Adopting and incorporating crowdsourced traffic data in advanced transportation management systems

by

Mostafa Amin-Naseri

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Industrial and Manufacturing Systems Engineering

Program of Study Committee:
Stephen B. Gilbert, Co-Major Professor
Anuj Sharma, Co-Major Professor
Mingyi Hong, Co-Major Professor
Sigurdur Olafsson
Mack C. Shelley

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

DEDICATION

This dissertation is dedicated to my wife, Asiyeh, who has always been a constant source of support and encouragement during the challenges of life. This work is also dedicated to my parents, who have always loved me unconditionally and whose good examples have taught me to work hard for the things that I aspire to achieve.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

ATMS	Advanced Traffic Management System
CSTIR	Crowdsourced Traffic Incident Repost
ITS	Intelligent Transportation System
DBSCAN	Density Based Spatial Clustering Applications with Noise
HDBSCAN	Hierarchical DBSCAN
ST-DBSCAN	Space-Time DBSCAN

ACKNOWLEDGMENTS

I would like to thank my co-advisors Stephen Gilbert, Anuj Sharma, and Mingyi Hong and my committee members, Sigurdur Olafsson and Mack Shelley, for their guidance and support throughout the course of this research. In addition, I would also like to thank my friends, colleagues, the department faculty and staff for making my time at Iowa State University a wonderful experience.

I exalt God for the blessing of introducing Himself and His messengers to me, and for His undeserved watch and care on my entire life. I praise God's last vicegerent on earth, our Contemporary Imam, through whom God blesses human kind. Finally, I would like to thank Mostafa Sattarvandi, who gave me the taste of the kind blessings of the Ahlu- Bayt (Peace be upon Them) and changed the path of my life.

ABSTRACT

The widespread availability of internet and mobile devices has made crowdsourced reports a considerable source of information in many domains. Traffic managers, among others, have started using crowdsourced traffic incident reports (CSTIRs) to complement their existing sources of traffic monitoring. One of the prominent providers of CSTIRs is Waze. In this dissertation, first a quantitative analysis was conducted to evaluate Waze data in comparison to the existing sources of Iowa Department of Transportation. The potential added coverage that Waze can provide was also estimated.

Redundant CSTIRs of the same incident were found to be one of the main challenges of Waze and CSTIRs in general. To leverage the value of the redundant reports and address this challenge, a state-of-the-art cluster analysis was implemented to reduce the redundancies, while providing further information about the incident. The clustered CSTIRs indicate the area impacted by an incident and provide a basis for estimating the reliability of the cluster. Furthermore, the challenges with clustering CSTIRs were described and recommendations were made for parameter tuning and cluster validation.

Finally, an open-source software package was offered to implement the clustering method in near real-time. This software downloads and parses the raw data, implements clustering, tracks clusters, assigns a reliability score to clusters, and provides a RESTful API for information dissemination portals and web pages to use the data for multiple applications within the DOT and for the general public.

With emerging technologies such as connected vehicles and vehicle-to-infrastructure (V2I) communication, CSTIRs and similar type of data are expected to grow. The findings and recommendations in this work, although implemented on Waze data, will be beneficial to the analysis of these emerging sources of data.

CHAPTER 1. INTRODUCTION

The role of road transportation and its challenges

Road transportation constitutes the distinct majority of passenger traveled miles in the United States and an indispensable part of our daily lives. According to the American Bureau of Labor Statistics, on average each individual spends about 79.5 minutes on road travel in weekdays (Bureau of Transportation statistics, 2016). Moreover, transportation forms a considerable portion of personal and governmental expenditures. $1,184 billion of all personal expenditures were on transportation, making it the fourth largest category in personal expenditures (U.S. Department of Transportation, Bureau of Transportation Statistics, 2016). Similarly, in the freight sector, road transportation is the most common mode among all others. Both the value and weight of the total shipments through road transportation are considerably higher than all other modes and are projected to remain the highest to 2040 (Bureau of Transportation Statistics, 2015). The significant influence of road transportation in the economy and personal life indicates the broad impact of research in this field.

The challenges in the field of transportation are proportionate to its significance. Crashes are one of the major sources of fatality and are projected to remain in the top ten causes of death and injury worldwide, by 2030 (Mathers & Loncar, 2006). Plans for timely detection of crashes and responding to them, allocating first responders appropriately, and plans for reducing the risk of crashes are some of the life-saving research problems in this field.

In addition to crashes, traffic congestion is another significant issue in road transportation. Road congestions cause travel time delays and adversely impact the reliability

of transportation systems. In the urban areas in the United States, 6.9 billion hours of commuters' time was spent in congestions resulting in the waste of 3.1 billion gallons of fuel in year 2014 (U.S. Department of Transportation, Bureau of Transportation Statistics, 2016). These are some statistics that indicate the direct impact of traffic related issues. There are several indirect aspects to this as well. For instance, the environmental impact of the road transportation is significant. 27% of the greenhouse gas emissions (GHG) are produced by the transportation sector, from which passenger cars have the largest share ("United States Environmental Protection Agency," n.d.). Another example is that studies have shown that congestions lead to higher stress levels (Gyawali & Sharma, 2013; Hennessy & Wiesenthal, 1999), which in long term can cause heart disease (Steptoe & Kivimäki, 2012). With the extensive impact of congestion on several aspects of our public health, economy, environment, and general life quality, any improvements will benefit the public greatly.

Intelligent Transportation Systems (ITS)

In order to manage traffic incidents (crashes and congestion/jams), real-time information about traffic conditions is necessary. The technological advancements in telecommunication networks and computer science in the 20th century sparked the ideas of leveraging these technologies for broad situational awareness of traffic conditions on national highways. ITS Canada defines ITS as: "the application of advanced and emerging technologies (computers, sensors, control, communications, and electronic devices) in transportation to save lives, time, money, energy and the environment" (Ministry of Transportation Ontario, 2007). Based on this definition, improved mobility, safety, and productivity of the transportation system are the main goals of ITSs. These goals are in

alignment with the main challenges in road transportation, based on the statistics discussed in the prior section.

The US Department of Transportation (USDOT) has defined its four-year strategic plan for 2015-2019. The priorities are based on five main themes (Barbaresso et al., 2014):

- Enable Safer Vehicles and Roadways
- Enhance Mobility
- Limit Environmental Impacts
- Promote Innovation
- Support Transportation System Information Sharing

The strategic plan and the ITS goals are designed to serve several stakeholders. Drivers and passengers expect to experience safer transit and lower travel times. In addition, real-time information and reliable prediction of traffic patterns are highly desirable. For the businesses, reliable delivery time of their procurement as well safety are some of the main expectations. The themes and strategic plans for the traffic managers and ITS re derived based on mutual interest of the several stakeholders.

The goals of ITS cover a breadth of applications which contains several subsystems. Each subsystem is designed to help traffic managers with actionable decision in accomplishing their goals. One of the most effective subsystems of an ITS is an Advanced Traffic Management System (ATMS). The first computerized traffic signal management systems were implemented about 60 years ago (Ministry of Transportation Ontario, 2007). However, new technology and machine learning techniques have made new features possible. Some of the main objectives of ATMS are (Ministry of Transportation Ontario, 2007):

- Information collection: To be able to inform the passenger and first responders of the road conditions and potential hazards, it is crucial to broad, real-time, and reliable coverage of the traffic conditions. Data is collected on metrics such as traffic flow or volume, vehicle speed, traffic density, occupancy, incidents, and weather.
- Controlling traffic and highways: Leveraging the collected data and camera feeds, traffic managers decide on actions such as lane control signs or ramp metering on the highways, to improve the traffic flow. On the other hand, the historical data collection will be valuable in defining the design requirements of surface street control systems.
- Traffic information dissemination: This functionality directly serves the motorists by providing them real-time information about traffic conditions and advisory messages. This information provides motorists to select the best road based on the current traffic situations, either directly based on the information provided by the ATMS or through navigation applications that leverage the publicly available ATMS report. Furthermore, information about road closures and construction plans help travelers and fleet managers to plan for the upcoming events, accordingly. To inform the motorists on the road, roadside Dynamic Message Signs (DMS), are one of the common means of communicating. Apart from the information provided to traveler, the detected incidents are reported to appropriate response agency (e.g., police, ambulance, fire) with the necessary information to help with the incident. This feature is particularly valuable for rural or less congested areas, where there might not be a direct phone call to 911 to report incidents.

Some of the common sources of data at the Iowa DOT and at the DOTs across the U.S. include:

- Camera imagery: Camera images installed at multiple locations throughout the state provide real-time and visible understanding of traffic situations.
- Radio calls: Incidents reported to 911 or highway helpers are recorded to the ATMS.
- Radar: Using radio wave signals to detect presence, speed, and size of the vehicle.
- Speed and occupancy data acquired from third party providers, e.g. INRIX.

In addition, to the conventional sources that are run and maintained by the DOT, more recently crowdsourced data has become available. Detecting events from social media (e.g., Twitter) or from mobile phone applications both have been shown to be viable (D'Andrea, Ducange, Lazzerini, & Marcelloni, 2015; Gu, Qian, & Chen, 2016; Santos, Davis Jr., & Smarzaro, 2016a; Steiger, de Albuquerque, & Zipf, 2015; Van Dyke, Walton, & Ballinger, 2016). Some of the mobile phone navigation applications, provide crowdsourced traffic incident reports to traffic agencies. A review study on the main providers of crowdsourced traffic incident reports indicated that Waze, a mobile navigation application, is among the most popular applications among the users and provides most coverage (Van Dyke et al., 2016). Waze offers a partnership with cities and traffic agencies called Connected Citizens Program (CCP). This partnership is an information exchange where the agencies provide Waze with road closures and detected incidents, and in return, Waze shares their data with the agencies ("Waze Connected Citizens Program (CCP)," n.d.). The increasing number of state and city traffic agencies that have started using Waze (Pack & Ivanov, 2017) has raised an interest in understanding the characteristics of this source of data.

Crowdsourced data is available for free or at low cost, however, the quality, coverage and validity of these reports need to be studied.

Challenges of using crowdsourced data in the ATMS

Before using crowdsourced data for traffic operations and in the ATMS, the traffic managers must study the characteristics of this new source of data. Once the characteristics of these data are known, the raw crowdsourced data require significant preprocessing to fit the ATMS requirements. Some of the main challenges that traffic agencies are faced with regarding crowdsourced data, particularly from Waze, are discussed in this section.

Characteristics of crowdsourced data

The pros and cons of conventional traffic data sources (e.g., sensors, cameras, and probe data) has been studied by many researchers. However, there are very few works that have quantitatively compared the characteristics of crowdsourced data with these existing sources. Knowing the strength and weaknesses as well as the potential additional coverage that crowdsourced data would provide to the existing sources is critical for interpreting the derived findings. This challenge is to determine the coverage and value of crowdsourced reports in general.

Dealing with redundant reports and reliability of reports

Crowdsourced data is notorious for redundant reports (Gu et al., 2016). The traffic operations managers are already loaded with incident reports from multiple sources. Overwhelming the operators with redundant reports of the same incident, not only makes the operators prone to error, it is also detrimental to their trust of the system (Dixon, Wickens, & McCarley, 2007; Madhavan, Wiegmann, & Lacson, 2006a, 2006b). On the other hand, the

multiple reports of a single incident provide valuable information about the intensity and reliability of the reported incident. Moreover, it provides information about the impacted area by the incident. Therefore, a real-time solution to handle redundant reports is while leveraging the redundant reports for further information is one of the most critical challenges of crowdsourced data.

Once the potential value is crowdsourced data is generally accepted and the redundant reports are handled, the critical decision for the traffic managers is whether the incident report they have at hand is reliable. Various sources provide information about incidents on the road. The challenge is to find ways to leverage the existing sources to provide traffic managers and ATMS administrators with an estimate the validity of the crowdsourced reports. This challenge has been recognized by researchers in the field (Amin-Naseri, Chakraborty, Sharma, Gilbert, & Hong, 2018; Pack & Ivanov, 2017; Van Dyke et al., 2016).

Fusing this new feed with existing sources

For each source of data, diverse types of instruments are used for data collection. Data coming from several sources, each with their own format, accuracy, and limitations, poses additional challenges to the ITS administrators to match these diverse sources. Data fusion in the ATMS application is one of the active fields of research. The attempt is to match the sources of data together, to provide a broader perspective of the traffic conditions and improve the reliability in the detected incidents.

Research topics

To this point, some of the main challenges and active fields of research have been discussed. In this section the three research topics with regards to these challenges are introduced, each composing a chapter of this dissertation.

Chapter 2 - Evaluating the Reliability, Coverage, and Added Value of Crowdsourced Traffic Incident Reports from Waze

To study the characteristics of Waze data, this work evaluated the quality of crowdsourced incident reports from Waze and compared Waze as a data source with the existing sources of data in the Iowa ATMS. The findings described the reliability of Waze reports and the additional coverage that it can provide to the ATMS. Moreover, the factors that impact the Waze coverage such as time and location were investigated.

Chapter 3 – Online clustering of Waze reports and reliability estimation

The findings in Chapter 2 confirmed the issue of redundant reports in Waze data. This chapter applied a state of the art near real-time spatiotemporal clustering technique to efficiently group redundant Waze reports. In addition to addressing redundancies, this grouping provided insight about the impact area of the incident, as well as the duration of the impact. Finally, a reliability score was assigned to each cluster of crowdsourced reports to assist ATMS administrators in prioritizing their operations. Cluster validation is known as the most difficult step in cluster analysis. To this end, this work offers customized suggestions and metrics for evaluating the quality of clusters using the conventionally sourced data in the ATMSs.

Chapter 4 – WazeClustR: An R-based program to enhance Waze crowdsourced traffic reports for traffic management applications

This chapter provides a software package that allows traffic agencies to implement the clustering and reliability estimation described in Chapter 3. To mitigate the challenge of data fusion from multiple sources, the enhanced Waze feed is presented in a format that suits

the requirements of ATMSs. This chapter describes the software architecture as well as the functions for estimating reliability of clusters over time.

Chapter 5 – Discussion and conclusion

This work targeted three of the main challenges facing ATMSs, each with state of the art data used in the Iowa Department of Transportation (IDOT). This research quantifies the value in one of the major sources of crowdsourced incident reports and implements a customized unsupervised learning method to clean the data. Moreover, methods for validating the clusters and tuning the parameters were offered. Finally, an open-sourced software package is presented to implement the proposed clustering method on Waze crowdsourced data. Although this work was particularly applied to Waze data for the state of Iowa, the methods as well as some of the general findings are applicable to many other sources of data and locations. Regardless of the actual source of data, validation and fusion techniques implemented in this research are generalizable to other spatiotemporal data sets.

CHAPTER 2. EVALUATING THE RELIABILITY, COVERAGE, AND ADDED VALUE OF CROWDSOURCED TRAFFIC INCIDENT REPORTS FROM WAZE

In press at the *Journal of Transportation Research Record*

Authors: Mostafa Amin-Naseri, Pranamesh Chakraborty, Anuj Sharma, Stephen Gilbert, Mingyi Hong

This work was performed and described by Mostafa Amin-Naseri with guidance by Anuj Sharma, Stephen Gilbert, and Mingyi Hong. Co-author Chakraborty provided the INRIX incident detection data and drafted the INRIX incident detection section.

Abstract

Traffic managers strive to have the most accurate information on road conditions, normally by using sensors and cameras, to act effectively in response to incidents. The prevalence of crowdsourced traffic information that has become available to traffic managers brings hope and yet raises important questions about the proper strategy for allocating resources to monitoring methods. Although many researches have indicated the potential value in crowdsourced data, it is crucial to quantitatively explore its validity and coverage as a new source of data. This research studied crowdsourced data from a smartphone navigation application called Waze to identify the characteristics of this social sensor and provide a comparison with some of the common sources of data in traffic management. Moreover, this work quantifies the potential additional coverage that Waze can provide to existing sources of the Advanced Traffic Management System (ATMS). One year of Waze data was compared with the recorded incidents in the Iowa's ATMS in the same timeframe. Overall, the findings indicated that the crowdsourced data stream from Waze is an invaluable source of information for traffic monitoring with broad coverage (covering 43.2% of ATMS crash

and congestion reports), timely reporting (on average 9.8 minutes earlier than a probe-based alternative), and reasonable geographic accuracy. Waze reports currently make significant contributions to incident detection and were found to have potential for further complementing the ATMS coverage of traffic conditions. In addition to these findings, the crowdsourced data evaluation procedure in this work provides researchers with a flexible framework for data evaluation.

Introduction

Traffic managers aim for increased mobility and safety on the roads. Real-time information on road conditions is necessary for taking proper actions. However, relying on the sensors and cameras for monitoring traffic conditions at all locations and times is neither possible nor economically justifiable (Yoon, Noble, & Liu, 2007). Moreover, many sensors detect incidents based on speed changes, while in less populated areas, a crash may present a high-risk zone for secondary crashes without an immediate significant speed drop. These circumstances point to the insufficiency of the existing means for full road condition monitoring.

Recent research has demonstrated the potential value in leveraging social media to detect traffic incidents (D'Andrea et al., 2015; Gu et al., 2016; R. Li, Lei, Khadiwala, & Chang, 2012; Seeger, Lillehoj, Wilson, & Jensen, 2014). Thus, crowdsourced data, have recently gained attention in traffic management. To this end, many cities and departments of transportation (DOTs) have incorporated data from a crowdsourced smartphone application called Waze into their ATMS. Using crowdsourced data, however, poses several questions to the traffic managers. In this research, a quantitative analysis is implemented to provide data-

driven answers to some of the common concerns of traffic managers with regards to Waze data.

Iowa Department of Transportation (IDOT) has used Waze data as a source of incident detection since September 2015. One year of data (2016) was used to address questions in three primary areas.

- How does Waze compare to existing sources?
    - Are Waze reports reliable?
    - What percentage of the current recorded incidents were detected by Waze?
    - How does Waze compare to other common sources of data collection in the ATMS?
- What are the characteristics of Waze data?
    - How does Waze coverage compare to other sources?
    - How does Waze coverage vary by time and location?
- What is the estimated potential additional coverage that Waze can provide to the ATMS?
    - In the locations where ATMS is unable to verify Waze reports, can Waze be trusted?

This last question is a critical topic. In current ATMS settings, crowdsourced data needs validation by a second source before being trusted. This is not available in all locations and times, however. Thus, an estimation of the potential added coverage in Waze provides a ground for justifying allocating resources to developing methods that assess crowdsourced reports using historical data. One of the ultimate goals of studying crowdsourced data is to understand its characteristics profoundly enough to know when and where to rely on

crowdsourced reports in locations where there are no other means for validation. Hence, this work seeks answers to the above questions in the process of finding the response to Question c. Moreover, some of the main challenges in utilizing Waze data for traffic monitoring were identified and discussed for future work.

