Evaluating the Temporal Coverage, Reliability, and Contribution of Incident Detection Sources Using Big Data Analysis

Fatemeh Khalilzadeh

Follow this and additional works at: https://lib.dr.iastate.edu/creativecomponents

Part of the Transportation Engineering Commons

Recommended Citation
https://lib.dr.iastate.edu/creativecomponents/518

This Creative Component is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Creative Components by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Evaluating the Temporal Coverage, Reliability, and Contribution of Incident Detection Sources Using Big Data Analysis

by

Fatemeh Khalilzadeh

A Creative Component submitted to the graduate faculty in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Transportation Engineering

Program of Study Committee:
Dr. Anuj Sharma, Major Professor
Dr. Christopher Day, Member
Dr. William Meeker, Member
Dr. Jing Dong, Member

Iowa State University
Ames, Iowa
2020

Copyright © Fatemeh Khalilzadeh, 2020. All rights reserved.
DEDICATION

I would like to dedicate this work to my fiance, Vamsi Krishna Jagarlamudi, whose unconditional love has always been a source of motivation and enthusiasm to face new challenges and keep me moving forward. This work is also dedicated to my parents, Sedigheh and MohammadReza; my lovely twin sister; Zahra, my brothers; Mahdi and Mostafa, My sister-in-law; Mona, and my little angel niece, Diana, for their invaluable support and encouragement throughout my life.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vi</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>vii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER 1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 2. RELATED WORK</td>
<td>3</td>
</tr>
<tr>
<td>CHAPTER 3. DATA DESCRIPTION</td>
<td>5</td>
</tr>
<tr>
<td>3.1 ATMS data</td>
<td>6</td>
</tr>
<tr>
<td>3.2 Infrastructure-mounted sensors</td>
<td>6</td>
</tr>
<tr>
<td>3.3 Probes-based data resource</td>
<td>7</td>
</tr>
<tr>
<td>3.4 Community-based source</td>
<td>7</td>
</tr>
<tr>
<td>CHAPTER 4. DATA ANALYSIS</td>
<td>8</td>
</tr>
<tr>
<td>4.1 Data Processing</td>
<td>8</td>
</tr>
<tr>
<td>4.1.1 Waze Incident detection process</td>
<td>8</td>
</tr>
<tr>
<td>4.1.2 Inrix and Wavetronix incident detection process</td>
<td>10</td>
</tr>
<tr>
<td>4.2 Methodology</td>
<td>12</td>
</tr>
<tr>
<td>4.2.1 Temporal Comparison</td>
<td>12</td>
</tr>
<tr>
<td>4.2.2 Spatio-temporal matching function</td>
<td>14</td>
</tr>
<tr>
<td>4.2.3 Start Time Latency</td>
<td>14</td>
</tr>
<tr>
<td>4.3 Spatio-temporal Matching</td>
<td>15</td>
</tr>
<tr>
<td>4.4 Incidents Detection Latency</td>
<td>17</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Incident Matching Matrix</td>
<td>16</td>
</tr>
<tr>
<td>4.2</td>
<td>Matching and Contributions by the data sources</td>
<td>16</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure 3.1</th>
<th>Study Location</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 3.2</td>
<td>Iowa DOT ATMS Incident Detection Sources</td>
<td>6</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Number of clusters and Non-clustered incidents for different combinations of $\epsilon_t$ and $\epsilon_d$ with $m_p=2$</td>
<td>10</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Threshold Algorithm</td>
<td>11</td>
</tr>
<tr>
<td>Figure 4.3</td>
<td>Incident Detection Algorithm</td>
<td>12</td>
</tr>
<tr>
<td>Figure 4.4</td>
<td>Daily average Incident frequency by different sources over hours of the day</td>
<td>13</td>
</tr>
<tr>
<td>Figure 4.5</td>
<td>Daily average Incident frequency by different sources over hours of the day</td>
<td>19</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to extend my heartiest gratitude to my major professor Dr. Anuj Sharma, for being a wonderful and supportive advisor. I am fortunate that I had the chance to collaborate with his all-star team at Institute for Transportation. I would also like to extend my sincere thanks to my committee members, Dr. Christopher Day, Dr. William Q Meeker, and Dr. Jing Dong for their time, effort, and advice devoted to the completion of my degree. In addition, I would like to thank Pranamesh Chakraborty for his guidance throughout the course of this research. I would also like to extend my appreciation to Skylar Knickerbocker for coordinating with DOT and giving regular feedback on the project. I am thankful for all the support and kindness that everyone has shown me along the way.
ABSTRACT

