



NYSERDA

**Department of
Transportation**

Reducing Incident-Induced Emissions and Energy Use in Transportation:

Use of Social Media Feeds as an Incident Management Support Tool

Final Report



Reducing Incident-Induced Emissions and Energy Use in Transportation: Use of Social Media Feeds as an Incident Management Support Tool

Final Report

Prepared for

New York State Energy Research and Development Authority

Albany, NY

Joseph Tario
Senior Project Manager

and

New York State Department of Transportation

Albany, NY

Robert Ancar
Project Manager

Prepared by

City College of New York

Camille Kamga, Ph.D.

Stony Brook University

M. Anil Yazici, Ph.D.
Seyedamirmasoud Almotahari

and

The University Transportation Research Center, Region II

Sandeep Mudigonda, Ph.D.,

Notice

This report was prepared by City College of New York and Stony Brook University in the course of performing work contracted for and sponsored by the New York State Energy Research and Development Authority and the New York State Department of Transportation (hereafter the “Sponsors”). The opinions expressed in this report do not necessarily reflect those of the Sponsors or the State of New York, and reference to any specific product, service, process, or method does not constitute an implied or expressed recommendation or endorsement of it. Further, the Sponsors, the State of New York, and the contractor make no warranties or representations, expressed or implied, as to the fitness for particular purpose or merchantability of any product, apparatus, or service, or the usefulness, completeness, or accuracy of any processes, methods, or other information contained, described, disclosed, or referred to in this . The Sponsors, the State of New York, and the contractor make no representation that the use of any product, apparatus, process, method, or other information will not infringe privately owned rights and will assume no liability for any loss, injury, or damage resulting from, or occurring in connection with, the use of information contained, described, disclosed, or referred to in this report.

NYSERDA makes every effort to provide accurate information about copyright owners and related matters in the reports we publish. Contractors are responsible for determining and satisfying copyright or other use restrictions regarding the content of the reports that they write, in compliance with NYSERDA’s policies and federal law. If you are the copyright owner and believe a NYSERDA report has not properly attributed your work to you or has used it without permission, please email print@nyserda.ny.gov

Information contained in this document, such as web page addresses, are current at the time of publication.

New York State Department of Transportation Disclaimer

This report was funded in part through grant(s) from the Federal Highway Administration, United States Department of Transportation, under the State Planning and Research Program, Section 505 of Title 23, U.S. Code. The contents of this report do not necessarily reflect the official views or policy of the United States Department of Transportation, the Federal Highway Administration or the New York State Department of Transportation (DOT). This report does not constitute a standard, specification, regulation, product endorsement, or an endorsement of manufacturers.

1. Report No. C-14-11	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle		5. Report Date: January 2018	
Reducing Incident-Induced Emissions and Energy Use in Transportation: Use of Social Media Feeds as an Incident Management Support Tool		6. Performing Organization Code	
7. Author(s): Anil Yazici, Camille Kamga, Sandeep Mudigonda, and Seyedamirmasoud Almotahari		8. Performing Organization Report No. 55865-01-26	
9. Performing Organization(s) Name Address:		10. Work Unit No.	
University Transportation Research Center, Suite MR 910, The City College of New York, 160 Convent Avenue, New York, NY 10031		Stony Brook University Department of Civil Engineering, 2434 Computer Science, Stony Brook, NY 11794	
12. Sponsoring Agency Name and Address		13. Type of Report and Period Covered: Final Report 8/2015-6/2017	
NYS Energy Research and Development Authority 17 Columbia Circle, Albany, New York 12203		NYS Department of Transportation 50 Wolf Road Albany, New York 12232	
14. Sponsoring Agency Code			
15. Supplementary Notes: Project funded in part with funds from the Federal Highway Administration.			
16. Abstract			
<p>Ubiquitous connected devices and microblogging platforms, such as Twitter, are providing a huge amount user-generated information that has a great potential for applications in transportation incident management (TIM) with minimal infrastructure required. In this study publicly posted Twitter posts were gathered using relevant keywords. While organizational Twitter accounts (e.g., DOT, news outlets) disseminate traffic information after an incident is reported and confirmed, tweets of personal accounts are more likely to contain previously unreported traffic information, and therefore are particularly valuable for TIM. A variety of information such as location, time, severity, extent of damage, presence of debris, and evolution of congestion can be extracted from the Twitter's text. Such information is especially useful for TIM as the traditional sources for gathering traffic information, such as loop detectors and sensors, are expensive to construct and maintain for local and rural roads. Accident delay as well as emissions and fuel consumption were calculated using comprehensive incident data from California Highway Patrol to demonstrate the benefits of using Twitter for TIM. As a result of the early detection, 4,046 vehicle-hours of delay savings, reduction in 5.9 kg of ROG, 133 kg of CO, 16.3 kg of NOx and 0.3 kg of PM 2.5 and 1,939 gal of gasoline and 622 gal of diesel were estimated to be saved – total monetary value of \$75,600 i.e., \$0.5 per mile per week in California. For incidents in NYS, for each accident recorded, accident delay as well as emissions and fuel consumption were estimated in order to benchmark the potential delay and savings due to early incident detection. The study concludes with recommendations for the application of social media for TIM.</p>			
17. Key Words		18. Distribution Statement	
Incident Management; Social Media; Text mining; Emission, Delay, Fuel consumption reduction		No restrictions	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 104	22. Price

Abstract

Ubiquitous connected devices and microblogging platforms, such as Twitter, are providing a huge amount of user-generated information that has a great potential for applications in transportation incident management (TIM) with minimal infrastructure required. In this study publicly posted Twitter posts were gathered using relevant keywords. While organizational Twitter accounts (e.g., DOT, news outlets) disseminate traffic information after an incident is reported and confirmed, tweets of personal accounts are more likely to contain previously unreported traffic information, and therefore are particularly valuable for TIM. A variety of information such as location, time, severity, extent of damage, presence of debris, and evolution of congestion can be extracted from the Twitter's text. Such information is especially useful for TIM as the traditional sources for gathering traffic information, such as loop detectors and sensors, are expensive to construct and maintain for local and rural roads. Accident delay as well as emissions and fuel consumption were calculated using comprehensive incident data from California Highway Patrol to demonstrate the benefits of using Twitter for TIM. As a result of the early detection, 4,046 vehicle-hours of delay savings, reduction in 5.9 kg of ROG, 133 kg of CO, 16.3 kg of NO_x and 0.3 kg of PM 2.5 and 1,939 gal of gasoline and 622 gal of diesel were estimated to be saved – total monetary value of \$75,600 i.e., \$0.5 per mile per week in California. For incidents in NYS, for each accident recorded, accident delay as well as emissions and fuel consumption were estimated in order to benchmark the potential delay and savings due to early incident detection. The study concludes with recommendations for the application of social media for TIM.

Keywords

Incident Management; Social Media; Text mining; Emission, Delay, Fuel consumption reduction

Acknowledgements

We would like to recognize and thank the advisory committee for their support, including John Tiplado, Mike Marsico, Jaelyn Whitney of the New York City Department of Transportation, Judith Peter of the New York State Department of Transportation, and Emilio Sosa of Greenman-Pedersen. We would also like to thank staff of the Region 11 of the New York State Department of Transportation, for providing the incident and sensor data for Long Island and Gowanus Expressways. We are thankful for the support and guidance offered by project managers Joseph Tario and Robert Ancar and would like to acknowledge the assistance of Wei Hao, Nathalie Martinez, and Ellen Thorson at the University Transportation Research Center at the City College of New York. We would also like to thank Anthony Cabrera, whose valuable support in setting up computers for data collection was much appreciated, as well as the students involved in data processing, namely, Jia (Jim) Yue Mo, Jiaqi Li, Ching Fang Cheng, and Bahman Moghimidarzi.

Table of Contents

Notice.....	ii
New York State Department of Transportation Disclaimer	ii
Abstract	iv
Keywords.....	iv
Acknowledgements	v
List of Figures	viii
List of Tables.....	x
Summary	1
1 Introduction.....	1
2 Traffic Incident Information Extraction via Twitter	3
2.1 Classification Methodologies Traffic Incident Detection	4
2.2 Analysis of Data Collected Through Twitter's Public API	7
2.2.1 Preliminary Analysis.....	8
2.3 Classification of Preliminary Tweet Data	14
2.3.1 Discussion	15
2.4 Analysis of Purchased Comprehensive Twitter Data for Auxiliary Incident Management Related Information.....	16
2.4.1 Statistics on Tweets from Different Types of Accounts.....	18
2.4.2 Extracting Geographic Information.....	23
2.4.3 Extracting Debris and Other Incident Management Related Information	24
2.4.3.1 Extracting Information from Personal Accounts.....	24
2.4.3.2 Information Regarding Incidents on Local Roads.....	25
2.4.3.3 Information Regarding Debris on Roadways	26
2.4.3.4 Supplementary Information on Incidents	29
2.4.3.5 Information Regarding Incidents from Other Nonagency Sources	32
3 Accident and Traffic Data Analysis	34
3.1 Accident Duration Analysis	35
3.1.1 Data.....	35
3.1.2 Descriptive Analysis	36
3.1.2.1 Weekdays versus Weekends.....	40
3.1.2.2 Automobiles versus Heavy Vehicles.....	42
3.1.2.3 East Direction versus West Direction.....	44
3.1.2.4 Lanes Affected by Accidents.....	46

3.1.2.5	Temporal Comparison of Accident Durations	49
3.1.3	Synthesis of Accident Duration Analysis.....	53
3.2	Traffic Volume/Flow and Speed Analysis	55
3.2.1	Data.....	55
3.2.2	Descriptive Analysis	57
3.2.2.1	Volume/Flow and Speed Profiles During the Day.....	57
3.2.2.2	Decrease in Capacity During Accidents.....	60
4	Early Incident Detection and other Information Extraction using Social Media with California as a Model	62
4.1	Identification of Early Incident Detection Success	62
4.2	Success Rate for Incident Detection through Twitter Feeds.....	63
4.3	Potential of Information Contained in Incident-Related Tweets	67
5	Savings Due to Early Accident Detection	68
5.1	Delay Reduction	68
5.2	Reduced Emissions	68
5.3	Reduced Fuel Consumption.....	70
5.4	Calculated Delay, Emissions, and Fuel Savings for Early Incident Detections using California Data	71
5.5	Potential Savings through Use of Twitter Feeds in New York State.....	73
6	Conclusions and Recommendations	79
7	Important Remark: Review of Hazards Associated with Using Mobile Devices in Vehicles	81
8	Statement on Implementation	86
9	References	88

List of Figures

Figure 1. Commonly occurring words in (a) agency accounts and (b) personal accounts.....	9
Figure 2. Flowchart of the collection and cleaning of tweets from twitter	10
Figure 3. Schematic for tweet classification and geocoding	15
Figure 4. Tweets in New York metropolitan area each month	19
Figure 5. Tweets by day of week.....	19
Figure 6. Proportion of agency and personal tweets	21
Figure 7. Class of tweet from each type of account.....	21
Figure 8. Corridors identified for matching incident data with personal tweets.....	22
Figure 9. Example of geolocation information extraction	24
Figure 10. Tweet about a crash on a local road in Rotterdam, NY.....	25
Figure 11. Tweet about a crash on a local road in Rochester, NY	26
Figure 12. Tweet regarding a downed tree on Hutchinson River Parkway.....	27
Figure 13. Tweet regarding a downed tree on a local road in Westfield, NJ	28
Figure 14. Tweet regarding debris from a building in midtown Manhattan	29
Figure 15. Tweets regarding the evolution of a road blockage in Brooklyn, NY	30
Figure 16. Tweet providing information on specifics of a road closure on New York Thruway in Lackawanna, NY	31
Figure 17. Tweet providing information on road closure in Depew, NY	32
Figure 18. Tweet with information from a local business regarding incidents and road conditions	33
Figure 19. Schematic representation of traffic flow and delay during an accident ⁵⁰	34
Figure 20. Percentage of accidents occurring in each 30 minutes during a 24-hour period in GE and LIE	36
Figure 21. Duration of accidents based on time of day on GE and LIE.....	37
Figure 22. Comparison of log-normal, Gamma, Weibull and log-logistic distributions for accident durations in GE	39
Figure 23. Comparison of Log-normal, Gamma, Weibull and Log-logistic distributions for accident durations in LIE	39
Figure 24. Distribution of accident durations during weekdays and weekends in GE and LIE.....	41
Figure 25. Comparison of accident durations in weekdays and weekends	42
Figure 26. Comparison of accident durations for automobile and heavy vehicle crashes in GE and LIE.....	43
Figure 27. Comparison of durations distributions for automobile and heavy vehicle accidents ..	44
Figure 28. Accident durations with respect to direction for GE and LIE	45
Figure 29. Comparison of duration distributions for west and east directions	46
Figure 30. Accident durations with respect to blocked lane(s) in GE	47
Figure 31. Accident durations with respect to blocked lane(s) in LIE	48
Figure 32. Comparison of duration distributions for different blocked lane(s)	49
Figure 33. Temporal comparison of accident durations on GE	50
Figure 34. Temporal comparison of accident durations on LIE.....	51
Figure 35. Comparison of duration distributions for different times of day	53

Figure 36. Locations of stations at the Gowanus and Long Island Expressways	57
Figure 37. Average speed profile during 24 hours of a day at station GE1	58
Figure 38. Average volume profile during 24 hours of a day at station GE1	59
Figure 39. Theoretical speed-flow diagram	60
Figure 40. Speed-flow diagram for station GE1	60
Figure 41. Fuel economy by speed ⁶⁴	70
Figure 42. A sample accident early detection using twitter	71
Figure 43. Flow and speed during sample accident #1 at GE.....	74
Figure 44 Flow and speed during sample accident #2 at GE.....	75
Figure 45. Flow and speed during sample accident #3 at GE.....	76
Figure 46. Impact of early incident detection on delay saving with respect to varying levels of twitter feed accuracy.....	77
Figure 47 Impact of early incident detection on fuel saving with respect to varying levels of twitter feed incident detection accuracy	77
Figure 48. Impact of early incident detection on fuel saving with respect to varying levels of twitter feed incident detection accuracy in GE	78
Figure 49. Impact of early incident detection on fuel saving with respect to varying levels of twitter feed incident detection accuracy in LIE	78
Figure 50. Amber Alerts	84

List of Tables

Table 1. Normalized tf-idf scores for relevant tweets using different sets of keywords.....	13
Table 2. Tweet classification categories.....	20
Table 3. Regular expressions for geolocation information extraction.....	23
Table 4. Basic summary statistics for accident durations data.....	37
Table 5. Information about accidents with more than 300 minutes duration	38
Table 6. Comparison of log-normal, Gamma, Weibull and log-logistic distributions for accident durations in GE and LIE	40
Table 7. Comparison of accident durations in weekdays and weekends	41
Table 8. Comparison of accident durations for automobile and heavy vehicle accidents.....	43
Table 9. Comparison of accident durations for west and east directions in GE and LIE	45
Table 10. Fitted distributions for durations based on affected lane(s).....	48
Table 11. Temporal comparison of accident durations	52
Table 12. New categories based on significance of difference between old categories.....	54
Table 13. Number of days in each month for which volume and speed data is available.....	56
Table 14. Number of accidents occurring close to each station on GE and LIE.....	56
Table 15. The empirical approach used for calculating capacity drop and delay	61
Table 16. Comparison of accident-related tweets and official accident records – PeMS and SWITRS for highway and local accidents, respectively, in the state of California.....	64
Table 17. Emission factors by speed.....	69
Table 18. Delay, emissions, and fuel consumption savings due to early incident detection through twitter for incidents in the state of California.....	72
Table 19. Potential benefits of five-minute early detection for sample accident #1 at GE	73
Table 20. Potential benefits of five-minute early detection for sample accident #2 at GE	74
Table 21. Potential benefits of five-minute early detection for sample accident #3 at GE	75

Summary

This study investigates the potential use of Twitter feeds as a Transportation Incident Management (TIM) support tool for transportation management and operations. The premise of the study relies on the following two facts: 1) social media users disseminate various types of information, including traffic incidents and 2) minimal infrastructure investments are required to gather the feeds and use them for TIM purposes. Extracted traffic information from tweets can be used to inform TIM practices such as early incident detection. Utilizing incident related information (e.g., fatality, injury) can inform emergency vehicle dispatch operations and personnel who activate hazard warnings on roadways (e.g., presence of debris). One straightforward benefit is early incident clearance (as a result of early detection), which has the potential to reduce, fuel consumption, traffic delays, and emissions. In the case of life threatening accidents, early detection, as well as post-incident information regarding severity, can save lives. Furthermore, road hazard information has the capacity to prevent possible traffic incidents.

Twitter feeds gathered through a Twitter API (Application Programming Interface) were analyzed using a predefined set of keywords related to incidents on roadways. While organizational Twitter accounts (e.g., DOT, news outlets) disseminate traffic information after an incident is reported and confirmed, tweets of personal accounts are more likely to contain previously unreported traffic information, and are therefore more useful to TIM. Researchers discovered that in order to obtain useful information from personal tweets, they had to adapt to the idiosyncrasies of casually written personal tweets (e.g., non-perfect grammar, abbreviations, use of transitive verbs). Overall, the study demonstrated that incident information can be gathered from Twitter, confirming the similar findings in the literature. The conclusions for this project were presented at the Transportation Research Board's 96th Annual Meeting and published in the *Transportation Research Record Journal*.¹

To investigate the possibility of early incident detection and calculate potential delay as well as fuel consumption and emission benefits, accident and traffic count data sets from two sections of the Gowanus Expressway (GE) and the Long Island Expressway (LIE) were obtained from the New York State Department of Transportation (DOT). The accident records were matched with corresponding

¹ Yazici, M. A., Mudigonda, S., & Kamga, C. (2017). Incident Detection through Twitter: Organization vs. Personal Accounts, *Transportation Research Record* (in Press)

Traffic flow data (unless the data was missing). Accident delay, emissions, and fuel consumption were calculated for each accident record in order to benchmark the potential delay, fuel savings, and emission reduction due to early incident detection.

Using Twitter required a particular process. A search for tweets with specific keywords through Twitter's public API limits the results to about 1% of the total number of actual tweets that meet the search criteria, producing a low yield for the number of personal tweets. To address the issue of low yield, tweets matching several search queries were purchased (corresponding to the timeframe of GE and LIE accident data sets), and further analysis was performed. Although multiple accidents could be identified in the Twitter feeds, none of those accidents could be matched with an accident record in the GE and LIE data sets. To demonstrate and illustrate the benefits of using social media for TIM, more spatially- and temporally-detailed incident data from the California Highway Patrol (CHP) were used to match tweets collected in California. Using tweet and incident data from six weeks in total, 21 traffic incident tweets were matched to the recorded incidents. Three tweets were able to precede the incident reported time by 19, 23, and 4 minutes respectively. For those early detected accidents, reductions in accident delay, emissions, and fuel consumption were calculated using the flow and speed data from the Performance Measurement System (PeMS) database. As a result of the early detection, 4,046 vehicle-hours were saved. Reduction in emissions amounted to 5.9 kg of ROG, 133 kg of CO, 16.3 kg of NO_x and 0.3 kg of PM 2.5. Fuel savings amounted to 1,939 gal of gasoline and 622 gal of diesel—a total monetary value of \$75,600 or \$0.5 per mile per week.

Due to the lack of other detailed incident records in the State of New York, the potential benefits of early detection could not be illustrated as previously anticipated. Instead, potential economic and environmental savings were analyzed using hypothetical scenarios based on the percentage of accidents detected through Twitter and the level of early detection. The purchased tweet records for both New York and California were further analyzed in terms of additional traffic information content, i.e., accident severity, debris, and geographic location of incidents. Similar to the incident detection, the tweets were shown to include relevant traffic related information, which can be used to help TIM.

Overall, the study succeeds in illustrating the use of Twitter feeds for extracting incident information and provides important guidelines for future studies regarding efficient approaches to obtaining traffic information from social media. The study shows that TIM for local, rural, and less instrumented roadways can benefit from information gathered from Twitter feeds. Additionally, supplementary information on incidents can be gathered to monitor the evolution of incidents, from time created to the time cleared. On

the other hand, due to the limited number of available accident records to match accident information extracted from Twitter feeds in New York State, the potential benefits of early detection and the economic and environmental savings could not be directly quantified in the State. The calculation of the potential benefits of early detection require a comprehensive and complete traffic incident records database (including non-crash incidents, e.g., disablements), such as the CHP incident data. Potential delay and fuel and emission benefits were estimated for a sample of the CHP data to demonstrate the utility of social media for TIM. A similar effort for the State of New York will be pursued by the project team with much more extensive incident data for future research.

In the light of the findings, the following recommendations for the efficient use of Twitter feeds for gathering incident information were presented:

1. Tweets from individual accounts: Compared to already known information disseminated by organizational Twitter accounts, personal account tweets are more likely to include useful TIM information, but require more specialized search keywords to be utilized. In order to extract such significant information, queries should incorporate the relatively casual use of language and grammar by individual Twitter users.
2. Use of structured hashtags: With the provision of structured hashtags, highly specific location information can be provided to the agencies without users worrying about their privacy in providing/revealing the exact geolocation in their tweets. These structured hashtags can also provide means of defining incident type. In addition, the collection of tweet data using specific hashtags is much easier than scraping Twitter feeds for specific information which requires Twitter APIs, text mining, etc.
3. Safety Concerns: Despite some positive aspects of using Twitter as a traffic information source, there are safety concerns from Twitter users. Distracted driving is one of the major causes of traffic accidents and the New York State vehicle and traffic law for distracted driving, talking, and texting, clearly restricts the way one can use cell phones and similar smart devices in traffic. This study does not, in any way, encourage unsafe driving and traffic violations for the sake of disseminating traffic information. Such social media information dissemination should be done by nondrivers, or transmitted in a safe manner by the drivers, possibly by stopping on the side of the road and making a phone call or text transmission, or by using a hands-free mobile telephone as indicated in New York State traffic law article 33, sections 1225-c and -d. (VTL 1225-c, VTL 1225-d).
4. Partnership with data providers: It became apparent during the performance of this study that access to real-time social media data will facilitate the implementation of such a tool for TIM. A partnership for real-time, crowd-sourced data sharing between transportation agencies, such as DOT with data providers (Twitter Inc, Waze, etc.) is recommended.

1 Introduction

The increase in digital technologies, particularly internet and smart phones, creates easier ways to collect, access, and analyze large amounts of data for various purposes. Specifically, the concept of Web 2.0 and social media emphasizes user generated content, which makes every user a potential source of information, ranging from factual information to personal opinions. Businesses have been one of the first entities to embrace the potential of social media feeds, mainly for marketing and customer relations.^{1,2,3} Political organizations have also used social media to assess public support or dissent^{4,5} and social media has been cited as a gathering platform for social change.⁶

The public sector has employed user-created information for policy and planning purposes, such as social media to monitor disease outbreaks^{7,8} and to gather information during disasters.^{9,10} The potential of social media has also been recognized by transportation agencies and researchers, particularly for public transportation. A 2012 Transit Cooperative Research Program (TCRP) report titled *Uses of Social Media in Public Transportation*¹¹ is a good example of the field's response to this growing phenomenon in its early stages.

In the transportation field, information provided by social media users allow researchers/practitioners to monitor certain trends/events in real time, as well as to use the historical feed data for planning/policy purposes. Public transportation agencies use social media to collect opinions on long-term plans and policies, e.g., user satisfaction and level of service.¹¹ As another planning and policy application, social media has been utilized as a supplementary source for transportation demand surveys, which helps reduce the survey cost while increasing the reliability and accuracy of the estimates.^{3,12,13} In terms of real-time applications, large-scale events (such as disasters) receive more attention than smaller-scale events (such as traffic incidents and congestion).^{14,15} Public transportation agencies use real-time interaction with their customers during service disruptions;¹⁰ however, such interactions are in large part cultivated to disseminate information and maintain customer satisfaction. Extracting real-time information from social media and implementing tools for road transportation are still under development due to several challenges that will be discussed in following sections. Nevertheless, there are efforts to provide frameworks for such implementations^{66,16} and some applications have been tested to supplement real-time traffic information with social media-based information.¹⁷

The attractiveness of social media emerges from the ability to harvested information that is widely available and up-to-date. There is no need for expensive instrumentation or costly infrastructure investments, and software applications can handle the information collecting and reporting with negligible costs. The feature has the potential to help transportation agencies gather information from parts of the road network that lack sufficient instrumentation.

The main topics of interest in the field of transportation are traffic congestion information and traffic incidents, as one of the major contributors to nonrecurrent delay. The duration of the incident is the primary factor that determines the magnitude of the delay and thus the level of fuel waste and excess emissions. An I-95 Corridor Coalition report¹⁸ estimates that reducing an incident duration by only 5 minutes can save up to 44.5 gallons of fuel and 3.5 kg, 44.36 kg, 6.49 kg of HC, CO and NO emissions respectively, per incident. Furthermore, the probability of surviving an accident with early detection is also quantified in *The Impact of Rapid Incident Detection on Freeway Accident Fatalities*.¹⁹ Much of the literature on social media and transportation shows that tweets can be used to detect traffic incidents (and at times earlier than officially reported).^{14, 15, 20, 21, 23}

This study also analyzed accident records and traffic count data sets from the Gowanus Expressway (GE) and Long Island Expressway (LIE) in New York State and contrasted the information with purchased filtered tweets from surrounding areas to identify potential delay and measure fuel consumption and emission reductions.

The following sections of this report describe traffic incident information extraction using Twitter. A literature review is presented on the general topic of event detection with social media, along with studies on traffic incident detection in the second section. The researchers describe the preliminary data collection methods applying Twitter's public API, the data analysis, as well as a larger data collection effort and extraction from various types of specific information from Twitter accounts in the second section. The third section presents the analysis of the accident and traffic count data sets at selected corridors (Long Island and Gowanus Expressways) in order to discuss the potential economic and environmental impacts of using Twitter feeds as a traffic information source. Fourth section presents the early incident detection and likely benefits. Fifth section describes potential benefits in the State of New York. The concluding remarks are then followed by remark on using mobiles phones and a statement on implementation in the sixth, seventh and eighth sections respectively.

2 Traffic Incident Information Extraction via Twitter

Social media users continuously provide a wide range of information that make social media data feeds a part of the “big data” phenomenon. Topics of social media are often referred to as “events,” such as a concert, sporting event, or building fires. Communication scholars launch digital journalistic websites in which identified events can be used as news information sources.²⁵ Naaman et al.²⁶ coins two terms, “meformers” and “informers,” based on the nature of the posted information. Based on their Twitter feed analysis, researchers cluster 80% of the users as “meformers” who post about their opinions or self-promote. Informers who post about others constitute 20% of the analyzed user samples, and 53% of these tweets are informational in nature. Another study²⁷ reports that 40% of the tweets relate to the personal sphere. Both types of users are valuable depending on who is interested and the purpose of their analysis. For instance, the so-called “sentiment analysis” targets the personal reflection of social media users about certain products or services and can be employed by companies to measure customer satisfaction, public reaction to a new product, and other similar topics of interest. A transportation related use of such analysis is performed by public transportation agencies to gauge their customers’ view of the institution, the provided services, and reactions to service disruptions.^{28, 29} Sentiment analysis requires advanced semantics processing (such as questionnaires and surveys) and comes with challenges in identifying positive or negative feelings solely based on text. For instance, researchers need to distinguish “sarcastic” positive comments from actual positive comments.

The subject matter of the current report is mainly the informer type of user or tweets, as factual information is the kind of data needed for issues relating to traffic and accident reporting. In this perspective, the social media users ostensibly behave like “social sensors”^{30, 31} who disseminate traffic related information. Despite the factual content, there are multiple challenges to gathering the facts in an automated and efficient way. As discussed in Grant-Muller et al.,³ there is a “needle in a haystack” problem to identify relevant information from a massive amount of data. Moreover, the social media feeds are generally unstructured, written in colloquial language with no predetermined wording structure, such as abbreviations and shorthand expressions which makes it difficult to identify and extract relevant information. Ayalet Gal-Tzur et al.²⁴ also discuss the potentially ungrammatical nature of the feeds and lack of “context” as other challenges. Based on the findings of Liu et al.,³² 80% of social media data is unstructured. Ayalet Gal-Tzur et al.²⁴ argue that overlooking unstructured feeds would result in underutilization of the vast information source. In these respects, the information must be “harvested” rather than simply queried.³ However, the identification of traffic incidents exhibits an additional computational challenge when real-time “harvesting” is sought. In addition to the computational

challenge, traffic incident information requires geospatial details for the greatest potential benefit. Most social media platforms, including Twitter, require user permission to include geocoding within the post. Since there is a significant percentage of users who opt out of providing their location, researchers^{33,34} have explored methods to associate geospatial information to social media feeds with only partial success.²⁴ Despite the challenges, social media (particularly Twitter) has been successfully employed by researchers to detect traffic incidents.

Along similar lines, Pereira et al.²⁰ identify three major challenges for using social media data for ITS purposes:

- Information retrieval: “The task of obtaining the list of documents that best matches a given query.”
- Information extraction: Converting an unstructured/structured, inexplicit text into relevant information.
- Prediction: Use of extracted information to predict future transportation issues.

The third challenge, referring to predicting future traffic congestion, will not be discussed as it does not fall within the scope of this report.

2.1 Classification Methodologies Traffic Incident Detection

Event detection in Twitter has been a popular topic as the acceptance of Twitter increases. For the interested reader, Atefeh and Khreich³⁵ provide a good survey of the techniques for event detection in Twitter. Overall, event detection through text mining first requires “tokenization.” For this purpose, a set of words are deduced from the raw text data by removing punctuation, stop words (i.e., “I,” “at”), and suffixes (i.e., “s” and “-ing.” Eventually, a set of words are compiled to form a working “dictionary” for further processing and keyword selection. Words which do not appear frequently are removed from the dictionary. Keyword selection can be also done systematically with various methods, for example, entropy (representing how well a word is suited to separate documents by keyword search),⁴⁰ term frequency–inverse document frequency (tf-idf).⁴¹ The final list of keywords is used to classify the events as relevant or not depending on the desired information.. The literature includes a variety of methodologies for this purpose.³⁵

As one of the earliest studies on Twitter and traffic incident detection, Mai and Hranac²¹ analyzed more than 5 million tweets in California, extracted incident information using a Twitter API, and compared it with the California Highway Patrol (CHP) database for verification. For extraction, their query predetermined incident-related keywords (i.e., “accident,” “crash,” “wreck,” and “car”)

and additional words for connotation (i.e., “saw,” “terrible,” and “just”). They determined the relevance of tweets by using an intensity score in which the number of keywords and observational connotations were used to assign a score for relevance-ranking. These scores increased if connotations (i.e., “saw” or “just”) were also present. However, the authors did not provide the methodology of how the intensity scores were calculated. Mai and Hranac²¹ utilized geotagged tweets (1.3% of all collected tweets) with a tweet time stamp while matching harvested incidents to CHP's database. They reported that incident related tweets were posted within five hours and within 10 to 25 miles of the incident location. Although they did not provide a percentage, successful early detection of incidents was reported. Hai and Hranac²¹ argued that Twitter feeds can be more efficiently utilized if a standardized messaging format or #hashtag was used. The hastags help the ease of search and classification of information, hence the relevant information can be identified through hastags rather than complex text mining algorithms. They also pointed out that many relevant tweets include specific freeway names, which can be exploited to determine and refine the location information.

Along these lines, another study by Wanichayapong et al.³⁶ restricted the analysis to tweets with location information in Bangkok. The researchers. focused on the classification of traffic information rather than focusing solely on incidents. Their aim was to extract information that could be re-tweeted to public. Hence, the identification of “what” a tweet was about and where it was located were crucial targets. They used four different query “dictionaries” to “tokenize” tweets: place, verb, ban, and preposition. The place dictionary included an extensive list (total of 46,241 names) of roads, places, crossroads, and alleys in Bangkok. The verb dictionary included traffic related terms, similar to Hai and Kranac,²¹ but with a count of 1093 words and phrases. Since the aim was to disseminate relevant information to the public, the ban dictionary included vulgar and profane words as well as interrogatives—as questions were assumed not to be factual information. The preposition dictionary included the road direction (start and end points of roads) with a total of 192 words. They reported a high-level of accuracy for their classification methodology. The researchers discussed that after a certain number of words were in the dictionaries, further additions yielded marginal improvement. Verb words had a higher impact on tokenization than place words, and a good selection of connotations improved the prediction.

