Exploring the Potentials of Using Crowdsourced Waze Data in Traffic Management: Characteristics and Reliability

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Zhihua Zhang

August 2020

Copyright © 2020 by Zhihua Zhang
All rights reserved.

DEDICATION

This dissertation is dedicated to my parents, Gucai Zhang and Fengmei Lu, and my girlfriend Lu Yang, for all their endless love and support.

ACKNOWLEDGMENTS

Over the past four years, there are so many people I would like to thank for making my Ph.D. study meaningful and memorable.

I would like to express my sincere gratitude to my advisor, Dr. Lee Han. He not only provided guidance and support in my research but gave valuable advice on my life. It was fortunate of me to have a nice advisor, who kindly listens to me talking about my research and provides needed encouragement and insights. I've learned a lot from working with Dr. Han. I am appreciative that he taught me how to conduct research, how to communicate and present ideas, and how to be professional.

Many thanks to my dissertation committee of Dr. Rachel Chen, Dr. Candace Brakewood, and Dr. Russell Zaretzki for their guidance in making an initial proposal to a completed study. It was a great honor of me to collaborate with Dr. Rachel Chen, and I benefited much from her profound knowledge and experience in turning ideas into compelling papers. Many thanks to Dr. Candace Brakewood for her generously sharing her time and ideas on my dissertation, and to Dr. Zaretzki for his valuable advice on my dissertation from the perspective of statistics.

I would also like to express my appreciation to my parents. They are always understandable and supportive of whatever decision I make. Special thanks to my girlfriend, Lu Yang, for all her endless love and support in the past four years. I feel so lucky and blessed to have her in my life.

I also want to thank everyone in Dr. Han's Team: Dr. Cheng Liu, Dr. Stephanie Hargrove, Hyeonsup Lim, Yang Zhang, Bumjoon Bae, Pankaj Dahal, Yuandong Liu, Nima Hoseinzadeh, Harry Zhang, Yangsong Gu, and my amazing lab mates: Kay Boakye, Ziwen Ling, Mohsen Kamrani, Xiaobing Li, Meng Zhang, Behram Wali, Ali Boggs, Abdul Rashid Mussah, Ramin Arvin, Mojdeh Azad, Yi Wen, Jing Guo, and many others.

ABSTRACT

Real-time traffic information is essential to a variety of practical applications. To obtain traffic data, various traffic monitoring devices, such as loop detectors, infrastructure-mounted sensors, and cameras, have been installed on road networks. However, transportation agencies have sought alternative data sources to monitor traffic, due to the high installation and maintenance cost of conventional data collecting methods. Recently, crowdsourced traffic data has become available and is widely considered to have great potential in intelligent transportation systems. Waze is a crowdsourcing traffic application that enables users to share real-time traffic information. Waze data, including passively collected speed data and actively reported user reports, is valuable for traffic management but has not been explored or evaluated extensively. This dissertation evaluated and explored the potential of Waze data in traffic management from different perspectives.

First, this dissertation evaluated and explored Waze traffic speed to understand the characteristics and reliability of Waze traffic speed data. Second, a calibration-free incident detection algorithm with traffic speed data on freeways was proposed, and the results were compared with other commonly used algorithms. Third, a spatial and temporal quality analysis of Waze accident reports to better understand their quality and accuracy was performed. Last, the dissertation proposed a network-based clustering algorithm to identify secondary crashes with Waze user reports, and a case study was performed to demonstrate the applicability of our method and the potential of crowdsourced Waze user reports.

TABLE OF CONTENTS

INTRODUCTION 1

CHAPTER 1 EXPLORATION AND EVALUATION OF PROBE-BASED WAZE TRAFFIC SPEED 3

Abstract 4

Introduction 4

Related work 6

Study Area and Data 7

Results 11

Conclusion 23

References 25

CHAPTER 2 EVALUATING THE RELIABILITY OF WAZE SPEED DATA IN INCIDENT DETECTION ON FREEWAYS 27

Abstract 28

Introduction 28

Related work 30

Data and Methodology 31

Results 35

Conclusion 40

Reference 41

CHAPTER 3 SPATIAL-TEMPORAL QUALITY ANALYSIS OF CROWDSOURCED WAZE INCIDENT REPORTS 44

Abstract 45

Introduction 45

Related work 47

Data and methods 48

Results 52
Conclusion 62
References 64
CHAPTER 4 SECONDARY CRASH IDENTIFICATION USING CROWDSOURCED WAZE USER REPORTS 67
Abstract 68
Introduction 68
Literature Review 70
Network-based spatial-temporal clustering 71
Case Study: The City of Knoxville, Tennessee 75
Conclusion 83
References 85
CONCLUSION 88
VITA 90

LIST OF FIGURES

Figure 1-1 The location of selected RTMS stations and corresponding Waze links along I-40 Eastbound in Knoxville, Tennessee 9
Figure 1-2 Speeds (mph) on I-40 Eastbound from both Waze and RTMS, July 24, 2019 (left) and August 3, 2019 (right) 12
Figure 1-3 RTMS speed vs. Waze speed data collected at several radar stations on both weekday (July 24, 2019, Wednesday) and weekend (August 3, 2019, Saturday) 13
Figure 1-4 Scatter plot for all speed observations from both Waze and RTMS 15
Figure 1-5 Boxplots of the speed difference for three different ranges of Waze speed ... 17
Figure 1-6 Speed comparison at the same stations on July 24, 2019, with highlighting the repeated Waze speed samples persisting at least 5 minutes 19
Figure 1-7 The distribution of length of time a report is repeated in terms of the number of consecutive Waze samples per hour for each link with the entire two months' worth of data. 21
Figure 1-8 The percentage of the number of speed change per hour for each link with the entire two months' worth of data. 22
Figure 2-1 Adaptive thresholds from the proposed algorithm for the incident case on 11/5/2019, MM 3759 37
Figure 2-2 Adaptive thresholds from the proposed algorithm for the incident-free case, on 10/29/2019, MM 379 38
Figure 2-3 Comparison of the performances of AID algorithms 39
Figure 3-1 Spatial distribution of TDOT Crashes (left) and Waze accident reports (right), in Nashville, TN, 2018 49
Figure 3-2 Number of crash matches with varying time and distance difference 51
Figure 3-3 The spatial distribution of matched crashes and the total crashes 53
Figure 3-4 The joint hexagonal histogram of the time difference and distance difference between Waze incident reports and TDOT crash records 55
Figure 3-5 Typical travel time and Actual travel time for Waze accidents (sorted by actual travel time) 60

Figure 3-6 The percentage of reliable Waze accident reports with varying level of significance α 61
Figure 4-1 ST-DBSCAN implementation 74
Figure 4-2 some examples of the obtained clusters 77
Figure 4-3 An example of a primary-secondary crash relationship captured by the proposed method 79
Figure 4-4 An example of the speed contour plot without (a) and with (b) accounting for the recurrent congestion 82

LIST OF TABLES

Table 1-1 Description of selected RTMS stations and corresponding Waze links along I-40 Eastbound in Knoxville, Tennessee 10
Table 1-2 Factors affecting the speed difference at different Waze speed range 17
Table 3-1 The descriptive analysis of time and distance difference for matched TDOT crashes 53
Table 3-2 Significant factors in the likelihood of TDOT crash record being matched by Waze accident reports 57
Table 3-3 Results of Mann-Whitney U-test 60
Table 4-1 The results of secondary crash identification with different methods 81

INTRODUCTION

Road networks are indispensable parts of transportation infrastructures, playing a crucial role in the transport and movement of people, goods, and services. However, as road networks become increasingly complex, there are many concerns for traffic and incident management. Particularly, traffic incidents and traffic jams challenge roadway system efficiency and public safety. Currently, many transportation agencies monitor traffic information with infrastructure-mounted sensors, but several limitations exist such as high installation and maintenance costs and limited geographical coverage. Therefore, we must find alternative data sources that can be integrated into traffic management.

Several emerging data sources, such as crowdsourced data, are available through technological developments. Increasingly, researchers are studying crowdsourced data in traffic management, which demonstrates their potential to improve traffic management by disseminating real-time traffic information and serving as a complementary data source.

Waze is one notable example of traffic information crowdsourcing. It is a crowdsourced platform that enables people to share traffic information (e.g., incidents, traffic jams, and construction reports), efficiently and in a timely manner. Every driver is both a traffic sensor and a beneficiary of the crowdsourced intelligence. Waze collects two types of data, Waze traffic speed and Waze user reports. Waze traffic speed data is passively collected, considering vehicles on road to be sensors, and Waze user reports data are actively reported by users when they encounter traffic incidents such as accidents, traffic jams, and construction areas. The available crowdsourced Waze data helps in traffic monitoring and incident management.

Therefore, it is valuable to efficiently integrate Waze data into traffic management strategies. Nevertheless, crowdsourced Waze data has received little independent evaluation and exploration in the extant literature. To address these issues, this dissertation focuses on using crowdsourced Waze data in traffic management in an efficient way, which is composed of the following four chapters.

- Chapter 1 evaluates the probe-based Waze traffic speed from different perspectives. To understand the characteristics, Waze traffic speed is compared with widely used infrastructure sensor speed.
- Chapter 2 proposes a calibration-free algorithm to detect incidents with Waze traffic speed data on freeways. The results of the proposed algorithm are compared with other widely used algorithms.
- Chapter 3 presents a spatial and temporal quality analysis of Waze accident reports, attempting to fully realize the potential of Waze accident reports.
- Chapter 4 introduces a network-based clustering algorithm to identify secondary crashes using Waze user reports. The results are compared with one of the commonly used secondary crash identification methods.

CHAPTER 1
EXPLORATION AND EVALUATION OF PROBE-BASED WAZE TRAFFIC SPEED

Abstract

Real-time traffic information such as traffic speed is essential to a variety of practical applications. Because of the high installation and maintenance cost of convention data collecting methods, transportation engineers have sought alternative data sources to monitor traffic. Probe-based traffic data, such as Waze produces, could serve as alternative data sources in traffic management, but this source has not been thoroughly explored nor evaluated. Using the 10.8 mile stretch of I-40 in Knoxville, Tennessee, we compared the speed measurements from both Waze and Remote Traffic Microwave Sensors (RTMS) over two months and explored the characteristics of Waze traffic speed data. These are the main findings: 1) These two datasets showed a similar pattern with slight differences. Waze speeds tend to be higher than RTMS speeds for high speed, while Waze speeds are more likely to be similar or even lower than RTMS speeds for low speed; 2) several factors affecting the speed differences between RTMS speeds and Waze speeds were identified, such as Waze speed value, time of day (peak hour vs. non-peak hour), AADT (Annual Average Daily Traffic), and segment length; and 3) Waze reported the same speed for several successive reporting periods if the real-time speed is not available, and Waze speeds had more real-time speed observations during congested times, indicating that Waze speeds are more reliable for congested scenarios. The findings may lead to a better understanding of this source of data and any resulting analysis.

Introduction

The real-time traffic information (e.g., traffic speed and travel time) is valuable for a variety of practical applications, such as incident identification, congestion detection, route choice decision (*1*). To obtain traffic data, various traffic monitoring devices, such as loop detectors, infrastructure-mounted sensors, and cameras, have been installed on road networks. Many state departments of transportation, including Tennessee DOT (TDOT), have used infrastructure-mounted radar sensors to collect real-time traffic information, such as vehicle occupancy and traffic speed. The data collected from these

devices benefit both the public and transportation agencies by informing their decisions. However, because of high installation and maintenance costs, these technologies have only limited coverage of major arterials and highways. Moreover, infrastructure-mounted radar sensors are prone to errors or malfunctions that may cause missing or unreliable traffic information (*2*).

To address the above issues, transportation engineers seek alternative data sources to monitor traffic. Traffic data from several new and promising technologies have become available, such as Bluetooth devices, probe vehicles, cellular devices, automated license plate recognition (LPR), and even social media (*3*). Especially with the increasing use of mobile phones, crowdsourced probe traffic data like Waze traffic data are now available. However, it is not a simple task to extract, collect, and evaluate the traffic data from these technologies since they are not created for collecting traffic data. Also, it is difficult for us to know the computation algorithms, such as data processing, filtering, aggregation, and imputation because private vendors are unwilling to disclose that information. This unwillingness makes it difficult to evaluate, improve, and deploy the collected traffic data.

Waze traffic data is one notable example of crowdsourced, probe-based traffic data, which is estimated by taking users' mobile phones as sensors. Waze traffic speed has the potential to be an alternative data source; however, it has not been sufficiently explored or evaluated. The purpose of this study is to learn about Waze speed data from different perspectives. Specifically, we compared the traffic speed measurements collected from both radar sensors and Waze over two months in Knoxville, Tennessee, and explored the characteristics of Waze traffic speed reports. This rest of this study is structured as follows: Section 2 describes related work about comparing and evaluating traffic data from different data sources. Section 3 describes the data used in this study. Section 4 illustrates the main results obtained from the case study, and the conclusions and discussions are presented in Section 5.

Related work

The current technologies used to collect traffic data are loop detectors and radar sensors, which measure speed at a specific point along the roadway. However, these technologies have limitations such as high installation and maintenance fees, limited coverage, and malfunction issues. Therefore, transportation professionals in both academia and industry have sought alternative approaches to collect traffic data. For example, with the increasing penetration of cellular phones, researchers have been attempting to use cellular phones as sensors to obtain traffic data. Bar-Gera (*1)* compared the speed and travel time data obtained from cellular phones and dual magnetic loop detectors and demonstrated the usefulness of cellular phone-based traffic data for a variety of practical applications. Herrera et al. (*2)* performed a field experiment to demonstrate the effectiveness of GPS-enabled mobile phones system to collect traffic data and found that a 2-3% penetration of cell phones in the driver population is capable of providing accurate measurements of traffic speed. Later on, probe vehicle traffic data, measuring traffic speed using the vehicles along a certain route, has been actively collected and used. (*4; 5*).

Much research explored probe-based traffic data, especially for measuring accuracy and reliability (*6-8*). For example, Lattimer and Glotzbach (*9)* measured the accuracy of third-party travel time data by comparing them against ground truth data obtained through floating car methodology. Kim and Coifman (*6)* compared the INRIX speed data against the loop detector data and found that INRIX speeds can have up to six minutes' latency compared with the loop detector measurements. Ahsani et al. (*8)* investigated the speed bias, coverage, and congestion detection accuracy of INRIX data.

Waze is a social navigation application where people can share traffic information. Waze provides two kinds of data, passively collected traffic speed data and actively reported user reports data such as incident reports and jam reports. Limited yet increasing studies have explored the possibility of using Waze data as an alternative source in the transportation field thanks to its low cost, real-time capacity, and reasonable accuracy (*10*). For example, to explore the potential of integrating Waze incident data into the official incident data, dos Santos, Davis Jr and Smarzaro (*11)* matched the two traffic accident datasets from Waze and BHTRANS (Belo Horizonte Transport and

Transit Company). Amin-Naseri et al. (*12)* explored the characteristics of the Waze incident data and compared it with several other common data sources in traffic management.

However, previous studies explored only the crowdsourced Waze user reports, including accident reports or jam reports; few explored and evaluated the crowdsourced probe-based Waze speed data. To fill this gap, this study examined the characteristics of Waze traffic speed and evaluated it by comparing Waze traffic speed against radar sensor traffic speed.

Study Area and Data

The study uses traffic speed data from Waze, NPMRDS (National Performance Management Research Data Set), and TDOT RTMS (Remote Traffic Microwave Sensors) for highway segments in Knoxville, Tennessee. Waze speed data is available at a one-minute interval, RTMS speed data is available at a 30-second interval, and NPMRDS speed data is available at a five-minute interval. In this study, we compared RTMS speeds with Waze speeds, and the NPMRDS speeds were used only for visualization purposes.

RTMS collects traffic information such as traffic count, speed, and occupancy for each lane every 30 seconds. In Tennessee, over 200 detector stations are installed on interstate highways for both directions, including two major highways, I-40 and I-75. Twenty-six stations installed along the 10.8 miles long I-40 eastbound segment, ranging from mile marker 374.3 (west end) to 385.1 (east end), were selected. The traffic speed data for the selected stations were collected every 30 seconds and were aggregated to one minute for consistency of the temporal scale of Waze speed data.

Waze traffic data was collected from Waze API, a localized JSON GeoRSS feed (*13*). The JSON file contains traffic data for each Waze road link such as traffic speed, road segment length, and travel time, and it was downloaded at a one-minute interval. Waze provides the functionality to customize the link: Namely, the user can specify the start point and end point for each link, which facilitates extracting Waze speed for specific road segments. Then, we customized the corresponding Waze links on Interstate

40 to compare the traffic speed measurements from Waze and RTMS. Sixteen links were found to be associated with the above selected RTMS stations. Table 1-1 presents the description of the selected RTMS stations and the corresponding Waze links along I-40 Eastbound in Knoxville, and Figure 1-1 shows the locations of selected RTMS stations and corresponding Waze links, also along with I-40 Eastbound in Knoxville.

In this study, we analyzed the two month's worth of traffic speed measurements in July and August 2019, along the I-40 interstate highway in Knoxville. Several ways have been explored to examine the difference between two large speed datasets (*3; 14; 15*); however, more detailed examination (e.g., graph presentation at the time-space dimensions) may reveal additional valuable insights, when comparing two huge datasets that represent complex phenomena (*1*). Therefore, in this study, we first compared the traffic speed measurements from both Waze and RTMS by time and location using heatmap and scatter plot; then, we investigated the factors affecting the speed difference between Waze and RTMS with regression analysis. Last, we explored the frequency of Waze reporting real-time traffic speed.

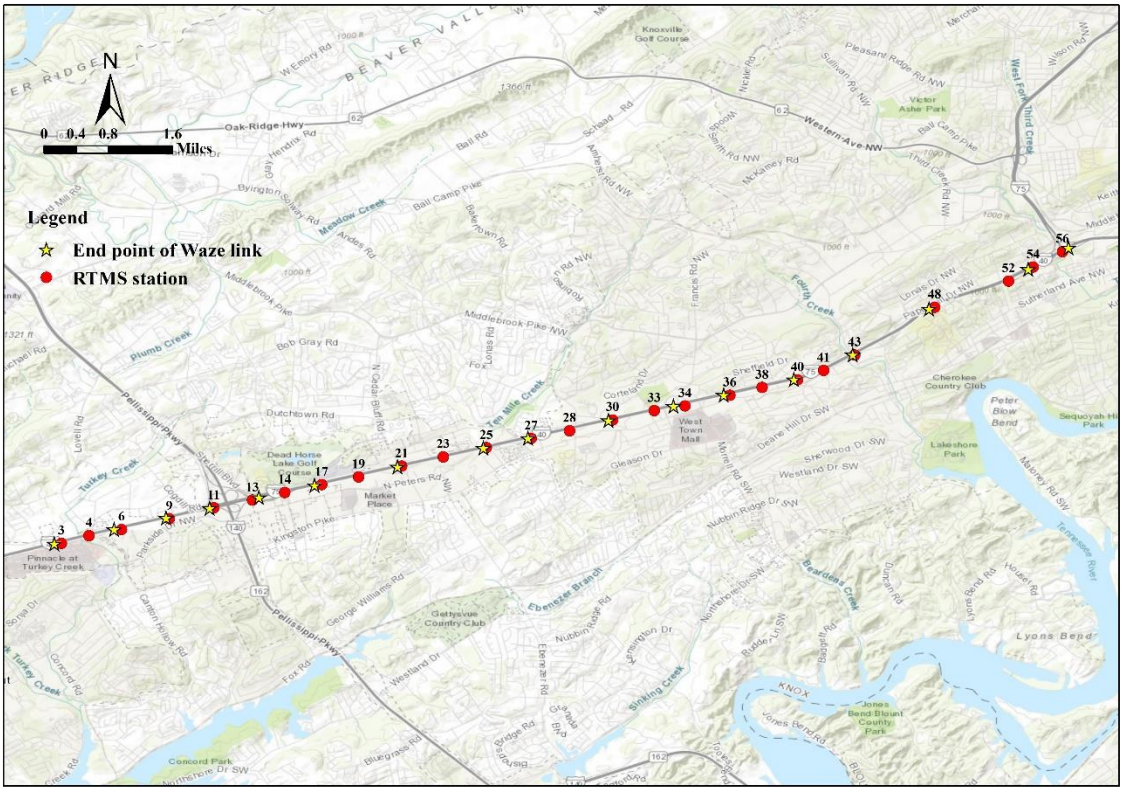


Figure 1-1 The location of selected RTMS stations and corresponding Waze links along I-40 Eastbound in Knoxville, Tennessee

Table 1-1 Description of selected RTMS stations and corresponding Waze links along I-40 Eastbound in Knoxville, Tennessee

RTMS Station ID	Mile Marker	Direction	Waze Link ID	Link length (mile)
3	374.3	Eastbound	E3-4	0.6
4	374.6	Eastbound		
6	374.9	Eastbound	E6	0.5
9	375.5	Eastbound	E9	0.5
11	375.9	Eastbound	E11-13	0.7
13	376.3	Eastbound		
14	376.6	Eastbound	E14	0.4
17	377.0	Eastbound	E17-19	0.9
19	377.4	Eastbound		
21	377.9	Eastbound	E21-23	0.9
23	378.3	Eastbound		
25	378.8	Eastbound	E25	0.4
27	379.2	Eastbound	E27-28	0.8
28	379.6	Eastbound		
30	380.0	Eastbound	E30-33	0.8
33	380.4	Eastbound		
34	380.8	Eastbound	E34	0.5
36	381.1	Eastbound	E36-38	0.7
38	381.5	Eastbound		
40	381.9	Eastbound	E40-41	0.6
41	382.2	Eastbound		
43	382.5	Eastbound	E43	1
48	383.6	Eastbound	E48-52	1.1
52	384.4	Eastbound		
54	384.7	Eastbound	E54-56	0.4
56	385.1	Eastbound		

Results

Speed difference

To compare the speed measurements from Waze and RTMS, we visualized the data at the spatial and temporal dimensions. Figure 1-2 shows the speed data by time and location on I-40 Eastbound in Knoxville, Tennessee. The figure shows the speed data for two different days, in which the left part and right part show the speed for the entire day of Wednesday, July 24, 2019, and Saturday, August 3, 2019, respectively. We chose these two dates because they are two atypical patterns with traffic incident occurred for both weekday and weekend, respectively. The top part shows the speeds from RTMS, while the bottom part shows the corresponding speeds from Waze. The horizontal axis represents the time of day, from 00:00 to 24:00, and the vertical axis represents the selected RTMS stations along I-40 interstate highway. White areas in the figures show the missing values, and other colors indicate the corresponding speed. The RTMS speed is location-based speed; we aggregated and averaged the speed of all vehicles on all lanes at the detector's location, for a one-minute interval. While the Waze speed is link-based, we collected the speed data directly from Waze API. The Waze speed was also collected at a one-minute interval.

