-identification and Classification," 2013.
- [65] S. S. Kanhere, "Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces," in *2011 IEEE 12th International Conference on Mobile Data Management*, 2011, vol. 2, pp. 3-6: IEEE.
- [66] D. J. Nair, F. Gilles, S. Chand, N. Saxena, and V. Dixit, "Characterizing multicity urban traffic conditions using crowdsourced data," *PLoS one*, vol. 14, no. 3, p. e0212845, 2019.
- [67] N. Hoseinzadeh, Y. Liu, L. D. Han, C. Brakewood, A. J. C. Mohammadnazar, Environment, and U. Systems, "Quality of location-based crowdsourced speed data on surface streets: A case study of Waze and Bluetooth speed data in Sevierville, TN," vol. 83, p. 101518, 2020.

- [68] V. SINGH, N. GORE, A. CHEPURI, S. ARKATKAR, G. JOSHI, and S. PULUGURTHA, "Examining Travel Time Variability and Reliability on an Urban Arterial Road Using Wi-Fi Detections-A Case Study," *Journal of the Eastern Asia Society for Transportation Studies*, vol. 13, pp. 2390-2411, 2019.
- [69] H. B. J. C. A. C. Celikoglu and I. Engineering, "An approach to dynamic classification of traffic flow patterns," vol. 28, no. 4, pp. 273-288, 2013.
- [70] P. Sekuła, N. Marković, Z. Vander Laan, and K. F. Sadabadi, "Estimating historical hourly traffic volumes via machine learning and vehicle probe data: A Maryland case study " *Transportation Research Part C: Emerging Technologies*, vol. 97, pp. 147-158, 2018.
- [71] B. Marwah and B. Singh, "Level of service classification for urban heterogeneous traffic: A case study of Kanpur metropolis," in *fourth international symposium on Highway Capacity, Hawaii*, 2000.
- [72] O. Altintasi, H. Tuydes-Yaman, and K. Tuncay, "Detection of urban traffic patterns from Floating Car Data (FCD)," *Transportation research procedia*, vol. 22, pp. 382-391, 2017.
- [73] J. Geistefeldt et al., "Assessment of level of service on freeways by microscopic traffic simulation," *Transportation Research Record*, vol. 2461, no. 1, pp. 41-49, 2014.
- [74] N. Wu and K. Lemke, "A new model for level of service of freeway merge, diverge, and weaving segments," *Procedia-social behavioral sciences*, vol. 16, pp. 151-161, 2011.
- [75] D. Jolovic, A. Stevanovic, S. Sajjadi, and P. T. Martin, "Assessment of Level-Of-Service for Freeway Segments Using HCM and Microsimulation Methods," *Transportation Research Procedia*, vol. 15, pp. 403-416, 2016.
- [76] H. B. Celikoglu and M. A. J. T. S. Silgu, "Extension of traffic flow pattern dynamic classification by a macroscopic model using multivariate clustering," vol. 50, no. 3, pp. 966-981, 2016.
- [77] M. A. Aljamal, H. M. Abdelghaffar, and H. A. Rakha, "Developing a Neural-Kalman Filtering Approach for Estimating Traffic Stream Density Using Probe Vehicle Data," *Sensors*, vol. 19, no. 19, p. 4325, 2019.
- [78] T. Wassantachat, Z. Li, J. Chen, Y. Wang, and E. Tan, "Traffic density estimation with on-line SVM classifier," in *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009, pp. 13-18: IEEE.
- [79] S. M. Khan, "Real-time traffic condition assessment with connected vehicles," 2015.
- [80] F. Kurniawan, H. Sajati, and O. Dinaryanto, "Image Processing Technique for Traffic Density Estimation."
- [81] T. Lomax and R. Margiotta, *Selecting travel reliability measures*. Citeseer, 2003.
- [82] M. Martchouk, F. L. Mannering, and L. Singh, "Travel time reliability in Indiana," 2010.

