TransInfo
Tier I University Transportation Center

FINAL REPORT

Mining Transportation Information from Social Media for Planned and Unplanned Events

Date of report: May 1, 2016

Zhenhua Zhang, Graduate Student, University at Buffalo, SUNY
Ming Ni, Graduate Student, University at Buffalo, SUNY
Qing He, PhD, Stephen Still Assistant Professor, University at Buffalo, SUNY
Jing Gao, PhD, Assistant Professor, University at Buffalo, SUNY

Prepared by:
University at Buffalo, SUNY

Prepared for:
Transportation Informatics Tier I University Transportation Center
204 Ketter Hall
University at Buffalo
Buffalo, NY 14260

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Mining Transportation Information from Social Media for Planned and Unplanned Events		**5. Report Date** May 1, 2016 **6. Performing Organization Code**
7. Author(s) Zhenhua Zhang, Ming Ni, Qing He and Jing Gao		**8. Performing Organization Report No.**
9. Performing Organization Name and Address University at Buffalo, The State University of New York		**10. Work Unit No. (TRAIS** **11. Contract or Grant No.** DTRT13-G-UTC48
12. Sponsoring Agency Name and Address US Department of Transportation Office of the UTC Program, RDT-30 1200 New Jersey Ave., SE Washington, DC 20590		**13. Type of Report and Period Covered** Final March 1 2014 – Feb 28 2016 **14. Sponsoring Agency Code**

15. Supplementary Notes

16. Abstract

The objective of this project is on mining social media data to deduce useful traveler's information with a special emphasis under events, including both planned events (such as sporting games), and unplanned events (such as traffic accidents). Specifically, the project proposes to develop effective and efficient techniques to collect, extract and mine social media data to support advanced traveler information systems and traffic operators. By mining social media based semantics, especially text semantics, this project aims to achieve the following aims: 1) Forecast transit ridership under large sporting games; 2) Identify causality between abnormal traffic flow pattern and social media data; 2) Detect traffic accident using online social media data and traffic loop-detector data.

17. Key Words Social media; Event identification; Subway passenger flow prediction; Social sensing; Transit ridership; Traffic surge; Tweet analysis; Twitter concentration; Traffic accident detection; Association rules; Traffic signature		**18. Distribution Statement** No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161	
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 68	**22. Price**

Acknowledgements

University at Buffalo, Virginia Department of Transportation, New York City Department of Transportation

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Table of Contents

EXECUTIVE SUMMARY **5**

1 FORECASTING THE SUBWAY PASSENGER FLOW UNDER EVENT OCCURRENCES WITH SOCIAL MEDIA **6**

1.1 INTRODUCTION 8
1.2 RELATED WORKS 9
1.3 DATASET 10
1.4 HASHTAG-BASED EVENT IDENTIFICATION 12
1.5 EVENTS CHARACTERISTICS 16
1.6 PREDICTION MODELING 17
1.7 CONCLUSIONS 21

2 AN EXPLORATORY STUDY ON THE CORRELATION BETWEEN TWITTER CONCENTRATION AND TRAFFIC SURGE **23**

2.1 INTRODUCTION 23
2.2 DATA AND INCENTIVES 25
2.3 TRAFFIC CLUSTERING ALGORITHM 26
2.4 TRAFFIC SURGE DEFINITION IN DETECTOR LEVEL 29
2.5 TWITTER CONCENTRATION EXTRACTION AND FILTERING 30
2.6 TWITTER CONCENTRATION CLASSIFICATION AND LABELING 33
2.7 CORRELATIONS BETWEEN TWITTER CONCENTRATIONS AND TRAFFIC OPERATIONS 34
2.8 CONCLUSIONS AND DISCUSSIONS 37

3 TRAFFIC ACCIDENT DETECTION WITH BOTH TRAFFIC AND SOCIAL MEDIA DATA **39**

3.1 INTRODUCTION 39
3.2 DATA DESCRIPTION 42
3.3 CLASSIFICATION BY SVMS 45
3.4 COMPARING WITH THE CLASSIFICATION BY SLDA 52
3.5 IMPROVEMENTS OF CLASSIFICATION BY TRAFFIC-RELATED INFORMATION 53
3.6 COMPARISON WITH GROUND TRUTH 60
3.7 CONCLUSIONS AND DISCUSSIONS 62

REFERENCE **64**

Summary

This report focuses on the study of extracting useful information from social media and exploring its potentials in improving the transportation management and control. As a newly-emerged communication revolution, Twitter founds a perfect stage for users to communicate, share, and follow the interestingness happened in their daily life in an instantaneous channel. It brings about a newly revolutionary method for information diffusion and the huge volume of messages and information by Twitter has aroused the interests in many research areas such as geographic study, urban planning, movie reviews, public opinion polls, etc. Most of these researches prove promising results which both enhance the traditional methodologies and broaden the new research visions. The enlightening results also in turn provide new ideas in mining and extracting useful information from Twitter.

There are mainly two kinds of data used in our study: tweets and traffic data. We collect the tweets through Twitter Streaming API with geo-location filter. All tweets are paired with time and location information. The traffic data are real-time collection by the lane-based loop detectors.

