




2021

ASSESSING FREEWAY CRASH RISK USING CROWDSOURCED WAZE INCIDENT ALERTS

Eugene Boasiako Antwi

University of Kentucky, antwi.eugene@uky.edu

Author ORCID Identifier:

 <https://orcid.org/0000-0002-1638-952X>

Digital Object Identifier: <https://doi.org/10.13023/etd.2021.145>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Boasiako Antwi, Eugene, "ASSESSING FREEWAY CRASH RISK USING CROWDSOURCED WAZE INCIDENT ALERTS" (2021). *Theses and Dissertations--Civil Engineering*. 108.
https://uknowledge.uky.edu/ce_etds/108

This Master's Thesis is brought to you for free and open access by the Civil Engineering at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Civil Engineering by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

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

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

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

REVIEW, APPROVAL AND ACCEPTANCE

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

Eugene Boasiako Antwi, Student

Dr. Mei Chen, Major Professor

Dr. Tim Taylor, Director of Graduate Studies

ASSESSING FREEWAY CRASH RISK USING CROWDSOURCED WAZE
INCIDENT ALERTS

THESIS

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

By

Eugene Boasiako Antwi

Lexington, Kentucky

Director: Dr. Mei Chen, Professor of Civil Engineering

Lexington, Kentucky

2021

Copyright © Eugene Boasiako Antwi 2021
<https://orcid.org/0000-0002-1638-952X>

ABSTRACT OF THESIS

ASSESSING FREEWAY CRASH RISK USING CROWDSOURCED WAZE INCIDENT ALERTS

Traffic data obtained through crowdsourcing are becoming more accessible to traffic agencies due to advancements in smartphone technology. Traffic managers aim to use this data to complement their conventional sources of data and provide additional context in their analysis. In this study, Waze incident alerts are integrated with GPS-Probe speed data and Kentucky State Police (KSP) crashes to assess their impact on traffic flow and safety on freeways in Kentucky. The analysis showed that the presence of a vehicle on the shoulder is associated with about 36.7% of freeway crashes in Kentucky. The presence of a vehicle on the shoulder coupled with congestion were 11.7% of the crashes. As such, the correlation between vehicle on shoulder, congestion and crashes was significant. Albeit present within the vicinity of 7.4% of crashes, the presence of a vehicle in the travel lane did not show as having a significant correlation with crashes. Linking Waze crash alerts with crashes and assessing their spatiotemporal patterns, it is found that Waze crashes are spatially accurate and hence could be used as an alternate source for identifying crashes, sometimes earlier, in Kentucky and hence cutting down incident response and clearance times. The data used in this study and the analytical methods employed offer much needed insight into the potential of crowdsourced traffic incident data for traffic monitoring to ensure safety.

KEYWORDS: Crowdsourced Data, Vehicle on shoulder, Data integration, Association Rule Mining, Vehicle stopped in road, Waze data

Eugene Boasiako Antwi

(Name of Student)

04/22/2021

Date

ASSESSING FREEWAY CRASH RISK USING CROWDSOURCED WAZE
INCIDENT ALERTS

By
Eugene Boasiako Antwi

Dr. Mei Chen

Director of Thesis

Dr. Timothy Taylor

Director of Graduate Studies

04/22/2021

Date

DEDICATION

To Angene Wilson

TABLE OF CONTENTS

LIST OF TABLES	vi
LIST OF FIGURES	vii
CHAPTER 1. INTRODUCTION	1
1.1 Background.....	1
1.2 Research Objectives.....	3
1.3 Chapter Organization	3
CHAPTER 2. LITERATURE REVIEW.....	4
2.1 Road shoulder and safety	4
2.2 Vehicles on Shoulder	4
2.3 Effect of Congestion on Safety	5
2.4 Characterizing Crowdsourced Waze Alerts.....	6
2.4.1 False Waze Alerts	6
2.4.2 Data Redundancy in Waze Alerts.....	7
2.5 Integration of Waze Data with other Datasets	8
2.6 Safety studies using Waze data.....	9
CHAPTER 3. DATA AND METHODOLOGY	12
3.1 Data Sources	12
3.2 Data Exploration	12
3.2.1 Spatial Distribution	12
3.2.2 Spatial and Temporal Depiction of the Data.....	16
3.3 Data Integration	20
3.4 Assessing Correlation between Factors	23
3.4.1 Contributory Factors Considered.....	24
3.4.2 Association Rule Mining	25
3.4.3 Applying Association Rule Mining	26
CHAPTER 4. RESULTS AND DISCUSSION	28
4.1 Spatiotemporal Pattern.....	28
4.2 Frequent Crash Contributory Factors.....	32
4.3 Association Rules.....	34

4.3.1	Correlation Between Individual Crash Factors and Crashes	34
4.3.1.1	Correlation Between Vehicles on Shoulder, Congestion and Crashes .	36
4.3.1.2	Correlation Between Vehicles Stopped in the Road and Crashes	37
4.3.2	Correlation Between Multiple Crash Factors and Crashes	37
4.4	Potential Additional Crash Coverage Provided by Waze	39
4.4.1	Waze Crash Alerts Linked to Crashes	39
4.4.2	Temporal Patterns in Waze Crash Clusters and Crashes	40
4.4.3	Spatial Accuracy of Waze Crash Reports	42
CHAPTER 5. CONCLUSION.....		47
5.1	Summary	47
5.2	Applications	48
5.3	Future Work	48
REFERENCES		50
VITA.....		53

LIST OF TABLES

Table 3.1 Percentage crashes with a vehicle on shoulder within their vicinity	21
Table 3.2 Human and environmental factors considered	24
Table 4.1 Frequently occurring crash contributory factors.....	33
Table 4.2 Two-item association rules indicating correlation between single factors.....	35
Table 4.3 Three-item association rules	38
Table 4.4 Waze crash alerts and crash integration.....	40
Table 4.5 Waze crash alert clusters related and unrelated to crashes	44

LIST OF FIGURES

Figure 3.1 Statewide distribution of Waze “vehicle on shoulder” alerts	13
Figure 3.2 Statewide distribution of Waze “vehicle stopped in road” alerts	13
Figure 3.3 Statewide distribution of crashes	14
Figure 3.4 Distribution of Waze “vehicle on shoulder” alerts within Louisville Metropolitan area	15
Figure 3.5 Distribution of Waze “vehicle stopped in road” alerts within Louisville Metropolitan area	15
Figure 3.6 Distribution of crashes within Louisville Metropolitan area.....	16
Figure 3.7 Waze “vehicle on shoulder” alerts preceding congestion and crashes.....	18
Figure 3.8 Waze “vehicle on shoulder” alerts succeeding crash report.....	19
Figure 3.9 Congestion succeeding a crash	20
Figure 4.1 Temporal analysis of crashes by hour	28
Figure 4.2 Statewide spatial distribution of "vehicle on shoulder" related crashes.....	29
Figure 4.3 Statewide spatial distribution of congestion related crashes	30
Figure 4.4 Statewide spatial distribution of “vehicle on shoulder” and congestion related crashes.....	30
Figure 4.5 Spatial distribution of “vehicle on shoulder” related crashes in Louisville Metropolitan Area.....	31
Figure 4.6 Spatial distribution of congestion related crashes in Louisville Metropolitan Area.....	31
Figure 4.7 Spatial distribution of “vehicle on shoulder” and congestion related crashes in Louisville Metropolitan Area.....	32
Figure 4.8 Time of day analysis.....	41
Figure 4.9 Day of week analysis.....	41
Figure 4.10 Distribution of distance of early reports in Waze from crash location.....	42
Figure 4.11 Distribution of distance of all Waze crash alerts from corresponding crash location.....	43
Figure 4.12 Temporal distribution of Waze crash alert report times for Waze crash alerts linked to crashes.....	44

CHAPTER 1. INTRODUCTION

1.1 Background

Around the world, road traffic crashes are a leading cause of injury and death. In the United States, an estimated 95% of transportation deaths and 99% of transportation injuries are attributable to highway crashes. The economic losses due to crash injuries and deaths, coupled with the delays to traffic resulting from crashes are undesirable. In particular, a crash on the road shoulder can reduce roadway capacity by up to 19% (Transportation Research Board, 2016) and for every 20 minutes the roadway remains uncleared, increases the likelihood of a secondary crash by up to 7% (Goodall, 2017). As such, considerable research efforts have been geared towards finding effective countermeasures to reduce crash risk and crash severity on highways. While the road shoulder is an important cross-sectional element of highways specifically designed for purposes which include, but not limited to, serving as a recovery area for driver errors and emergency stop (AASHTO, 2011), the use of the road shoulder for the latter poses an additional crash risk (Stamatiadis et al., 2009). As noted by Hauer (2000) an estimated ten percent of fatal freeway crashes are related to vehicles stopped on the shoulder. A similar estimate was obtained by Agent & Pigman (1989) who reported eleven percent of all fatal freeway crashes to be related to vehicles on the shoulder. Hence, it becomes imperative to understand the relationship between vehicles parked on the shoulders of highways and crashes if an optimal crash mitigation level is to be achieved.

In recent times, information technology advancements and rapid digital adoption have facilitated the collection of transportation related data and traffic monitoring. Conventional transportation systems management entails using ITS infrastructure such as

CCTV cameras and induction loops to monitor traffic conditions and collect data in locations where they are deployed. Consequently, traffic managers relied on emergency services dispatch to fill in data gaps outside of the ITS infrastructural network. Much recently, however, crowdsourced data generated actively and passively by road users has provided an inexpensive alternative to traffic monitoring. Thus, providing information such as traffic speed and traffic incidents including the near real-time location of vehicles on road shoulders, disabled vehicles in the road and objects in the road. Researchers have studied crowdsourced traffic incident data from Waze, one such application that affords road users the ability to actively report traffic incidents characterized by low false alarm rates (Amin-Naseri et al., 2018; Goodall & Lee, 2019; Liu et al., 2019).

Waze, through its Connected Citizens Program (CCP) provides the Kentucky Transportation Cabinet (KYTC) with real-time traffic incident alerts and traffic jam reports. KYTC uses this information to improve its traffic incident management operations and provide situational awareness to travelers. While traffic data from Waze has been used in literature for traffic crash estimation (Flynn et al., 2018), traffic crash monitoring (Young et al., 2019) and freeway traffic risk assessment (Turner et al., 2020), no previous researchers have attempted to use Waze “vehicle on shoulder”, “vehicle stopped in road” and “object in road” alerts to assess how traffic safety and flows are impacted, particularly of limited access highways. A better understanding of this relationship will aid the development of operational strategies and policies to reduce if not prevent future crashes and fatalities.

1.2 Research Objectives

Given that more and more traffic agencies are adopting crowdsourced traffic data sources which provide the locations of stationary objects and vehicles in traffic lanes and on the road shoulders, their impact on traffic flow and crashes. As such, the primary goals of this study are:

- To establish a spatial and temporal link between each crash, Waze "vehicle on shoulder" alert, and speed.
- To determine the correlation between vehicles on the shoulder, traffic slowdowns, and crashes.
- To evaluate the effect of vehicles stopped in traffic lanes and objects in traffic lanes on traffic safety and congestion.

It is hoped that this will provide a better understanding of the events leading up to the crash. Additionally, Waze crash alerts are linked with crash data to assess the potential additional coverage they provide.

1.3 Chapter Organization

This document consists of five chapters organized as follows. Chapter one introduces and provides a brief background to the topic as well as defines the research goals of the study. Chapter two presents an overview of relevant literature related to this study. Chapter three presents the data sources and methods of analysis employed in this study. Chapter four presents the results obtained following the analysis of the data and a discussion of the implication of the results. Finally, Chapter five presents a summary of the study and future work in this regard.

CHAPTER 2. LITERATURE REVIEW

2.1 Road shoulder and safety

Numerous studies have assessed the effect of road shoulder characteristics and occupancy on traffic safety. Narrow shoulders increase off-road collision risk (Kraus et al., 1993). This can be attributed to the inadequacy in driver recovery area provided by narrow shoulders should lane deviation occur. As noted by Hauer (2000) and Stamatiadis et al. (2009), wider shoulders give drivers a sense of security and space for making correctional maneuvers. As such, wider shoulders were associated with a decrease in crash rates (Choueiri et al., 1994; Gross & Donnell, 2011; Zegeer et al., 1980). Using 540 rural two-lane segments in America, Labi (2006) developed a crash prediction model which showed that wider shoulder widths had a substantial negative effect on the incidence and severity of crashes. In contrast, wider shoulders are associated with higher travel speeds (Mecheri et al., 2017) contributing to reckless driving. Labi (2011) attributes this to a false sense of security provided by wider shoulders.

2.2 Vehicles on Shoulder

To determine the impact of vehicles on the shoulders of limited access highways on crash incidence and severity, Agent & Pigman (1989) conducted observational surveys and analyzed crash data over a three-year period, from 1985 to 1987. They manually searched crash records to identify related crashes and discovered that on average, 1.9 crashes per 100 million vehicle miles were caused by a vehicle on an interstate or parkway shoulder. Agent & Pigman (1989) also found that eleven percent of all fatal freeway crashes were related to vehicles that had stopped on the shoulder. Similarly, Hauer (2000)

report that approximately ten percent of all fatal freeway crashes are related to vehicles stopped on the shoulder.

Crashes involving a vehicle stopped on the shoulder are more common at night when visibility is low and are more severe than all other types of crashes (Agent & Pigman, 1989). Vehicles parked on the shoulders of limited access highways also pose a higher risk of secondary collision. In a study using seven years of incident and crash data on freeways in Tennessee, Chimba & Kutela (2014) sought to identify secondary crashes that resulted from disabled and abandoned vehicles on freeway shoulders. They found 76% of the incidents involved a disabled or abandoned vehicle on the shoulder.