Background

Crowdsourced data and social media have been widely used in many areas. For instance, tweets have been used to detect earthquakes in real-time (Sakaki, Okazaki, & Matsuo, 2010) or predict influenza outbreaks (Aramaki, Maskawa, & Morita, 2011; Signorini, Segre, & Polgreen, 2011). More closely related to traffic, the Twitter-based Event detection and Analysis System (TEDAS) has been proposed by Li and colleagues (R. Li et al., 2012). Another work utilized Twitter to detect traffic incidents in real time (D'Andrea et al., 2015). To increase the percentage of useful tweets, Gu et al. have implemented a method to extract geolocation from the text of traffic-related tweets (Gu et al., 2016). Furthermore, the validity of the traffic information acquired from social media was approved by comparing to the recorded traffic situation in London (Steiger, Resch, de Albuquerque, & Zipf, 2016a). These applications demonstrate the potential wealth of information in crowdsourced data. Regardless of how the data are collected, however, there are challenges in using crowdsourced data that require consideration.

Although crowdsourced data usually come at a relatively inexpensive price, there are challenges in understanding and interpreting this type of data. The crowdsourced data are reported by users who might be slightly inaccurate in time or location. For users traveling on the roads at the speed of 60 miles per hour, 30 seconds' delay in reporting an incident is a 0.5-mile distance. Moreover, users might falsely assume the causes of irregular congestion and report a crash while simply stuck in traffic.

Creating a clean dataset by reconciling the variation in crowdsourced user reports of the same incident and matching these reports to incidents recorded in the ATMS represents one of the primary challenges. The matching procedures as explored in the literature are known as matching or conflation methods (Goh et al., 2012; Ruiz, Ariza, Ureña, & Blázquez, 2011; Sester, Arsanjani, Klammer, Burghardt, & Haunert, 2014; Xavier, Ariza-López, & Ureña-Cámara, 2016; Yang & Zhang, 2015). As summarized Xavier et al. (Xavier et al., 2016), similarity measures for point data (like the incident data in this study) are generally a combination of the following:

Geometric: Distance or area overlap

Semantic: Measures of non-geometric properties.

Context: the special relationship between objects.

Ruiz et al. added the temporal criteria into their categories as well. For point matching, using geographic distance (Euclidian distance is most common) is the most classic approach (Beeri, Doytsher, Kanza, Safra, & Sagiv, 2005; Beeri, Kanza, Safra, & Sagiv, 2004; Safra, Kanza, Sagiv, Beeri, & Doytsher, 2010). Adding extra information about the points when available, such as road names and direction, adds additional power to the matching function. The hybrid approach of geographic and semantic information has shown high accuracy in matching crowdsourced information (McKenzie, Janowicz, & Adams, 2014). Considering the problem at hand and the available data in this research, a hybrid approach was used to leverage geographic as well as semantic matching methods.

Data

Waze Data

Waze is a navigation application that leverages crowdsourced user reports for providing service. Users can report traffic crashes, congestion, hazards, or police traps on the

road (www.waze.com/about). The Iowa Department of Transportation (IDOT) joined the Connected Citizen Program (CCP), which is an agreement in which the city or state managers provide Waze with information on road closures and constructions and, in return, Waze provides user reports to the managers. However, since the raw Waze data contain duplicate reports for a single incident and all reports may not have high reliability, data preprocessing is necessary (Pack & Ivanov, 2017). IDOT's ATMS implements stringent acceptance criteria for Waze reports before considering them for validation (filtering criteria: type = crash or reliability>=6 or report rating >=4). The reports that meet the criteria are sent to ATMS operators to verify the incident. If the incident is verified, it will be recorded in the ATMS database.

ATMS Data

The Iowa ATMS records all incidents, hazards, and congestion detected by various sensors and cameras or the reports by the highway helpers or police. The incidents in this dataset are validated by ATMS operators and thus serve as a reference for evaluating other sources of data. However, not all incidents, particularly congestion, are recorded in this dataset.

Incidents Detected from Third-Party Traffic Services Vendors

Third-party traffic services vendors such as INRIX (www.inrix.com) gather anonymized position data, which in turn provide rural and urban system-wide traffic data with reasonable accuracy (Haghani, Hamedi, & Sadabadi, 2009). Iowa ATMS applies a state of the art method for detecting incidents from INRIX data. This method utilizes interquartile range (IQR) of the historical speed data in each timeframe to detect outliers as described by Chakraborty and colleagues (Chakraborty, Hess, Sharma, & Knickerbocker, 2017). Threshold speeds are computed for each segment, day of the week, and 15-minute period of

the day utilizing the last 8 weeks of data. More specifically, threshold = (Median - 2 × IQR) is computed for each period and an incident alarm is triggered when the real-time speed is below the corresponding threshold. The data generated from this process are another feed of data to the ATMS and a basis for comparison with Waze data.

Traffic Camera Images

Cameras mounted in various locations across Iowa are one of the main means for traffic monitoring in the ATMS. To estimate false alarms in Waze reports, this study uses screenshots of the camera video feed that are captured every five minutes. Cameras in the Des Moines, Iowa metropolitan area (56 cameras) were selected for manual labelling of road conditions. Since labeling the road conditions (particularly congestion) based on a single image is a subjective decision, the images were labelled "clear" when the road was obviously clear and no congestion or incidents were observed. The labelled road images were used to detect the false alarms in Waze reports.

Anticipated Coverage of Data Sources

In practice, each of these sources cover a portion of the true incidents; they have some overlaps, and may have false alarms as well. The Venn diagram of our data sources depicted in Figure 1 illustrates this relationship (circles are not drawn to scale); the characteristics of the overlapping areas are of primary interest. Iowa ATMS captures a subset of the true incidents which is validated and free from false alarms. Waze and INRIX are expected to cover some of the true incidents while having a portion of false alarms. This study is mainly focused on estimating the potential additional contribution of Waze to the ATMS (region D). It is worth noting that the exact findings of this work are applicable to states and locations like Iowa, and that depending on the number of Waze users and penetration rates, the results may vary.

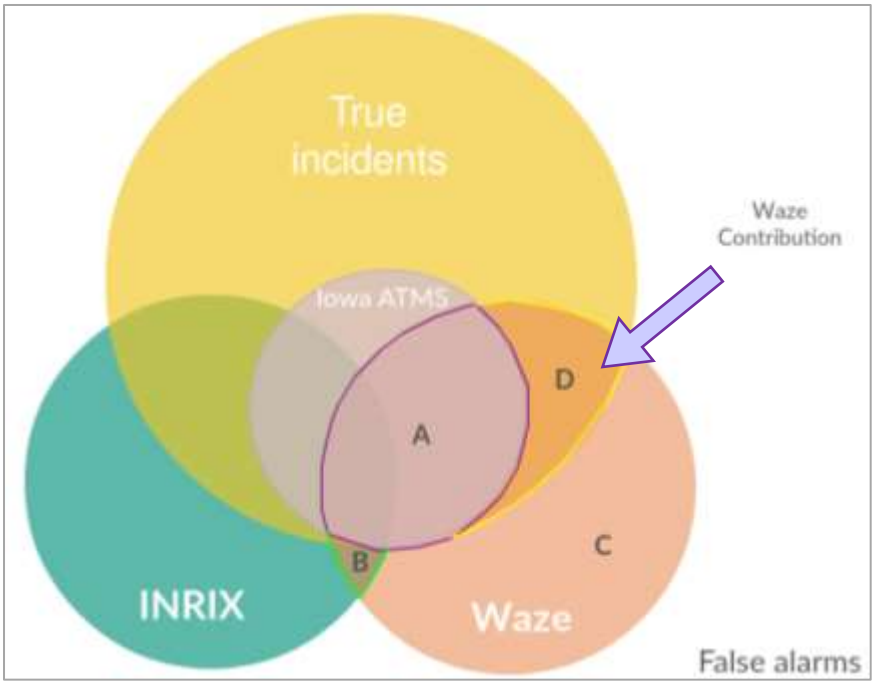


Figure 1 - Venn diagram of the sources of traffic monitoring data, pointing to region of interest (D), the potential contribution of Waze.

Evaluation Procedure

Region (D) on the Venn diagram of our data sources (Figure 1) marks the potential contribution of Waze to the ATMS. However, since data on true incidents in all locations and times are not available, the existing sources were used to quantify the potential contribution and value in Waze feed. Hence, the estimation of (D) was achieved in four main steps which are explained in this section using notations from Figure 1. The four steps are:

1. Match Waze and ATMS incidents (A)
2. Match Waze and INRIX incidents (B)
3. Estimate the false alarms (C)
4. Estimate Waze's contribution D = Waze – (A ∪ B ∪ C)

This study focused on two main type of incidents, congestion and crashes, as the sources that most directly impact traffic. To accomplish these steps, a matching function was necessary, which is described below.

Table 1 - Event Matching Procedure for Step 1 (ATMS and Waze matching)

Matching Levels	Criterion	Logic	Matching method	Action category
First	Time	Waze reports 20 minutes before the start and after the end time of an ATMS record	Temporal	Preprocessing
Second	Location	Crashes in a 2.5-mile radius, Congestion in 1-mile radius	Geographic	Preprocessing
Third	Road name and direction	Grouped into: Matching both and opposite direction	Semantic	Preprocessing
Fourth	Type of incident	Type, road name, and direction match	Semantic	Full/exact Match
	Type of incident	ATMS event is a crash, Jam reported in Waze, No full match exists	Semantic	Secondary Jam of a crash
	Road direction	Everything matches, Opposite direction, 1-mile radius	Semantic	Opposite direction

Matching Function

For matching incidents between sources, a hybrid method leveraging geographic and semantic matching methods was implemented. In both data sources, the road name and direction, as well as the type of the incident (i.e., crash, congestion, or stalled vehicle) were recorded. Table 1 presents the levels of the matching function as well as the criteria and method used in each level. The matching function first selects incidents in the temporal vicinity, then the geographic distance is examined. From spatiotemporal neighboring incidents, semantic information such as road names, direction, and type of the incident were used to mark matching incidents. The matching function introduced for this step (Match Waze and ATMS incidents) is the most comprehensive one. In the next steps, when matching with INRIX data and detecting false alarms, the match function was slightly modified to fit the semantic features of the respective data fields. Table 2 provides a summary of the evaluation procedure in this work and the data used in each step.

Table 2 -Summary of the Waze Evaluation Procedure Steps

Step	Name	Venn diagram segment	Research motivation	Data	
				Time	Location
0	Exploratory analysis	-	Waze and ATMS reports based on: - Time of day - Region - Road type - Etc.	2016 entire year	entire state of Iowa
1	match Waze and ATMS	A	- Waze and ATMS overlap - Redundancies - Influential factors in Waze coverage	2016 entire year	entire state of Iowa
2	match Waze and INRIX	B	- ATMS and INRIX overlap - Waze vs INRIX contribution to ATMS	October 2016	entire state of Iowa
3	Estimate the false alarms in Waze	C	- % of Waze reports when road is clear (False alarms)	October 2016	Des Moines Area
4	Estimate Waze's contribution	D	- The information that Waze can add		

Results

Exploratory Waze Data Analysis

To initiate the evaluation, an exploratory data analysis was performed to better understand the Waze and ATMS data. The exploratory analysis looked into the pure number of reports regardless of the matching percentages or potential duplicates, to provide a high-level understanding of the two sources of data.

Sources of Incident Detection in the ATMS

Waze has been used as a source of incident detection in the IDOT ATMS since September 2015. As depicted in Figure 2, part (a), among the 23 sources of detection in the Iowa ATMS, law enforcement (which includes 911 calls, County Sheriff, State Patrol, etc.) contributes the highest number of incidents in the ATMS. Interestingly, Waze reports

(detection source for 13.4% of the ATMS records) rank fourth in detection sources, after law enforcement, CCTV, and highway helpers. Comparing the operation and maintenance cost of each of the first three sources, Waze has a considerable contribution as a "free" detection source. However, in the current ATMS settings, the Waze reports need to be verified, usually by one of the top three sources before being trusted.

Incident Reports in Distinct Locations and Road Types

The location of each report was mapped to the demographics of the region based on 2010 census data (United States Census Bureau, 2010). Every county is grouped by their population as either metropolitan (>50,000), micropolitan (10,000-50,000), urban cluster (2,500-50,000), or rural (any non-urban region is considered rural). This analysis provides an insight into the spread and coverage of each source of data. As depicted in Figure 2 part (b), the ATMS has recorded almost no congestion incidents (jams) outside of the metro area. This is while there are many congestion incidents reported in Waze from the urban clusters and rural areas (even off the interstates). In addition, the considerably larger numbers of reports on the interstates show the concentration of reports in both sources. This chart indicates the type of incident and locations where Waze could best contribute to the ATMS.

Impact of Time on the Waze Reports

To evaluate how the crowdsourced data reflect the reality on the roads, the number of reports in each hour of the day were compared and it was expected that the crowdsourced data resemble the ATMS records. As observed in Figure 2 part (c), both data sources tend to have a higher frequency of crash records during the rush hours. However, between midnight and 6 a.m., although ATMS shows 50-100 crash records, there are less than 10 Waze crash reports in the same time. The proportion of the number of Waze to ATMS crash reports during these hours (mean 9%) showed a statistically significant difference from the same

proportion for other hours of the day (mean 37%). This indicates that Waze is not be a reliable detection source during midnight to 6 a.m. This observation aligns with the fact that during these hours there are fewer drivers on the roads and consequently fewer Waze users that might observe and report an incident.

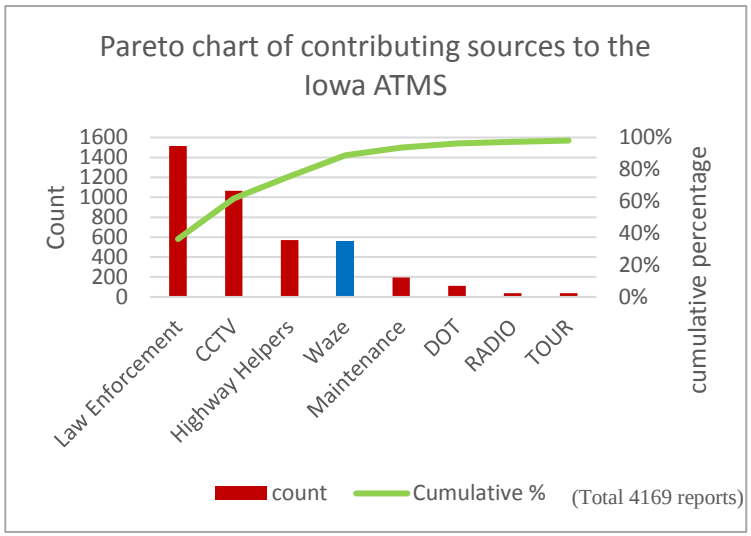
Otherwise, the number of crashes reported in each hour of the day (from 6 a.m. to 11 p.m.) was highly correlated (R^2=0.9) between ATMS and Waze; as depicted in Figure 2part (d). Thus, the number of Waze crash reports during the day follow the reality of the roads.

Evaluation and Comparison

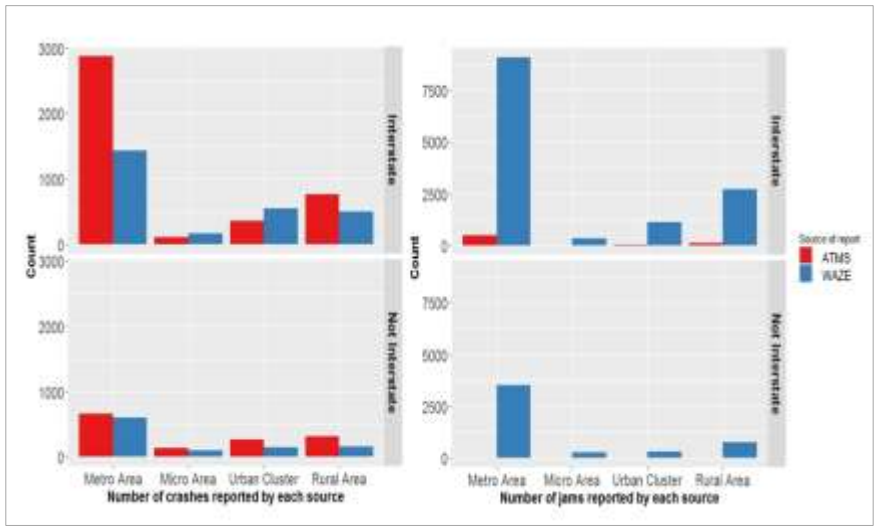
Step 1: The ATMS incidents that were reported in Waze (Estimating A: Waze ∩ ATMS)

This step compares Waze reports to the ATMS reports as source of validated events. The percentage of matching incidents in both sources answers questions regarding the reliability of Waze reports, while leading to the estimation of the potential contribution of Waze.

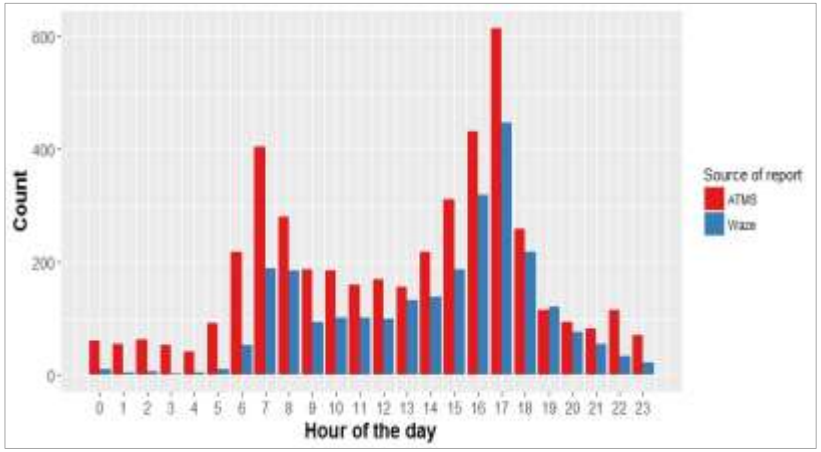
Using the described matching function, overall the congestion and crashes reported in Waze covered 43.2% of the ATMS records. The matching percentage by each type of incident is presented in Table 3.