Traffic incidents impact traffic system components and have a major contribution to congestion, delays, and secondary incidents. In order to avoid these issues, all traffic control systems are trying to detect incidents in a timely manner using different sources of traffic monitoring. This research has developed a comparison method among different sources of real-time incident detection in terms of temporal and spatial characteristics. Incident reports from Waze and ATMS and generated automated incidents from the Inrix, and Wavetronix were used for this analysis. The result shows that all data sets have a similar pattern in recording incidents. The number of incidents recorded by Waze, Wavetronix, and Inrix tracks each other closely during different times of the day. ATMS is a validated set of incidents reports, but still, there are more incidents that were not recorded in ATMS. A state-of-the-art matching function is built to get the number of overlapping incidents from all four data sets. Waze and Inrix have the highest overlapping with ATMS, about 57%, which affirms their reliability. The matching function also was used to get the estimated contribution of each source to ATMS. Due to the lack of validation source, there will be some false alarms in the total number of contributions that were not identified.
CHAPTER 1. INTRODUCTION

Traffic incidents are generally identified as non-recurrent events such as a vehicle involved crash accidents, debris on roadways, slow traffic, stalled vehicle, earlier crash, emergency vehicles, vehicle fire, and towing operation events [28]. Thomas F. Golob et al. [25] showed that there is a strong relationship between types of incident collisions and speed drop in different lanes of the roadway. Incidents have bearing effects on reducing traffic flow, accordingly, capacity reduction as a result of direct impact, namely lane closure or indirect impact caused by drivers slowing down to watch the incident scene [11]. Congestion and subsequent delays are critical repercussions of incidents. Lindley [16] demonstrates that incidents led to around 60-70% of all urban freeway delays in 1998 and 2005. Also, based on the study in 75 urban areas in the United States, Schrank et al. [23] concluded between 52 and 58 percent of the total delay caused by incidents in the year 2000. Just as incidents are the major cause of delay, providing sophisticated sources for detecting incidents can help Traffic Management Center (TMC) and the decision-makers to control the traffic and the system operation in such a way that the responding process will get more efficient and quicker as a result of shortening the time of the detection. Also, informing the travelers with instant information about incident detection can help them in terms of route selection, preplanning, and rescheduling. Moreover, Khattak et al. [3] illustrate that the longer the incident, the more likely of secondary incidents to happen. Detecting, responding to, and clearing incidents in a most efficient way are three major concerns in the area of traffic management that can lead to mitigate negative impacts of incidents namely incident-induced delay [32], incidents costs (Fuel and time consumption) [15], and the risk of secondary crashes [2]. A significant amount of research in this area has been directed toward developing fast and accurate incident detection algorithms [21]. Moreover, a multitude of research has been conducted to provide TMC with incident management programs that improve the efficiency of clearing and responding to incidents [29, 9, 20].
The primary focus of this research is to evaluate different data sources for real-time incident detection, namely probe-based data (Inrix), Infrastructure sensor (Wavetronix), community-based (Waze) data, and Advanced Traffic Management System (ATMS) data source. We considered the validated events from the ATMS dataset as the ground truth data. Since the Inrix and Wavetronix don’t directly report the incidents, in section 4.1, a big-data analysis was performed on these datasets, and an incident detection algorithm was developed to extract the incidents. In section 4.2.1, the time of the day when to rely on Inrix, Wavetronix, and Waze incident reports are extracted. This analysis can be useful to identify which source is more reliable based on the time of the day. In section 4.3, we performed a Spatio-temporal comparison algorithm to match the incidents from Inrix, Wavetronix, and Waze with the ATMS incidents. It is true that ATMS is a validated source, includes the majority of incidents, but still, it is a subset of all incidents, and each of the remaining three sources can have a contribution to the ATMS. We showed these contributions in section 4.3. Matching rate with the benchmark dataset and incident start time latency as two key factors to measure the reliability of these datasets are explained in section 4.4.
CHAPTER 2. RELATED WORK