2.3 Effect of Congestion on Safety

The influence of traffic slowdowns on safety has been studied in the past with mixed conclusions. However, it is widely accepted that it is an important variable affecting traffic safety. Veh (1937) in his study concluded that the number of accidents per million vehicle miles was directly proportional to average daily traffic (ADT) up to an ADT of 7000 vehicles, after which there is a steady reduction in accident rates. This could be explained by increasing congestion resulting in decreases in speed (Raff, 1953). Similarly, Shankar et al. (1997) developed an accident frequency model for local arterials in Washington State, defining road sections by homogeneous characteristics including the annual average daily traffic (AADT). One of the study's main findings was that the frequency of crash incidence increases as the AADT per lane increases. Persaud & Dzbik (1993) investigated the nonlinear relationship between crash frequency and traffic volume. They discovered that on roadways with comparable traffic volumes, the number of crashes on a congested roadway was higher than for an uncongested roadway. Additionally, to

model traffic crash incidence and involvement on a sample freeway, Abdel-Aty & Radwan (2000) employed both negative binomial and Poisson regressions. According to the findings of their study, using AADT per lane as a measure of congestion, an increase in AADT per lane resulted in increased probabilities for higher crash frequencies. While increasing congestion increases traffic crash risk and a positive linear relationship has been found in literature (Head, 1959; Raff, 1953; Schoppert, 1957; Woo, 1957), one may argue that it decreases fatal crash risk as was found by Shefer (1994). Shefer (1994) reports that traffic fatalities were greatest at median levels of congestion and lowest when congestion was high or low. As such a greater understanding of the complex effects of congestion or traffic slowdowns on crashes is desired.

2.4 Characterizing Crowdsourced Waze Alerts

Crowdsourced data has been investigated as an alternative data source in the transportation industry due to the limited nature of traditional intelligent transportation infrastructure's traffic network coverage, as well as their high installation and maintenance costs (Jia et al., 2013; Pack & Ivanov, 2017; Yoon et al., 2007). Integrating crowdsourced datasets into traditional data sets generated by public agencies has also been shown to have benefits. Generally, understanding the characteristics of crowdsourced reports aids in the assessment of its reliability and the potential additional traffic coverage it provides.

2.4.1 False Waze Alerts

An inherent challenge with using actively crowdsourced traffic data is the possible presence of false incident reports in the dataset. While Waze attempts to reduce the incidence of false reports by prompting its users within the vicinity of an alert to confirm

or deny the report, false reports are nonetheless present in the data. As such, researchers have attempted to quantify the false alarm rate in Waze. Amin-Naseri et al. (2018) and Goodall & Lee (2019) compared Waze reports to screenshots of traffic camera video feeds taken at time intervals of five minutes and one minute respectively. They discovered that false alarm rates were significantly low. Of 319 Waze reports in the month of October, Amin-Naseri et al. (2018) report only one, representing 0.3%, was a false alarm. Similarly, Goodall & Lee (2019) report 5% false alarm rates for crash reports and 23% for disabled vehicle reports. The variance in false alarm rates may be attributed to the differences in the frequency with which they collect their ground truth for Waze incident validation, as well as differences in study area.

2.4.2 Data Redundancy in Waze Alerts

Also inherent in crowdsourced data is the issue of redundancy. That is, multiple reports of the same incident. As such, various approaches leveraging spatiotemporal as well as semantic information including incident type, road name and direction (Amin-Naseri et al., 2018; Eriksson, 2019; Lenkei, 2018) have been proposed in literature to minimize redundancy in Waze data by aggregating multiple reports that refer to the same incident. Amin-Naseri (2018) developed an R tool for the purposes of reducing redundancy, based on user specified constraints, using density-based clustering methods. As demonstrated by Amin-Naseri et al. (2018), Lenkei (2018) and Eriksson (2019), the intuition is to match crowdsourced alerts based on their semantic attributes and spatiotemporal proximity. In particular, specifying space-time proximity constraints is more effective at matching alerts (Eriksson, 2019). The result is a cluster of related alerts referring to the same incident and independent alerts not related to any alerts.

Consequently, a cluster of related alerts is represented as one alert thus reducing redundancy.

2.5 Integration of Waze Data with other Datasets

While crowdsourced data is a cost-effective alternative data source for traffic monitoring, it is frequently desired for traffic management purposes to integrate it with traffic incident data obtained from traditional data sources. To that end, the methods proposed in the literature match Waze incident alerts with traditional traffic data sources using spatiotemporal proximity constraints. For example, Goodall & Lee (2019) used space and time thresholds of 0.5 miles and 30 minutes to match Waze incident alerts to Virginia Department of Transportation official data. Aside from the spatiotemporal constraints, the two events had to occur on the same road and in the same direction of travel. The limitation of this approach is that it cannot distinguish between distinct incidents that are close in space and time.

Since the output of the various integration tools and methodologies developed by researchers is dependent on spatiotemporal constraints, its efficiency is affected by the spatial and temporal accuracy of the of the crowdsourced Waze incident alerts. When compared to their corresponding incident reports in official datasets, Waze incident alerts were found to be reported 2.2 to 9.8 minutes earlier (Amin-Naseri et al., 2018; Lenkei, 2018; Liu et al., 2019; Young et al., 2019). Liu et al. (2019) investigated the spatial accuracy of Waze incident alerts in Tennessee and discovered that the spatial difference between Waze crash and stopped vehicle alert locations and their corresponding official traffic data locations on Interstates was less than 0.001 mile and 0.0025 mile, respectively.

As a result, integrating Waze incident data into traffic management systems could be accomplished with reasonable accuracy.

Researchers can assess the additional benefit Waze provides to traffic managers after integrating Waze data with official data sources. Amin-Naseri et al. (2018) investigated the additional coverage that Waze could provide to the Advanced Traffic Management Systems (ATMS) and concluded that Waze data could provide an additional 34.1% coverage. Generally, the various studies report that Waze detects more than 40% of official records, with the exception of (dos Santos et al., 2017), who reported that Waze detected 7% of official records in his study in Belo Horizonte, Brazil. However, only a small percentage of Waze data is reported in official records (Amin-Naseri et al., 2018; Eriksson, 2019; Flynn et al., 2018). This indicates the potential additional data Waze could provide to traffic management systems, particularly on low severity crashes which are underrepresented in police crash reports and the location of disabled or abandoned vehicles on the shoulder.

2.6 Safety studies using Waze Data

Waze data has been employed in literature for safety related studies. Flynn et al. (2018) used six months of Waze incident data, as well as historical crash data, weather data, traffic volume data, and socio-economic data, to develop a crash prediction model to estimate the number and severity of crashes in Maryland. Based on these datasets, they employed random forest models and classification and regression trees for prediction. According to their findings, 57.05% of crashes were associated with at least one Waze event, while 5.98% of Waze events could be associated with crashes. On model

performance, the model could estimate the number of crash incidents with sufficient accuracy with spatiotemporal patterns close to ground truth official crash data.

Also, Turner et al. (2020) used Waze incident reports in their crash risk prediction studies. Employing spatiotemporal approaches to reduce redundancy in the Waze dataset and merge the Waze incident data with police crash reports, they suggest that Waze incident reports and predicted crashes are significant predictors for estimating police crash reports. They also report that more high-risk road segments can be obtained by combining Waze incident reports and police crash reports than using Waze incident reports alone, police crash reports alone and predicted crashes alone. As such, integrating Waze data with crash data was better at estimating traffic crash risk.

From the preceding discussions, the road shoulder is an important cross-sectional element with respect to traffic safety. Its use as an emergency stop location for vehicles on freeways increases the likelihood of fatal crashes on average. Congestion is also regarded as an important factor influencing road safety. While there have been a few studies that have analyzed and characterized crashes involving vehicles on the shoulder, no studies have been found that investigate the effect of a vehicle on the shoulder on traffic flow. Furthermore, as more transportation agencies use crowdsourced Waze data, the near real-time location of vehicles on the shoulders can be obtained from this data. Based on the preceding discussions, this data has been shown to be spatially and temporally accurate, with low false alarm rates, and when combined with other data sources, can provide additional insights into the circumstances leading up to crashes. The limitation of searching crash records to identify vehicle on shoulder related crashes is that it only captures crashes directly involving vehicle on shoulder crashes. However, using

crowdsourced reports of vehicles on the shoulder, crashes indirectly involving vehicles on the shoulder may be captured. The discovery of the relationship will assist traffic managers who have access to the location of vehicles on the shoulder in developing operational strategies to improve safety. The methodology used, as well as a brief description and exploration of data, are shown in the following chapter to assess the impact of vehicles on the shoulder on congestion and safety.

CHAPTER 3. DATA AND METHODOLOGY

This chapter provides an overview of the data sources used and the methods used in achieving the objective of this research.

3.1 Data Sources

The study is based on three data sources, all of which cover the period from July to December 2018 for all mainline Interstates in Kentucky. The data sources used are official Kentucky State Police (KSP) crash data, GPS-based speed data obtained from HERE Technologies, and Waze incident data obtained through the Waze Connected Citizens Program (CCP) by the Kentucky Transportation Cabinet (KYTC). The HERE Technologies data and Waze data used for this study had been pre-conflated with KYTC's road network. As such, each data point had a distinct route identifier attribute that defined the county, road name and direction of travel. For each travel direction, the archived GPS-based speed data were available at two-minute epochs.

3.2 Data Exploration

3.2.1 Spatial Distribution

Figure 3.1, Figure 3.2 and Figure 3.3 show the statewide spatial distribution of Waze “vehicle on shoulder” alerts, Waze “vehicle stopped in road” alerts and crashes respectively.