As is shown, some values are missing both for RTMS data and Waze data. But the missing values are not at the same time and location, allowing us to impute missing values for one data source using the other data source (*16*). For July 24, 2019, the speed pattern suggested traffic congestion occurred around 8 AM at milepost 378.8, lasting for about one hour. Another instance of severe congestion was also found starting around 4:00 PM at milepost 379.2, and later another instance of congestion occurred (perhaps an incident) at around 5:30 PM at milepost 382.5. The first congestion dissipated around 6:00 PM while the second one eventually dissipated at 7:00 PM. Both the RTMS speeds and Waze speeds show a similar pattern. Similarly, the right side of Figure 1-2 shows the speed pattern for August 3, 2019. Two datasets show a similar pattern, indicating light congestion around 10:00 AM, which may have been caused by an incident.

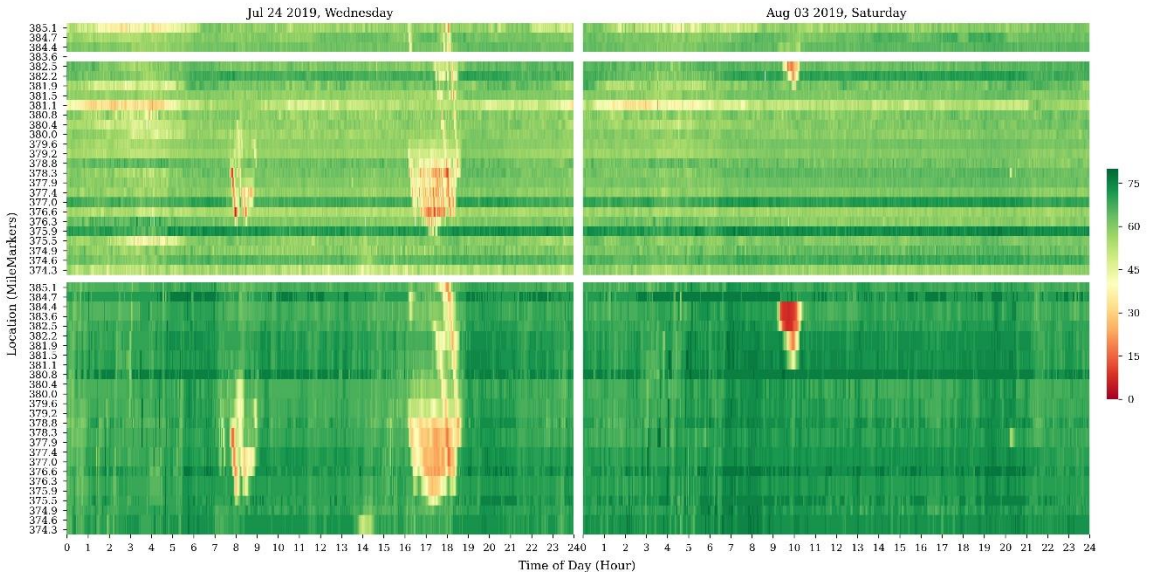


Figure 1-2 Speeds (mph) on I-40 Eastbound from both Waze and RTMS, July 24, 2019 (left) and August 3, 2019 (right)

Next, we compared and plotted the time-series speeds at several radar sensor stations to investigate if there was any difference between Waze and RTMS speeds. Figure 1-3 shows the time-series RTMS and Waze speeds for four randomly selected stations for both a weekday (July 24, 2019; Wednesday) and a weekend day (August 3, 2019; Saturday). The NPMRDS speeds are shown in the figure as well for visualization. From the figures, RTMS speeds and Waze speeds show a similar pattern but with a significant difference. Both on weekday and weekend, Waze speeds are always higher than the RTMS speeds when the speed value is high, while for low speeds, Waze speeds are similar or even slightly less than RTMS speeds. It may indicate Waze speeds are more reliable for congestion scenarios because of the considerable number of sample vehicles in the scenario. These findings may suggest that RTMS and Waze have different methods of computing velocity, and they have their measurement errors. Meanwhile, Waze data is affected significantly by the sample size, and when there are few sample vehicles, the Waze speed may have significant measurement error.

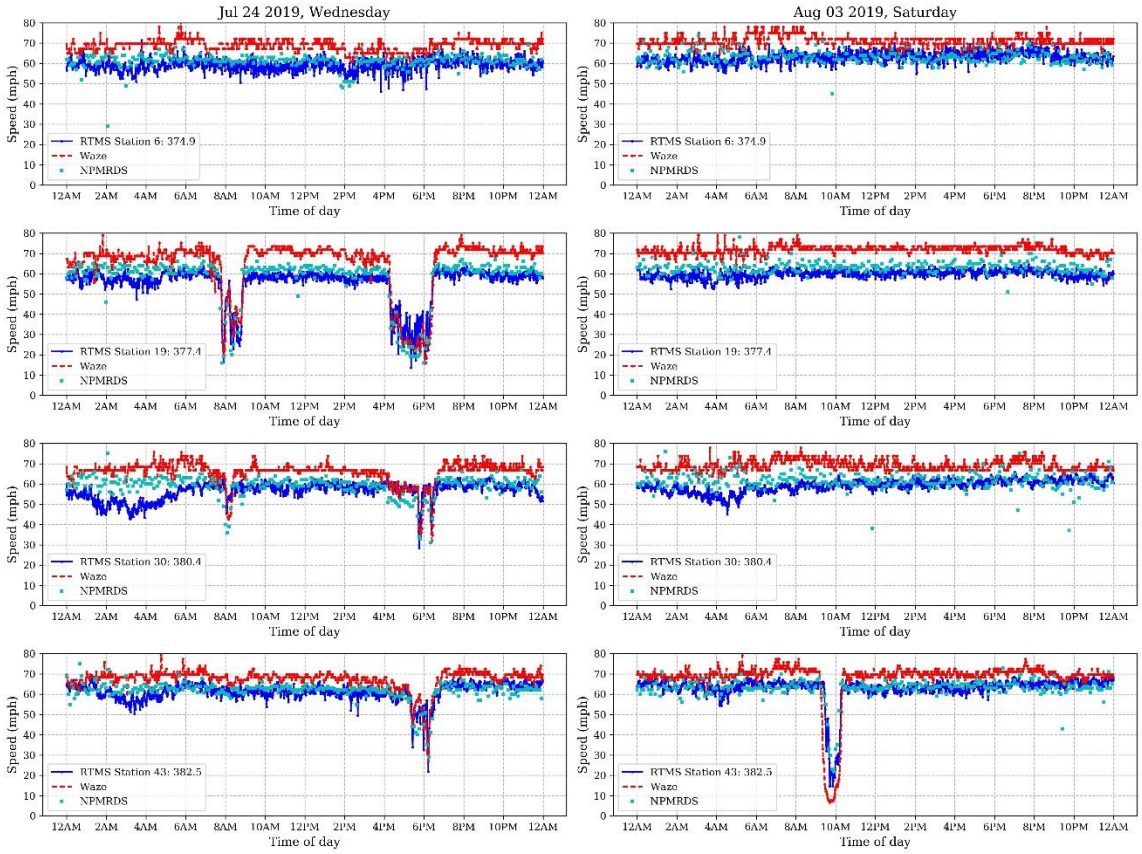


Figure 1-3 RTMS speed vs. Waze speed data collected at several radar stations on both weekday (July 24, 2019, Wednesday) and weekend (August 3, 2019, Saturday)

We extracted all the speed observations for the sixteen Waze road links from Waze and RTMS in those two months and obtained 1,506,414 observations. Figure 1-4 shows the scatter plot for the observations. These two datasets, RTMS and Waze, have a relatively high correlation ($r = 0.65$). From the figure, most of the speed observations in both Waze and RTMS speed data are near 60 mph and show a circular shape, and a considerable number of repeated values for Waze can be found, suggesting that Waze may not able to report real-time speed every minute. Besides, Waze speeds tend to be higher than RTMS speeds for high speeds, while RTMS speeds are similar to or even higher than Waze speeds for low speeds. It may be caused by sample bias in calculating the speed for Waze. Figure 1-5 shows the speed difference (RTMS speed minus Waze speed) for three different ranges (0–45 mph, 45–55 mph, and greater than 55 mph) of Waze speed. 45 mph is widely considered as the breakdown speed on highways (*8; 17*), and 55 mph is the speed limit for the road segments. From the figure, we can observe that as the Waze speed increase, the interquartile ranges become smaller, meaning less variation in speed difference for high Waze speeds. Also, this figure reaffirms that Waze speed tends to be greater than RTMS speed for high speeds while it is more likely to be less than RTMS speed for low speeds. To determine the effect of Waze speed levels in speed difference statistically, we performed the one-way analysis of variance (ANOVA) on five predefined groups of Waze speed. Based on ANOVA analysis results ($F = 220{,}995, p < 0.0001$), the mean speed difference for all groups differed significantly.

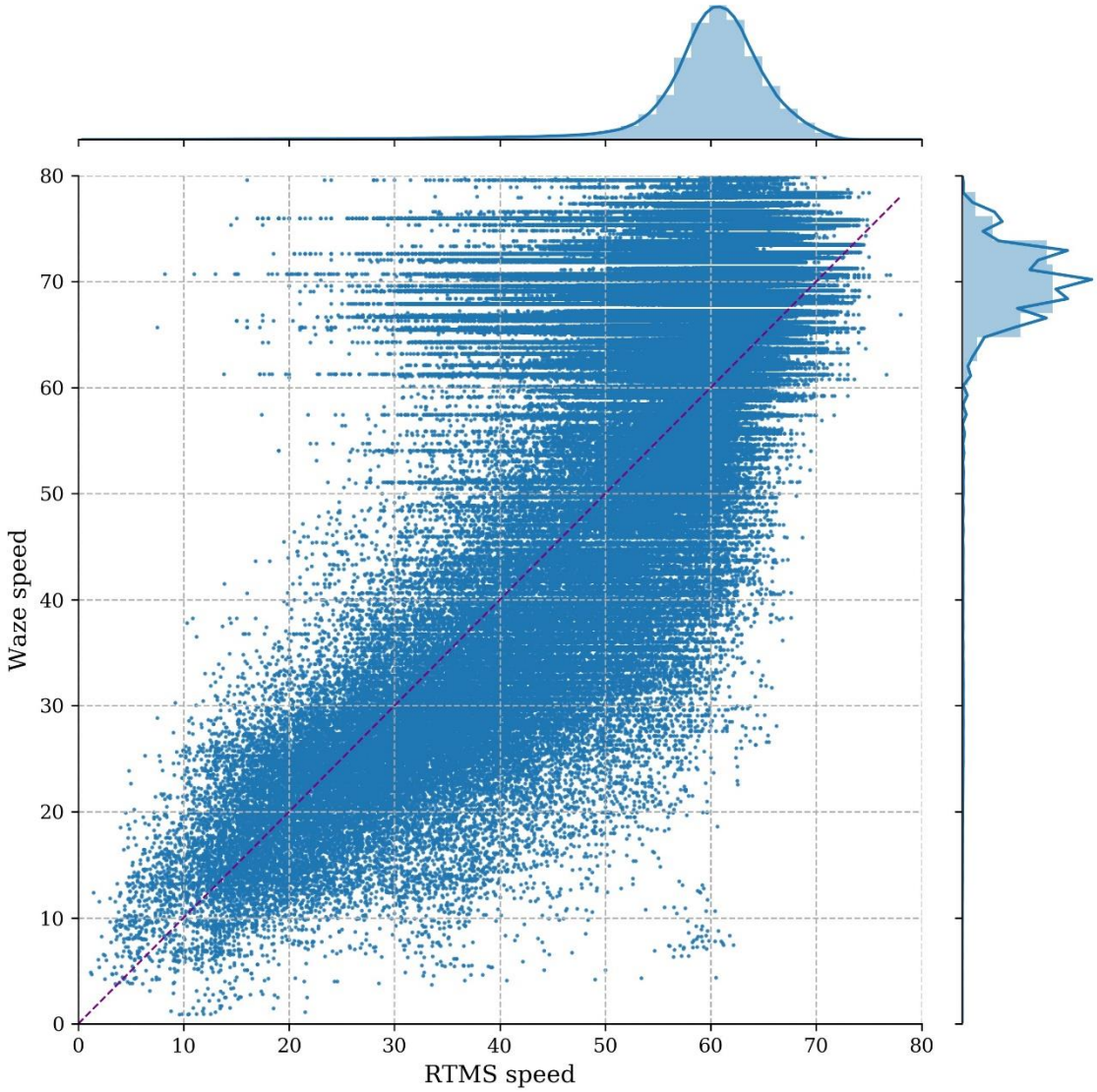


Figure 1-4 Scatter plot for all speed observations from both Waze and RTMS

A linear regression model was performed to identify the factors affecting speed difference. Since the speed difference for Waze speeds greater than 55 mph has a small variation (less than 3 mph), we only performed regression analysis to ascertain the effects of Waze speed value, time of day (peak hour vs. non-peak hour), AADT (Annual Average Daily Traffic), and segment length on the magnitude of the speed difference for Waze speeds less than 55 mph. The results are shown in Table 1-2. From the results, we can observe that Waze speed value negatively affected the speed difference for two models, namely, as Waze speed value increases, the speed difference will decrease, confirming the finding observed in Figure 1-5. The effect of traffic volume in terms of AADT was examined, and the results showed that Waze links with high traffic volume have less speed difference. It can be attributed partly to the fact that high traffic volume means many Waze users on the road, thus resulting in a more accurate Waze speed. The time of day was also found to negatively affect the speed difference; namely, the smaller difference in speed was found during peak hours. Moreover, the effect of time of day on speed difference is higher for Waze speeds that are less than 45 mph, which may be because observations with high Waze speeds are not significantly affected by peak hours. For road segment length, longer road segments tend to have higher speed differences for observations with Waze speeds less than 45 mph. We speculate that Waze speed would be more sensitive to the change of road segment length for lower Waze speed observations.

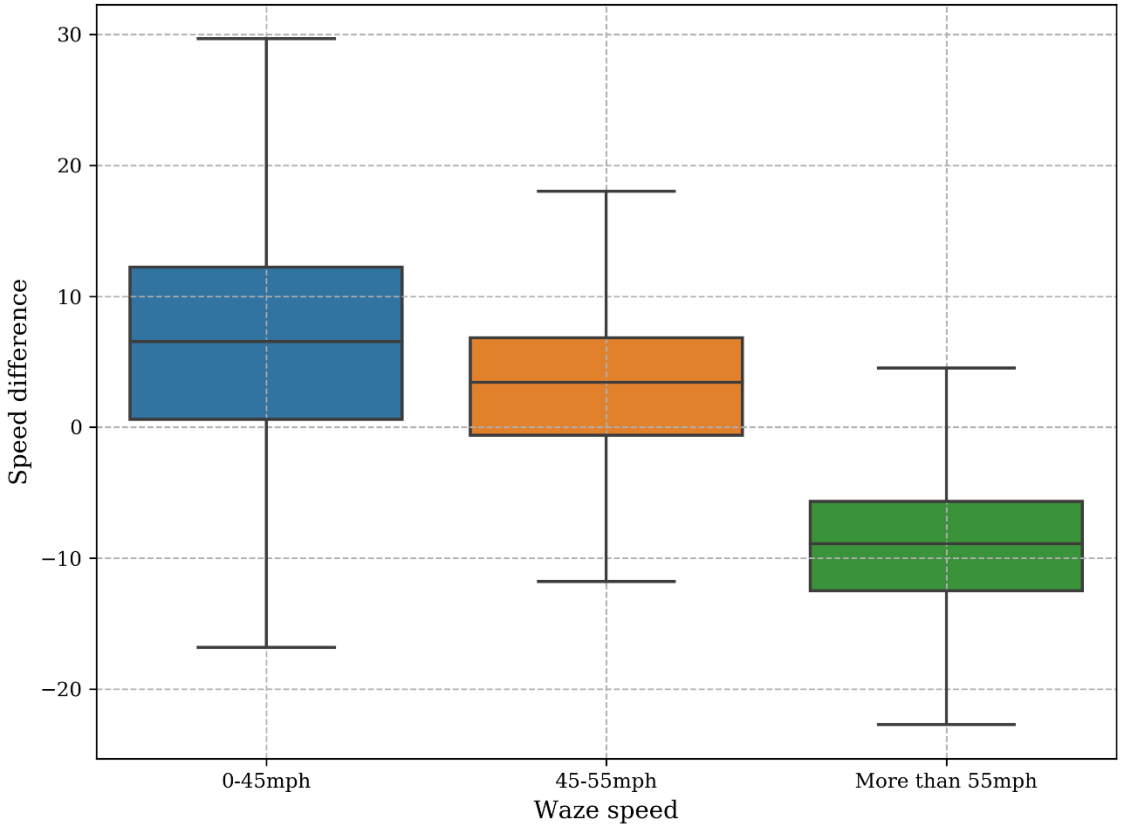


Figure 1-5 Boxplots of the speed difference for three different ranges of Waze speed

Table 1-2 Factors affecting the speed difference at different Waze speed range

	Model 1: <=45 mph	Model 2: 45-55 mph
Waze speed value	-0.62***	-0.79***
AADT	-0.0002***	-0.0001***
Time of the day (peak hour vs. non-peak hour)	-3.06***	-0.37***
Segment Length	1.64**	0.36
R-squared	0.26	0.28
Number of observations	44,997	17,212

Note: * p<.1, ** p<.05, ***p<.01

Repeated Waze speed

While Waze reports speed for a given Waze link every minute, Waze seems to report the same speed for several successive reporting periods if the real-time speed is not available. Therefore, we investigated the pattern of repeated Waze speed, namely, how often Waze would report the real-time speed. Figure 1-6 shows the speed comparison between Waze and RTMS speeds at the same RTMS stations on July 24, 2019, and Waze speed samples persisting at least five minutes are highlighted. The figure indicates that although Waze reports speeds around every minute, there are a significant number of Waze speed observations merely repeating from the previous sample. Additionally, more repeated Waze speed observations are found to occur during off-peak hours, which makes sense since there may not be enough samples to obtain real-time speed during off-peak hours.

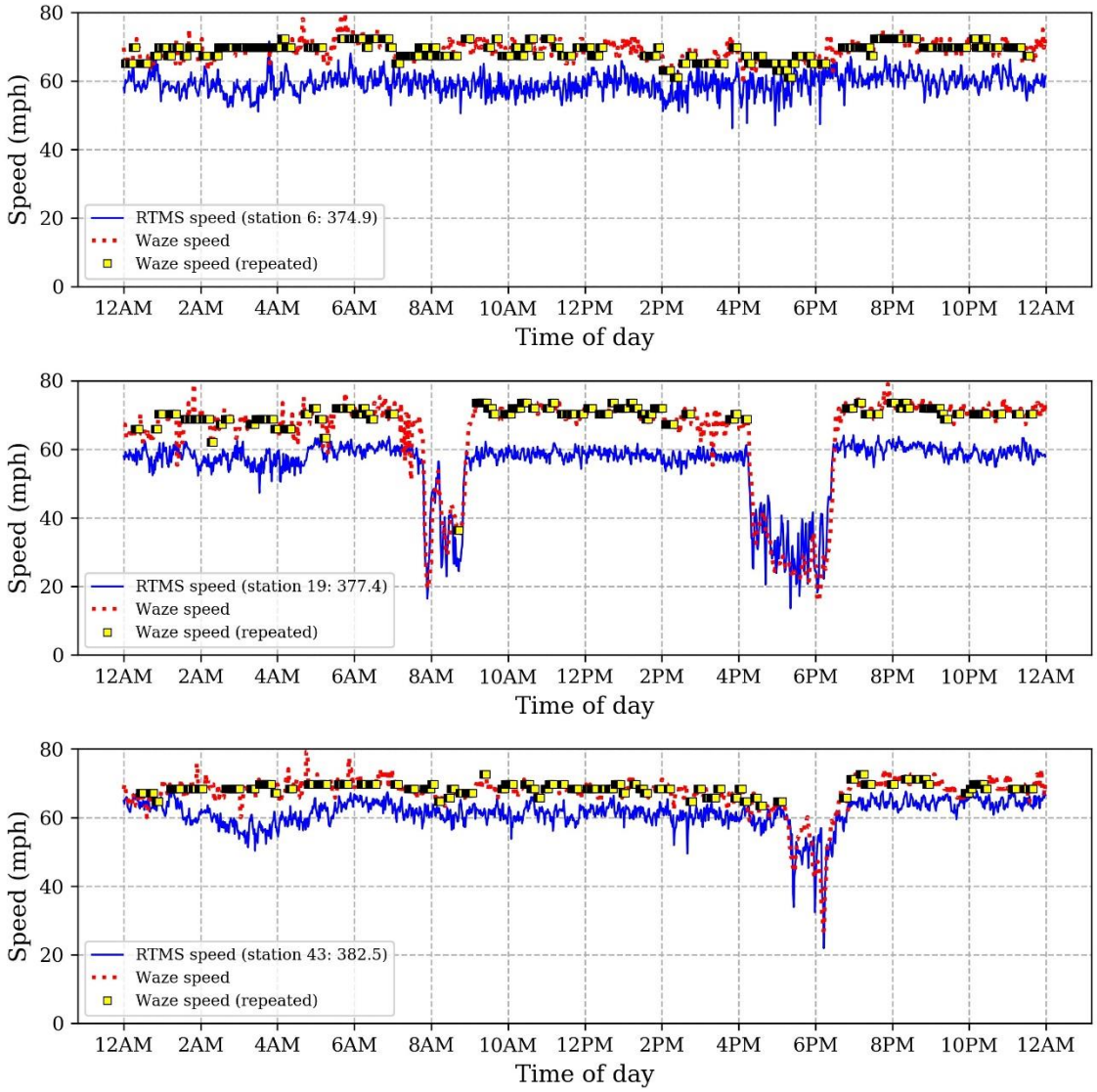


Figure 1-6 Speed comparison at the same stations on July 24, 2019, with highlighting the repeated Waze speed samples persisting at least 5 minutes

For any Waze road link, there should be 60 reported speeds per hour since we downloaded the Waze speed every minute, although some speeds may be missing. To thoroughly investigate the pattern of Waze repeated speeds, we computed the duration of repeated Waze speed observations, namely, the number of consecutive same Waze speed observations. Similar to (*6*), if the repeated Waze speed observations fall in different hours, then the duration would be the total number of repeated speed observations. For example, assume there are ten repeated Waze speed observations with five observations in the previous hour and five observations in the following hour, then each hour has five observations with a duration of ten minutes.