Traffic incident detection sources are mainly classified into two groups, human-based methods and intelligence-based technologies. Typical examples of human-based incident detection sources are traffic navigation devices, state patrolling, highway helpers, and police. Intelligence-based sources may include information obtained from cameras, sensors, and segments. Crowd-sourced data (data gained from the general public and commercial traffic devices such as navigations apps, connected vehicles, etc.) and social media are the cost-effective alternatives for collecting the traffic data to traditional intelligent transportation system technologies such as cameras and sensors which have a limited coverage over the traffic network due to the high installation and maintenance costs[19, 31, 13]. Also, Crowd-sourced data and social media are used to integrate the available data sets. Smarzaro et al. [24] showed that integrating Location-Based Social Network (LBSN) data such as Yelp, Foursquare, etc., into the official data generated by public governmental agencies provides significant insights for the sake of urban planning and decision-making process. Amin-Naseri et al.[17] determined the pure contribution that Waze can provide to ATMS. They determined the percentage of Waze contribution to ATMS by subtracting Waze-ATMS as well as Waze-Inrix matching incidents, using a start time in conjunction with the location as matching criteria, and false alarm reports in Waze, using camera images, from the whole Waze incident reports. Belzowski et al. [7] constructed a field test method to compare the accuracy of traffic jam reporting from the Inrix app and the other 5 data provider devices. All devices did a better job of reporting longer traffic jams (i.e., more than 10 minutes), and those ones which were located on highways than surface streets.

Various Automatic Incident Detection (AID) algorithms have been developed to detect incidents based on the traffic data from intelligent transportation system technologies that don’t require human-based traffic monitoring. Ahsani et al. [26] utilized a dynamic threshold algorithm provided
by Chakraborty et al. [22] to detect both recurring and nonrecurring congestions. Based on the speed data, provided by Inrix and Wavetronix, recurring congestions happen when speed is less than 45 mph and more than the threshold. Nonrecurring congestions happen when speed is less than 45 mph as well as the corresponding dynamic threshold and should be reported in ATMS as well. Then they applied the same procedure to Wavetronix sensors speed data and got its congestion detection results as the ground truth source. They came to the conclusion that Inrix is more reliable in detecting recurring congestions than the non-recurring ones. Also, it does a better job on longer segments. Adu-Gyamfi et al. [30] applied a pattern recognition algorithm on both Inrix, and Wavetronix speed data to detect short-trend and medium-trend congestions that takes between 15 to 30 minutes, and between 1 to 3 hours respectively. Based on the synchrony between Inrix trends and Wavetronix ones as the ground truth data set, and considering different combinations of freeway, non-freeway, Short-term, as well as medium-term congestion, the Inrix accuracy of medium-Term congestion detection is the highest (95%) on freeway with on average 12 minutes latency.

The data from the intelligence-based sources itself should be trustworthy to build useful analytic. Kim et al. [14] used time-series graphs, showed that three factors could affect the accuracy and reliability of Inrix data set, namely latency (the time lag between speed reports of Inrix and corresponding report of the loop detector sensors as the ground truth data), repetitive speed reports, and the disability of Inrix confidence measures (Confidence score, and C-value which will be explained more later) in detecting both latency and repeated speed reports. Also, Haghani et al. [12] stated that the confidence score is not representative of the data quality, and only shows how the data has been calculated.