- [83] A. Chepuri, A. Wagh, S. S. Arkatkar, and G. Joshi, "Study of travel time variability using two-wheeler probe data—an Indian experience," in *Proceedings of the Institution of Civil Engineers-Transport*, 2018, vol. 171, no. 4, pp. 190-206: Thomas Telford Ltd.
- [84] W. Kittelson and M. Vandehey, "Incorporation of Travel Time Reliability into the HCM," 2013.
- [85] C. Chen, A. Skabardonis, and P. Varaiya, "Travel-time reliability as a measure of service," *Transportation Research Record*, vol. 1855, no. 1, pp. 74-79, 2003.
- [86] S. S. Pulugurtha and M. S. Imran, "Modeling basic freeway section level-of-service based on travel time and reliability," *Case Studies on Transport Policy*, vol. 8, no. 1, pp. 127-134, 2020.
- [87] S. R. Kodupuganti and S. S. Pulugurtha, "Link-level travel time measures-based level of service thresholds by the posted speed limit," *Transportation research interdisciplinary perspectives*, vol. 3, p. 100068, 2019.
- [88] S. Turner *et al.*, "Exploring Crowdsourced Monitoring Data for Safety," 2020.
- [89] G. V. A. Perez, J. C. Lopez, A. L. R. Cabello, E. B. Grajales, A. P. Espinosa, and J. L. Q. Fabian, "Road Traffic Accidents Analysis in Mexico City through Crowdsourcing Data and Data Mining Techniques," *International Journal of Computer Information Engineering*, vol. 12, no. 8, pp. 604-608, 2018.
- [90] R. Sanchez *et al.*, "Dynamic Traffic Lights and Urban Mobility: An application of Waze data to the city of Medellin," 2019.
- [91] Al-Sobky, Al-Sayed Ahmed, and M. R. M, "Traffic density determination and its applications using smartphone," *Alexandria Engineering Journal*, vol. 55, no. 1, pp. 513-523, 2016.
- [92] J. O. Ensley, "Application of Highway Capacity Manual 2010 Level-of-Service Methodologies for Planning Deficiency Analysis," 2012.
- [93] U. M. J. I. Scorecard, USA, "The Texas A&M Transportation Institute and Inrix," vol. 9, no. 2015, p. 10, 2015.
- [94] E. Parkany and C. Xie, "A complete review of incident detection algorithms & their deployment: what works and what doesn't," 2005.
- [95] M. M. Castro-Neto, L. D. Han, Y.-S. Jeong, and M. K. J. T. r. r. Jeong, "Toward Training-Free Automatic Detection of Freeway Incidents: Simple Algorithm with One Parameter," vol. 2278, no. 1, pp. 42-49, 2012.
- [96] P. Chakraborty, C. Hegde, and A. J. T. r. p. C. e. t. Sharma, "Data-driven parallelizable traffic incident detection using spatio-temporally denoised robust thresholds," vol. 105, pp. 81-99, 2019.
- [97] P. Chakraborty, J. R. Hess, A. Sharma, and S. Knickerbocker, "Outlier mining based traffic incident detection using big data analytics," in *Transportation Research Board 96th Annual Meeting Compendium of Papers*, 2017, pp. 8-12.