We could then plot the distribution of length of time a report is repeated in terms of the number of consecutive Waze speed observations per hour for each link. We categorized the length of time a Waze speed measurement is repeated into four categories, which are one minute (*Unique sample*), two to five minutes (*2-5 minutes*), greater or equal to six minutes (*6+ minutes*), and missing value (*Missing value*). We then cumulatively computed the duration of repeated samples for each category for the same hour of the day at the same Waze link for the entire two months.

Figure 1-7 shows the distribution plot for several Waze links for the entire two months' worth of Waze speed data. From the figure, we can observe that the pattern of Waze repeated speed observations varies by Waze link, but with several similarities. A significant number of unique samples are found during peak hours. The median of repeated Waze speed duration is primarily two to five minutes, indicating that Waze may not collect the real-time data every minute, but every two to five minutes.

Another way to measure the sampling period of Waze data is to compute the count of speed change in one hour, namely, how many unique speed values occur in one hour, including the same, but not consecutive, speed value. Figure 1-8 depicts the percentage of speed change per hour plot for several Waze links with the whole two months Waze speed data. We found that the percentage is between 25%-50%, which means the effective Waze sampling period tends to be two to four minutes.

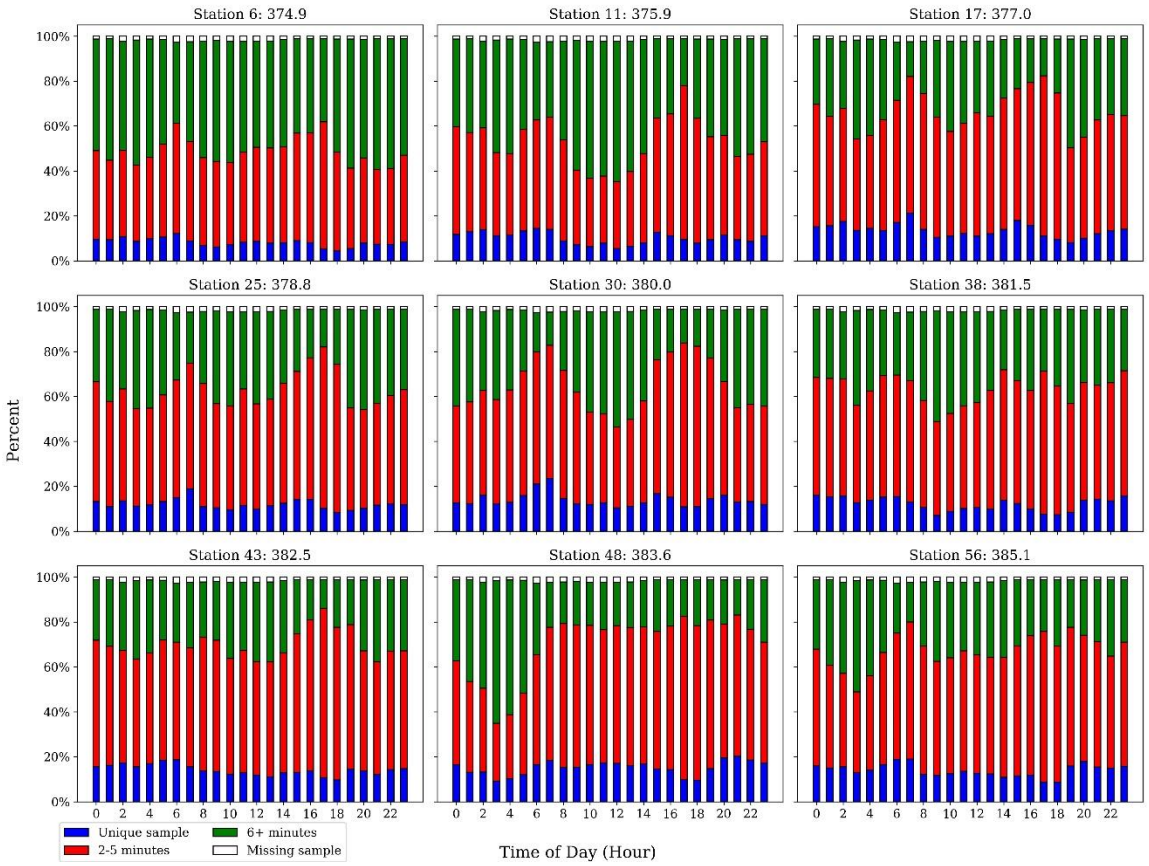


Figure 1-7 The distribution of length of time a report is repeated in terms of the number of consecutive Waze samples per hour for each link with the entire two months' worth of data.

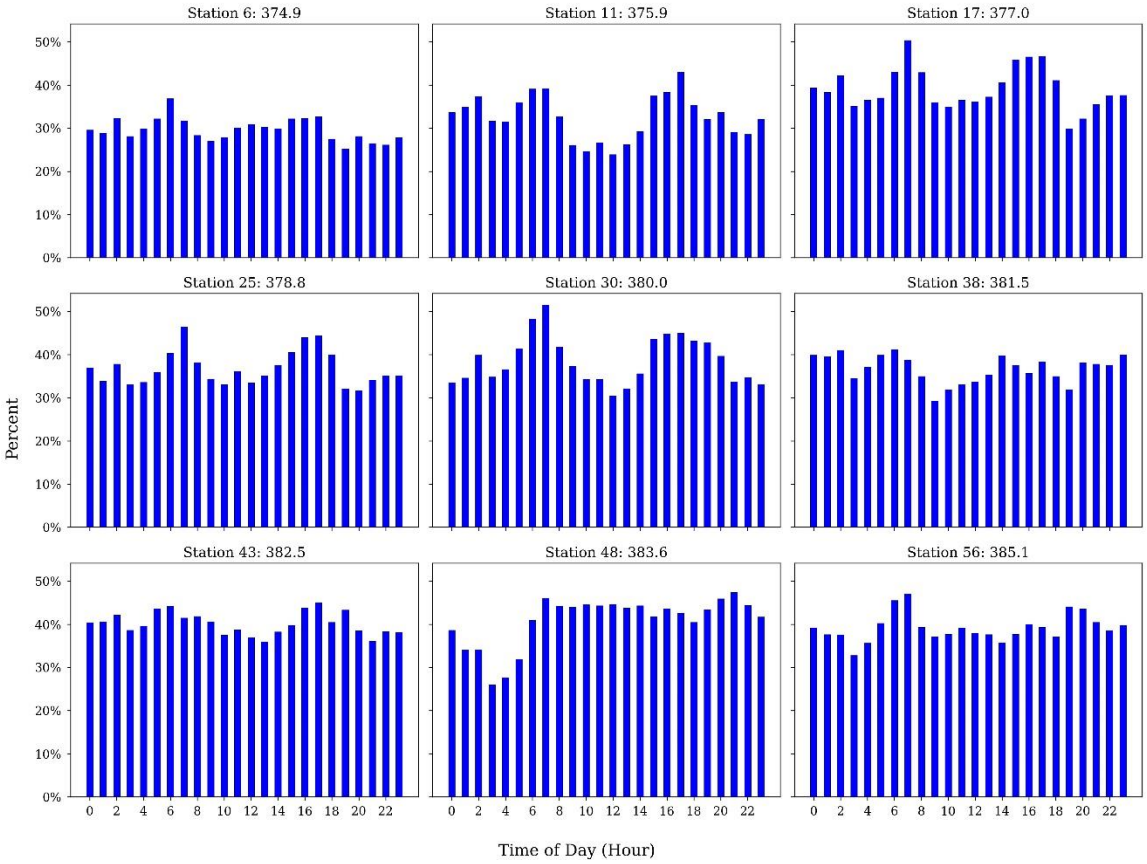


Figure 1-8 The percentage of the number of speed change per hour for each link with the entire two months' worth of data.

Conclusion

Real-time traffic information such as traffic speed is essential to a variety of practical applications, such as incident identification, congestion detection, and route choice decision. Because of the high installation and maintenance cost of conventional data collecting methods, transportation engineers have sought alternative data sources to monitor traffic. Crowdsourced, probe-based traffic data like Waze traffic data could well serve as alternative data sources in traffic management, yet these sources have not been well explored or evaluated. Over two months, this study compared the speed measurements from both Waze and Remote Traffic Microwave Sensors (RTMS) for a segment of 10.8 miles of I-40 in Knoxville, Tennessee, and explored the characteristics of Waze traffic speed data.

For the speed comparison, we found these two datasets showed a similar pattern with slight differences. Waze speeds tend to be higher than RTMS speeds for high speeds, while Waze speeds are more likely to be similar or even lower than RTMS speeds for low speeds. Several factors affecting the speed differences between RTMS speeds and Waze speeds were identified, such as Waze speed value, time of day (peak hour vs. non-peak hour), AADT, and segment length. Moreover, Waze reported the same speed for several successive reporting periods if the real-time speed was not available, and Waze may not collect the real-time data every minute, but every two to four minutes. Also, Waze speeds had more real-time speed observations during congested times, indicating that Waze speeds are more reliable for congested scenarios.

The goal of this study is not to identify the most accurate measurement method since we do not have the "ground truth". Waze data provide sampled speed data with a high coverage area, influenced by the sample size and measurement equipment accuracy. The RTMS data have a limited coverage area, influenced by missing values and the speed aggregation method. However, it is expected that the integration of both static (RTMS) and mobile sensors (Waze) should be more accurate than each of them individually (*2; 18*). Also, the integration of multiple speed datasets can be used for many transportation applications and would achieve better performance. Therefore, the exploration and evaluation of Waze data are essential to better understand this source of data and any

resultant analysis. Moreover, given the positive benefits provided by Waze traffic data, such as high coverage, low missing value rate, and further improvements and enhancements in data collection and computation, we anticipate the increasing application of Waze traffic data in traffic management in the near future.

References

[1] Bar-Gera, H. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C-Emerging Technologies,* Vol. 15, No. 6, 2007, pp. 380-391.

[2] Herrera, J. C., D. B. Work, R. Herring, X. G. 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, 2010, pp. 568-583.

[3] Hargrove, S. R., 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. 2594, 2016, pp. 73-82.

[4] Nanthawichit, C., T. Nakatsuji, and H. Suzuki. Application of probe-vehicle data for real-time traffic-state estimation and short-term travel-time prediction on a freeway. *Transportation Data Research,* Vol. 1855, No. 1855, 2003, pp. 49-59.

[5] Altintasi, O., H. Tuydes-Yaman, and K. Tuncay. Detection of urban traffic patterns from Floating Car Data (FCD). *19th Euro Working Group on Transportation Meeting (Ewgt2016),* Vol. 22, 2017, pp. 382-391.

[6] Kim, S., and B. Coifman. Comparing INRIX speed data against concurrent loop detector stations over several months. *Transportation Research Part C-Emerging Technologies,* Vol. 49, 2014, pp. 59-72.

[7] Feng, W., 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. 2178, 2010, pp. 67-78.

[8] Ahsani, V., 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, 2019, pp. 103-119.

[9] Lattimer, C., and G. Glotzbach. Evaluation of third party travel time data.In *ITS America 22nd Annual Meeting & ExpositionITS America*, 2012.

[10] Gu, Y., Z. 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, 2016, pp. 321-342.
[11] dos Santos, S. R., C. A. Davis Jr, and R. Smarzaro. Integration of data sources on traffic accidents.In *GeoInfo*, 2016. pp. 192-203.
[12] Amin-Naseri, M., P. Chakraborty, A. Sharma, S. B. Gilbert, and M. Y. Hong. Evaluating the Reliability, Coverage, and Added Value of Crowdsourced Traffic Incident Reports from waze. *Transportation Research Record,* Vol. 2672, No. 43, 2018, pp. 34-43.
[13] WAZE. *Connected Citizens Program*. https://www.waze.com/ccp. Accessed December 6, 2017.
[14] Liu, X., S. Chien, and K. Kim. Evaluation of floating car technologies for travel time estimation. *Journal of Modern Transportation,* Vol. 20, No. 1, 2012, pp. 49-56.
[15] Smith, B. L. Purchasing travel time data - Investigation of travel time data service requirements. *Traffic Signal Systems and Regional Systems Management 2006,* Vol. 1978, No. 1978, 2006, pp. 178-183.
[16] Bae, B., H. Kim, H. Lim, Y. D. Liu, L. D. Han, and P. B. Freeze. Missing data imputation for traffic flow speed using spatio-temporal cokriging. *Transportation Research Part C-Emerging Technologies,* Vol. 88, 2018, pp. 124-139.
[17] TRB. Highway capacity manual 6th edition. A guide for multimodal mobility analysis.In, Transportation Research Board of the National Academy of Science Washington, DC, 2016.
[18] Westerman, M., R. Litjens, and J.-P. Linnartz. Integration of probe vehicle and induction loop data: Estimation of travel times and automatic incident detection. 1996.

CHAPTER 2

EVALUATING THE RELIABILITY OF WAZE SPEED DATA IN INCIDENT DETECTION ON FREEWAYS

Abstract

Early detection of incidents is valuable for incident management, motivating studies to develop quick and accurate automatic incident detection (AID) algorithms. As the availability of probe traffic data increases, it can be used to detect traffic incidents. In this study, we explored and evaluated the reliability of Waze speed data in incident detection on freeways. Specifically, we proposed a new calibration-free algorithm to detect incident with Waze speed data, and we compared the performance with other AID algorithms, in terms of detection rate (DR), false alarm rate (FAR), and mean time to detect (MTTD). From the results of the case study on the I-40 freeway in Knoxville, we found that Waze speed data is accurate enough to be used in incident detection with a high DR of 90.0%. Also, our proposed algorithm performed better in terms of DR and FAR, but with a slightly high MTTD. Overall, the results showed the applicability of our proposed algorithm and the reliability of Waze speed data in incident detection on freeways, which can improve incident management systems operated by transportation agencies.

Introduction

Road networks are indispensable components of transportation infrastructures that are crucial to the transport and movement of people, goods, and services. Traffic incidents have been intensively studied by researchers and traffic engineers, due to the negative impacts of traffic incidents on public safety and traffic operation. Consequently, accurate incident detection is valuable and a primary objective of the Intelligent Transportation System (ITS), which can help reduce congestion, increase safety, and improve daily operation efficiency (*1; 2*).

Many transportation agencies and state departments of transportation (DOTs) have installed fixed-mounted sensors, loop detectors, or cameras to monitor traffic. However, the fixed-mounted equipment usually has high installation and maintenance costs, thus limiting their coverage in transportation road networks. Researchers and practitioners continuously seek alternative data sources to use in traffic monitoring.

Advanced technologies have produced applicable probe traffic data, such as INTRIX, HERE, and WAZE. Much research has explored and evaluated INRIX and HERE data in various aspects (*3-5*). In incident detection, several studies have found that INRIX has relatively high reliability in incident detection (*3; 6*). This is a compelling reason to revisit AID algorithms with the new data source, and Waze speed data is one kind of emerging probe vehicle datasets, although it is under-explored and -evaluated.

Various automatic incident detection (AID) algorithms have been developed by researchers, such as pattern recognition algorithms, outlier mining methods, artificial neural networks, comparative methods, wavelet transformation, and other machine learning methods. However, many of these algorithms are hard to implement by TMCs because algorithm calibrations are usually problematic since it is difficult to get the historical incident information (*7*). Moreover, the vast majority of AID algorithms are limited in universality or transferability, lacking the ability to obtain satisfactory results at different traffic scenarios with little or no recalibration efforts (*8; 9*). Therefore, it needs to revisit the AID algorithms and propose a new calibration-free AID algorithm that can perform well universally.

The main goal of this study is to explore and evaluate the reliability of Waze speed data in incident detection on freeways. We proposed a new calibration-free algorithm to detect incident with Waze speed data, and compared the performance with several other calibration-free AID algorithms, in terms of detection rate (DR), false alarm rate (FAR), and mean time to detect (MTTD).

The next section gives an overview of related work on AID algorithms and corresponding performance measures. Section 3 describes the data and methodology used in this study. The detailed results are shown in section 4, and the conclusions and future work are presented in section 5.

Related work

AID algorithms have been researched intensively, and many AID algorithms have been developed, such as outlier mining methods (*10-14*), wavelet theory algorithms (*15*), artificial neural networks (*16; 17*), fuzzy set theory (*18; 19*), and machine learning methods (*20-22*). The outlier mining methods are simple yet effective calibration-free methods to detect incidents. These have several advantages over other methods: it is a segment/station-specific algorithm; namely, each segment and station has its parameters of the algorithm, thus making calibration and tuning of the parameters quite easy. This algorithm does not need historical incident information, which makes it more attractive since it may be difficult to collect historical incident data (*23*). For example, the California algorithm and its derivatives are commonly used outlier mining methods that compute three values based on the occupancy data of vehicle detection stations and then compare these three values with the predefined thresholds to determine an incident. Dudek, Messer and Nuckles (*12)* proposed and evaluated the standard normal deviates (SND) method to detect incidents, and demonstrated the effectiveness of this method in freeway incident detection. Castro-Neto et al. (*9)* proposed a new, simple, and calibration-free incident detection algorithm with traffic occupancy data and achieved better performance compared with several other incident detection algorithms. The method can also be applied with traffic speed data.

However, the data used by most of the existing AID algorithms are either inductive loop detectors or radar sensors. Limited studies have been conducted to detect incident with probe vehicle traffic data, which can compensate for several limitations of loop detectors and radar sensors. For example, Balke, Dudek and Mountain (*11)* used the standard normal deviates (SND) to generate incident-free traffic speed thresholds for every segment, every time of day and every day of the week and declared traffic incidents if the speed observations exceed the computed thresholds. Further studies have improved the SND method by considering other information, such as incident reports and spatial-temporal relationships (*24*). Chakraborty, Hegde and Sharma (*6)* detected lane-blocking incidents with INRIX data with univariate threshold methods, such as SND (Standard Normal Deviates) method and IQD (Inter-Quartile Distance) method. Ahsani et al. (*3)*

investigated the accuracy and reliability of INRIX data in congestion detection and investigated the factors affecting the performance of congestion detection. Though INRIX is an important probe data source, other alternative data sources, such as Waze, also have the potential of being used in incident management.

Waze is a crowdsourcing platform where people can share traffic information like traffic incident reports and traffic jam reports. Also, it can gather and collect the speed data from vehicles on the road. Increasing studies have been conducted to use Waze data as an alternative source in traffic management, but limited studies have explored and evaluated the reliability of Waze traffic speed in incident detection. Therefore, to evaluate the reliability of Waze traffic speed data in incident detection, we proposed a new, calibration-free algorithm to detect incidents with Waze speed data, and we compared its performance with several AID algorithms. We also compared the performance of Waze traffic speed data and radar sensor speed data in incident detection with the same AID algorithms.

Data and Methodology

Data

Multiple datasets were used in this study, including traffic speed data, traffic occupancy data, and incident data from eastbound Interstate 40 in Knoxville, Tennessee, USA. I-40 is one of the major freeways in and out of the city of Knoxville, carrying a large volume of traffic, especially during peak hours. It is important to detect incidents early to mitigate their effects, though it may be difficult to separate the traffic incidents from recurring congestion and develop a reliable AID framework.

Five months' worth of Waze speed data, collected from June 1, 2019, to November 30, 2019, in Knoxville, Tennessee was used in this study. The Waze speed data was collected from Waze API, and at a one-minute interval, the XML file containing real-time traffic data for each segment was downloaded, totaling 1440 observations per day for each road link. In this study, the Waze speed data for 17 segments on I-40 Eastbound were collected and analyzed, covering 10.81 miles. The length of the segments varied from 0.3 miles to 1.1 miles. Waze speed data from July to September (12

weeks) 2019 was the historic data used to compute the parameters in AID algorithms. The remaining two months' of data were used as the validation dataset to measure the performance of AID algorithms.

Five months of traffic data, including speed and occupancy data, were collected from TDOT (Tennessee Department of Transportation) RTMS (Remote Traffic Microwave Sensors). RTMS collects traffic information (e.g., traffic count, speed, and vehicle occupancy) for each lane every 30 seconds. In this study, 26 RTMS stations, ranging from mile marker 374.3 (west end) to 385.1 (east end), were associated with the 10.8 miles long Eastbound I-40 segment. RTMS occupancy data for these stations were extracted and averaged to be used in AID algorithms for comparison. Also, RTMS speed data were extracted and aggregated to one minute for analysis.