In this research Inrix, and Wavetronix, as two intelligent data sources, as wells as Waze, as a community-based dataset, are used and their reliability, temporal coverage, and added values to the existing source of incident detection are discussed.
CHAPTER 3. DATA DESCRIPTION

The main goal of this study is to compare four kinds of valuable sources of real-time incident detection, namely, Infrastructure sensor (Wavetronix), Probes-based (Inrix), community-based (Waze), and Advanced Traffic Management System (ATMS) data resources. These data sets were collected from the Iowa Department of Transportation (DOT) from April to October 2018 due to avoiding the weather effects on the analysis and having stable weather during those months. Based on the Automatic Incident Detection Algorithm, which will be explained later, the last seven weeks of speed data are needed to generate incidents for each day. So that after applying the AID algorithm on Inrix as well as Wavetronix datasets, we got the incidents from May 28th to October 26th. The case study area is a section of I-35, I-80, and I-235 located in Des Moines, Iowa which is shown in Figure 3.1.

Figure 3.1: Study Location
3.1 ATMS data

Advanced Traffic Management System (ATMS) data is a complete source of incident detection records all over the Iowa state. It consists of all incidents recorded by different sources such as sensors, cameras, County Sheriff, state patrol, and highway helpers, or police at the incident scene. Then these records should be confirmed by ATMS operators. ATMS dataset contains important information of incidents, namely date, start time, end time, incident duration, type of accidents, and locations. Figure 3.2 [10] shows Traffic Incident Management (TIM) and highway helper performance report for the Iowa department of transportation (DOT). In ATMS dataset, 1 vehicle collision, 2 vehicle collision, 3+ vehicle collision, Debris on Roadway, Slow Traffic, Grass Fire, Flooding, Vehicle Fire, Stalled Vehicle and Emergency Vehicles are included inside the Incident category.

![Figure 3.2: Iowa DOT ATMS Incident Detection Sources](image)

3.2 Infrastructure-mounted sensors

Wavetronix sensors are non-intrusive automatic traffic counting devices installed above the road. These sensors provide us with volume, speed, and vehicle classifications, including passenger vehicle, single-unit vehicle, and combination vehicle using radar technology every 20 seconds [5, 1, 18]. In this case study, 130 Wavetronix sensors are available.
3.3 Probes-based data resource

Iowa DOT resumed boosting the DOT-owned traffic sensors with probe data through the Inrix TMC monitoring platform in 2013 [4]. Inrix provides us with real-time and historical traffic flow containing speed, travel time, and location every minute [30]. Confidence Score, C-Values, and the type of Segments are three criteria in getting the Inrix dataset. Confidence score shows the real-time, historical, and combination of them by taking the value of 30, 20, and 10, respectively. C-Values are reported only when the confidence score takes on 30. TMC and XD are two segment types defining road parts on which the data are gathered. XD segments can take on up to 1.5-mile length so that they are more constant than TMC segments, which can even take on length values more than 15 miles. The selected study area contains 281 Inrix XD-segments.

3.4 Community-based source

Waze is a GPS navigation smartphone application that provides drivers with a lot of information about on-going and up-coming road traffic, construction works, and police locations. Real-time traffic data, such as traffic jams, weather hazards, incidents, etc., are shared by drivers. Iowa DOT also has begun to augment its real-time traffic data with Waze since September 2015 [17]. Report Rating and reliability score are two parameters provided by Waze to indicate the data quality reported by Waze users. Report Rating indicates the user rank between 1 to 6 based on the reporting accuracy, Wazers experience, and the Wazers map contribution. A Waze user with rate 6, has the highest rank. Waze assigns a 1 to 10 reliability score to each incident based on other users’ reactions and experience level of the reporter. The higher score shows the more reliable data [27].
CHAPTER 4. DATA ANALYSIS

Detecting incidents in a timely fashion can reduce the impact of congestion in form of travel delays reduction and lesser number of secondary crashes. This chapter first describes the methodology used for automatic congestion detection based on the speed data from Inrix, and Wavetronix. Also, we present an approach to remove redundancies from Waze dataset. Then, the temporal coverage of different incident detection sources is discussed. And finally, overlapping rate with ATMS, and incident start time latency are used as two key factors to measure the reliability of all data sources used in this article.