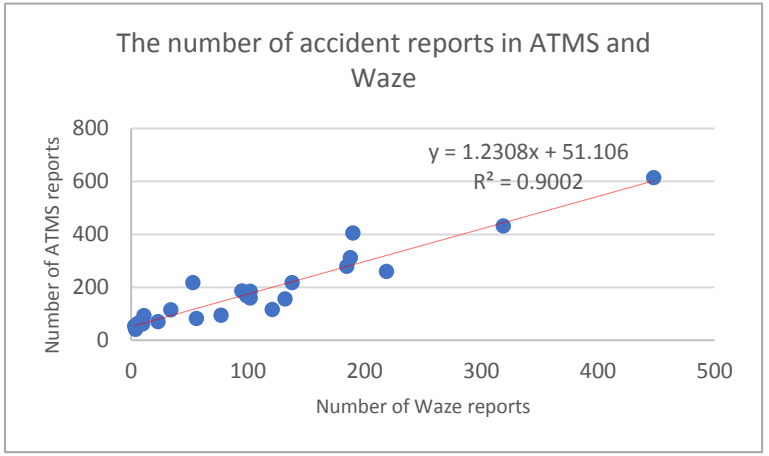
(a)



(b)



(c) Total crash counts in all weekdays of year 2016 per hour of the



(d)

Figure 2 - Exploratory data analysis results, comparing number of reports in Waze and ATMS. All data are from 2016

In Iowa, similar to many other Midwestern U.S. states, traffic is not a daily concern for most people, and thus fewer people are familiar and active users of Waze, compared to more populated cities and states. Yet, the number of matched reports are interesting, considering a single crowdsourced feed of data has captured 43.2% of ATMS records.

Table 3 - ATMS-Waze Matching Percentage by Report Type

Type of incident	Total reports in ATMS	% matched with Waze
Crashes	3713	42.1 %
Congestion	456	58.5 %
Stalled vehicles	12552	43.0 %

What factors contribute to an incident being reported in Waze in the Metro area?

To find the variables which have a statistically significant influence in determining whether an ATMS incident is reported in Waze, a binomial logistic regression was conducted. The binomial logistic regression was performed to ascertain the effects of day of the week, hour of the day, incident type, and the road type on the likelihood that an event covered by an ATMS record would be covered by Waze as well. The logistic regression model was statistically significant, $\chi^2(31) = 450.2$, $p << .001$. The model explained 20.0% (Nagelkerke R^2) of the variance in the matched instances and correctly classified 63.6% of cases. Of the thirty-one predictor variables (factors converted to dummy variables), the statistically significant ones were related to time and road type (as shown Table 4). The incident type did not indicate a significant impact in this model.

Since the road type turned out to be a significant contributing variable to the model, another logit model was tested using the interstate road names (9 variables) in the metro area

as new variables, to investigate if a certain road significantly impacts the chance of an ATMS report being covered in Waze. None of the major interstates indicated a significant impact.

Table 4 - Significant Influencers in ATMS-Waze Matching (** indicates significance level of 0.001)

Variable group	Variable	Estimate	P-Value	**Variable definition**	
Time of the Day	07:00-08:00	1.5444	< .0001 **	07:00 – 09:00	Morning rush hour
	08:00-09:00	0.9143	.0003		
	11:00-12:00	0.6435	.0267	11:00 – 13:00	Lunch time
	12:00-13:00	0.7137	.0135		
	14:00-15:00	0.9792	.0004	14:00 – 19:00	Afternoon
	15:00-16:00	0.8815	.0006		
	16:00-17:00	1.5484	< .0001 **		
	17:00-18:00	1.5602	< .0001 **		
	18:00-19:00	0.8350	.0015		
	20:00-21:00	0.7376	.0333	20:00 – 21:00	Evening
Road type	Interstate or not	0.9083	< .0001 **	Interstate/Freeway or not**	

What Percentage of Waze Was Covered in ATMS? And Were There Redundant Reports?

Only 14.6% of the total Waze reports were matched with incidents in the ATMS records (36.8% for the crashes and 10.0% of the congestion). Thus, it is critical to investigate the unmatched Waze data to estimate the potential added coverage of Waze.

It was also found that on average, each ATMS report matched to 1.9 Waze reports, indicating the redundancy rate in Waze data. The median is 1 report, mean is 1.9, and 80% of the reports have two or fewer matches in Waze.

To examine the accuracy of the matching in distance, the 95% confidence interval for the distance between the matched Waze report and the ATMS record was calculated as .36 to .39 miles. Evaluating the time accuracy of the matches, the time difference (latency of the reports) was calculated. As depicted in Figure 3 (a), the time difference forms a bell-shaped distribution around –0.22 minutes (95% CI, –1.3 to .8 minutes), which is slightly skewed to

left. Slightly more than half of the matched incidents were detected earlier in Waze than the ATMS record.

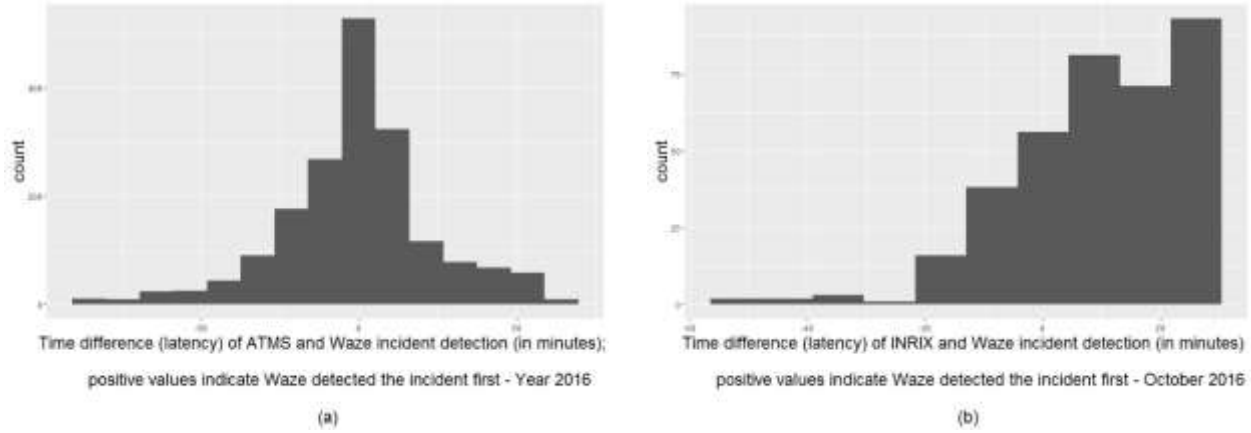


Figure 3 - Waze incident detection time compared with ATMS and INRIX.

Step 2: Estimating the Common Incidents in INRIX and Waze (B)

Although the INRIX reports are not all validated, the overlap of Waze and INRIX reports increases the plausibility of an actual incident occurrence in the same time and location. To control for weather effects in our results, one month with relatively stable weather and about average matching percentages from Waze and ATMS incidents, was desired. October fulfilled the desired properties; therefore, October 2016 data was used for this part. Having applied incident detection method in Iowa ATMS, as described by Chakraborty et al. (Chakraborty et al., 2017), the incidents were detected from INRIX.

Using the described matching function in Table 1, 48% of Region A of Figure 1 (Waze ∩ ATMS) was also matched with INRIX. This result implies that the INRIX feed had detected about half of the common incidents in Waze and ATMS, adding to the validity of the INRIX detected incidents.

To estimate Region B on the Venn diagram, the overlap of the Waze reports with the INRIX data was evaluated. The results indicated 16.8% of Waze reports were matched to INRIX. The time difference between Waze reports and matched incidents demonstrated that on average, INRIX reports were detected 9.8 minutes later (95% CI, 8.25 to 11.36) than Waze reports

(Figure 3 (b)).

Step 3: Estimating the False Alarms in the Metro Area (C)

Region C of Figure 1 represents false alarms from Waze, i.e., reports of incidents that did not actually exist. To estimate the number of false alarms in Waze, manually labelled images from IDOT cameras in the Des Moines metro area were used. The results indicated that overall, only one of the 319 Waze reports in October 2016 and locations was a false alarm. This accounts for 0.3% of the reports.

Although our false alarm definition is not strict (a false alarm is when the road is visibly clear and there is a Waze incident report), the false alarm rate is interestingly lower than expectations. It is worth mentioning a great portion of Waze reports are congestion reports that DOT is not particularly interested in recording. Yet, this is an important finding to understand the validity of these crowdsourced Waze reports.

Step 4: Estimating the Waze Contribution (D)

The final step in the process is to estimate the Waze contribution, or Region D on the Venn diagram of FIGURE 1. Based on the following calculations, 68.3% of the Waze incidents were estimated to be the additional information that Waze can contribute. Once accounting for the number of redundant reports (1.9 redundant reports was rounded up to 2.0 for a more conservative estimation), 34.1% of the Waze's crash and congestion reports (7387

instances which are mainly congestion reports) were potential incidents that were not recorded by the current sources of the ATMS.

$$D = (A \cup B \cup C)' = 100\% - (14.6\% + \ 16.8\% + 0.3\%) = 68.3\%$$

$$Accounting\ for\ redundancies: \frac{68.3\%}{2} = 34.1\%$$

$$Number\ of\ incidents: 34.1\%\ (total\ congestion\ and\ crash\ reports\ in\ Waze) = .341\ \times 21662 \cong 7{,}387$$

To further estimate the potential additional crash coverage in Waze data, the proportion of crash reports among Region D within the 2016 data was 12% of all Region D incidents. Assuming this percentage is uniform in the unmatched Waze reports, this yields about 904 crashes in year 2016 (12% × 7,387 reports) which are either potentially missed or recorded with different labels by the ATMS. These numbers provide an estimate of Waze's potential contribution to traffic coverage in the state of Iowa.

Note that the Waze congestion reports don't come with the recurring or non-recurring labels. Thus, many of the congestion reports might be recurring traffic patterns. Although the ATMS operators are not concerned with the recurrent congestions, the Waze reports still provide invaluable information about the traffic conditions. Moreover, records on all types of traffic incidents provide training data for classification models that can distinguish recurring and non-recurring congestion.

Comparing Waze with Findings about Twitter

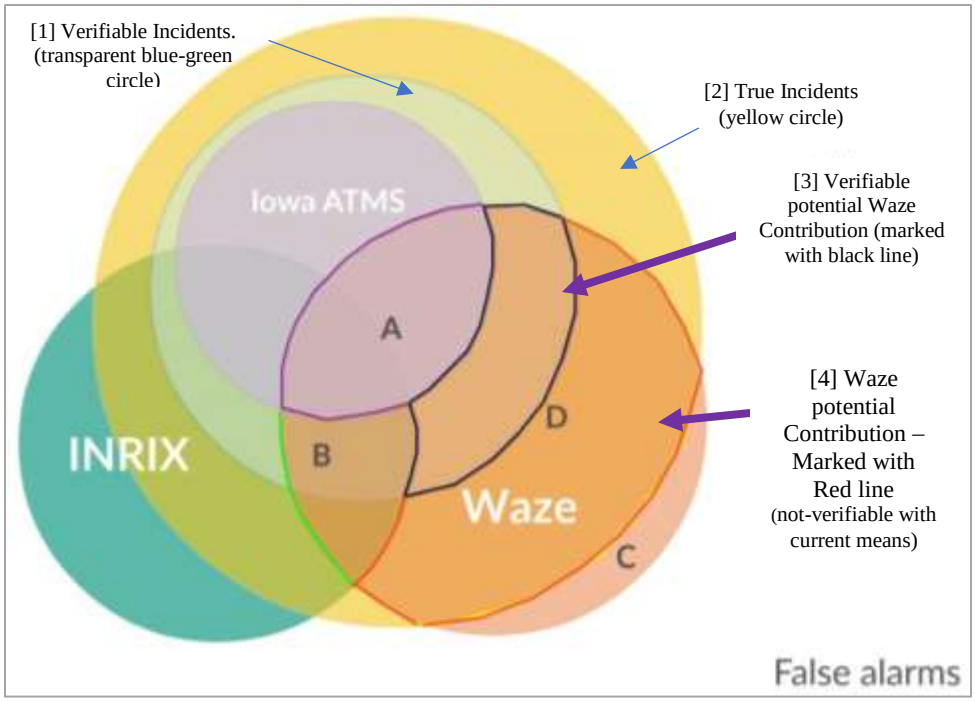
Now that the contribution of Waze has been estimated, it is worth examining its performance with other data sources of data. The work of Gu et al. (Gu et al., 2016) provided information about traffic incidents extracted from Twitter in Pennsylvania. Comparing some of the findings about Twitter with Waze was insightful. Like the present results with Waze,

Gu et al.'s analysis showed Twitter to be less reliable during night hours. Also, most of the tweets were during the peak traffic hours. Gu et al. reported an average of 1.6 Twitter-reported incidents per unique incident. This number was estimated 1.9 reports for Waze, indicating that redundant reports are a common challenge in other crowdsourced data feeds.

Summary of the Findings

Based on the quantitative analysis of Waze data, Figure 4 is an updated view of the Venn diagram that better illustrates the relationship and overlap of the three sources of data. In this another aspect of the challenge is demonstrated. Although there exists a set of true incidents (the yellow circle), not all of them are known through the existing means. Thus, when evaluating the potential of Waze this challenge should be acknowledged. Note that the (D) region in the figure is now split into sections [3] and [4]. The overlap of (D) and Verifiable incidents [3] shows the incidents that are verifiable through other existing means (particularly CCTV cameras). Part [4] in region (D) are reports that can potentially be valid incidents, and there are currently no cameras or other means to verify their accuracy. Based on this work, it is believed that a considerable percentage of the potential incidents in (D) provide invaluable information to the ATMS.

Figure 4 - Updated Venn diagram based on the analysis, the regions are drawn closer to scale. Region D, the estimated contribution of Waze to the ATMS, is divided into verifiable and non-verifiable regions.



Discussion and Conclusion

This research evaluated crowdsourced traffic incident reports from Waze, to study its characteristics as a data source. This section provides a summary of the findings.

How does Waze compare to the existing sources?

The reliability of crowdsourced incident reports from Waze was affirmed with the matching percentages between Waze and validated ATMS (42.3% of ATMS records) and INRIX data. In the Iowa ATMS, 13.4% of the recorded congestion and crashes were initially detected by Waze reports, making it the fourth most contributing source of incident detection. These findings indicate the reliability and competent coverage of crowdsourced traffic incident reports like Waze.

What are the characteristics of Waze data?

Waze incident reports indicated a wide spread coverage of instances in most locations and road types, particularly for reported congestion. The quality of the reports did not depend on the day of week or a specific roadway. On the other hand, the analysis indicated in the less crowded hours of the day (12 a.m. to 6 a.m.), Waze reports are not a reliable source for monitoring road conditions.

What is the estimated potential additional coverage that Waze can provide to the ATMS?

The potential additional coverage that Waze can provide to the ATMS was estimated to be 34.1% of Waze reports, which accounts for 7387 incidents per year (from which 904 were estimated to be crash reports), making it a valuable source for traffic managers to invest.

Overall, it can be concluded that crowdsourced reports like Waze are invaluable sources of information for traffic monitoring with broad coverage, timely response time, and reasonable accuracy. Integrating this source of data into the ATMS feeds provides significant contributions to the traffic monitoring coverage.

However, there are challenges in working with this crowd-based data, including redundancies, inaccuracies, and mismatches in report types, as well as the need for report reliability estimation. Therefore, preprocessing and validating such data is necessary and requires resource investment. The crowdsourced data, on the other hand, are typically provided freely (or at a low cost) to the ATMS managers. Compared to the immense cost of installation and maintenance of other data sources (sensors, third party probe data, or even law enforcement reports), raw Waze data is available for free. This analysis indicated potential valuable incident information from cleaned and processed Waze data. Therefore, a

short-term investment in human resources to establish an infrastructure for eliciting valuable information from Waze data seems economically justifiable. This infrastructure would include models to address the redundancy issue and to automatically estimate the reliability of the reports, which are directions for future work.

Although the exact value of Waze data would vary for different regions and over time, these numbers in a less congested U.S. state seem impressive, and the techniques used in this research for Waze data evaluation could be applied to any region. Moreover, knowing the number of active Waze users in different regions would add a valuable basis for comparative across multiple regions.

Acknowledgment

This material is based on the work funded by Iowa Department of Transportation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Iowa Department of transportation.

References

Aramaki, E., Maskawa, S., & Morita, M. (2011). Twitter Catches The Flu : Detecting Influenza Epidemics using Twitter. In Proceedings of the conference on empirical methods in natural language processing (Vol. 2011, pp. 1568–1576). Association for Computational Linguistics.

Beeri, C., Doytsher, Y., Kanza, Y., Safra, E., & Sagiv, Y. (2005). Finding corresponding objects when integrating several geo-spatial datasets. In Proceedings of the 13th annual ACM international workshop on Geographic information systems (pp. 87–96).

Beeri, C., Kanza, Y., Safra, E., & Sagiv, Y. (2004). Object fusion in geographic information systems. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30 (pp. 816–827).

Chakraborty, P., Hess, J. R., Sharma, A., & Knickerbocker, S. (2017). Outlier Mining Based Traffic Incident Detection Using Big Data Analytics. In Transportation Research Board 96th Annual Meeting. Washington DC.

D'Andrea, E., Ducange, P., Lazzerini, B., & Marcelloni, F. (2015). Real-Time Detection of Traffic from Twitter Stream Analysis. IEEE Transactions on Intelligent Transportation Systems, 16(4), 2269–2283. http://doi.org/10.1109/TITS.2015.2404431

Goh, C. Y., Dauwels, J., Mitrovic, N., Asif, M. T., Oran, A., & Jaillet, P. (2012). Online map-matching based on hidden markov model for real-time traffic sensing applications. Itsc'12, 117543. http://doi.org/10.1109/ITSC.2012.6338627

Gu, Y., Qian, Z. (Sean), & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. Transportation Research Part C: Emerging Technologies, 67, 321–342. http://doi.org/10.1016/j.trc.2016.02.011

Haghani, A., Hamedi, M., & Sadabadi, K. F. (2009). I-95 Corridor coalition vehicle probe project: Validation of INRIX data. I-95 Corridor Coalition.

Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C. C. (2012). TEDAS: A twitter-based event detection and analysis system. Proceedings - International Conference on Data Engineering, 1273–1276. http://doi.org/10.1109/ICDE.2012.125

McKenzie, G., Janowicz, K., & Adams, B. (2014). A weighted multi-attribute method for matching user-generated Points of Interest. Cartography and Geographic Information Science, 41(2), 125–137. http://doi.org/10.1080/15230406.2014.880327

Pack, B. M., & Ivanov, N. (2017). Are You Going My Waze ? Practical Advice for Working. ITEjournal, 87(2), 28–35. Retrieved from https://mydigitalpublication.com/publication/?i=380807

Ruiz, J. J., Ariza, F. J., Ureña, M. a., & Blázquez, E. B. (2011). Digital map conflation: a review of the process and a proposal for classification. International Journal of Geographical Information Science, 25(9), 1439–1466. http://doi.org/10.1080/13658816.2010.519707

Safra, E., Kanza, Y., Sagiv, Y., Beeri, C., & Doytsher, Y. (2010). Location-based algorithms for finding sets of corresponding objects over several geo-spatial data sets. International Journal of Geographical Information Science, 24(1), 69–106. http://doi.org/10.1080/13658810802275560

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. Proceedings of the 19th International Conference on World Wide Web, 851–860. http://doi.org/10.1145/1772690.1772777

Seeger, C., Lillehoj, C., Wilson, S., & Jensen, A. (2014). Facilitated-VGI, Smartphones and Geodesign: Building a Coalition while Mapping Community Infrastructure. In Digital Landscape Architecture 2014 - Landscape Architecture and Planning: Developing Digital Methods in GeoDesign. Zurich, Switzerland (pp. 300–308).

Sester, M., Arsanjani, J. J., Klammer, R., Burghardt, D., & Haunert, J.-H. (2014). Integrating and Generalising Volunteered Geographic Information. In D. Burghardt, C. Duchene, &

W. Mackaness (Eds.), Abstracting geographic information in a data rich world (pp. 119–155). Springer International Publishing. http://doi.org/10.1007/978-3-319-00203-3

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS ONE, 6(5). http://doi.org/10.1371/journal.pone.0019467

Steiger, E., Resch, B., de Albuquerque, J. P., & Zipf, A. (2016). Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. Transportation Research Part C: Emerging Technologies, 73, 91–104. http://doi.org/10.1016/j.trc.2016.10.010

United States Census Bureau. (2010). Urban Area relationship files. Retrieved January 1, 2017, from https://www.census.gov/geo/maps-data/data/ua_rel_download.html

Xavier, E. M. A., Ariza-López, F. J., & Ureña-Cámara, M. A. (2016). A Survey of Measures and Methods for Matching Geospatial Vector Datasets. ACM Computing Surveys, 49(2), Article 39. http://doi.org/10.1145/2963147

Yang, B., & Zhang, Y. (2015). Pattern-mining approach for conflating crowdsourcing road networks with POIs. International Journal of Geographical Information Science, 8816(June), 1–20. http://doi.org/10.1080/13658816.2014.997238

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. http://doi.org/10.1145/1247660.1247686

CHAPTER 3. CLUSTERING CROWDSOURCED TRAFFIC REPORTS FROM WAZE: CHALLENGES AND RECOMMENDATIONS

To be submitted to the *IEEE Access* journal

Authors: Mostafa Amin-Naseri, Vesal Ahsani, Skylar Knickerbocker, Anuj Sharma, Stephen Gilbert, Neal Hawkins

This work was conducted mainly by Mostafa Amin-Naseri. Author Ahsani provided Wavetrinix and INRIX incident data and drafted the corresponding section about incident detection. Skylar Knickerbocker and Neal Hawkins provided valuable directions as subject matter experts. Anuj Sharma and Stephen Gilbert guided this work.

Abstract

Crowdsourced traffic incident reports (CSTIRs) have been shown to complement existing sources in traffic management systems. However, crowdsourced data has its limitations. Its greatest limitation is the redundant reports. The redundancies overload the traffic managers with redundant information, increasing the chance of error. Clustering CSTIRs is a solution which reduced the redundancies, while providing extra information about the impacted area by an incident and the reliability of the cluster. The greatest challenge with clustering is validation. This work explores a procedure for selecting the pertinent clustering method, selecting the validation measures, and tuning the parameters for the desired clusters. The external sources of cluster valuation which are commonly available to traffic agencies are discussed, and customized measures are offered to validate clusters. Moreover, a discussion of the challenges and tradeoff decisions are provided to assist decision making. An implementation of the clustering is demonstrated using Waze CSTIRs in the state of Iowa. The measures and challenges discussed in this work are applicable to CSTIRs as well as to the emerging data from connected vehicles in the transportation domain.

Introduction

Real-time information about traffic conditions is the fuel for a successful Intelligent Transportation System (ITS) (Barbaresso et al., 2014). While sensors and cameras provide valuable inputs to the ITS, it is not economically feasible to have data collection means in every location across the roads. The increasing availability of crowdsourced traffic data (generally through Twitter and Waze) have provided traffic agencies with a free source of data to complement their coverage. Technologies like the Internet of Things, connected vehicles, wearable devices, and voice commands facilitate reporting incidents, which is expected to increase the availability of crowdsourced traffic information. The value of crowdsourced data has been shown in a variety of studies (Amin-Naseri et al., 2018; D'Andrea et al., 2015; Gu et al., 2016; Santos et al., 2016a; Steiger, Resch, de Albuquerque, & Zipf, 2016b). More specifically, Waze data in Iowa was found to cover 43% of the Iowa ATMS records (Amin-Naseri et al., 2018). In addition, Waze had a consistent coverage across various locations, while DOTs' or city officials' incident coverage might be focused in certain locations (Amin-Naseri et al., 2018; Santos, Davis Jr., & Smarzaro, 2016b), making Waze a proper complement to the existing means of data collection. To this end, many cities and state agencies have partnered with Waze (www.waze.com) to gain access to this source of data ("Waze," 2017). The findings, as well as the fact that more than 72 cities and state agencies in North America have joined Waze, reiterate the value in this data (Pack & Ivanov, 2017).

Despite the value in Waze, there are challenges in using this data in practice. One of the main challenges with crowdsourced data (Gu et al., 2016) and particularly with Waze (Amin-Naseri et al., 2018; Pack & Ivanov, 2017) is redundant reports, meaning multiple people reporting the same incident (average 1.9 per recorded incident in Iowa (Amin-Naseri

et al., 2018)). Traffic managers are already loaded with several feeds of information; exposure to an overwhelming feed of redundant data would not only cause distraction but would also be detrimental to managers' trust in Waze reports (Dixon et al., 2007; Madhavan et al., 2006a). Addressing the redundancy issue is thus necessary for using Waze data in practice.

Cluster analysis (also referred to as unsupervised classification) is a well-known approach for grouping similar observations based on their similarities. A successful implementation of clustering both reduces redundancies and provides information about the area which is impacted by an incident. Moreover, a cluster of reports about the same incident increases the reliability in the reported incident. However, cluster validation is known to be the most challenging step in the process (Craenendonck & Blockeel, 2015; Jain & Dubes, 1988). Thus, parameter tuning and cluster validation have been heavily reliant on human assistance (Rosalina, Salim, & Sellis, 2017). Selecting the right clustering method, suitable cluster validation measures, and making trade-off decisions in setting clustering parameters are the main challenges in clustering crowdsourced traffic incident reports (CSTIRs). Although several works have proposed general cluster validations measures to reduce the human role in the process e.g., (Craenendonck & Blockeel, 2015; Jaskowiak et al., 2016; Moulavi, Jaskowiak, Campello, Zimek, & Sander, 2014; Rodríguez, Medina-Pérez, Gutierrez-Rodríguez, Monroy, & Terashima-Marín, 2018), the unique shape of clusters in traffic applications (line centered) limits the applicability of such measures.

To this end, this work explored applicable validation measures for crowdsourced traffic report data particularly from Waze. External sources were used to validate clusters using the commonly available data to traffic agencies. Finally, the main challenges and

tradeoff decisions in the process are discussed to provide the decision maker with objective inputs. The process is demonstrated in an implementation of clustering on Waze crowdsourced incident reports from the state of Iowa.

Background

Clustering methods

Determining the clustering method is key to the cluster analysis. Prior knowledge about the shape of the true clusters as well as the characteristics of the data help narrow down the suitable clustering methods. In the case of crowdsourced traffic data, the clusters are expected to form along the roads and thus be elongated shapes. Moreover, since crowdsourced data is expected to have noise (false or inaccurate reports) the clustering method should be able to handle it.

Density based methods are less sensitive to outliers and are flexible with various cluster shapes. Among the density-based methods, Density Based Spatial Cluster of applications with noise (DBSCAN) (Ester, M., Kriegel, H. P., Sander, J., & Xu, 1996) or DBSCAN-based methods have been widely applied or proposed for various applications including crowdsourced data (Kwak, Liu, Kim, Nath, & Iftode, 2016; M Roriz Junior, Endler, & Silva, 2014; Rosalina et al., 2017). DBSCAN does not require a priori knowledge of the number of clusters. The method takes two inputs: m_{pts}, which defines the minimum number of observations that can form a cluster, and ε, which is the scanning neighborhood. Points that have at least m_{pts} other points in their ε-neighborhood are designated part of a cluster. To expand DBSCAN to spatiotemporal data, Space-Time DBSCAN (ST-DBSCAN) was offered as an extension of DBSCAN (Birant & Kut, 2007). This method takes two epsilon inputs, one for time and another one for location.

DBSCAN results are highly sensitive to the value of both parameters, and while a general heuristic is suggested for $m_{pts} = ln(N)$ where N is the number of observations (Birant & Kut, 2007), the optimal ε needs to be found. To this end, modifications have been made to reduce the number of parameters or simplify the process, such as OPTICS (Ankerst, Breunig, Kriegel, & Sander, 1999), HDBSCAN (Campello, Moulavi, & Sander, 2013; Sun, 2012), and RNN-DBSCAN (Bryant & Cios, 2017). Moreover, DBSCAN uses a single ε parameter for the entire space. However, for non-homogeneous density of data, adaptable epsilon values are more desirable in some applications. Thus, some extensions have been made to accommodate for varying density (Campello et al., 2013; Elbatta & Ashour, 2013; Liu, Zhou, & Wu, 2007; Sun, 2012). Among these methods, Hierarchical-DBSCAN (HDBSCAN) has attracted more attention in application. HDBSCAN runs DBSCAN with all possible epsilon values and generates a density-based clustering hierarchy of all points. This clustering method, applies an optimization algorithm to find the optimal cut in the dendrogram based on a stability measure to gain best cluster solution. This method allows for detecting nested clusters with higher granularity (Rosalina et al., 2017). Campello, et al. have described the HDBSCAN algorithm in detail (Campello et al., 2013). Based on our data, cluster types, and similar works (Kwak et al., 2016; Rosalina et al., 2017), DBSCAN, ST-DBSCAN, and HDBSCAN were considered for further analysis.

Tuning parameters and cluster validations measures

To decide on the exact clustering method and the values of the parameters, proper cluster validation measures are needed. Cluster validation measures are generally divided into three groups: external, Internal, and relative measures (Jain & Dubes, 1988; Moulavi et al., 2014; Rosalina et al., 2017). External measures use pre-specified information about the data (i.e., ground truth) which is outside the original dataset to validate clusters (e.g., in this

application domain, camera images or speed sensor data). The most common external measures are Rand Index (Hubert & Arabie, 1985), Adjusted Rand Index (ARI), and Jaccard (Jain & Dubes, 1988; Rousseeuw & Kaufman, 1990; Vendramin, Campello, & Hruschka, 2010a), which were used in this research.

$$Rand\ Index = \frac{a+b}{a+b+c+d}$$

$$Jaccard = \frac{a}{a+c+d}$$

$$Adjusted\ Rand\ Index = \frac{index - expected\ index}{\max index - expected\ index}$$

$a\ (True\ Positive)$: $the\ number\ of\ paris\ of\ observations\ that\ are\ in\ the\ same\ cluster\ in\ both\ sets$

$b\ (True\ \ Negative)$: $the\ number\ of\ paris\ of\ observations\ that\ are\ in\ distinct\ clusters\ in\ both\ sets$

$c\ (False\ Positive)$: $the\ number\ of\ paris\ of\ observations\ that\ are\ in\ the\ same\ cluster\ in\ the\ ground\ truth\ and\ in\ different\ clusters\ in\ the\ clustering\ results$

$d\ (Flase\ Negative)$: $the\ number\ of\ paris\ of\ observations\ that\ are\ in\ the\ different\ cluster\ in\ the\ ground\ truth\ and\ are\ in\ the\ same\ cluster\ in\ the\ clustering\ results$

However, external validation sources are not always available. In cases where external data sources of validation are not available, internal and relative measures use intrinsic information about the data without any external labels (e.g., dissimilarity matrices) for evaluating clusters. Internal measures are used for determining the best clustering method, while relative measures are used for tuning the parameters of the same method. Some of the common internal and relative measures are Dunn's measure (Dunn, 1974) and Silhouette Width Criterion and its variations (Hruschka, Campello, & de Castro, 2004; Hruschka, Campello, & De Castro, 2006; Rousseeuw & Kaufman, 1990; Rousseeuw, 1987; Vendramin, Campello, & Hruschka, 2009) (refer to (Vendramin, Campello, & Hruschka, 2010b) for further reference). However, most of the mentioned internal measures were found

insufficient for evaluating density-based clusters. Density Based Cluster Validation (DBCV) (Moulavi et al., 2014) has been proposed to address this challenge and was implemented in multiple applications (Bryant & Cios, 2017; Cagnini & Barros, 2016; Craenendonck & Blockeel, 2015). Yet, Rosalina et al. found DBCV insufficient for the specific characteristics of spatial urban data clustering (Rosalina et al., 2017). Thus, the authors suggest customized internal validation criteria for evaluating clusters when external data set is not available.

Final decision

To compare clustering methods, three strategies were explored by Rosalina, et al. (Rosalina et al., 2017). More variations have been reviewed in reference (Jaskowiak et al., 2016). Three forms of comparison were used. The first was overall comparison, in which comparing all combinations of the method and parameters are explored and the best is selected for each validation measure. Second was indices best comparison, in which for each approach, only the best run according to the validation measures is considered for further analysis. Third was default comparison, in which methods are compared with the default parameter values and the best method is tuned for the best parameter. In this research, the best indices comparison was used for comparing the clustering methods.

Despite all the proposed internal and relative measures, there is no measure that out performs all others. Each measure captures certain aspects of a cluster while missing others. Therefore, to coalesce the findings from all measures, multiple strategies have been proposed. In applications with limited knowledge about the true clusters, ensemble methods (Jaskowiak et al., 2016) and voting schemas have been proposed (Rosalina et al., 2017). However, when general information about the desired clusters, as well as the implications of each validation measure are available, using multi-criteria decision making (MCDM) techniques has been applied to find the solutions that maximized the utility for the decision

makers (Kou, Peng, & Wang, 2014; Peng, Zhang, Kou, Li, & Shi, 2012). In evaluating Waze clusters, the knowledge about the clusters and the trade-off decisions exist. Thus, the authors recommend MCDM methods for this application.

Method

The process of clustering is an iterative approach of modeling, exploring, evaluating, and making decisions. Throughout this process, the clustering method, the distance calculations, and the parameters for the method are defined. Moreover, to properly validate the data, the external sources as well as informative internal measures need to be selected. This section describes the process of this work.

The raw Waze data feed is first preprocessed to add three primary features to the raw data. Next, the stream clustering method is implemented to enhance the feed. Finally, the clusters are evaluated, and parameter tuning is performed.

Data preprocessing

The raw data was downloaded as a JSON or XML file and was parsed into a data frame. The end time for each report is extracted based on the last time it appeared in the feed. In the analysis, the end time of the event is also critical for determining real-time road conditions.

On the other hand, information about the road name or direction is occasionally missing in the Waze feed. Moreover, the type of roads (municipal or freeways) are not reported explicitly in the feed. Linear Referencing System (LRS) is commonly used among the traffic agencies, particularly for highways and freeways. Therefore, the geo-coordinates provided by Waze were mapped to the LRS information to provide information about the mile markers as well as the missing road name and directions. The processed data is ready for distance calculations and clustering.

Clustering method

In the literature background section, a breadth of pertinent clustering methods were discussed. According to the desired cluster shapes and the functionalities of the clustering methods, DBSCAN, ST-DBSCAN, and HDBSCAN were considered as the most desirable methods for this type of data.

Distance calculation

Another important factor in clustering crowdsourced reports is distance calculations. The incident reports have space and time dimensions. Some methods like Spatial-Temporal DBSCAN (ST-BDCAN) consider two epsilon values, one for time and another one for location (Birant & Kut, 2007). Others use combined distance measure of time and location, such as DBSCAN (Rosalina et al., 2017) and HDBSCAN.

Distance between CSTIRs must be calculated from a variety of features. In this work we used an additive distance measure which adds distance in several features. These features can be grouped into three categories: spatial, temporal, and contextual.

Spatial features

Spatial features are: geolocation (latitude and longitude), road name, and the direction of the road. First, the Euclidian distance was calculated between the reports and was converted to miles. Then a penalty was added for a mismatch between the road names and directions. The value of the distance should be defined based on the epsilon in the DBSCAN method. Lower penalty values allow for road or direction mismatch clustering, while higher values would prevent the algorithm from clustering such reports with each other. In this work, direction mismatch was not allowed, thus a large value was added to distance for direction mismatch. However, road mismatch was considered with different penalties in cluster validation. Moreover, in the DBSCAN and HDBSCAN method, time and spatial

distances were combined, while in ST-DBSCAN they were considered independently. Further details about the distance calculation and the clustering method are provided in Chapter 4.

Temporal distance

Each CSTIR, after pre-processing, has a start time as well as an estimate of its end time. Thus, when calculating the time distance, there needs to be a calculation of distance between time intervals. In general, events may completely overlap with one another, partially overlap, or not overlap at all, as depicted in Figure 5. When events overlap, a larger time distance between the start or end of events is allowed than when not overlapping.

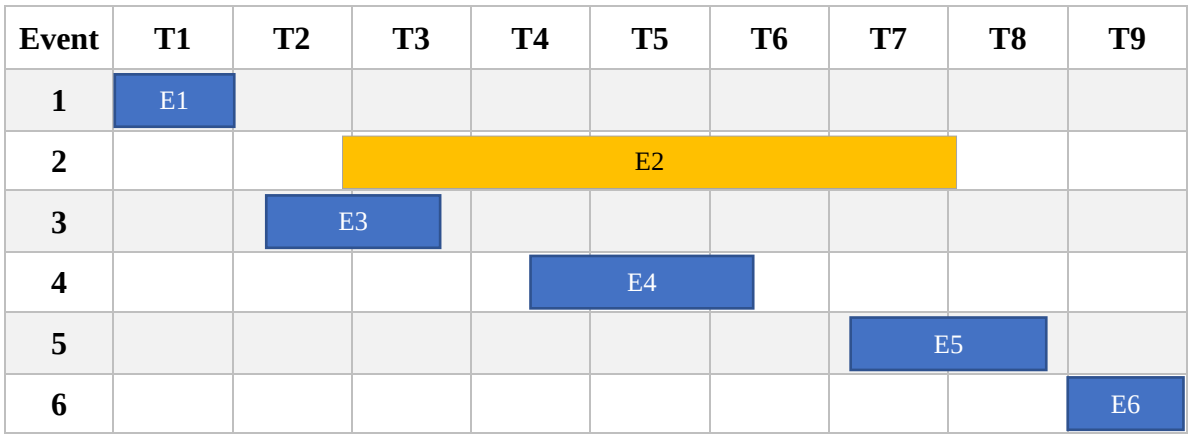


Figure 5: Events across time may completely overlap (E2, E4), partially overlap (E2, E3), or have no overlap (E2, E6). The overlap status affects the temporal distance metric.