To compare the performance of various calibration-free AID algorithms, we needed to obtain the incident data. The incident data on selected I-40 segments from September to October 2019 were collected from TDOT's Region 1 Traffic Management Center (TMC) through a web-based archiving tool, LOCATE/IM. The incident data are well structured, containing detailed incident information, such as incident duration, location, incident type, incident start time, response time, and lane blocked. Since the calibration-free AID algorithms used in this study rely on unique traffic variables, traffic speed, or occupancy, we chose the incident/crash that caused lane blockage. Finally, a total of 20 lane-blocking crashes disrupting traffic were collected in the study.

Methodology

Proposed algorithm

In this study, we proposed a new, unique parameter, calibration-free algorithm to detect incidents with speed data; it is a simple modification of the Castro-Neto's algorithm (*9*). Instead of using occupancy data as in Castro-Neto's algorithm, we modified the algorithm so that it can be used with speed data. For occupancy data, a significant increase in occupancy would trigger an incident, while for speed data, a significant decrease in speed should be detected to declare an incident. Also, we used multiple values of time intervals rather than just a unique value.

In the proposed algorithm, we compute the mean and standard deviation of the speed difference between two adjacent road links for a specific time interval and a specific day. For example, for specific time window (t), and day of the week (d), we define $Speeddiff1min_{(i)}(t,d)$ as the ith difference of one-minute speed between two adjacent road links inside the time window (t). Assume that for a particular (t,d),

$$Speeddiff1min_{(i)}(t,d) \sim N(\mu_{1min,d}, \sigma^2_{1min,d})$$

Where N represents the normal distribution and i starts from 1 to t. The normality of speed difference will be tested when we perform the model.

Then, if we can estimate the $\mu_{1min,d}$ and $\sigma^2_{1min,d}$ from historical data, we can declare an incident if the current observation value of speed difference falls outside the one-sided region of normal distribution. For example, for specific time window (t), and day of the week (d), let $Speeddifftmin(t,d)$ as sample observations of one-minute speed differences between two adjacent links for a particular period t. We can calculate the mean $\mu_{tmin,d}$ and variance $\sigma^2_{tmin,d}$ directly from historical observations, then we can estimate the population mean and standard deviation from the sample observations of the speed difference. We can simply have $\mu_{1min,d} = \mu_{tmin,d} = \overline{\mathrm{X}}_{tmin,d}$, and $\sigma^2_{1min,d} = t\sigma^2_{tmin,d} = tS^2_{tmin,d}$. Thus, the above formula can be converted to,

$$Speeddiff1min_{(i)}(j,d) \sim N\left(\mu_{tmin,d}, t\sigma^2_{tmin,d}\right) \sim N\left(\overline{\mathrm{X}}_{tmin,d}, tS^2_{tmin,d}\right)$$

We then can define the one-sided region that contains $(1-\alpha)*100\%$ of the $Speeddiff_{(i)}1min(t,d)$, and the upper-limit value is the threshold (Thr) used to determine if an incident will be triggered. The α is the significance level, which controls the threshold of declaring an incident. If a new observation value exceeds the threshold, an incident alarm is declared. The threshold is defined as

$$Thr = N^{-1}(\bar{X}_{tmin,d}, tS^2_{tmin,d}, 1-\alpha)$$

Where N^{-1} is the inverse of the normal distribution, $\bar{X}_{tmin,d}$ is the estimated $\mu_{tmin,d}$, and S^2_{tmin} is the estimated $\sigma^2_{tmin,d}$.

In this study, for the proposed algorithm with speed data, the time window was chosen from five different values: 6,8,10,12, and 15, and the false alarm rate α was chosen from nine different values: 0.0005, 0.00075, 0.001, 0.0025, 0.005, 0.0075, 0.01, 0.015, 0.02, 0.025, totaling 50 models to be performed.

California Algorithm

California algorithm is one of the earliest developed AID algorithms that compares three variables based on vehicle occupancy with predefined thresholds. First, It computes values of three variables based on the difference of occupancy between two adjacent vehicle detection stations, namely OCCDF (spatial differences in occupancies), OCCRDF (relative spatial differences in occupancies), DOCCTD (relative temporal differences in downstream occupancies). Then the values are compared with three predefined thresholds (Thr1, Thr2, Thr3); if all three values exceed the thresholds, then an incident alarm is triggered.

In this study, each threshold was tested from 0.05 through 0.5, with increments of 0.05, which resulted in a total of 1,000 combinations of thresholds. Given some combinations of thresholds may have the same DR, we selected the model result with the minimum FAR for each same level of DR ranging from 0.6 to 1.0. If two models have the same FAR, the one with the lowest MTTD was chosen. The model performance selection process was also applied in the following AID algorithms, and then the performances of various AID algorithms were compared.

Minnesota Algorithm

Minnesota algorithm is another commonly used AID algorithm that computes the statistical variables based on vehicle occupancy and compares the variables with predefined thresholds. It computes the moving average of OCCDF (spatial differences in occupancies) between two adjacent vehicle detection stations before (y_b, 3 minutes) and after a particular time interval (y_a, 5 minutes) (*25*). Then, y_a and y_b are normalized by the pre-incident occupancy (m_t), which is the maximum value of the 5-min moving average of occupancy on both downstream and upstream vehicle detection stations. The normalized y_a and y_b are then compared with pre-defined two thresholds (Thr1 and Thr2): if y_a/m_t exceeds the first threshold (Thr1), then congestion is detected; If the second threshold is exceeded by $(y_a - y_b)/m_t$, an incident alarm is triggered.

For the Minnesota algorithm, as suggested by (*25*), the time intervals for y_a and y_b were five minutes (ten observations) and three minutes (six observations), respectively. Each threshold (Thr1, Thr2) was tested from 0.05 through 0.5, with increments of 0.05, which resulted in a total of 100 combinations of thresholds.

SND Algorithm

The standard normal deviate (SND) algorithm is based on the detection of outliers or anomalies in the continuous data stream that declares the incident. It precomputes the mean and standard deviation from the historical dataset, and SND is calculated based on the mean and standard deviation with traffic variable observations. Then the SND is compared with the predefined threshold, and if SND is larger than the predetermined threshold (Thr), an incident alarm is triggered. In classical SND method, the SND was derived from two parameters: reference value ($\hat{x}$) and variation (S). For specific segment s, time window (t), and day of the week (d), the SND can be expressed as

$$SND_s^{t,d} = \frac{|x_s^{t,d} - \hat{x}_s^{t,d}|}{S_s^{t,d}}$$

Where $x_s^{t,d}$is the traffic variable observation, $\hat{x}_s^{t,d}$ is the reference value, such as mean, and $S_s^{t,d}$ is its variation, e.g., standard deviation. From the literature, the SND algorithm can be applied both with traffic occupancy data and speed data, but a slight difference exists in this method since an incident will increase the occupancy but decrease the speed. In this study, we used the SND algorithm with Waze speed, RTMS speed, and RTMS occupancy to detect incidents.

For the SND algorithm with RTMS occupancy data, the threshold (Thr) was chosen in a range from 0.5 to 1.5 with 0.1 increments, and the time window was chosen with four different values (4, 6, 8,10, and 12 minutes), totaling 50 models to be performed. For the SND method with Waze speed and RTMS speed, the threshold was also chosen in a range from 2 to 4 with 0.2 increments, and the time window was chosen with five different values (6, 8, 10, 12, and 15 minutes), totaling 50 models to be performed.

Results

For the proposed algorithm, we assume that for a particular time window (t) and day of the week (d), the speed differences of one-minute speed data (speeddiff1min) are normally and independently distributed. To validate our consumption, for each time window (t) and each pair of adjacent road links, we performed $1440/t * 7$ chi-square

goodness-of-fit tests with a level of significance of 5%; The results showed that the null hypothesis can not be rejected in any of them, which demonstrates the applicability of our proposed algorithm.

From the results of our proposed algorithm, we found the best performance with Waze traffic speed was with time window $t = 8$ minutes and $\alpha = 1.5\%$, which gave us the highest DR of 90%, with a low FR of 0.58% and an MTTD of 2.40 minutes. Figure 2-1 shows an example of time-varying thresholds obtained from the proposed algorithm and the actual speed difference observations for the Waze link close to the incident location with time window $t = 8$ minutes and $\alpha = 1.5\%$. As shown, the proposed algorithm can accurately detect the incident that occurred on November 5, 2019 (Tuesday) near Mile Marker (MM) 379. The incident was detected around 8:08 PM and lasted for about 45 minutes. To compare, we also plotted the speed thresholds from the proposed algorithm for the incident-free case, on 10/29/1019, MM 379 (Figure 2-2), and found that the Waze speed was always below the computed thresholds, indicating our algorithm are accurate enough not to trigger false alarms.

We compared the performances of our proposed algorithm (Waze speed data and RTMS speed data) with the previously developed SND algorithm (Waze speed data, RTMS speed data and RTMS occupancy data), the California algorithm (RTMS occupancy data) and the Minnesota algorithm (RTMS occupancy data), in terms of DR, FAR, and MTTD. For each algorithm, we chose the model with the minimum FAR for each level of DR from 0.6 to 1. If two models have the same FAR, the one with the lowest MTTD was chosen. For our proposed algorithm, we found the best time windows for Waze speed and RTMS speed were eight minutes and six minutes, respectively; For the SND algorithm, the best time windows for Waze speed, RTMS speed, and RTMS occupancy were six minutes, six minutes, and eight minutes, respectively.

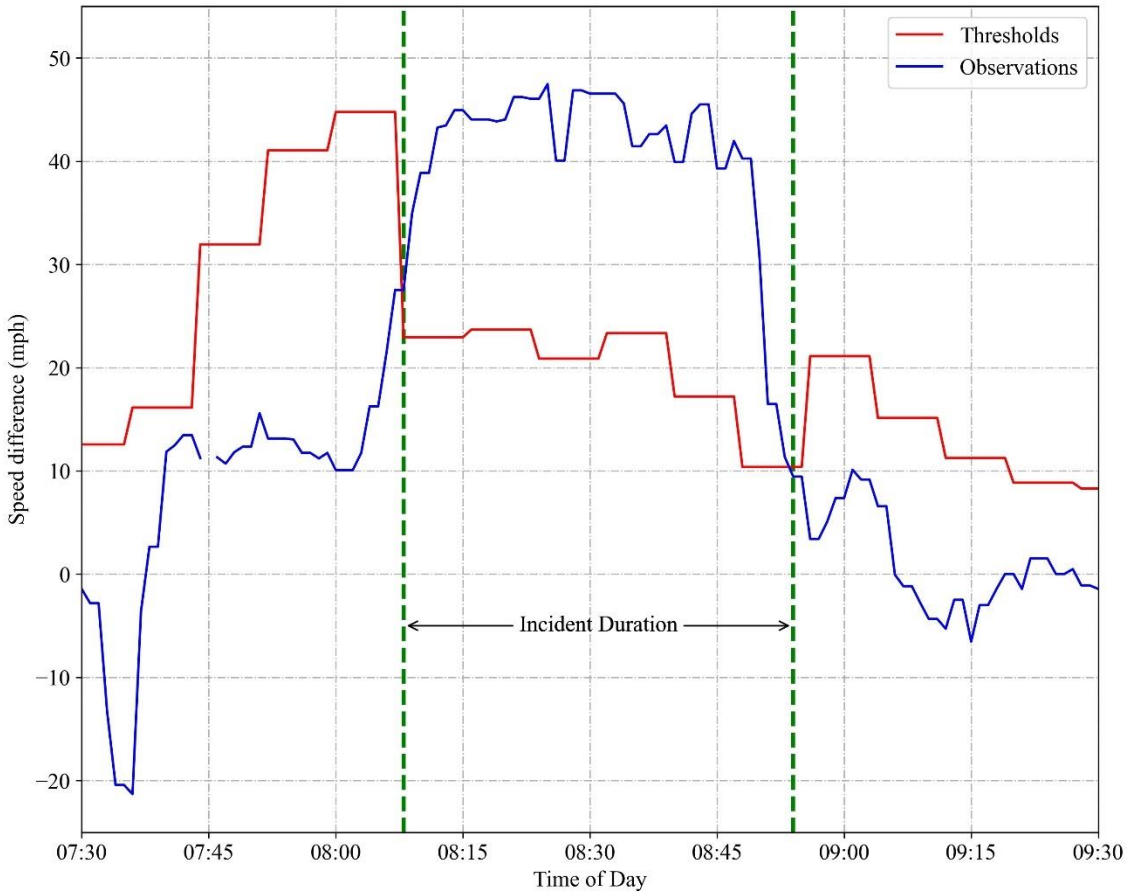


Figure 2-1 Adaptive thresholds from the proposed algorithm for the incident case on 11/5/2019, MM 3759

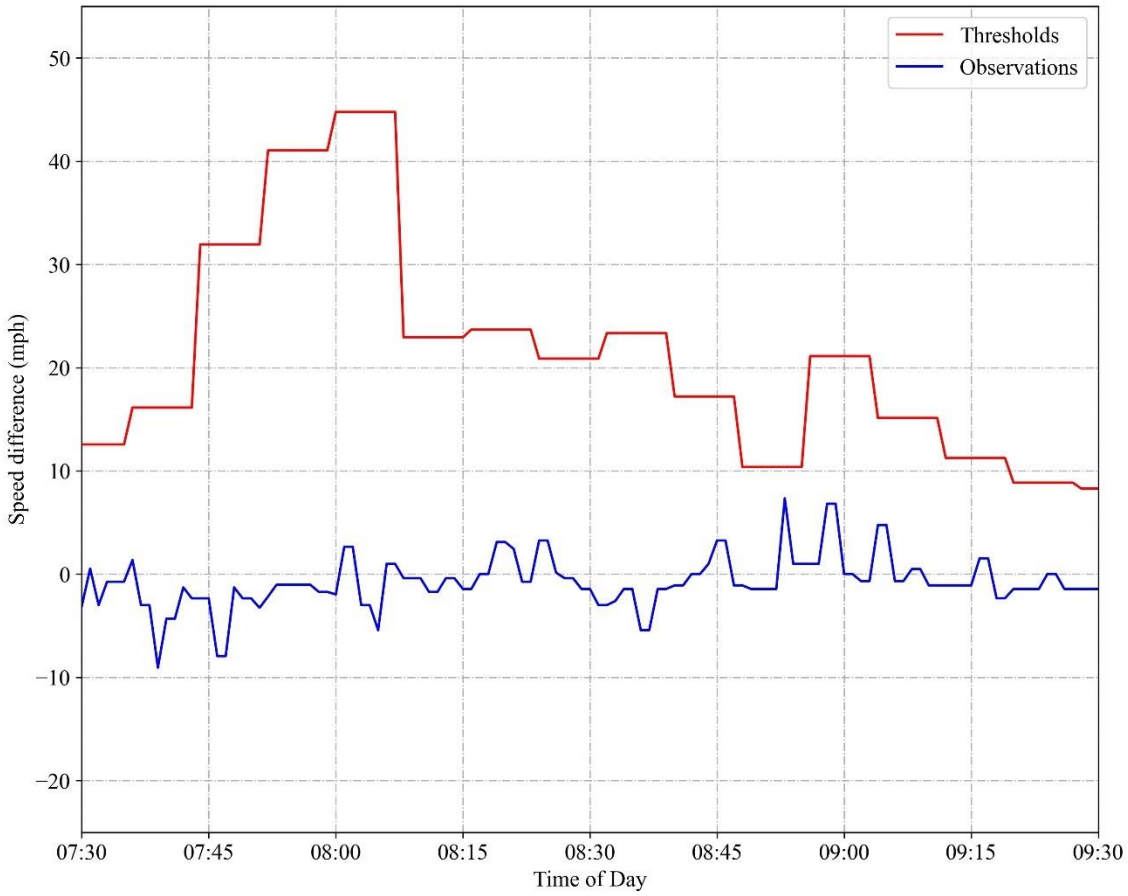


Figure 2-2 Adaptive thresholds from the proposed algorithm for the incident-free case, on 10/29/2019, MM 379

We obtained the algorithm performance with the prime time window and various thresholds. From the results, the highest DR achieved by all algorithms is 90% (Figure 2-3), which is acceptable for transportation agencies, demonstrating the reliability of Waze speed data in incident detection. The relatively low highest DR can be partly attributed to the insensitivity of algorithms to detect incidents that occurred during peak hours since we found that the uncaptured incidents occurred during peak hours.

For DR, our proposed algorithm with Waze speed presented the lowest FAR at all levels of DR, followed by the proposed algorithm with RTMS speed. For FAR, the highest FAR obtained is around 1.4% by the SND algorithm with RTMS occupancy, with

the high DR of 90%, but our proposed algorithm can achieve a FAR smaller than 0.6%. For MTTD, it seems that the algorithms with occupancy data presented lower MTTD, which can be partly attributed to the higher data resolution compared with Waze speed data.

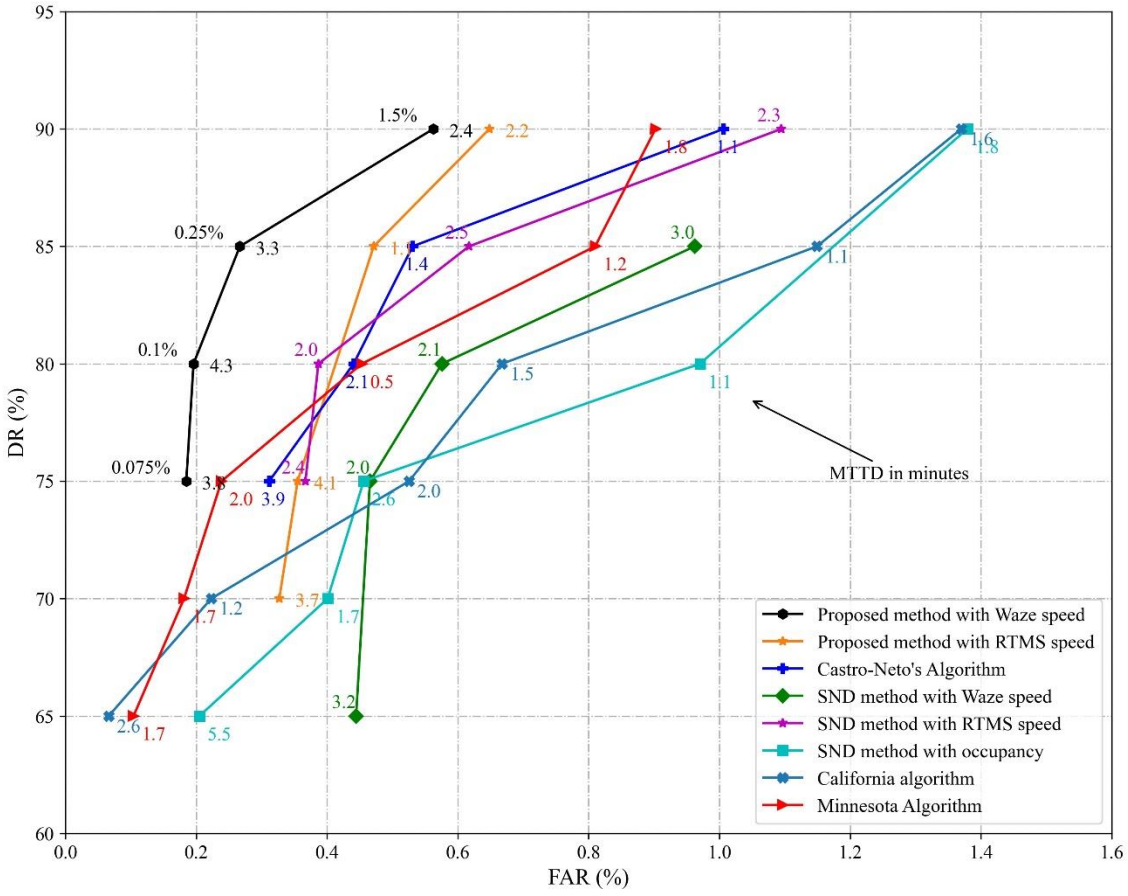


Figure 2-3 Comparison of the performances of AID algorithms

Conclusion

In this study, we evaluated the reliability of using Waze speed data to detect incidents on freeways. We proposed a new calibration-free algorithm based on the speed difference between two adjacent road links to detect incidents on freeways, and we compared them with several common, other calibration-free AID algorithms. We conducted a case study on I-40 Eastbound freeway in Knoxville, Tennessee in which we collected Waze speed data, RTMS speed data, and RTMS occupancy data for the 10.8-mile long segment on I-40 Eastbound.

From the results, we found the Waze speed data is accurate enough to be used in incident detection with a high DR of 90%. Our proposed algorithm achieved better performance in terms of DR and FAR compared with other methods. Our proposed algorithm with Waze speed presented the lowest FAR at all levels of DR, followed by the proposed algorithm with RTMS speed. However, the MTTD for our proposed algorithm seems to be slightly higher than that of algorithms with RTMS occupancy, which may be related to the data resolution. Overall, the results showed the applicability of our proposed algorithm in incident detection with speed data and the reliability of Waze speed data in incident detection, which can be helpful for incident management systems operated by transportation agencies.

Note that several limitations should be addressed for future studies. First, the proposed algorithm is based only on traffic speed; other traffic flow fundamentals can be incorporated to augment performance. Second, the proposed algorithm can also be tested with other speed datasets, and the combination of multiple datasets can be investigated in the future to improve performance. Last, the proposed algorithm considered only the difference in speed to detect an incident. In future work, the spatial-temporal relationship among Waze speed for difference links should be considered to get better performance.

Reference

[1] Seo, T., and T. Kusakabe. Probe vehicle-based traffic state estimation method with spacing information and conservation law. *Transportation Research Part C: Emerging Technologies,* Vol. 59, 2015, pp. 391-403.