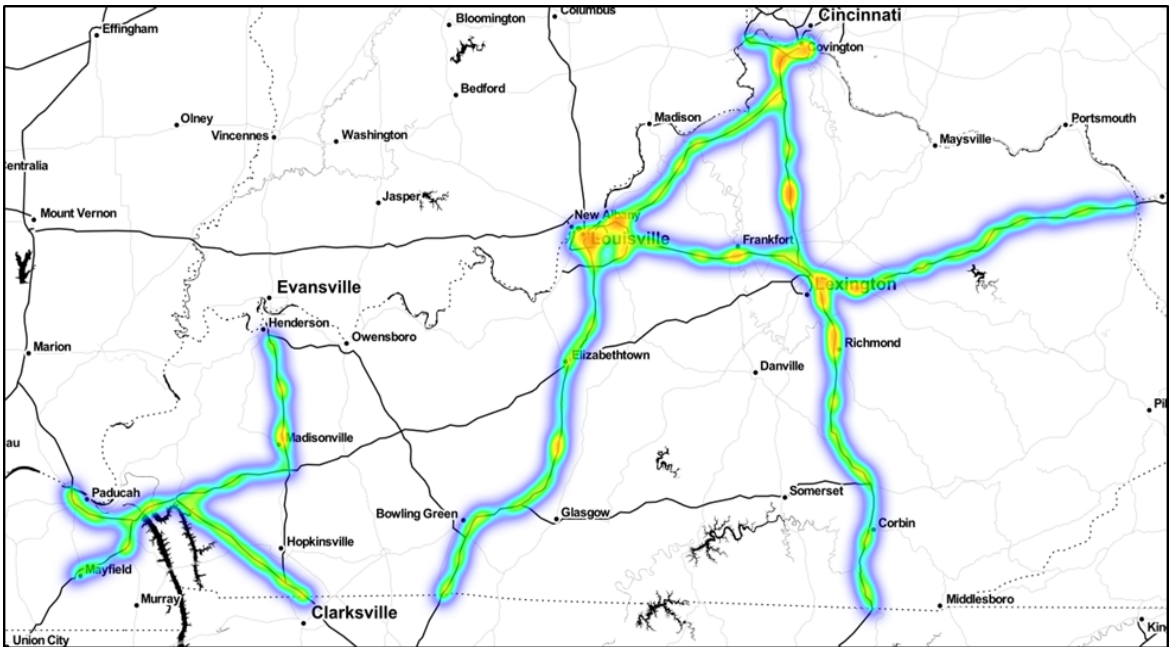


Figure 3.1 Statewide distribution of Waze “vehicle on shoulder” alerts

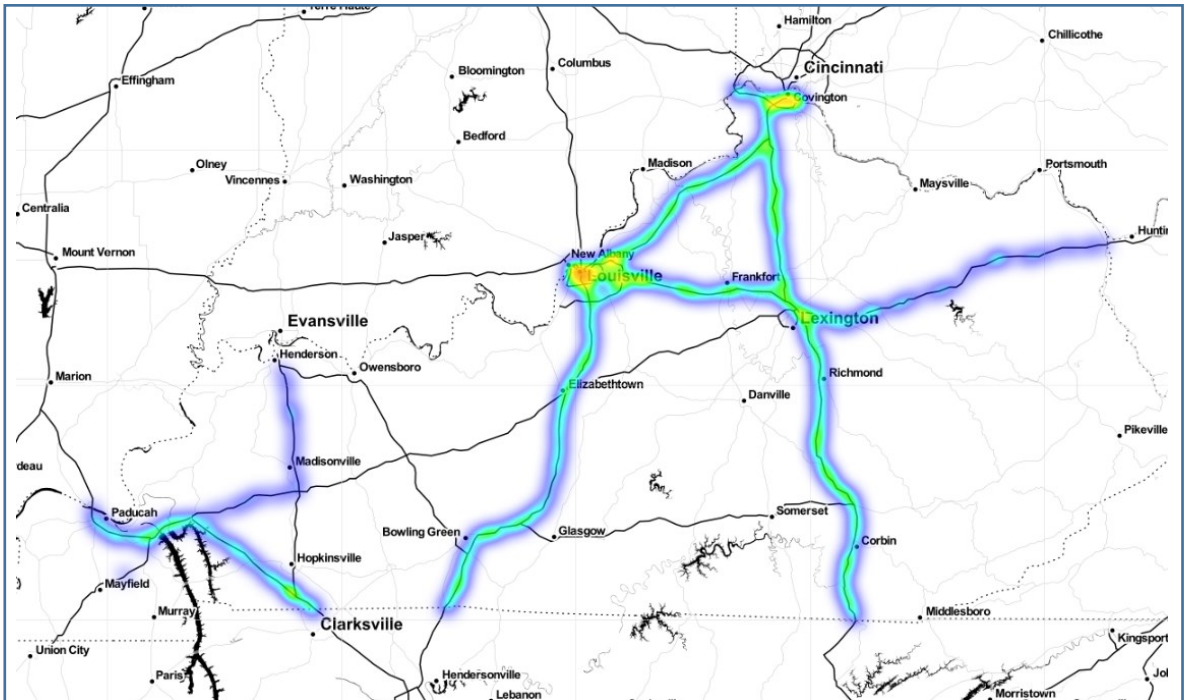


Figure 3.2 Statewide distribution of Waze “vehicle stopped in road” alerts

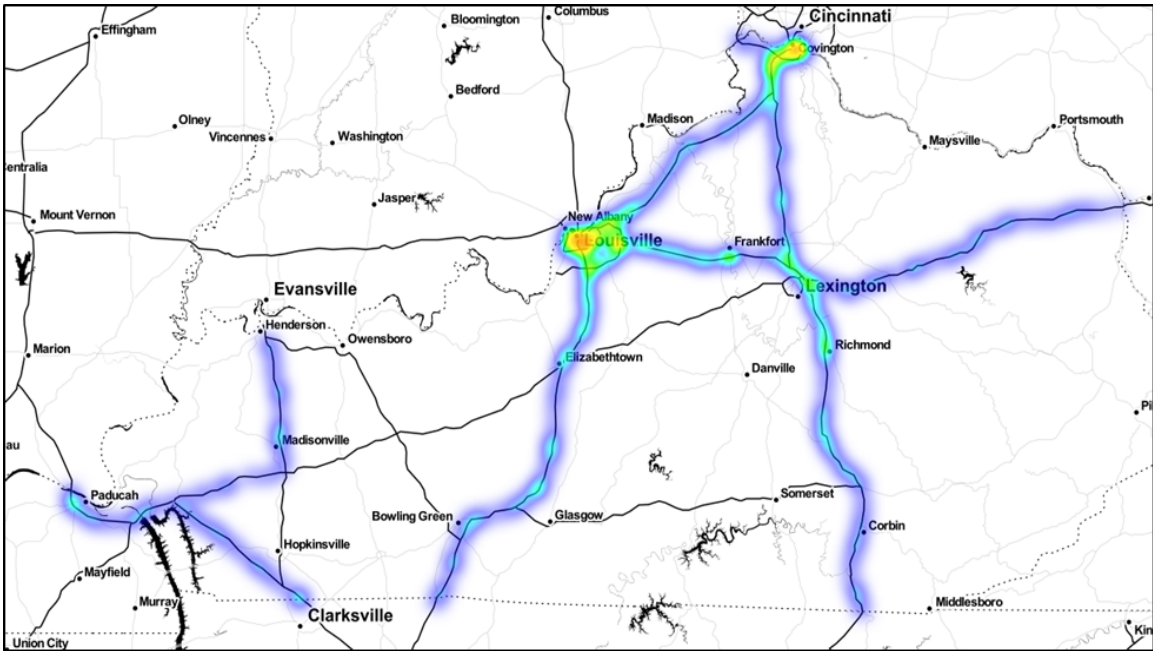


Figure 3.3 Statewide distribution of crashes

It is seen from the spatial distributions presented in Figure 3.1, Figure 3.2 and Figure 3.3 that there are more Waze “vehicle on shoulder” and “vehicle stopped in road” alerts in the urban areas than in rural areas. Particularly in Jefferson county and Northern Kentucky. The statewide distribution of traffic crashes on freeways in Kentucky shows a similar distribution.

Similarly, Figure 3.4, Figure 3.5 and Figure 3.6 shows the spatial distribution of these datasets within the Louisville Metropolitan Area. This is to depict a more detailed overview of the distribution of Waze “vehicle on shoulder” and “vehicle stopped in road” alerts as well as crashes.

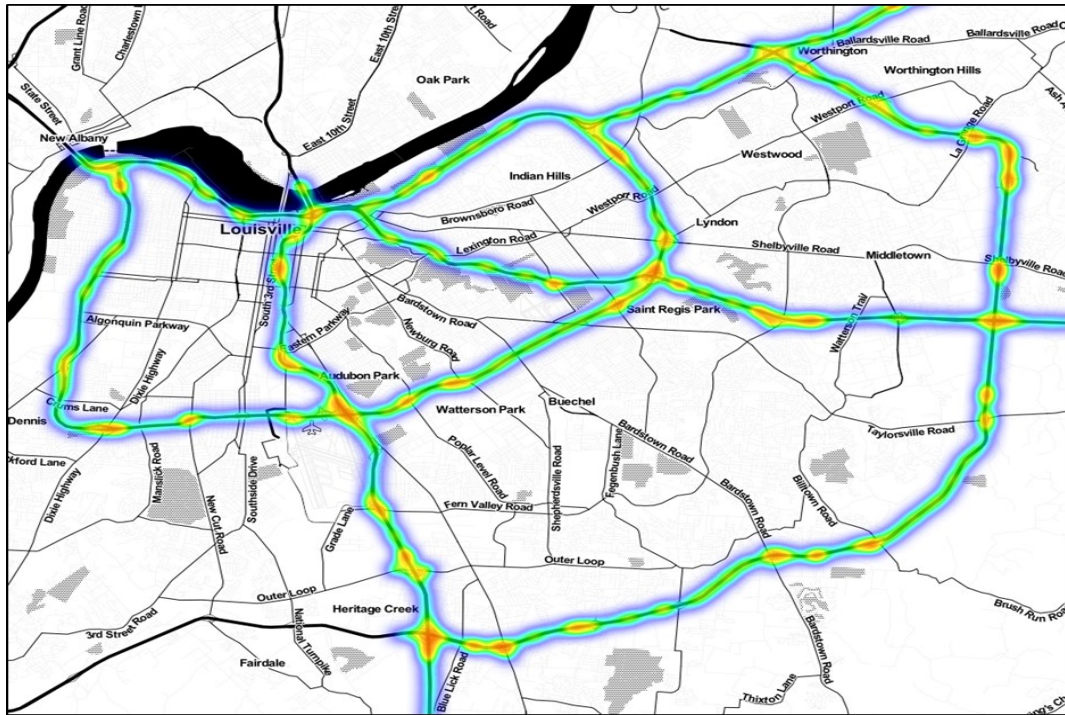


Figure 3.4 Distribution of Waze “vehicle on shoulder” alerts within Louisville Metropolitan area

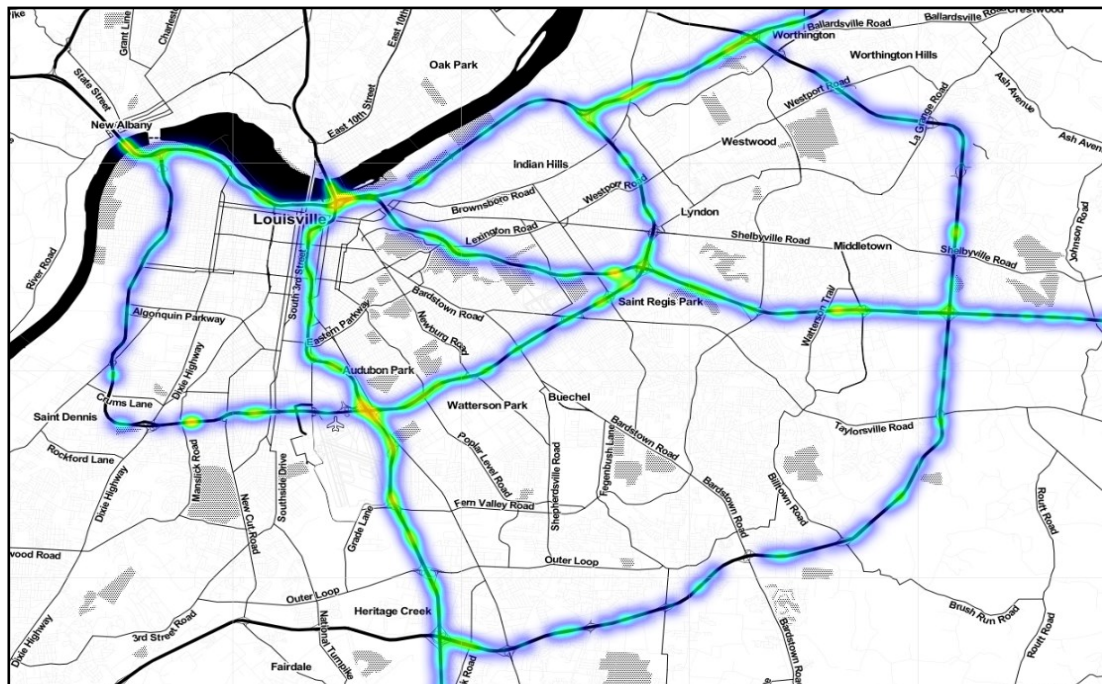


Figure 3.5 Distribution of Waze “vehicle stopped in road” alerts within Louisville Metropolitan area

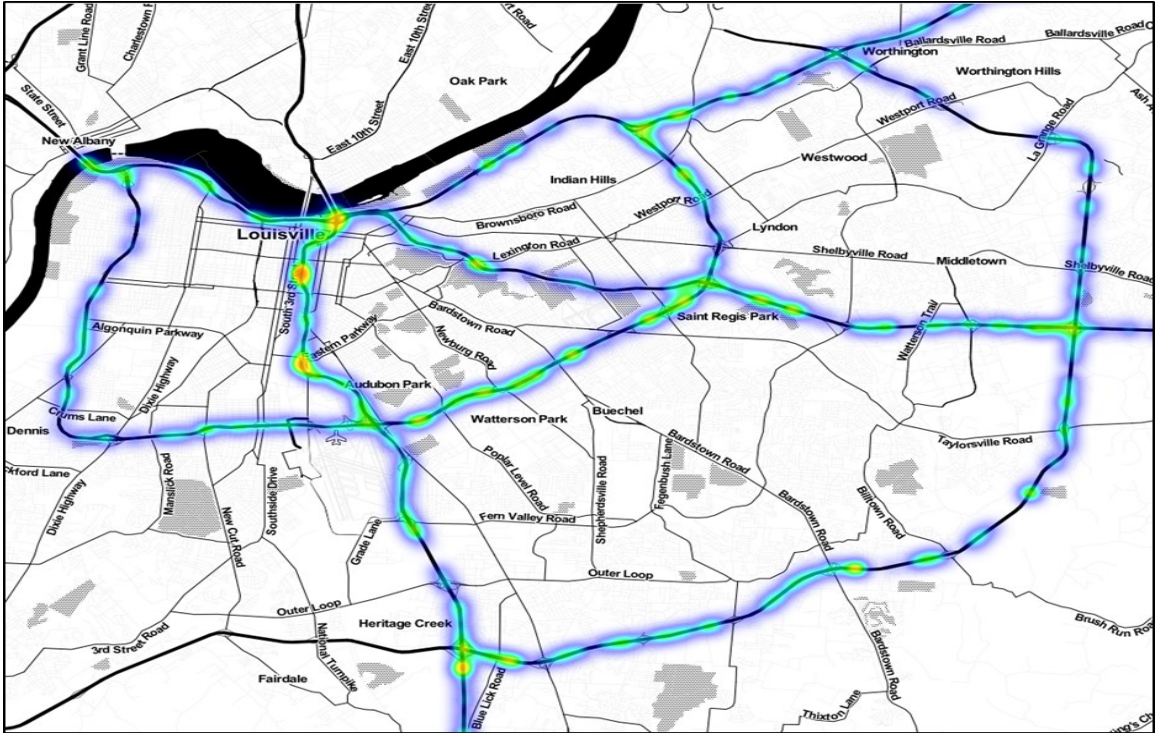


Figure 3.6 Distribution of crashes within Louisville Metropolitan area

It is seen from Figure 3.4, Figure 3.5 and Figure 3.6 that the “vehicle on shoulder” and “vehicle stopped in road” hotspots coincide with crash hotspots and these hotspots are highly concentrated near freeway intersections within the urban area.

3.2.2 Spatial and Temporal Depiction of the Data

To achieve the objectives of this research, the three data sources had to be integrated and consistent with previous research using crowdsourced data with other data sets, a spatiotemporal approach will be adopted. However, in order to set reasonable spatial and temporal integration thresholds, it is necessary to understand the interrelationships between the datasets. Here, heatmaps of the three datasets combined were used as a tool for this purpose. A heatmap was generated using the data sources for each day a crash occurred in the second half of 2018. This provided a visual representation of the interaction

between the datasets. First, speed data were queried and plotted and then the crashes, Waze crash alerts, Waze “vehicle on shoulder” alerts, Waze “vehicle stopped in road” alerts and Waze “object in road” alerts were subsequently overlaid to assess their interrelationship. All Waze alerts were charted based on their start times, with elongated symbology to depict their duration within the Waze data stream. Figures 3.7, 3.8, and 3.9 show examples of heatmaps generated. Each of these Figures represents only one travel direction — either cardinal or non-cardinal. Mile points increase in the cardinal direction (Northbound or Eastbound) and decrease in the non-cardinal direction (Southbound or Westbound). Lower mile points represent upstream in the cardinal direction and downstream in the non-cardinal direction of flow. A descriptive legend of the symbology is provided below for the figures:

- ★ – Crash
- ✚ – Waze “vehicle on shoulder” alert
- – Waze crash alert
- – Waze “object in road” alert
- ◆ – Waze “vehicle stopped in road” alert

Figure 3.7 illustrates a sequence of events in which reports of parked, disabled, or abandoned vehicles on the shoulder were followed by congestion and crashes.