4.1 Data Processing

4.1.1 Waze Incident detection process

In this study, only incidents reports were considered. This data provided us with information such as start time and incidents locations. The raw data was downloaded as 40 GB XML files, then was parsed and transformed into one single CSV file with the size of almost 10 GB all over Iowa during June to October 2018. This 10 GB of data was imported into Tableau to get the incidents for only the desired location of this study. The final dataset is almost a 421 KB CSV file in which 80% of the data has ReportRating between 0 and 2, and only 20% of the data has ReportRating between 3 and 5. On the other hand, almost all of the data has reliability score of more than 5. Also, the data for August 24th, August 4th, 5th, and 6th were missed. So that for the sake of fair comparison, all the incidents that happened in those days were also removed from all four incident detection sources.

Waze dataset contains incidents that are reported by the motorists so that there is no end time can be found in these reports. Moreover, the drivers are traveling with a specific speed while reporting the incidents, so that the start time and the location of reports are not precise. In the
Waze data set, redundancy is possible because many drivers can report the same incident. Amin-Naseri et al. [6] showed that Space-Time DBSCAN (ST-DBSCAN) clustering algorithm [8] has the best performance to find similar reports in the Waze data set based on the desired cluster shapes, the functionalities of the clustering methods, and the cluster validation. In this study, the ST-DBSCAN clustering algorithm was used as well. The algorithm takes on three parameters to create clusters. The minimum number of points in each cluster ($m_p$), the maximum time ($\epsilon_t$) as well as the maximum distance among points in the cluster ($\epsilon_d$). First, the Waze data set was grouped by road names and directions. Then, the algorithm applied to each group of the same road and direction, to get a single incident corresponds to many reports. The start time of each incident was extracted based on the first reports start time in its corresponding cluster. All combinations of $t$ in the range of (10,130) minutes with 10-minute time step, and $d$ in the range of (0.2,2.1) mile with 0.1-mile step were examined. The number of clustered, and Non-clustered incidents for $\epsilon_t = \{0.2,0.5,0.7,1,1.5\}$ mile, and all combinations of $\epsilon_t$ are shown in figure 4.1. The maximum number of clustered incidents was found with $\epsilon_t = 20$ minutes, $\epsilon_d = 0.5$ mile, and $m_p = 2$ so that these values were chosen to cluster incidents and remove repetitive reports from our Waze data set.
Figure 4.1: Number of clusters and Non-clustered incidents for different combinations of $\epsilon_t$ and $\epsilon_d$ with $m_p=2$

4.1.2 Inrix and Wavetronix incident detection process

**Wavetronix** Totally 5 GB of sensors speed data over the desirable time and location were analyzed.

**Inrix** Approximately 700 GB of XD-Segments all over the Iowa state during 29 weeks was filtered based on Confidence score of 30, C-values equals to 30 or more, and the desirable study location. To do so, the filtering function was applied for each day which contained around 4 GB dataset, then all the days were attached. After the filtering was applied, approximately 1 GB dataset was used for incident detection analysis.

**Incident detection algorithm** Automatic Incident Detection Algorithm (AID) offered by Chakraborty et al. [22] was used to detect incidents using the 1-minute interval of Inrix and 5-minute aggregated interval of Wavetronix speeds. According to his methodology, incidents are detected based on computing a speed dynamic threshold (Median-2×Interquartile Range (IQR)) to distinguish outliers as incidents when the real-time speed is less than that threshold for a constant 3
intervals. The false alarm is recorded then will be checked by false alarms within 1 hour and 2 miles of nearby segments or sensors then report all as an incident. The algorithm flowchart for generating thresholds as well as detecting incidents are shown in the Figure 4.2, and 4.3 respectively.

Figure 4.2: Threshold Algorithm
4.2 Methodology

The primary purpose of data analysis was to compare the four main incident detection data resources Inrix, Wavetronix, Waze, and ATMS, with each other in terms of temporal and spatial coverage from June to October 2018 in some parts of Des Moines area. The most reliable hours of the day for each source are determined in the Temporal comparison part. Then the Spatio-Temporal matching function was used to get matching incidents.