[2] An, S., H. Yang, J. Wang, N. Cui, and J. Cui. Mining urban recurrent congestion evolution patterns from GPS-equipped vehicle mobility data. *Information Sciences,* Vol. 373, 2016, pp. 515-526.

[3] Ahsani, V., 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, 2019, pp. 103-119.

[4] Bae, B. Real-time Traffic Flow Detection and Prediction Algorithm: Data-Driven Analyses on Spatio-Temporal Traffic Dynamics. 2017.

[5] Park, H., and A. Haghani. Real-time prediction of secondary incident occurrences using vehicle probe data. *Transportation Research Part C-Emerging Technologies,* Vol. 70, 2016, pp. 69-85.

[6] Chakraborty, P., C. Hegde, and A. Sharma. Data-driven parallelizable traffic incident detection using spatio-temporally denoised robust thresholds. *Transportation Research Part C-Emerging Technologies,* Vol. 105, 2019, pp. 81-99.

[7] Williams, B. M., and A. Guin. Traffic management center use of incident detection algorithms: Findings of a nationwide survey. *IEEE Transactions on intelligent transportation systems,* Vol. 8, No. 2, 2007, pp. 351-358.

[8] Abdulhai, B., and S. G. Ritchie. Enhancing the universality and transferability of freeway incident detection using a Bayesian-based neural network. *Transportation Research Part C-Emerging Technologies,* Vol. 7, No. 5, 1999, pp. 262-280.

[9] Castro-Neto, M. M., L. D. Han, Y. S. Jeong, and M. K. Jeong. Toward Training-Free Automatic Detection of Freeway Incidents Simple Algorithm with One Parameter. *Transportation Research Record,* Vol. 2278, No. 2278, 2012, pp. 42-49.

[10] Stephanedes, Y. J., and A. P. Chassiakos. Application of Filtering Techniques for Incident Detection. *Journal of Transportation Engineering-Asce,* Vol. 119, No. 1, 1993, pp. 13-26.

[11] Balke, K., C. L. Dudek, and C. E. Mountain. Using probe-measured travel times to detect major freeway incidents in Houston, Texas. *Transportation Research Record,* Vol. 1554, No. 1, 1996, pp. 213-220.
[12] Dudek, C. L., C. J. Messer, and N. B. Nuckles. Incident detection on urban freeways. *Transportation research record,* Vol. 495, 1974, pp. 12-24.
[13] Zhu, T., J. Wang, and W. Lv. Outlier mining based automatic incident detection on urban arterial road.In *Proceedings of the 6th International Conference on Mobile Technology, Application & Systems*, ACM, 2009. p. 29.
[14] Castro-Neto, M. M., L. D. Han, Y.-S. Jeong, and M. K. Jeong. Toward Training-Free Automatic Detection of Freeway Incidents: Simple Algorithm with One Parameter. *Transportation research record,* Vol. 2278, No. 1, 2012, pp. 42-49.
[15] Jeong, Y. S., M. Castro-Neto, M. K. Jeong, and L. D. Han. A wavelet-based freeway incident detection algorithm with adapting threshold parameters. *Transportation Research Part C-Emerging Technologies,* Vol. 19, No. 1, 2011, pp. 1-19.
[16] Dia, H., and G. Rose. Development and evaluation of neural network freeway incident detection models using field data. *Transportation Research Part C-Emerging Technologies,* Vol. 5, No. 5, 1997, pp. 313-331.
[17] Cheu, R. L., D. Srinivasan, and W. H. Loo. Training neural networks to detect freeway incidents by using particle swarm optimization. *Freeway Operations and Traffic Signal Systems 2004,* Vol. 1867, No. 1867, 2004, pp. 11-18.
[18] Hawas, Y. E. A fuzzy-based system for incident detection in urban street networks. *Transportation Research Part C-Emerging Technologies,* Vol. 15, No. 2, 2007, pp. 69-95.
[19] Teng, H. L., and Y. Qi. Application of wavelet technique to freeway incident detection. *Transportation Research Part C-Emerging Technologies,* Vol. 11, No. 3-4, 2003, pp. 289-308.
[20] Mak, C. L., and H. S. L. Fan. Algorithm fusion for detecting incidents on Singapore's Central Expressway. *Journal of Transportation Engineering,* Vol. 132, No. 4, 2006, pp. 321-330.
[21] Payne, H. J., and S. C. Tignor. Freeway incident-detection algorithms based on decision trees with states. *Transportation Research Record,* No. 682, 1978.

[22] Zhang, K., and M. A. P. Taylor. Towards universal freeway incident detection algorithms. *Transportation Research Part C-Emerging Technologies,* Vol. 14, No. 2, 2006, pp. 68-80.
[23] Chakraborty, P., 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.
[24] Chung, Y., and W. W. Recker. A Methodological Approach for Estimating Temporal and Spatial Extent of Delays Caused by Freeway Accidents. *Ieee Transactions on Intelligent Transportation Systems,* Vol. 13, No. 3, 2012, pp. 1454-1461.
[25] Stephanedes, Y. J., and A. P. Chassiakos. Application of filtering techniques for incident detection. *Journal of Transportation Engineering,* Vol. 119, No. 1, 1993, pp. 13-26.

CHAPTER 3

SPATIAL-TEMPORAL QUALITY ANALYSIS OF CROWDSOURCED WAZE INCIDENT REPORTS

Abstract

Crowdsourced transportation data, such as Waze user reports, have been generated with more and more people using mobile phones; these data could help traffic managers make better-informed decisions. To understand these traffic data sources, we conducted a spatial-temporal quality analysis of crowdsourced Waze accident reports by comparing the Waze accident reports with the TDOT crash records from Nashville, Tennessee in 2018, and explored the reliability of Waze accident reports not found in crash records. From the results, we found that 32.8% of TDOT crash records can be found in Waze accident reports when allowing for a reasonable time and distance variation. For matched crashes, the mean distance difference is 0.08 miles and the mean time difference is -4.0 minutes, suggesting the relatively high accuracy of Waze accident reports. Several factors affecting the likelihood of matching were identified, including the time of day, day of the week, weather, and the number of injuries. Moreover, about 56% of unmatched Waze accident reports were found to have a significantly higher travel time with the presence of accidents at a significant level of 5%, demonstrating the contributions and potential of Waze accident reports as an alternative data source in incident management.

Introduction

Road networks are indispensable components of transportation infrastructures that are crucial to the transport and movement of people, goods, and services. Traffic incidents have significant negative effects on the smooth operation of road networks, challenging roadway system efficiency and public safety. For example, every minute a freeway lane is blocked as a result of an incident can result in four minutes of travel time delay (*1*), over 1.25 million people die each year by road traffic crashes (*2*), and approximately 37 thousand people die on U.S. roads as a result of road traffic accidents (*3*). Therefore, early accident detection can help transportation agencies make quick and timely responses to reduce and mitigate the effects of an incident.

Transportation agencies have various systems to identify and manage incidents, and various datasets are used in incident management systems, mainly including radar

sensor data, loop detector data, probe vehicle data, and video data. However, these datasets may have limitations, such as high installation and maintenance fees, limited coverage, and malfunction issues. Nowadays, crowdsourced transportation data has become an essential alternative data source in roadway incident management with a massive input and output data flow, and this emerging data source has been investigated by many researchers. For example, Gu, Qian and Chen (*4)* developed a methodology to extract traffic incident information from Twitter, and they applied the methods in two regions, the Pittsburgh and Philadelphia Metropolitan Areas, in September 2014, demonstrating that social media data could well be a cost-effective alternative incident data source. Crowdsourced transportation data has been explored in depth within various topics in the transportation field, such as human mobility tracing (*5; 6*), sentiment analysis (*7-10*), and incident detection (*11-13*). Many of these researchers were using data extracted from social media platforms such as Twitter, which often contain typos or grammatical errors, making it difficult to separate accurate information from noise. Fortunately, crowdsourced Waze reports (e.g., accident reports, stopped vehicle reports, and jam reports) have become available and provide a large amount of cost-effective, real-time, traffic-related information. This new source of data has the potential of being an alternative data source that can be used in incident management systems, but it needs to be explored and evaluated.

The objective of this study is to better understand Waze accident reports by comparing Waze accident reports with crash records collected by agency officials to explore its potential in incident management. This study first measured the spatial-temporal quality of crowdsourced Waze accident reports, both on and off interstate highways. Then, we investigated the factors affecting the matching likelihood between the two datasets. Last, we explored the reliability of Waze accident reports not found in crash records. In the remainder of this paper, Section 2 reviews extant related studies and Section 3 introduces the data and methods used in this study. The main results are presented in Section 4, and Section 5 presents the conclusion and future work.

Related work

Waze is a crowdsourcing traffic application used for both navigation and for users to share the real-time traffic information, such reports of accidents, stopped vehicles, traffic jams, road construction, and police reports. Waze accident reports can be helpful in incident management since they may detect or identify accident faster than the existing methods. There are several sequential processes in incident management, including incident detection, incident verification, incident response, incident clearance (*14*). The reduction of incident detection and verification time can lead to a quick response from transportation agencies. Using conventional methods, an incident may not be instantly detected and reported to the transportation agency. Thus, early incident detection using a variety of datasets allows for timely response to reduce and mitigate the effects of an incident. Moreover, as road networks become more complex and incidents may occur at any time and location, transportation agencies need more efficient and effective ways to detect incidents.

Much research has been devoted to exploring the potential of using crowdsourced traffic data in incident management because it provides large amounts of cost-effective and real-time traffic-related information (*15-17*). Mai and Hranac (*17),* for example, investigated the relationships between the occurrence of traffic incidents and the related social media message numbers and found that they are highly associated, demonstrating the power of social media to analyze traffic-related information. Waze data have been explored in areas such as user behavior (*18*), traffic conditions (*19*), and incident detection (*20; 21*). For example, to explore the potential of integrating Waze incident data into the official incident data, dos Santos, Davis Jr and Smarzaro (*22)* matched the two traffic accident datasets from Waze and Belo Horizonte Transport and Transit Company (BHTRANS). Amin-Naseri et al. (*23)* investigated the validity and coverage of crowdsourced Waze incident reports and found that Waze helps monitor traffic on the road with broad coverage, timely reporting, and reasonable geographic accuracy.

Despite the invaluable information that previous works provide, they often do not address the spatial-temporal quality of crowdsourced Waze accident reports on interstates and other roadways and highways. Also, the Waze accident reports that do not have the

corresponding official crash records received little exploration but could be a major contribution of crowdsourced Waze data. In response, this study explored Waze accident reports from these less considered perspectives to gain an in-depth understanding of Waze data. We compared the Waze accident reports with the crash records for Nashville, Tennessee in 2018 to see if the crash records can be found in Waze accident reports, allowing for small variations of incident time and location. In addition, we investigated the factors affecting the matching likelihood of these two datasets and explored the reliability of unmatched Waze accident reports.

Data and methods

Data

Waze accident reports were collected from a localized XML feed (Waze API). This feed is not publicly accessible but is available for Waze Connected Citizens Program partners (*24*). At a one-minute interval, the XML file containing real-time accident reports information is downloaded. Given that XML file collection is re-executed frequently, the series of XML files were processed to eliminate duplicate incident reports. After removing the duplicate incident data, we obtained 29,802 Waze accident reports for Nashville in 2018; because of missing data, 326 days were logged. These accident reports covered both highways and local streets, which is valuable to rural areas since incidents in a rural area cannot be detected quickly by agency officials. The obtained reports contain rich information, such as the accident location coordinates, the accident start time, and the accident type.

The official crash data were obtained from TDOT's Tennessee Enhanced Tennessee Roadway Information Management System (E-TRIMS, https://e-trims.tdot.tn.gov). We collected the City of Nashville crash records for the corresponding days of the year 2018. The crash data are well structured, containing detailed crash information, such as crash location, date and time of the crash, road type, type of crash, total injuries, total vehicles, weather, and light conditions. TDOT crash data also covers the crashes off the roadway, which may not have corresponding Waze accident reports; thus, the crashes not on the roadway were removed. Finally, we obtained 13,547 crashes

in Nashville for 2018. Since the same road crash might be reported multiple times, the number of Waze accident reports is much higher than the number of E-TRIMS crash records.

Figure 3-1 presents the spatial distributions of both TDOT crash records and Waze accident reports. As shown, both TDOT crashes and Waze accident reports are concentrated along with the major road segments, which makes sense since those segments are always carrying heavy traffic. But a significant number of TDOT crashes are also concentrated in the city center, and these receive fewer Waze accident reports.

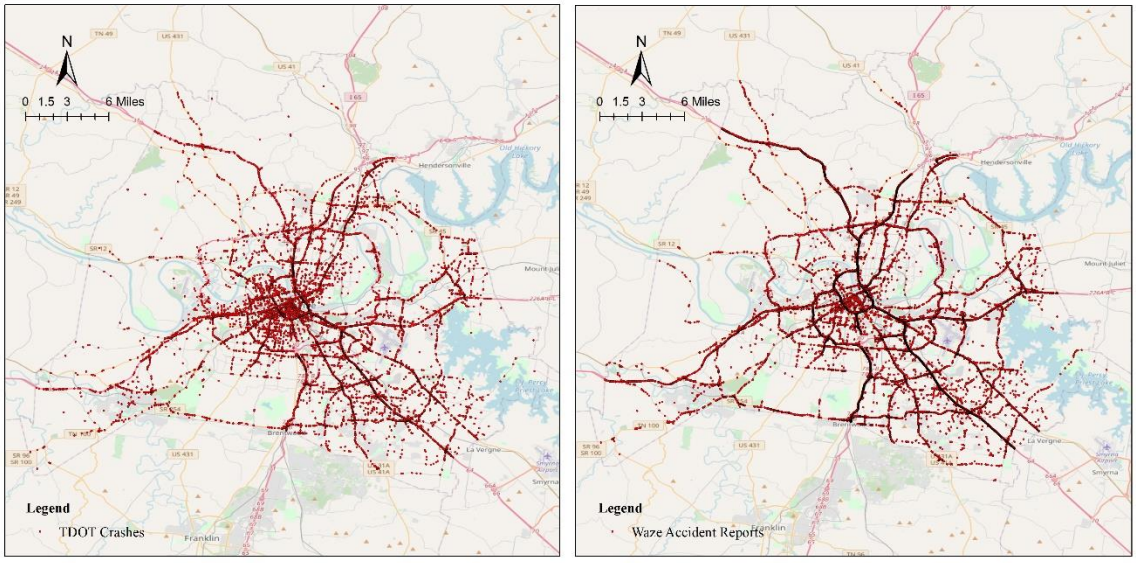


Figure 3-1 Spatial distribution of TDOT Crashes (left) and Waze accident reports (right), in Nashville, TN, 2018

Methods

Based on the assumption that if a traffic crash occurs, there should be corresponding Waze accident reports, we compared the two crash datasets to determine the relationships between them, if any. Note that the time and location of an accident from Waze data and TDOT crash data may not be precisely identical; thus, when matching the two datasets, we referred to the accident records from two datasets as the same accident if they were reported within a certain time interval of each other and occurred within a certain distance from each other.

Previous studies have used different time and distance thresholds (*21; 22*), for example, 20 minutes and 2.5 miles, or 60 minutes and 150 meters, which are determined subjectively. In this study, we attempted to obtain suitable thresholds semi-subjectively. To get the suitable distance and time threshold for matching, we obtained the number of matches by allowing distance varying from 0 to 1 mile and time varying from 0 to 30 minutes (Figure 3-2). From the figure, we can observe that for each level of distance difference, the number of matches would not increase significantly after 20 minutes. For distance difference, it seems that the number of matched crashes increases as the distance difference increases, but it would not increase significantly after 0.3 miles. Therefore, we may assume 0.3 miles and 20 minutes as the suitable thresholds for matching the two datasets, and 4,452 TDOT crash records were found in Waze accident reports with the selected thresholds.

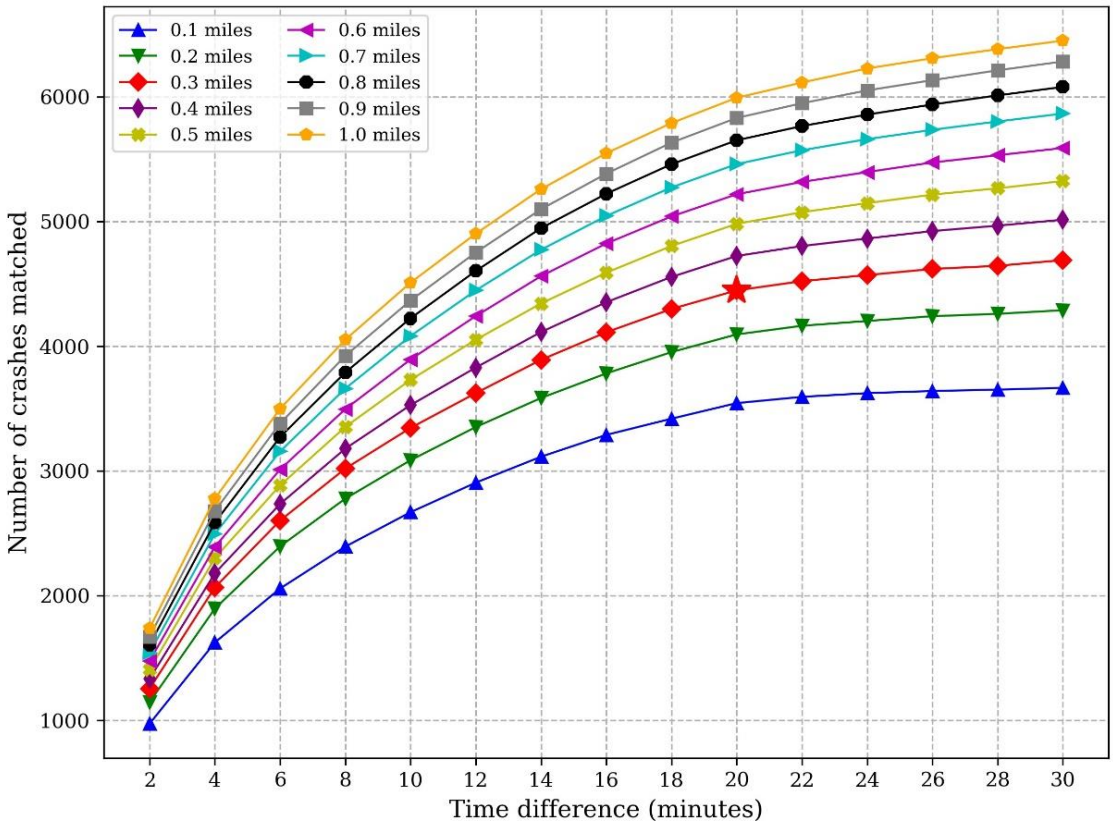


Figure 3-2 Number of crash matches with varying time and distance difference

Results

Allowing for a small variation of incident time and location (20 minutes and 0.3 miles), we generated a set of all possible matched crashes. Out of 13,574 TDOT crash records, 4,452 (32.8%) were found in Waze accident reports, of which 1,776 out of 3,293 (53.9%) TDOT crash records were on interstate highways and 2,676 out of 10,281 (26.0%) TDOT crash records were on other roadways and highways. This means that the crashes that occurred on highways are more likely to be reported by Waze users, which may due to the large volumes of traffic on interstate highways. For Waze accidents, 7,019 out of 29,802 (23.6%) Waze accident reports were matched to TDOT crash records, and more than one Waze accident report can be matched to the same crash. Figure 3-3 shows the spatial distribution of TDOT crash records and matched crashes between these two datasets. The matched accidents are more likely to be concentrated along the major arterials in Nashville, which is expected since Waze accident reports are concentrated along the major arterials because of the relatively high traffic on major arterials.

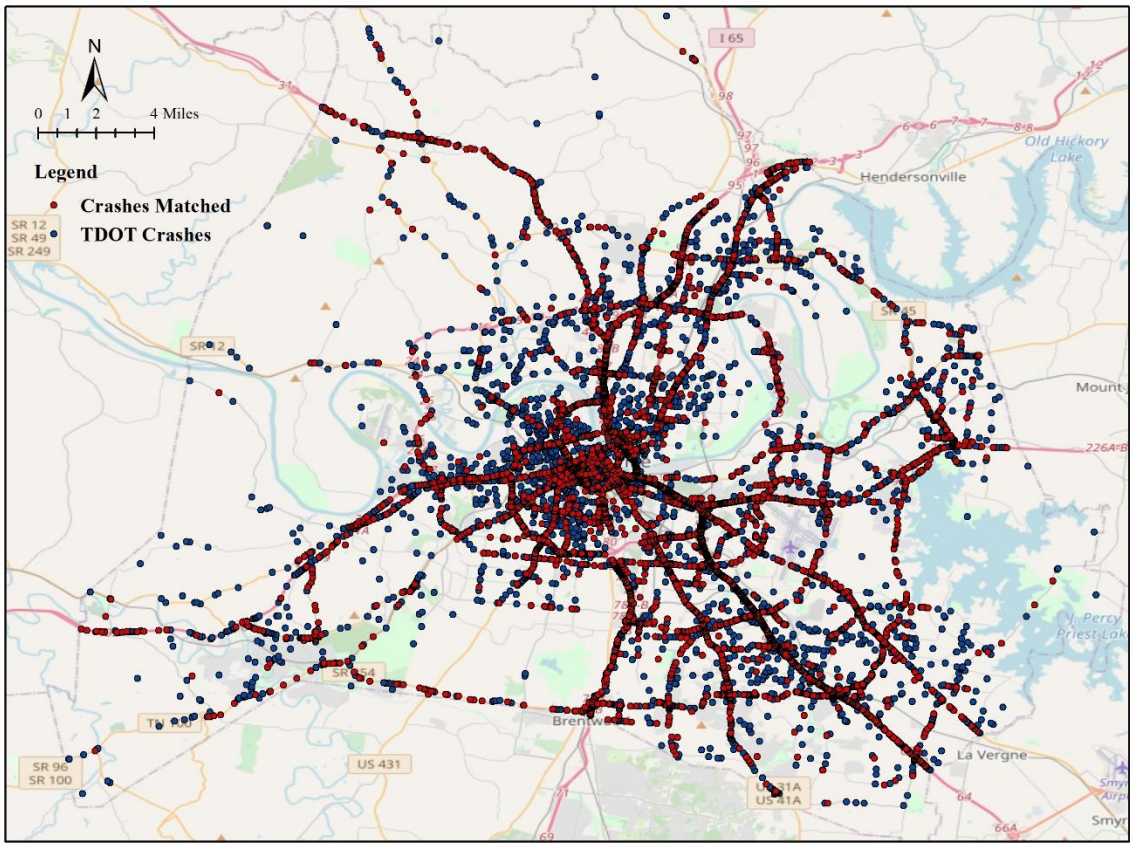


Figure 3-3 The spatial distribution of matched crashes and the total crashes

Table 3-1 The descriptive analysis of time and distance difference for matched TDOT crashes

Matched TDOT Crashes	Distance Difference (miles)					Time Difference (minutes)			
	Number	Mean	Std	50%	75%	Mean	Std	50%	75%
All crashes	4452	0.08	0.08	0.05	0.13	-3.97	8.93	-3.93	1.53
Crashes on interstate highways	1776	0.09	0.08	0.06	0.14	-2.91	8.88	-2.23	2.45
Crashes not on interstate highways	2676	0.08	0.08	0.04	0.14	-4.67	8.90	-4.92	0.63

Matched crashes between TDOT crash records and Waze accident reports

To investigate the spatial and temporal quality of Waze accident reports, we conducted an in-depth exploration of the spatial and temporal patterns of matched crashes. Figure 3-4 presents the joint hexagonal histogram of the time difference and distance difference between Waze accident reports and TDOT crash records for the matched crashes. Since each matched crash may have more than one Waze accident report, the Waze accident report with the smallest time difference was selected to compute the time difference and distance difference. From the figure, we can observe that the majority matched accidents have a small distance difference and a negative time difference, showing the high accuracy of Waze accident reports.