D’Andrea et al.¹⁶ provided a general framework of real-time traffic information detection via Twitter feeds. They defined an “event” to be “a real-world occurrence that happens in a specific time and space”^{35, 37} and tested different classification methods using Status Update Messages (SUMs)—basically the tweets—to identify traffic events in Italy. Their classifications were not

confined to incidents, covering nontraffic, traffic due to congestion, traffic due to crashes, and traffic due to external events (such as a sports game). The researchers employed a bag-of-words representation that broke down the text with respect to words and their frequency, and employed the following methods for classification: support vector machine (SVM), Naïve Bayes (NB) classifier, C4.5 decision tree algorithm, k-nearest neighbors (kNN) algorithm, and PART algorithm. They concluded that SVM had the best performance for classification, although other methods also yielded high accuracy.

Kurkcu et al.¹⁵ studied the feasibility of using tweets to detect incidents and incident duration. They used a set of incident related keywords to collect tweets through a Twitter API. The feed was pre-processed to eliminate mentions, replies, and retweets (to avoid duplicate information), and words such as “the” “be” and “along” were deleted to improve classification performance. The researchers employed Naïve Bayes (NB) classification to identify incident related tweets and compared it with a 511NY incident database. Their results showed that Twitter feeds can detect incidents earlier, although no percentage of success was reported.

In a paper related to the District of Columbia Department of Transportation’s implementation of social media for incident detection, Fu et al.²² developed a set of key incident related words and their association rules. The Twitter feeds were aimed at detecting incidents earlier than traditional methods and filling in missing details for incidents. The authors cited the presence of young people (who are mostly social media users), high-pedestrian traffic, and largely urbanized areas as the conditions for the implementation.³⁸ For this purpose, four influential Twitter accounts that actively post incident information were selected, and their tweets were used to determine the influential keyword set. They chose 50 keywords using “term frequency—inverse document frequency” (tf-idf). Inverse document frequency measures the importance of a word based on its frequency among words in selected tweets as well as the specificity, or the level of information each word provides. Using the tf-idf scores, the authors calculated “weights” for each keyword and ranked the tweets in terms of relevance using the sum of tf-idf score. In addition, as queries with single keywords can create “noisy” data, the authors identified association patterns so that combination of keywords (“word sets”) can be used so that more relevant tweets are identified. This study confirmed the potential benefits provided by an early detection of incidents via Twitter feeds. Data quality, inconsistency between the location of tweets and incidents, reluctance to embrace Twitter as a data source, and staff training were identified as the main challenges.³⁸

2.2 Analysis of Data Collected Through Twitter’s Public API

Extracting incident related information from raw tweets has two main components. The first is to collect tweets with the highest “potential” to be “relevant”. The second is to rank or classify the “potentially relevant” tweets and narrow them down for practical use in incident management. Gathering social media posts utilizes a public API provided by the host. Twitter provides an API that allows searching the public posts based on various criteria.⁶⁵ This allowed researchers in the study to form the raw text data by mining Twitter and scraping for public posts.

One important issue with regards to incident management is identifying whether a detected incident through social media has already been brought to the attention of agencies such as emergency response personnel. Due to information dissemination power, organizations and agencies use social media to inform the public about traffic conditions, for instance, 511 service or state and city New York State Department of Transportation (DOT) Twitter accounts. These accounts disseminate information in a structured manner (with proper grammar and spelling), which makes it relatively easy to extract necessary information. Organizational tweets are also rich in content, providing details of the incident (type of incident, location via road and/or exit number, etc.). Unlike agency accounts, individuals do not share information in a structured manner, which is one of the main challenges for event detection in social media.²⁴ Personal account holders mostly report an accident after they witness it and use active or transitive verbs and adverbs, whereas organizations use nouns and intransitive verbs. For instance, an organization tweet reads “Closed due to accident in #Summit on Rt-24 EB between X9a and I-78, stopped traffic back to X8, delay of 20 mins #traffic. As a contrast, an individual account tweets reads “Motor vehicle accident just happened in front of us!! @ Belt Pkwy.” However, organizational accounts are not as current as social media—many times conveying information already dispatched to incident management units. In this respect, individual user accounts are the main targets for traffic incident detection. In this report, based on an analysis of a sample of tweets, strategies to better identify events from the contents of targeted personal tweet accounts are discussed.

In light of the issues identified above, the investigation focused on the impact of keyword lists by analyzing the contents of messages in organization and personal accounts. First geo-tagged tweets were collected using Twitter’s API with generic keywords (listed below) related to traffic accidents in the New York metropolitan area:

accident, crash, traffic, road, freeway, highway, lane, wreck, car, cars, delay, NB, northbound, SB, southbound, EB, eastbound, WB, westbound, blocking, blocked, block, road, rd, street, st, parkway, pkwy, highway, ave, incident, collision

After the querying phase, rather than combining all tweets into a single sample, the tweets from public and private organizations providing traffic information (i.e., 511NY, NYCDOT, DOT, TotalTrafficNYC, traffic4NY, news outlet accounts) and tweets from personal accounts were manually identified and separated into two different samples. For both samples, the tweet contents were also manually coded as “relevant” or “irrelevant” with respect to incident management. Each tweet was coded with at least two annotators and the relevancy of the tweet was assigned only when two annotators agreed. A total of approximately 6,900 randomly selected tweets were used for the analysis.

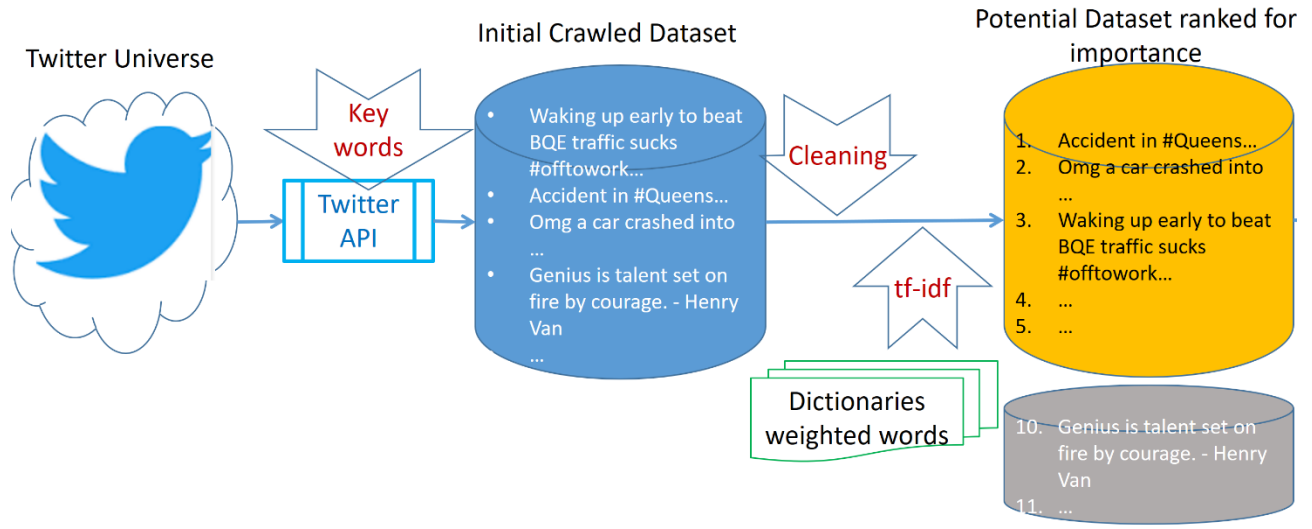
2.2.1 Preliminary Analysis

As a part of the text mining process, a series of tasks were performed using the R statistical software package. First, the raw text was converted into a corpus for text mining analysis in R., and then cleaning operations on the text were performed involving the following:

- Punctuation removal
- Conversion of text into lower case
- Removal of commonly occurring words (stop words such as "I," "the," "and," etc.)
- Stemming of words to reduce them to base words (“saw” and “seen” are converted to the word “see”)

As a preliminary step, word content was analyzed in the tweets to identify the overall frequency of words used by organization and personal accounts. For this purpose, commonly occurring words in the cleaned-up version of the text for both agency-type accounts and personal accounts are illustrated as word clouds (Figure 1). A word cloud is a powerful visual representation technique which shows the most common words along with their frequency (reflected by font size).

Figure 2. Flowchart of the collection and cleaning of tweets from twitter



As shown in Figure 1, the occurrence of the words “traffic” and “accident” in a tweet implies a higher possibility of being a relevant traffic incident tweet. The word “disabled” is a more formal word which is used by organizational accounts but personal tweets do not use this word frequently. In other words, looking for the word “disabled” is unlikely to help identify a personal tweet with relevant information. As an opposite example, the expression “just,” “got” or “omg” are found in personal tweets as individuals mostly report after they witnessed an accident, at times using exclamation words. Thus, it is more common to observe active or transitive verbs and adverbs in the tweets from personal accounts, as individuals try to report events as they experience them, e.g., “just saw an accident.” The words from organizations are more commonly nouns and intransitive verbs, e.g., “one lane blocked.” Since the organizational accounts use a more formal language, active or transitive verbs and adverbs do not appear in their tweets. The word “crash,” on the other hand, serves as a perfect example of context ambiguity in social media event detection mentioned previously. Since “crash” is one of the key words used for the Twitter API query, the sample includes multiple tweets (both organizational and personal) with the same word. When the word “crash” appears in an organizational account tweet, more often than not, it is a relevant tweet. However, individuals use the word “crash” frequently in other contexts, such as “crashing a party” or “crashing to bed.” In other words, the existence of the word “crash” does not necessarily imply as strong a relevance probability in personal accounts as it does for organizational accounts. Another example of the different nature of organizational and individual accounts is the word

“httpcocljikjqwv.” It is actually the standard truncated version of a URL (<http://t.co/cljikjqwv>) that has been subjected to some basic cleaning. The URL points to the traffic information page that shows the incident location and appears only in agency tweets with a substantial frequency. Individual accounts almost never share any supporting multimedia. These idiosyncrasies relate to the issue of differences in text structure, word selection (“tokenization”), and the difficulty of weighing those words.

One of the aims of the current project is to study the importance of the social media account type (individual and organizational) and corresponding information content. The literature on reducing traffic instances with social media reports an overall-high accuracy for various prediction methods; therefore, the motivation of the study is not to provide additional analysis in this vein. Taking these factors into consideration, two methodologies were utilized as representative approaches to serve the purpose of the study. First, a tf-idf approach was used as it is employed widely in previous literature and is also utilized by the District of Columbia Department of Transportation as one of the few real-world agency implementations. Second, a Naïve Bayesian approach was utilized in this study as the other classification methodology.

To create a word list and calculate weighs, Fu et al.²² identified a few organizational accounts and used a tf-idf approach. The set of all tweets were assigned as set T_i and then term frequency $tf(t,d)$ and $idf(t,d)$ were calculated using

Equation 1

$$tf(t, d) = \frac{f(t, d)}{\max\{f(w, d) : w \in d\}}$$

$$idf(t, D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

where $f(t, d)$ is the number of times that term t occurs in document d .

- D is a corpus of all documents
- N is the number of documents in the corpus
- $|\{d \in D : t \in d\}|$ is the number of documents where word t appears

The tf-idf scores were later used to score the tweets and tweets above a certain score cut-off were set aside as raw data. In brief, the contents of organizational account tweets were used to extract the incident related information from all tweets that were queried through the Twitter API.

In this study, to improve upon the methodology of Fu et al.²² and to understand the aforementioned differences between organizational and personal accounts, the tweets from organization and individual accounts were assigned as separate sets T_l^o and T_l^i respectively. Accordingly, the 20 and 15 words, respectively for T_l^o and T_l^i , with the highest tf-idf are identified and listed as follows:

$\left\{ \begin{array}{l} \text{ORGANIZATION ACCOUNTS} \\ \text{exit, ave, accident, lane, block} \\ \text{delay, min, pkwy, traffic,} \\ \text{right, back, stop, crash, clear,} \\ \text{close, left, vehicle, road, disable} \end{array} \right.$	$\left\{ \begin{array}{l} \text{PERSONAL ACCOUNTS} \\ \text{accident, just, car, traffic} \\ \text{got, bridge, block, crash} \\ \text{highway, thank, get} \\ \text{road, today} \end{array} \right.$
--	--

A quick observation reveals that there are words that were not used for querying in the Twitter API. Because of this, the results can be fed back to the querying phase to expand the list of relevant words and obtain a potentially more relevant set of raw tweets.

Since tf-idf scores are calculated based on the processed raw tweets, the selection of the tweet sample can affect the frequency of certain words and thus the weights and the tweet scores. Some frequent words for personal accounts do not even appear in the list for organization accounts. Due to such discrepancies, some actually relevant personal tweets can score low and be discarded. To investigate the impacts of the tf-idf weights on tweet scoring and ranking, calculated tf-idf scores were based on

- both the organization and personal account tweets
- only the organizational account tweets
- only the personal account tweets

Then tweets were ranked and compared accordingly. Since the tf-idf scores are dependent on the number of words, the absolute values do not provide clear interpretations. The tf-idf scores shown in Table 1 are normalized by dividing the tf-idf score by the total number of keywords in set $S \in \{agency, agency + personal, personal\}$ as shown.

Equation 2 $Normalized\ tfidf(S) = \frac{\sum_{for\ all\ t\ in\ d\ tfidf(t,d)}{\sum_{t \in S} t}$

This normalization helps scale the values and compare the relative importance given to each tweet scored from different keyword lists in S . For instance, tweet #1 is given much more importance than tweet #2 and even more than tweet #3 using tf-idf with only organizational keywords. However, this happens to a much lesser extent for scores estimated using the other two keyword lists.

Table 1. Normalized tf-idf scores for relevant tweets using different sets of keywords

	Relevant tweet	Account type	Using organizational + personal keywords	Using only organizational keywords	Using only personal keywords
#1	State troopers just blocked the ramps leading from route 138 in Canton onto 93 due to serious crash #WCVB	Agency	0.27	0.27	0.8
#2	Omg a car crashed into the paramus Wendy's @amandabootsy http://t.co/C4DwTElyHN	Personal	0.2	0.16	0.4
#3	@crosatto it was a bad wreck that a car went straight into the wall and went up in flames. http://t.co/XCvA7QkAF8	Personal	0.04	0	0.1
#4	car on fire on Lower level of Verrazano Bridge. 🚒🚑🚓 @ Verrazano Bridge Tolls https://t.co/lpEPEGGXWn	Personal	0.34	0	1.5

Table 1 reveals that when the set of keywords changes, the score ranking of the same tweet can also change. For example, tweet #4 is ranked the highest both under “organizational + personal” and “personal-only” categories, but ranked the lowest when scored with organizational-only keywords. These rankings are particularly important since tweets under certain cut-off values are discarded as irrelevant. It means that relevant personal account tweets #3 and #4 can be discarded when the word list is derived only from organizational accounts.

As previously noted, if one aims to detect an unknown incident, the above elimination of a personal tweet is not advised. To demonstrate the potential impacts of different tf-idf scoring, the score threshold was set as the twentieth percentile of the maximum tf-idf score estimated using each set of keywords in S . All the tweets with scores below the threshold were discarded. When only agency accounts were used for scoring, a total of 439 tweets were retained. Among those tweets, 435 (99%) were from organizational accounts and 4 (1%) were from personal accounts. The same analysis for only personal accounts yielded a total of 458 tweets, of which 409 (89%) and 49 (11%) were from organizational and personal accounts respectively. When both organizational and personal accounts were used for scoring, 469 (96%) organizational and 18 (4%) personal account tweets were retained. Overall, not incorporating personal account tweets in the scoring causes a loss of relevant personal tweets, which are potentially more important for early incident detection. However, the use of only personal accounts does not result in a dramatic decrease in the percentage of organizational tweets that are retained for relevancy.

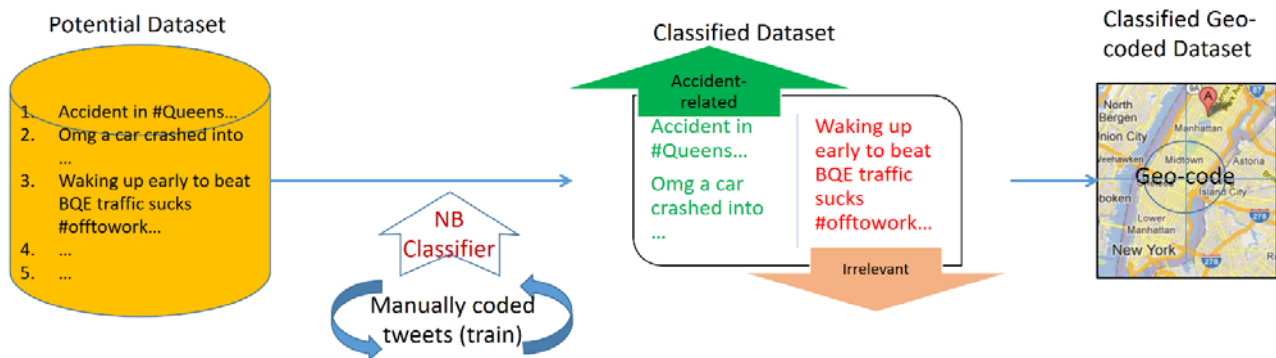
2.3 Classification of Preliminary Tweet Data

To substantiate the above discussions, a popular method like the Naïve Bayesian (NB) classification was also utilized in the study. NB classification is performed by calculating the probability of a class value c given a test document d as

Equation 3.
$$P_{NB}(c|d) = \frac{(p(c) \prod_{i=1}^k p(f_i|c))^{n_i(d)}}{P(d)}$$

In the tweet analysis context, c refers to the relevance of a tweet (e.g.. incident-related or irrelevant) and d refers to word content of a particular tweet, and $n_i(d)$ represents the count of feature or keyword f found in tweet d . Intuitively, NB assigns a relevance probability based on the presence of keywords f in the tweet. While doing this, the algorithm is trained on a sample called the training data set and the conditional contribution of each word to the relevance probability is calculated. Subsequently, the predictive capability of the classifier is tested on a test data set for validation. A schematic representation of the classification process is shown in Figure 3.

Figure 3. Schematic for tweet classification and geocoding



For the purposes of this study, NB classifiers are estimated based on organization-only (NB_{org}), personal-only NB_{per} and agency+personal (NB_{all}) accounts. The training data set is chosen from a random sample of about 5,300 tweets. The test data set is a random sample of about 1,300 tweets. Overall, all NB classifiers achieve more than 75% accuracy while classifying relevant tweets. When the NB classifiers are used to classify only personal account tweets, NB_{org} performs with 50.5% accuracy. When NB_{all} is used, the accuracy only improves to 54%. When NB_{per} is employed, 74.4% accuracy is achieved. In short, similar to the tf-idf analysis, the NB_{per} which is based on personal accounts performs the best in predicting the relevance of personal tweets, while its overall accuracy is still on par with NB_{org} and NB_{all} .

2.3.1 Discussion

For event detection in social media, such as Twitter, first the publicly available feeds are queried with a predetermined list of key words. The resulting data set can be used as raw data. However, this data set could contain many tweets that may offer low value. Alternatively, raw data are refined for further analysis of event detection using various scoring/ranking methods. The scoring methods are used to discard many tweets that have a low potential of being relevant. The scoring methods utilize a “dictionary” of frequently occurring words to weigh/score the tweets based on relevancy. The available feeds can be generated by social media accounts owned/operated by transportation agencies, news outlets, individual citizens, etc. Depending on the account type, the wording and structure of the tweets vary. The “dictionaries” derived from feeds also vary based on the type of accounts.

Some of the approaches used in the literature such as Fu et al.²² involve extracting a dictionary from a few prominent Twitter accounts and using it for classification. In this study, an alternative path is pursued. The raw data is manually processed to identify the account types and relevancy of the tweets, the dictionaries are derived, and classifications are performed accordingly. It was shown that a classification based on the dictionary derived from a few prominent accounts may not be the best method. For instance, when it comes to identifying relevant personal tweets, the classification based on a few prominent accounts performs poorer than the classification based on personal account data or a combination of the two.

In particular, Fu et al.²² used a dictionary derived from selected organizational accounts. While, this yielded a more solid dictionary with well-structured, consistent, and frequent words, a personal account dictionary included a more diverse list of words with limited consistency among the data. However, their dictionary was more likely to discard relevant personal tweets.

Mathematically speaking, this result is of no surprise as the classification algorithms perform best for the data type for which they are “trained.” However, the implications of the findings are important in terms of gathering additional information. For example, if one seeks early incident detection, the information detected from organizational feeds are more likely not to be useful as most organizations depend on the Transportation Incident and Management Center for their information. The findings imply that a “customized” dictionary will result in a more efficient classification than an analysis ignoring account types, *if* the target is early incident detection. Therefore, as a different type of trade-off, it requires extra pre-processing work to manually identify each tweet source as organizational or personal. Nevertheless, tweets from organizational accounts can be seen repeatedly in the data set and the workload decreases as more of these tweets are identified and labeled.

2.4 Analysis of Purchased Comprehensive Twitter Data for Auxiliary Incident Management Related Information

The literature on the use of Twitter for traffic incident information is mainly focused on the time of the incident (e.g., early detection) with less importance given to the identification of the incident's spatial location. Identifying spatial location is a challenging task on its own because revealing geolocation in a tweet is based on user preference. Moreover, there is no guarantee that tweet location is in close proximity to the actual incident location (users may send tweets about an incident after they reach their destinations). On the other hand, by using spatially related keywords, e.g., street and freeway names, it may be possible to infer incident location. As a matter of fact, this consideration is one of the

reasons for pursuing a more detailed investigation of keyword selection in the current study. The other reason is related to a second consideration: extracting information related to the potential causes for future incidents. Debris, dead animals, and similar roadway hazards can cause traffic accidents. If such information can be harvested from tweets, the incidents and related impacts can be prevented. In other words, aside from reducing the existing delay due to incidents, tweets can be used to eliminate delays due to future potential incidents.

The search for tweets with specific keywords through Twitter's public API limits the results to about 1% of the total number of actual tweets that meet the search criteria. Thus, the number of personal tweets from Twitter's public API has a low yield. However, Twitter offers the option of purchasing the whole set of tweets that match determined search criteria. Hence, to address the issue of low yield, tweets matching several search queries were purchased. The set of keywords chosen for framing the search queries were

accident, crash, lane, street, st, road, rt, dr, ave, us, expy, expwy, block, closed, bridge, tunnel, eb, wb, nb, sb, highway, parkway, truck, bus,

police, tow, wreck, ambl (short for ambulance), cop, cops, debris, shoulder (short)

deer, road, lane, tire, flat,

near, at, before, after, exit,

I495, FDR , RFK , Lincolntunnel , Hollandtunnel , #nyctrffic , #traffic , GSP , BQE , GWB , tunl , tpke, LIE, tpk

Though each of these keywords seem to be relevant, there could be many tweets that are irrelevant, but still contain any of these keywords. For example, a tweet such as "Genius is talent set on fire by courage. – Henry van Dyke" will be processed if the keyword 'fire' is included exclusively. For this purpose, various combinations of keywords called bigrams were used. Bigrams are combinations of two words among the keyword list, for example, "accident road," "lane closed," "street blocked." In addition to the keyword bigrams, location can also be mentioned. A location filter was expressed in two ways. The first is by setting the geolocation boundaries for the tweets. Geolocation was specified for only about 5% of tweets in general, hence, a second way of filtering location was by the inclusion of "NY, NJ, CT" in the 'place' field. The keyword bigrams and location filters were arranged according to a set format for Twitter to process the search queries. Examples of such queries are listed below:

(lane closed) OR (lane blocked) OR (road closed) OR (road blocked) OR (rd closed) OR (rd blocked) OR (street closed) OR (street blocked) OR (st closed) OR (st blocked) OR (ave closed) OR (ave blocked)) bounding_box:[-74.707528 40.396628 -74.236528 40.746628] -is:retweet

((accident bridge) OR (accident tunnel) OR (accident nb) OR (accident sb) OR (accident wb) OR (accident eb) OR (accident rd) OR (accident road) OR (accident st) OR (accident street) OR (accident highway) OR (accident parkway) OR (accident pkwy) OR (accident ave)) bounding_box:[-74.62188 40.17 -74.31152 40.40094] -is:retweet

((accident dr) OR (accident car) OR (accident lane) OR (accident lanes) OR (accident bus) OR (accident truck) OR (accident br)) bounding_box:[-74.62188 40.17 -74.31152 40.40094] -is:retweet

((accident us) OR (accident at) OR (accident near) OR (accident rt) OR (accident us) OR (accident police) OR (accident cop) OR (accident cops) OR (accident ambulance) OR (accident amb) OR (accident tow) OR (accident exit)) bounding_box:[-74.62188 40.17 -74.31152 40.40094] -is:retweet

((disabled lane) OR (disabled car) OR (disabled truck) OR (dead animal) OR (dead deer) OR I495 OR turnpike OR FDR OR RFK OR LincolnTunnel OR Hollandtunnel OR #nyctrffic OR #traffic OR GSP OR BQE OR GWB OR tunl OR tpke OR tpk OR LIE) bounding_box:[-73.29452 41.09662 -72.98526 41.450] -is:retweet

2.4.1 Statistics on Tweets from Different Types of Accounts

The tweets based on the previously mentioned keyword bigrams and location filters were queried from June 2015 to May 2016. Following the data collection, the tweets were subjected to data cleaning and tf-idf scoring as described in the Preliminary Analysis Section. Then, the tweets were filtered for a tf-idf score of five or more. The preliminary filtering of data was similar to the depiction shown in Figure 2. The resultant tweet database consisted of 150,468 tweets. These tweets plotted by month of the year and day of the week can be seen in Figure 4 and Figure 5.

Figure 4. Tweets in New York metropolitan area each month

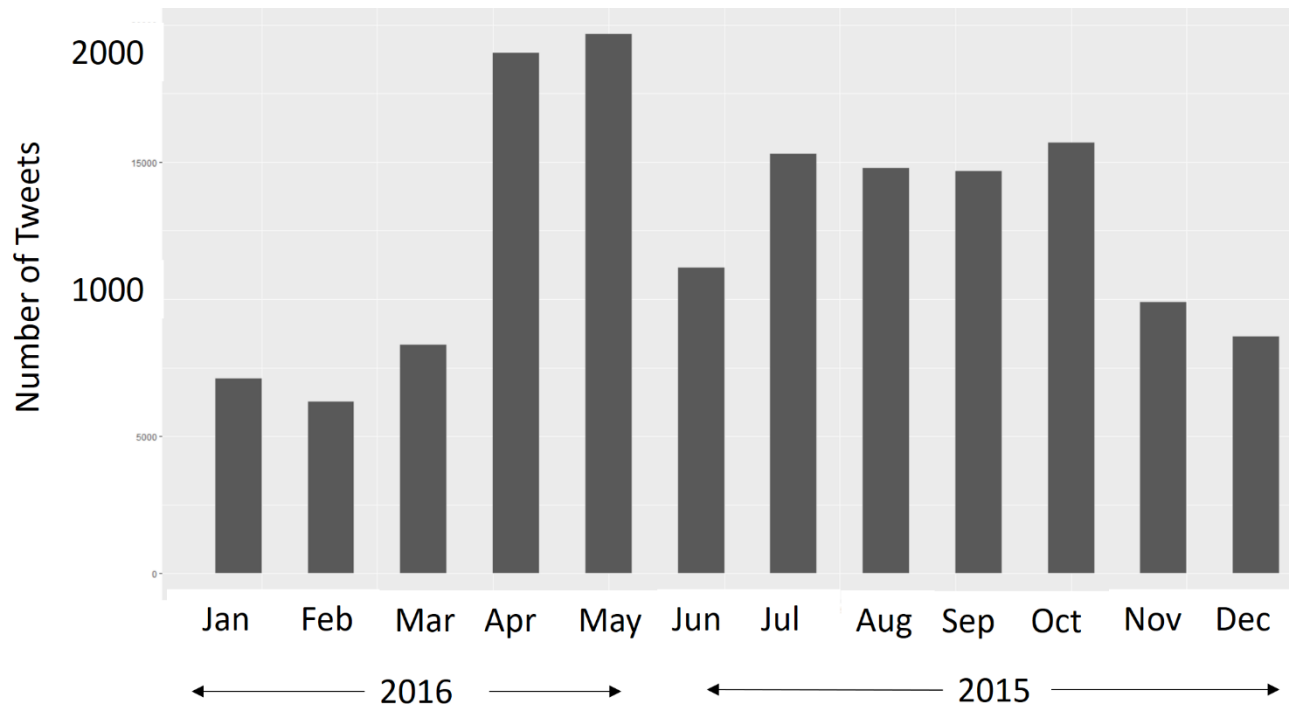
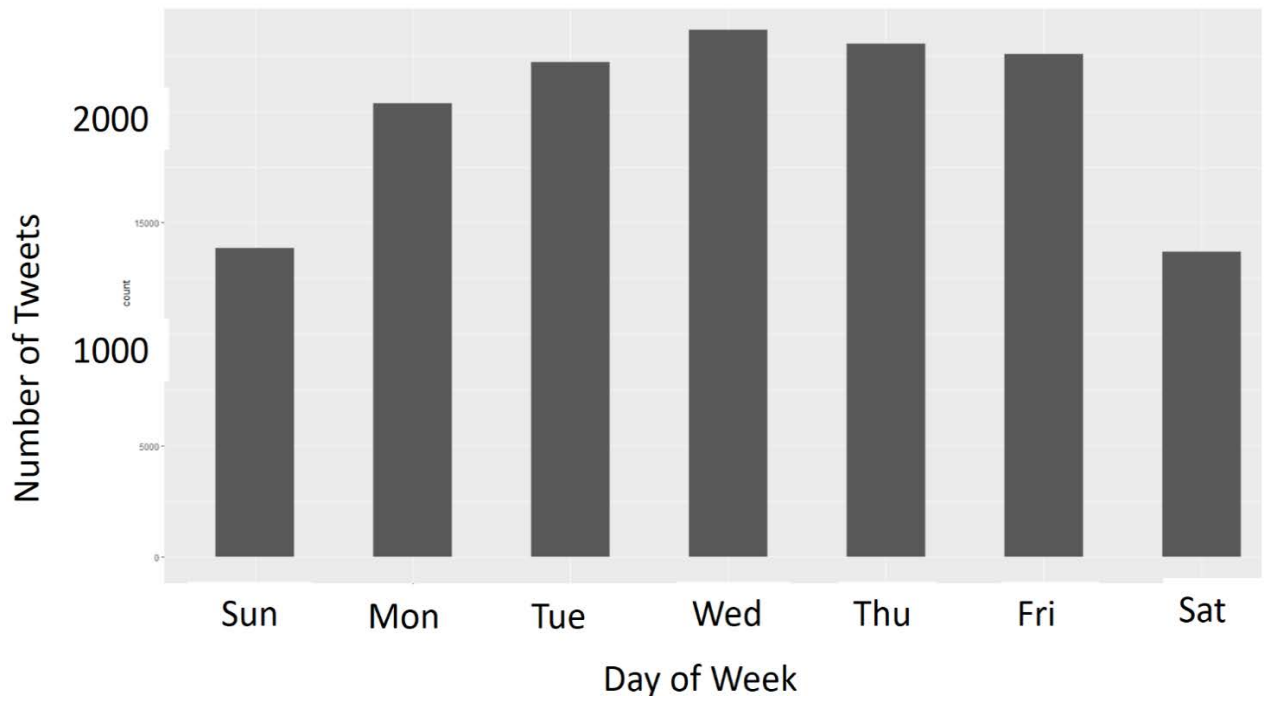


Figure 5. Tweets by day of week



Although, in this study, the tweets were collected using keyword combination bigrams, there is still a possibility that some tweets may be irrelevant. Thus, in order to study the nature of the tweets, several tweets were manually classified into the categories shown in Table 2.