In addition to the calculation method, when using combined space time distance, time distance needs to be scaled. In this research, 10 minutes between the end and start of non-overlapping events constructed one unit of distance in time. Moreover, for overlapping events, 1-hour time difference between start or end of incidents was considered a unit of distance in time.

In this research, only crash and congestion reports were considered for analysis. There is a minor distance penalty (0.5 units) added for mismatch in type to avoid clustering

two irrelevant incidents. In addition to the general type of the incident, Waze users can provide subtypes that determine the severity of the incident. However, since these reports are subjective and may vary by different users' perception of the incident, the subcategories were not considered in the distance calculations. Based on the report types in each cluster, a cluster can be marked with a label of Congestion, Crash, or Crash/Congestion. The category of Crash/Congestion is valuable for exploring secondary crashes from congestion.

Thus, the total distance between incidents was measured according to the following equation:

$$d = d_{time} + d_{space} + w_r\, d_{road\ and\ direction} + w_t\, d_{incident\ type}$$

w_r : $Penalty\ value\ for\ road\ and\ direction\ mismatch$

$d_{road\ and\ direction}$: $A\ binary\ matrix\ indicating\ the\ road\ and\ direction\ mismatches$

w_t : $Penalty\ value\ for\ incident\ type\ mismatch$

$d_{incident\ type}$: $A\ binary\ matrix\ indicating\ the\ incident\ mismatches$

External sources of validation

In the application of traffic incident clustering, there are external sources to validate clusters. However, generally none of the sources provide definitive cluster labels. Rather, each provide partial information about road conditions and incidents, which together construct a basis for external cluster validation. The external sources used to create a validation set, as well as their limitations, are briefly discussed below.

a) CCTV cameras

Closed-circuit television camera (CCTV) recordings provide valuable information about the traffic conditions. Recent improvements in deep learning and object detection have

unlocked further opportunities to automatically detect incidents or estimate speed from videos (Chakraborty et al., 2018; Fung, Yung, & Pang, 2003; Poddar et al., 2018; Redmon, Divvala, Girshick, & Farhadi, 2016). However, cameras are not available in all locations. Moreover, the camera might be faced at the opposite direction at the time of an incident. Thus, cameras are a valuable yet not sufficient source for validation.

b) Speed sensors

Most traffic agencies and state DOTs have speed sensors mounted along major roads to track the traffic speed (Haghani et al., 2009; Sharma, Ahsani, & Rawat, 2017). Like cameras, sensors are not available in all locations.

c) Probe-based speed data

Probe-based data consists of traffic speed estimation by a third-party provider based on probe vehicles in the area. Unlike cameras and sensors, probe data is not limited to a certain area, however, probe-based speed data does not provide real-time data for all locations and times (Adu-Gyamfi, Sharma, Knickerbocker, Hawkins, & Jackson, 2017; Kim & Coifman, 2014; Sharma et al., 2017).

d) Congestion reports detected by Waze

In addition to the crowdsourced reported incidents, Waze reports congestion based on their own models of travel times. However, these congestion reports only consider incidents with a certain level of severity and may not include all clusters of incidents reported.

e) DOT or traffic agency incident management records

Lastly, the database of traffic incidents recorded by DOTs or traffic agencies provides another source for cluster validation. It has been shown that there are incidents reported in Waze that are not recorded in these sources (Amin-Naseri et al., 2018; Santos et al., 2016b). Yet they provide information on a portion of the ground truth.

Cluster validation measures and parameter tuning

Prior to tuning the parameters, the plausible ranges for the time and location distance were determined using a subject matter expert's knowledge and historical data. These two questions defined this range: 1) if two CSTIRs appear at the same time, what is the furthest distance for considering them reports of a single event? 2) For two reports of the same type and location, how distant in time can they be before being considered two separate events?

Using these guiding questions, the furthest apart two reports can be to be considered the same incident were defined between 1.5-2.0 miles. The m_{pts} was set to two points to capture smallest clusters while filtering the noise or less notable reports. This was also proposed by Kwak, et al. (Kwak et al., 2016) for clustering visual navigation tweets within their proposed platform of social vehicular navigation.

Comparison strategies

To tune the clustering parameters, the most common external validation measures, ARI and Jaccard coefficient, were used to compare the performance of clustering methods. These indexes are further described in (Prieto, Rodríguez-Triana, Kusmin, & Laanpere, 2017). Moreover, to compare methods, the best indices approach was utilized, meaning the best performing method from each clustering method was compared with other methods.

Data

Several sources of data were used in this study. Data was collected in December 2017 from freeways surrounding the Des Moines, IA metro area (Figure 6). The sources and data types are introduced in this section.

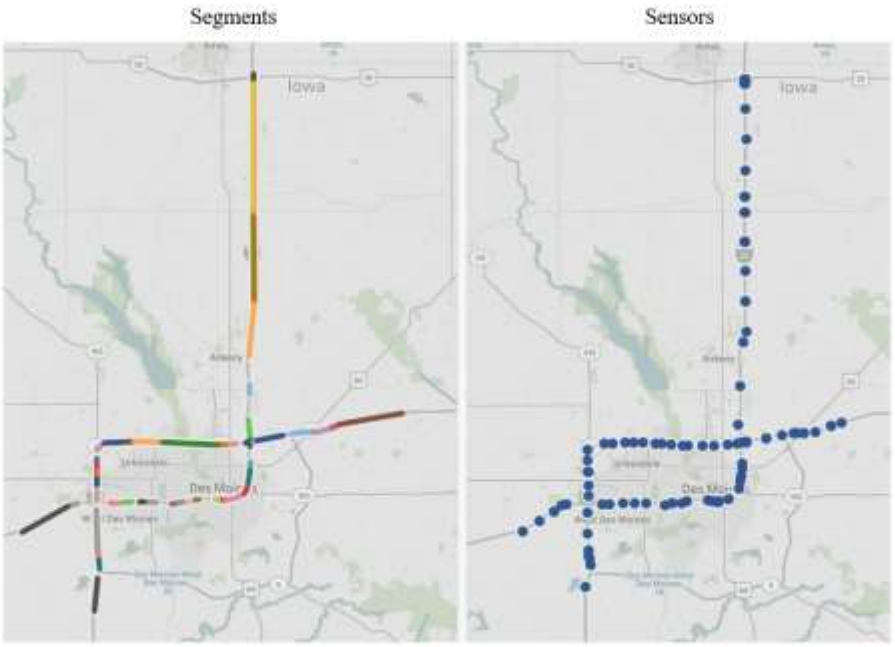


Figure 6 - The location of study and the segment and sensors used for validation.

Waze data

Waze raw data was obtained through Iowa DOT's partnership with Waze, called connected citizens partnership (CCP). The raw data were downloaded every 5 minutes and processed in real-time.

CCTV Camera recordings

For each clustered report in the enhanced Waze feed, the two closest cameras were located. The video feed was manually observed and marked for the observability of the reported incident, the start, and the end of the incident based on the camera feed. These manually labeled video recordings were used as a source of validation and evaluation for the clustered events. Data collection for this source was conducted from Dec. 1-19th, 2017.

Sensor data from Wavetronix sensors

The sensor dataset used in this research was obtained from Wavetronix smart sensors, which utilizes radar technologies for data collection. Although we acknowledge that sensors might have some inherent errors, Wavetronix Smart Sensors have been commonly utilized as

ground truth for comparison purposes, e.g. (Sharifi, Hamedi, Haghani, & Sadrsadat, 2011; Sharma et al., 2017). Each Wavetronix sensor unit consists of a Doppler radar sensor, a wireless modem, solar panel, and on-board processors for real-time processing of traffic data such as speed, volume, etc. High-resolution (20 second) traffic speed data was provided by Wavetronix sensors.

Congestion detection method

After data processing, a congestion detection method was implemented to detect and classify the onset of congestion throughout the network for the study period. Congestion was identified as when the speed data of the segment or the mean of the 1-minute aggregated speed data of the Wavetronix sensor for that location indicated that the speed dropped below 45 mph. According to the Highway Capacity Manual (version 6) [65], LOS (level of service) on basic freeway segments is defined by density. Although speed, as it relates to service quality, is a major concern of drivers, describing LOS on the basis of speed is difficult, as it remains constant up to high flow rates [i.e., 1,000 to 1,800 pc/h/ln for basic freeway segments (depending on the free flow speed)]. There are six levels of service defined for basic freeway segments (levels A–F). The minimum speed of around 50 mph for LOS E is almost constant for different free flowing speeds (from 75 to 55 mph). With an approximately 5 mph average speed bias, 45 mph is considered the threshold for traffic congestion.

The congestion detection process is depicted in Figure 7.The blue line represents the original traffic speed, and red line represents the fixed threshold of 45 mph. The congestion start time is when the speed drops below 45 mph, and the congestion end time is when the speed rises above 45 mph.

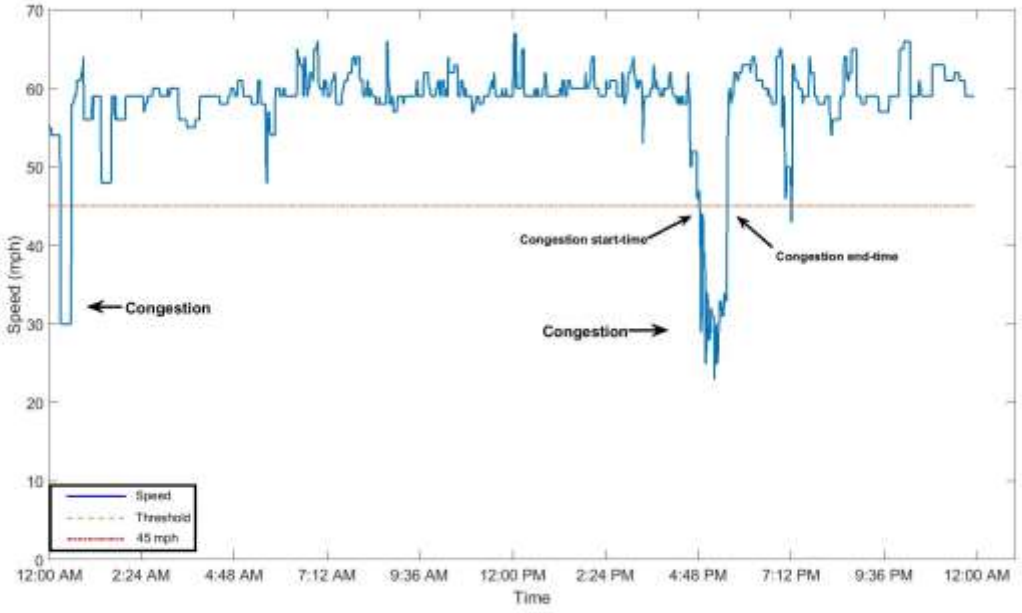


Figure 7 - Congestion detection example

Cluster analysis implementation

Utilizing the external data sources, a validation set was created for comparison. For each epsilon, the ARI and Jaccard coefficients were calculated as presented in Figure 8 and Figure 9. As presented in Table 5, the best performing DBSCAN result was achieved with epsilon value of 1.75 using the combined distance measure and the ARI was equal to 0.81. The ST-DBSCAN method with 1.6 and 0.8 for space and time epsilon respectively, performed closest to the validation set. The ARI for STDBSCAN was higher than DBSCAN (ARI=0.87).

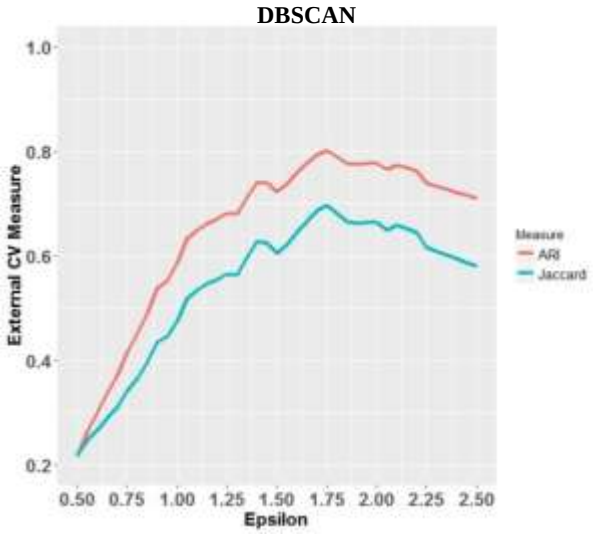


Figure 8 - Parameter tuning for ST-DBSCAN and DBSCAN. For DBSCAN, the ARI and Jaccard coefficient measures were calculated for epsilon values between 0.5 and 2.5.

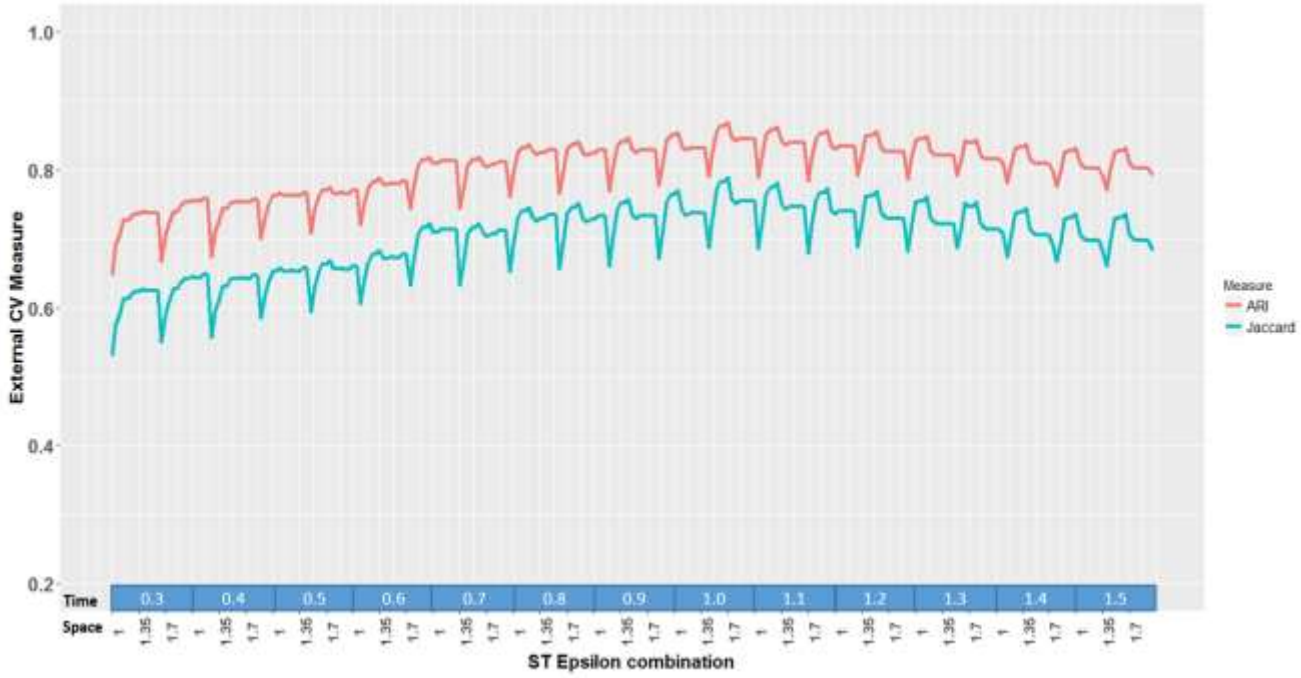


Figure 9- Parameter tuning for ST-DBSCAN. 273 combinations of space epsilon between 1 and 2 (21 values), as well as time epsilon of values between 0.3 and 1.5 (13 values) increments were tested and the CV measures were calculated.

HDBSCAN does not require tuning epsilon. However, the ARI value for HDBSCAN was 0.24, which is significantly less than the other two alternatives. HDBSCAN partitioned the 99 ground truth clusters into 168 clusters, which contributed to the low ARI (Table 5).

Table 5 - Summary of the best performing method from each clustering method.

DBSCAN		ST-DBSCAN		HDBSCAN	
# of clusters	102	# of clusters	101	# of clusters	168
Epsilon	1.75	Space Eps	1.6	Epsilon	NA
		Time Eps	0.8		
ARI	.81	ARI	.87	ARI	.24

As explained in the introduction of HDBSCAN, using the hierarchy of clusters, it detects nested clusters. In other words, while DBSCAN and ST-DBSCAN can merge nearby clusters, the clusters from HDBSCAN are all dense and do not allow merging neighboring clusters when there is not enough density in time and location.

Characteristics of each method

In general, both DBSCAN and ST-DBSCAN performed well in detecting the true clusters, and ST-DBSCAN performed better. Yet it is important to know the characteristics of each method. Figure 10 is a useful illustration of the characteristics of these three methods. In each case, a single cluster according to DBSCAN and the validation set was divided into three and two clusters by HDBSCAN and ST-DBSCAN, respectively. Exploring the clusters, it was observed that HDBSCAN forms clusters with less time variation, while the other two methods were more flexible with time and distance variations. Thus, the tradeoff decision is between the space-time accuracy of the clusters and the number of redundant reports. HDBSCAN provides clusters with higher density, which consequently are more accurate in time and location. However, the feed might still contain redundant reports. On the other

hand, DBSCAN and ST-DBSCAN clusters might be grouping two distinct incidents and inaccurately over estimating the impacted area of an incident. Thus, each method can be beneficial in certain applications.

Figure 10 - A single cluster in the validation data and DBSCAN which has been partitioned into three and two clusters by HDBSCAN and ST-DBSCAN. HDBSCAN has also marked one report as noise (far right). The clusters in HDBSCAN have less variation in time or location; the other two methods allow for more variation in a cluster.

Internal cluster validation measures

As discussed in the literature, most of the common internal validation measures do not perform well with CSTIR data (Craenendonck & Blockeel, 2015; Rosalina et al., 2017). Figure 11 demonstrates the performance of the DBSCAN model using Average Silhouette Width and Dunn2 method, using the same range of epsilon values with the external measures. Contrary to the external measures (both with a max value of 1) these measures deem the results incompetent (close to 0.0), while the match with the true clusters is considerable, thus not very helpful for validating the quality of clusters.