4.2.1 Temporal Comparison

In terms of temporal comparison, for each time of the day, incident frequency has been evaluated to achieve the more reliable time interval of the day for each source for the sake of real-time incident detection. The daily average of incidents over all hours of the day are illustrated in Figure 4.4.
Each time corresponds to incidents that happen within a one-hour interval. For instance, 10 AM shows all incidents between 10 AM and 11 AM.

![Daily average Incident frequency by different sources over hours of the day](image)

**Figure 4.4**: Daily average Incident frequency by different sources over hours of the day

The temporal comparison seeks to find the efficiency of reporting incidents during time of the day for all incident detection resources. In other words, it helps us to find the time interval of the day in which each source can be more reliable in terms of number of detected incidents.

The results indicate that the more reports from Waze happened only from 6 to 9 AM, and 5 to 7 PM. Also, it can be interpreted that people are more likely to report incidents during 6 to 9 AM as well as 5 to 7 PM. As you can see there is no reports or very less (at most 1 during all 5 months) from 12 to 5 AM. As Waze reports are from only drivers on that road, it is most probably because of the limited number of drivers drive at those time of the day.

All four sources have almost a similar pattern to find incidents. Besides ATMS more number of incidents were detected by Waze. We can also find that Inrix has a better detection than the Wavetronix because Inrix network has a huge number of XD-segments with a length of 0.2 to 1.5 mile that were able to cover all the roads in the network. Also, in some incidents, no need for
highway helpers or police to attend the incident scene so that no more reports of those incidents can be found in the ATMS reports. 7 to 8 AM, 4 to 5 PM, and 8 to 9 PM are the most critical time of the day found by all incident resources. So that more facilities, and patrolling should be provided by the TMCs during those hours.

4.2.2 Spatio-temporal matching function

This part seeks to compare all four incident detection data sources Inrix, Wavetronix, Waze, and ATMS in terms of both spatial and temporal matching. The previous part of data analysis was just focusing on time of the day in which each source can be reliable in terms of number of incidents they are able to report. In this part, both location and start time are considered to find those incidents in each of four data sources reports that can be said are a good match to incidents from other sources. For matching incidents between sources, 2 matching criteria were applied, Geographic as well as temporal matching. First, all the incidents from all four different sources were combined in one CSV file. Then one incident from a source, and all the incidents in other sources which were within +/- 60 minutes interval, and 1.5 miles were identified. Once nearby incidents were identified, then to find the best possible matching incident, there should be synchrony between road name as well as direction.

4.2.3 Start Time Latency

Start time latency is defined as the difference between two incident detection sources start time. It is used as a second factor to measure the reliability of incident detection sources in detecting incidents. Kernel Density Estimation (KDE) is used to estimate the probability density function of Start time latency. This is a bit like histogram, but the idea is that it is actually a summation of placing a Kernel function on each data point. So that it is a smooth function of the probability density function of Start time. The graphs and the analysis of using this function is explained in section 5.2.
\[ f(x) = \frac{1}{N} \sum_{i=1}^{N} K(x - x_i) \]

where:

- \( f(x) \): kernel density estimator
- \( x_1, x_2, \ldots, x_n \): is a time latency variable with an unknown density \( f \).
- \( K \): is the kernel function, we used Gaussian
- \( \frac{1}{N} \): Normalizes the estimate

### 4.3 Spatio-temporal Matching

This section aims to analyze the Spatio-temporal comparison among all four sources of incident detection. The pure contribution of each source to ATMS regardless of its overlap with other sources is of interest. The final pure contribution is also containing false alarms because the incidents recorded by Waze, Inrix, and Wavetronix were not validated. To get the contribution of each source the following formula was used:

\[
S \leftarrow \{\text{ATMS, Inrix, Wavetronix, Waze}\}
\]

\[
Inc(s) = \text{Incidents detected by source } s \text{ where } s \in S
\]

\[
Match(s) = Inc(\text{ATMS}) \cap Inc(s)
\]

\[
\text{Contribution}(s) = Inc(s) - \forall t \in S - \{s\}(Inc(s) \cap Inc(t))
\]

\[
+ \forall (u,v) \in S - \{s\} (Inc(s) \cap Inc(u) \cap Inc(v))
\]

\[
- \forall t \in S Inc(t) - FP(s)
\]

In Table 4.1, the overlap incidents, using the matching function, between Waze, Inrix, and Wavetronix are shown. In Table 4.2 matching and contribution of each source with ATMS are shown. To get the matching percentage, the total number of matching incidents should be divided by the total number of incidents detected by that source. As you can see more than 50 percent of
Waze and Inrix incidents have overlap with ATMS which shows the reliability of these two sources in detecting incidents.