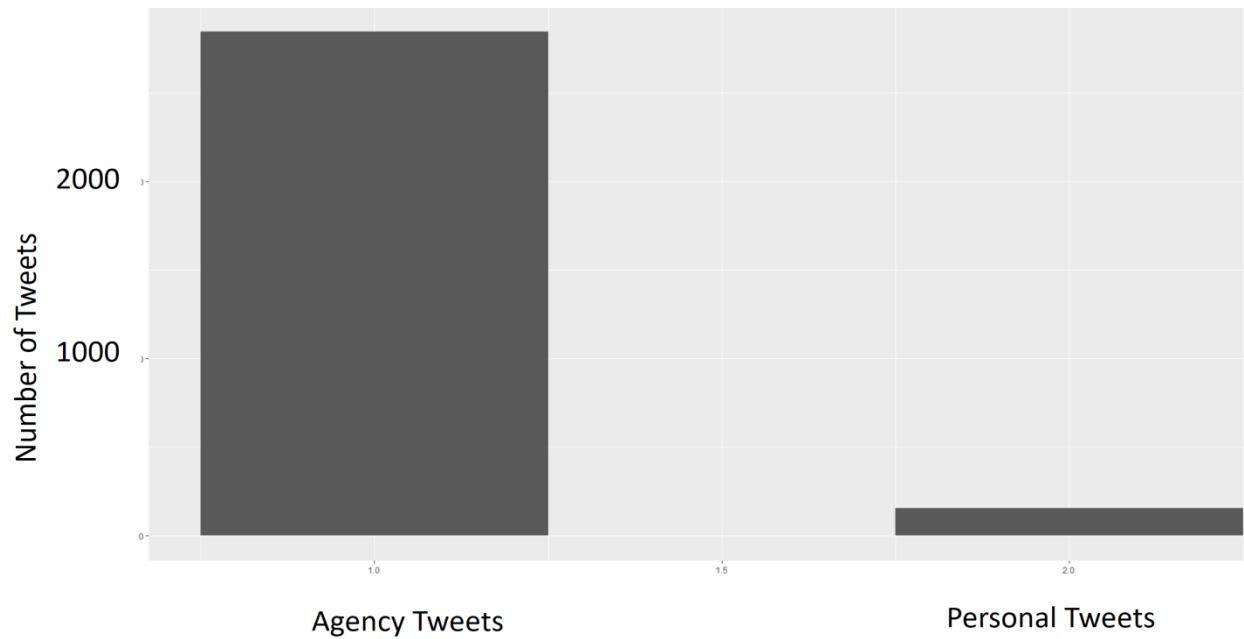
Table 2. Tweet classification categories

Account type classification	1 – agency; 2 – personal
Tweet type classification	0 – irrelevant, 1 – incident , 2 – traffic delay-related, 3 – physical pavement condition- or debris-related, 4 – other (roadwork, planned events, lane restrictions, etc.)

Two sets of randomly selected samples were generated for the purpose of classification. The first set is a general sample of 3,000 tweets. The second set is a sample of tweets that were *not* posted by commonly observed agencies. The most commonly observed agencies include 511-related accounts (@511NY, @511NYC, @511NYNJ, @511NYLongIsland, @511NYAlbany, etc.), Total Traffic-related (@TotalTrafficNYC, @TotalTrafficPHL, @TotalTrafficALB, @TotalTrafficSYR, etc.; @NYSDOT, @NYC_DOT, news agencies such as NBC, WGRZ, WBFO, etc. The purpose of the second sample is to analyze the nature of tweets posted by personal accounts, since tweets posted by personal accounts are much more sparse compared to those of agencies.

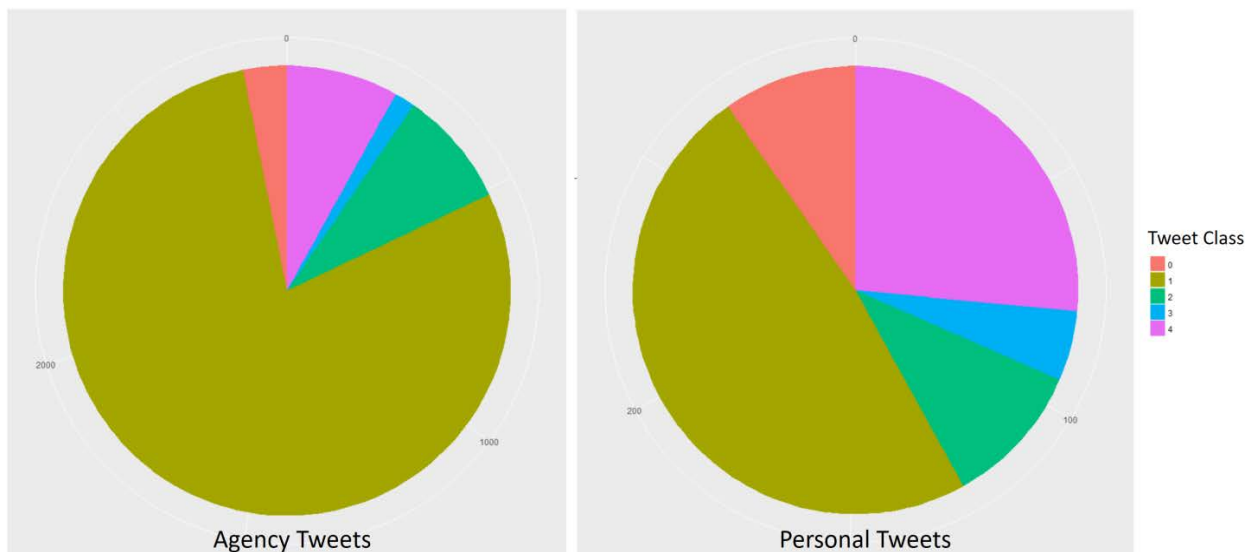
In this study, there were about 3,112 tweets from personal accounts that provide information about incidents—approximately 2% of tweets obtained after querying. In addition, the breakdown of the tweet categories of the samples is shown in Figure 6 which indicates that about 95% of the tweets in the queried data set are from agency-type accounts.

Figure 6. Proportion of agency and personal tweets



The tweet type classification shown in Table 2 was performed on both of the samples. The classification for personal tweets is based on the second sample in which tweets from most common agencies were removed. Since only a small proportion (5%) of the total data set were personal tweets, it should be noted that the second sample represented a small portion of the data set. While more than 75% of agency tweets were incident-related, less than 50% of personal tweets were incident-related as can be seen from Figure 7.

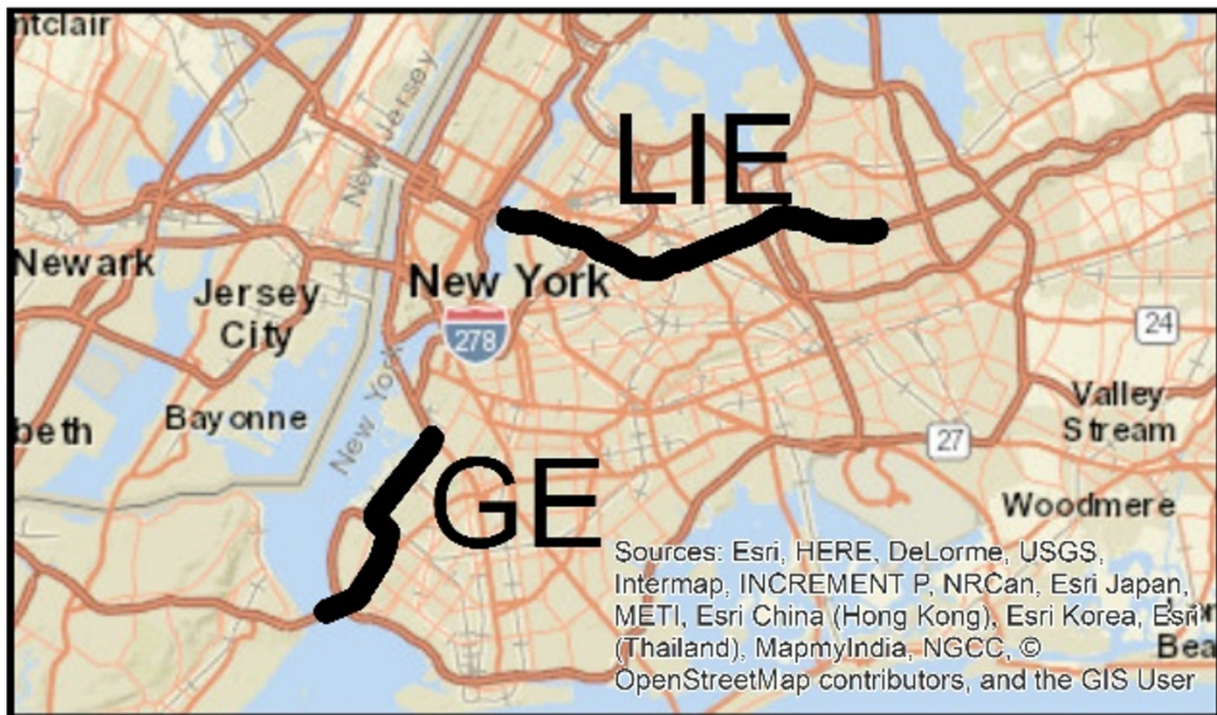
Figure 7. Class of tweet from each type of account



To make use of social media for incident detection, incident-related tweets from personal accounts have to be harvested. Therefore, it is important to know the number of incident-related tweets from personal accounts versus agency accounts. In addition, it is important to know the number of tweets from personal accounts that match any given incidents. For this purpose, the project team (within guidance and advice of the Advisory Committee) identified two corridors, the Gowanus Expressway (GE) and the eastern section of the Long Island Expressway (LIE), to analyze and match the incident data for these corridors with any personal tweets. These two corridors can be seen in Figure 8. The incident data from LIE and GE were obtained for the 12 months for which the tweet data was available – June 2015 to May 2016. However, during this period of 12 months, there were only 970 incidents in the database. The study team could not find any tweets from personal accounts that matched the incidents in the database for the Long Island and Gowanus Expressways. However, to demonstrate and illustrate the benefits of using social media for TIM, a sample of more spatially and temporally detailed incident data from the CHP was used to match tweets collected in California, which can be seen in section 4.

Despite not finding enough personal tweets to match the limited incident records available in New York State, various types of information that can be extracted from personal account tweets are discussed in subsection 2.4.3.

Figure 8. Corridors identified for matching incident data with personal tweets



2.4.2 Extracting Geographic Information

Twitter provides the option of geocoding each tweet. However, including geocoded information in each tweet is at the discretion of the user. Most accounts do not include geolocation information for privacy reasons. It is generally observed that less than 5% of all tweets include geolocation information. However, geolocation information can be extracted from the text content of the tweet. The extraction of geolocation information is performed in this study by identifying commonly occurring words that provide geolocation such as street, avenue, parkway, highway, exit, etc. These words are termed as regular expressions. Once these regular expressions are identified, names of streets, highways and exits are obtained by extracting the words adjacent to the regular expressions.

The regular expressions used in this study are shown in Table 3.

Table 3. Regular expressions for geolocation information extraction

Location	Regular Expressions
Highways	495, 278, i-495, i-278, i495, i278, #i495, #i278, I-95, I95, 95, I-80, i80, i76, i-76, i78, i-78, pkwy, between, highway, hwy, fdr, WB, EB, NB, SB, west, east, north, south, exit, bridge, rt, route, parkway, turnpike, NJTurnpike, tpk
Local roads	street, st, road, rd, ave

Two examples of extracting geolocation information are shown in Figure 9. The first tweet is “Accident cleared in #Queens on The L.I.E. WB at Douglaston Pkwy, stop and go traffic back to x34, delay of six mins #traffic.” Using regular expressions such as “WB,” “Pkwy,” “at,” and the location information such as “L.I.E. WB at Douglaston Pkwy” are extracted. By adding the entry in ‘location’ field, the location information “L.I.E. WB at Douglaston Pkwy Queens, NY” is extracted. The expression of this location information is entered into the Google Location API to obtain the longitude/latitude i.e., the geolocation information of the incident.

Figure 9. Example of geolocation information extraction

	Tweet text	Location
	Accident cleared in <u>#Queens</u> on <u>The L.I.E. WB</u> at <u>Douglaston Pkwy</u> , stop and go traffic back to x34, delay of 6 mins #traffic	Queens, NY
	@KTVU there was a high speed crash on <u>Thornton ave</u> in Newark car flipped several times before bursting into flames	Newark, CA

The second example is the tweet “@KTVU there was a high speed crash on Thornton ave in Newark car flipped several times before bursting into flames.” Using regular expressions, the location information results in “Thornton ave, Newark, CA.” This information is not sufficient to obtain an exact geolocation of the tweet. Thus, the geolocation algorithm proposed in this study, may not always result in the exact geolocation.

2.4.3 Extracting Debris and Other Incident Management Related Information

One of the main findings in the earlier sections is that gathering additional information from tweets is important. One of the aims of Fu et al.²² was to identify missing information in existing incident databases. Organizational accounts offer more details of an accident. For example, a news outlet’s after-the-fact tweet may provide details of incidents, which are not in the police report. Hence the trade-off due to discarded relevant personal tweets can be desirable.

2.4.3.1 Extracting Information from Personal Accounts

From an agency’s perspective, scraping tweets from personal accounts can yield useful information. Take for instance the several such examples from personal accounts taken from this study. Figure 7 shows that more than 80% of personal tweets provide useful information regarding, incidents, traffic delays, debris

or pavement, condition or planned events or road work. There were about 3,112 tweets from personal accounts that provided information about incidents—about 2% of the tweets that were obtained after querying.

2.4.3.2 Information Regarding Incidents on Local Roads

Many incident-related tweets posted by organizational accounts were based on information available to the traffic management center from sources such as traffic video cameras and 911 calls verified by patrols. However, for such information to be available, it is imperative that the roadways on which incidents occur be well instrumented. This may not be possible on all roadways—particularly local streets and rural roads. Examples of such tweets are shown in Figure 10 and Figure 11. These tweets are from local roads in suburban New York State, where the information and location of incidents may not be readily available to agencies via standard instrumentation. Location information for these tweets was extracted using the geolocation information extraction routine described earlier.

Figure 10. Tweet about a crash on a local road in Rotterdam, NY

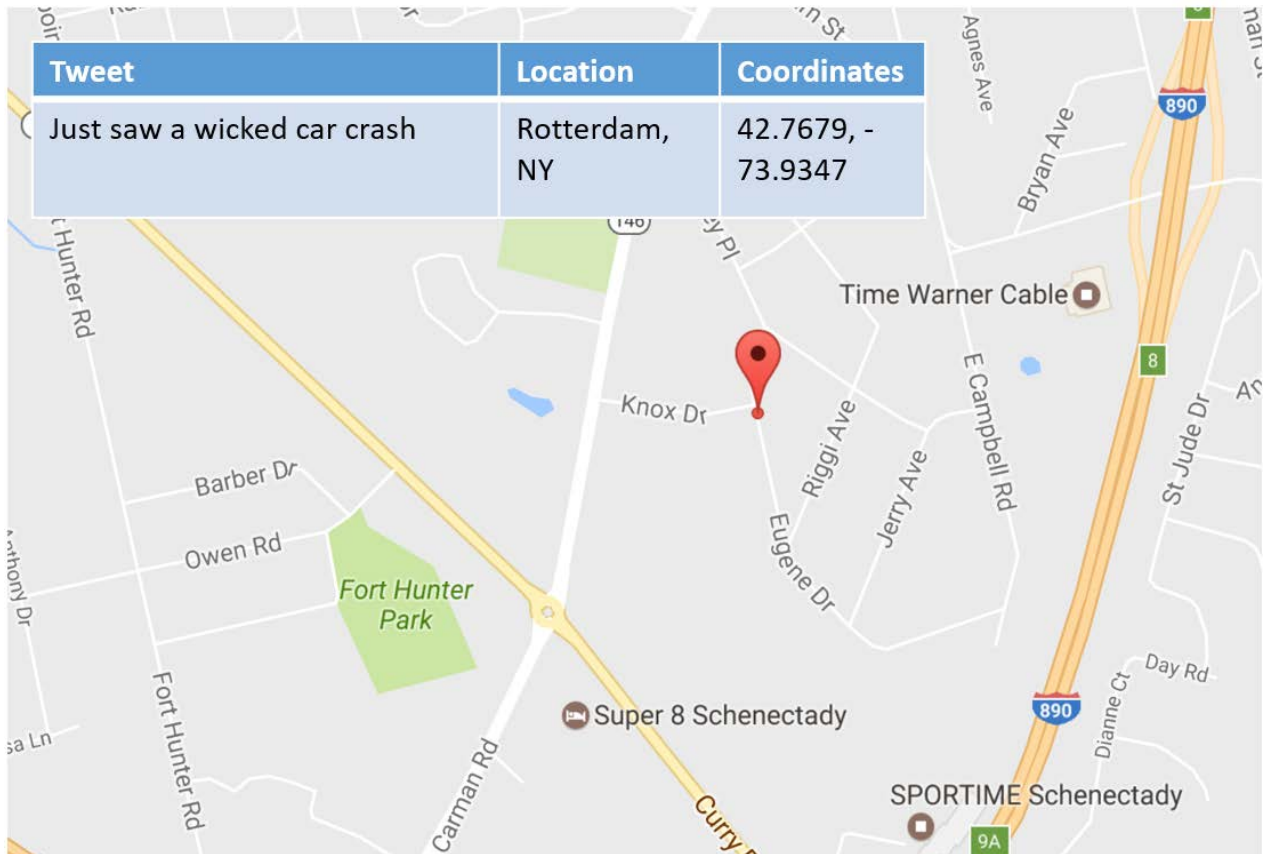
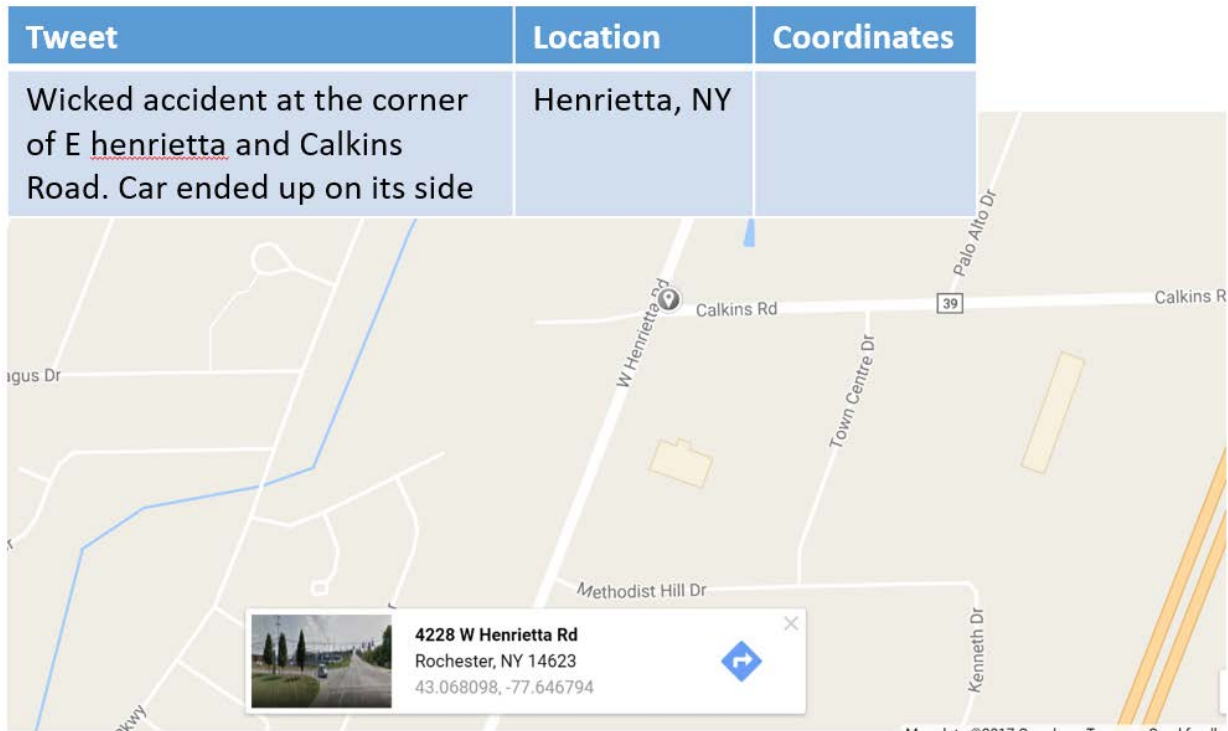


Figure 11. Tweet about a crash on a local road in Rochester, NY



2.4.3.3 Information Regarding Debris on Roadways

Some of the tweets from personal accounts provided information on debris and fallen trees on roadways. This is useful in providing additional information on the direction, location, and number of lanes that were affected. Examples of these tweets are shown in Figure 12, Figure 13, and Figure 14.

Figure 12. Tweet regarding a downed tree on Hutchinson River Parkway

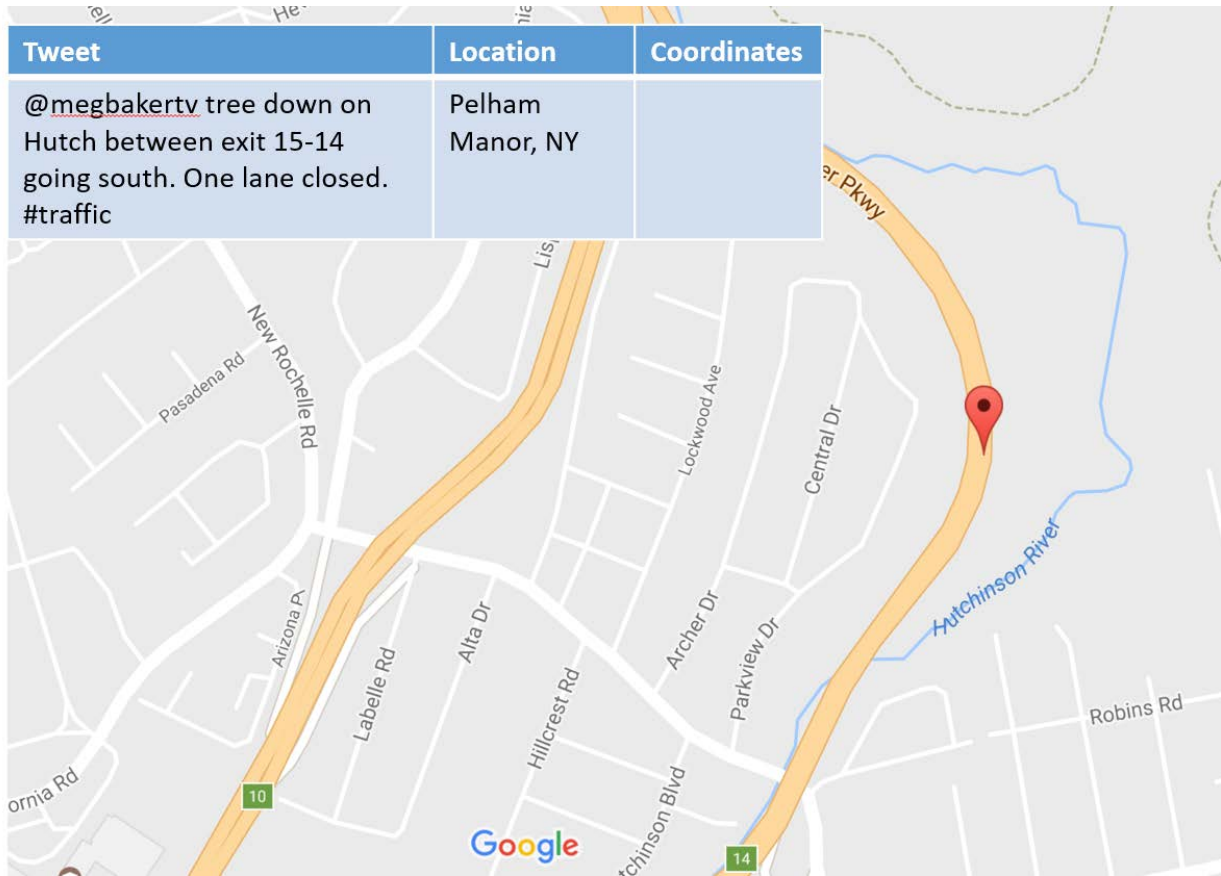


Figure 13. Tweet regarding a downed tree on a local road in Westfield, NJ

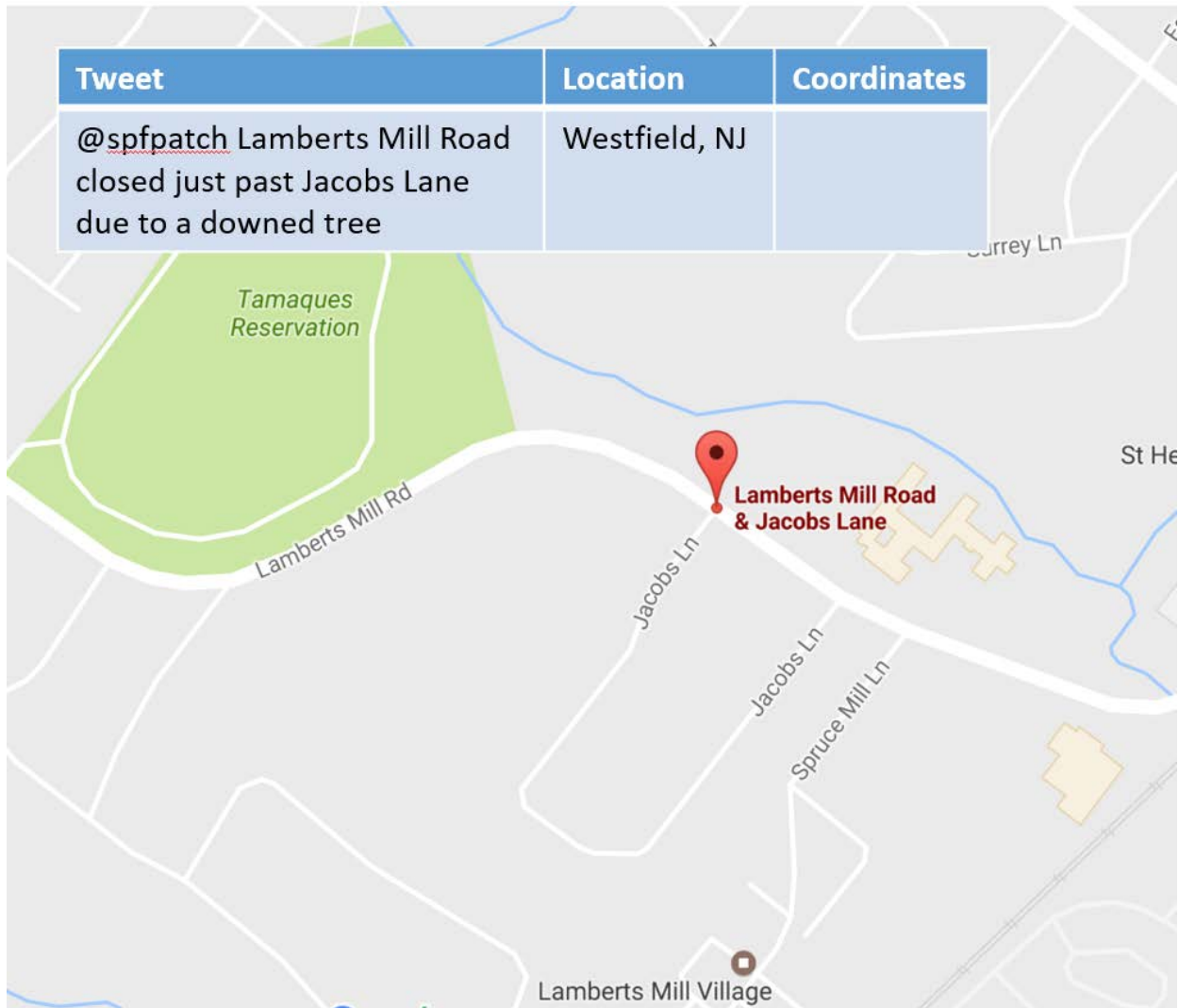
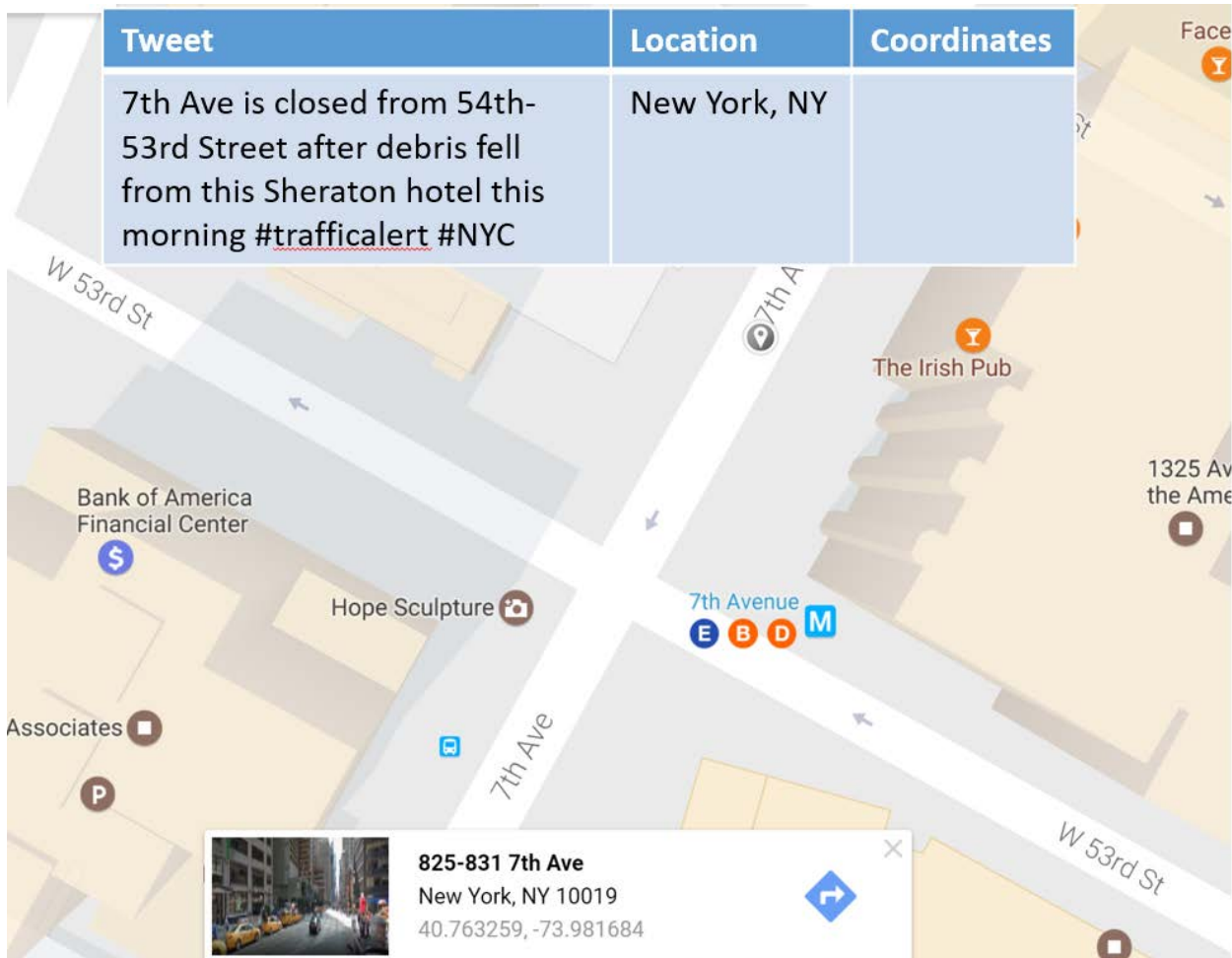


Figure 14. Tweet regarding debris from a building in midtown Manhattan



2.4.3.4 Supplementary Information on Incidents

Another important contribution of tweets from personal accounts is additional information from incidents that could be used in supplementing already available information to agencies. Furthermore, incidents evolve dynamically and information regarding their evolution can help disseminate information to users, in addition to improving the response to them. Examples of such tweets can be seen in the events of an urban road blockage (Figure 15), a road closure on a freeway (Figure 16) and a road closure and response on a local road (Figure 17).