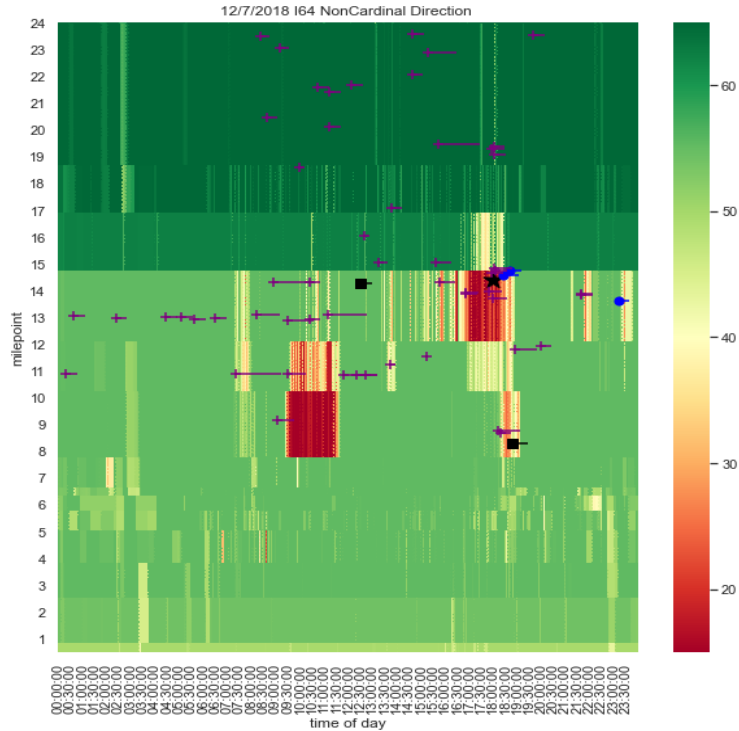


Figure 3.7 Waze “vehicle on shoulder” alerts preceding congestion and crashes

Figure 3.7 depicts congestion on I-64 westbound on a weekday in December 2018 between 9 a.m. and 12 p.m. and during the evening peak hours (between 4pm to 7pm). On this weekday, several reports of Waze "vehicle on shoulder" alerts had been received around mile point 13, prior to the evening peak. Following traffic congestion during the evening peak hours, a crash occurs, which is also reported in Waze. Five hours earlier, around the same mile point, a Waze “object in road” alert was received. Albeit not as frequent, the data shows this chain of events in which reports of parked, disabled, or abandoned vehicles on the shoulder were followed by congestion and crashes.

Similarly, Figure 3.8 shows instances where Waze “vehicle on shoulder” alerts were received following a crash report.

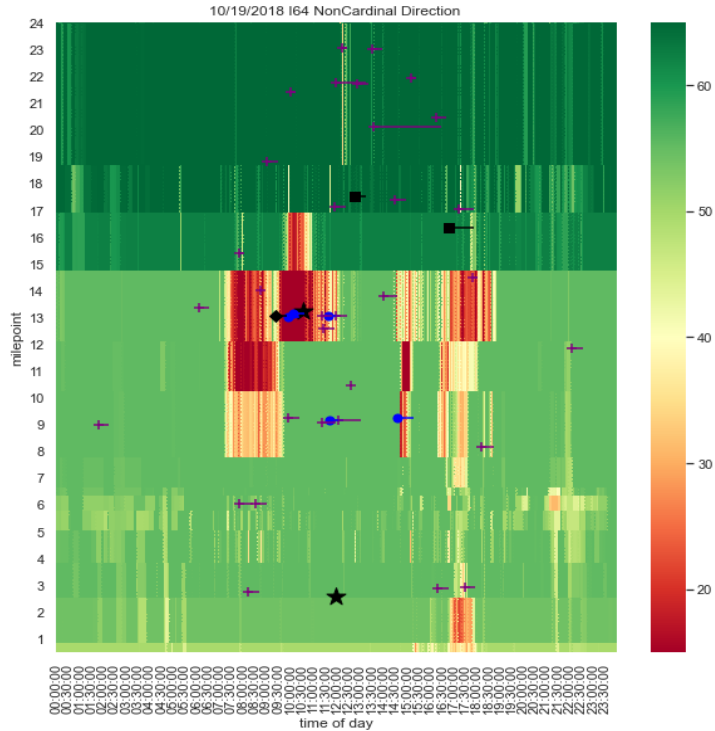


Figure 3.8 Waze “vehicle on shoulder” alerts succeeding crash report

In Figure 3.8, a Waze “vehicle stopped in road” alert is received during the morning peak on I-64 westbound, followed by Waze crash alerts, and then a crash report. The crash was reported earlier in Waze. This is followed by reports of Waze “vehicle on shoulder” incidents, which may refer to the vehicle involved in the crash being moved to the shoulder. To ensure the measured effects of this study were of vehicles on the shoulder or stationary vehicles and objects in the travel lanes leading to crashes, only Waze alerts received prior to a crash were considered for analysis.

Figure 3.9 also shows a scenario where there was congestion succeeding reports of a crash on a weekday on I-75 southbound.

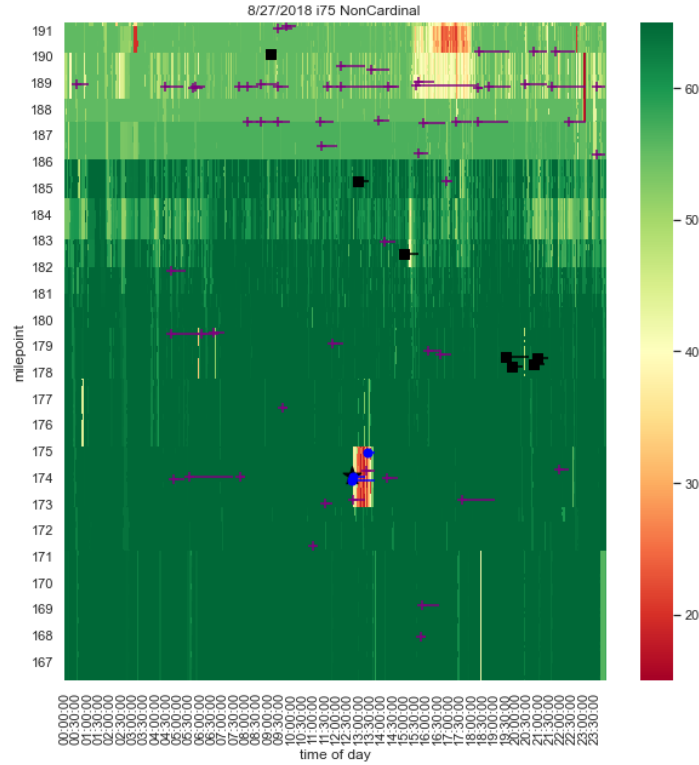


Figure 3.9 Congestion succeeding a crash

3.3 Data Integration

Having a visual representation of the interaction between the datasets, different space-time thresholds were tested to assess the impact presence of a vehicle on the road shoulder of a limited access highway prior to crash occurrence. This step is illustrated in Table 3.1, showing the percentage of crashes that had a Waze “vehicle on shoulder” alert within their spatiotemporal vicinity defined by the distance and time thresholds.

Table 3.1 Percentage crashes with a vehicle on shoulder within their vicinity

		Statewide	Jefferson	Fayette	Kenton	Boone	Campbell
Total Crashes (Jul-Dec 2018)		5768	1608	240	598	315	272
% Crashes with vehicle on shoulder alerts within the spatiotemporal vicinity							
30 min before	0.25 mi	36	48	28	32	35	48
	0.50 mi	54	66	46	48	52	64
	1.0 mi	72	83	65	66	70	82
30 min before and after	0.25 mi	47	59	41	44	45	59
	0.50 mi	64	76	60	59	62	76
	1.0 mi	80	89	77	75	79	89

In Table 3.1, statewide statistics for the spatiotemporal integration between Waze “vehicles on shoulder” alerts and crashes are presented. Also presented in Table 3.1 are statistics for some urban counties in Kentucky. For example, 1,608 crashes occurred in Jefferson County between July and December 2018, with 66 percent of these crashes having active “vehicle on shoulder” alert(s) 30 minutes prior to the crash and within 0.5 miles upstream and downstream of the crash site. As shown in Table 3.1, increasing the thresholds significantly increases the number of matches between Waze “vehicle on shoulder” alerts and crashes. However, for this study, a spatiotemporal threshold of 0.25 miles upstream and downstream of a crash site and 30 minutes before crash occurrence was used. This was deemed a reasonable threshold for identifying crashes caused by the presence of a parked, disabled, or abandoned vehicle on a limited access highway's shoulder. Additionally, to obtain a match, the two events had to be on the same road and

in the same direction of travel. Based on the data shown in Table 3.1, about 36% freeway crashes statewide had “vehicle on shoulder” alert(s) in their vicinities. The percentages were 48% for Jefferson and Campbell Counties, much higher than the statewide rate.

Having established a reasonable threshold for integrating Waze “vehicle on shoulder” alerts with crash data and speed data, the same spatiotemporal thresholds were similarly used to integrate Waze “object in road” and Waze “vehicle stopped in road” alerts with crash and speed data. Comparing these two types of the alerts to the Waze “vehicle on shoulder” alerts, over ten times more “vehicle on shoulder” alerts were received than these two alerts combined. The numbers of “vehicle stopped in road” alerts received however were only a third of the number of “object in road” alerts received. Based on the spatiotemporal threshold of 0.25 miles upstream and downstream of a crash site and 30 minutes prior to crash occurrence, only 7.4% and 4.2% of “vehicle stopped in road” and object in road alerts respectively had a crash within their vicinity.

Additionally, to ascertain the presence of congestion prior to crash occurrence, GPS-based speed data at the location of the crash were queried two minutes prior to crash time. Congestion was considered present if the query returned a speed value less than 45 miles per hour.

Crowdsourced incident data has the potential to capture traffic crash events that otherwise would not have been captured through conventional traffic data sources. To assess this potential additional coverage, Waze crash alerts need to be linked to crash data. However, the crowdsourced Waze crash incident alerts contained redundant records.

To mitigate redundancy, duplicate records of the same incident within the Waze dataset reported by the same Waze user were removed using their unique IDs (UUID).

However, at this stage the Waze crash data still contained multiple reports of the same incident reported by multiple Waze users and hence having different UUIDs. As such, to cluster together Waze crash incident alerts of the same incident from different users, a spatiotemporal clustering approach was used as proposed in literature. Here, similarly a spatial and temporal threshold of 0.25 miles upstream and downstream and 30 minutes respectively is used. The result of this is a dataset containing Waze crash incident clusters that are likely to be reports of the same incident. As such each cluster may be considered as a single incident. Then using the same spatial and temporal thresholds to integrate this data with crash data, crashes that were also reported in Waze were identified and Waze crash alerts that were found in crash data were identified.

Using the six months of Waze data there were 313,953 Waze crash alerts prior to filtering out Waze crash reports with duplicate reports of the same UUID. That is, it includes duplicate Waze crash reports of the same incident made by the same Waze user that appear in the data set more than once due to KYTCs data pulling frequency. Filtering out duplicate UUIDs, the resultant dataset contained 15,859 unique Waze crash reports made by different users that may still contain reports of the same incident made by different users. After clustering to filter out Waze reports made by different users referring to the same incident, there were 13,279 unique Waze crash report clusters in the data set as compared to 5,768 crashes.

3.4 Assessing Correlation between Factors

Crashes happen for a variety of reasons, including human factors, environmental factors, and vehicle factors. Crash occurrences are sometimes caused by the interaction of multiple of these factors. As a result, in order to assess the relationship between vehicles

on the shoulder and stationary vehicles and objects in the travel lanes on congestion and crashes on limited access highways, their interaction with a variety of human and environmental factors must be considered as well as the various factors may not act in isolation. The relationships were assessed using data mining techniques as described in this section.

3.4.1 Contributory Factors Considered

Human factors extracted for this study include driver impairment, distraction, inattention, driving too fast for conditions, improper vehicle maneuvers, failure to yield right of way, and following too closely. Environmental factors included roadway character — presence of curves and grades, inclement weather, poor visibility based on the lighting condition field in crash reports, animal/debris, water pooling, slippery road surface, and construction work zone. Table 3.2 provides a summary of the human and environmental factors considered in this study.

Table 3.2 Human and environmental factors considered

Human factors	Environmental factors
Driver impairment	Curves and Grades
Distraction	Inclement weather
Inattention	Poor visibility
Driving too fast for conditions	Waterpooling
Improper maneuvers	Slippery road surface
Following too close	Construction work zone
Failure to yield right of way	Presence of animals/Debris

Additionally, GPS speed data and Waze data were used to determine the presence of traffic congestion, vehicles on the shoulder, and stationary objects or vehicles in travel lanes prior to a crash.