<table>
<thead>
<tr>
<th>Detection Source</th>
<th>Waze</th>
<th>Inrix</th>
<th>Wavetronix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waze</td>
<td>1019</td>
<td>232</td>
<td>83</td>
</tr>
<tr>
<td>Inrix</td>
<td>232</td>
<td>834</td>
<td>156</td>
</tr>
<tr>
<td>Wavetronix</td>
<td>83</td>
<td>156</td>
<td>392</td>
</tr>
</tbody>
</table>

Table 4.1: Incident Matching Matrix

As you can see in Table 4.2, Wavetronix has the lowest number of matching incidents with ATMS. The reason is due to a few numbers of Wavetronix sensors on our selected network that dont cover all the network. Sensors need to be installed in a large number to get accurate traffic information which requires deploying additional infrastructures. The high cost of deploying and implementing sensors across large networks led DOTs to consider other sources of traffic monitoring as well. Table 2 shows that more than 50 percent of Waze and Inrix incidents have overlap with ATMS, which shows the reliability of these two sources in detecting incidents. High reliable results for Inrix is directly related to its full spatial coverage all over the network as well as the availability of real-time speed data (Score 30) during all time of the day by Inrix segments.

$Contribution(s) + FP(s)$ is the potential contribution of each source to the ATMS. $FP(s)$ is the number of false alarms that are unavoidable due to the inherent errors of sources. Waze data are recorded by people which are not accurate neither in terms of time nor location. Most of the time, people who report incidents dont arrive at the incident scene exactly when it occurs. Moreover,
vehicles are moving while people are reporting, so that they cant report the exact location of the incidents. Sometime, people might simply get stuck in traffic, but wrongly report crashes. Waze reports are not reliable during less crowded hours (9 PM to 5 AM) because it is a community-based source and fewer people are traveling during those hours. Also, Inrix and Wavetronix speed report reliability is lower during low rate of volume late night hours. Inrix underestimates speeds over 60 mph, and overestimates speeds below 45 mph. Moreover, in some segments such as those that are located on interchanges speeds are typically lower. These inherent errors of Inrix and Wavetronix affect the accuracy of results from a fixed threshold speed incident detection algorithm.

### 4.4 Incidents Detection Latency

Incident detection Latency calculates the time delay between the benchmark (ATMS)incidents start time, and each of three other sources (Waze, Inrix, Wavetronix) incidents start time. Latency shows how quickly each of these three sources is able to detect incidents compares to ATMS. In section 5.1, both Inrix and Waze have the highest matching rate with ATMS. However, the matching rate is not the only measure of reliability of sources for incident detection. In other words, the ability to detect incidents for different incident detection sources is not only dependent on the overlapping with the benchmark data set (ATMS in this study), but also their accuracy in time. As discussed earlier, start time is a very important term in traffic management which avoids secondary crashes by dispatching highway helpers, and informing the en-route traffic in a timely fashion. The exact location of the incident is not much important and is not considered as the measure of reliability because the congested area caused by incidents is not a single spot, and dependent on incident severity, it spreads throughout the road around the incident location.