Figure 15. Tweets regarding the evolution of a road blockage in Brooklyn, NY

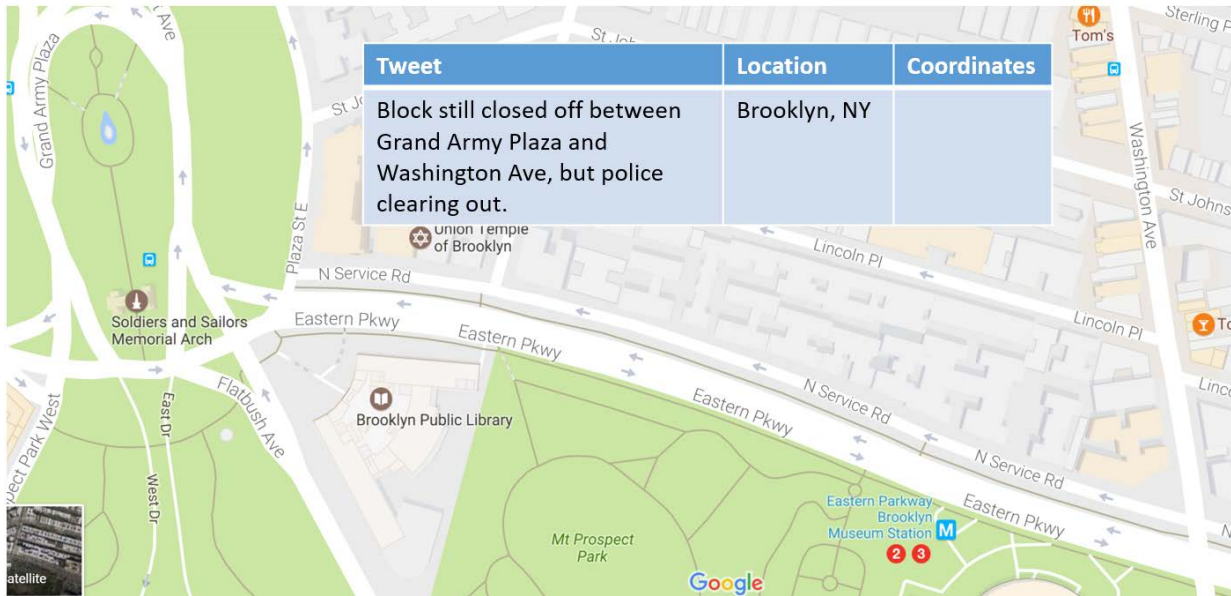


Figure 16. Tweet providing information on specifics of a road closure on New York Thruway in Lackawanna, NY

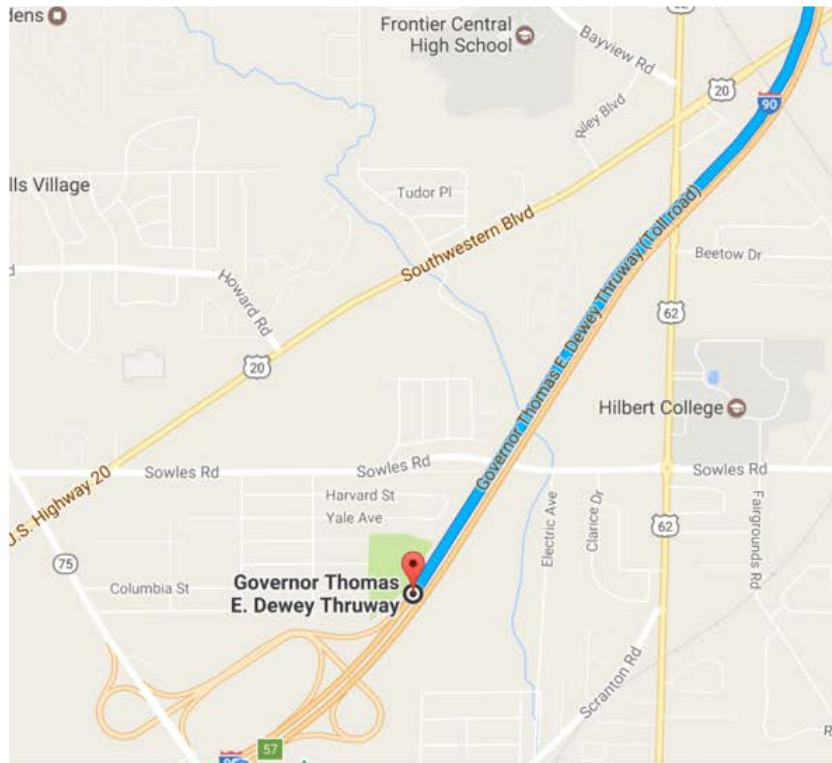
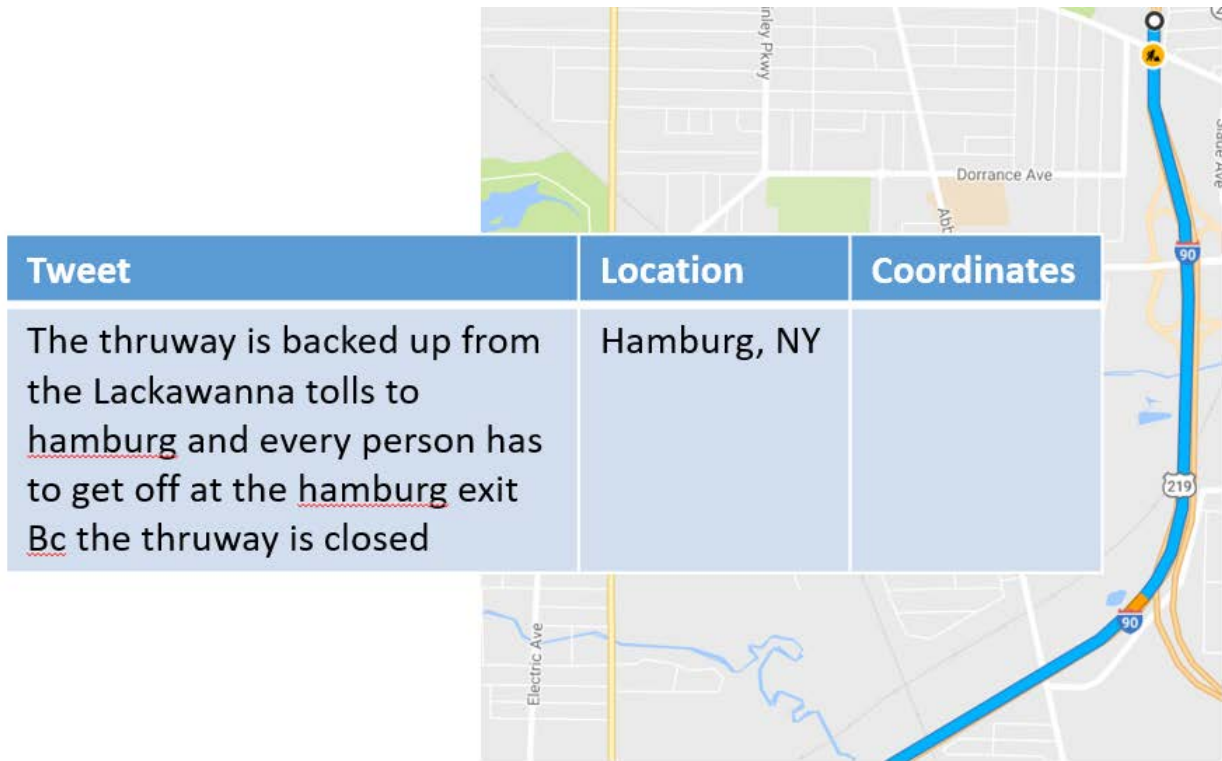


Figure 17. Tweet providing information on road closure in Depew, NY

Fire is out. Asst Chief says road will be closed for awhile. A lot of water on road, it needs to be salted. @WKBW



RETWEET 1 LIKE 1




9:12 AM - 20 Jan 2016 from Depew, NY

2.4.3.5 Information Regarding Incidents from Other Nonagency Sources

In addition to tweets from personal accounts, other sources could also provide useful and reliable information. For instance, tweets by businesses located along the route on which incidents occur can provide information regarding the severity of accidents, and can also be used to alert drivers of dynamically changing road conditions. An example is the tweet posted by a user who follows the social media feeds of a local gas station, “Everyone please be careful driving. Two cars just smashed into guard rail then another car accident 2... <https://t.co/e8mb5WtFVm>.” The gas station posted a picture of the accident and the dangerous road conditions as shown in Figure 18.

Figure 18. Tweet with information from a local business regarding incidents and road conditions

Tweet	Location	Coordinates
<p>Everyone please be careful driving. Two cars just smashed into guard rail then another car accident 2... https://t.co/e8mb5WtFVm</p>	<p>West Babylon, NY</p>	




mikessupercitgo
Mike's Super Citgo
Follow

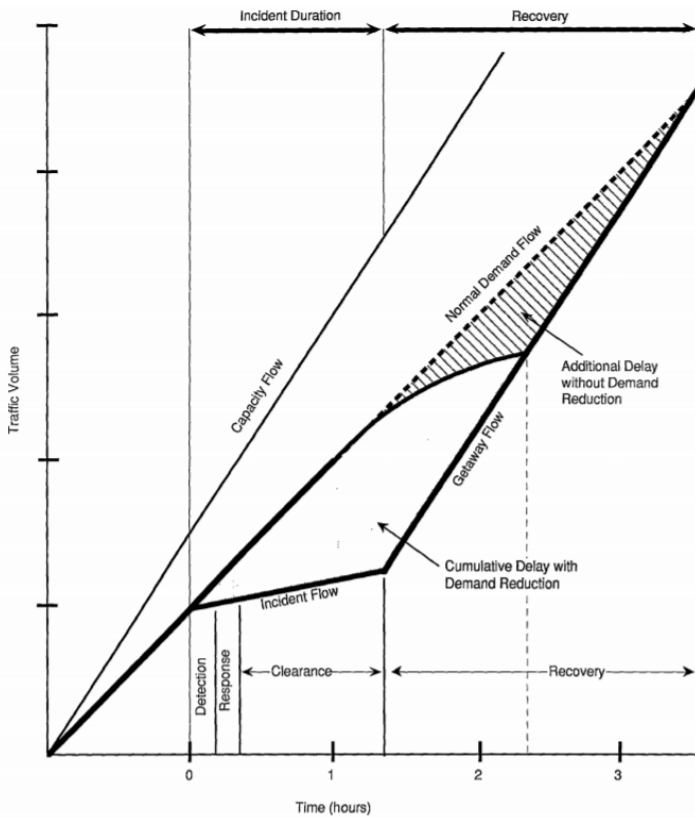
mikessupercitgo Everyone please be careful driving. Two cars just smashed into guard rail then another car accident 2 minutes later. Drive safe, save lives.

mikessupercitgo We are open 24/7 including holidays. Car wash is open daily 5am-11pm rain or shine. Deli opens daily 5am. Ask us about car wash discounts. #mikessupercitgo #westbabylon #citgo #fuelinggood #triclean #deli #e85 #carwash #fuel #gas #grocery #lotto #lottery #lunch #breakfast #dinner #snacks #diesel #carporn #coffee

3 Accident and Traffic Data Analysis

One of the goals of this project is to reduce incident-induced delay and related emissions and fuel waste through using Twitter for early detection and gathering incident specific information for emergency response. To calculate the benefits of auxiliary information obtained through a Twitter feed, there is a need to understand the existing delay conditions. For this purpose, the incident data were again used from the two previously identified corridors, namely the Gowanus Expressway (GE) and Long Island Expressway (LIE), and analyzed for the incident delay characteristics and potential improvements through Twitter information.

Figure 19. Schematic representation of traffic flow and delay during an accident ⁵⁰



DOT provided crash/accident data (not covering disablements and other non-crash incidents) for the selected corridors along with traffic flow and speed data. The accident and volume data sets were used to calculate the accident duration and the flow reduction during accidents (e.g., bottleneck capacity), which are two crucial components of delay as shown in Figure 19. The speed data was used to calculate the emissions. The following two sections provide a descriptive analysis of accident duration and traffic flow, which will then be used in the delay calculation.

3.1 Accident Duration Analysis

3.1.1 Data

The incident data provided by DOT includes only highway accidents (crashes) and covers portions of the GE (I-278) and the LIE (I-495) for years 2015 and 2016. The full data set contains 1446 accident observations with 566 records for the GE and 880 for the LIE. The data contains information, which falls into three main categories.

Temporal Data: The temporal data include the following data fields:

- Created date/time
- Year
- Month
- Cleared date/time

The temporal information is used to evaluate the accident, e.g., the difference between the cleared time and the created time gives the accident duration.

Spatial Data: The data also contains information about the location in which accidents happened.

The relevant data fields are as follows:

- Direction
- Main street
- Cross street
- Lanes affected
- Latitude and longitude coordinates

Based on the accident data set, these geographic coordinates do not show the exact location for each accident. The records show the same coordinates for each group of accidents that occurred close to each other. This could be because the coordinates are related to count stations or other specific locations on the GE and LIE. This fact must be considered for interpreting flow and speed profiles for each accident.

Vehicle Data: The data set specifies the types of vehicles that are involved in the accident. The existing categories are as follows:

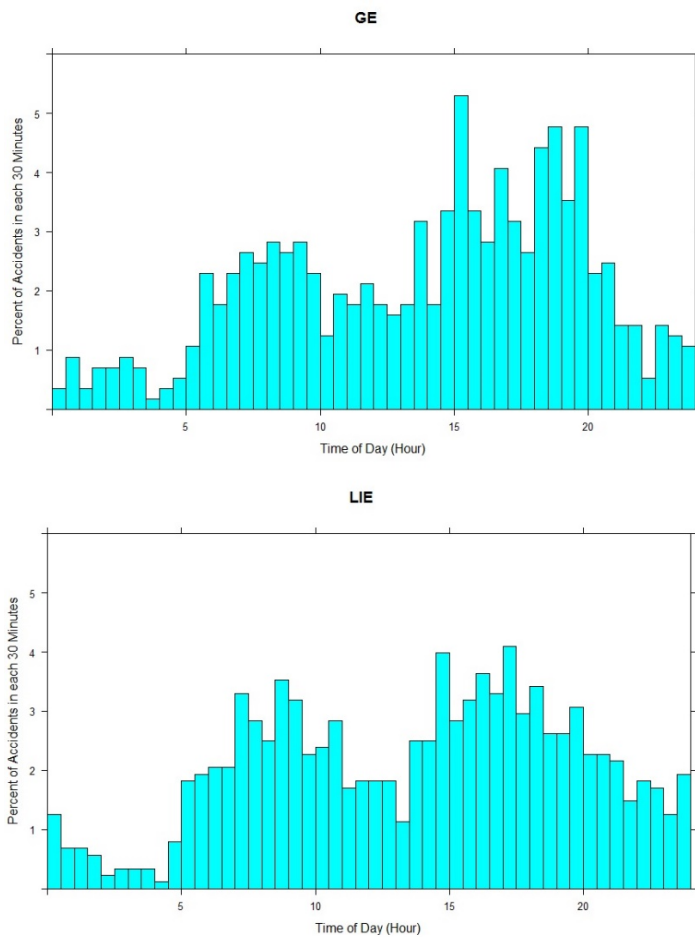
- Automobile
- Motorcycle
- Pickup van
- Bus

- Light truck
- Tractor trailer

3.1.2 Descriptive Analysis

The percentage of accidents for each 30-minute period over the 24-hour day for both the Gowanus and Long Island Expressways are shown in Figure 20. As expected, the records show that the accident frequency is higher during peak hours due to higher levels of traffic flow.

Figure 20. Percentage of accidents occurring in each 30 minutes during a 24-hour period in GE and LIE



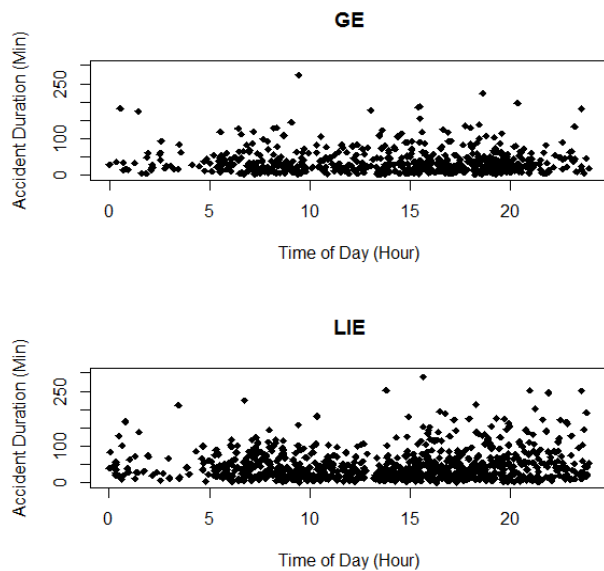
For evaluating accident induced delay, one of the most important parameters is accident duration. The accident duration is calculated by subtracting “created time” from “cleared time” for each accident record. Basic summary statistics for accident durations on the Gowanus and the Long Island Expressways is shown in Table 4. The 95% accident duration for both facilities is about 120 minutes, which implies

there are only a small number of accidents with a duration of more than two hours. In addition, the majority of accidents (75%) are cleared within one hour.

Table 4. Basic summary statistics for accident durations data

Gowanus Expressway		Long Island Expressway	
Statistic	Value	Statistic	Value
Sample Size	566	Sample Size	878
Average	41.05	Average	50.53
Std. Deviation	51.3479	Std. Deviation	65.9235
50% Percentile (Median)	28	50% Percentile (Median)	37
75% Percentile (Q3)	50	75% Percentile (Q3)	64
95% Percentile	114	95% Percentile	135.4

Figure 21. Duration of accidents based on time of day on GE and LIE



In Figure 21, accident duration for all observations is plotted with respect to the time of day. There were seven observations with a duration of more than 300 minutes, which are out of range and are not shown in Figure 21. Information about these outlier records is given in Table 5.

Table 5. Information about accidents with more than 300 minutes duration

Fac.	Dir.	Cross street	Date	Created	Cleared	Lane(s)	Veh. Inv.	Dur.
LIE	East	Woodhaven Blvd.	20-Feb-15	11:03 p.m.	11:19 p.m.	Center Lane	-	1456
GE	East	Prospect Expressway	1-Feb-15	10:14 a.m.	10:22 p.m.	Right Lane	1A	728
GE	West	33rd Street	1-Jan-16	11:04 p.m.	7:21 a.m.	Left Lane	1A	497
LIE	West	I-678	5-Feb-16	1:52 p.m.	8:59 p.m.	Left Lane	2A	427
LIE	East	I-278	22-Dec-15	1:33 a.m.	7:27 a.m.	Left Lane	-	354
GE	East	Hamilton Avenue	27-Oct-15	6:37 a.m.	12:28 p.m.	Left Lane	1LT	351
LIE	East	I-678	24-Feb-16	12:42 p.m.	6:31 p.m.	Right & Center lanes	-	349

For the accident duration analysis, the data set is divided into several categories, and the distributions of duration for each category are compared. For this purpose, distribution fitting is used to predict the probability for the frequency of occurrence of accident durations. There are many probability distributions, which can be fitted to the duration data. The Kolmogorov–Smirnov, Cramer-von Mises, and Anderson-Darling tests are used to evaluate the goodness of fit for the distributions. These tests are nonparametric tests of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution.

Different distributions, which are based on different data sources of traffic incidents, have been used to estimate and predict traffic incident duration in previous studies, including log-logistic distribution,^{51,52,53,54} Weibull distribution,^{55,56} log-normal distribution,⁵⁷ and gamma distribution.⁵⁸ In this study, the four most common distributions, namely Gamma, Weibull, log-normal and log-logistic distributions were utilized. The parameters for each distribution are calculated by the maximum likelihood estimation method, using the “*fitdistrplus*” package in R. R is an open source programming language for statistical computing that is widely used among statisticians and data miners for data analysis. The capabilities of R are extended through user-created packages, which allow specialized statistical techniques, graphical devices, import/export capabilities, reporting tools, etc. Figure 22 and Figure 23 show a comparison of these four different distributions that were used for analyzing accident durations on the Gowanus and the Long Island Expressways.

Figure 22. Comparison of log-normal, Gamma, Weibull and log-logistic distributions for accident durations in GE

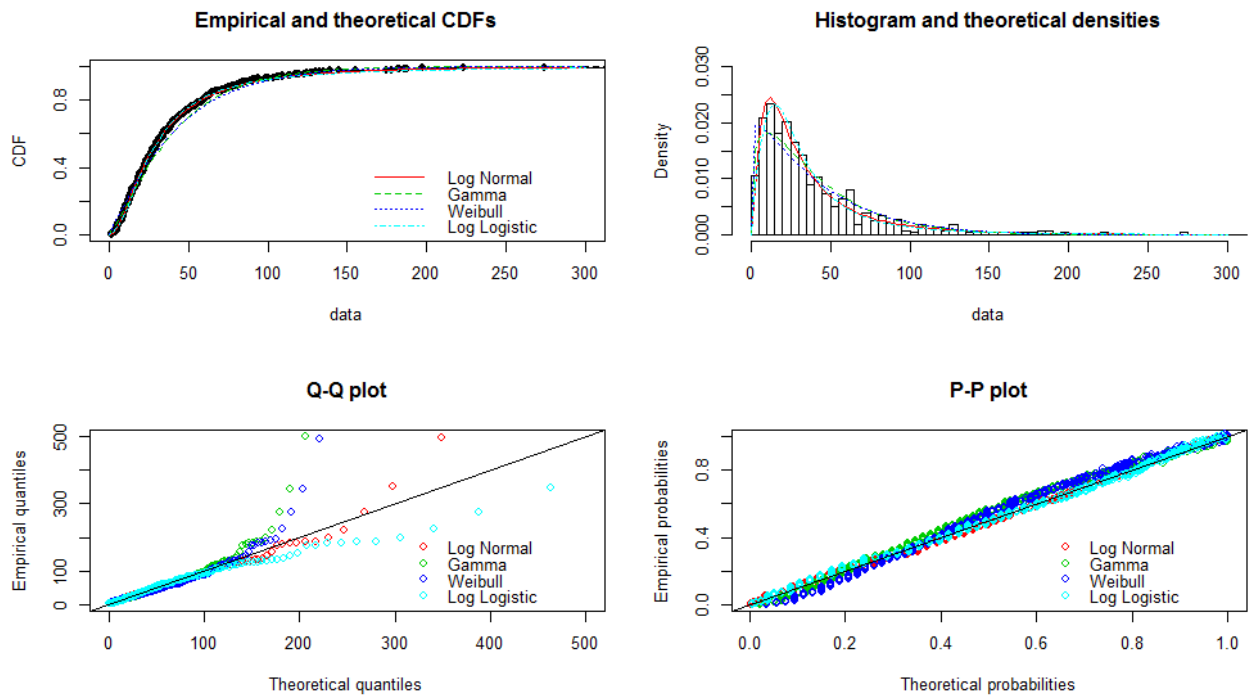


Figure 23. Comparison of Log-normal, Gamma, Weibull and Log-logistic distributions for accident durations in LIE

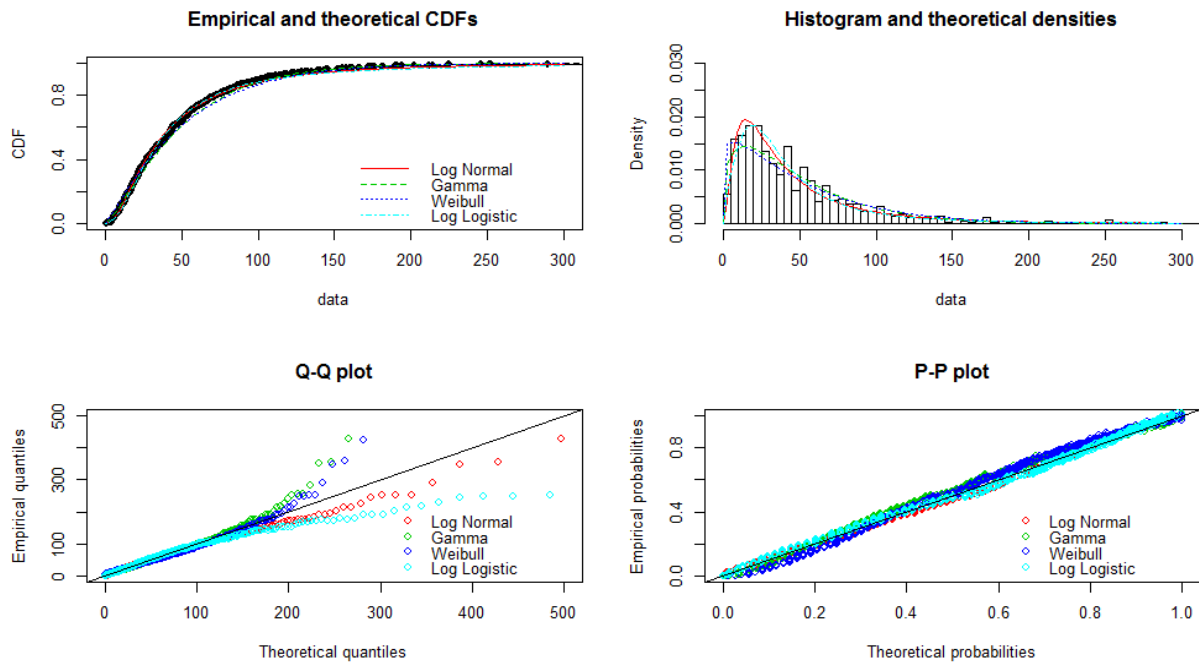


Table 6. Comparison of log-normal, Gamma, Weibull and log-logistic distributions for accident durations in GE and LIE

Facility	Distribution	Parameters		Standard Error	Kolmogorov-		
Gowanus Expressway	Log Normal	Mean log	3.2953437	0.03856471	0.03480405	0.08002155	0.54497733
		Sd log	0.9174835	0.02726922			
	Gamma	Shape	1.33480725	0.071487429	0.07213901	0.69721546	4.10684222
		Rate	0.03253171	0.002103363			
	Weibull	Shape	1.080201	0.03133465	0.07315143	0.80464283	5.99514350
		Scale	42.494863	1.75248063			
	Log Logistic	Shape	1.914413	0.06670491	0.03758173	0.10014951	0.80262284
		Scale	27.367129	1.04976002			
Long Island Expressway	Log Normal	Mean log	3.5212920	0.03097548	0.0485220	0.2535932	1.5498408
		Sd log	0.9178364	0.02190285			
	Gamma	Shape	1.38949965	0.059899987	0.04918059	0.60265877	Inf
		Rate	0.02749776	0.001420256			
	Weibull	Shape	1.104605	0.02531076	0.06296214	0.86083850	Inf
		Scale	52.756752	1.70547004			
	Log Logistic	Shape	1.922966	0.0537179	0.04269366	0.25438875	1.82765674
		Scale	34.795589	1.0676516			

As shown in Table 6, the log-normal and log-logistic distributions had a better fit in comparison with the other two distributions for the accident duration data based on goodness of fit measures. Furthermore, as illustrated in Figure 22 and Figure 23, theoretical quantiles were closer to empirical quantiles for the log-normal distribution than the log-logistic distribution. Thus, in the following sections in which accident durations are divided into different categories based on different aspects, log-normal distributions were used to compare the different categories.

3.1.2.1 Weekdays versus Weekends

In Figure 24, accident durations on weekends and weekdays are shown for each facility. The fitted log-normal distribution for each category on both facilities is also shown.

Figure 24. Distribution of accident durations during weekdays and weekends in GE and LIE

Four separate graphs show the distribution of accident durations during weekdays and weekends in GE and LIE. The fitted log-normal distributions are also represented for each category.

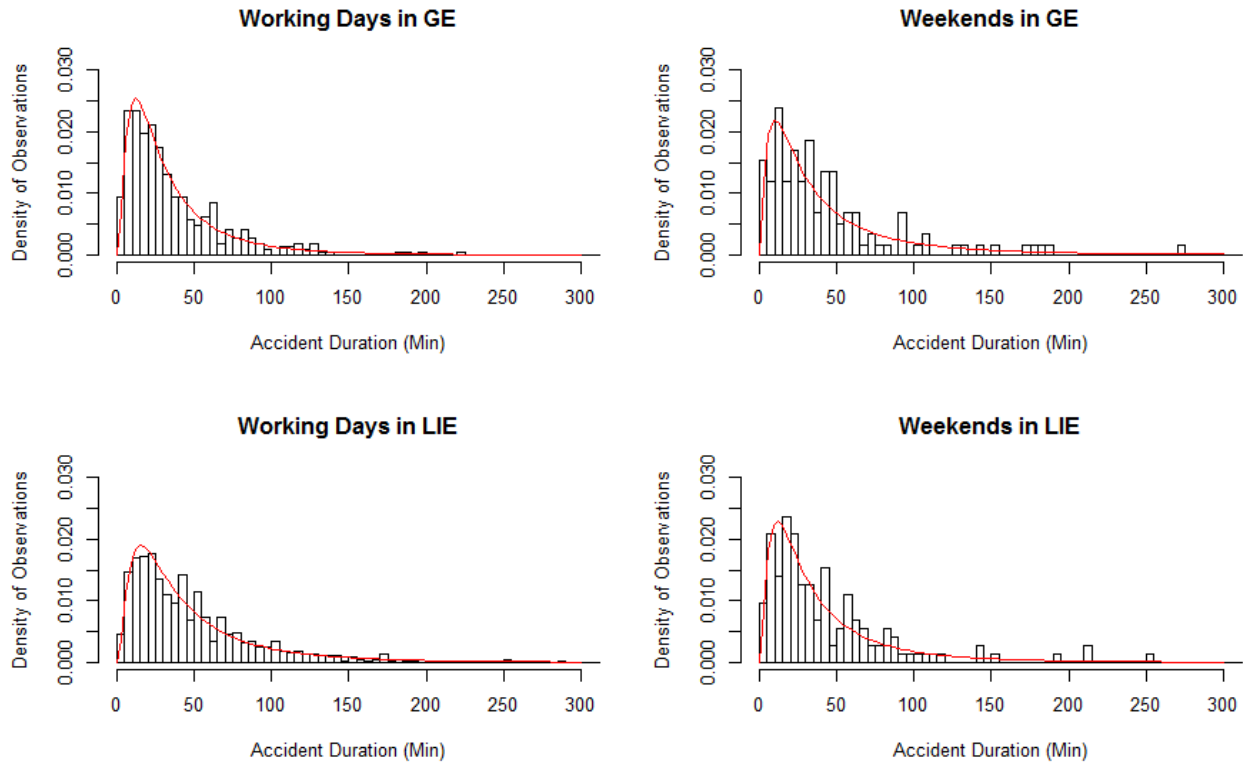
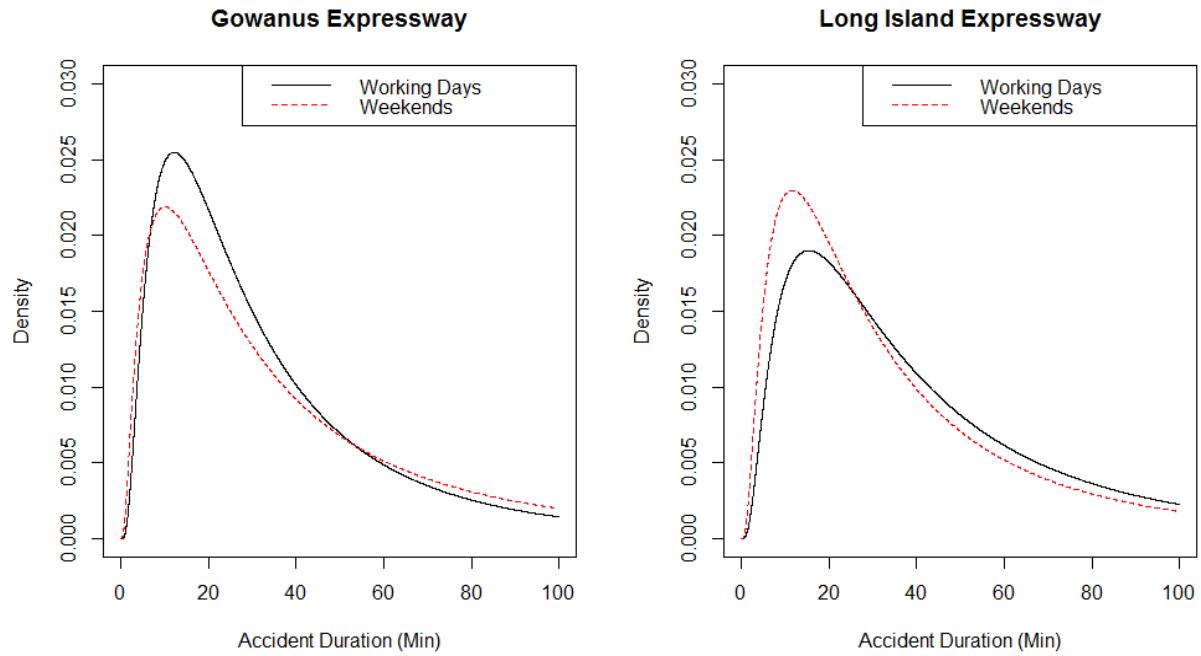


Table 7 and Figure 25 show that, on the LIE, accident durations on weekdays are longer than accident durations on weekends. On the contrary, accident durations on the GE are longer on weekends. In addition, it can be concluded that the accident durations on weekends are more dispersed, e.g., have higher standard deviations.