3.4.2 Association Rule Mining

Much recently, data mining techniques have been adopted extensively in transportation research. In the past, statistical models, which have their inherent assumptions, were used to analyze road crashes and their causative factors (Lee et al., 2002). However, due to the limitation of statistical models for large dimensional datasets and the need to specify the functional form of statistical models prior to application, data mining algorithms such as Association rule mining have gained attention among the transportation safety research community in recent times. The basic idea is to identify frequent item sets within a large relational database using frequent item search algorithms such as Apriori (Agrawal et al., 1993) or FP-Tree ((Han et al., 2000) and identify relationships between these item sets based on measures such as the support-confidence framework (Agrawal et al., 1993).

In relation to transportation safety studies, Geurts et al. (2003) used association rules to analyze high-frequency crash locations in Belgium in order to identify frequently occurring crash patterns and the extent to which crash characteristics at these high-frequency crash locations differed from those at low-frequency crash locations. They concluded that, while human and behavioral factors played a significant role in the occurrence of crashes, the main difference in accident patterns between high-frequency and low-frequency accident locations could be found in infrastructure- or location-related circumstances. Similarly, Kumar & Toshniwal (2016) used association rules to identify

and characterize high-accident locations in India. They demonstrated that association rules were effective at uncovering relationships between crash factors, and that having more attributes in the data allows association rules to uncover more relationships.

In a case study using work zone crash data between 2004 and 2008, Weng et al. (2016) used association rules to assess the characteristics of work zone fatalities and to obtain an in-depth understanding of the contributory factors to such fatalities. They further emphasize that the use of association rule mining in other areas of traffic safety research will be beneficial and provide guidance in the selection of effective countermeasures.

Much recently, Das et al. (2020) used association rules to assess the contributory factors to flood related crashes in Louisiana. Originally developed for market basket analysis, association rule mining has become a good algorithm for analyzing traffic crashes to identify key contributory factors. The aforementioned studies were able to identify key contributory factors and recommend countermeasures to reduce, if not mitigate, crashes using association rules.

3.4.3 Applying Association Rule Mining

According to (Agrawal et al., 1993), association rule mining is defined as follows: Let $I = \{i_1, i_2, \dots, i_m\}$ be a universal set of crash-related factors, including human, environmental, and vehicle factors. Let $D = \{c_i, c_{i+1}, \dots, c_n\}$ be a set of the crashes from the crash data, where each crash has a unique crash ID (Cid) and an item set (C-itemset) consisting of the factors related to this specific crash. Let $X \subseteq I$, $Y \subseteq I$ each be a subset of the universal crash contributory factors. An association rule is the implication $X \rightarrow Y$ such that $X \cap Y = \emptyset$, $p(X) \neq 0$ and $p(Y) \neq 0$ where X is the antecedent and Y the consequent.

Indicators such as support, confidence, conviction, lift, leverage, and other measures can be used to assess the significance or effectiveness of a rule. However, for the purposes of this study, the three measures used are support, confidence, and lift. The frequency with which the antecedent and consequent of a rule occur together in crash data is referred to as rule support. It is computed using Equation 1.

$$\text{Support } (X \rightarrow Y) = \frac{P(X \cap Y)}{N} \quad (1)$$

Confidence (see Equation 2) refers to the strength of a rule's implication and is the proportion of crashes involving contributing factor X that also contain Y.

$$\text{Confidence } (X \rightarrow Y) = \frac{P(X \cap Y)}{P(X)} \quad (2)$$

Although the support-confidence framework is a common model for mining association rules, it does not provide a test for identifying the correlation between two item sets (Zhang & Zhang, 2002). As such, the lift measure, which measures the interdependence of factors, was used as a third measure. With values ranging from 0 to ∞ , a lift value of 1 indicates factors are independent, values greater than 1 denote positive correlation, and values less than 1 indicate negative correlation between factors. Mathematically, lift is computed as:

$$\text{Lift } (X \rightarrow Y) = \frac{P(X \cap Y)}{P(X) \times P(Y)} \quad (3)$$

Using the 'MLxtend' python package (Raschka, 2018), the Apriori algorithm was applied with a minimum support of 5% and a minimum lift of 1 so that only positive correlations between factors were reported.

CHAPTER 4. RESULTS AND DISCUSSION

This chapter presents the results obtained from the analysis performed to assess the correlation between the presence of vehicles on the shoulder, stationary vehicles or objects in travel lanes, congestion, and crashes. It also presents the spatiotemporal distribution of crashes involving vehicles on shoulder and congestion as well as the additional crash coverage Waze can provide.

4.1 Spatiotemporal Pattern

Following the integration of Waze “vehicle on shoulder” alerts with crashes and speed data as described in the previous chapter, the subsequent exploratory analysis to assess the temporal pattern of the crashes – including crashes with Waze “vehicle on shoulder” alerts present in their vicinity – indicates that more crashes occurred during the peak hours of the day on limited access highways. Crashes with “vehicle on shoulder” alerts within their spatial and temporal vicinity also followed this trend as shown in Figure 4.1.

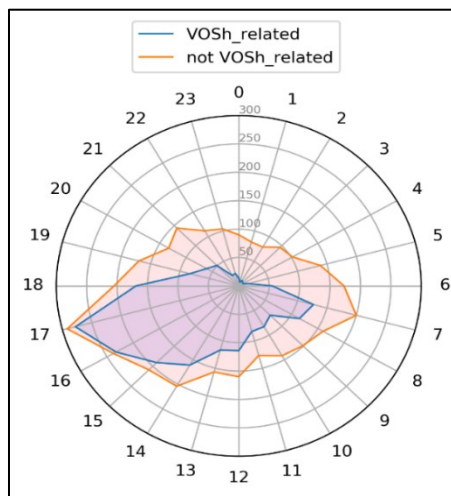


Figure 4.1 Temporal analysis of crashes by hour

Assessing the spatial distribution of crashes that had a “vehicle on shoulder” alert in their spatiotemporal vicinity based on the thresholds set as described in the previous chapter, more of such crashes were observed to have occurred in the urban areas, particularly in northern Kentucky and the Louisville Metropolitan area. The same is true for crashes which’s occurrence were preceded by congestion. The spatial distributions of the “vehicle on shoulder” related crashes statewide and within the Louisville Metropolitan area are presented in Figure 4.2 and 4.5 respectively. Similarly, the distribution of congestion related crashes statewide and within the Louisville Metropolitan area are presented in Figure 4.3 and Figure 4.6 respectively. Also, the spatial distribution of crashes that had both congestion prior to their occurrence and vehicles on shoulder present within their vicinity is presented in Figure 4.4, for statewide, and Figure 4.7 for the Louisville Metropolitan area.

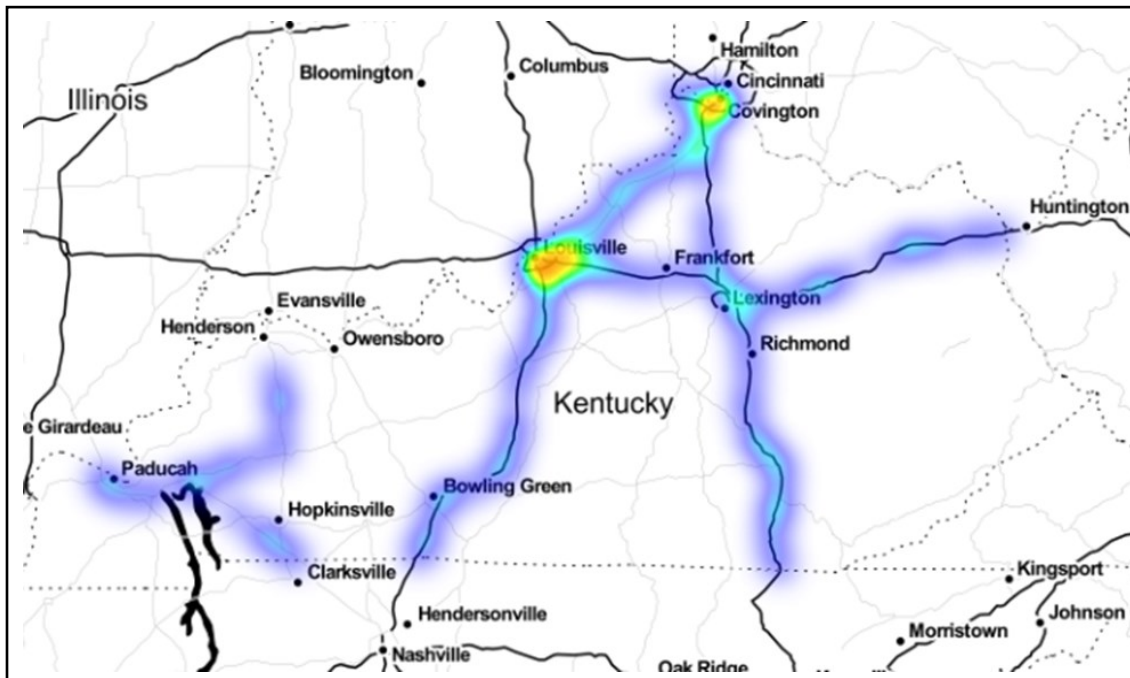


Figure 4.2 Statewide spatial distribution of "vehicle on shoulder" related crashes

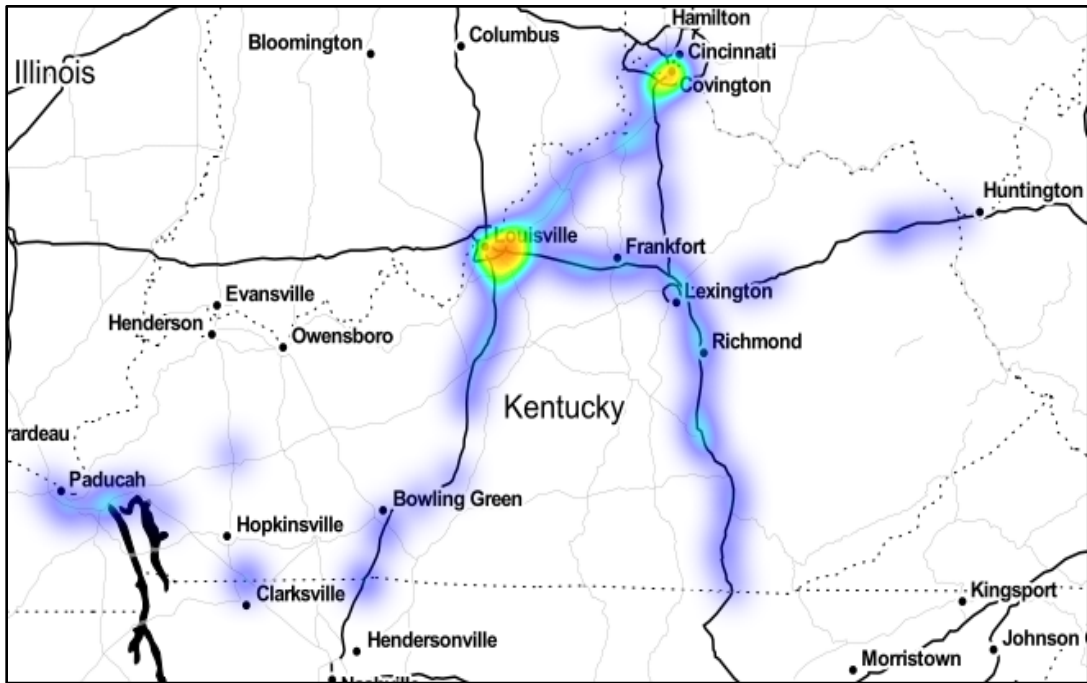


Figure 4.3 Statewide spatial distribution of congestion related crashes

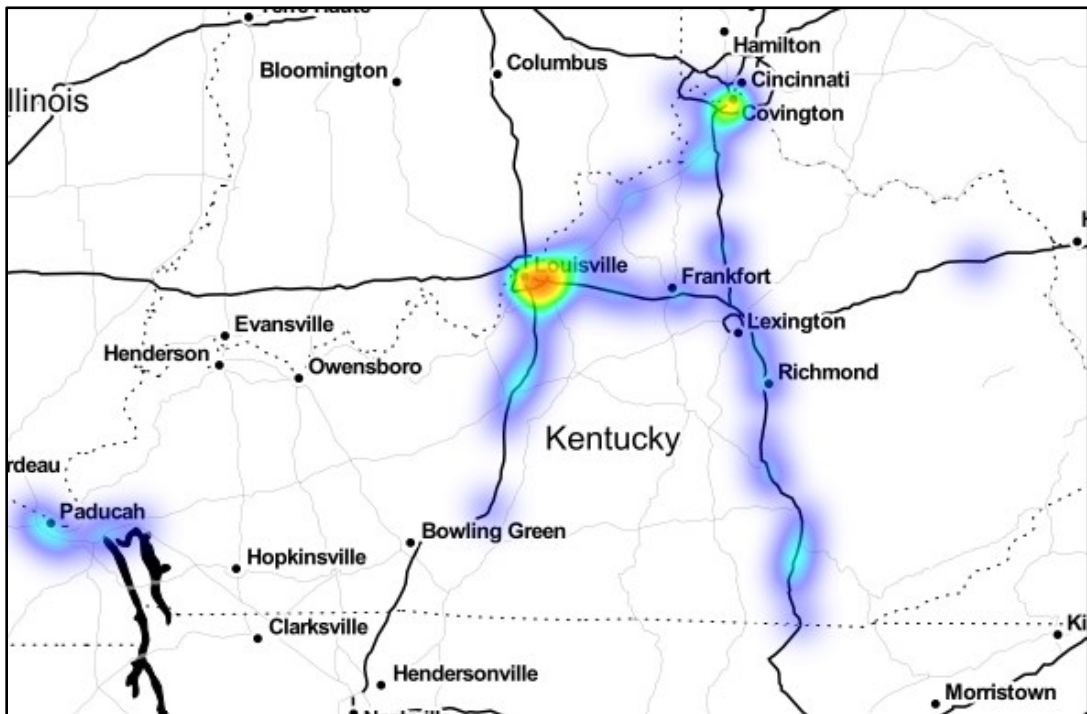


Figure 4.4 Statewide spatial distribution of "vehicle on shoulder" and congestion related crashes