Spatially, the distance difference has a highly skewed distribution, and most of the matched crashes have a distance difference of fewer than 0.1 miles. The mean distance difference is 0.08 miles, showing the relatively high spatial accuracy of Waze accident reports. Also, the mean distance difference for matched crashes on highways is slightly higher than that of matched crashes not on highways (Table 3-1), which makes sense since the vehicles on interstate highways travel at high speeds. Temporally, for the majority of the matched crashes (about 67%), the time differences are negative, which means that Waze accident reports detect the accident earlier than the TDOT crash records. The mean time difference is -4.0 minutes, indicating that Waze reports seem to be more accurate than TDOT crash records in terms of accident reporting time. Additionally, the time difference for matched crashes on interstate highways (-2.91 minutes) is slightly greater than that of matched crashes not occurring on interstate highways (-4.67 minutes). It can be attributed to the quicker response by transportation agencies for crashes on interstate highways, thus making the time difference smaller for crashes on interstate highways.

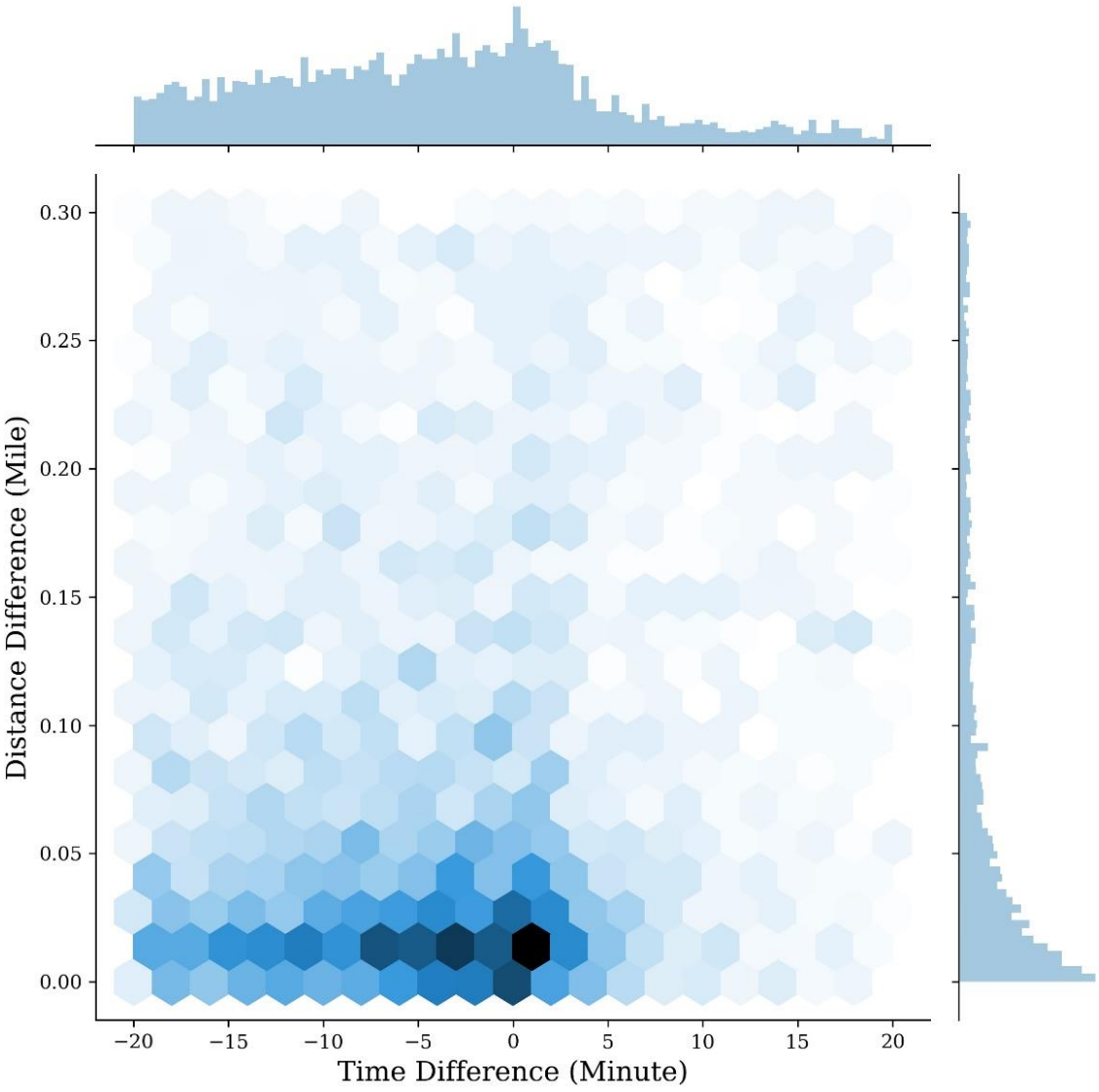


Figure 3-4 The joint hexagonal histogram of the time difference and distance difference between Waze incident reports and TDOT crash records

To investigate the factors affecting the matching likelihood between TDOT crash records and Waze accident reports, we performed a logistic regression analysis. The following variables were selected: time of day (peak-hours and non-peak hours), day of the week (weekday and weekend), location (interstate highway and other roadways), light conditions, weather, and the number of injuries. Table 3-2 presents the estimated results for the logistic regression. From the results, the number of injuries positively affects the likelihood of matching, which makes sense since the number of injuries is indicative of the severity of the injury, and the more severe the crash is, the more likely that users will report it. The interstate highway yields a higher likelihood of matching, owing to the huge volumes of traffic on interstate highways. Poor light conditions and bad weather yield a lower likelihood of matching. This may be because drivers need to be more focused in bad weather conditions, making it difficult for them to report accidents on Waze. Crashes with injuries or fatalities are more likely to match, compared to crashes with only property damage. The time of day and day of the week also significantly affect the likelihood of matching, as do peak hours and weekdays, compared to non-peak hours and weekends, which may relate to a large number of Waze users on the road during peak hours on weekdays.

Table 3-2 Significant factors in the likelihood of TDOT crash record being matched by Waze accident reports

Variable		Estimate	P-value	Interpretation
Number of injuries		0.1702	0.0000	Higher number of injuries yields higher matching likelihood
Road type (Base: Not interstate highway)	Interstate highway	1.0861	0.0000	On the interstate highway, the likelihood of matching is higher compared to other types of road
Light conditions (Base: Daylight)	Dark	-0.8253	0.0000	Compared to daylight, bad light conditions negatively affect the incident matching likelihood
	Dawn	-0.8319	0.0001	
	Dusk	-1.2743	0.0000	
	Other	-0.5119	0.2937	
Time of day (Base: Non-peak hour)	Peak hour	0.6461	0.0000	Peak hour yields higher matching likelihood
Weather (Base: Clear)	Snow	-0.3913	0.5551	Compared to clear weather, bad weather conditions negatively affect the incident matching likelihood
	Fog	-2.1579	0.0364	
	Cloudy	-0.2372	0.0001	
	Rain	-0.1033	0.1104	
	Other	-1.3203	0.0069	
Crash type (Base: Property damage)	Injury	0.3895	0.0000	Severe crash yields higher matching likelihood
	Fatal	0.7489	0.0292	
Day of week (Base: weekend)	weekday	0.3519	0.0000	Weekday yields higher matching likelihood
Log-Likelihood: -6923.0; p-value: 0.0000; sample size: 13,574				

Unmatched Waze accident reports

Exploring unmatched Waze accident reports can help in understanding how to use them as an alternative data source in incident management. We found 21,613 Waze accident reports that could not be matched to TDOT crash records. Considering that multiple Waze accident reports can refer to the same accident, we kept only one report for each accident and obtained 16,057 unique Waze accident reports by allowing a small variation of time and distance (0.3 miles and 20 minutes).

Unmatched Waze accident reports were validated by investigating the travel time near the accident location using the National Performance Management Research Data Set (NPMRDS). The assumption here is that, when a traffic accident occurs, the actual travel time will vary significantly from the typical travel time. The NPMRDS data were downloaded from INRIX (https://npmrds.ritis.org/). By comparing the actual travel time near the location of an accident with the typical travel time at the same place, the same time-of-day and the same day-of-week, we can determine if the travel time shows a significant change with the presence of accidents, thus inferring whether the Waze accident reports are reliable. The typical travel time near the location of an accident was computed by averaging the travel times at the same place, the same time-of-day, and the same day-of-week over eight weeks (four weeks before and after the accident). We obtained 10,079 unmatched Waze accident reports within 10 meters of NPMRDS road segments, in which 1,817 were Waze major accident reports, 5505 were Waze minor accident reports, and 2,757 were Waze accident reports without accident type information.

Figure 3-5 depicts the actual travel time and typical travel time for each Waze accident report where the actual travel time of each accident is sorted in ascending order. As is shown, the majority of travel times with the presence of an accident are substantially higher than the travel times without the presence of an accident. To statistically test the hypothesis, we conducted the Mann-Whitney U-test, which compares the means of two groups that do not follow a normal distribution to test if the mean of travel times is significantly different with or without the presence of an accident (Table 3-3). The mean actual travel time for all accidents was 3.4 minutes, while the mean typical

travel time was 1.6 minutes. The Mann-Whitney U-test result (p-value = 0.0000) suggests a significant difference between these two travel times.

Moreover, for each Waze accident report, we performed a hypothesis test with the null hypothesis that the actual travel time is not significantly different from the typical travel time. Assume the typical travel time population follows a normal distribution, and the eight travel times extracted are samples from the population. For each Waze accident, consider t_i the typical travel time inside the eight weeks' travel times T, the typical travel time t should have the following distribution.

$$t \sim N(\bar{T}, 8S^2)$$

Where N represent the normal distribution, $\bar{T}$ is the mean of the sampled travel times, and S^2 is the variance of sampled travel times. Then, for the desired level of significance α, a one-sided region that contains $(1-\alpha) * 100\%$ of the typical travel times can be defined, and the upper-limit value is the threshold (Thr) used to determine if actual travel time t' is significantly different from typical travel time. The threshold can be computed as

$$\text{Thr} \sim N^{-1}(\bar{T}, 8S^2, 1-\alpha)$$

If actual travel time t' exceeds the threshold, we consider the actual travel time is significantly higher than the typical travel time, suggesting that the accident report is reliable. Figure 3-6 shows the percentage of reliable Waze accident reports with varying levels of significance α. From the figure, we observe that at least 56% of Waze accident reports have a significantly higher travel time with the presence of accidents. Furthermore, for Waze major accident reports, the percentage can be up to 72%. The results suggest that many accidents are reported by Waze users, especially Waze major accident reports, yet the accidents receive no response from transportation agencies; this demonstrates the contributions and potential of Waze accident reports in traffic incident management.

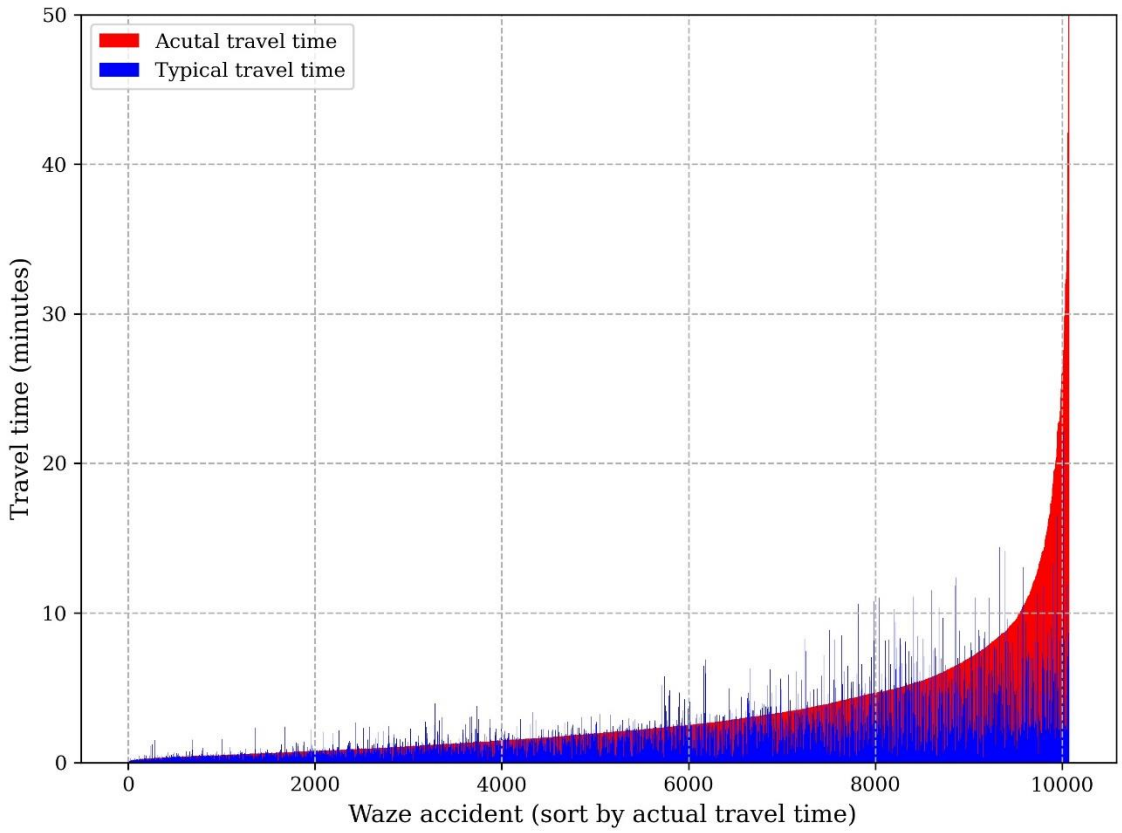


Figure 3-5 Typical travel time and Actual travel time for Waze accidents (sorted by actual travel time)

Table 3-3 Results of Mann-Whitney U-test

	Actual travel time	Typical travel time
Mean	3.339	1.616
Standard deviation	4.659	1.613
Observations	10,079	10,079
Pearson Correlation	0.540	
U Statistic	22480415.0	
P-value	0.0000	
Alternative hypothesis	True location shift is not equal to 0	

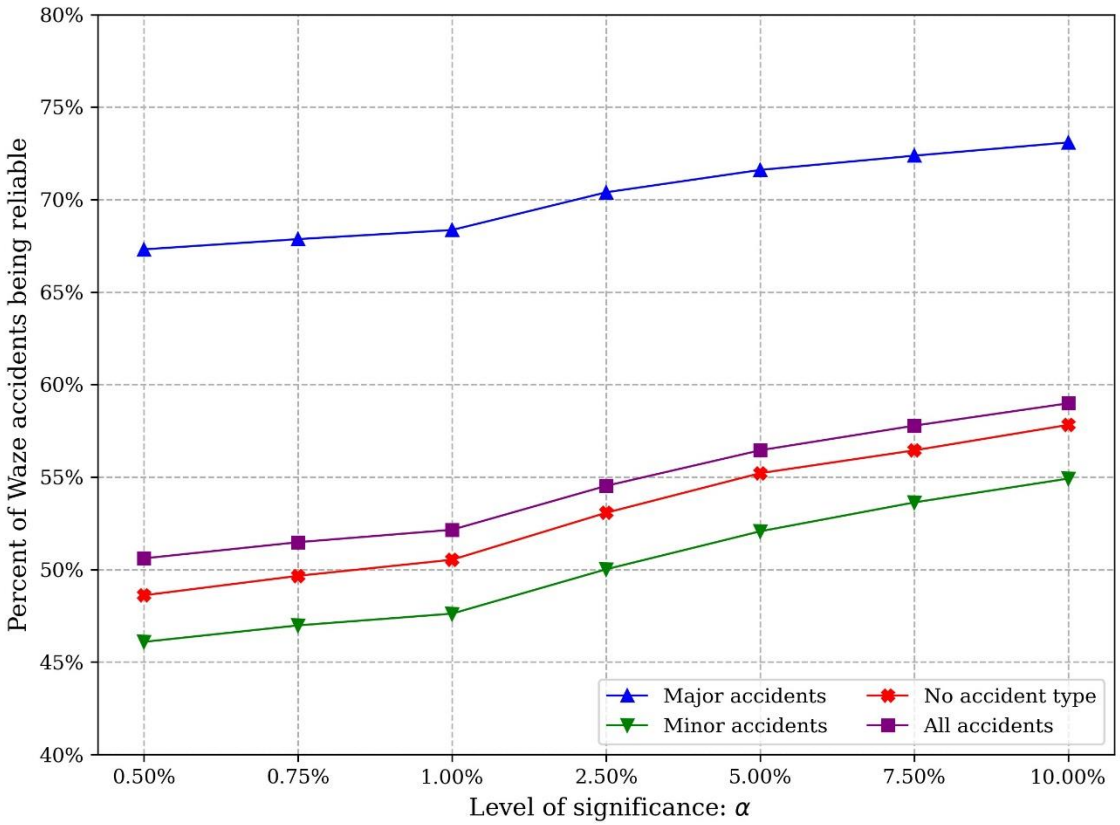


Figure 3-6 The percentage of reliable Waze accident reports with varying level of significance α

Conclusion

In this study, we conducted a spatial-temporal quality analysis of Waze accident reports to better understanding this source of data and any resultant analysis. First, we compared the Waze accident reports and TDOT crash records spatially and temporally. We collected and matched these two datasets (13,574 TDOT crash records and 29,802 Waze accident reports in Nashville, Tennessee in 2018) by allowing a reasonable variation in time and distance. Then, we investigated the factors affecting the likelihood of TDOT crash records also being reported in Waze. In addition, we measured the reliability of unmatched Waze accident reports by comparing the travel time with and without the presence of accidents obtained from NPMRDS travel time data.

From the results, when allowing a small variation of 0.3 miles and 20 minutes to suggest the same accident, 32.8% of TDOT crash records can be matched in Waze accident reports, in which 53.9% of TDOT crash records on interstate highways and 26.0% of TDOT crash records on other roadways were matched. A large number of matched crashes have a small distance difference with a negative time difference. The distance difference has a highly skewed distribution, and most of the matched crashes have a distance difference smaller than 0.1 miles. The mean distance difference is 0.08 miles, demonstrating the relatively high spatial accuracy of Waze accident reports. The mean time difference for the matched crashes is -4.0 minutes, indicating that Waze reports seem more accurate than TDOT crash records in terms of accident reporting time. Several factors affecting the likelihood of TDOT crash records being matched in Waze accident reports were identified, including the number of injuries, the time of day, day of the week, weather, and location. Moreover, the unmatched Waze major accident reports were verified using NPMRDS travel time data, and a significant increase in travel time was found with the presence of accidents. We observed that at least 56% of Waze accident reports have a significantly higher travel time with the presence of accidents at a significant level of 5%. This shows the contributions and potential of Waze accident reports as an alternative data source in incident management.

Several limitations should be noted for future major efforts. First, the matching methodology can be further improved, as well as the determination of the matching

threshold; for example, the matching algorithm can be improved by considering the road directions. However, we believe that the thresholds may not meaningfully impact our findings. Second, incident data covering a broader region and a larger time range should be analyzed in the future to gain a more in-depth understanding of the relationships between Waze accident reports and official crash data. Last, but not least, we compared only the accident reports with TDOT crash data, and multiple other incident data sources can be used together to measure the accuracy and reliability of Waze accident reports. Besides, the integration of multiple incident datasets would increase the accuracy of incident detection, thus assisting transportation agencies and road users to make timely responses that could reduce and mitigate the effects of an incident.

References

[1] Haas, K. Benefits of traffic incident management. *National traffic incident management coalition*, 2006.

[2] WHO. *Road traffic injuries*. http://www.who.int/mediacentre/factsheets/fs358/en/. Accessed December 6, 2017.

[3] NHTSA. *USDOT Releases 2016 Fatal Traffic Crash Data*. https://www.nhtsa.gov/press-releases/usdot-releases-2016-fatal-traffic-crash-data. Accessed December 6, 2017.