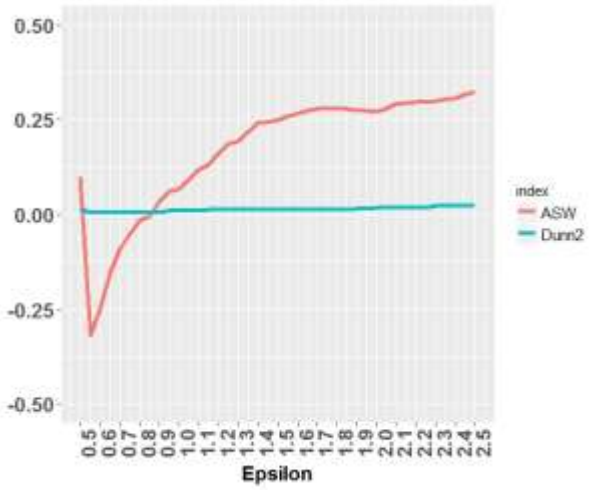


Figure 11 - Classic internal cluster validation measures for epsilon values in DBSCAN. The results confirm the mismatch between these measures and the external measures in Figure 8.

To explore clusters with internal measures, the summary statistics in each cluster of the time and space distance between consecutive CSTIRs in a cluster were explored. Large distances between consecutive CSTIRs indicate potential chaining of clusters, i.e., two unrelated clusters are merged. Thus, large values of mean, standard deviation, and the maximum value are signs of potential undesirable chaining of clusters.

Challenges

Considering the external validation measures, although a significant majority of the clusters were matched the desired ground truth, none of the clustering methods were able to perform exactly as the validation set. Part of this mismatch is due to the complicated nature of nested clusters and the limited means to verify them. Yet there are some special cases that regardless of the parameters can impact the quality of clusters. It is important to understand the characteristics of these cases to be able to inform decision making based on the organization’s priorities. Some of these special cases are discussed in this section.

a) Reports on adjacent roads

As depicted in Figure 12- Part (a), a severe crash on a major interstate (I-80 E) can cause congestion on a ramp and the adjacent road (I-235 E). Therefore, it is desirable to cluster these reports as the same event. However, allowing events on adjacent roads to be clustered can be problematic in cases like Figure 12 – Part (b). Severe congestion was reported on I-80 E and a crash was reported on US-6 E as well. Although they were close in time and location, there isn't a ramp between US-6 E and I-80 E.

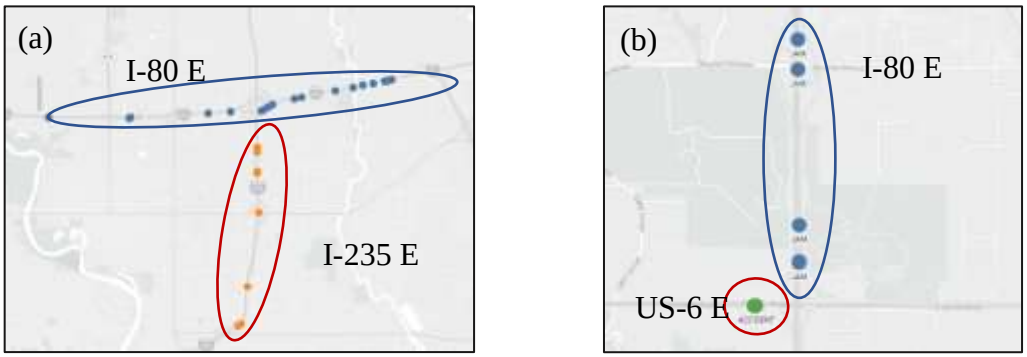


Figure 12 – Challenges with reports on adjacent roads. Part (a) shows severe crash and congestion on I-80E that has impacted I-235E as well. It is desirable to cluster these two incidents together. Part (b), severe congestion is reported on I-80E and a minor crash is reported on Highway 6E. These two incidents had no relation to one another and should not be clustered together.

The challenge of adjacent roads as well as its pros and cons must be well considered for a clustering analysis. In this work, since the cluster results were meant to be used for the traffic managers, the adjacent roads were not allowed to be in a cluster. The assumption was that human managers are able to connect two adjacent clusters that are meaningfully connected. However, we don't want to deceive the managers by reporting a cluster of incidents over multiple roads which are in fact not related. To include this feature, knowledge of the road network is necessary to adjust the distance calculations accordingly.

b) Inaccurate incident end/recovery time

The end time of Waze reports (last time reported in the feed) is calculated using feedback from other users as well as Waze's internal models. Since Waze's model depends on the contribution of active users in the time, there might not be sufficient real-time data for Waze's model to accurately estimate the end time. This means the end times of reports are not always reliable. The uncertainty with the end times is the root cause of cluster mismatch in many cases. For instance, Figure 13 is a depiction of clusters in time and location. Each of the green and purple arrows represent a CSTIR. The arrows of same color have close time overlap. However, report 1 is the last CSTIR reported among the purples. Shortly before the purple is removed from the feed, CSTIR 2 is reported and a new set of reports flow into the feed. In this case it is hard to decide whether the CSTIR 1 was actually cleared from the road and had falsely stayed longer on the map or the green cluster was truly a continuation of the purple CSTIRs and no active Waze users had been there to report. Such cases make the clusters complicated, and the traffic agency must decide based on their priorities to accept the risk of chaining unrelated clusters or avoid it at the risk of receiving multiple clusters from a single incident.

c) Reports of an incident on the opposite direction

It was observed that when an incident significantly impacts the road, drivers on the opposite side report the same incident as well. Thus, a cluster of the same incident is created on both directions. Generally, from the size of the cluster, the original direction of the incident is detectable. However, this is a rare case where clustering incidents from the opposite directions of a freeway is desirable.

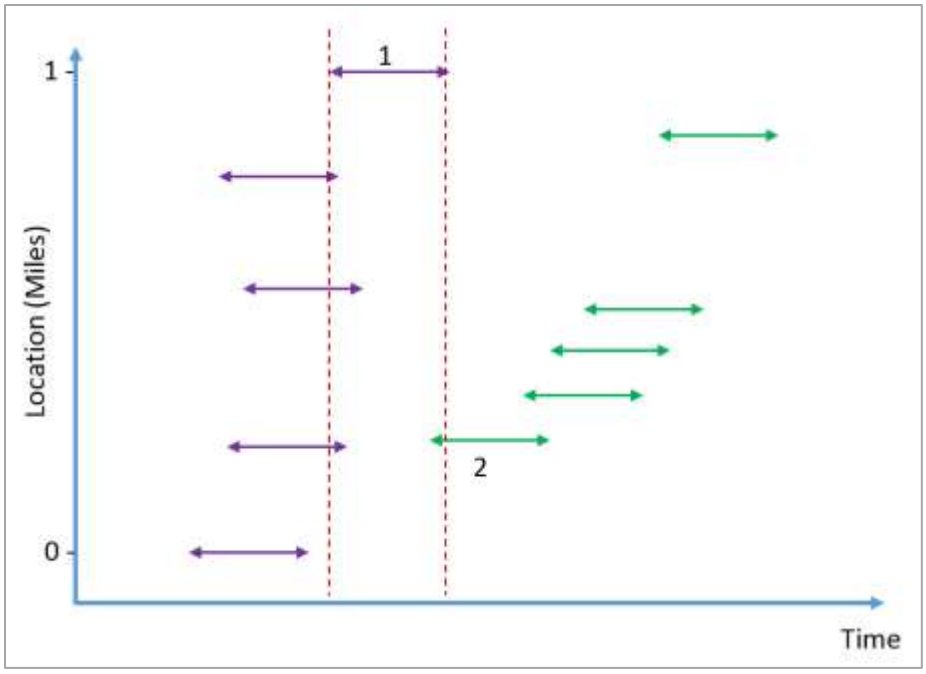


Figure 13 - A depiction of report clusters over time. This example illustrates the challenge of inaccurate end time of incidents from Waze that could falsely group two distinct clusters.

d) Space and time on separate scales or on the same scale?

When clustering spatiotemporal events, the distance can be calculated as a single measure (time and location combined) or two distinct distances used in the ST-DBSCAN. Each have pros and cons. When using combined distance, time and location can compensate for the other, i.e., in the same epsilon CSTIRs which are closer in time can be further apart in location. Similarly, events which are distant in time need to be closer in location to be considered as a cluster. However, in some applications this might not be a desirable or meaningful feature. For instance, a stalled vehicle or pothole might be present for several days and thus, there will be several Waze reports which may be days apart. If time and location could compensate one another. In this case, two pothole reports in distant locations which were reported around the same time could be falsely clustered together. Therefore, in such cases using distinct time and location measures is preferred.

Conclusion

In this work CSTIR clustering was demonstrated and some of its main challenges were discussed. The resulting cluster characteristics, as shown in Table 6, have enhanced the feed, significantly reduced the redundant reports, and provided valuable information to the traffic managers. Feed enhancement on Waze data allows DOTs and traffic agencies to further benefit this valuable source of information to improve their operations in the interest of the public.

Table 6 - The enhanced feed characteristics

	Raw feed	**Enhanced feed**
1	No event end time	Time when removed from the map
2	No LRS location	LRS added
3	Redundant reports (many Unique IDs referring to the same incident)	# of reports reduced to 39% through clustering (61% were redundant)
4	No impact area for an incident	Impact area defined by the cluster shapes
5	Reliability based on a single report	Reliability based on a group of reports (higher confidence)

Furthermore, similar reports from connected vehicles and images from wearable devices will soon become an indispensable part of the ATMS data sources (Budde, De Melo Borges, Tomov, Riedel, & Beigl, 2014; Joy, Rabsatt, & Gerla, 2018; Kwon, Park, & Ryu, 2017; Lee, Gerla, Pau, Lee, & Lim, 2016). Therefore, solutions to reliable CSTIR clustering are necessary utilizing this emerging source of data.

The parameter tuning and validation measures and procedures discussed in this work are applicable in domains of connected vehicles, vehicular social networks(Kwak et al., 2016), and disaster relief organizations (Barbier, Zafarani, Gao, Fung, & Liu, 2012).

Clustering efficiency for real-time implementation

The analysis in this work was implemented in batches for tuning and validation. However, once the clustering method is selected and the parameters are tuned, the model is set to run in near real-time (batch processing every 1 minute). A near real-time implementation of this work is presented in Chapter 4 of this dissertation. Moreover, to further speed the clustering, a plethora of density based models have been proposed that can be utilized for this problem (Amini, Saboohi, Herawan, & Wah, 2016; Amini, Wah, & Saboohi, 2014; M Roriz Junior et al., 2014).

Limitations and future directions

The region of this study encompassed mostly urban areas where cameras and sensors were more densely located. Although, studying this region provided the basis for better validation of cluster parameters, to expand the results to less populated regions, further parameter tuning and data collection is needed.

Moreover, methods to automatically fuse crowdsourced reports with the existing sourced data are an interesting direction for exploration. Finally, once the clusters are validated and satisfactory, it is important to investigate approaches to integrate this feed of data into the traffic management operations, considering the users' mental workload and constraints to best inform the process.

Acknowledgment

This material is based on the work funded by Iowa Department of Transportation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Iowa Department of transportation.

References

Adu-Gyamfi, Y. O., Sharma, A., Knickerbocker, S., Hawkins, N., & Jackson, M. (2017). Framework for Evaluating the Reliability of Wide-Area Probe Data. Transportation Research Record: Journal of the Transportation Research Board, (2643), 93–104. http://doi.org/10.3141/2643-11

Amini, A., Saboohi, H., Herawan, T., & Wah, T. Y. (2016). MuDi-Stream: A multi density clustering algorithm for evolving data stream. Journal of Network and Computer Applications, 59, 370–385. http://doi.org/10.1016/j.jnca.2014.11.007

Amini, A., Wah, T. Y., & Saboohi, H. (2014). On density-based data streams clustering algorithms: A survey. Journal of Computer Science and Technology, 29(1), 116–141. http://doi.org/10.1007/s11390-013-1416-3

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.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). Optics: Ordering points to identify the clustering structure. ACM Sigmod Record, 49–60. http://doi.org/10.1145/304182.304187

Barbaresso, J., Cordahi, G., Garcia, D., Hill, C., Jendzejec, A., & Wright, K. (2014). USDOT's Intelligent Transportation Systems (ITS) Strategic Plan 2015-2019.

Barbier, G., Zafarani, R., Gao, H., Fung, G., & Liu, H. (2012). Maximizing benefits from crowdsourced data. Computational and Mathematical Organization Theory, 18(3), 257–279. http://doi.org/10.1007/s10588-012-9121-2

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial--temporal data. Data & Knowledge Engineering, 60(1), 208–221.

Bryant, A. C., & Cios, K. J. (2017). RNN-DBSCAN: A Density-based Clustering Algorithm using Reverse Nearest Neighbor Density Estimates. IEEE Transactions on Knowledge and Data Engineering, 1–14. http://doi.org/10.1109/TKDE.2017.2787640

Budde, M., De Melo Borges, J., Tomov, S., Riedel, T., & Beigl, M. (2014). Leveraging Spatio-Temporal Clustering for Participatory Urban Infrastructure Monitoring. Proceedings of the The First International Conference on IoT in Urban Space, 1–6. http://doi.org/10.4108/icst.urb-iot.2014.257282

Cagnini, H. E. L., & Barros, R. C. (2016). PASCAL: An EDA for parameterless shape-independent clustering. 2016 IEEE Congress on Evolutionary Computation, CEC 2016, 3433–3440. http://doi.org/10.1109/CEC.2016.7744224

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In Pacific-Asia conference on knowledge discovery and data mining (pp. 160–172). Berlin, Heidelberg: Springer. http://doi.org/10.1007/978-3-642-37456-2_14

Chakraborty, P., Adu-Gyamfi, Y. O., Poddar, S., Ahsani, V., Sharma, A., & Sarkar, S. (2018). Traffic Congestion Detection from Camera Images Using Deep Convolution Neural Networks. In 97th Annual Transportation Research Board Meeting.

Craenendonck, T. Van, & Blockeel, H. (2015). Using Internal Validity Measures to Compare Clustering Algorithms. In Benelearn 2015 Poster presentations (pp. 1–8).

D'Andrea, E., Ducange, P., Lazzerini, B., & Marcelloni, F. (2015). Real-Time Detection of Traffic from Twitter Stream Analysis. IEEE Transactions on Intelligent Transportation Systems, 16(4), 2269–2283. http://doi.org/10.1109/TITS.2015.2404431

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the Independence of Compliance and Reliance: Are Automation False Alarms Worse Than Misses?, 49(4), 564–572. http://doi.org/10.1518/001872007X215656

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics, 4(1), 95–104.

Elbatta, M. T. H., & Ashour, W. M. (2013). A dynamic method for discovering density varied clusters. Int. Journal of Signal Processing, Image Processing, and Pattern Recognition, 6(1), 123–134.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In KDD (Vol. 96, pp. 226–231). http://doi.org/10.1016/B978-044452701-1.00067-3

Fung, G. S. K., Yung, N. H. C., & Pang, G. K. H. (2003). Camera calibration from road lane markings. Optical Engineering, 42(10), 2967–2978.

Gu, Y., Qian, Z. (Sean), & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. Transportation Research Part C: Emerging Technologies, 67, 321–342. http://doi.org/10.1016/j.trc.2016.02.011

Haghani, A., Hamedi, M., & Sadabadi, K. F. (2009). I-95 Corridor coalition vehicle probe project: Validation of INRIX data. I-95 Corridor Coalition.

Hruschka, E. R., Campello, R. J. G. B., & de Castro, L. N. (2004). Improving the efficiency of a clustering genetic algorithm. In Ibero-American Conference on Artificial Intelligence (pp. 861–870).

Hruschka, E. R., Campello, R. J. G. B., & De Castro, L. N. (2006). Evolving clusters in gene-expression data. Information Sciences, 176(13), 1898–1927.

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193–218.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data.

Jaskowiak, P. A., Moulavi, D., Furtado, A. C. S., Campello, R. J. G. B., Zimek, A., & Sander, J. (2016). On strategies for building effective ensembles of relative clustering validity criteria. Knowledge and Information Systems, 47(2), 329–354. http://doi.org/10.1007/s10115-015-0851-6

Joy, J., Rabsatt, V., & Gerla, M. (2018). Internet of Vehicles: Enabling safe, secure, and private vehicular crowdsourcing. Internet Technology Letters.

Kim, S., & Coifman, B. (2014). Comparing INRIX speed data against concurrent loop detector stations over several months. Transportation Research Part C: Emerging Technologies, 49, 59–72. http://doi.org/10.1016/j.trc.2014.10.002

Kou, G., Peng, Y., & Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Information Sciences, 275, 1–12. http://doi.org/10.1016/j.ins.2014.02.137

Kwak, D., Liu, R., Kim, D., Nath, B., & Iftode, L. (2016). Seeing Is Believing: Sharing Real-Time Visual Traffic Information via Vehicular Clouds. IEEE Access, 4, 3617–3631. http://doi.org/10.1109/ACCESS.2016.2569585

Kwon, D., Park, S., & Ryu, J.-T. (2017). A study on big data thinking of the internet of things-based smart-connected car in conjunction with controller area network bus and 4g-long term evolution. Symmetry, 9(8), 152.

Lee, E.-K., Gerla, M., Pau, G., Lee, U., & Lim, J.-H. (2016). Internet of Vehicles: From intelligent grid to autonomous cars and vehicular fogs. International Journal of Distributed Sensor Networks, 12(9), 1550147716665500.

Liu, P., Zhou, D., & Wu, N. (2007). VDBSCAN: varied density based spatial clustering of applications with noise. In Service Systems and Service Management, 2007 International Conference on (pp. 1–4).

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids, 48(2), 241–256. http://doi.org/10.1518/001872006777724408

Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., & Sander, J. (2014). Density-Based Clustering Validation. In Proceedings of the 2014 SIAM International Conference on Data Mining (pp. 839–847). Society for Industrial and Applied Mathematics. http://doi.org/10.1137/1.9781611973440.96

Pack, B. M., & Ivanov, N. (2017). Are You Going My Waze ? Practical Advice for Working. ITEjournal, 87(2), 28–35. Retrieved from https://mydigitalpublication.com/publication/?i=380807

Peng, Y., Zhang, Y., Kou, G., Li, J., & Shi, Y. (2012). Multicriteria decision making approach for cluster validation. Procedia Computer Science, 9, 1283–1291. http://doi.org/10.1016/j.procs.2012.04.140

Poddar, S., Ozcan, K., Chakraborty, P., Ahsani, V., Sharma, A., & Sarkar, S. (2018). Comparison of Machine Learning Algorithms to Determine Traffic Congestion from Camera Images. 97th Annual Transportation Research Board Meeting.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779–788).