In this section, Kernel Density Estimation (KDE) technique was used to estimate the probability density function of start time latency. KDE is a data smoothing technique that represents our dataset into a continuous function. This helps us show the shape of start time latency distribution between two different data sources.
In Figure 4.5, the Kernel Density Estimation of start time latency between ATMS and the other three sources as well as between Waze and Inrix are shown. Throughout Figure 4.5 A to C, Start time latency between ATMS and other sources is equal to any of other sources start time minus ATMS start time. Positive latency values show that ATMS detected incidents earlier. As you can see in Figure 4.5 A, the Waze latency curve resembles the Normal distribution with a deviation of 20 minutes. This means most of the matching incidents are detected between +/- 20 minutes, and if we compare with other data sources Waze is more in-line with the ATMS. The Inrix latency curve in Figure 4.5 B is more like a uniform distribution, where the Inrix detected incidents as early as 60 minutes before, and as late as 60 minute later than ATMS. For further exploration, we need to identify the factors where Inrix can detect the incident very early, and using these factors to build a smart incident detection by integrating the Inrix data source with the ATMS. Those incidents that Inrix detected later than ATMS might not have an immediate impact on the traffic, but slowly builds up the congestion and speed drop on the roadway. The Wavetronix distribution in Figure 4.5 C is skewed towards the negative part, which tells us that most of the matching incidents by Wavetronix are detected before the ATMS.

In Figure 4.5 D, start time latency between Inrix and Waze are compared. This figure shows us that some of the incidents detected earlier by Waze and others by Inrix. To take the advantage provided by each source, we can integrate both of the sources into ATMS.
Figure 4.5: Daily average Incident frequency by different sources over hours of the day
CHAPTER 5. CONCLUSION

This work has some limitations that need to be observed. First, the location of the study is limited to urban areas where so many sensors can be found. The availability of incident detection sources used in this work, as well as good alternatives for them in less populated areas should be considered. The second limitation regards the lack of ground truth data set to verify the incidents reports especially from Waze, Inrix, and Wavetronix. For future work, images from cameras can be used to verify all the true incidents. One more limitation is that in this research, only a fixed speed threshold algorithm was used to detect incidents based on speed reports from Inrix, and Wavetronix. This method has its own limitations and errors which directly affect the number of detected incidents. For future work, developing new robust methods for detecting incidents are recommended to decrease the number of false detected incidents.

The significant goals of Traffic management and operation are to mitigate congestion, increase safety, and decrease incidents impacts such as secondary incidents, and delays. These are lofty but attainable goals require getting real-time, and reliable data from every road, all over the time. ATMS, Waze, Inrix, and Wavetronix are four valuable sources of information. ATMS and Waze give us the incidents, and traffic jams information directly. Inrix and Wavetronix are two intelligent transportation systems (ITS) technologies for the traffic monitoring purpose, and the incidents were detected by using anomaly detection algorithms on their speed reports. ATMS is the most powerful dataset to report incidents. Waze, Inrix, and Wavetronix are tracking each other closely in terms of recording incidents during a different time of the day. All four sources follow the same pattern in recording incidents. The more incident frequency happened from 7 to 8 AM, 4 to 5 PM, and 8 to 9 PM based on peaks were seen in all resources incident reports. So that the incident management and operation should be more prepared in terms of providing extra patrolling and facilities during those hours. High matching percentage of Waze and Inrix with ATMS represents
their high reliability in detecting incidents. Wavetronix shows the lowest matching with ATMS because its coverage doesn’t stay constant over the network, while Inrix segments cover all over the network. In Table 4.2, C(s)+FP(s) column represents the total number of incidents that were not recorded by ATMS. These incidents are the potential contribution of each source to ATMS, but due to the lack of validation source the exact amount of their contribution without false alarms is not specified. Incident start time latency is another factor to measure the reliability of the data sources used in this research. For Waze, the latency distribution plot’s peak happened at zero latency which affirms the fact that ATMS already uses Waze in its incident reports. Some of Waze incidents are not exactly happened at ATMS incidents start time because ATMS used another source to detect them, but still most of them happened within +/- 20 minutes of ATMS incidents. According to the previous studies which considered Wavetronix as the benchmarked events for the Inrix detected events, the average latencies were 8 min for freeways and 12 minutes for non-freeway. As you can see, even though Wavetronix has less matching with ATMS, but most of the incidents were detected earlier than ATMS compares to Inrix.
REFERENCES