[4] Gu, Y., Z. 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, 2016, pp. 321-342.

[5] Maghrebi, M., A. Abbasi, T. H. Rashidi, and S. T. Waller. Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time.In *2015 Ieee 18th International Conference on Intelligent Transportation Systems*, Ieee, New York, 2015. pp. 208-213.

[6] Huang, W., and S. N. Li. Understanding human activity patterns based on space-time-semantics. *Isprs Journal of Photogrammetry and Remote Sensing,* Vol. 121, 2016, pp. 1-10.

[7] Pournarakis, D. E., D. N. Sotiropoulos, and G. M. Giaglis. A computational model for mining consumer perceptions in social media. *Decision Support Systems,* Vol. 93, 2017, pp. 98-110.

[8] Ali, F., D. Kwak, P. Khan, S. M. R. Islam, K. H. Kim, and K. S. Kwak. Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling. *Transportation Research Part C: Emerging Technologies,* Vol. 77, 2017, pp. 33-48.

[9] Georgiou, T., A. El Abbadi, X. F. Yan, and J. George. Mining Complaints for Traffic-Jam Estimation: A Social Sensor Application. *Proceedings of the 2015 Ieee/Acm International Conference on Advances in Social Networks Analysis and Mining (Asonam 2015)*, 2015, pp. 330-335.

[10] Bridgelall, R. A participatory sensing approach to characterize ride quality.In *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2014, No. 9061*, Spie-Int Soc Optical Engineering, Bellingham, 2014.
[11] Zhang, S., J. Tang, H. Wang, and Y. Wang. Enhancing Traffic Incident Detection by Using Spatial Point Pattern Analysis on Social Media. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2528, 2015, pp. 69-77.
[12] Kurkcu, A., E. F. Morgul, and K. Ozbay. Extended Implementation Method for Virtual Sensors Web-Based Real-Time Transportation Data Collection and Analysis for Incident Management. *Transportation Research Record*, No. 2528, 2015, pp. 27-37.
[13] Kosala, R., E. Adi, and Steven. Harvesting Real Time Traffic Information from Twitter. *Procedia Engineering,* Vol. 50, 2012, pp. 1-11.
[14] Amer, A., E. Roberts, U. Mangar, W. H. Kraft, J. T. Wanat, P. C. Cusolito, J. R. Hogan, and X. Zhao. Traffic Incident Management Gap Analysis Primer.In, United States. Federal Highway Administration. Office of Operations, 2015.
[15] Kosala, R., and E. Adi. Harvesting real time traffic information from twitter. *Procedia Engineering,* Vol. 50, 2012, pp. 1-11.
[16] Grant-Muller, S. M., A. Gal-Tzur, E. Minkov, S. Nocera, T. Kuflik, and I. Shoor. Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data. *IET Intelligent Transport Systems,* Vol. 9, No. 4, 2014, pp. 407-417.
[17] Mai, E., and R. Hranac. Twitter interactions as a data source for transportation incidents.In *Proc. Transportation Research Board 92nd Ann. Meeting, No. 13,* 2013. p. 1636.
[18] Neto, V. R., D. S. Medeiros, and M. E. M. Campista. Analysis of mobile user behavior in vehicular social networks.In *Network of the Future (NOF), 2016 7th International Conference on the,* IEEE, 2016. pp. 1-5.
[19] Silva, T. H., P. O. V. de Melo, A. C. Viana, J. M. Almeida, J. Salles, and A. A. Loureiro. Traffic condition is more than colored lines on a map: characterization of waze alerts.In *International Conference on Social Informatics,* Springer, 2013. pp. 309-318.
[20] dos Santos, S. R., C. A. Davis Jr, and R. Smarzaro. Integration of data sources on traffic accidents.In *GeoInfo,* 2016. pp. 192-203.

[21] Amin-Naseri, M., P. Chakraborty, A. Sharma, S. B. Gilbert, and M. Y. Hong. Evaluating the Reliability, Coverage, and Added Value of Crowdsourced Traffic Incident Reports from waze. *Transportation Research Record*, Vol. 2672, No. 43, 2018, pp. 34-43.

[22] dos Santos, S. R., C. A. Davis Jr, and R. Smarzaro. Analyzing traffic accidents based on the integration of official and crowdsourced data. *Journal of Information and Data Management*, Vol. 8, No. 1, 2017, pp. 67-67.

[23] Amin-Naseri, M., 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, 2018, pp. 34-43.

[24] WAZE. *Connected Citizens Program*. https://www.waze.com/ccp. Accessed December 6, 2017.

CHAPTER 4

SECONDARY CRASH IDENTIFICATION USING CROWDSOURCED WAZE USER REPORTS

Abstract

Secondary crashes are considered to be crashes that occur as a result of the noncurrent congestion originating from primary crashes, which always has a greater impact on safety and traffic than a single crash. A better understanding of secondary crashes would benefit traffic incident management, and this requires accurate identification of secondary crashes. In this study, we explored using crowdsourced Waze user reports to identify secondary crashes. A network-based clustering algorithm was proposed to extract the primary crash cluster, including all user reports originating from the primary crash, and any crash that occurred within the cluster would be the secondary crash. This method worked as a filter to select accurate primary-secondary relationships, thus identifying the exact secondary crashes. Then, we performed a case study for crashes occurring from June to December 2019 on a 30-mile stretch of I-40 in Knoxville. A static threshold method (crash duration and 10 miles), was used to pre-select the potential primary-secondary crash pairs. We pre-selected 75 out of 708 crashes as potential secondary crashes. Based on the pre-selected primary-secondary crash pairs, 17 secondary crashes were obtained with our method. We compared the results of our method with one of the commonly used methods, the speed contour plot method. Though our method captured fewer secondary crashes, it did identify several secondary crashes that could not be observed with the speed contour plot method. The results showed the applicability of our method and the potential of crowdsourced Waze user reports.

Introduction

Traffic crashes are a critical issue that harms people's lives and negatively affects daily traffic operation (*1*). Secondary crashes occur as a result of primary crashes, which always result in more severe traffic congestion and road safety issues than a single crash. Therefore, an in-depth understanding of secondary crashes that benefits traffic incident management requires the accurate identification of secondary crashes.

Identifying secondary crashes is not a trivial task. Much research has been conducted and various methods have been developed to identify secondary crashes (*2-8*).

However, the data exploited in most of the previous studies are either collected from fixed infrastructure sensors placed on the road, such as video recognition cameras, inductive loop sensors, and radar sensors or obtained from sensor technology within the vehicle. These days, traffic information can be collected in a variety of ways owing to the development of advanced technologies. Among the existing traffic information collecting methods, crowdsourcing is a relatively reliable and cost-effective tool to collect traffic information covering a wide range of road networks, and it could be used as a complementary data source in addressing traffic issues.

Waze (https://www.waze.com/) is a widely-used example of traffic information crowdsourcing. It is a platform that enables people to share traffic information (e.g., incident reports, jam reports, construction reports) efficiently and in a timely manner. Waze user reports, generated when users encounter traffic congestion or road accidents, can provide insights into secondary crash identification. When a crash occurred on road, Waze users would report corresponding accident reports; if a secondary crash occurred as a result of the primary crash, there would be traffic jam reports reported to Waze. Thus, Waze user reports can be clustered to define the impact area of the primary crash, and any crash within the cluster of the primary crash can be considered as the secondary crash.

The objective of this study is to develop a framework for secondary crash identification with crowdsourced Waze user reports. Specifically, a network-based spatial-temporal clustering approach was proposed that adopts the knowledge of map-matching algorithms, ST-DBSCAN, and Dijkstra's algorithm. the methodology was then validated through a case study in Knoxville, Tennessee with crowdsourced Waze user reports. The results were also compared with a commonly used secondary crash identification method, the speed contour plot method.

In the remainder of this study, Section 2 presents the extant literature related to secondary crashes identification. Section 3 describes the proposed method for secondary crashes identification. To validate the proposed approach, a case study of Knoxville, TN was performed in Section 4. Section 5 presents the conclusion and future work.

Literature Review

Various methods have been developed for secondary crash identification such as static methods, dynamic methods, and speed contour plot methods. In the early stage, researchers used straightforward static and predefined temporal and spatial thresholds to identify secondary crashes. The assumption was that secondary crashes would occur within a certain spatial and temporal range of the primary crash. For example, Raub (*9)* defined the secondary incident as any incident that occurred within one mile upstream and incident duration of the primary incident plus 15 minutes. Hirunyanitiwattana and Mattingly (*10)* explored the characteristics of the secondary crash in which they defined the secondary crash as any crash that occurs in the same direction within 60 minutes and 2 miles upstream of the primary crash. Khattak, Wang and Zhang (*11)* investigated the relationships between primary incident duration and secondary incident occurrence, and they considered secondary incidents as incidents that occur on the same road segment and within the actual incident duration of the primary incident. Similarly, Moore, Giuliano and Cho (*12)* defined the secondary incident as any incident that occurs within two hours and two miles upstream in both directions of the primary incident. Despite the difference in the threshold used in the abovementioned studies, they are static and predefined regardless of the specific crash characteristics. These static approaches are not capable of accurately identifying secondary crashes with varying characteristics.

To overcome the limitations of static methods, many studies have developed diverse dynamic methods for secondary crash identification (*3; 4; 6; 13*), such as incident progression curve, queue length estimations, and shock wave theory. For example, based on the cumulative arrival and departure queuing model, Zhan, Gan and Hadi (*4)* proposed a model to estimate the maximum queue length and queue dissipation time and assumed that any incident that occurring within the above spatial and temporal range of the primary incident to be the secondary incident. Sun and Chilukuri (*13)* used the incident progression curve to dynamically identify secondary incidents and found that the incident progression curve method has a higher performance than the static method. Zheng et al. (*3)* applied shock-wave theory to estimate the dynamic impact of a primary incident, and

then an incident that occurred within the impact area was considered a secondary incident.

Much research has been conducted to develop a method to define the impact area of primary crashes dynamically and assume the crashes within the impact area of primary crashes to be the secondary crashes (*8; 14-17*). For example, Yang, Bartin and Ozbay (*8)* used the speed contour plot to identify secondary crashes on freeways and explored the characteristics of secondary crashes. Junhua et al. (*6)* determined the spatial-temporal impact area of the primary crash with a shock wave boundary filtering (SWBF) method and assumed that any crash that occurred within the impact area was a secondary crash. Based on the speed contour plot, Xu et al. (*15)* identified the secondary crashes and found that about 1.23% of the crashes were secondary crashes. Park and Haghani (*17)* used the bayesian structure equation model to define the impact area of the primary incident thus identifying the secondary incidents.

However, most previous secondary crash identification studies are based mainly on traffic speed data or travel time data, collecting from loop detectors, radar sensors, or probe vehicles, which may have several limitations such as limited coverage or malfunction issues. Hence, new data sources, such as crowdsourced data, are sought to identify secondary crashes. The integration of different data sources would results in more accurate secondary crash identification than each of them individually. Therefore, this study used the crowdsourced Waze user reports to identify secondary crashes with the proposed new network-based spatial-temporal clustering method.

Network-based spatial-temporal clustering

In this section, we propose a novel, network-based spatial-temporal clustering framework to cluster crowdsourced Waze user reports. The proposed approach is based on the knowledge of the map matching algorithm, ST-DBSCAN (Spatial-temporal density-based spatial clustering of applications with noise), Dijkstra's shortest path algorithm.

First, positioning technologies like GPS may produce different kinds of errors, making the location data not entirely accurate. In this study, the position of a user report may not be located exactly on the road network, so a process known as map matching is

used to determine the actual position of the report on the road. In the literature, various map-matching algorithms have been developed, which can be divided into four categories: geometric map-matching algorithms, topological map-matching algorithms, probabilistic map-matching algorithms, and other advanced map-matching algorithms (*18*). Two common datasets, location data and spatial road network data, are required in the majority of map-matching algorithms. The geometric map-matching algorithms are simple and easy to implement, and point-to-curve matching is a commonly used geometric map-matching algorithm that matches the position of a point onto its closest curve in the road networks. In this study, the point-to-curve matching methodology was applied. Basically, for each point obtained from a navigation system, a buffer zone is first created, and the road segments that intersect with the buffer zone are considered to be the candidate segments. Next, the distance from the point to the candidate segments are calculated. Finally, we project the point to the road segment with the shortest distance.

Next, the ST-DBSCAN clustering algorithm was used to cluster the Waze user reports. DBSCAN (Density-based spatial clustering of applications with noise), introduced by Ester et al. (*19)* in 1996, is one of the most widely used density-based clustering algorithms. It requires two parameters, neighborhood radius (ε) and the minimum number of points (*minPts*). DBSCAN defines clusters by examining the neighborhood points of a point p within the neighborhood radius (ε) iteratively. A point is considered to be a core point if it has at least *minPts* neighbors. A point q is defined as directly reachable from p if p is a core point and q is in the ε-neighborhood of p. A point q is defined as density-reachable from point p if there exists a path $p_1, ..., p_n$ with $p_1 = p$ and $p_n = q$, where each p_{i+1} is directly reachable from p_i. A density cluster contains the core point and all its density connected neighbors.

The ST-DBSCAN algorithm is a variation and extension of DBSACN, taking into account both spatial and non-spatial (e.g., time) aspects (*20; 21*). The difference between ST-DBSCAN and DBSCAN is that the neighborhood radius ε in DBSCAN is separated into two radii: the spatial neighborhood radius ε_s and temporal neighborhood radius ε_t . Therefore, a point q is the ε-neighborhood of point p if and only if the point q is within the ε_s-neighborhood and ε_t-neighborhood of point p. Similarly, the other concepts in ST-DBSCAN should be also extended accordingly based on DBSCAN.

Figure 4-1 depicts the pseudocode of ST-DBSCAN implemented in this study. First, for each data points, the (ε_t, ε_s)-neighborhood is obtained. If a point has at least *minPts* neighbors including itself, this point is considered a core point, which is qualified for starting up a cluster. Then, the core point is expanded with its directly reachable core points on the neighboring graph, ignoring the non-core points; Lastly, the non-core points within the (ε_t, ε_s)-neighborhood of a cluster are assigned to the nearby cluster, and other non-core points are assigned to noise.

Obtaining ε-neighborhoods is a major cost for density-based clustering algorithms, especially in transportation-related applications where road network distance is often required rather than Euclidean distance. In this study, gathering the ε_t-neighborhood is quite simple. The ε_t-neighborhood of *p* can be obtained by filtering out the points that are within the ε_t the point *p*. However, getting the ε_s-neighborhood may be complex. The computational complexity for obtaining road network distance using a shortest-path algorithm is much higher than that of euclidean distance. Hence, the modified Dijkstra's algorithm is used to improve the efficiency of obtaining ε_s-neighborhood in the algorithm (*22; 23*). For each point *p*, we want to find the shortest path between *p* and every other point. But, instead of traversing the entire road network, we control the algorithm by comparing the most lately determined shortest distance with the distance threshold ε_s. Because if the shortest distance between a point *q* and the point *p* is greater than ε_s, there is no need to evaluate other points since the distance to the source is increasing. Therefore, the modified Dijkstra's algorithm returns exactly the ε_s-neighborhood that is required in ST-DBSCAN.

Finally, we obtained the cluster for each primary crash, and we could check if there are traffic jam reports associated with the primary crash and if another crash is in the cluster due to the impact of the primary crash. If so, the latter crash could be the secondary crash.

Algorithm: ST-DBSCAN

```
1  ST-DBSCAN (D, ε_t, ε_s, minPts)
2      Cluster_id = 0
3      for each unvisited point P in dataset D:
4          mark P as visited
5          ε_s-neighborhood = spatialNeighbors(G, P, ε_s)
6          ε_t-neighborhood = timeNeighbors(D, P, ε_t)
7          N = ε_s-neighborhood ∩ ε_t-neighborhood
8          if sizeof (N) < minPts:
9              mark P as Noise
10         else
11             Cluster_id = Cluster_id + 1
12             expandCluster(P, N, ε_t, ε_s, Cluster_id, minPts)

13 expandCluster(P, N, ε_t, ε_s, minPts)
14     add P to cluster Cluster_id
15     for each point P' in N
16         if P' is not visited
17             mark P' as visited
18             ε_s'-neighborhood = spatialNeighbors(P', ε_s)
19             ε_t'-neighborhood = timeNeighbors(P', ε_t)
20             N' = ε_s'-neighborhood ∩ ε_t'-neighborhood
21             if sizeof (N') >= minPts
22             N = N joined with N'
23         if P' is not yet member of any cluster
24             add P' to cluster Cluster_id
```

Figure 4-1 ST-DBSCAN implementation

Case Study: The City of Knoxville, Tennessee

Study Area and Data

In this study, multiple datasets were obtained from the 30-mile (MM368.0 to MM398.0) segment on I-40 in Knoxville, Tennessee from June to December 2019, including Waze traffic jam and accident reports, traffic speed data, and shapefile of the road network. The shapefile of the I-40 freeway was obtained from Topologically Integrated Geographic Encoding and Referencing (TIGER) shapefile data developed by United States Census Bureau. The crash data were obtained from TDOT's Region 1 Traffic Management Center (TMC) through a web-based archiving tool, LOCATE/IM. The crash data are well structured, containing detailed incident information, such as incident duration, incident location (milepost), incident type, incident start time, response time, and the number of lanes blocked. A total of 708 crashes was obtained and used for the analysis, in which 337 crashes were on I-40 Eastbound and 471 crashes were on I-40 Westbound.

The high-resolution traffic data was obtained from the Remote Traffic Microwave Sensors (RTMS), maintained by TDOT. RTMS collects traffic information (e.g., traffic count, speed, and occupancy) for each lane every 30 seconds. Ninety-two RTMS stations are installed along the 30-mile long I-40 segment in both directions in which 47 RTMS stations are on I-40 Westbound and 45 RTMS stations are on I-40 Eastbound. The traffic speeds were aggregated into one-minute interval values for the analysis in this study.

The Waze user reports were obtained from Waze API, which is not publicly accessible but is available for Waze Connected Citizens Program partners. Once each minute we downloaded the XML file containing the real-time Waze user reports. Given that the XML file collection is re-executed frequently, the series of XML files need to be processed to eliminate duplicate user reports. After removing the duplicate reports, 113,508 Waze user reports were obtained within five meters of the 30-mile long I-40 freeway from June to December 2019. The user reports have four major categories: accident reports, traffic jam reports, weather hazard reports, and road construction reports. The four categories can then be further divided into subgroups, such as major accident reports, minor accident reports, weather reports (fog, rain, flood, snow), and construction reports. Since this study aims to identify secondary crashes with Waze user

reports, accident reports, traffic jam reports, weather reports, and construction reports were used, totaling 49,833 user reports. Each report contains detailed information including location (longitude and latitude), timestamp, a unique report ID, report type (accident or traffic jam report), and other information.

Secondary crash identification

For ST-DBSCAN, the important yet difficult task is to determine the parameters. The easier-to-set parameter is the *minPts* parameter. As a rule of thumb, the *minPts* should be set to at least twice the dataset dimensionality, but for high-dimensional data, noisy data, or for data has many duplicates, the *minPts* need to set larger (*19; 24*). In our study, the *minPts* was chosen as four. The radius parameter ε is often harder to set. It is preferred that this parameter is chosen based on the application domain knowledge (*25*). Therefore, one mile and 30 minutes were chosen as the distance radius and time radius to perform ST-DBSCAN clustering in our dataset because we were clustering the Waze traffic jam and accident reports on the freeway, and these reports disappeared after 30 minutes without further user feedback.

From the clustering results, we obtained 795 clusters with at least one accident report inside, which can be considered as crash events. In these clusters, 31 crash events contained weather reports, 51 crash events contained construction reports, and 366 crash events had at least two accident reports and jam reports. These 366 crash events could have had secondary crashes and were studied further for secondary crash identification. Figure 4-2 shows some examples of the clusters, demonstrating the capability and high routing flexibility of our proposed approach.

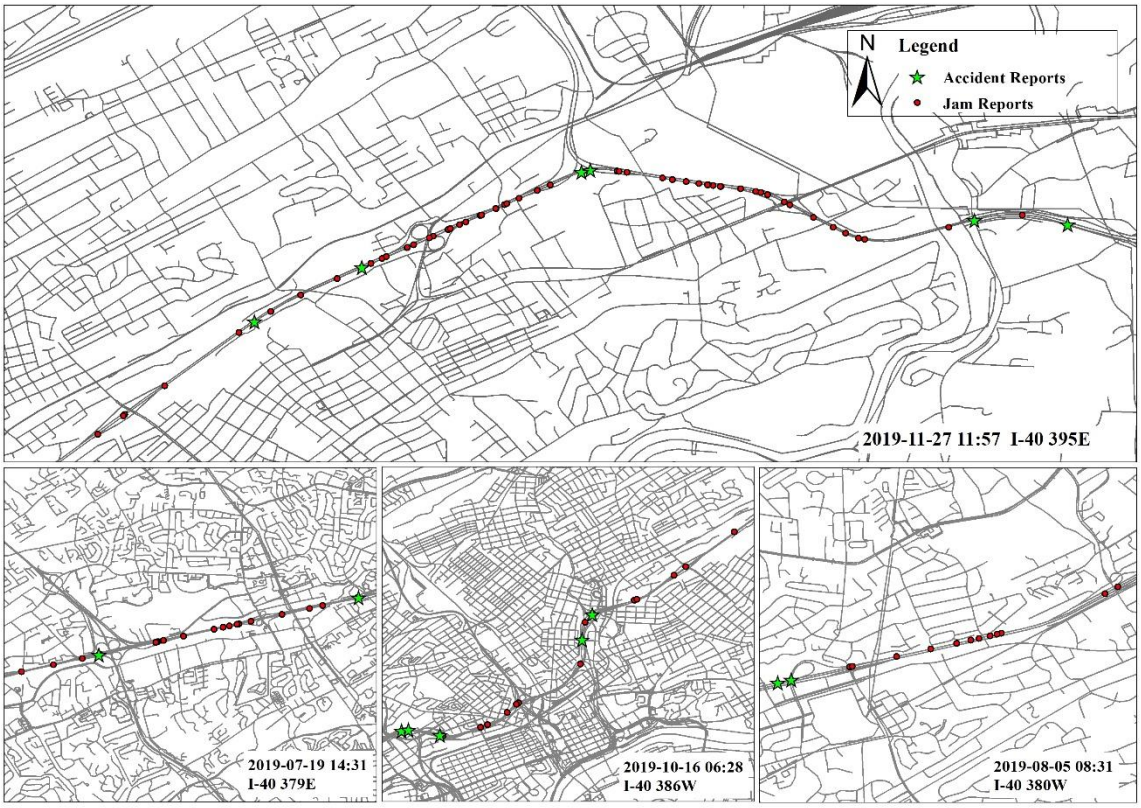


Figure 4-2 some examples of the obtained clusters

Since we obtained many crashes records from TDOT crash data, we performed a pre-selection process to get the potential primary-secondary relationships. Based on what we learned in previous research, we choose the static thresholds, actual incident duration, and 10 miles, to pre-select the possible primary and secondary crash pairs, and we obtained 62 possible primary-secondary crash pairs containing 75 secondary crashes from the original 708 crashes. Each crash pair may include multiple secondary crashes.