- [98] H. J. Payne and S. C. J. T. R. R. Tignor, "Freeway incident-detection algorithms based on decision trees with states," no. 682, 1978.
- [99] C. Liu, M. Zhao, A. Sharma, and S. J. J. o. B. D. A. i. T. Sarkar, "Traffic Dynamics Exploration and Incident Detection Using Spatiotemporal Graphical Modeling," vol. 1, no. 1, pp. 37-55, 2019.
- [100] B. N. Persaud and F. L. J. T. R. P. A. G. Hall, "Catastrophe theory and patterns in 30-second freeway traffic data—implications for incident detection," vol. 23, no. 2, pp. 103-113, 1989.
- [101] Y.-K. Ki, J.-H. Kim, T.-K. Kim, N.-W. Heo, J.-W. Choi, and J.-H. Jeong, "Method for Automatic Detection of Traffic Incidents Using Neural Networks and Traffic Data," in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2018, pp. 184-188: IEEE.
- [102] Y.-K. Ki, W.-T. Jeong, H.-J. Kwon, and M.-R. Kim, "An Algorithm for Incident Detection Using Artificial Neural Networks," in *2019 25th Conference of Open Innovations Association (FRUCT)*, 2019, pp. 162-167: IEEE.
- [103] L. Li, S. He, J. Zhang, and F. Yang, "Bagging-SVMs algorithm-based traffic incident detection," in *CICTP 2016*, 2016, pp. 1461-1469.
- [104] B. Anbaroglu, B. Heydecker, and T. Cheng, "Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks," *Transportation Research Part C: Emerging Technologies*, vol. 48, pp. 47-65, 2014.
- [105] Y. Chung, "Quantification of nonrecurrent congestion delay caused by freeway accidents and analysis of causal factors," *Transportation research record*, vol. 2229, no. 1, pp. 8-18, 2011.
- [106] S. Hasan and S. V. Ukkusuri, "Urban activity pattern classification using topic models from online geo-location data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 363-381, 2014.
- [107] C. Cottrill, P. Gault, G. Yeboah, J. D. Nelson, J. Anable, and T. Budd, "Tweeting Transit: An examination of social media strategies for transport information management during a large event," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 421-432, 2017.
- [108] Z. Zhang, Q. He, J. Gao, and M. Ni, "A deep learning approach for detecting traffic accidents from social media data," *Transportation research part C: emerging technologies*, vol. 86, pp. 580-596, 2018.
- [109] Y. Gu, Z. S. Qian, and F. Chen, "From Twitter to detector: Real-time traffic incident detection using social media data," *Transportation research part C: emerging technologies*, vol. 67, pp. 321-342, 2016.
- [110] S. Zhang, J. Tang, H. Wang, and Y. Wang, "Enhancing traffic incident detection by using spatial point pattern analysis on social media," *Transportation Research Record*, vol. 2528, no. 1, pp. 69-77, 2015.

- [111] A. Salas, P. Georgakis, and Y. Petalas, "Incident detection using data from social media," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 751-755: IEEE.
- [112] M. Amin-Naseri, "Adopting and incorporating crowdsourced traffic data in advanced transportation management systems," 2018.
- [113] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial-temporal data," *Data knowledge engineering*, vol. 60, no. 1, pp. 208-221, 2007.
- [114] (2020). *Naming the coronavirus disease (COVID-19) and the virus that causes it*. Available: [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it/](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it/)
- [115] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, and P.-R. Hsueh, "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges," *International journal of antimicrobial agents*, p. 105924, 2020.
- [116] (2020, July 28). *COVID-19 CORONAVIRUS PANDEMIC*. Available: <https://www.worldometers.info/coronavirus/>
- [117] E. K. Stokes *et al.*, "Coronavirus Disease 2019 Case Surveillance—United States, January 22–May 30, 2020," vol. 69, no. 24, p. 759, 2020.
- [118] K. Glanz and D. B. J. A. r. o. p. h. Bishop, "The role of behavioral science theory in development and implementation of public health interventions," vol. 31, pp. 399-418, 2010.
- [119] M. P. Kelly and M. J. P. h. Barker, "Why is changing health-related behaviour so difficult?," vol. 136, pp. 109-116, 2016.
- [120] C.-H. Lin, T.-H. J. I. j. o. e. r. Wen, and p. health, "Using geographically weighted regression (GWR) to explore spatial varying relationships of immature mosquitoes and human densities with the incidence of dengue," vol. 8, no. 7, pp. 2798-2815, 2011.
- [121] B. Meng, J. Wang, J. Liu, J. Wu, and E. J. P. H. Zhong, "Understanding the spatial diffusion process of severe acute respiratory syndrome in Beijing," vol. 119, no. 12, pp. 1080-1087, 2005.
- [122] K. Al-Ahmadi, S. Alahmadi, and A. Al-Zahrani, "Spatiotemporal clustering of Middle East respiratory syndrome coronavirus (MERS-CoV) incidence in Saudi Arabia, 2012–2019," *International journal of environmental research public health*, vol. 16, no. 14, p. 2520, 2019.
- [123] M. U. Kraemer *et al.*, "The effect of human mobility and control measures on the COVID-19 epidemic in China," vol. 368, no. 6490, pp. 493-497, 2020.
- [124] R. Li *et al.*, "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)," vol. 368, no. 6490, pp. 489-493, 2020.