Table 7. Comparison of accident durations in weekdays and weekends

Facility	Category	Num. of Rec.	Actual Mean	Fitted Mean	Mean Log	SD Log	K-S Statistic
GE	Weekdays	447	38.33	26.213	3.2662833	0.8784567	0.03489509
	Weekends	118	51.45	30.100	3.404531	1.048429	0.06043293
LIE	Weekdays	734	51.97	34.936	3.5535177	0.9071777	0.04568283
	Weekends	144	43.19	28.704	3.3570306	0.9535728	0.05969318

Figure 25. Comparison of accident durations in weekdays and weekends



3.1.2.2 Automobiles versus Heavy Vehicles

Figure 26 illustrates the durations for accidents involving heavy vehicles and automobiles. A comparison of the fitted distributions for these categories is given in Table 8 and Figure 27.

Figure 26. Comparison of accident durations for automobile and heavy vehicle crashes in GE and LIE

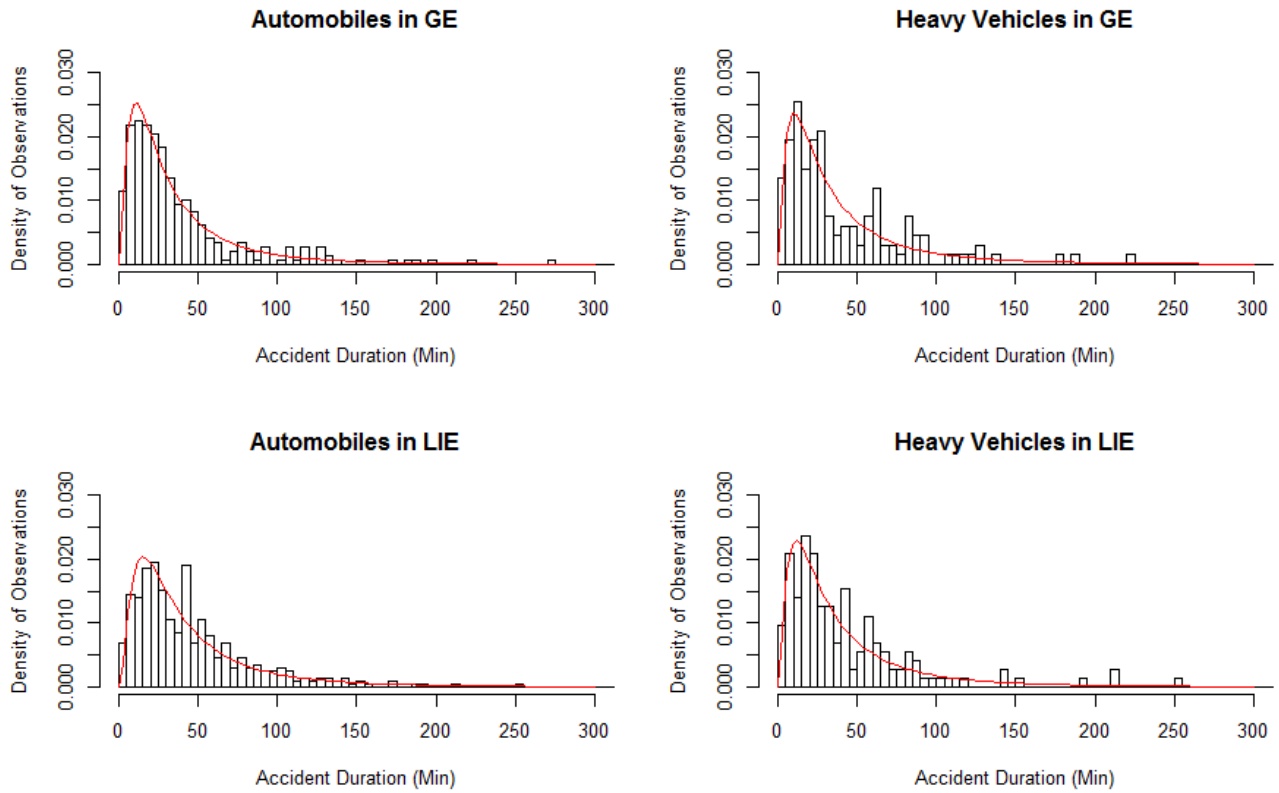
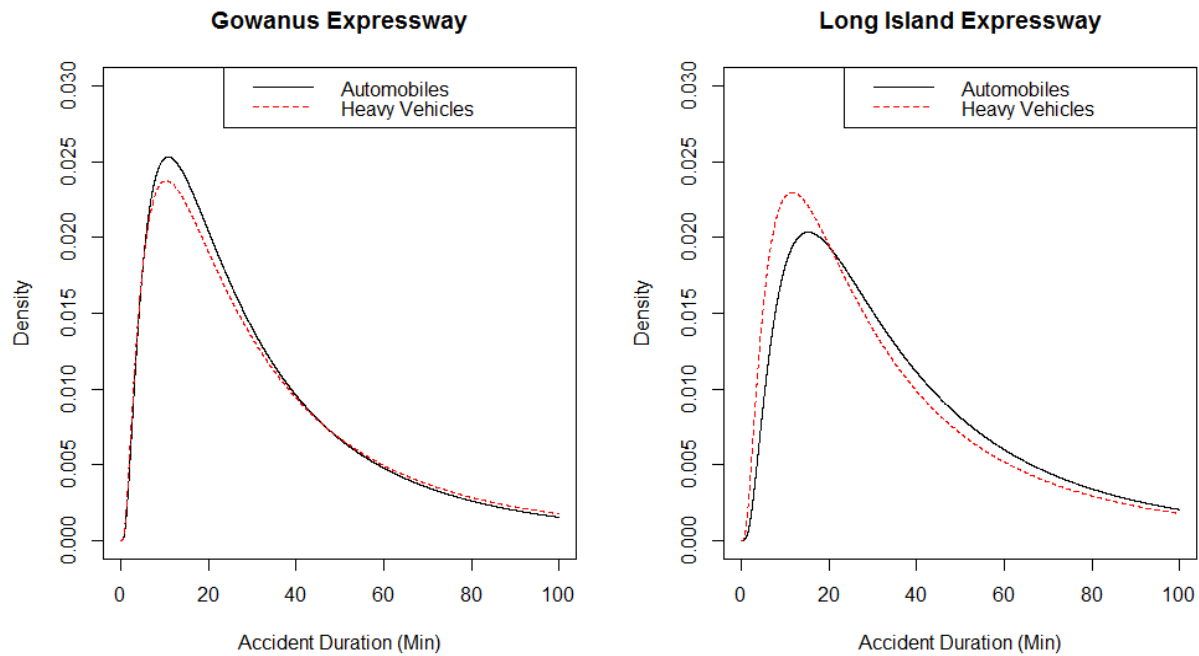


Table 8. Comparison of accident durations for automobile and heavy vehicle accidents

Facility	Category	Num. of Rec.	Actual Mean	Fitted Mean	Mean Log	SD Log	K-S Statistic
GE	Automobile	295	42.02	26.08	3.2611054	0.9413162	0.03607816
	Heavy Vehicle	134	43.92	27.71	3.3218794	0.9941494	0.06227671
LIE	Automobile	473	46.16	32.84	3.4916818	0.8773969	0.06671558
	Heavy Vehicle	144	43.19	28.70	3.3570306	0.9535728	0.05969318

Figure 27. Comparison of durations distributions for automobile and heavy vehicle accidents



Although it was expected that accidents involving heavy vehicles would generally take longer to clear in comparison with accidents involving automobiles, the analysis showed that the durations for these two categories were not so different.

3.1.2.3 East Direction versus West Direction

Figure 28, Table 9, and Figure 29 give a comparison of the accident durations for accidents occurring in the east and west bound directions.

Figure 28. Accident durations with respect to direction for GE and LIE

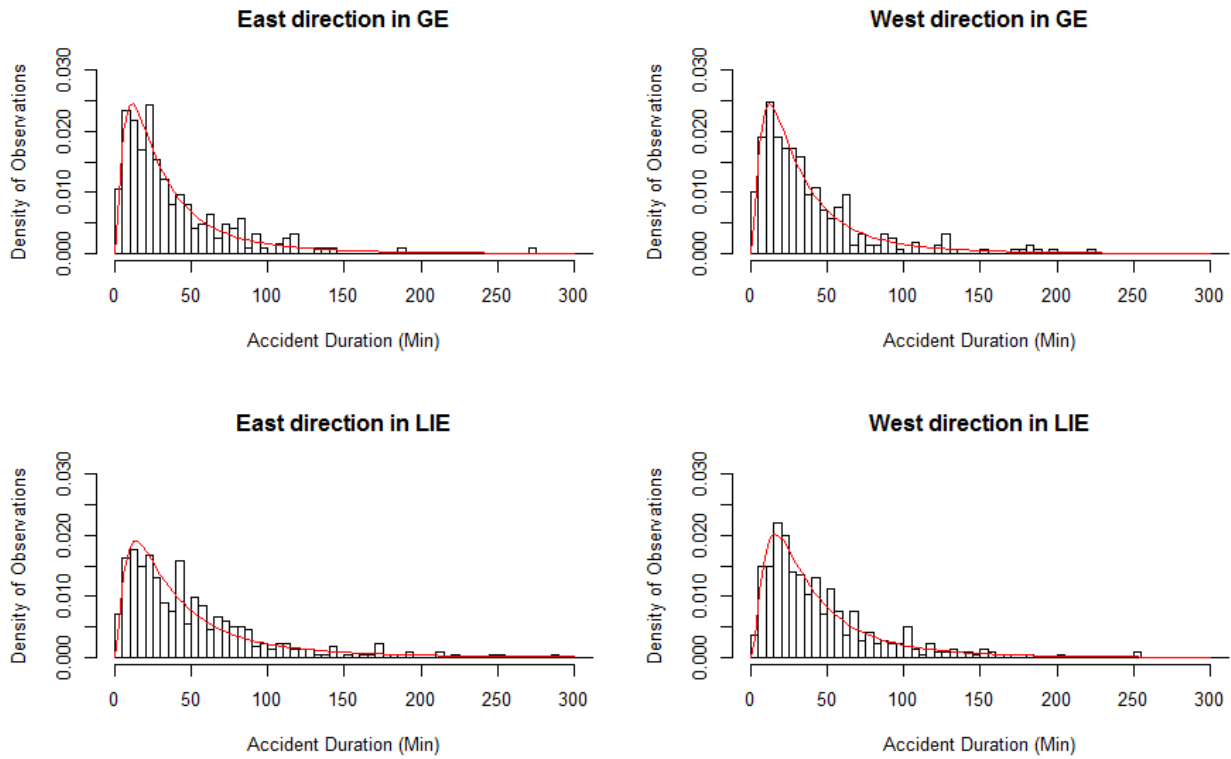
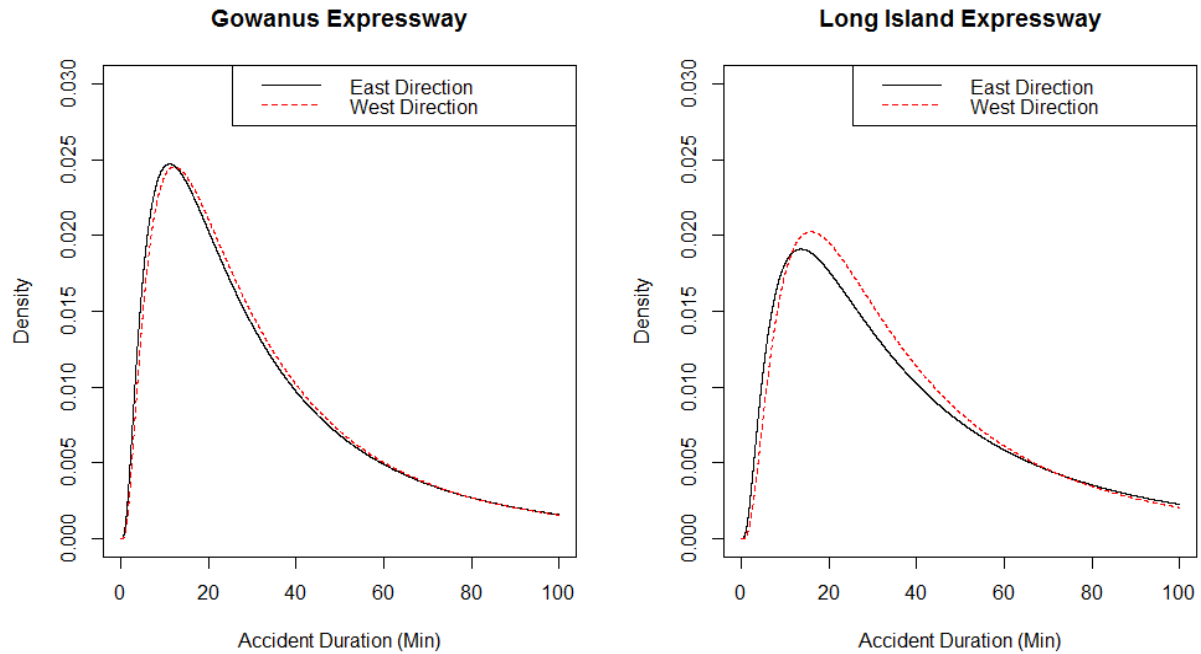


Table 9. Comparison of accident durations for west and east directions in GE and LIE

Facility	Category	Num. of Rec.	Actual Mean	Fitted Mean	Mean Log	SD Log	K-S Statistic
GE	East Direction	248	41.96	26.71	3.2850870	0.9390004	0.03983725
	West Direction	314	40.15	27.13	3.3006023	0.8955085	0.03609362
LIE	East Direction	444	54.32	34.48	3.5403938	0.9672869	0.06550542
	West Direction	429	46.52	33.18	3.5018411	0.8597135	0.03421397

Figure 29. Comparison of duration distributions for west and east directions



From this data, the researchers in the study concluded that the accident durations for both directions had similar characteristics.

3.1.2.4 Lanes Affected by Accidents

Differences between accident durations for different blocked lanes are shown in Figure 30 and Figure 31. As shown in Table 10 and Figure 32, the accident duration was higher for accidents that blocked all lanes. Furthermore, accidents in the right lane had shorter durations than accidents in the left lane, but longer durations in comparison to accidents that blocked the center lane.

Figure 30. Accident durations with respect to blocked lane(s) in GE

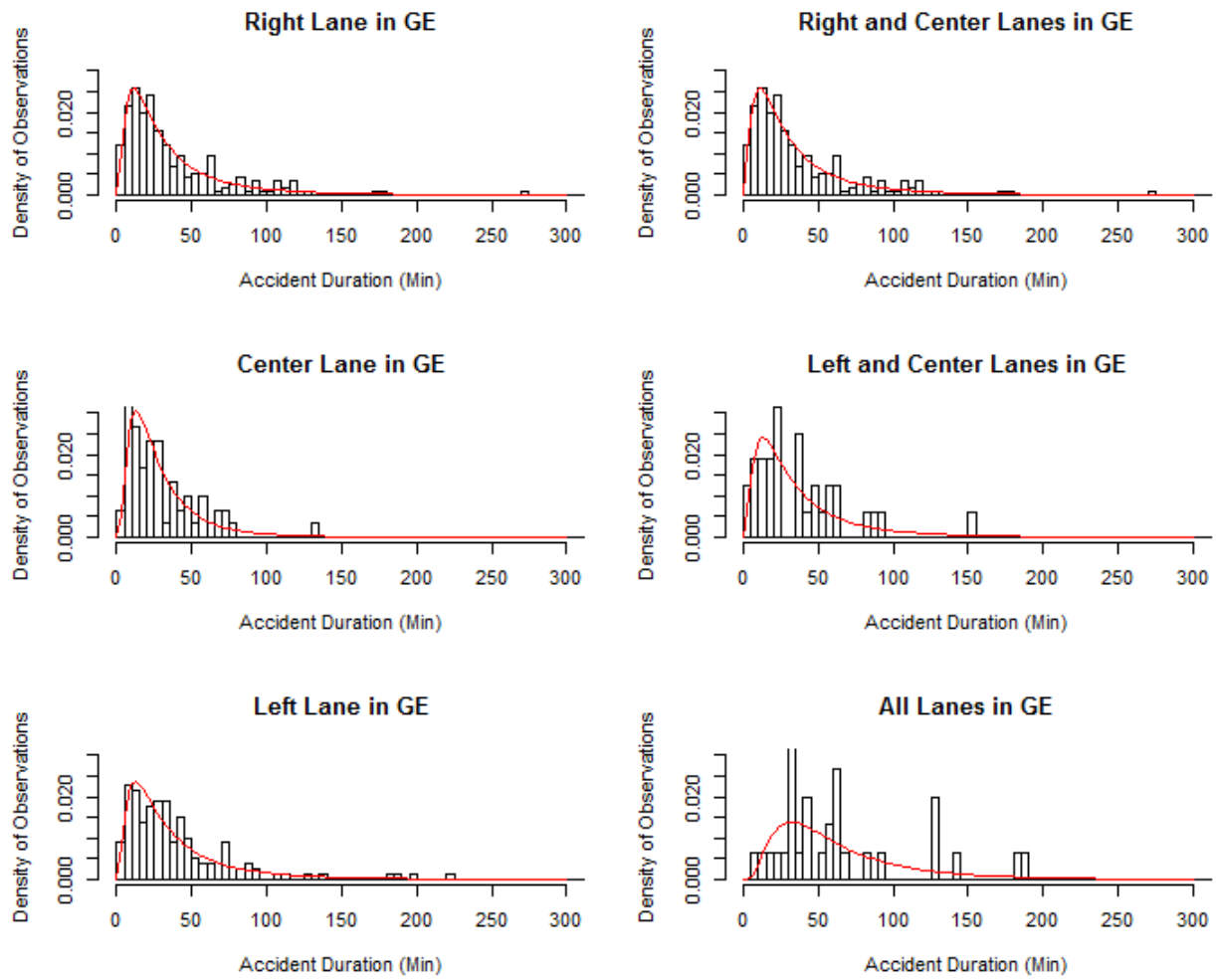


Figure 31. Accident durations with respect to blocked lane(s) in LIE

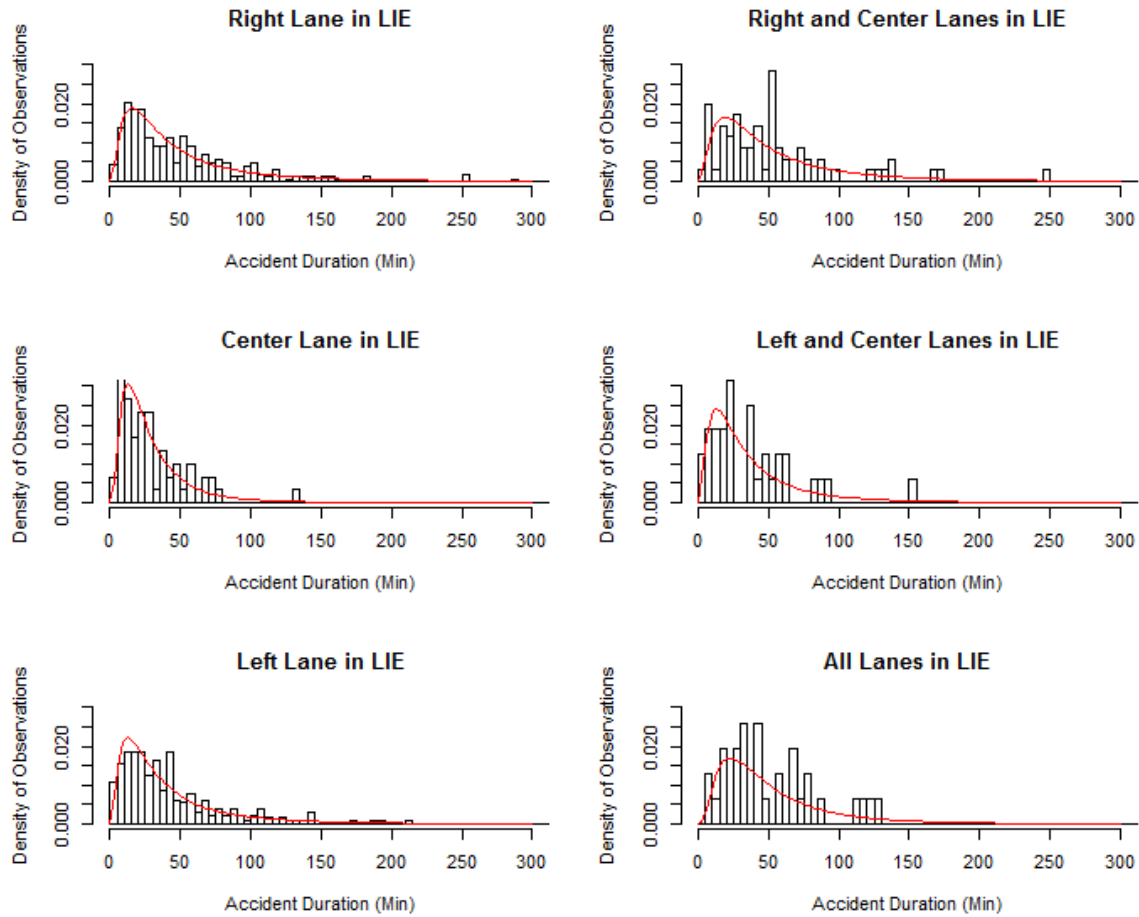
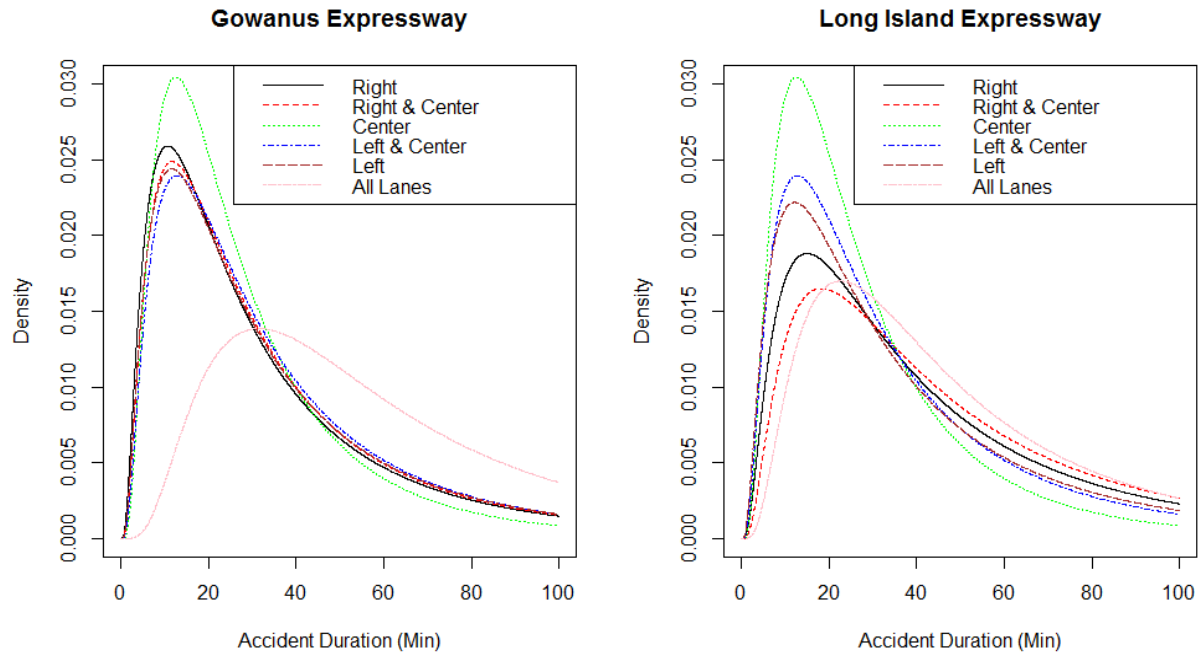


Table 10. Fitted distributions for durations based on affected lane(s)

Facility	Category (Lane)	Num. of Rec.	Actual Mean	Fitted Mean	Mean Log	SD Log	K-S Statistic
GE	Right	233	40.8	25.50	3.2388015	0.9360421	0.03274934
	Right & Center	39	38.38	26.68	3.2840527	0.9079344	0.12569773
	Center	60	30.47	22.83	3.1281624	0.7760619	0.07755687
	Left & Center	32	39.16	27.85	3.3269272	0.8852936	0.12581412
	Left	158	43.42	27.84	3.3266092	0.9249157	0.06278057
	All Lanes	30	66.43	52.21	3.9552126	0.7103493	0.10401901
LIE	Right	324	51.8	35.18	3.560557	0.925662	0.0508827
	Right & Center	70	58.4	40.39	3.6985668	0.8901647	0.08906900
	Center	60	30.47	22.83	3.1281624	0.7760619	0.07755687
	Left & Center	32	39.16	27.85	3.3269272	0.8852936	0.12581412
	Left	259	44.82	29.73	3.3922385	0.9486483	0.04825204
	All Lanes	31	52.52	41.01	3.713747	0.770743	0.11674042

Figure 32. Comparison of duration distributions for different blocked lane(s)



3.1.2.5 Temporal Comparison of Accident Durations

Figure 33 and Figure 34 show the comparison between accident durations based on time of day on weekdays and weekends on the GE and LIE respectively. For temporal comparison, the five following time intervals during weekdays as well as weekends were considered: a.m. peak (6:00 a.m.–9:00 a.m.), Midday (9:00 a.m.–4:00 p.m.), p.m. peak (4:00 p.m.–7:00 p.m.), Evening (7:00 p.m.–12:00 a.m.), Midnight (12:00 a.m.–6:00 p.m.).

Figure 33. Temporal comparison of accident durations on GE

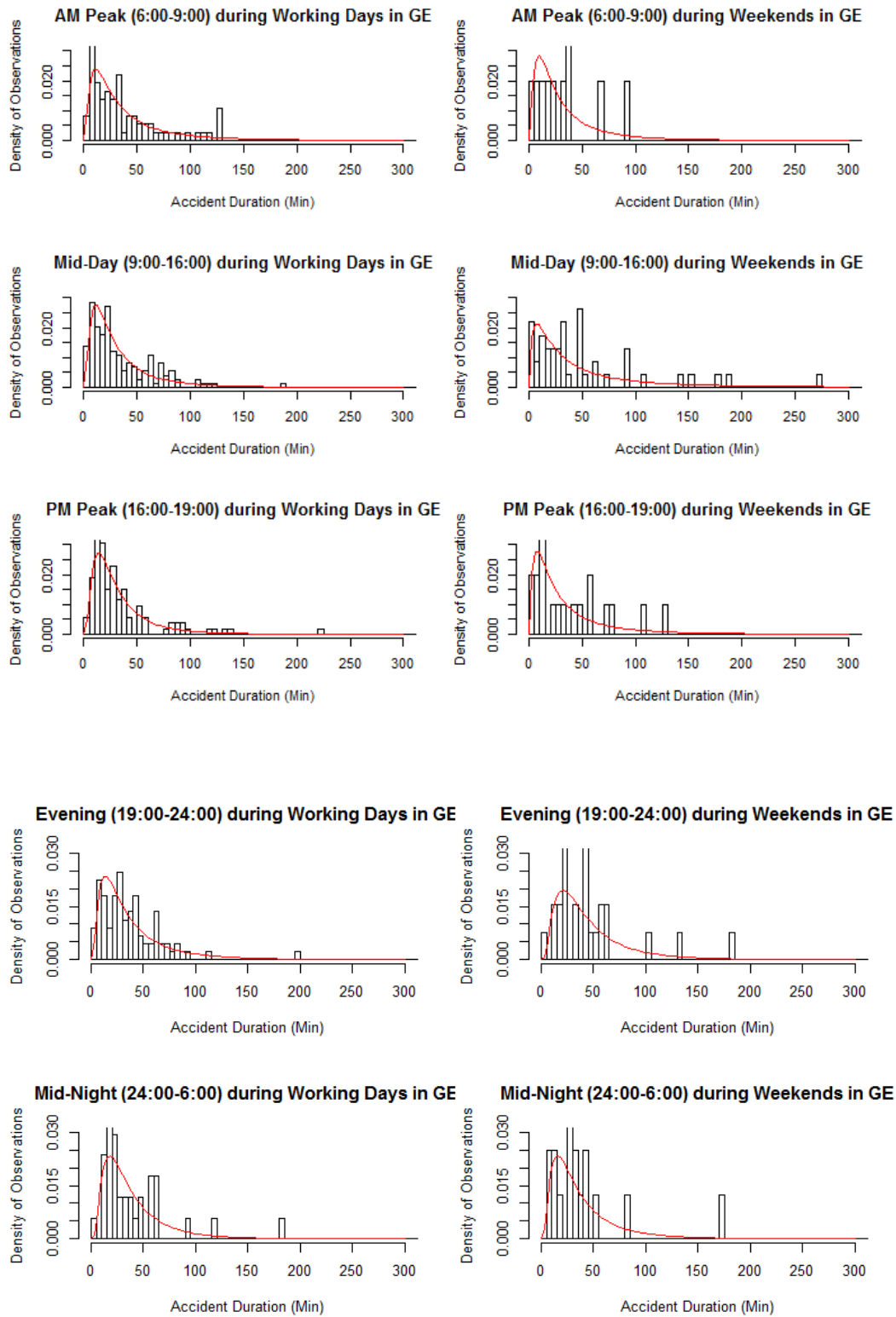


Figure 34. Temporal comparison of accident durations on LIE

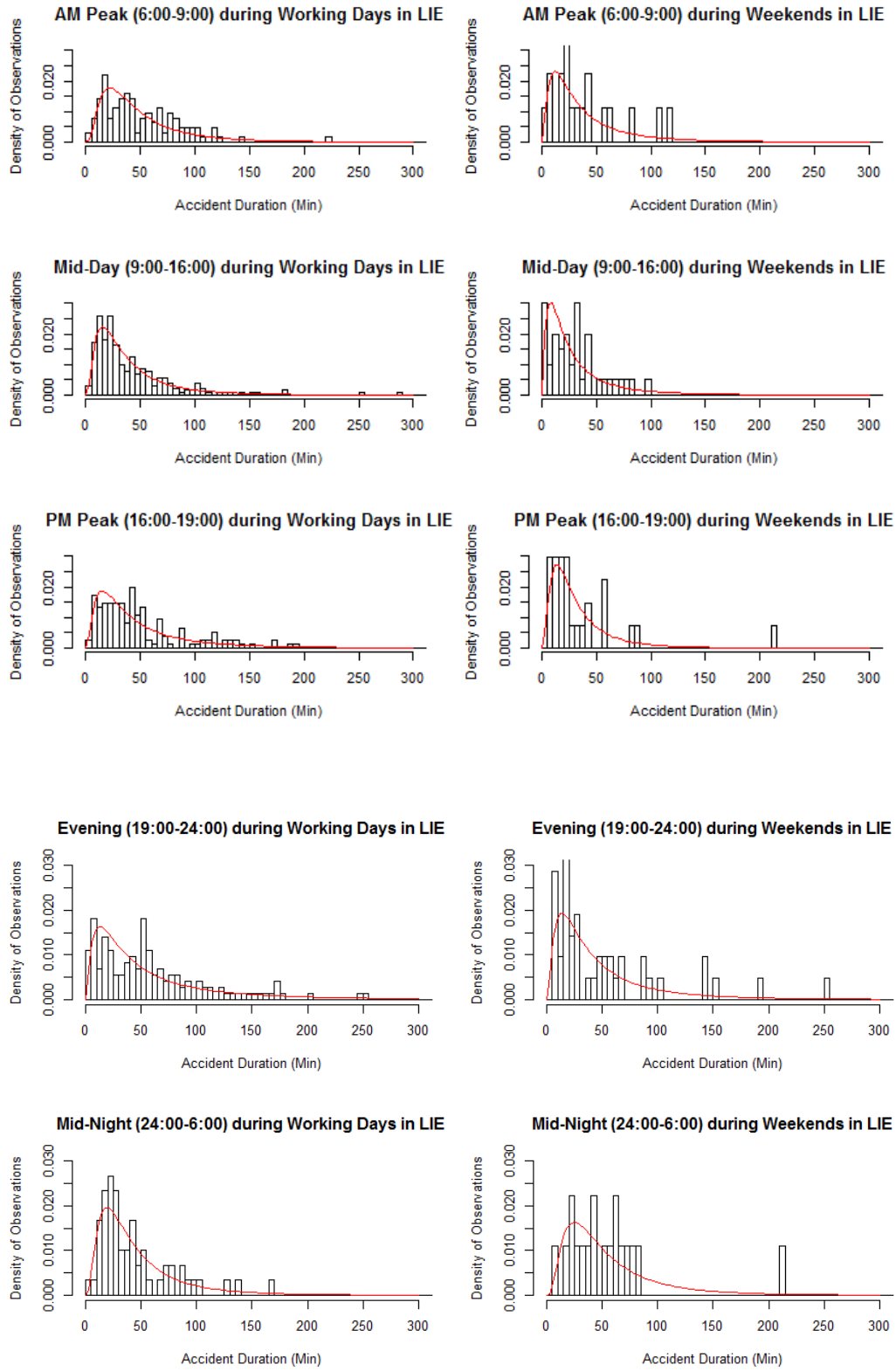


Table 11. Temporal comparison of accident durations

Facility	Category		Num. of Rec.	Actual Mean	Fitted Mean	Mean Log	SD Log	K-S Statistic
GE	Week Days	AM Peak (6 a.m.-9 a.m.)	73	43.79	27.33	3.3080676	0.9856962	0.07591581
		Midday (9 a.m.-4 p.m.)	147	34.82	24.08	3.1815932	0.9021301	0.06298722
		PM Peak (4 p.m.-7 p.m.)	105	35.06	25.11	3.2232353	0.7964227	0.06231656
		Evening (7 p.m.-12 a.m.)	89	42.38	28.57	3.3524156	0.8711519	0.08084968
		Midnight (12 a.m.-6 a.m.)	34	40.41	30.62	3.4218644	0.7338247	0.07955123
	Weekends	AM Peak (6 a.m. -9 a.m.)	10	33.8	23.18	3.1431646	0.9652739	0.16531469
		Midday (9 a.m. -4 p.m.)	46	67	32.26	3.473958	1.228119	0.08448975
		PM Peak (4 p.m.-7 p.m.)	20	38.4	23.59	3.160677	1.078236	0.11267842
		Evening (7 p.m.-12 a.m.)	36	47.23	35.98	3.5830071	0.7485167	0.11174995
		Midnight (12 a.m.-6 a.m.)	16	40.94	29.46	3.3830814	0.8007418	0.12548245
LIE	Week Days	a.m. Peak (6 a.m. -9 a.m.)	127	50.93	39.22	3.6692604	0.7768092	0.07850914
		Midday (9 a.m. -4 p.m.)	257	43.93	30.16	3.4066301	0.8477276	0.04081644
		PM Peak (4 p.m.-7 p.m.)	150	50.59	35.28	3.5633305	0.9388593	0.07730185
		Evening (7 p.m.-12 a.m.)	144	69.66	40.34	3.697467	1.073312	0.1208975
		Midnight (12 a.m.-6 a.m.)	60	49.28	35.42	3.5673740	0.7829738	0.06737256
	Weekends	a.m. Peak (6 a.m. -9 a.m.)	18	41.11	28.53	3.3509336	0.9465783	0.13151592
		Midday (9 a.m. -4 p.m.)	40	31.93	21.71	3.077584	1.021346	0.1430627
		PM Peak (4 p.m.-7 p.m.)	27	36.59	25.17	3.2255493	0.7990659	0.13745858
		Evening (7 p.m.-12 a.m.)	42	53.14	34.03	3.527361	0.951505	0.09060566
		Midnight (12 a.m.-6 a.m.)	18	57.88	43.34	3.7691282	0.7462775	0.13207753