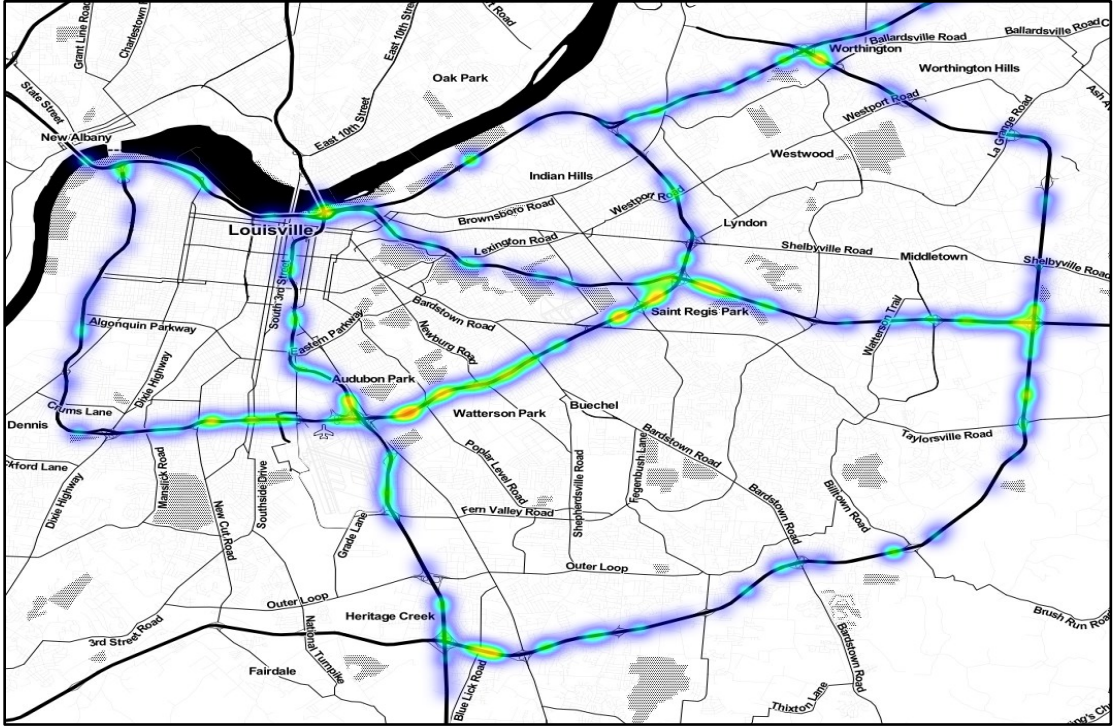


Figure 4.5 Spatial distribution of “vehicle on shoulder” related crashes in Louisville Metropolitan Area

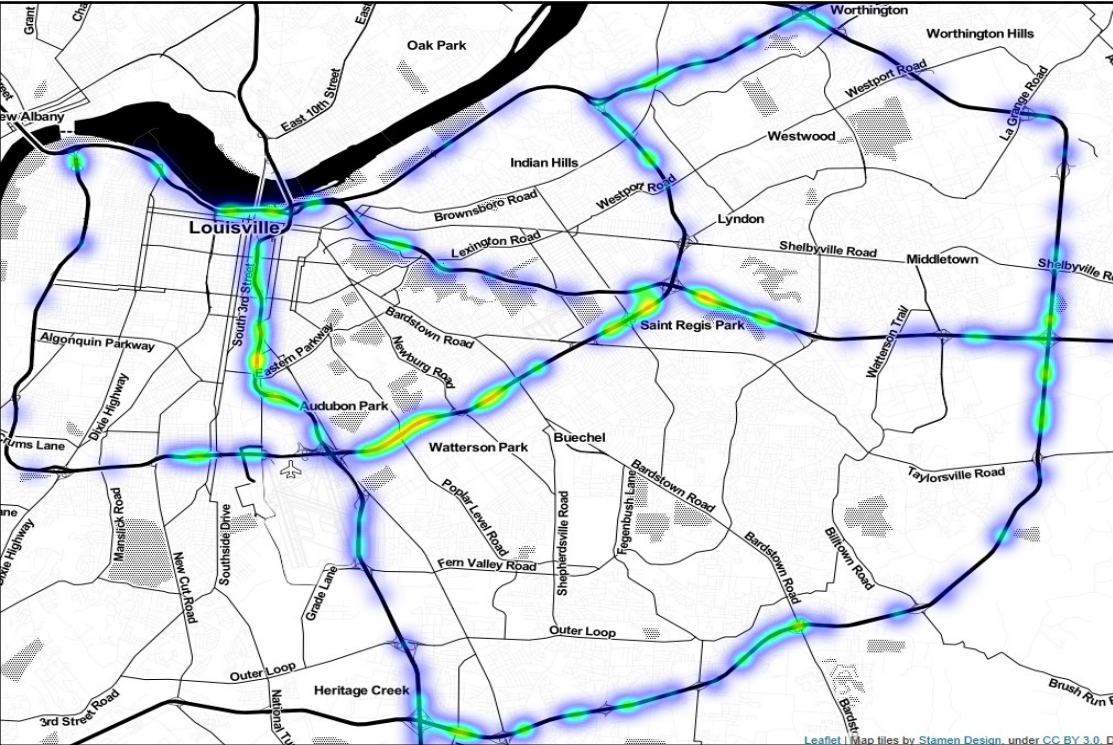


Figure 4.6 Spatial distribution of congestion related crashes in Louisville Metropolitan Area

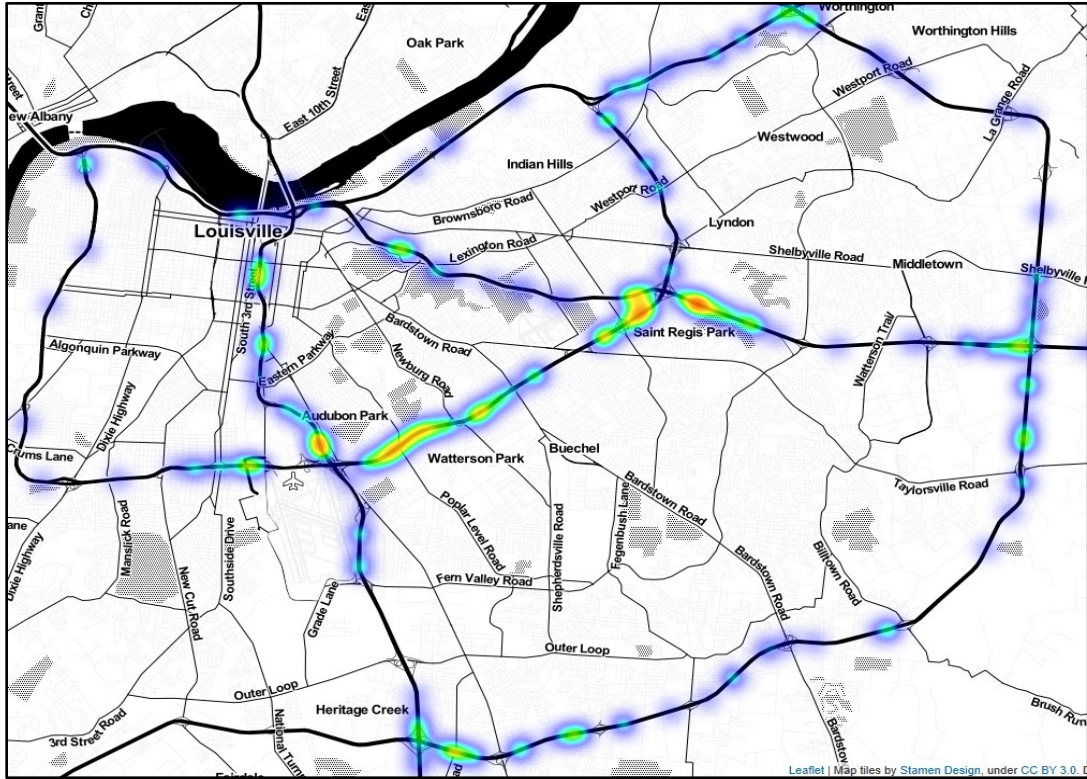


Figure 4.7 Spatial distribution of “vehicle on shoulder” and congestion related crashes in Louisville Metropolitan Area

4.2 Frequent Crash Contributory Factors

Based on the association rule mining procedure discussed in the previous chapter, among the set of crash contributory factors, the most frequently occurring crash contributory factors are presented in Table 4.1 in order of decreasing support (frequency of occurrence).

Table 4.1 Frequently occurring crash contributory factors

Factor	Support
Bad Visibility	37.49
Presence of vehicle on the shoulder	36.72
Bad Weather	32.29
Improper Maneuver	27.31
Driver Inattention	26.80
Congestion	25.67
Slippery Surface / Water Pooling	23.10
Grade Present	22.27
Curve Present	16.19
Driver Following Too Close	9.76
Presence of Animal/Debris	8.51
Driving too fast for conditions	7.91
Presence of vehicle stopped in road	7.38
Driver Impairment	5.94

From Table 4.1, The top five crash contributory factors in terms of support indicate that if a vehicle is present on the shoulder of the limited access highway at night or during inclement weather conditions, a crash is likely if the driver is inattentive or makes an inappropriate steering maneuver. The risk is compounded if a curve or grade is present. While the presence of a vehicle stopped in the travel lane showed up as a frequent item within the spatiotemporal vicinity of approximately 7.4% of crashes, the presence of an object in the roadway did not appear to be a major contributory factor to freeway crashes. It is however seen here that the presence of a vehicle on the shoulder prior to crash occurrence is a highly frequent. Since approximately 11% of freeway crashes involving a vehicle on the shoulder in Kentucky were fatal (Agent & Pigman, 1989) much attention

should be given to the removal of such hazards, particularly disabled vehicles on the road shoulders.

4.3 Association Rules

Association rules show the interdependence or correlation between the crash contributory factors. They may be classified based on the number of items in the rules. In this study, the association rules are classified into two-item rules and three-item rules showing the correlations between the crash factors and crashes.

4.3.1 Correlation Between Individual Crash Factors and Crashes

The top two-item association rules, which show the correlation between individual crash factors, are presented in Table 4.2 sorted by lift in decreasing order.

Table 4.2 Two-item association rules indicating correlation between single factors

Antecedents	Consequents	Antecedent Support	Consequent Support	Support	Confidence	Lift
{'DrivingTooFast'}	{'SlipperySurf/WaterPool'}	0.079	0.231	0.051	0.646	2.796
{'BadWeather'}	{'DrivingTooFast'}	0.323	0.079	0.060	0.185	2.338
{'Congestion'}	{'FollowTooClose'}	0.257	0.098	0.052	0.203	2.081
{'Animal/Debris'}	{'BadVisibility'}	0.085	0.375	0.058	0.679	1.811
{'Congestion'}	{'Inattention'}	0.257	0.268	0.112	0.435	1.624
{'SlipperySurf/WaterPool'}	{'ImproperManeuver'}	0.231	0.273	0.086	0.373	1.366
{'ImproperManeuver'}	{'BadWeather'}	0.273	0.323	0.114	0.418	1.295
{'Vehicle on Shouder'}	{'Congestion'}	0.367	0.257	0.117	0.318	1.237
{'Vehicle on Shouder '}	{'Inattention'}	0.367	0.268	0.111	0.303	1.130

Antecedent support and consequent support in Table 4.2 refer to the proportion of crashes involving the antecedent and consequent, respectively. The lift measure, as explained in the previous chapter, is a measure of factor dependence or correlation whereas confidence measures the proportion of crashes involving the antecedent that also involve the consequent. The higher lift association rules from Table 4.2 indicate that human and environmental factors are highly correlated with crashes. The first rule in Table 4.2, for example, states that when a driver drives too fast on slippery road surfaces or in areas where water has pooled on the road, a crash is very likely.

4.3.1.1 Correlation Between Vehicles on Shoulder, Congestion and Crashes

As noted in the previous section, the two-item set of rules assists us in comprehending the relationships between individual contributing factors. While human and environmental factors both play a role in the occurrence of a crash, their interaction with the presence of either congestion or vehicles on the shoulder raises the risk of a crash. This is demonstrated in Table 4.2, where the interaction between congestion and inattentive driving and inadequate following distance accounts for slightly more than 11% and 5% of freeway crashes, respectively showing a high chance of crashes occurring involving driver inattention during congested periods. Additionally, Waze “vehicle on shoulder” alerts are correlated with congestion in crash occurrence. Inferring from the correlation, the presence of a vehicle on the freeway shoulder could negatively impact traffic flow leading to congestion and subsequently crashes. However, as noted by Chen et al. (2020), most crash narratives associated with congestion do not specify the reason for congestion. As such, while there may be crashes that fit this chain of events, it is difficult to tell how many there are.

4.3.1.2 Correlation Between Vehicles Stopped in the Road and Crashes

Though about 7.4% of crashes showed up as having a vehicle stopped in road alert within their spatiotemporal vicinity from the frequent crash factor set shown in Table 4.1, their interaction with other predominant factors were insignificant and as such none of the association rules generated had this factor. A possible explanation is that situations involving a vehicle stopped in road, particularly on limited access highways are rare and as such do not occur with as much frequency as other crash factors. In situations where they do occur, they may be moved over to the shoulder or towed away from the road.