Rodríguez, J., Medina-Pérez, M. A., Gutierrez-Rodríguez, A. E., Monroy, R., & Terashima-Marín, H. (2018). Cluster validation using an ensemble of supervised classifiers. Knowledge-Based Systems, 145, 134–144. http://doi.org/10.1016/j.knosys.2018.01.010

Roriz Junior, M., Endler, M., & Silva, F. J. D. S. E. (2014). An On-line Algorithm for Cluster Detection of Mobile Nodes through Complex Event Processing. Information Systems, 64, 303–320. http://doi.org/10.1016/j.is.2015.12.003

Rosalina, E., Salim, F. D., & Sellis, T. (2017). Automated Density-Based Clustering of Spatial Urban Data for Interactive Data Exploration, (May). http://doi.org/10.1109/INFCOMW.2017.8116392

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

Rousseeuw, P. J., & Kaufman, L. (1990). Finding groups in data. Wiley Online Library Hoboken.

Santos, S. R. dos, Davis Jr., C. A., & Smarzaro, R. (2016a). Integration of data sources on traffic accidents. Brazilian Symposium on Geoinformatics - GeoInfo, 192–203. Retrieved from http://www.geoinfo.info/geoinfo_series.htm

Santos, S. R. dos, Davis Jr., C. A., & Smarzaro, R. (2016b). Integration of data sources on traffic accidents. Brazilian Symposium on Geoinformatics - GeoInfo, 192–203.

Sharifi, E., Hamedi, M., Haghani, A., & Sadrsadat, H. (2011). Analysis of vehicle detection rate for bluetooth traffic sensors: A case study in maryland and delaware. In 18th World Congress on on Intelligent Transport Systems.

Sharma, A., Ahsani, V., & Rawat, S. (2017). Evaluation of Opportunities and Challenges of Using INRIX Data for Real-Time Performance Monitoring and Historical Trend Assessment.

Steiger, E., Resch, B., de Albuquerque, J. P., & Zipf, A. (2016). Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. Transportation Research Part C: Emerging Technologies, 73, 91–104. http://doi.org/10.1016/j.trc.2016.10.010

Sun, Z. (2012). A Hierarchical Clustering Algorithm Based on Density for Data Stratification, (Icsai), 2208–2211.

Vendramin, L., Campello, R. J. G. B., & Hruschka, E. R. (2009). On the comparison of relative clustering validity criteria. In Proceedings of the 2009 SIAM International Conference on Data Mining (pp. 733–744).

Vendramin, L., Campello, R. J. G. B., & Hruschka, E. R. (2010a). Relative Clustering Validity Criteria: A Comparative Overview. Statistical Analysis and Data Mining: The ASA Data Science Journal, 3(4), 209–235. http://doi.org/10.1002/sam.10080

Vendramin, L., Campello, R. J. G. B., & Hruschka, E. R. (2010b). Relative Clustering Validity Criteria: A Comparative Overview. Statistical Analysis and Data Mining: The ASA Data Science Journal, 3(4), 209–235. http://doi.org/10.1002/sam.10080

Waze. (2017). Retrieved January 1, 2017, from https://www.waze.com/about

CHAPTER 4. *WAZECLUSTR*: AN R-BASED CODE TO ENHANCE WAZE CROWDSOURCED TRAFFIC REPORTS FOR TRAFFIC MANAGEMENT APPLICATIONS

To be submitted to the SoftwareX journal

Authors: Mostafa Amin-Naseri, Vamsi Krishna, Skylar Knickerbocker, Anuj Sharma, Stephen Gilbert, Neal Hawkins

This work was conducted mainly by Mostafa Amin-Naseri. Author Krishna developed the third module and the web service. Skylar Knickerbocker and Neal Hawkins provided valuable directions as subject matter experts. Anuj Sharma and Stephen Gilbert guided this work.

Abstract

Traffic agencies have recently started using crowdsourced traffic incident report (CSTIR) data to complement their existing information on road conditions. However, there are several challenges in using CSTIRs in traffic management operations, such as the redundant CSTIRs and the unknown reliability of each group of reports. Moreover, the raw data from CSTIR data providers is usually not compatible with the requirements of traffic agencies. This work provides an open-sourced code that enables traffic agencies to download and parse raw data from a prominent CRTISR provider called Waze. Moreover, it allows for near real-time clustering of the reports to address the redundant report issue. The shape and characteristics of clusters are then used to provide information on the reliability, area of impact, and duration of incidents. The results are presented in a compatible format with the common needs of traffic agencies through a RESTful API.

Motivation and significance

Improving traffic safety and operations have long been areas of motivation among researchers and traffic engineers. Traffic incidents, are of great interest due to the huge delay and costs that traffic injuries and fatalities impose on society. To this end, traffic agencies

around the world thrive for collecting the most accurate data on traffic conditions to respond optimally. Crowdsourced traffic incident reports have been shown to complement the existing sourced of traffic agencies, particularly in locations where there are fewer conventional means for collecting traffic data (Amin-Naseri et al., 2018; Santos et al., 2016a). CSTIRs are available from multiple applications as reviewed in (Van Dyke et al., 2016). Many cities and traffic agencies around the world (72 city and state agencies in North America) have partnered with the navigation application provider Waze (www.waze.com) to access the CSTIRs in their jurisdiction, making Waze one the most popular vendors for CSTIRs. In this partnership, traffic agencies share their traffic records with Waze for the public benefit and in return, receive the CSTIRs from Waze users ("Waze Connected Citizens Program (CCP)," n.d.).

The raw data from Waze, although very informative, requires significant preprocessing to suit the regular activities of traffic agencies (Pack & Ivanov, 2017). This has hindered traffic agencies from benefitting from the potential in Waze data. The main challenge with using CSTIR data are redundant reports (Amin-Naseri et al., 2018; Gu et al., 2016; Pack & Ivanov, 2017). Presenting redundant reports to the already loaded traffic operators not only increases the risk of human error, it also reduces human trust (Dixon et al., 2007; Madhavan et al., 2006a) in Waze reports. Moreover, features like the end time of an incident, the impacted area by an incident, and the reliability of a group of reports is not explicitly available in the raw data. Table 7 presents a summary of the enhancements made to Waze raw data by the current work.

This work presents a web-based program tailored to the raw Waze data feed, which downloads and parses the data, applies clustering to leverage the redundant reports to

enhance the feed, reduce redundancies, and remove noise or less significant reports in near real-time. The clusters are updates and tracked in a server which provide a feed of data to the online traffic information dissemination portals such as 511 traffic websites in the United States.

Theoretical basis

Clustering CSTIRs have been suggested by several works (Amin-Naseri et al., 2018; Kwak et al., 2016; Marcos Roriz Junior, Endler, & Silva, 2017; Rosalina et al., 2017). Most of these works suggest density-based clustering algorithms for this application. The models that best fit Waze data were considered for this application; namely Density Based Spatial Analysis of Clusters with Noise (DBSCAN) (Ester, M., Kriegel, H. P., Sander, J., & Xu, 1996), Space-Time DBSCAN (ST-DBSCAN) (Birant & Kut, 2007), and Hierarchical DBSCAN (HDBSCAN) (Campello et al., 2013; L. Li & Xi, 2011; Sun, 2012). DBSCAN does not require a priori knowledge of the number of clusters. The method takes two inputs: m_{pts}, which defines the minimum number of observations that can form a cluster, and ε, which is the scanning neighborhood. The general idea is that points which have at least m_{pts} points in their ε-neighborhood are considered part of a cluster. For the purpose of clustering CSTIRs, $m_{pts} = 2$ is recommended in the literature (Kwak et al., 2016), however other values can be used by the user. The other two models are based on the notion of DBSCAN. ST-DBSCAN takes two epsilon values, one for space and another one for time. HDBSCAN does not require an epsilon value; rather it finds the optimal epsilon value based on the hierarchy of all clusters. The present code allows all three methods for clustering.

Table 7 - Waze feed enhancement

	Raw feed	Added in the enhanced feed
1	No event end time	Time when removed from the map
2	No Linear Referencing System (LRS) information	LRS added
3	Several redundant reports of a single incident	Reducing the number of raw reports by clustering
4	No impact area for a report	Impact area defined by the cluster shapes
5	Reliability based on a single report	Reliability based on a group of reports (higher confidence)

For the clustering there are decisions to be made regarding the distance calculations. The presented code allows for combining space and time as a single measure, as well as using space and time independently. The distance matrix is an additive model which accumulates space distance (in miles) with other distances as penalties. Mismatch in road name, direction, and the report type impose penalties to restrict certain reports from being clustered (e.g., reports on opposite sides of a freeway should not be clustered together, thus road direction penalty should be selected greater than the epsilon to avoid these events from being clustered). Finally, the time distance is calculated can be added to the distance matrix for DBSCAN clustering.

$$d = d_{time} + d_{space} + w_r\, d_{road\ and\ direction} + w_t\, d_{incident\ type}$$

To apply ST-DBSCAN, space and time distance are treated separately. The preferred applications of each clustering method and distance penalty are discussed in Section 4.

Validating cluster qualities is the most challenging step in the cluster analysis process (Jain & Dubes, 1988; Moulavi et al., 2014). For some of the most common cluster validations, refer to these works (Moulavi et al., 2014; Rosalina et al., 2017; Vendramin et

al., 2010b). Moreover, the authors have discussed the challenges and tailored measures for validating Waze CSTIRs in their work, as discussed in chapter 4 of this dissertation.

Waze provides reliability measures for the CSTIRs. Once the clusters are defined, a general score is generated for each cluster based on the reliability scores of each report. The cluster score uses the average of the reliability score from each report in the cluster. Moreover, to account for the number of reports in a cluster, a logarithmic function increases the score of the cluster. On the other hand, a decay function reduces the reliability score of a cluster once the cluster stops receiving new members. Finally, since there are clusters that contain congestion and crashes, a separate reliability score is assigned to accident reports in each cluster. The cluster reliability score function can be customized based on the user's needs. Equations (1) and (2) explain the function that estimates the reliability score of the cluster.

$$\overline{r_j} = \frac{\sum_{i=1}^{k} r_{ij}}{k} \qquad (1)$$

$$R_j = \min\left(10\ , \overline{r_j} + \log(k, K_{max}\ ) \times \left(10 - \overline{r_j}\right) - \log(\ \max(t_{now} - \ t_{last} + 20, 1)\right)\ /2) \qquad (2)$$

r_i: $The\ reliability\ score\ of\ each\ incident\ report, provided\ by\ Waze. Between\ 1-10$

k: $number\ of\ reports\ in\ each\ cluster$

$\overline{r_j}$: $The\ mean\ reliability\ score\ of\ the\ CSTIRs\ in\ cluster\ j$

R_j: $The\ reliability\ score\ of\ cluster\ j$

K_{max}: $User\ specified\ number\ of\ CSTIRs\ in\ a\ cluster\ that\ makes\ it\ a\ completely\ reliable\ cluster$

t_{now}: $The\ current\ time$

$t_{last} + 20$: $Twenty\ minutes\ after\ the\ last\ CSTIR\ in\ the\ cluster\ was\ reported$

Software description

WazeClustR (V1.0) is an open source code of which its core is written in R. The code consists of three main modules. The first module is for downloading and preprocessing. The second module applies the clustering method and distance calculation approach as provided by the user and posts the clusters to a Mongo DB database. The third module keeps track of clusters and updates them overtime. The results are posted to the fourth module, which is a Shiny application that visualizes the clusters. Several R packages were used in this work (Chang, Cheng, Allaire, Xie, & McPherson, n.d.; Hahsler & Piekenbrock, 2017; Kahle & Wickham, 2013; Munir, 2015; Vavrek, 2011; Wickham & Francois, 2015).

Module 1: Data preprocessing

The user inputs the download link that provides access to Waze data, as well as the interval at which she/he wishes to update the data. The WazeDownloadR function downloads the data and parses the XML into a data frame. The end time for each report is updated using the last time when the incident was included in the feed. Moreover, for visualization purposes, points which have been removed from the feed are marked as inactive in the data.

A portion of Waze CSTIRs have missing values on the road names and direction. To unify the road naming convention and add LRS information, a Python script is used to add the route IDs and the mile markers. The Python code is tailored to Iowa DOTs application; however, it uses ArcGIS conventions, and thus is applicable to other agencies. In case that a linear referencing system is not available to the agency, the code removes reports with missing road or directions from the analysis.

Module 2: Cluster implementation

This module calculates the distance matrix between CSTIRs and implements clustering based on the user's preference. The CSTIRs are updated, the distances are

calculated, and the clustering is implemented. Once the clusters are defined, based on the reliability of the reports in each cluster, a general reliability score is assigned to the cluster. The clusters with the new labels are posted to the Mongo DB. The clusters are defined in three forms: Congestion, Crash, or Crash/Congestion. In a Crash/Congestion cluster, the location of the crash reports are also marked on the map.

Module 3: Track and store clusters

One of the challenges when implementing clustering models on batches of data is to track clusters. The cluster labels generated for each cluster may change over time. Moreover, two clusters might merge as new reports emerge. Thus each cluster must be tracked using the members of the clusters. The server consumes Waze reports with cluster labels provided by the R code to track them over time.

The tracking on the server is composed of three main tasks:

a) Data Storage

For tracking, the server maintains two tables. The first table stores all Waze reports with their unique IDs and updates the information on end time or reliability score. The second table maintains information about the clusters. It generates a unique ID for each cluster that is formed, as well as the member of that cluster. Moreover, the reliability score of the clusters is stored in the table as well. If two clusters were merged together, the smaller (or more recent) is closed and the members are added to the larger or older cluster.

b) Processing

When a new update is received from the R code and stored to the Waze reports table, if any other member of a cluster matches with an existing cluster, all members of that cluster are added to the existing cluster. If all members of the cluster are new to the clustered tables, a new cluster is generated and added to the cluster table.

c) Data Feed

The server provides a RESTful service to retrieve the active clusters and un-clustered information along with the points (active points are CSTIRs which are live in Waze feed). Moreover, a supplementary Python code is provided that enables connection to ArcGIS LRS services to connect the points in the cluster according to the shape of the road.

Visualization

This part is mainly used for parameter tuning, cluster validation, and exploring the cluster shapes. The cluster results are visualized in an R Shiny application. Figure 14 demonstrates a general overview of the code architecture and each module.

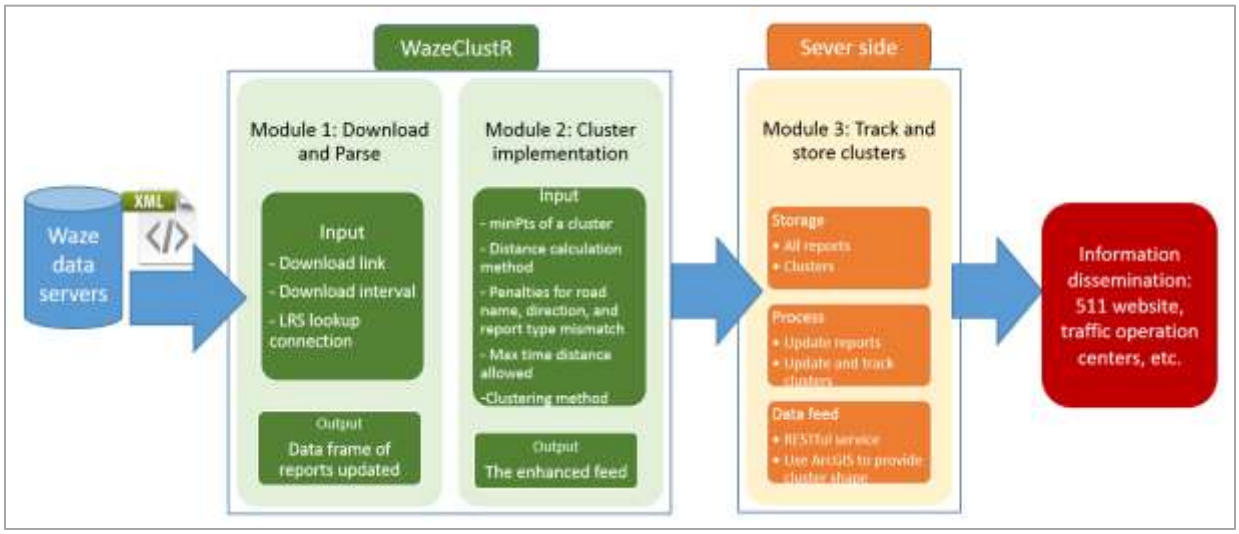


Figure 14 - The architecture of the Waze feed enhancement code

Illustrative examples

Iowa Department of Transportation (IDOT) uses Waze data as a source of incident detection. This example demonstrates the application of the presented code to download and process the data. The data was collected for the Des Moines, IA during Dec 1-21, 2017. The validity of these clusters has been verified using IDOT highway cameras and speed sensors. Examples of the cluster results using different distance measures as well as different clustering methods are presented.

Distance calculations

The code allows the user to decide on decision metrics. In the following some of the decisions and their implications are demonstrated along with the resultant clusters.

Allow adjacent roads to be clustered

If the penalty for the road mismatch is set to a value smaller than epsilon, reports which are in a certain vicinity on different roads can be clustered together. As depicted in Figure 15, both desirable and undesirable cases for clustering CSTIRs on adjacent roads are discussed. The decision must be made based on user's application and priorities in this regard.

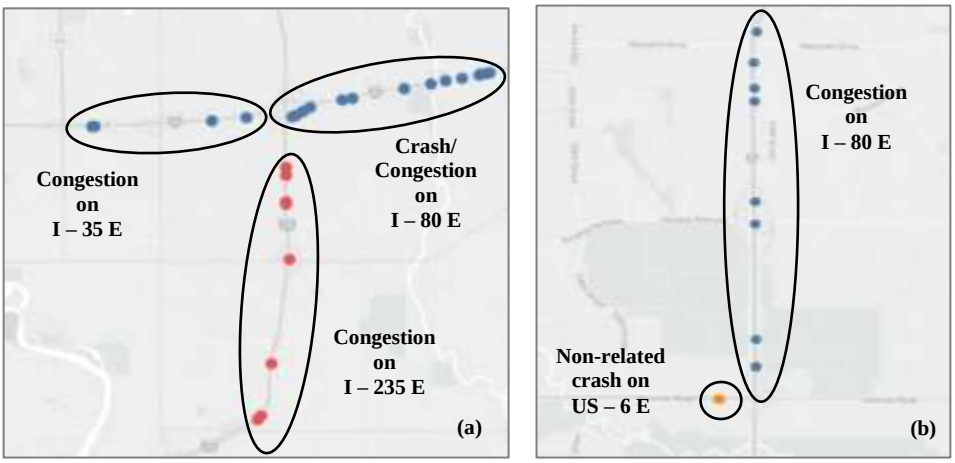


Figure 15 - The distance penalty for road name mismatch is determined such that incident reports on adjacent roads can be clustered. As observed in (a) A major incident on I-80 East has caused congestion on I-35 E as well as I-235 E, thus the clustering method has correctly been able to cluster these reports. On the other hand, in (b) congestion was reported on I-80 E, and an unrelated minor crash was reported on US-6 E. The current penalty allows the model to falsely cluster these reports together.