Figure 35. Comparison of duration distributions for different times of day

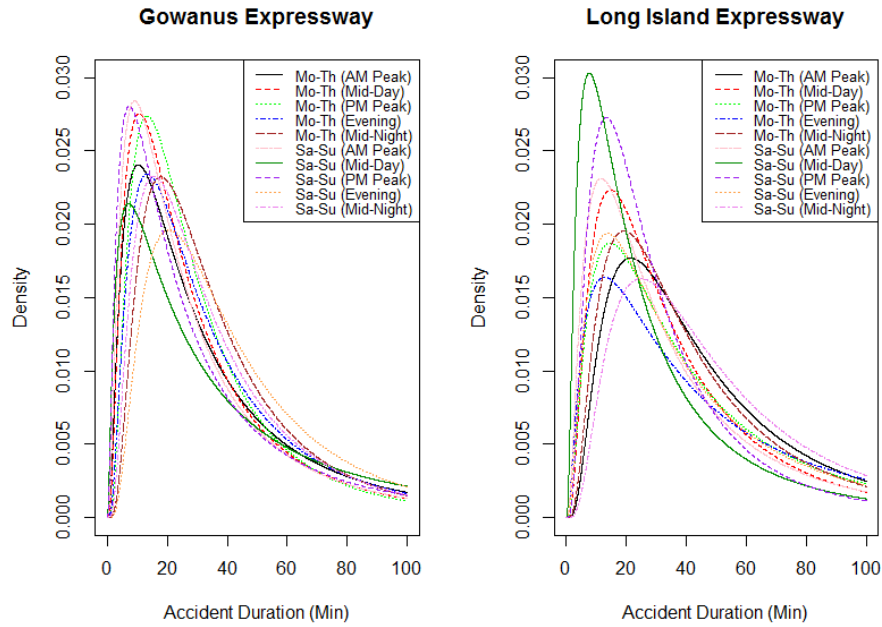


Table 11 and Figure 35 indicate that, on both the GE and LIE the durations for accidents occurring during peak hours were longer than those occurring in midday. However, there was an exception: on the GE on weekends, accident durations were relatively longer during midday than during peak hours.

3.1.3 Synthesis of Accident Duration Analysis

In analyzing accident durations several commonly used distributions were evaluated to find the distribution that best fits the duration data. Then, in sections 3.1.2.1 to 3.1.2.5 accidents were categorized with respect to different aspects (such as time, type of vehicle, and direction of traffic) to find which types of accidents result in longer accident durations and whether the differences between categories were significant. Based on these analyses, categories with similar results were recognized and combined. Table 12 illustrates new categories defined based on the significance of differences.

Table 12. New categories based on significance of difference between old categories

Facility	Feature	Old Categories		New Categories	
Gowanus Expressway	Time of Day	Week Days	AM Peak (6 a.m. -9 a.m.)	Week Days	AM Peak
			Midday (9 a.m. -4 p.m.)		Midday PM Peak
			PM Peak (4 p.m.-7 p.m.)		Evening Midnight
			Evening (7 p.m.-12 a.m.)		
			Midnight (12 a.m.-6 a.m.)		
		Weekends	AM Peak (6 a.m.-9 a.m.)	Weekends	AM Peak PM Peak
			Midday (9 a.m.-4 p.m.)		Midday Evening Midnight
			PM Peak (4 p.m.-7 p.m.)		
			Evening (7 p.m.-12 a.m.)		
			Midnight (12 a.m.-6 a.m.)		
	Affected Lane(s)	Right		Right Right and Center Left Left and Center Center All Lanes	
		Right and Center			
		Center			
		Left and Center			
		Left			
		All Lanes			
	Direction	East Direction		No Difference	
		West Direction			
	Vehicle	Automobiles		No Difference	
		Heavy Vehicles			

Table 12. continued

Facility	Feature	Old Categories		New Categories	
Long Island Expressway	Time of Day	Week Days	AM Peak (6 a.m. -9 a.m.)	Week Days	AM Peak
			Midday (9 a.m. -4 p.m.)		PM Peak
			PM Peak (4 p.m.-7 p.m.)		Midday
			Evening (7 p.m.-12 a.m.)		Evening
			Midnight (12 a.m.-6 a.m.)		Midnight
		Weekends	AM Peak (6 a.m. -9 a.m.)	Weekends	AM Peak
			Midday (9 a.m. -4 p.m.)		Midday
			PM Peak (4 p.m.-7 p.m.)		PM Peak
			Evening (7 p.m.-12 a.m.)		Evening
			Midnight (12 a.m.-6 a.m.)		Midnight
	Affected Lane(s)	Right		Right	
		Right and Center		Right and Center	
		Center		Left	
		Left and Center		Left and Center	
		Left		Center	
		All Lanes		All Lanes	
	Direction	East Direction		No Difference	
		West Direction			
	Vehicle	Automobiles		No Difference	
		Heavy Vehicles			

3.2 Traffic Volume/Flow and Speed Analysis

3.2.1 Data

DOT volume and speed data consists of volume counts and speed data for every 15 minutes for two sections of I-495 (LIE) and I-278 (GE), during an 18-month period from January 2015 to May 2016. The count station also recorded speed profiles which are crucial for calculating emissions. The volume data was collected at nine stations on the Gowanus Expressway and 21 stations on the Long Island Expressway, which are depicted in Figure 36. The data set included missing months, which corresponded to 150 days or 35% of the data. The available volume and speed data is shown in Table 13.

Table 13. Number of days in each month for which volume and speed data is available

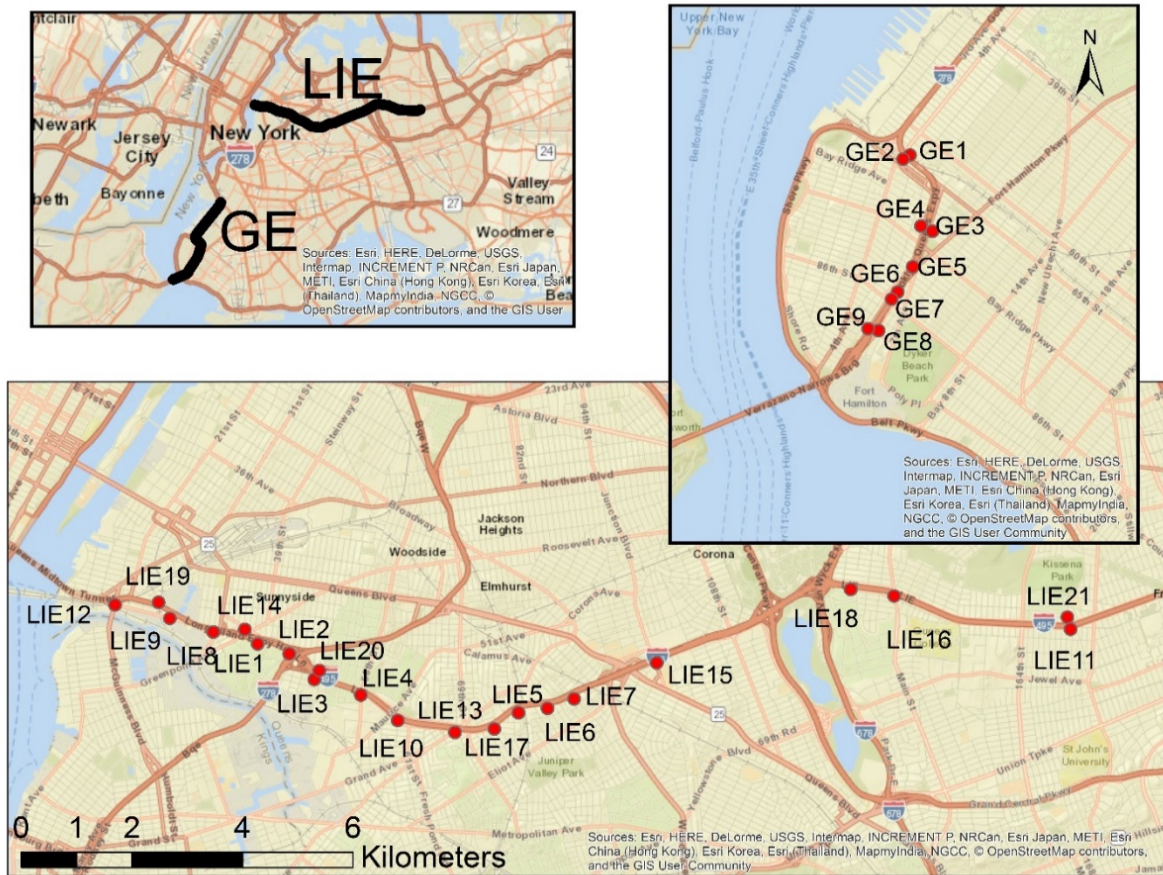
Fac.	Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
GE	2015	28	16	0	1	4	30	7	0	0	1	0	27
	2016	21	23	19	27	21	0	0	0	0	0	0	0
LIE	2015	31	28	29	29	6	30	22	1	5	10	19	29
	2016	30	29	29	30	31	0	0	0	0	0	0	0

Table 14. Number of accidents occurring close to each station on GE and LIE

St.I D	GE 1	GE 2	GE 3	GE4	GE5	GE6	GE7	GE8	GE9	LIE1	LIE2	LIE3	LIE4	LIE5	LIE6
#Ac c.	3	0	1	2	1	0	0	3	3	0	0	0	0	4	0
St.I D	LIE 7	LIE 8	LIE 9	LIE1 0	LIE1 1	LIE1 2	LIE1 3	LIE1 4	LIE1 5	LIE1 6	LIE1 7	LIE1 8	LIE1 9	LIE2 0	LIE2 1
#Ac c.	0	0	0	0	0	0	0	7	0	24	0	0	0	0	0

For the analysis, the data was divided into day of the week and time of day periods. For the preliminary analysis, the flow data for one sample station (GE1 at 65th Street East) was analyzed. This analysis will be extended to all other stations. Since the aim of the flow analysis was to calculate the flow conditions during an accident, the flow data set was also matched with accident records based on proximity. Table 14 shows the total number of records in the accident database, which correspond to each count station.

Figure 36. Locations of stations at the Gowanus and Long Island Expressways



3.2.2 Descriptive Analysis

3.2.2.1 Volume/Flow and Speed Profiles During the Day

The 24-hour speed and volume profiles are shown in Figure 37 and Figure 38. As shown in Figure 37, the patterns of travel speed during weekdays and weekends were different. There was a drop-in speed on weekdays during peak hours, while there was no significant drop in speed for weekends. According to Figure 38, on weekdays, the maximum volume occurred during morning peak hours while, on weekends, the maximum volume was in the afternoons. There were also more variations in traffic volume on weekdays than on weekends.

Figure 37. Average speed profile during 24 hours of a day at station GE1

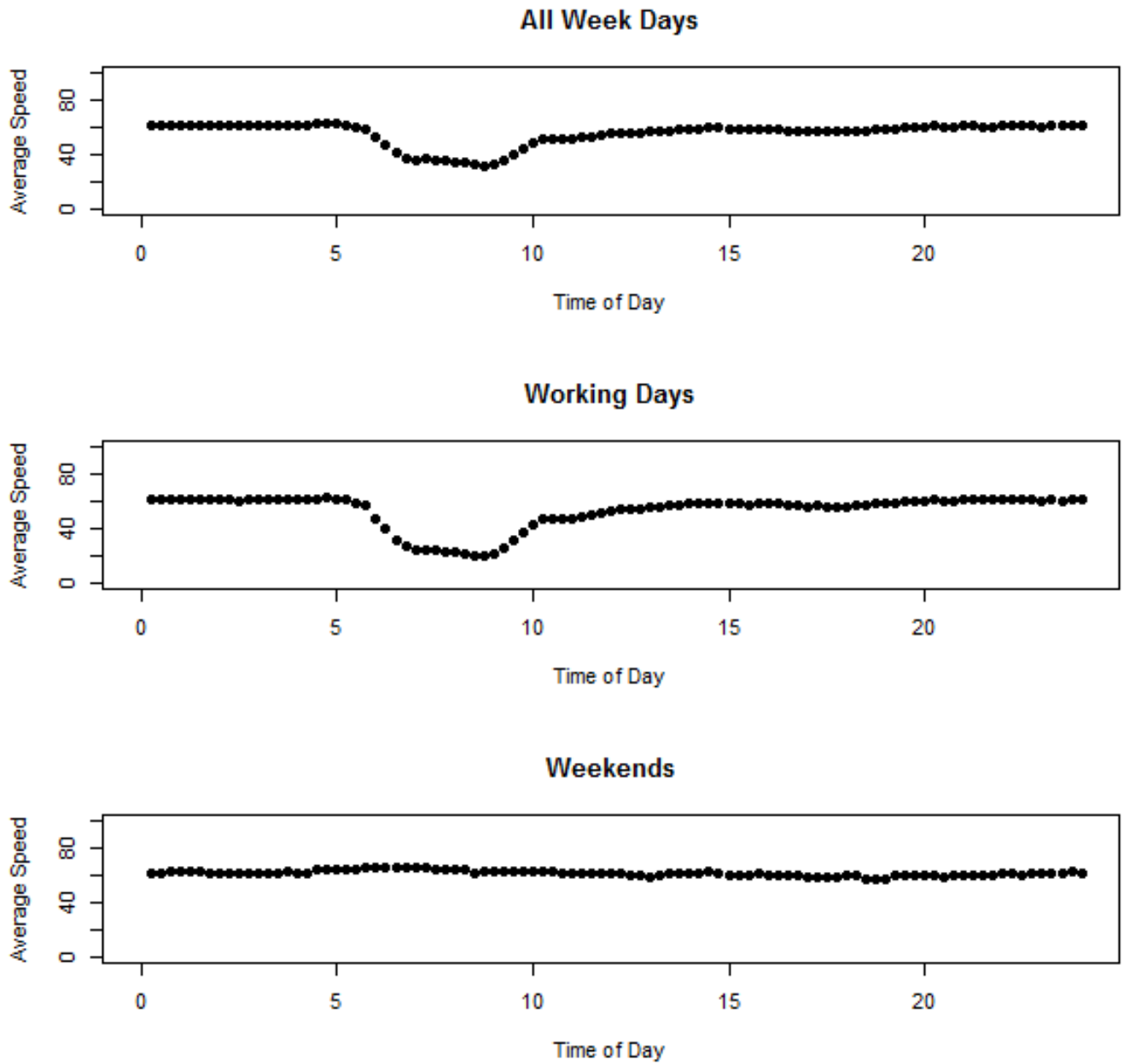
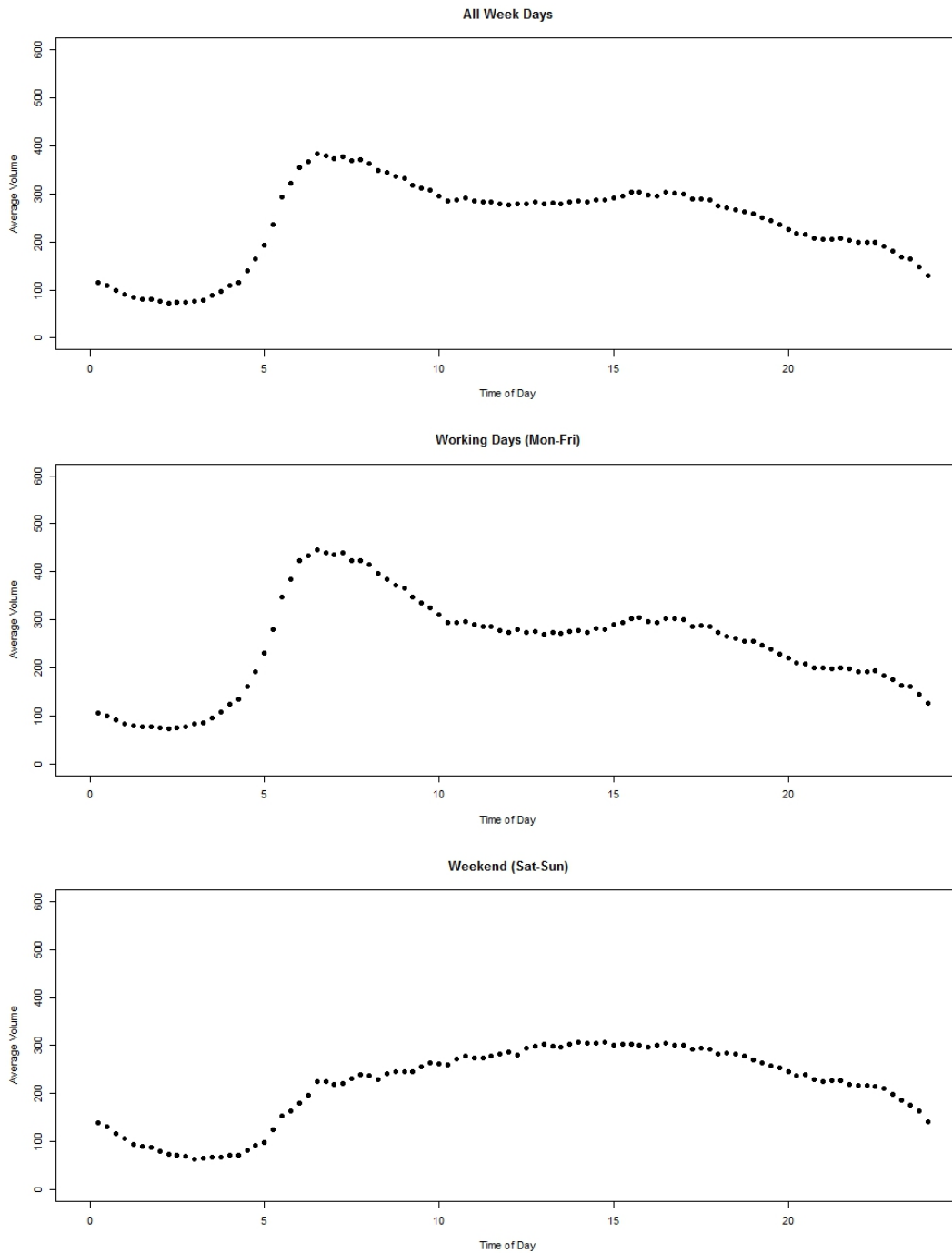


Figure 38. Average volume profile during 24 hours of a day at station GE1



3.2.2.2 Decrease in Capacity During Accidents

According to the Highway Capacity Manual, the capacity of a facility is defined as the maximum hourly rate at which vehicles can reasonably be expected to traverse a point or uniform section of a lane or roadway during a given time period under prevailing roadway, traffic, and control conditions.⁵⁹ Capacity under prevailing conditions can be estimated by calibrating a speed-volume curve for a given segment of highway. The peak of this curve defines capacity. This method is based on the fundamental models describing the speed-volume relationship.

Figure 39. Theoretical speed-flow diagram

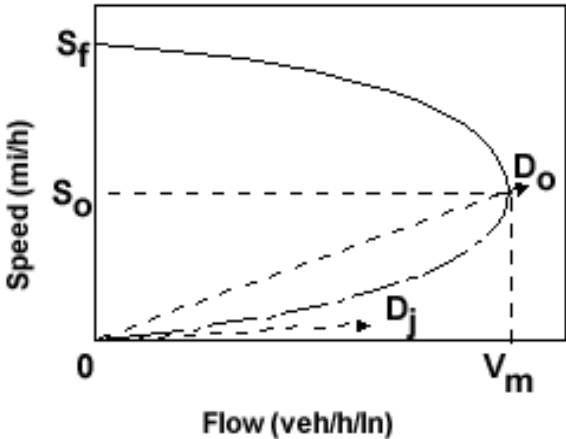
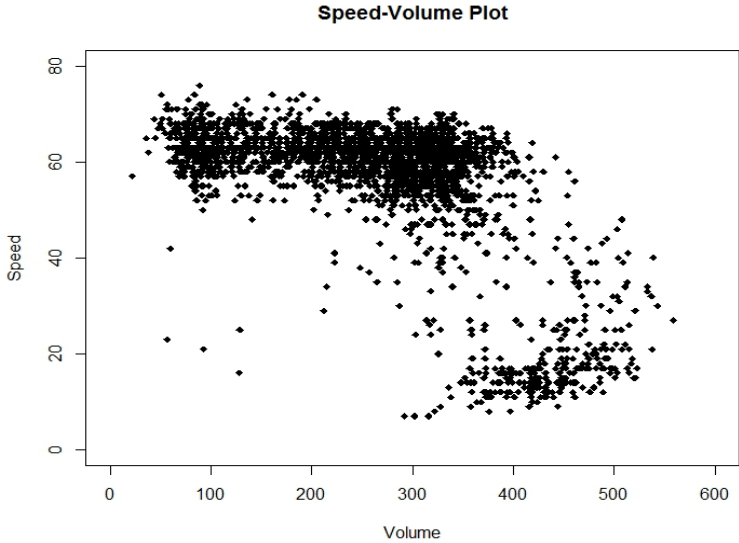


Figure 40. Speed-flow diagram for station GE1



A theoretical speed-flow diagram is shown in Figure 39. Figure 40 shows the speed-flow diagram based on the data sets in this study. When an accident occurs, a bottleneck is formed, which in turn creates additional delay. For the two facilities analyzed in this study, the traffic flow is at high levels most of the time so when an accident occurs, it can be assumed that demand exceeds capacity during the accident. Thus, road capacity during an accident can be measured directly as bottleneck traffic flow. Due to the significant divergence of the actual speed-flow data from the theoretical depiction, an empirical approach given in Table 15 is used to estimate the accident capacity and the traffic flow during non-accident conditions.

Table 15. The empirical approach used for calculating capacity drop and delay

Maximum Capacity	Regular Flow	Accident Capacity	Regular Speed	Accident Speed
Maximum traffic flow during the whole analysis period	Average traffic flow from 1 hour before accident creation to 2 hours after clearance during the day in which accident occurred	Minimum flow of 15-minutes intervals from 30 minutes before accident to 30 minutes after	Average speed of two 15-minutes intervals before accident creation in the accident day	Average Speed During Accident

For this analysis, delay is calculated as follows:

Equation 4
$$Delay = (Acc. Dur.)^2 \times \frac{(Max.Cap. - Acc.Cap.) \times (Reg.Flow - Acc.Cap.)}{2(Max.Cap. - Reg.Flow)}$$

It should be noted that the available flow data is given in 15-minute intervals, so the accident duration used for the delay calculation is the aggregated 15-minute intervals which cover the actual accident duration. This approach is used for calculating the capacity drop and delay for accidents occurring at stations for which the volume and speed data are available.

4 Early Incident Detection and other Information Extraction using Social Media with California as a Model

In Section 2.3, the analyses showed that the limited incident data of 970 records for the LIE and GE from June 2015 to May 2016 did not match any personal tweets for the two corridors. Nevertheless, a few studies have shown the utility of Twitter for incident detection and management.^{22,67} Particularly, Gu et al. (2016) showed that among tweets collected in Pittsburgh and Philadelphia using several keywords, about 4-5% were from incident-related personal accounts.⁶⁷ However, as compared to the current study, Gu et al. (2016) had access to data sets that included RCRS (Road Condition Report System) incident data maintained by Pennsylvania DOT for all state-owned roads and 911 Call for Service (CFS) data provided by the City of Pittsburgh.⁶⁷ Hence, the methodology identified in the current study can optimistically be applied for productive purposes and verified with the availability of incident-related data that is richer and applicable to a greater number of roadways.

For the purpose of illustrating the benefits of earlier-incident detection using social media, incident data with larger geographical and temporal range available from CHP were used to match the incident-related tweets in California. The California Department of Transportation (CalTrans) provides all historical incident data (incident start time, location, milepost, description, TIM response log, etc.) for the state highways, interstate, and freeways, along with traffic flow and speed data on the Performance Measurement System (PeMS) website.⁶⁸ In addition, to match accidents from local streets in California, accident data available through the Statewide Integrated Traffic Records System (SWITRS) were used.⁶⁹ The flow and speed data were used to estimate the reduction in delay as well as emissions and fuel consumption for early detection of highway incidents through Twitter.

4.1 Identification of Early Incident Detection Success

In order to identify potential early detection, the tweets for the month of May 2016 and October 1-15, 2015 from California were processed, mined, and geo-identified using the framework described in the earlier section. The resultant tweet database was further filtered by removing organizational tweets from accounts such as CalTrans, SF311, Total Traffic, and other news agencies. Subsequently, the databased is scored based on tf-idf weighting of keywords consisted of 19,040 tweets. The tweets with tf-idf score over 5.0 were chosen for further analysis. Despite filtering out tweets with low tf-idf scores, there still existed irrelevant tweets (e.g. *"I got a Mac Safari crash with an exception at address*

Oxfbadbeef, which I thought was the coolest coincidence ever until I Googled it”). The relevant tweets were identified manually and compared with the CalTrans’ PeMS database and CHP’s SWITRS database for highway and local streets, respectively. The matching accidents were recorded, and early detections noted. For the early detections in highways, the corresponding flow and speed data were extracted from PeMS for delay and emissions and fuel consumption benefits analysis. The results of corresponding benefits are shown in section 5. Since SWITRS database does not include flow and speed data, the local accidents were not included in the benefits analysis but were still considered for early detection success rate.

4.2 Success Rate for Incident Detection through Twitter Feeds

For the analysis, a total of 19,040 personal tweets were processed for May 2016 and October 1-15, 2015. From this total, 1465 tweets made the tf-idf cut-off score of five. Among these 1465 potentially relevant tweets, 549 were identified as accident-related tweets. When compared with PeMS and SWITRS databases—containing about 64,000 incidents—21 of these tweets were uniquely matched with an actual accident record (9 highway and 12 local road accidents) within two hours of official accident time. Meanwhile, three accidents that were detected through Twitter could not be matched with official accident records. The three tweets with early detection through Twitter preceded the recorded accident time by 19, 23, and 4 minutes. A description of these tweets and the corresponding incidents is shown in Table 16.