4.3.2 Correlation Between Multiple Crash Factors and Crashes

The three-item item rules clarify the interaction between more than two factors, especially if they have a higher lift value, indicating a stronger correlation between the interaction of those factors and crashes. These three-item rules are presented in Table 4.3.

Table 4.3 Three-item association rules

Antecedents	Consequents	Antecedent Support	Consequent Support	Support	Confidence	Lift
{'BadWeather', 'CurvePresent'}	{'SlipperySurf/WaterPool'}	0.071	0.231	0.051	0.726	3.142
{'BadWeather'}	{"Vehicle on Shoulder ', 'SlipperySurf/WaterPool'}	0.323	0.070	0.064	0.200	2.833
{'BadWeather'}	{'ImproperManeuver', 'SlipperySurf/WaterPool'}	0.323	0.086	0.078	0.242	2.812
{'Congestion'}	{"Vehicle on Shoulder ', 'Inattention'}	0.257	0.111	0.052	0.201	1.809
{"Vehicle on Shoulder ', 'Congestion'}	{'Inattention'}	0.117	0.268	0.052	0.443	1.652

Though the presence of a vehicle on the shoulder and congestion increases crash risk and contributed to 11.7% of freeway crashes in Kentucky, from the three-item rules it is seen that when a driver is inattentive or loses concentration when there is the presence of these two factors, the risk of getting involved in a crash is compounded. In 44.3% of crashes where a vehicle on shoulder and congestion may have contributed to the crash, driver inattention was also a factor as depicted by the confidence measure which measures the percentage of crashes involving the antecedent that also involved the consequent. Also, while congestion was associated with 25.7% of crashes as depicted by its support measure, the confidence measure for the three-item rule with congestion as antecedent in Table 4.3 shows about 20.1% of crashes involving congestion also had vehicles on the shoulder and driver inattention as contributory factors. As such, it is seen that the presence of vehicles on the shoulder is correlated with crashes and correlated with traffic slowdowns.

4.4 Potential Additional Crash Coverage Provided by Waze

4.4.1 Waze Crash Alerts Linked to Crashes

Filtering out duplicate UUIDs, the resultant dataset contained 15,859 unique Waze crash reports made by different users that may still contain reports of the same incident made by different users. Down from 313,953 alerts. After clustering to filter out Waze reports made by different users referring to the same incident, there were 13,279 unique Waze crash report clusters in the data set. More Waze alerts are reported from urban areas than from rural areas and with more duplicate reports of the same incident by different users. After clustering, it is seen that slightly more unique Waze alerts are made from urban

areas. Table 4.4 shows the results of the data integration between crash data and Waze crash alert clusters.

Table 4.4 Waze crash alerts and crash integration

	Total	Waze crash alert cluster related	Not Waze crash alert cluster related
Interstate Crashes	5881	2304	3577

	Total	Crash related	Not related to Crash
Waze alert clusters	13279	2608	10671

Waze crash alerts captured 39.18% of all interstate crashes. The proportion of all unique Waze crash alert clusters that were related to a crash, based on the spatiotemporal thresholds set, was 19.64%. 376 crashes were earlier reported in Waze and 395 Waze crash alert clusters were earlier reports of crashes.

4.4.2 Temporal Patterns in Waze Crash Clusters and Crashes

A temporal analysis of Waze crash alert clusters and crashes was performed by time of day and weekday. Figure 4.8 depicts time of day analysis, whereas Figure 4.9 depicts weekday analysis. In Figure 4.9, the number “0” represents Monday and the number “6” represents Sunday.

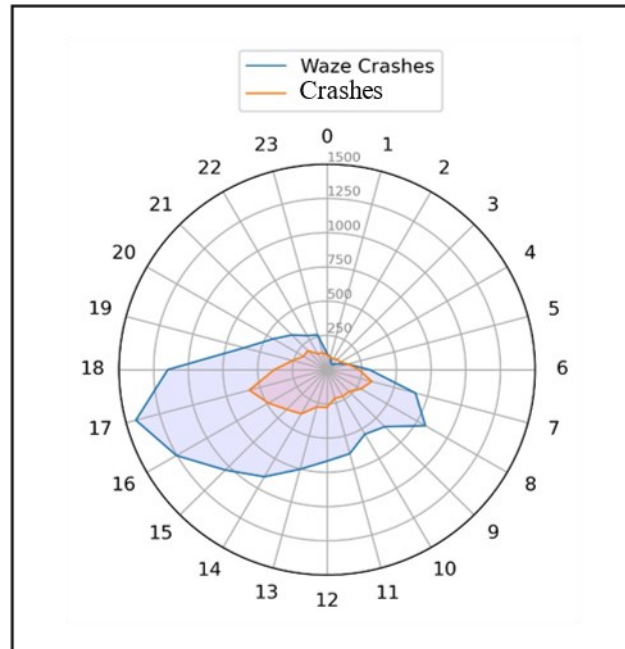


Figure 4.8 Time of day analysis

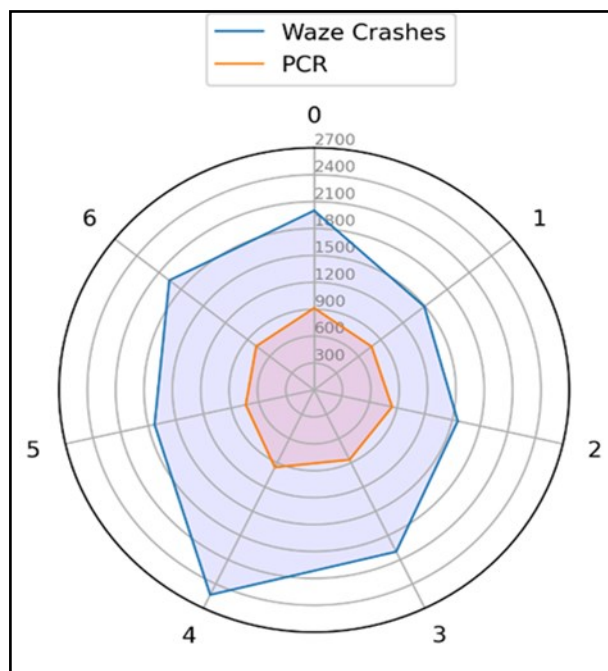


Figure 4.9 Day of week analysis

According to the analysis, the day of the week when the most crashes are recorded – Fridays – coincides with the day of the week when the most Waze crash reports are

received. The same is true for the hour of day analysis, as the hour of day with the highest crash numbers also has the highest Waze crash alerts. As such, Waze crashes have a similar temporal distribution as crashes.

4.4.3 Spatial Accuracy of Waze Crash Reports

Having linked Waze crash alerts to crashes, Figures 4.10 and 4.11 show the distribution of the distance between a crash and its corresponding Waze crash alert(s). Figure 4.10 represents the distribution of the distances between a crash report and their corresponding linked Waze crash alert(s) for crashes that were earlier reported in Waze.

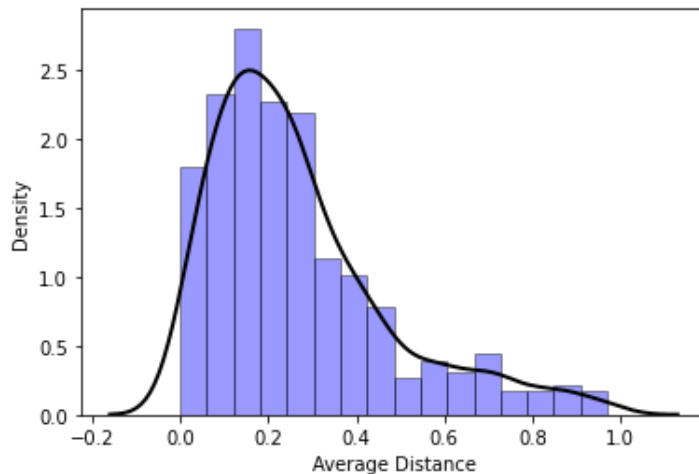


Figure 4.10 Distribution of distance of early reports in Waze from crash location

From Figure 4.10 it is seen that majority of the early reports of crashes in Waze are within 0.2 miles of their corresponding crash location.

Similarly, Figure 4.11 shows the distribution of distances between Waze crash alerts and their corresponding crashes for all Waze crash alerts linked to crashes.

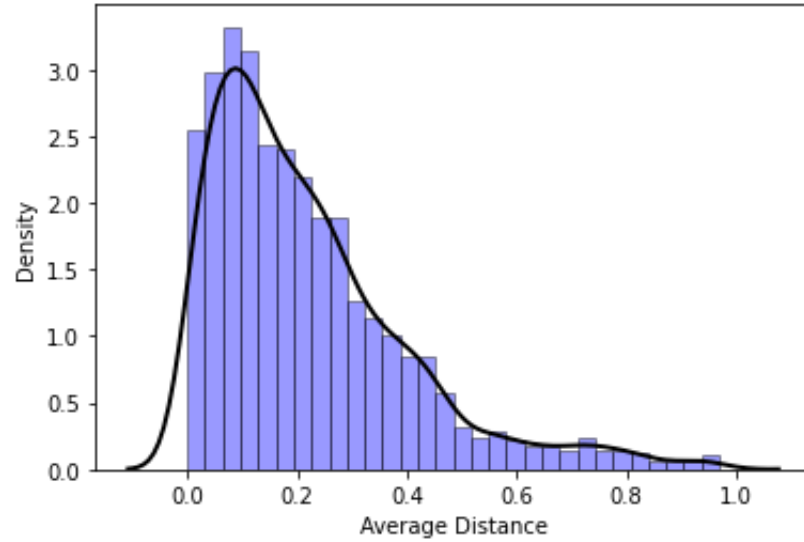


Figure 4.11 Distribution of distance of all Waze crash alerts from corresponding crash location

Figure 4.11 illustrates that the majority of Waze alerts associated with crashes were within 0.2 miles of the corresponding crash sites. The average distance from the crash site for early Waze reports of crashes was 0.26 miles. However, the average distance for all Waze crash alerts linked to crashes was 0.22 miles. As such consistent with Liu et al. (2019), the Waze crash alerts were reasonably spatially accurate and hence, Waze crash alerts could present an alternative for identifying crashes, particularly minor property damage only crashes that otherwise may go unreported.

Figure 4.12 shows the temporal distribution of Waze crash alerts for those Waze crash alerts that were linked to crashes.

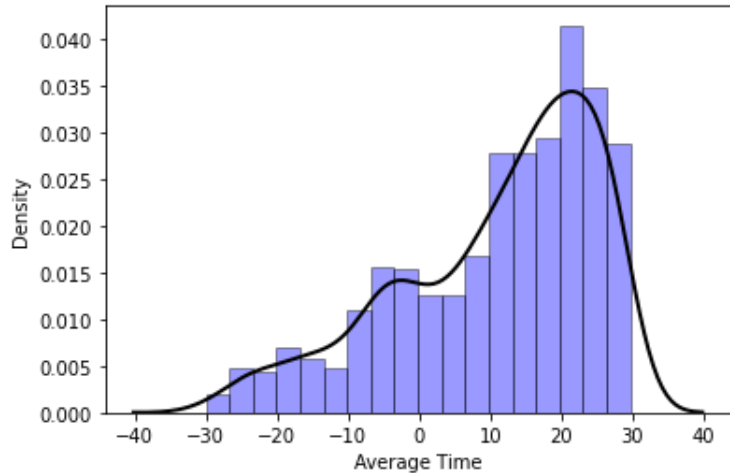


Figure 4.12 Temporal distribution of Waze crash alert report times for Waze crash alerts linked to crashes

From Figure 4.12 it is seen that there were earlier reports of crashes in Waze, some 30 minutes earlier before their corresponding crash record, which could be taken advantage of to reduce incident response and clearance times. Reducing incident response and clearance times would mean a reduction in the likelihood of secondary crashes thus improving safety on the road.

Aggregating the Waze crash incident data by county, the data is as presented in Table 4.5. The data only covers 44 counties out of the 47 counties with interstate highways in Kentucky.

Table 4.5 Waze crash alert clusters related and unrelated to crashes

COUNTY	Total Waze	Waze: Crash Related	Waze: Crash Unrelated
Barren	86	12	74
Bath	51	9	42
Boone	848	164	684
Boyd	25	8	17
Bullitt	522	108	414
Caldwell	16	2	14
Campbell	454	172	282
Carroll	218	39	179
Carter	172	40	132

Table 4.5 (Continued) Waze alert clusters related and unrelated to crashes

COUNTY	Total Waze	Waze: Crash Related	Waze: Crash Unrelated
Christian	222	37	185
Clark	79	10	69
Edmonson	27	1	26
Fayette	568	116	452
Franklin	138	31	107
Gallatin	165	25	140
Grant	222	38	184
Graves	1	0	1
Hardin	370	65	305
Hart	176	19	157
Henderson	10	0	10
Henry	145	19	126
Hopkins	68	1	67
Jefferson	3835	819	3016
Kenton	1501	358	1143
Larue	26	2	24
Laurel	355	50	305
Livingston	36	2	34
Lyon	84	7	77
Madison	377	69	308
Marshall	129	8	121
McCracken	154	10	144
Montgomery	47	10	37
Oldham	343	64	279
Rockcastle	399	64	335
Rowan	56	15	41
Scott	281	57	224
Shelby	331	72	259
Simpson	82	4	78
Trigg	29	3	26
Trimble	15	2	13
Warren	242	17	225
Webster	20	1	19
Whitley	294	51	243
Woodford	60	7	53

From Table 4.5, a similar spatial temporal pattern is observed as seen in the Waze “vehicle on shoulder” alerts. More Waze crash alerts are concentrated in the urban areas where they also match a significant proportion of crashes. However, within the urban areas, a majority of the alerts are not linked to any crash. While a few of these unmatched alerts may be false alarms consistent with literature, a significant proportion of them may be minor crashes that are not reported and hence the adoption of Waze as an alternate source of data, particularly in the urban areas where they are prevalent could help reduce underreporting of crashes.