Time distance penalty

Since CSTIRs have time durations, the distance calculation is calculated as the time between the end and start of two consecutive reports when the durations don't overlap. The maximum allowable time distance (time epsilon) should be specified by the user to calculate

the time distance unit. For example, as demonstrated in Figure 16, changing the maximum time between the start and end of two consecutive events from 10 minutes to 14 minutes results in one and two clusters respectively.

Figure 16 - Sensitivity of clusters to maximum allowable time distance. When the allowable time is set to 10 minutes, all CSTRIRs (which are on the same road and direction) are considered a single incident. With 14 minutes (shown above), there are two distinct clusters.

Clustering method

The results of the clusters can vary when using DBSCAN or ST-DBSCAN. In general, when the variation in time distance in the true clusters is significantly larger than location distance, it is better to use ST-DBSCAN.

Cluster reliability score

The reliability score is calculated for a Crash/Congestion cluster over lifetime of the cluster, as presented in Figure 17. The function demonstrated the way that accounting for the cluster size as well as the time decay makes the reliability score more smooth and meaningful. As observed in Figure 17 (a), using the mean reliability score for the cluster does not reflect the number of CSTIRs on the cluster. Moreover, towards the end when the number reports in the cluster drops from 6 to 4, undesirably the mean reliability score is increased. Part (b) of the figure, demonstrates the reliability score calculated using the proposed reliability score function which reflects the cluster size and the decay in reliability score as reports stop to join the cluster.

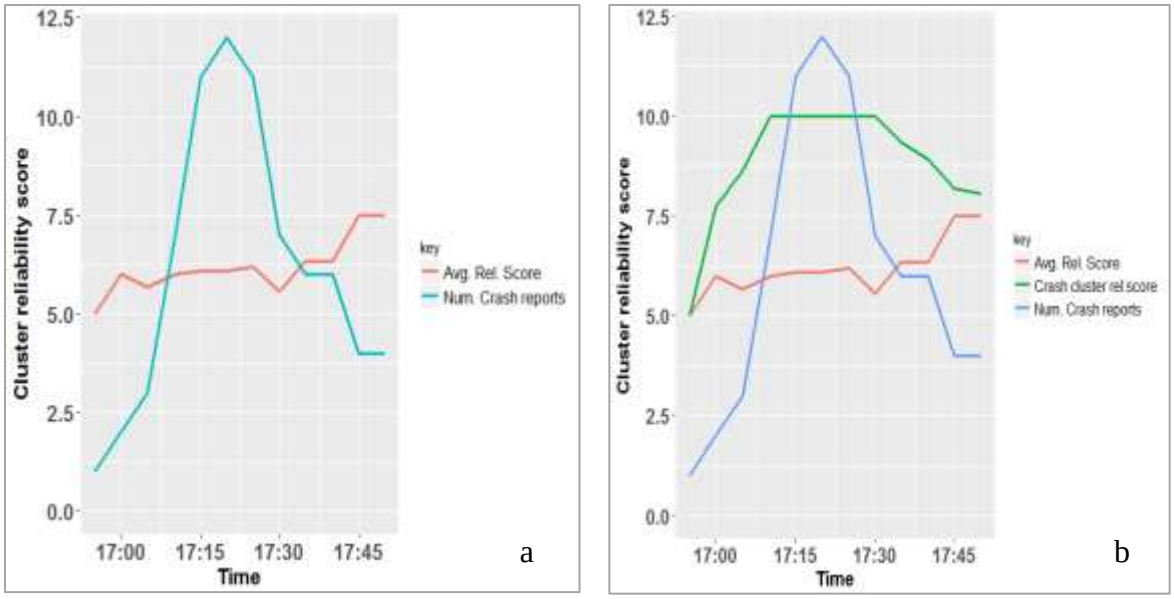


Figure 17 - Comparing the reliability score of a congestion cluster over time using mean CSTIR reliability and the customized reliability score. Part (a) compares the average reliability score of the cluster with the number of active CSTIRs in that cluster. Part (b) shows the reliability score using the proposed reliability estimation function. As observed, the number of CSTIRS in the cluster over time, indicates the impact of size in the score. Moreover, the overall score of the cluster decreases when new CSTIRs stop appearing in the feed.

Cluster visualization

The shape and length of clusters can change over time. The feed for cluster shapes only reports the live events in the Waze feed. That is, if a cluster lasts longer than some of its reports, only the live CSTIRs are posted to the live visualization system (e.g., 511).

Conclusion and future directions

This code presents a remedy to some of the main challenges in adopting Waze CSTIRs in traffic monitoring. The enhanced data reduced the raw data significantly (61% for Iowa) and added information on the duration as well as the impacted area and the reliability of the cluster over time. Further directions and discussions about validation and parameter tuning strategies have been discussed by the authors in chapter 4.

Although the number of clusters is sensitive to the parameters in the algorithm, using the reliability score, mitigates the risk of chaining two distinct incidents as a single cluster. If the reliability score of a cluster starts increasing significantly after it had dropped, an alert is sent to the operator, to relook into the incident, as a new incident might have happened. This feature, reduces the risk in parameter tuning so that decision makers can choose their strategies more easily.

The current DBSCAN method work is updated every minute and for 1000 CSTIRs the processing time is less than a second on a regular PC. However, there are options to improve the efficiency of the distance calculation for large scale data. Several grid-based stream clustering methods have been proposed in the literature that are claimed to closely resemble the DBSCAN results with significantly less calculation time (Amini et al., 2016, 2014; Marcos Roriz Junior et al., 2017; Roriz & Endler, 2014). Applying such methods or using geo-hashes would allow the code to scale better to much larger data sets.

The existing version of this code is tailored to the characteristics of Waze data. However, expansion of this work to other CSTIR data providers such as Here, INRIX, and Beat the Traffic are possible and encouraged.

Acknowledgment

This material is based on the work funded by Iowa Department of Transportation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Iowa Department of transportation.

References

Amini, A., Saboohi, H., Herawan, T., & Wah, T. Y. (2016). MuDi-Stream: A multi density clustering algorithm for evolving data stream. Journal of Network and Computer Applications, 59, 370–385. http://doi.org/10.1016/j.jnca.2014.11.007

Amini, A., Wah, T. Y., & Saboohi, H. (2014). On density-based data streams clustering algorithms: A survey. Journal of Computer Science and Technology, 29(1), 116–141. http://doi.org/10.1007/s11390-013-1416-3

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.

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial--temporal data. Data & Knowledge Engineering, 60(1), 208–221.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In Pacific-Asia conference on knowledge discovery and data mining (pp. 160–172). Berlin, Heidelberg: Springer. http://doi.org/10.1007/978-3-642-37456-2_14

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (n.d.). Shiny: web application framework for R. R package version 0.13. 2, 2016.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the Independence of Compliance and Reliance: Are Automation False Alarms Worse Than Misses?, 49(4), 564–572. http://doi.org/10.1518/001872007X215656

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In KDD (Vol. 96, pp. 226–231). http://doi.org/10.1016/B978-044452701-1.00067-3

Gu, Y., Qian, Z. (Sean), & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. Transportation Research Part C: Emerging Technologies, 67, 321–342. http://doi.org/10.1016/j.trc.2016.02.011

Hahsler, M., & Piekenbrock, M. (2017). dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms. R Package Version, 0–1.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data.

Kahle, D., & Wickham, H. (2013). ggmap: A package for spatial visualization with Google Maps and OpenStreetMap. R package version 2.3. R Foundation for Statistical Computing, Vienna.

Kwak, D., Liu, R., Kim, D., Nath, B., & Iftode, L. (2016). Seeing Is Believing: Sharing Real-Time Visual Traffic Information via Vehicular Clouds. IEEE Access, 4, 3617–3631. http://doi.org/10.1109/ACCESS.2016.2569585

Li, L., & Xi, Y. (2011). Research on clustering algorithm and its parallelization strategy. Proceedings - 2011 International Conference on Computational and Information Sciences, ICCIS 2011, 325–328. http://doi.org/10.1109/ICCIS.2011.223

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids, 48(2), 241–256. http://doi.org/10.1518/001872006777724408

Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., & Sander, J. (2014). Density-Based Clustering Validation. In Proceedings of the 2014 SIAM International Conference on Data Mining (pp. 839–847). Society for Industrial and Applied Mathematics. http://doi.org/10.1137/1.9781611973440.96

Munir, F. (2015). ST-DBSCAN. github. http://doi.org/https://github.com/fitrahmunir/Web-Clustering-ST-DBSCAN

Pack, B. M., & Ivanov, N. (2017). Are You Going My Waze ? Practical Advice for Working. ITEjournal, 87(2), 28–35. Retrieved from https://mydigitalpublication.com/publication/?i=380807

Roriz Junior, M., Endler, M., & Silva, F. J. da S. e. (2017). An on-line algorithm for cluster detection of mobile nodes through complex event processing. Information Systems, 64, 303–320. http://doi.org/10.1016/j.is.2015.12.003

Roriz, M., & Endler, M. (2014). DG2CEP: A Density-Grid Stream Clustering Algorithm based on Complex Event Processing for Cluster Detection. VI Simpósio Brasileiro de Computação Ubíqua E Pervasiva - SBCUP. http://doi.org/10.13140/RG.2.1.4667.4640

Rosalina, E., Salim, F. D., & Sellis, T. (2017). Automated Density-Based Clustering of Spatial Urban Data for Interactive Data Exploration, (May). http://doi.org/10.1109/INFCOMW.2017.8116392

Santos, S. R. dos, Davis Jr., C. A., & Smarzaro, R. (2016). Integration of data sources on traffic accidents. Brazilian Symposium on Geoinformatics - GeoInfo, 192–203. Retrieved from http://www.geoinfo.info/geoinfo_series.htm

Sun, Z. (2012). A Hierarchical Clustering Algorithm Based on Density for Data Stratification, (Icsai), 2208–2211.

Van Dyke, C. W., Walton, J. R., & Ballinger, J. (2016). Synthesis of Kentucky's Traveler Information Systems.

Vavrek, M. J. (2011). Fossil: palaeoecological and palaeogeographical analysis tools. Palaeontologia Electronica, 14(1), 16.

Vendramin, L., Campello, R. J. G. B., & Hruschka, E. R. (2010). Relative Clustering Validity Criteria: A Comparative Overview. Statistical Analysis and Data Mining: The ASA Data Science Journal, 3(4), 209–235. http://doi.org/10.1002/sam.10080

Waze Connected Citizens Program (CCP). (n.d.). Retrieved March 31, 2018, from https://www.waze.com/ccp

Wickham, H., & Francois, R. (2015). dplyr: A grammar of data manipulation. R Package Version 0.4, 3.

CHAPTER 5. CONCLUSION AND FUTURE WORK

This work targeted some of the main challenges facing ATMSs regarding the use of crowdsourced traffic data, each with state of the art data used in the Iowa Department of Transportation (IDOT). This work quantified the value in one of the major sources of crowdsourced incident reports through comparison with existing sources of IDOT. Moreover, to tackle the challenge of redundant reports, a tailored unsupervised learning method to clean the data was implemented, while offering approaches to validate the quality of clusters and tune the parameters. The clusters enhanced the raw data feed by adding information about the impacted area by an incident. Furthermore, based on the shape and number of reports in each cluster, the reliability of each cluster was estimated to assist the operators in prioritizing their decisions. Finally, an open-sourced software package was presented to implement the proposed clustering method on Waze crowdsourced data. Although this work was particularly applied with Waze data for the state of Iowa, the methods as well as some of the general findings are applicable to many other sources of data and locations. Regardless of the actual source of data, validation and fusion techniques implemented in this work are generalizable to other spatiotemporal data sets. A summary of the general finding of this dissertation are presented in the following.

Characteristics of Waze data

Crowdsourced traffic incident reports from Waze demonstrated a considerable coverage (43%) of the existing record in the Iowa ATMS, while offering significant potential in additional coverage, particularly in incident types and location where ATMS has limited coverage. The exact percentage is subject to change over time as the market penetration rate

changes and is expected to improve. However, the additional coverage as well as the timely reports for incidents are rather confirmed for this source of data.

On the other hand, Waze was shown to have less coverage between midnight and 6 am, due to fewer users on the streets in these times. Moreover, similar to other crowdsourced data sources, Waze was found to have redundant reports for each incident. These redundant reports of a single incident, however, add valuable information about the reliability of the report and the impacted area by that incident. Understanding these characteristics enables the decision makers to allocate their resources to this source of data accordingly with reasonable expectations about its quality and coverage.

Dealing with redundancies and reliability

A near-real-time method was proposed for clustering the raw feed Waze to tackle the redundancy challenge. The results enhanced the feed by providing reliability score for each cluster as well as the impacted area by the incident. Moreover, approaches for validating the quality of clusters and tuning parameters were proposed and demonstrated using data from IDOT. Although, this work provides a basis for initiating the Waze clustering process, in more complex transportation networks, there are yet significant challenges that need to be discussed. This work elaborated on some of the challenges and the trade-off decision that need to be made in this process, as directions for future work.

Data feed integration

One of the challenges with adopting any new source of data in the ATMS is compatibility of the data feeds. In this work, an open-sourced software package was offered to implement the proposed clustering method, in a way that suits the needs of most applications in traffic agencies and DOTs. The aim is to provide traffic agencies with a tool that enables them to get started with adopting crowdsourced data like Waze into their

ATMSs and customize it to their own applications. This work was mainly targeted at Waze data which is the prominent provider of crowdsourced data to traffic agencies. However, most functionalities can be used for data from any other provider.

Future directions

Based on each chapter of this work, there are directions for future research and investigation that would add value to the understanding of characteristics of crowdsourced data.

Data coverage and quality

To learn more about what factors influence the quality and coverage of CSTIRs data such as Waze, it is interesting to find the impact of time, location, and market penetration rate on these findings. Thus, conducting similar studies in various locations and in multiple years is desired. Moreover, information about the number of active users in the location of study and the way it changes over time sheds light onto the characteristics of crowdsourced data reports.

Challenges with clustering CSTIRs

Although the demonstrated solution has shown promising results for adopting Waze data into the ATMS, there are yet challenges. Clustering reports on connecting roads was shown to be one of the challenges in this work. Practical strategies for this problem, particularly in urban areas, is a direction which needs further investigation.

Moreover, a reliability score for cluster has been offered which provides some insight to the ATMS operators. However, once the real-time clustering is implemented in the ATMS, the feedback from operators would provide an invaluable source of training data for tuning this function.

Furthermore, to scale this work to large datasets in real-time the computation of clusters can be further improved. Leveraging proper stream clustering algorithms that fit the characteristics of crowdsourced traffic incident reports is another direction future investigation.

User interaction and visualization

This work focuses on the methods and approaches in processing and analyzing the data. Yet, the findings of these methods are intended to inform operators in the traffic operations center. Thus, various aspects of information presentation, user experience, and human factors must be considered to best present these findings to the operators. An ideal presentation would reduce the mental workload of the operators while reducing the chance of error through understanding the needs of the users.

REFERENCES

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.

Barbaresso, J., Cordahi, G., Garcia, D., Hill, C., Jendzejec, A., & Wright, K. (2014). USDOT's Intelligent Transportation Systems (ITS) Strategic Plan 2015-2019.

Bureau of Transportation Statistics. (2015). Freight Facts and Figures 2015, 3.

Bureau of Transportation statistics. (2016). Passenger Travel Facts and Figures 2016. Retrieved from https://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/FFF_complete.pdf

D'Andrea, E., Ducange, P., Lazzerini, B., & Marcelloni, F. (2015). Real-Time Detection of Traffic from Twitter Stream Analysis. IEEE Transactions on Intelligent Transportation Systems, 16(4), 2269–2283. http://doi.org/10.1109/TITS.2015.2404431

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the Independence of Compliance and Reliance: Are Automation False Alarms Worse Than Misses?, 49(4), 564–572. http://doi.org/10.1518/001872007X215656

Gu, Y., Qian, Z. (Sean), & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. Transportation Research Part C: Emerging Technologies, 67, 321–342. http://doi.org/10.1016/j.trc.2016.02.011

Gyawali, S., & Sharma, A. (2013). Stress during Driving for Novice Truck Drivers. In Transportation Research Board 92nd Annual Meeting.

Hennessy, D. A., & Wiesenthal, D. L. (1999). Traffic Congestion, Driver Stress, and Driver Aggression. Aggressive Behavior, 25(6), 409–423. http://doi.org/10.1002/(SICI)1098-2337(1999)25

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006a). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids, 48(2), 241–256. http://doi.org/10.1518/001872006777724408

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006b). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids, 48(2), 241–256. http://doi.org/10.1518/001872006777724408

Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. PLoS Medicine, 3(11), 2011–2030. http://doi.org/10.1371/journal.pmed.0030442

Ministry of Transportation Ontario. (2007). Advanced Traffic Management Systems. In Ontario Traffic Manual, Book 19 (pp. 11–17).

Pack, B. M., & Ivanov, N. (2017). Are You Going My Waze ? Practical Advice for Working. ITEjournal, 87(2), 28–35. Retrieved from https://mydigitalpublication.com/publication/?i=380807

Santos, S. R. dos, Davis Jr., C. A., & Smarzaro, R. (2016). Integration of data sources on traffic accidents. Brazilian Symposium on Geoinformatics - GeoInfo, 192–203. Retrieved from http://www.geoinfo.info/geoinfo_series.htm

Steiger, E., de Albuquerque, J. P., & Zipf, A. (2015). An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. Transactions in GIS, 19(6), 809–834. http://doi.org/10.1111/tgis.12132

Steptoe, A., & Kivimäki, M. (2012). Stress and cardiovascular disease. Nature Reviews Cardiology, 9(6), 360–370.

U.S. Department of Transportation, Bureau of Transportation Statistics. (2016). Transportation Statistics Annual Report 2016, Chapter 5. Washington, DC. Retrieved from https://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/transportation_statistics_annual_report/2016/chapter_5

United States Enviromental Protection Agency. (n.d.).

Van Dyke, C. W., Walton, J. R., & Ballinger, J. (2016). Synthesis of Kentucky's Traveler Information Systems.

Waze Connected Citizens Program (CCP). (n.d.). Retrieved March 31, 2018, from https://www.waze.com/ccp