We then implemented the network-based clustering algorithm for secondary crashes identification as a filter to obtain accurate secondary crashes. Finally, from the pre-selected primary-secondary crash pairs, we observed 17 secondary crashes from 15 primary crashes. Figure 4-3 presents an example of the primary crash cluster in which queue formations and a secondary crash were observed after the primary crash. Though multiple reports were made hours after the primary crash, we could still observe that the primary crash occurred at the time of 14:14 on December 14, 2019, at milepost 391.0 of I-40 Eastbound and a queue was formed and propagated because of the primary crash. Any crash that occurred due to the primary crash was identified as a secondary crash. From the figure, one secondary crash was observed at the time of 14:29 on December 14, 2019, and the milepost of I-40 389.0 Eastbound. In addition, the primary crash and secondary crash occurred at 14:15 on December 14, 2019, and 14:32 on December 14, 2019, respectively in the TDOT crash data. The times of crashes in Waze accident reports are slightly earlier than the times in TDOT crash data, suggesting the temporal accuracy of Waze accident reports.

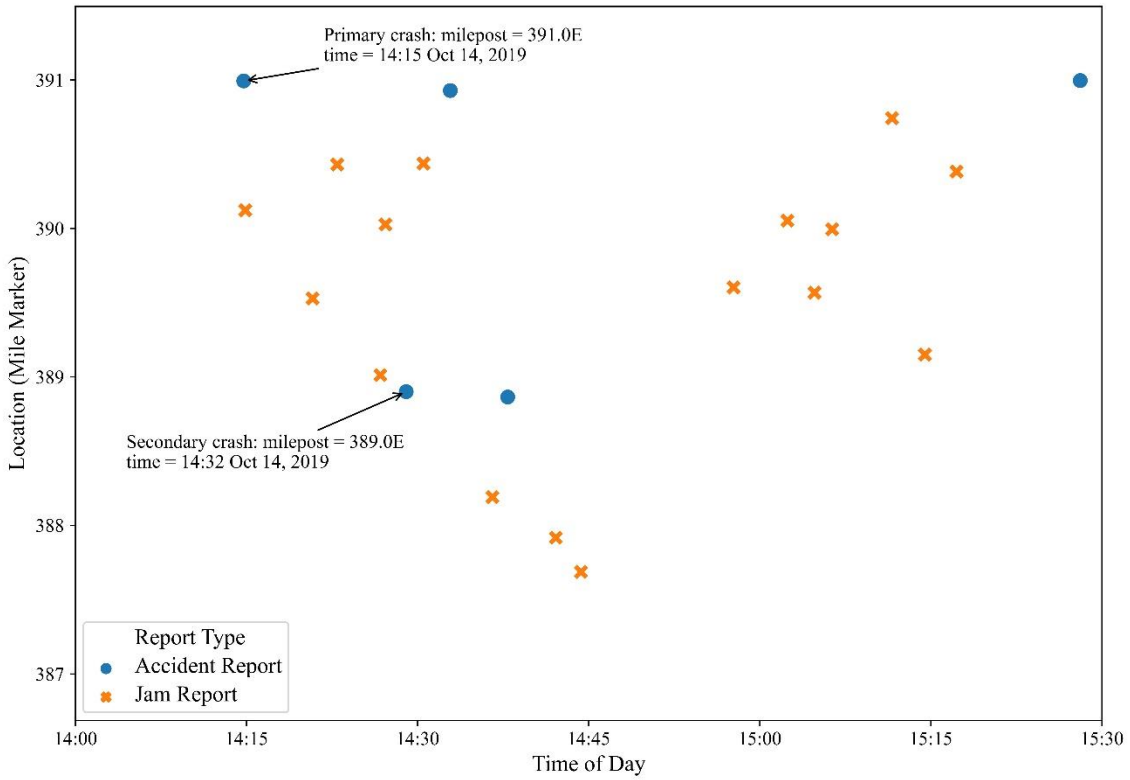


Figure 4-3 An example of a primary-secondary crash relationship captured by the proposed method

Comparison with the speed contour plot method

To validate our proposed method, we compared our results with the speed contour plot method for secondary crashes identification. First, if we used the static fixed spatial and temporal threshold, the above-mentioned incident duration and 10 miles upstream of the primary crash, we identified 75 secondary crashes from 708 crashes. This would be the rough pre-selection process, which can cause several false identifications.

Then, the speed contour plot method was used for secondary crashes identification to compensate for the limitation of static thresholds. The speed contour plot method has the following steps:

- We plotted the speed contour map for the primary crash with the RTMS one-minute speed data. In this study, the RTMS speed data was extracted at six-hour time intervals before and after the primary crash and five miles downstream and 10 miles upstream from the corresponding nearest RTMS station of the primary crash. Figure 4-4(a) demonstrates an example of the speed contour plot for a primary crash that occurred at 14:15 on December 14, 2019. We could observe that the congestion occurred around 14:15, but we were not able to determine if the congestion was recurrent or caused by the primary crash.
- To compensate for the effect of recurrent congestion, we further extracted the RTMS one-minute speed data from crash-free days in our study period for the same time intervals, locations, and day of the week. Then, for each time and location, we subtracted the average speed values of crash-free days from the extracted speed values of the crash day. A new speed contour plot was developed with the speed difference for various times and locations, which defined the spatial and temporal impact area of the primary crash. Figure 4-4(b) depicts the new speed contour plot for the same primary crash. From the figure, we can observe the spatial and temporal impact ranges of the primary crash, then if any crash occurred in the impact area of the primary crash, we could assume it to be the secondary crash. Using the speed contour plot method, we identified 39 secondary crashes associated with 32 primary crashes from the pre-selected primary-secondary crash pairs.

Table 4-1 summarizes the secondary crash identification results from different methods. As shown, using static thresholds, 75 (10.6%) out of 708 crashes were identified as secondary crashes; with the speed contour plot method, we obtained 39 (5.5%) secondary crashes out of 708 crashes, and 17 (2.4%) secondary crashes were obtained with our proposed network-based clustering algorithm. The speed contour plot method and our method worked as filters to eliminate false identification and obtain accurate secondary crashes. It seems that only a few secondary crashes were identified with crowdsourced Waze data, which may because of the insufficient Waze accident or jam reports of the crash since it depends on the number of Waze users on road. In the 17 secondary crashes identified by our method, 14 secondary crashes were also identified by the speed contour plot method, but three secondary crashes were only observed with Waze data using our method. The results demonstrated the potential of crowdsourced Waze data, which can serve as an alternative data source thus being incorporated into traffic management to improve the secondary crash identification performance.

Table 4-1 The results of secondary crash identification with different methods

Method	Data	Total number of crashes	Number of secondary crashes	Percentage of secondary crashes
Static Threshold: incident duration and 10 miles	NA	708	75	10.6%
Speed contour plot	High-resolution RTMS speed data	708	39	5.5%
ST-DBSCAN	Waze traffic accident and jam reports	708	17	2.4%

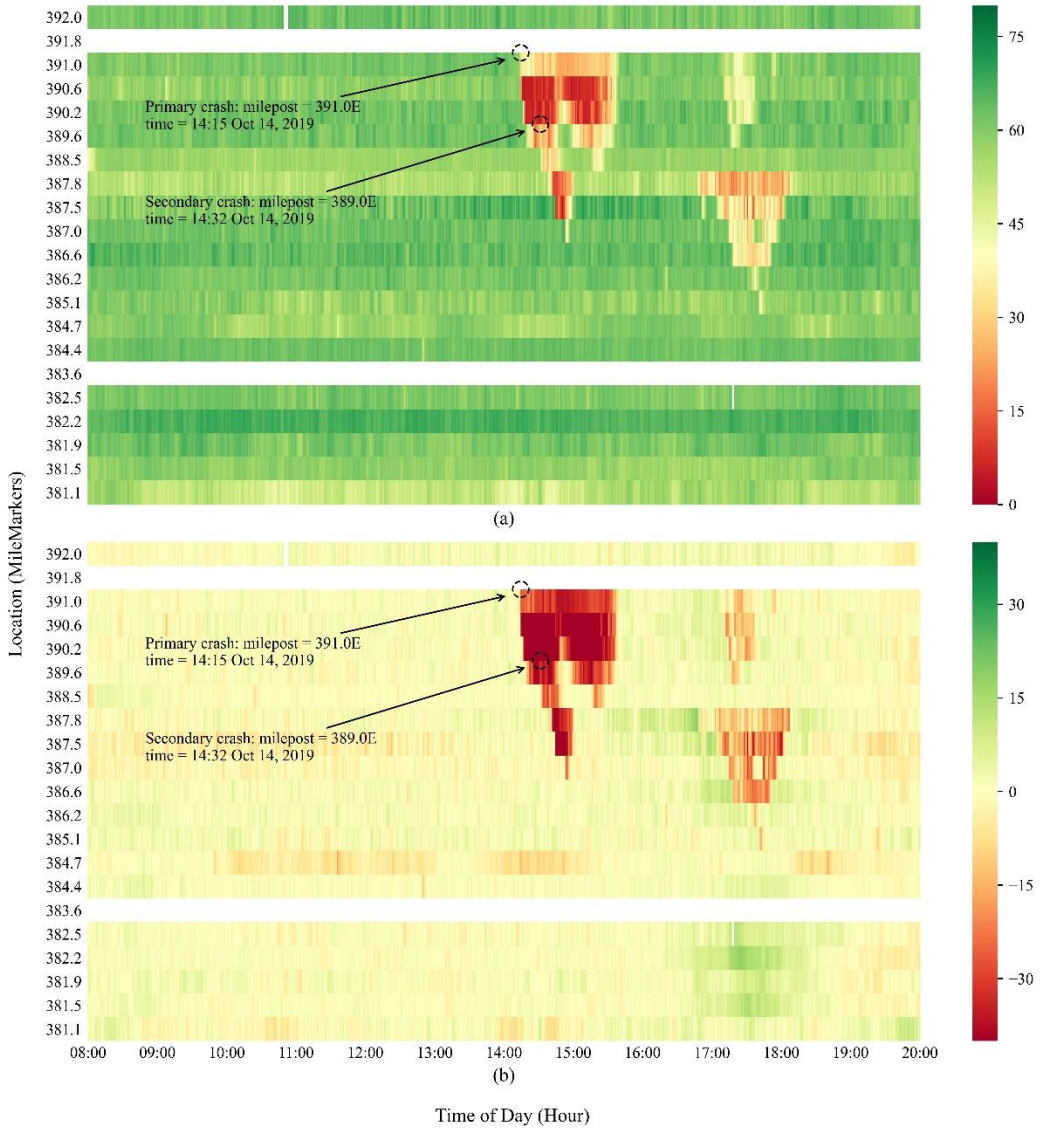


Figure 4-4 An example of the speed contour plot without (a) and with (b) accounting for the recurrent congestion

Conclusion

Secondary crashes are crashes that occur as a result of noncurrent congestion originating from primary crashes, which usually have a greater impact on safety and traffic than a single crash. A better understanding of secondary crashes would benefit traffic incident management, which requires accurate identification of secondary crashes. However, most previous studies focus on identifying secondary crashes with traffic speed or travel time data obtained either from fixed mounted sensors or probe vehicles, which may have several limitations such as limited coverage, missing data, and malfunction issues. To address these issues, this study explored using crowdsourced Waze user reports to identify secondary crashes.

We propose a network-based clustering algorithm to extract the primary crash cluster, including Waze user reports originating from the primary crash and assume any crash that occurs within the cluster of the primary crash is the secondary crash. This method filtered the data to select accurate primary-secondary relationships, thus identifying the correct secondary crashes. Then, we performed a case study of crashes occurring from June to December 2019 on a 30-mile segment of the I-40 freeway in Knoxville, Tennessee. A static threshold method (crash duration and 10 miles) was used to pre-select the possible primary-secondary crash pairs. Seventy-five out of 708 crashes were pre-selected as secondary crashes. Based on the pre-selected primary-secondary crash pairs, 17 secondary crashes were obtained with our method. Also, we compared the results of our method with the commonly used speed contour plot method. Though our method captured fewer secondary crashes, it also identified several secondary crashes that could not be identified using the speed contour plot method. Our method provides the potential of integrating these two datasets for secondary crash identification. The results showed the applicability of our method and the potential of crowdsourced Waze user reports in secondary crash identification.

Several limitations should be mentioned since these could be major efforts for future work. First, our methods used the pre-defined time and distance thresholds for ST-DBSCAN, which may be subjective and require advanced methods to determine the optimum parameters. Second, to comprehensively understand the secondary crashes,

more years of data and data from multiple locations should be collected to establish a greater sample of secondary crashes. Finally, crowdsourced Waze reports could be an important alternative data source in identifying secondary crashes and may be relevant for other transportation applications.

References

[1] Kong, X. J., Z. Z. Xu, G. J. Shen, J. Z. Wang, Q. Y. Yang, and B. S. Zhang. Urban traffic congestion estimation and prediction based on floating car trajectory data. *Future Generation Computer Systems-the International Journal of Escience,* Vol. 61, 2016, pp. 97-107.

[2] Vlahogianni, E. I., M. G. Karlaftis, J. C. Golias, and B. M. Halkias. Freeway Operations, Spatiotemporal-Incident Characteristics, and Secondary-Crash Occurrence. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 2178, No. 1, 2010, pp. 1-9.

[3] Zheng, D., M. V. Chitturi, A. R. Bill, and D. A. Noyce. Identification of Secondary Crashes on a Large-Scale Highway System. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 2432, No. 1, 2014, pp. 82-90.

[4] Zhan, C., A. Gan, and M. Hadi. Identifying Secondary Crashes and Their Contributing Factors. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 2102, No. 1, 2009, pp. 68-75.

[5] Karlaftis, M. G., S. P. Latoski, N. J. Richards, and K. C. Sinha. ITS Impacts on Safety and Traffic Management: An Investigation of Secondary Crash Causes. *ITS Journal - Intelligent Transportation Systems Journal,* Vol. 5, No. 1, 1999, pp. 39-52.

[6] Junhua, W., L. Boya, Z. Lanfang, and D. R. Ragland. Modeling secondary accidents identified by traffic shock waves. *Accident Analysis & Prevention,* Vol. 87, 2016, pp. 141-147.

[7] Park, H., A. Haghani, S. Samuel, and M. A. Knodler. Real-time prediction and avoidance of secondary crashes under unexpected traffic congestion. *Accident Analysis & Prevention,* Vol. 112, 2018, pp. 39-49.

[8] Yang, H., B. Bartin, and K. Ozbay. Mining the characteristics of secondary crashes on highways. *Journal of Transportation Engineering,* Vol. 140, No. 4, 2014, p. 04013024.

[9] Raub, R. A. Secondary crashes: An important component of roadway incident management. *Transportation Quarterly,* Vol. 51, No. 3, 1997.

[10] Hirunyanitiwattana, W., and S. P. Mattingly. Identifying secondary crash characteristics for California highway system.In, Washington, DC, 2006.

[11] Khattak, A., X. Wang, and H. Zhang. Are incident durations and secondary incidents interdependent? *Transportation Research Record,* Vol. 2099, No. 1, 2009, pp. 39-49.
[12] Moore, J. E., G. Giuliano, and S. Cho. Secondary accident rates on Los Angeles freeways. *Journal of transportation engineering,* Vol. 130, No. 3, 2004, pp. 280-285.
[13] Sun, C. C., and V. Chilukuri. Dynamic Incident Progression Curve for Classifying Secondary Traffic Crashes. *Journal of Transportation Engineering-Asce,* Vol. 136, No. 12, 2010, pp. 1153-1158.
[14] Chung, Y. Identifying primary and secondary crashes from spatiotemporal crash impact analysis. *Transportation research record,* Vol. 2386, No. 1, 2013, pp. 62-71.
[15] Xu, C., P. Liu, B. Yang, and W. Wang. Real-time estimation of secondary crash likelihood on freeways using high-resolution loop detector data. *Transportation Research Part C: Emerging Technologies,* Vol. 71, 2016, pp. 406-418.
[16] Park, H., and A. Haghani. Real-time prediction of secondary incident occurrences using vehicle probe data. *Transportation Research Part C: Emerging Technologies,* Vol. 70, 2016, pp. 69-85.
[17] Park, H., and A. Haghani. Real-time prediction of secondary incident occurrences using vehicle probe data. *Transportation Research Part C-Emerging Technologies,* Vol. 70, 2016, pp. 69-85.
[18] Quddus, M. A., W. Y. Ochieng, and R. B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C-Emerging Technologies,* Vol. 15, No. 5, 2007, pp. 312-328.
[19] Ester, M., H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise.In *Kdd, No. 96*, 1996. pp. 226-231.
[20] Birant, D., and A. Kut. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering,* Vol. 60, No. 1, 2007, pp. 208-221.
[21] Wang, M., A. Wang, and A. Li. Mining spatial-temporal clusters from geo-databases.In *International Conference on Advanced Data Mining and Applications*, Springer, 2006. pp. 263-270.

[22] Zhang, Y., L. D. Hang, and H. Kim. Dijkstra's-DBSCAN: Fast, Accurate, and Routable Density Based Clustering of Traffic Incidents on Large Road Network. *Transportation Research Record,* Vol. 2672, No. 45, 2018, pp. 265-273.
[23] Dijkstra, E. W. A note on two problems in connexion with graphs. *Numerische mathematik,* Vol. 1, No. 1, 1959, pp. 269-271.
[24] Sander, J., M. Ester, H. P. Kriegel, and X. W. Xu. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery,* Vol. 2, No. 2, 1998, pp. 169-194.
[25] Schubert, E., J. Sander, M. Ester, H. P. Kriegel, and X. Xu. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS),* Vol. 42, No. 3, 2017, p. 19.

CONCLUSION

This dissertation compiled a series of studies on evaluating and exploring crowdsourced Waze data and on using Waze data in traffic management. These studies were conducted to propose multiple applications to detect incidents on freeways with Waze traffic data, integrate Waze accident reports into incident management, and identify secondary crashes with Waze user reports.

First, we evaluated and explored the Waze traffic speed. Waze traffic speed is compared with commonly used infrastructure sensor speed data, and the effective sampling period of Waze speed is estimated. Waze traffic speed improves conventional traffic data in traffic monitoring with suggested control strategies. The results are important for understanding this data source and future resultant analysis.

Second, we proposed a calibration-free algorithm in automatic incident detection with Waze speed data. The algorithm is transferable and requires no calibration. The algorithm outperformed the benchmark algorithms in terms of detection rate (DR) and false alarm rate (FAR).

Third, we measured the quality of crowdsourced Waze accident reports spatially and temporally. The Waze accident reports were found to detect incidents earlier than the official incident dataset. Also, many of Waze accident reports are reliable but not found the corresponding record in the official crash dataset.

Last, a network-based clustering framework for secondary crashes identification with Waze user reports was proposed. The proposed framework captured several secondary crashes that cannot be observed by other secondary crashes identification methods, augmenting the accuracy of identifying accurate secondary crashes.

Overall, this dissertation provides multiple analysis frameworks and tools for practical applications with crowdsourced Waze data in traffic management. In terms of Waze traffic speed, Waze was found reliable to serve as an alternative dataset to augment the infrastructure sensor data. Knowing the characteristics and reliability of Waze traffic speed facilitates developing and building models to assist traffic managers to improve efficiency and effectiveness in traffic management. Waze user reports help in detecting incidents on road timely, thus mitigating traffic congestion and improving safety.

Moreover, the analysis frameworks in the dissertation are not limited to Waze data, but applicable to other crowdsourced data. For example, Google has been started collecting crowdsourced data, providing large volumes of data. Therefore, the future study can not only focus on investigating this data from different perspectives but on applying our analysis frameworks in the dissertation with other emerging datasets.

VITA

Zhihua Zhang was born in ShangRao City, Jiangxi Province, China. He received his B.S. and M.S. in Geography from Nanjing University, China. In 2020, he was granted a doctoral degree in Civil Engineering with a concentration in Transportation Engineering and M.S. degrees in both Statistics and Computer Science at the University of Tennessee, Knoxville (UT). As a Ph.D. student, Zhihua was the recipient of several scholarships, including John R. Harper Scholarship award from Tennessee Section Institute of Transportation Engineers (TSITE), TSITE student paper competition award, Lifesavers Traffic Safety Scholars award from the Committee of Lifesavers Conference on Highway Safety Priorities, Graduate Student Senate travel awards, and so on. His research interests include GIS in transportation, intelligent transportation systems (ITS), traffic operations, and transportation data and information systems.