CHAPTER 5. CONCLUSION

5.1 Summary

Using a spatiotemporal approach, this study aimed to quantitatively analyze the relationship between vehicles on the shoulder, traffic slowdowns, and crashes by integrating Waze alerts, GPS-based speed data, and crash data. Association rule mining was used to assess and quantify correlation. According to the analysis of limited access highways, 36% of crashes had a vehicle parked on the roadway shoulder within their spatiotemporal vicinity – 0.25 miles upstream and downstream of the crash site and 30 minutes prior to the crash occurrence. Also, approximately 25% of limited access highway crashes were associated with congestion. As such, there exists a high correlation between vehicles on the shoulder, congestion and crashes. Moreover, in 11.7% of the crashes, both a vehicle on shoulder and congestion were present immediately prior to crash occurrence corroborating the correlative relationship between these factors and crashes. The subsequent association rule mining analysis confirmed the association between vehicles on shoulder, congestion, and crashes was statistically significant. The level of significance ranked this relationship behind combinations of several important human and environmental factors, such as bad weather, slippery surface, driving too fast, following too closely, and executing an improper maneuver. While these human and environmental factors are inherently hard to remedy, incident management and operational strategies may be employed to alleviate congestion and vehicles on the shoulders of limited access highways to improve traffic safety on highways.

Also, this study sought to assess the correlation between Waze “vehicle stopped in road” alerts, Waze “object stopped in road” alerts and crashes. While “vehicle stopped in

road” alerts were present within the spatiotemporal vicinity of about 7.4% of crashes, it did not show up in the association rule mining process as significantly correlated with crashes or congestion. An analysis of the coverage of Waze shows only 39.18% crashes were reported in Waze whereas crashes found in Waze accounted for only 19.64% of all Waze crash alerts. Waze crash alerts were also found to be spatially accurate and as such could serve as an additional source of data for traffic safety related purposes.

5.2 Applications

The presence of vehicles on freeway shoulders for extended periods of time increase crash risk as has been shown in the study. Their interaction with other human and environmental factors compounds this risk. As such, Waze “vehicle on shoulder” alerts may be used as a tool to monitor freeway shoulders and consequently remove vehicles on freeway shoulders that remain there for extended periods of time.

Also, corroborating similar studies in other states, Waze crash incident alerts were found to be spatially accurate and hence can be used as a traffic monitoring source by traffic management centers to identify crashes thereby cutting down incidence response times and clearance times. Moreover, 376 out of the 5768 mainline interstate highway crashes, in the second half of 2018, were earlier reported in Waze.

5.3 Future Work

The data and analytical methods used in this indicate the potential of crowdsourced traffic data to offer much needed insights into the challenges posed by vehicles on the shoulder. While this study focused on Waze vehicle on shoulder alerts, vehicle in road

alerts, objects in road alerts and crash alerts, other hazard and jam alerts from Waze can provide additional context and therefore should be included in future analysis.

REFERENCES

- AASHTO. (2011). *A Policy on Geometric Design of Highways and Streets*. The American Association of State Highway and Transportation Officials.
- Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, 32(5), 633–642. [https://doi.org/10.1016/S0001-4575\(99\)00094-9](https://doi.org/10.1016/S0001-4575(99)00094-9)
- Agent, K. R., & Pigman, J. G. (1989). *Accident Involving Vehicles Parked on Shoulders of Limited Highways*. College of Engineering, University of Kentucky, Lexington.
- Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207–216.
- Amin-Naseri, M. (2018). *Adopting and incorporating crowdsourced traffic data in advanced transportation management systems*.
- Amin-Naseri, M., Chakraborty, P., Sharma, A., Gilbert, S. B., & Hong, M. (2018). Evaluating the Reliability, Coverage, and Added Value of Crowdsourced Traffic Incident Reports from Waze. *Transportation Research Record: Journal of the Transportation Research Board*, 2672(43), 34–43. <https://doi.org/10.1177/0361198118790619>
- Chen, M., Zhang, X., & Blandford, B. (2020). *Vehicle on Shoulder and Crash—Correlation or Causation?* (KTC-20-25/SPR20-598-1F; p. 32).
- Chimba, D., & Kutela, B. (2014). Scanning secondary derived crashes from disabled and abandoned vehicle incidents on uninterrupted flow highways. *Journal of Safety Research*, 8. <https://doi.org/10.1016/j.jsr.2014.05.004>
- Choueiri, E. M., Lamm, R., Kloeckner, J. H., & Mailaender, T. (1994). Safety aspects of individual design elements and their interactions on two-lane highways: International perspective. *Transportation Research Record: Journal of the Transportation Research Board*, 1445, 34–46.
- Das, S., Sun, X., Goel, S., Sun, M., Rahman, A., & Dutta, A. (2020). Flooding related traffic crashes: Findings from association rules. *Journal of Transportation Safety & Security*, 1–19. <https://doi.org/10.1080/19439962.2020.1734130>
- dos Santos, S. R., Junior, C. A. D., & Smarzaro, R. (2017). Analyzing-Traffic-Accidents-based-on-the-Integration-of-Official-and-Crowdsourced-Data.pdf. *Journal of Information and Data Management*, 8(1), 67–82.
- Eriksson, I. (2019). *Towards Integrating Crowdsourced and Official Traffic Data*. Uppsala Universitet.
- Flynn, D. F. B., Gilmore, M. M., & Sudderth, E. A. (2018). *Estimating Traffic Crash Counts Using Crowdsourced Data* (DOT-VNTSC-BTS-19-01; p. 48).
- Geurts, K., Wets, G., Brijs, T., & Vanhoof, K. (2003). Profiling of High-Frequency Accident Locations by Use of Association Rules. *Transportation Research Record: Journal of the Transportation Research Board*, 1840(1), 123–130. <https://doi.org/10.3141/1840-14>
- Goodall, N. J. (2017). Probability of Secondary Crash Occurrence on Freeways with the Use of Private-Sector Speed Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2635(1), 11–18. <https://doi.org/10.3141/2635-02>

- Goodall, N., & Lee, E. (2019). Comparison of Waze crash and disabled vehicle records with video ground truth. *Transportation Research Interdisciplinary Perspectives*, 1, 100019. <https://doi.org/10.1016/j.trip.2019.100019>
- Gross, F., & Donnell, E. T. (2011). Case-control and cross-sectional methods for estimating crash modification factors: Comparisons from roadway lighting and lane and shoulder width safety effect studies. *Journal of Safety Research*, 42(2), 117–129. <https://doi.org/10.1016/j.jsr.2011.03.003>
- Han, J., Pei, J., & Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation. *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 1–12. <https://doi.org/10.1145/342009.335372>
- Hauer, E. (2000). *Shoulder Width, Shoulder Paving and Safety*.
- Head, J. A. (1959). Predicting traffic accidents from roadway elements on urban extensions of state highways. *Highw. ^[SEP]Res. Board Bull.*, 208, 45–63.
- Jia, C., Li, Q., Oppong, S., Ni, D., Collura, J., & Shuldiner, P. W. (2013). *Evaluation of alternative technologies to estimate travel time on rural interstates*. Transportation Research Board 92nd Annual Meeting.
- Kraus, J. F., Anderson, C. L., Arzemanian, S., Salatka, M., Hemyari, P., & Sun, G. (1993). Epidemiological aspects of fatal and severe injury urban freeway crashes. *Accident Analysis & Prevention*, 25(3), 229–239. [https://doi.org/10.1016/0001-4575\(93\)90018-R](https://doi.org/10.1016/0001-4575(93)90018-R)
- Kumar, S., & Toshniwal, D. (2016). A data mining approach to characterize road accident locations. *Journal of Modern Transportation*, 24(1), 62–72. <https://doi.org/10.1007/s40534-016-0095-5>
- Labi, S. (2006). *Effects of Geometric Characteristics of Rural Two-Lane Roads on Safety* (FHWA/IN/JTRP-2005/02, 2664; p. FHWA/IN/JTRP-2005/02, 2664). Purdue University. <https://doi.org/10.5703/1288284313361>
- Labi, S. (2011). Efficacies of roadway safety improvements across functional subclasses of rural two-lane highways. *Journal of Safety Research*, 42(4), 231–239. <https://doi.org/10.1016/j.jsr.2011.01.008>
- Lee, C., Saccomanno, F., & Hellinga, B. (2002). Analysis of Crash Precursors on Instrumented Freeways. *Transportation Research Record: Journal of the Transportation Research Board*, 1784(1), 1–8. <https://doi.org/10.3141/1784-01>
- Lenkei, Z. (2018). *Crowdsourced traffic information in traffic management: Evaluation of traffic information from Waze*.
- Liu, Y., Hoseinzadeh, N., Han, L. D., & Brakewood, C. (2019). *Evaluation of Crowdsourced Event Reports for Real Time IMplementation—Spatial and Temporal Accuracy Analyses*.
- Mecheri, S., Rosey, F., & Lobjois, R. (2017). The effects of lane width, shoulder width, and road cross-sectional reallocation on drivers' behavioral adaptations. *Accident Analysis & Prevention*, 104, 65–73. <https://doi.org/10.1016/j.aap.2017.04.019>
- Pack, M., & Ivanov, N. (2017). Are You Gonna Go My WAZE? *Institute of Transportation Engineers. ITE Journal*, 87(2), 28.
- Persaud, B., & Dzbik, L. (1993). Accident Prediction Models for Freeways. *Transportation Research Record*, 1401, 55–60.

- Raff, M. S. (1953). Interstate highway—Accident study. *Highw. Res. Board Bull.*, 74, 18–45.
- Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python’s scientific computing stack. *The Journal of Open Source Software*, 3(24). <https://doi.org/10.21105/joss.00638>
- Schoppert, D. W. (1957). Predicting traffic accidents from roadway elements of rural two-lane highways with gravel shoulders. *Highw. Res. Board Bull.*, 158(4–26).
- Shankar, V., Milton, J., & Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. *Accident Analysis & Prevention*, 29(6), 829–837. [https://doi.org/10.1016/S0001-4575\(97\)00052-3](https://doi.org/10.1016/S0001-4575(97)00052-3)
- Shefer, D. (1994). Congestion, air-pollution, and road fatalities in urban areas. *Accid. Anal. Prev.*, 26, 501–509.
- Stamatiadis, N., National Research Council (U.S.), Transportation Research Board, National Cooperative Highway Research Program, American Association of State Highway and Transportation Officials, United States, & Federal Highway Administration. (2009). *Impact of shoulder width and median width on safety*. Transportation Research Board.
- Transportation Research Board. (2016). *Highway Capacity Manual 6th Edition: A Guide for Multimodal Mobility Analysis*. The National Academies Press. <https://doi.org/10.17226/24798>
- Turner, S., Martin, M., Griffin, G., Le, M., Das, S., Wang, R., Dadashova, B., & Li, X. (2020). *Exploring Crowdsourced Monitoring Data for Safety* (TTI-Student-05). Texas A&M Transportation Institute. https://www.vtti.vt.edu/utc/safe-d/wp-content/uploads/2020/04/TTI-Student-05_Final-Research-Report_Final.pdf
- Veh, A. (1937). Improvements to Reduce Traffic Accidents. *Meeting of the Highway Division: New York, NY, USA, 1775-1785*.
- Weng, J., Zhu, J.-Z., Yan, X., & Liu, Z. (2016). Investigation of work zone crash casualty patterns using association rules. *Accident Analysis & Prevention*, 92, 43–52. <https://doi.org/10.1016/j.aap.2016.03.017>
- Woo, J. C. (1957). Correlation of Accident Rates and Roadway Factors. *Purdue University: Lafayette, Indiana*, 2326–6325.
- Yoon, J., Noble, B., & Liu, M. (2007). Surface street traffic estimation. *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services - MobiSys '07*, 220. <https://doi.org/10.1145/1247660.1247686>
- Young, S. D., Wang, W., & Chakravarthy, B. (2019). Crowdsourced Traffic Data as an Emerging Tool to Monitor Car Crashes. *JAMA Surgery*, 154(8), 777. <https://doi.org/10.1001/jamasurg.2019.1167>
- Zegeer, C. V., Deen, R. C., & Mayes, J. G. (1980). Effect of Lane and Shoulder Widths on Accident Reduction on Rural, Two-Lane Roads. *Transportation Research Record*, 11.
- Zhang, C., & Zhang, S. (2002). *Association rule mining: Models and algorithms*. Springer.

VITA

Eugene Boasiako Antwi

Education:

BSc. Civil Engineering – Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana

Professional positions held:

- Graduate Research Assistant, University of Kentucky
- Teaching and Research Assistant, KNUST
- Intern, J. Adom Construction

Professional publications:

Chen, M., Zhang, X., Blandford, B., & Boasiako, E. A. (2020). Vehicle on Shoulder and Crash—Correlation or Causation? (No. KTC-20-25/SPR20-598-1F).

Osei, J. B., Adom-Asamoah, M., Ahmed, A. A. A., & Antwi, E. B. (2018). Monte Carlo Based Seismic Hazard Model for Southern Ghana. *Civil Engineering Journal*, 4(7).