- [125] M. Chinazzi *et al.*, "The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak," vol. 368, no. 6489, pp. 395-400, 2020.
- [126] K. Linka, M. Peirlinck, F. Sahli Costabal, E. J. C. M. i. B. Kuhl, and B. Engineering, "Outbreak dynamics of COVID-19 in Europe and the effect of travel restrictions," pp. 1-8, 2020.
- [127] B. Seibold, Z. Vucetic, and S. J. a. p. a. Vucetic, "Quantitative Relationship between Population Mobility and COVID-19 Growth Rate based on 14 Countries," 2020.
- [128] B. Klein *et al.*, "Assessing changes in commuting and individual mobility in major metropolitan areas in the United States during the COVID-19 outbreak," ed: Northeastern University. Network Science Institute. Retrieved from [https ...](https://...), 2020.
- [129] E. Pepe *et al.*, "COVID-19 outbreak response: a first assessment of mobility changes in Italy following national lockdown," 2020.
- [130] G. A. Wellenius *et al.*, "Impacts of state-level policies on social distancing in the united states using aggregated mobility data during the covid-19 pandemic," 2020.
- [131] C. Malone and K. Bourassa. (2020). *Americans Didn't Wait For Their Governors To Tell Them To Stay Home Because Of COVID-19*. Available: <https://fivethirtyeight.com/features/americans-didnt-wait-for-their-governors-to-tell-them-to-stay-home-because-of-covid-19/>
- [132] G. Dance and L. Gamio. (2020). *As Coronavirus Restrictions Lift, Millions in U.S. Are Leaving Home Again*. Available: <https://www.nytimes.com/interactive/2020/05/12/us/coronavirus-reopening-shutdown.html>
- [133] S. Sen, P. Karaca-Mandic, and A. J. J. Georgiou, "Association of Stay-at-Home Orders With COVID-19 Hospitalizations in 4 States," 2020.
- [134] Y. Chen, J. Jiao, S. Bai, and J. J. A. a. S. Lindquist, "Modeling the Spatial Factors of COVID-19 in New York City," 2020.
- [135] (2020). *StreetLight Data*. Available: <https://www.streetlightdata.com/>
- [136] C. Brunsdon, S. Fotheringham, and M. Charlton, "Geographically weighted regression," *Journal of the Royal Statistical Society: Series D*, vol. 47, no. 3, pp. 431-443, 1998.

VITA

Nima Hoseinzadeh was born and raised in Iran. In 2012 he graduated from Iran University of Science and Technology with a bachelor's degree in Electrical Engineering. In 2015, he graduated from Kharazmi University with an MBA degree. Nima joined the University of Tennessee, Knoxville in 2017, and was granted a doctoral degree in Civil Engineering with a concentration in Transportation Engineering and his master's degree in Statistics.

During his Ph.D. studies, Nima worked as a research assistant in several research projects, including travel time reliability analysis, Exploring the performance of highways and HOV lanes, assessment benefit and cost of federal safety service patrol (FSSP), and protect the queue (PTQ) program, evaluation the impacts of COVID19 on transportation, and crowdsourced data assessment. Nima's research interests include intelligent transportation systems (ITS), transportation operation, real-world crowdsourced data, big data analysis, and connected and automated vehicles.

Over the past four years, Nima received several scholarships and awards, including TSITE Bill D. Kervin, Jr. Scholarship, National Safety Scholarship-Lifesaver Conference, TSITE Scholarship - Bill Moore Jr. Scholarship, and Finalist of national Technology Transfer Tournament, and Graduate Student Senate travel award. He was also selected as president of the Institute of Transportation Engineers (ITE) student chapter at the University of Tennessee and served as an active member of the UTK Woman Transportation Seminar (WTS) student chapter. He was also active as senator of UTK Graduate School Senator and UTK College of Engineering Graduate Student Council Board member.