Table 16. Comparison of accident-related tweets and official accident records – PeMS and SWITRS for highway and local accidents, respectively, in the state of California

Accident	Early Detection	Twitter Feed		PeMS/ SWITRS Record			
		Tweet Time	Tweet	Start Time	Duration (mins)	Location	Description
#1 Highway	Yes	5/21/2016 14:48	@CHP_HQ 118Fwy East Bound between Yosemite & Sterns Ave is a Large Desk that's Broken Up going across Several Lanes!! My car is now Damage!	5/21/2016 15:09	8	Sr118 W Yosemite Ave	1125-Traffic Hazard
#2 Highway	No	5/16/2016 11:04	Not the best way to start a morning. #fire #accident #firemen #road #danger #freeway #monday #la; Geocode: (33.8456, -118.2058)	5/16/2016 10:52	89	I710 S Willow St	FIRE-Report of Fire
#3 Highway	No	5/1/2016 20:00	405 north left two lanes are closed for big car accident.	5/1/2016 19:34	175	I405 N Magnolia St Ofr	1179-Trfc Collision-1141 Enrt
#4 Highway	No	5/9/2016 8:30	If you need to take i80 west avoid it if you can a tweaker is tying to jump off the bridge. So there is gonna be hella traffic real soon	5/9/2016 8:27	12	I80 W Madison Ave	1125-Traffic Hazard
#5 Highway	Yes	10/2/2015 6:49	North of Sac- SB 99 at Elverta crash involving a semi and debris, impact is slow lane closure right now	10/2/2015 7:12	187	Sr99 S / G St	1183-Trfc Collision-Unkn In
#6 Highway	Yes	10/2/2015 14:03	Avoid the 215 guys, HUGE accident! It's at a dead stop	10/2/2015 14:07	60	I215 S / I215 S Mount Vernon Ave Onr	1182-Trfc Collision-No Inj
#7 Highway	No	10/12/2015 19:39	San Mateo Bridge CA 92 Westbound major accident at about 6:20 pm earlier #sanmateoaccident	10/12/2015 18:20	12	Sr92 W / High Rise	1183-Trfc Collision-Unkn Inj
#8 Highway	No	10/14/2015 22:40	Evening car crash cops ram car to shoulder side traffic slows way down @ Highway 101 https://t.co/hMRvexvldG	10/14/2015 20:15	239	Us101 N / Us101 N Millbrae Ave Ofr	1183-Trfc Collision-Unkn Inj

Table 16 continued

#9 Highway	No	10/14/2015 21:40	Car fire shuts down traffic on Golden Gate Bridge, http://t.co/PM5fTudRed	10/14/2015 17:25	25	Us101 N / Golden Gate Bridge	Fire
#10 Highway	No records??	10/15/2015 11:46	Accident on Highway 41/49, two cars, no major injuries http://t.co/5Rg1MjXBsG	No records??			
#11 Highway	No records??	10/16/2015 6:47	4 car crash 17 NB close to Glenwood cutoff. One lane blocked	No records??			
#1 Local	No	5/9/2016 8:52	Traffic accident at MLK and Delaware. Expect delays, congestion, gawkers, etc. @BerkeleyPatch #bpd	5/9/2016 8:30		Berkeley, Delaware St & MLK Dr	3 vehicle collision; 3 injured
#2 Local	No	5/4/2016 15:45	Damn bad ass accident highway 4 by the streets of brentwood	5/4/2016 15:13		SR 4 & Sand Creek Rd	4 vehicle crash; 1 killed; 3 injured
#3 Local	No	5/26/2016 10:32	Major injury accident reported at Highway 132 at North Blossom Road near Waterford.	5/26/2016 8:15		Modesto SR 132 @ McEwen	2 vehicle; 2 injured
#4 Local	No	5/4/2016 2:57	@ktlagingerchan ave 26 and Humboldt closed in Lincoln heights body's on ground	5/4/2016 1:30		Los Angeles Humboldt & 26 Ave	1 killed; 1 injured
#5 Local	No records in SWITRS crash data;	5/26/2016 12:56	Car crash just now on the corner of 7th & Hugo which I have told @sfmta_muni is pretty much designed for crashes. What a shocker.	No records			
#6 Local	No	10/11/2015 15:53	Fatality traffic collision: Veile Road / W 1st St. in Beaumont. More: http://t.co/ISOz1vGolm	10/11/2015 15:08		Veile Avenue / 1st Street, Riverside	1 killed; 1 injury
#7 Local	No local records for traffic hazards	10/13/2015 13:48	@KNX1070 traffic sign pushed into traffic traffic lane wilshire/fairfax area Car left it's bumper http://t.co/RTGEh4HO5Z	No records			
#8 Local	No local records for vehicle disablements	10/02/2015 10:52	My car broke down right outside where that car crashed into the eye care building off of lake chabot road Imao	No records			

Table 16 continued

#9 Local	No	10/14/2015 18:15	Huge traffic accident on PCH near our hospital involving a police vehicle so those picking up pets may be delayed. No worries, we are here.	No records			
#10 Local	No	10/15/2015 7:18	@kcbstraffic NB skyline closed at john muir, heavy traffic, police activity	10/15/2015 2:14		Skyline Blvd & John Muir, San Francisco	
#11 Local	No	10/11/2015 1:32	Car accident right in my front yard #fawk sux living in a corner house sometimes http://t.co/0uQcsCn1DG	No records			

4.3 Potential of Information Contained in Incident-Related Tweets

Incident-related tweets can potentially provide other benefits due to the rich user-generated information content. Early detection of incidents can potentially help in reducing the severity of accidents such as a fatality could be mitigated to an injury, major injury to minor injury, etc. In this way, the information content of incident-related tweets can provide benefits in addition to early detection savings discussed in previous sections. Among the 21 tweets listed in Table 16 the information regarding debris, progression of events, etc. can be useful for incident response or even verification.

For instance, for highway incident #1, the information regarding the large wooden debris is useful for informing drivers of potential hazard even before the police can reach the scene. The deployment of an appropriate debris collection crew can also be completed sooner. Similarly, the tweet for local incident #8 provides information and location of vehicle disablement. Such information is useful particularly on local or rural roadway, which may not be instrumented with cameras as is the case on highways. In other words, Twitter can function as an auxiliary information source for road facilities with no sensing infrastructure. The tweet information content can also be utilized to prevent potential accidents due to undetected road hazards. For instance, Tweet #7 provides information and location regarding a potentially dangerous local traffic sign and the exact nature of the hazard can be ascertained by using the picture attached to the tweet.

The information gathered through Twitter can also be used to help accident data archiving. For instance, the local incident reported in Tweet #5 does not have any official record entry. It is likely that the drivers involved in the collision have not reported the incident. However, the particular nature of information, i.e., the perceived potential hazard of the particular intersection located at 7th and Hugo Streets can be useful for the agency in making street design evaluations.

The potential benefits can also go beyond traffic management. Tweet #2 reports fire and congestion. The corresponding CHP TIM log reports of a brush fire on the side of the freeway and the spreading of this fire from a house to surface streets. Although, the tweet is eight minutes later than the CHP TIM log, it contains a geocode. The location of the “Brush fire” is reported almost three miles downstream of the geocoded point. Thus, the tweet provides information on the spread of fire that is extremely useful not only for TIM but also for saving lives and property. It could also help in providing information regarding freeway congestion.

5 Savings Due to Early Accident Detection

5.1 Delay Reduction

After the calculation of a delay due to an accident, the incident-delay saving (IDS) can then be calculated in terms of vehicle-hours (or vehicle-minutes). IDS is determined by the difference in time between the incident timeline with and without the aid of early detection.⁶⁰

Equation 5
$$IDS = Delay_{base} - Delay_{early\ detection}$$

Monetary cost savings due to this delay reduction can be calculated as:

Equation 6.
$$IDS_{Cost\ Saving} = IDS(Car_{\%} \times Car_{Occupancy} \times Car_{VoT} + Truck_{\%} \times Truck_{VoT})$$

In the equation, $Car_{\%}$ and $Truck_{\%}$ are the portion of cars and trucks in highways which are assumed to be 92% and 8%, respectively.⁶⁰ $Car_{Occupancy}$ is the average vehicle occupancy, which was assumed to be 1.15 persons per vehicle. For estimating the monetary value of time for California, Car_{VoT} and $Truck_{VoT}$ are a value of time per vehicle per hour which were assumed to be \$13.65 and \$31.4 for cars and trucks, respectively.⁷⁰

5.2 Reduced Emissions

Emission factors, available from the U.S. Environmental Protection Agency, are used to calculate the reduced amount of emissions released because of early accident detection.⁶¹ The emission factors, based on vehicle speed, provide the amount of emissions released in terms of grams/mile and are presented in Table 17. For this study, the emissions amount in grams is converted to grams/hour by multiplying the emission factor and average speed during an accident. Delay savings, in terms of vehicle-hours, can then be used to determine the reduction in emissions due to early detection.

Equation 7.
$$Reduced\ Emission_{(gr)} = IDS_{(veh.hr)} \times Emission\ Factor_{\left(\frac{gr}{mi}\right)} \times Accident\ Speed_{\left(\frac{mi}{hr}\right)}$$

Table 17. Emission factors by speed ⁶¹

Grams per Mile									
Speed (mph)	ROG	CO	NOx	PM2.5 Ex	Speed (mph)	ROG	CO	NOx	PM2.5 Ex
5	0.34	3.88	0.44	0.014	35	0.06	1.86	0.26	0.003
6	0.31	3.75	0.43	0.013	36	0.06	1.83	0.26	0.003
7	0.29	3.62	0.42	0.012	37	0.06	1.81	0.26	0.003
8	0.27	3.50	0.40	0.011	38	0.06	1.79	0.26	0.003
9	0.24	3.37	0.39	0.010	39	0.06	1.77	0.26	0.002
10	0.22	3.24	0.38	0.009	40	0.06	1.75	0.26	0.002
11	0.21	3.15	0.37	0.009	41	0.06	1.73	0.26	0.002
12	0.19	3.05	0.36	0.008	42	0.06	1.72	0.27	0.002
13	0.18	2.96	0.35	0.008	43	0.06	1.71	0.27	0.002
14	0.17	2.87	0.34	0.007	44	0.06	1.69	0.27	0.002
15	0.15	2.78	0.33	0.007	45	0.06	1.68	0.27	0.002
16	0.15	2.71	0.33	0.006	46	0.06	1.67	0.27	0.002
17	0.14	2.65	0.32	0.006	47	0.06	1.67	0.27	0.002
18	0.13	2.58	0.31	0.006	48	0.06	1.66	0.28	0.002
19	0.12	2.51	0.31	0.005	49	0.06	1.65	0.28	0.002
20	0.11	2.45	0.30	0.005	50	0.06	1.65	0.28	0.002
21	0.11	2.40	0.30	0.005	51	0.06	1.65	0.29	0.002
22	0.10	2.34	0.29	0.005	52	0.06	1.65	0.29	0.002
23	0.10	2.29	0.29	0.004	53	0.06	1.66	0.30	0.002
24	0.09	2.24	0.29	0.004	54	0.06	1.66	0.30	0.002
25	0.09	2.19	0.28	0.004	55	0.06	1.66	0.30	0.002
26	0.09	2.16	0.28	0.004	56	0.06	1.68	0.31	0.002
27	0.08	2.12	0.28	0.004	57	0.06	1.69	0.32	0.002
28	0.08	2.08	0.27	0.004	58	0.07	1.71	0.32	0.002
29	0.08	2.04	0.27	0.003	59	0.07	1.73	0.33	0.002
30	0.07	2.00	0.27	0.003	60	0.07	1.74	0.34	0.002
31	0.07	1.97	0.27	0.003	61	0.07	1.78	0.35	0.002
32	0.07	1.94	0.27	0.003	62	0.07	1.81	0.36	0.003
33	0.07	1.91	0.27	0.003	63	0.08	1.85	0.37	0.003
34	0.07	1.88	0.26	0.003	64	0.08	1.88	0.38	0.003
					65	0.08	1.92	0.39	0.003

Source: EMFAC2011LDV, average annual emissions, statewide vehicle fleet, 50% humidity, temperature 75 degrees F.
 ROG includes running exhaust and running evaporative emissions. PM2.5 Ex includes running exhaust emissions only.

The speed data during the accidents were gathered from the PeMS database for incidents relating to California data. The necessary emission factors (which provide the amount of emissions released in terms of grams/mile based on speed) were obtained from the U.S. Environmental Protection Agency database. ⁶¹ Based on reduced emissions, the monetary savings were calculated using emission cost parameters provided by California Department of Transportation. ⁷⁰ The associated costs for ROG, CO, NOx, and PM 2.5 were 1,305; 80; 18,700; and 151,100 dollars per ton, respectively. Accordingly, the cost savings due to emission reductions were calculated by multiplying the reduced amount of pollutant with the corresponding cost parameter.

5.3 Reduced Fuel Consumption

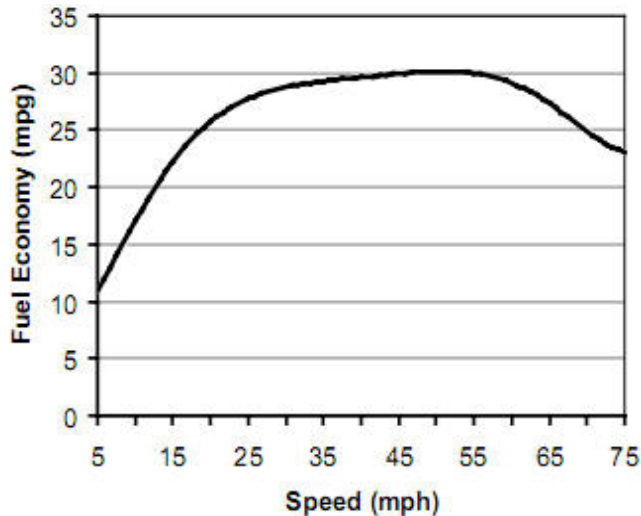
Delay reduction due to early detection results in a reduction in the time that vehicles are on the highway. The reduction in fuel consumption is estimated using IDS. Based on the average speed and vehicle makeup of the queue, the reduced fuel consumption can be determined.⁶⁰ The fuel consumption, in terms of gallons per mile, for cars and trucks can then be entered in the following equation to determine the total fuel reduction:

Equation 8

$$Fuel\ Reduction_{(gal)} = IDS_{(veh.hr)} \times Accident\ Speed_{\left(\frac{mi}{hr}\right)} \times (Car_{\%} \times Gas_{\left(\frac{gal}{mi}\right)} + Truck_{\%} \times Diesel_{\left(\frac{gal}{mi}\right)})$$

Figure 41 shows the fuel economy by speed for cars. The graph in this figure indicates that there would be a decrease in fuel economy when the speed drops below 55 mph. Based on previous studies,⁶⁰ 20 miles per hour is typically used as the upper bound of travel speed in severe congestion. Therefore, the value of 20 mph is chosen to calculate the rate of fuel consumption in congestion due to incidents. Using this value, fuel consumption is assumed to be 0.03875 gallons of gasoline per mile for cars and 0.1429 gallons of diesel per mile for trucks.^{62, 63}

Figure 41. Fuel economy by speed ⁶⁴

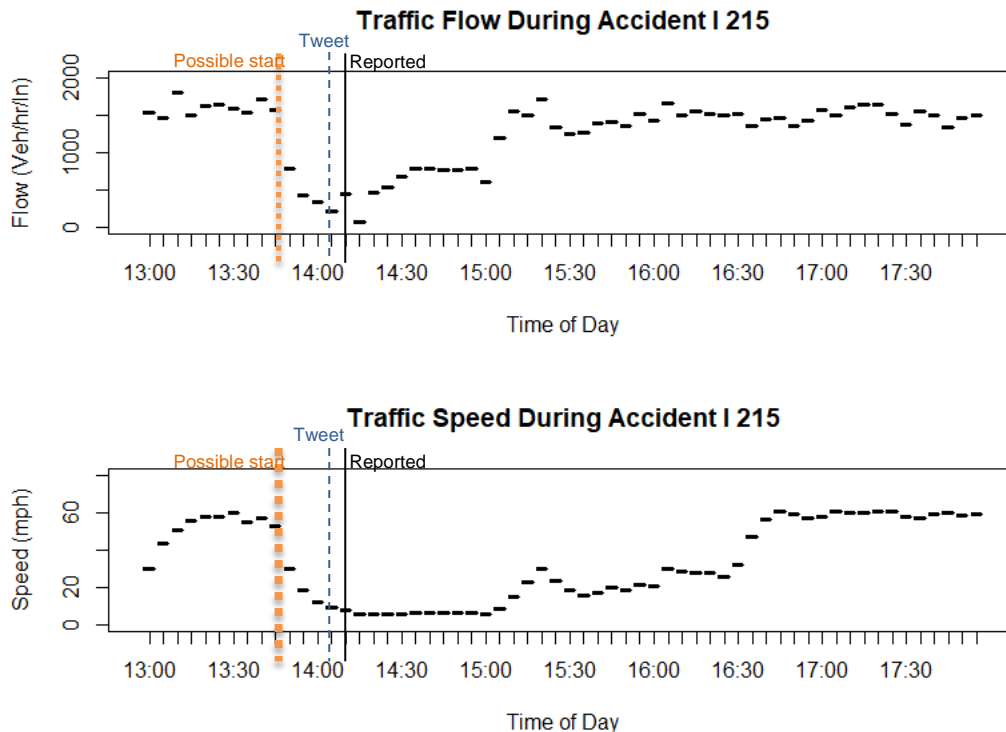


The monetary savings due to fuel consumption reduction for incidents in California can be calculated using fuel prices (\$2.65/Gal. for Gasoline and \$2.4/Gal. for Diesel) provided by California Department of Transportation.⁷⁰

5.4 Calculated Delay, Emissions, and Fuel Savings for Early Incident Detections using California Data

The benefits of using incident-related tweets from six weeks' Twitter data for early detection for TIM of incident data in California were estimated using the methodology described above using the flow and speed data from PeMS database. To illustrate the result of early detection due to tweets, the flow and speed and tweet and incident times for highway incident #6 on I-215S are shown in Figure 42. For this particular incident, the tweet preceded the incident time by four minutes. The premise is that had the information from tweets been used for early detection, the response time would be shortened by four minutes.

Figure 42. A sample accident early detection using twitter



The delay reduction, emission reduction, and fuel savings for early detecting tweets are estimated and shown in Table 18 for highway incidents #5 and #6. The monetary value of the savings is estimated using the latest life-cycle costs for California.⁷⁰ In total, 4,046 vehicle-hours of delay savings, reduction in 5.9 kg of reactive organic gases (ROG), 133 kg of carbon monoxide (CO), 16.3 kg of nitrous oxides (NOx), 0.3 kg of particular matter (PM 2.5), 1,939 gal of gasoline, and 622 gal of diesel were estimated to be saved due to the early detection by analyzing Twitter feeds. These savings amount to a monetary value of \$75,600 per six weeks. Given the fact that there are about 32,000 miles of highways in CA,⁶⁸ the savings could amount to about \$0.5 per highway mile per week.

Researchers noted that although highway incident #1 is detected earlier using Twitter due to minimal flow, the incoming demand at the incident location was able to pass through the available lanes. Therefore, although there are no obvious savings, there are secondary benefit from such tweets are highlighted in earlier section.

Table 18. Delay, emissions, and fuel consumption savings due to early incident detection through twitter for incidents in the state of California

Factor	Accident			
	I 215		SR99	
	Reduction	Monetary saving	Reduction	Monetary saving
Gasoline	380 (gal)	1,008 (\$)	1,559 (gal)	4,132 (\$)
Diesel	122 (gal)	293 (\$)	500 (gal)	1,200 (\$)
Fuel cost saving	1,300 (\$)		5,332 (\$)	
ROG	1,173 (gr)	2 (\$)	4,811 (gr)	6 (\$)
CO	26,135 (gr)	2 (\$)	107,168 (gr)	8 (\$)
NOx	3,200 (gr)	60 (\$)	13,122 (gr)	245 (\$)
PM2.5	53 (gr)	8 (\$)	218 (gr)	33 (\$)
Emission cost saving	72 (\$)		293 (\$)	
Delay	893 (veh.hr)	15,136 (\$)	3,153 (veh.hr)	53,467 (\$)
Total cost saving	16,507 (\$)		59,093 (\$)	
Grand Total	75,600 (\$)			

The monetary savings obtained in Table 18 are only for the traffic externalities. Additional savings can be expected due to severity reduction, improved information coverage, hazard detection, etc.

5.5 Potential Savings through Use of Twitter Feeds in New York State

To illustrate the benefits of early detection, three accidents on the GE (near count station #1) were individually analyzed, assuming five minutes of early detection via Twitter feeds. For the delay savings, reduced emissions, and fuel consumption calculations, it is also assumed that 8% of traffic consists of trucks and 92% consists of passenger cars. Figure 43 to Figure 45 show the variations of flow and speed, before, during, and after these accidents. Also, the accident characteristics along with the calculated delay for these sample accidents are given in Table 19 to Table 21.

Table 19. Potential benefits of five-minute early detection for sample accident #1 at GE

St.ID	Direction	Cross Street	Created Date	Created Time	Cleared Time	Lanes Affected	Vehicles Involved
GE1	East	65 th Street	Tuesday, January 26, 2016	6:14 p.m. (73)	6:32 p.m. (75)	Left & Center Lanes	1A
Acc. #	Regular Flow (veh/hr)	Accident Flow (veh/hr)	Flow Drop (%)	Delay (hrs)	Regular Speed (mi/hr)	Accident Speed (mi/hr)	Speed Drop (%)
1	843	588	30	1274	55	54	1.8
Ear. Det. (min)	IDS (hrs)	Red. ROG (gr)	Red. CO (gr)	Red. NOx (gr)	Red. PM2.5 (gr)	Red. Gas (gal)	Red. Dis (gal)
5	267	588	13100	1604	27	190	61

Figure 43. Flow and speed during sample accident #1 at GE

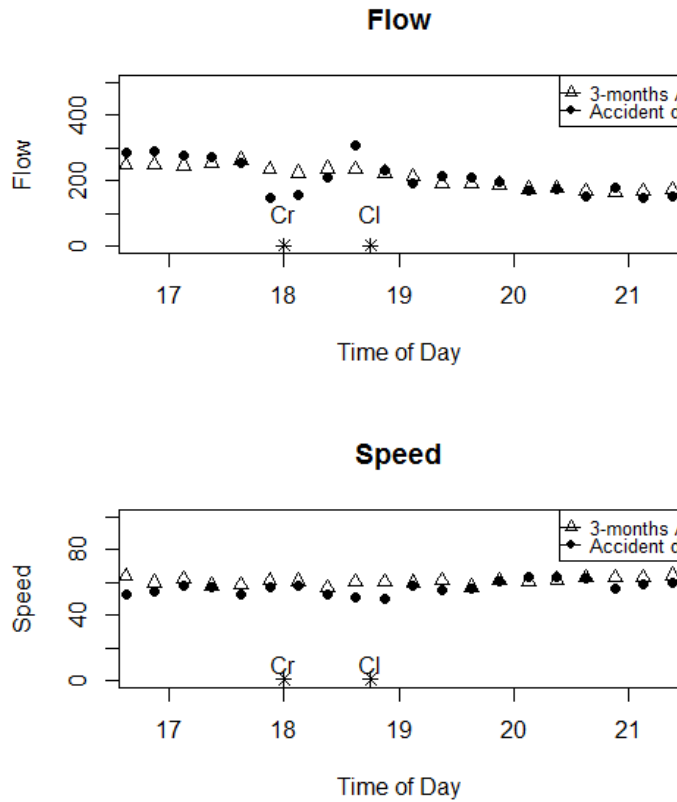


Table 20. Potential benefits of five-minute early detection for sample accident #2 at GE

St.ID	Direction	Cross Street	Created Date	Created Time	Cleared Time	Lanes Affected	Vehicles Involved
GE1	East	65th Street	Wednesday, February 24, 2016	6:14 a.m. (25)	7:04 a.m. (29)	Left Lane	1A 1TT
Acc. #	Regular Flow (veh/hr)	Accident Flow (veh/hr)	Flow Drop (%)	Delay (hrs)	Regular Speed (mi/hr)	Accident Speed (mi/hr)	Speed Drop (%)
2	1605	908	43	17187	42	50	0
Ear. Det. (min)	IDS (hrs)	Red. ROG (gr)	Red. CO (gr)	Red. NOx (gr)	Red. PM2.5 (gr)	Red. Gas (gal)	Red. Dis (gal)
5	2215	4873	108546	13291	221	1579	506

Figure 44. Flow and speed during sample accident #2 at GE

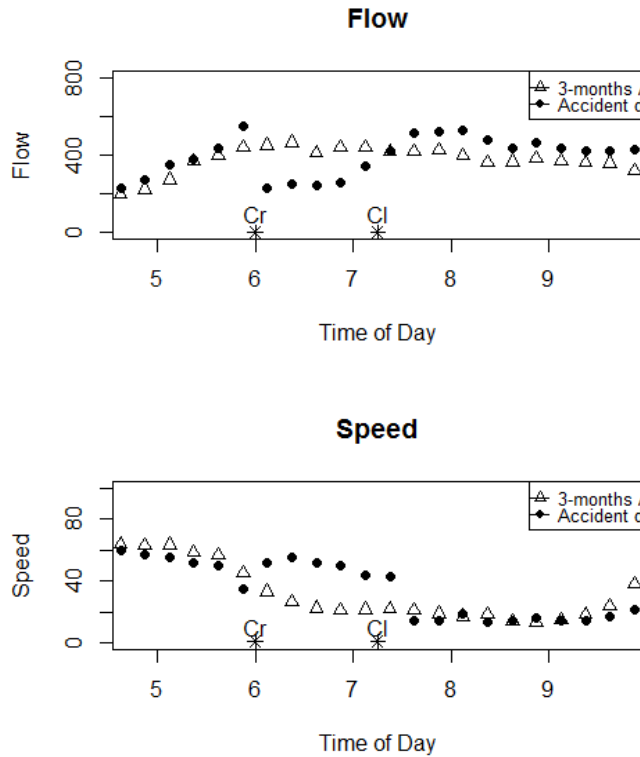
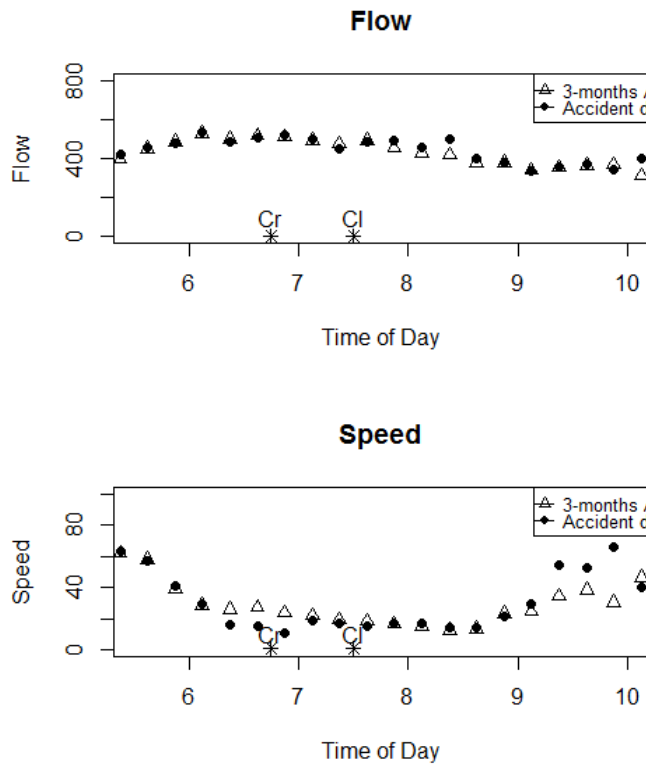


Table 21. Potential benefits of five-minute early detection for sample accident #3 at GE

St.ID	Direction	Cross Street	Created Date	Created Time	Cleared Time	Lanes Affected	Vehicles Involved
GE1	East	65th Street	Thursday, May 12, 2016	6:51 a.m. (28)	7:18 a.m. (30)	Left Lane	1A
Acc. #	Regular Flow (veh/hr)	Accident Flow (veh/hr)	Flow Drop (%)	Delay (hrs)	Regular Speed (mi/hr)	Accident Speed (mi/hr)	Speed Drop (%)
3	1833	1800	2	151	15.5	15.6	0
Ear. Det. (min)	IDS (hrs)	Red. ROG (gr)	Red. CO (gr)	Red. NOx (gr)	Red. PM2.5 (gr)	Red. Gas (gal)	Red. Dis (gal)
5	31	70	1552	190	3	23	7

Figure 45. Flow and speed during sample accident #3 at GE



These individual savings can be generalized using a Monte Carlo simulation for accident duration distributions calculated in Section 3.1. Since there is no matched Twitter information for early accident detection in the current database, a hypothetical scenario is considered to evaluate the impacts of early detection on delay, fuel consumption, and emissions. The Monte Carlo simulation is based on log-normal distributions for accident durations on both the GE and LIE. Regarding the accuracy of Twitter feeds for early detection, the percentage of accidents which were identified through Twitter was assumed to range from 0.1% to 1%. The maximum capacity, inflow rate, and accident capacity are assumed to be 2,250; 2,000; and 800 vehicles per hour, respectively. All other values related to fuel consumption and released emissions are the same as in Section 3.3. The results depicted in Figure 46 to Figure 49, show that even a small percentage of accident early detection creates considerable reductions in traffic delays, fuel consumption, and emissions.

Figure 46. Impact of early incident detection on delay saving with respect to varying levels of twitter feed accuracy

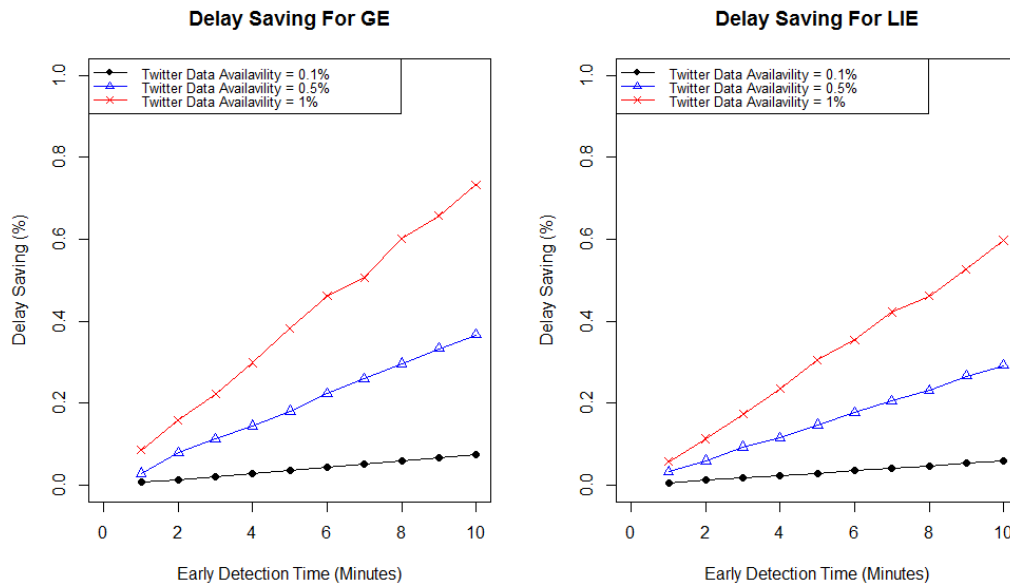


Figure 47. Impact of early incident detection on fuel saving with respect to varying levels of twitter feed incident detection accuracy

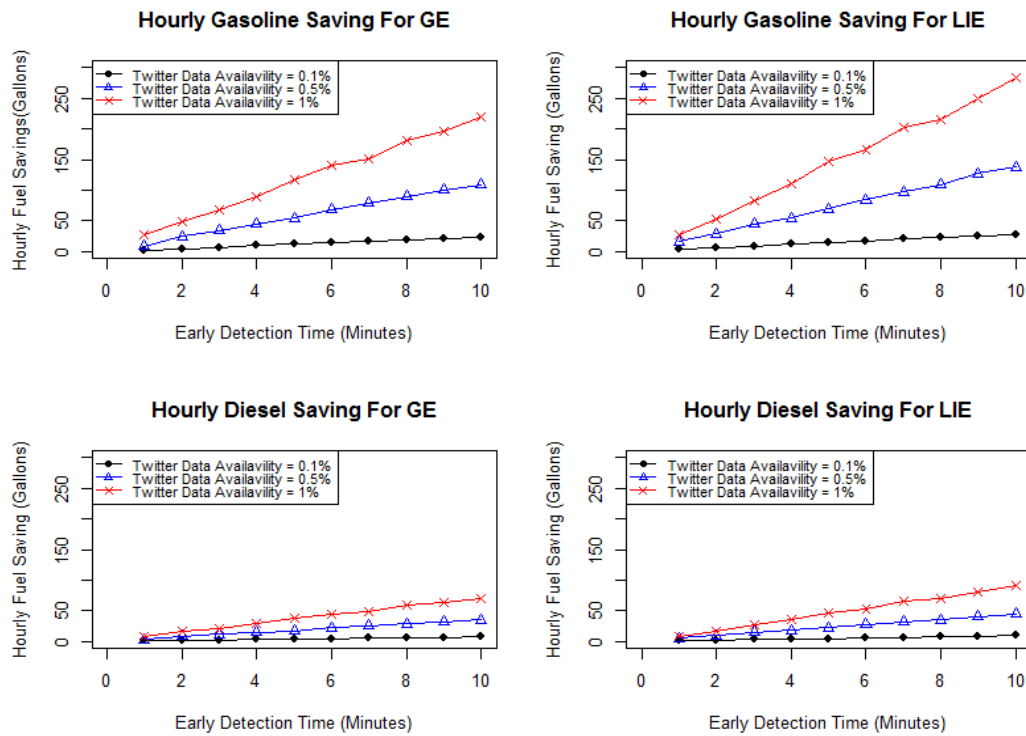


Figure 48. Impact of early incident detection on fuel saving with respect to varying levels of twitter feed incident detection accuracy in GE

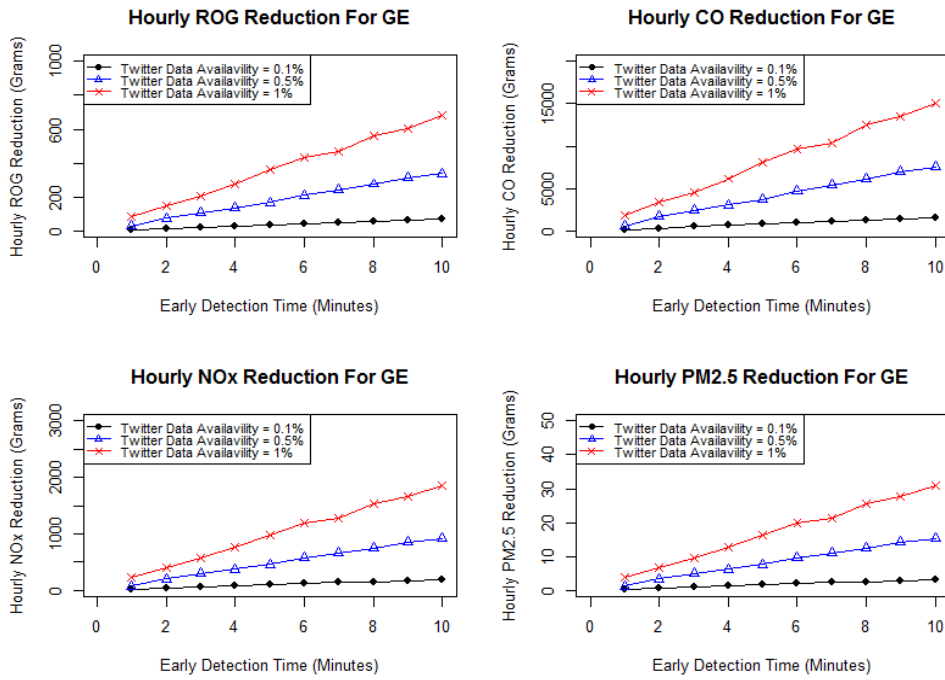
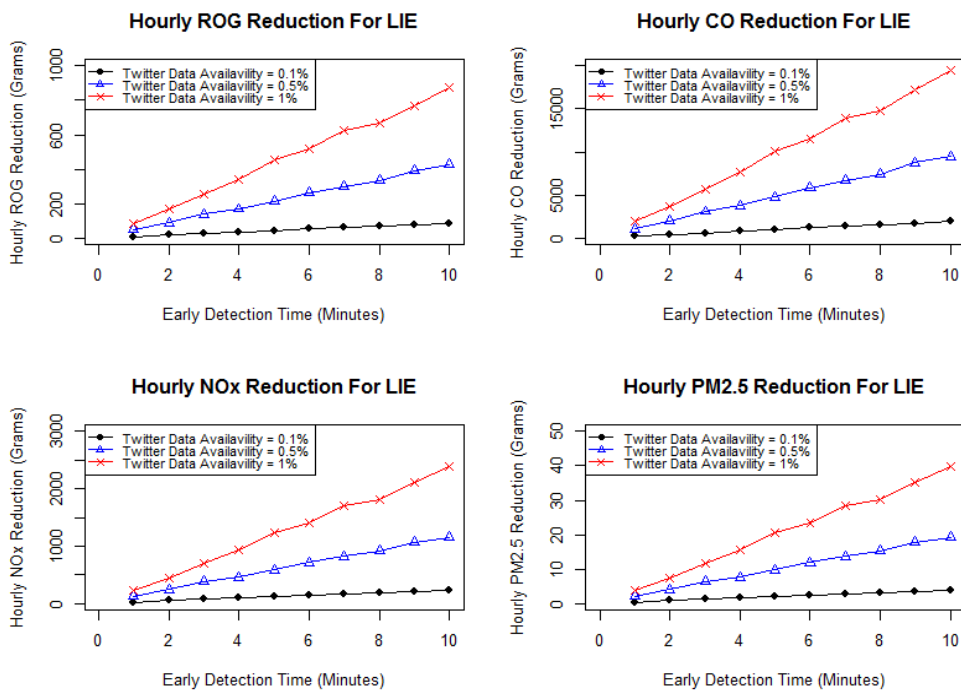


Figure 49. Impact of early incident detection on fuel saving with respect to varying levels of twitter feed incident detection accuracy in LIE



6 Conclusions and Recommendations

The usefulness of social media for incident management is presented in this report. One of the important findings is the need to target information from certain account types in order to extract useful incident-related information. On the one hand, organizational accounts (e.g., 511) disseminate traffic incident information in a structured manner (proper grammar and spelling) and provide important incident details (location, type of incident, severity, etc.). This information is already available to the agencies. On the other hand, personal account tweets are not structured, have language errors, and do not include many incident details. While the features of organizational accounts make it relatively easy to detect events, it was noted that the information disseminated by organizational accounts were more likely to have *already* been conveyed to the necessary incident management units. A personal tweet is more likely to report an event that “just-happened” than an organizational account tweet. Thus, the extracted information from personal accounts are more likely to be timely/useful.

For this reason, the personal and organizational tweets were treated separately and “dictionaries” to perform relevancy classification derived separately. Combinations of dictionaries (i.e., personal-only, organizational-only, personal and organizational) were used for “term frequency – inverse document frequency” (tf-idf) scoring. The Naïve Bayesian analysis following the tf-idf scoring has shown that the classifier is more accurate if trained using dictionaries that distinguish between personal and organizational accounts. Subsequently, personal tweets that were collected were more relevant when using the personal dictionary as compared to when using the organizational dictionary only. This finding implies that obtaining customized dictionaries for targeted account types is crucial for more efficient and effective incident detection for incident-management purposes.

In the report we provide several examples of tweets from personal accounts that can be mined to yield useful information. Information such as incidents on local, rural, and less instrumented roadways can be useful for incident management. Additionally, supplementary information on incidents can be gathered to monitor the evolution of incidents.

Due to the lack of a comprehensive incident database in New York State, the benefits of using social media for TIM could not be estimated for incidents in the State. To demonstrate and illustrate the benefits of using social media for TIM, more spatially and temporally detailed incident data from the CHP were used to match tweets collected in California. Using tweet and incident data from six weeks in total 21 traffic incident tweets were uniquely matched the recorded incidents. Three tweets preceded

the incident reported time by 19, 23, and 4 minutes, respectively. For those early detected accidents, reductions in accident delay, emissions, and fuel consumption were calculated using the flow and speed data from the Performance Measurement System (PeMS) database. As a result of the early detection, 4,046 vehicle-hours of delay savings, reduction in 5.9 kg of ROG, 133 kg of CO, 16.3 kg of NO_x, 0.3 kg of PM 2.5, 1,939 gal of gasoline, and 622 gal of diesel were estimated to be saved – a monetary value of \$75,600 or \$0.5 per mile per week.

The information gathered from social media for incident management can be enhanced by:

1. Use of structured hashtags: With the provision of structured hashtags, highly specific location information can be provided to the agencies without users worrying about their privacy in providing/revealing the exact geolocation in their tweets. These structured hashtags can also provide a means of defining incident type. Finally, collection of tweet data using specific hashtags is much easier than scraping Twitter feeds for specific information which needs Twitter APIs, text mining, etc.
2. Use of tweets from individual accounts: As mentioned earlier, individual tweets can be harvested and mined for useful information such as:
 - local events where instrumentation may be scarce
 - extracting supplementary information for incident verification, response, and monitoring evolution of incident dynamics.
 - information extraction about debris on roadways which can be useful in preventing incidents, if analyzed in time.
3. Use of tweets from local businesses: Local businesses located along the route can also provide useable information for incident response and driver information dissemination.

In summary, the study illustrated the potential of real-time crowd-sourced data (Twitter feeds) as a valuable incident information source, which agrees with previous literature on the topic. Due to the lack of a comprehensive incident database in New York State, the benefits of using social media for TIM could not be estimated for incidents in the State. However, in this study, we demonstrated and investigated the efficiency of the approach by matching incidents extracted from Twitter with actual incidents, calculating the rate of early detection and estimate the benefits for a sample of incidents and tweets in California. The potential impacts of using social media for TIM have been calculated for New York State incidents based on Twitter feed accident detection accuracy and early detection scenarios, and it was shown that the feeds can help achieve substantial economic and environmental benefits. The potential of social media for TIM could further be illustrated in the future by locating more comprehensive incident databases for the New York State so that more specific recommendations can be made regarding the use of Twitter in TIM.

7 Important Remark: Review of Hazards Associated with Using Mobile Devices in Vehicles

Given the increased use of devices providing connectivity, especially mobile phones, users have constant access to information from sources such as news, video, and social media. In the context of motor vehicle drivers, this access to information may prove to be a distraction to the primary activity of driving.

Generally speaking, distractions for motor vehicle drivers can be classified into three categories:²

- 1) manual distractions involving moving hands away from the task of controlling vehicles
- 2) visual distractions involving tasks that lead to drivers taking their eyes away from the road
- 3) cognitive distractions involving tasks that distract drivers' minds away from the task of driving

The risks of distracted driving are well-established in the literature. A study by the Fatality Accident Reporting Systems (FARS) showed that the proportion of distraction-related fatalities increased from 10.9% in 1999 to 15.8% in 2008. One reason behind this increase is the increased frequency of texting while driving, which is considered dangerous since it involves all three types of distractions listed above. Texting while driving results in injuries and fatalities.⁴² Driver distraction due to cell phone use increases crash risks by 2.8 to 5 times. The increased usage of cell phones has been accompanied by an increase in the number of traffic accidents.⁴³

The New York State vehicle and traffic law for distracted driving, talking, and texting, with regards to operating a phone or an electronic device while driving states the following:³ (VTL 1225-c, VTL 1225-d)

² <http://www.enddd.org/the-facts-about-distracted-driving/>

³ New York Vehicle Traffic Law 1225-c & -d Use of Mobile Telephone <http://www.safeny.ny.gov/phon-vt.htm> (accessed April 11, 2016).

Article 33, section 1225-c. Use of mobile telephones:

“(e) "Hands-free mobile telephone" shall mean a mobile telephone that has an internal feature or function, or that is equipped with an attachment or addition, whether or not permanently part of such mobile telephone, by which a user engages in a call without the use of either hand, whether or not the use of either hand is necessary to activate, deactivate or initiate a function of such telephone. Provided, however, that for purposes of this section, a mobile telephone used by a person operating a commercial motor vehicle shall not be deemed a "hands-free mobile telephone" when such person dials or answers such mobile telephone by pressing more than a single button.”

“...2. (a) Except as otherwise provided in this section, no person shall operate a motor vehicle upon a public highway while using a mobile telephone to engage in a call while such vehicle is in motion; provided, however, that no person shall operate a commercial motor vehicle while using a mobile telephone to engage in a call on a public highway including while temporarily stationary because of traffic, a traffic control device, or other momentary delays. Provided further, however, that a person shall not be deemed to be operating a commercial motor vehicle while using a mobile telephone to engage in a call on a public highway when such vehicle is stopped at the side of, or off, a public highway in a location where such vehicle is not otherwise prohibited from stopping by law, rule, regulation or any lawful order or direction of a police officer.

(b) An operator of any motor vehicle who holds a mobile telephone to, or in the immediate proximity of, his or her ear while such vehicle is in motion is presumed to be engaging in a call within the meaning of this section; provided, however, that an operator of a commercial motor vehicle who holds a mobile telephone to, or in the immediate proximity of, his or her ear while such vehicle is temporarily stationary because of traffic, a traffic control device, or other momentary delays is also presumed to be engaging in a call within the meaning of this section except that a person operating a commercial motor vehicle while using a mobile telephone to engage in a call when such vehicle is stopped at the side of, or off, a public highway in a location where such vehicle is not otherwise prohibited from stopping by law, rule, regulation or any lawful order or direction of a police officer shall not be presumed to be engaging in a call within the meaning of this section. The presumption established by this subdivision is rebuttable by evidence tending to show that the operator was not engaged in a call.

(c) The provisions of this section shall not be construed as authorizing the seizure or forfeiture of a mobile telephone, unless otherwise provided by law.

(d) No motor carrier shall allow or require its drivers to use a hand-held mobile telephone while operating a commercial motor vehicle as provided in this section.

3. Subdivision two of this section shall not apply to (a) the use of a mobile telephone for the sole purpose of communicating with any of the following regarding an emergency situation: an emergency response operator; a hospital, physician's office or health clinic; an ambulance company or corps; a fire department, district or company; or a police department, (b) any of the following persons while in the performance of their official duties: a police officer or peace officer; a member of a fire department, district or company; or the operator of an authorized emergency vehicle as defined in section one hundred one of this chapter, or (c) the use of a hands-free mobile telephone..."

Article 33, section 1225-d. Use of portable electronic devices:

"...1. Except as otherwise provided in this section, no person shall operate a motor vehicle while using any portable electronic device while such vehicle is in motion; provided, however, that no person shall operate a commercial motor vehicle while using any portable electronic device on a public highway including while temporarily stationary because of traffic, a traffic control device, or other momentary delays. Provided further, however, that a person shall not be deemed to be operating a commercial motor vehicle while using a portable electronic device on a public highway when such vehicle is stopped at the side of, or off, a public highway in a location where such vehicle is not otherwise prohibited from stopping by law, rule, regulation or any lawful order or direction of a police officer."

"...3. Subdivision one of this section shall not apply to (a) the use of a portable electronic device for the sole purpose of communicating with any of the following regarding an emergency situation: an emergency response operator; a hospital; a physician's office or health clinic; an ambulance company or corps; a fire department, district or company; or a police department, (b) any of the following persons while in the performance of their official duties: a police officer or peace officer; a member of a fire department, district or company; or the operator of an authorized emergency vehicle as defined in section one hundred one of this chapter.

4. A person who holds a portable electronic device in a conspicuous manner while operating a motor vehicle or while operating a commercial motor vehicle on a public highway including while temporarily stationary because of traffic, a traffic control device, or other momentary delays but not including when such commercial motor vehicle is stopped at the side of, or off, a public highway in a location where such vehicle is not otherwise prohibited from stopping by law, rule, regulation or any lawful order or direction of a police officer is presumed to be using such device, except that a person operating a commercial motor vehicle while using a portable electronic device when such vehicle is stopped at the side of, or off, a public highway in a location where such vehicle is not otherwise prohibited from stopping by law, rule, regulation or any lawful order or direction of a police officer shall not be presumed to be using such device. The presumption established by this subdivision is rebuttable by evidence tending to show that the operator was not using the device within the meaning of this section..."

The research team clearly recognizes the risks involved in distracted driving, especially involving the usage of mobile phones. The team also understands the New York State traffic law for distracted driving, talking, and texting and the provisions therein. This research study involves using user tweets that report road conditions, especially those involving road incidents. However, this study *does not* entail gathering a voluntary team of users and requesting that they tweet about incidents encountered. Nor does the study encourage or even suggest that users become involved in tweeting activities while driving. The source of information for the study is solely a set of tweets that were already posted, i.e., historical tweet databases.

However, there are many alerts posted on variable message signs by DOTs and authorities managing roadway facilities such as amber alerts, silver alerts, etc. (as seen in Figure 50) that request drivers to inform the authorities about any information pertaining to the alert. This information is supposed to be transmitted in a safe manner by drivers, possibly by stopping on the side of the road and making a phone call or text transmission or by using a hands-free mobile telephone as indicated in New York State traffic law article 33, sections 1225-c and -d. (VTL 1225-c, VTL 1225-d).

Figure 50. Amber Alerts



Procedures exist that provide a viable and safe alternative to distractions such as texting while driving, etc. Speech-based cell phone use is less disruptive to driving performance than handheld cell phone use, such as texting or typing.⁴⁶ Another study by Cao⁴⁷ showed that performing a secondary speech comprehension task may not affect the performance of primary driving activities such as lane keeping, although concurrent comprehension increased drivers' mental workload and reduced drivers' capability to comprehend speech correctly. In addition, a study by Owens⁴⁸ found that speech-based interaction reduced the number of glances, the total glance durations and subjective mental demand compared to

handheld interactions for dialing a phone. Crisler et al.⁴⁹ showed that a manual texting task requires the driver to look at the phone and press the correct buttons, while an audio cell phone interaction or conversation requires less visual and manual distraction.

In conclusion, it should be noted that this study does not involve active collection of tweets from drivers for the project; rather, the study uses a historical set of tweets that have already been posted. In other words, the researchers have not requested drivers to tweet during traffic incidents for the purpose of this project. Thus, all of the tasks in the study are in compliance with New York State Vehicle and Traffic law article 33, sections 1225-c and -d.

8 Statement on Implementation

As discussed throughout this report, social media tools have been identified as one of the top trends and technologies for transportation incident management. Social networks such as Twitter are growing in popularity and can be employed by transportation management agencies as a source for real-time information. Through the use of social media, members of the public who witness incidents can provide public safety organizations with timely, geographic-based information. Although the amount of information that can be collected and processed in real time by these systems still presents some significant challenges, this information can be used by decision-makers in planning response strategies, deploying resources in the field, and, in turn, providing updated and accurate information to the public.

Therefore, some transportation agencies are using or considering using data from social media to support decision-making in an operational context. For adoption in practice, the need for tools to search and filter social media content in real time has emerged. According to a report from Science and Technology of the U.S. Department of Homeland Security, there are now dozens of applications available to search and monitor social media information streams for specific keywords and mentions.⁴ The 2013 Homeland Security report documented that the majority of the major players in this area offer some kind of free service that allows organizations to monitor social media streams by offering free search capabilities from their web sites or free widgets that can be incorporated as part of a web page.

Transportation Management Centers (TMC) are recognizing the benefits of deploying social media tools to support transportation incident management. Some state DOTs are developing data sharing partnerships with crowd-sourced organizations like Waze. For example, Georgia DOT, Virginia DOT, Pennsylvania DOT, and North Carolina DOT have entered into a data sharing partnership with Waze, the real-time, crowd-sourced navigation application, to reduce congestion and improve travel information. Partnerships with third-party data providers are recommended given the fact that DOTs have limited experience in processing big data and designing mobile applications. A national survey on applying advanced technologies at TMC (TRB Paper, 15-0290)⁵ found that TMCs consider the main obstacle to deploying new technologies as institutional, technical, and financial issues.

⁴ Innovative Uses of Social Media in Emergency Management, Science and Technology, U.S. Department of Homeland Security, September 2013

⁵ Jin et al. (2015) Potential for Applying Advanced Technologies at TMCs – Results from a Nationwide Survey, TRB Paper # 15-0290

For some TMCs cybersecurity is the main issue, while for another it is legal issues, and for others it is staffing. Other factors to consider by agencies for implementation of social media tools are the direct and indirect costs of these tools. Agencies should consider developing new measures of effectiveness to better evaluate the benefits of these tools for their operations.

9 References

1. Kushal, D., Lawrence, S., Pennock, D.M., 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *Proc. of the 12th Int. Conf. on World Wide Web.* 519–528.
2. Pang, B., Lee, L., 2008. Opinion mining and sentiment analysis, *Found. Trends Inf. Retr.* 2, 1–135.
3. Grant-Muller, S. M., Gal-Tzur, A., Minkov, E., Nocera, S., Kuflik, T., Shoor, I., 2014. Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data. *IET Intelligent Transport Systems.* 9(4), 407-417.
4. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M., 2010. Predicting elections with twitter: what 140 characters reveal about political sentiment. *Proc. of the Fourth Int. AAAI Conf. on Weblogs and Social Media.*
5. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A., 2010. From tweets to polls: linking text sentiment to public opinion time series. *Proc. of the Fourth Int. AAAI Conf. on Weblogs and Social Media (ICWSM), Washington, DC,* 122–129.
6. Tufekci, Z., Freelon, D., 2013. Introduction to the special issue on new media and social unrest. *American Behavioral Scientist.*
7. Corley, C., Cook, D., Mikler, A., Singh, K., 2010. Text and structural data mining of influenza mentions in web and social media', *Int. J. Environ. Res. Public Health.* 7 (2), 596–615.
8. Grishman, R., Huttunen, S., Yangarber, R., 2002. Information extraction for enhanced access to disease outbreak reports. *J. Biomed. Inform.* 35 (4), 236–246.
9. Imran, M., Elbassuoni, S. M., Castillo, C., Diaz, F., Meier, P., 2013. Extracting information nuggets from disaster-related messages in social media. *Proc. of ISCRAM, Baden-Baden, Germany*
10. Goodchild, M. F., Glennon, J. A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth.* 3(3), 231-241.
11. Bregman, S., 2012. Uses of social media in public transportation. *Transit Cooperative Research Program (TCRP) Synthesis.*
12. Jin, P., Cebelak, M., Yang, F., Zhang, J., Walton, C., Ran, B., 2014. Location-based social networking data: exploration into use of doubly constrained gravity model for origin-destination estimation. *Transportation Research Record: Journal of the Transportation Research Board.* 2430, 72-82.
13. Efthymiou, D., Antoniou, C., 2012. Use of social media for transport data collection. *Procedia-Social and Behavioral Sciences.* 48, 775-785.

14. Schulz, A., Ristoski, P., Paulheim, H., 2013. I see a car crash: Real-time detection of small scale incidents in microblogs. In *The Semantic Web: ESWC 2013 Satellite Events*. Springer Berlin Heidelberg.
15. Kurkcu, A., Morgul, E., Ozbay, K., 2015. Extended Implementation Method for Virtual Sensors: Web-Based Real-Time Transportation Data Collection and Analysis for Incident Management, *Transportation Research Record: Journal of the Transportation Research Board*. 2528, 27-37.
16. D'Andrea, E., Ducange, P., Lazzerini, B., Marcelloni, F., 2015. Real-time detection of traffic from twitter stream analysis. *Intelligent Transportation Systems, IEEE Transactions on*. 16(4), 2269-2283.
17. Lécué, F., Tallevi-Diotallevi, S., Hayes, J., Tucker, R., Bicer, V., Sbodio, M. L., Tommasi, P., 2014. Star-city: semantic traffic analytics and reasoning for city. In *Proceedings of the 19th international conference on Intelligent User Interfaces*.179-188.
18. Haghani, A., Iliescu, D., Hamed, M., Yang, S., 2006. Methodology for quantifying the cost effectiveness of freeway service patrols programs. I-95 Coalition Report. University Of Maryland.
19. Evanco, W. M. (1996) *The Impact of Rapid Incident Detection on Freeway Accident Fatalities*. Miretek, FHWA Project No. 049518C10A.
20. Pereira, F. C., Bazzan, A. L., Ben-Akiva, M.,2014. The role of context in transport prediction. *IEEE Intelligent Systems*. 1, 76-80.
21. Mai, E., Hranac, R., 2013. Twitter interactions as a data source for transportation incidents. *Transportation Research Board 92nd Ann. Meeting*.
22. Fu, K., Nune, R., Tao, J. X., 2015. Social media data analysis for traffic incident detection and management. *Transportation Research Board 94th Annual Meeting*
23. 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.
24. Gal-Tzur, A., Grant-Muller, S. M., Minkov, E., Nocera, S., 2014. The impact of social media usage on transport policy: issues, challenges and recommendations. *Procedia-Social and Behavioral Sciences*. 111, 937-946.
25. Xia, C., Schwartz, R., Xie, K. E., Krebs, A., Langdon, A., Ting, J., Naaman, M., 2014. Citybeat: Real-time social media visualization of hyper-local city data. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*. International World Wide Web Conferences Steering Committee.
26. Naaman, M., Boase, J., Lai, C. H., 2010. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 189-192.
27. Weng, J., Lee, B. S., 2011. Event Detection in Twitter. *ICWSM*. 11, 401-408.

28. Collins, C., Hasan, S.,Ukkusuri, S. V.,2013. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. *Journal of Public Transportation*. 16(2), 21-45.
29. Pender, B., Currie, G., Delbosc, A., Shiwakoti, N., 2013. Social Media Utilisation during Unplanned Passenger Rail Disruption What's Not to Like?. In *Proc. Australasian Transport Research Forum*.
30. Anastasi, G., Antonelli, M., Bechini, A., Brienza, S., D'Andrea, E., De Guglielmo, D., Segatori, A., 2013. Urban and social sensing for sustainable mobility in smart cities. In *Sustainable Internet and ICT for Sustainability (SustainIT)*.
31. Rosi, A., Mamei, M., Zambonelli, F., Dobson, S., Stevenson, G., Ye, J., 2011. Social sensors and pervasive services: approaches and perspectives. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), IEEE International Conference*.
32. Liu, X., Lang, B., Yu, W., Luo, J., Huang, L.,2011. AUDR: an advanced unstructured data repository. In *Pervasive Computing and Applications (ICPCA), 6th International Conference*. 462-469.
33. Amitay, E., Har'El, N., Sivan, R., Soffer, A., 2004. Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*.
34. Paradesi, S. M.,2011. Geotagging Tweets Using Their Content. In *FLAIRS Conference*.
35. Atefeh, F., Khreich, W., 2015. A survey of techniques for event detection in Twitter. *Comput. Intell.* 31(1),132-164.
36. Wanichayapong, N., Pruthipunyaskul, W., Pattara-Atikom, W., Chaovalit, P., 2011. Social-based traffic information extraction and classification. In *ITS Telecommunications (ITST), 11th International Conference*.
37. J. Allan, 2002. *Topic Detection and Tracking: Event-Based Information Organization*. Norwell, MA, USA: Kluwer.
38. USDOT Intelligent Transportation Systems Joint Program Office, 2015. T3 Webinar: Using Crowdsourced Data from Social Media to Enhance TMC Operations, assessed on March 11 2015 at: https://www.pcb.its.dot.gov/t3/s150311_crowdsourced_data.asp.
39. Twitter Public Search API, <https://dev.twitter.com/streaming/public>.
40. Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß. "A brief survey of text mining." *Ldv Forum*. Vol. 20. No. 1. 2005.
41. G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513-523, 1988.
42. Caird, J.K., Willness, C.R., Steel, P., Scialfa, C., 2008. A meta-analysis of the effects of cell phones on driver performance. *Accident Analysis Prevention*. 40(4), 1282-1293.

43. Wilson, F.A., Stimpson, J.P., 2010. Trends in fatalities from distracted driving in the United States, 1999-2008. *Am J Public Health*. 100(11), 2213-2219.
44. New York Vehicle Traffic Law 1225-c Use of Mobile Telephone <http://www.safeny.ny.gov/phon-vt.htm> (accessed April 11, 2016).
45. New York Vehicle Traffic Law 1225-d Use of Portable Electronic Devices (includes texting while driving) <http://www.safeny.ny.gov/phon-vt.htm> (accessed April 11, 2016)
46. He, J., Chaparro, A., Nguyen, B., Burge, R.J., Crandall, J., Chaparro, B., Ni, R., Cao, S., 2014. Texting while driving: is speech-based text entry less risky than handheld text entry?. *Accident Analysis Prevention*. 72, 287-295.
47. Cao, S., Liu, Y., 2013. Concurrent processing of vehicle lane keeping and speech comprehension tasks. *Accident Analysis Prevention*. 59, 46-54.
48. Owens, J.M., McLaughlin, S.B., Sudweeks, J., 2011. Driver performance while text messaging using handheld and in-vehicle systems. *Accident Analysis Prevention*. 43, 939-947.
49. Crisler, M.C., Brooks, J.O., Ogle, J.H., Guirl, C.D., Alluri, P., Dixon, K.K., 2008. Effect of wireless communication and entertainment devices on simulated driving performance. *Transportation Research Record*. 2069, 48-54.
50. Cambridge Systematics, Inc. in association with JHK & Associates, Transmode Consultants, Inc., and Sydec, Inc., "Incident Management," 1990.
51. Y. Chung, "Development of an accident duration prediction model on the Korean Freeway Systems," *Accident Analysis & Prevention*, vol. 42, no. 1, p. 282–289, 2010.
52. Younshik Chung, Lubinda Walubita, Keechoo Choi, "Modeling Accident Duration and Its Mitigation Strategies on South Korean Freeway Systems," *Transportation Research Record*, vol. 2178, p. 49–57, 2010.
53. Kang, G. and Fang, S, "Applying Survival Analysis Approach to Traffic Incident Duration Prediction," in First International Conference on Transportation Information and Safety (ICTIS), Wuhan, China, 2011.
54. Wang Junhua, Cong Haozhe, Qiao Shi, "Estimating freeway incident duration using accelerated failure time modeling," *Safety Science*, vol. 54, p. 43–50, 2013.
55. Abdulla Alkaabi, Dilum Dissanayake, Roger Bird, "Analyzing Clearance Time of Urban Traffic Accidents in Abu Dhabi, United Arab Emirates, with Hazard-Based Duration Modeling Method," *Transportation Research Record*, vol. 2229, 2014.
56. Ahmad Tavassoli Hojatia, Luis Ferreira, Simon Washington, Phil Charles, "Hazard based models for freeway traffic incident duration," *Accident Analysis & Prevention*, vol. 52, p. 171–181, 2013.

57. Younshik Chung, Byoung-Jo Yoon, "Analytical method to estimate accident duration using archived speed profile and its statistical analysis KSCE Journal of Civil Engineering, vol. 16, no. 6, p. 1064–1070, 2012.
58. R. Li, "Traffic incident duration analysis and prediction models based on the survival analysis approach IET Intelligent Transport Systems, vol. 9, no. 4, pp. 351-358, 2014.
59. J. H. Banks, Introduction to transportation engineering, Tata Mc-Graw Hill, 2004.
60. Office of Traffic Operations, "Benefits Analysis for the Georgia Department of Transportation NaviGator Program Georgia Department of Transportation, Atlanta, GA, 2006.
61. C. A. R. Board, "Methods to Find the Cost-Effectiveness of Funding Air Quality Projects: Emission Factor Tables 2013.
62. Office of Transportation and Air Quality, "Average Annual Emissions and Fuel Consumption for Passenger Cars and Light Trucks United States Environmental Protection Agency (EPA), 2008.
63. Brodrick, C. J. et al, "Evaluation of fuel cell auxiliary power units for heavy-duty diesel trucks Transportation Research Part D: Transport and Environment, vol. 7, no. 4, p. 303–315, 2002.
64. "The official government source for fuel economy information [Online]. Available: www.fueleconomy.gov. [Accessed 23 4 2017].
65. Twitter Public Search API, <https://dev.twitter.com/streaming/public>.
66. Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., Tao, K., 2012. Twitcident: fighting fire with information from social web streams. In Proceedings of the 21st international conference companion on World Wide Web, ACM, 305-308.
67. Gu, Y., Qian, Z., Chen, F., From Twitter to detector: Real-time traffic incident detection using social media data, Transportation Research Part C 67 (2016) 321–342.
68. California Department of Transportation, Performance Measurement System (PeMS), accessible at pems.dot.ca.gov
69. California Highway Patrol, Statewide Integrated Traffic Records System (SWITRS) accessible at <http://iswitrs.chp.ca.gov/Reports/jsp/userLogin.jsp>
70. California Department of Transportation, Life-Cycle Benefit-Cost Analysis Economic Parameters 2016, accessible at: http://www.dot.ca.gov/hq/tpp/offices/eab/benefit_cost/LCBCA-economic_parameters.html

NYSERDA, a public benefit corporation, offers objective information and analysis, innovative programs, technical expertise, and support to help New Yorkers increase energy efficiency, save money, use renewable energy, and reduce reliance on fossil fuels. NYSERDA professionals work to protect the environment and create clean-energy jobs. NYSERDA has been developing partnerships to advance innovative energy solutions in New York State since 1975.

To learn more about NYSERDA's programs and funding opportunities, visit nyserda.ny.gov or follow us on Twitter, Facebook, YouTube, or Instagram.

**New York State
Department of Transportation**

50 Wolf Road
Albany, NY 12232

telephone: 518-457-6195

dot.ny.gov

**New York State
Energy Research and
Development Authority**

17 Columbia Circle
Albany, NY 12203-6399

toll free: 866-NYSERDA
local: 518-862-1090
fax: 518-862-1091

info@nyserda.ny.gov
nyserda.ny.gov



NYSERDA
Department of
Transportation

State of New York

Andrew M. Cuomo, Governor

New York State Energy Research and Development Authority

Richard L. Kauffman, Chair | Alicia Barton, President and CEO

New York State Department of Transportation

Paul A. Karas, Acting Commissioner