Our studies show that the Twitter may reflect some major social events that arouse enormous interests of the public. The number of tweets posted online related to a social event can somewhat represent its corresponding attention levels. As most of the social events involve some kinds of trip requirements, there is usually an obvious traffic increase in the surrounding area. To prove this, our first study focuses on using the tweets related to sporting games to predict the subway passenger flow, which is strategically important in metro transit system management. The prediction under event occurrences turns into a very challenging task in most of the previous studies. Our empirical results demonstrate that there exists a moderate positive correlation between passenger flow and the rates of social media posts. On this basis, we develop a hashtag based event detection algorithm to find extract the tweets that are related to the game in the nearby stadium. The core of this algorithm is a parametric and convex optimization based approach, called Optimization and Prediction with hybrid Loss function (OPL), to fuse the linear regression and the results of seasonal autoregressive integrated moving average (SARIMA) model jointly. The OPL hybrid model takes advantage of the unique strengths of linear correlation in social media features and SARIMA model in time series prediction. Experiments on events nearby a subway station show that OPL reports the best forecasting performance compared with other state-of-the-art techniques. In addition, an ensemble model is developed to leverage the weighted results from OPL and support vector machine regression together. The prediction of the subway passenger flow is improved both in the accuracy and precision and the method proves a good robustness. According to this study, social media data show the capability in passenger flow prediction under event conditions in a cost-effective way. It also proves the validity of social media as a good indicator of passenger flow in the public transit system.

The major social events reflected on Twitter cause not only an unexpected increase in the public transit flow, but also traffic flow surge on the entire road network. The general traffic operations may deteriorate around major social events including ceremony opening, celebrity death, festival parades, international conference, etc. Our second study focuses on the general public events and explore the potentials of Twitter that can broadcast these events. These tweets that are related to major public events are called "Twitter concentrations" in this study. The study fuses a set of tweets and traffic data collected during the whole year of 2014 in North Virginia Region, and mainly investigates the correlation between Twitter concentration and traffic surge in July. Our algorithm starts extracting the tweets that contain certain keywords that can indicate a major event. Those keywords that appear frequently over a certain days but less frequently in other days are our major interests. Then, we build a traffic surge detection algorithm on the "big data" analysis of previous data collections. The algorithm can precisely unveil the traffic patterns in a large road networks and even identify the anomaly traffic conditions. Finally, we compare the Twitter concentrations with the traffic patterns around the corresponding tweets and find that 77.4% of traffic-related Tweeter concentrations can be justified by local traffic surge. The results show that the public activities behind Twitter concentrations potentially pose more pressure on traffic network and cause traffic surge within a specified time and location. This study can help traffic operators understand the cause of traffic surge and improve short-term prediction of traffic congestion (especially non-recurrent congestion) on roadways in the future. Furthermore, monitoring the social media data may deliver useful traffic event information, including traffic accident, traffic jam, road construction, etc. and the Twitter concentration can broadcast the traffic-related events in a much more timely and quickly manner than traditional broadcasting media.

As compared to the studies on the passenger flow increase and traffic surge and the tweets that broadcasting a major social event, Twitter can even broadcast much more severe events such as traffic accident which is also a major concern for traffic operators. Our third study employs social media to detect on-site traffic accidents and employ a supervised prediction model: support vector machines (SVMs) to automatically classify the accident-related tweets. We first explore the features of keywords and their association rules inherent in the accident-related tweets and explore the potentials of these keywords in accident detection. We use two types of token features: single tokens and paired tokens that may correlate with the traffic accident labels. Second, we build a regression database based on the token features and employ the SVMs to detect the traffic accidents from tweets. Our results show that paired tokens can possibly capture the association rules inherent in the accident-related tweets and the results are better than that of individual tokens. The combination of individual and paired tokens may not bring any increase of accuracy to the detection and this means that paired tokens can alone provide a good prediction results in an efficient way. The results are even better than that of supervised topic models: supervised latent Dirichlet allocation (sLDA). Third, we study on the traffic data collected by the loop detectors and prove that the traffic flows over a certain range of occupancy in a given cluster are observed to follow a Gaussian distribution. Using large-scale data, a new traffic-related features can be derived about the relationships between traffic flow and occupancy based on the fundamental diagram. The derived traffic-related information may provide limited improvement for accident prediction. The reasons

for these may be found by some empirical studies. Empirical comparisons between the prediction results and the traffic management log maintained by VDOT show that there sometimes exist time lag between the starting time of the traffic accident and the corresponding tweet. This means that sometimes the users may tweet about an accident after they have already drive away from the site. Thus, traffic information around where the tweets are posted may not improve the prediction results. Besides these tweets with temporal and spatial lags, we also find that tweets can sometimes respond to traffic accident much more quickly than traditional detecting methods and can even find some un-documented accidents which make up for the deficiencies of VDOT records. This means tweets can sometimes capture those "mild" accidents that do not incur the attention of traffic police and make up for the deficiencies of traffic management log. It is concluded that integrating social media data into the traffic-related study opens up a wide range of possibilities for research in on-site traffic accident detection.

Following these findings, one can see that social media data are noisy and sometimes unreliable, and there is still room to improve the models and results. One may further pursue the study by tackling the limitations of the current approach. In our study, the tweets are collected through Twitter Streaming API with geo-location filter and cannot possibly cover all the traffic surge of the whole region. This may be due to the limited volume of geo-tagged tweets. By incorporating more non-geo-tagged tweets, the precision of the results may increase. Further studies can focus on the data fusion of different data sources to better realize the purposes of other research fields such as traffic jam detection, traffic emergency evacuation, etc. The spatial-temporal features of tweets are also worth studying for regional traffic operations as higher coverage of the tweets may result in a better scholar purpose.

1 Forecasting the Subway Passenger Flow under Event Occurrences with Social Media

Subway passenger flow prediction is strategically important in metro transit system management. The prediction under event occurrences turns into a very challenging task. In this paper, we adopt a new kind of data source -- social media to tackle this challenge. We develop a systematic approach to examine social media activities and sense event occurrences. Our initial analysis demonstrates that there exists a moderate positive correlation between passenger flow and the rates of social media posts. This finding motivates us to develop a novel approach for improved flow forecast. We first develop a hashtag based event detection algorithm. Further, we propose a parametric and convex optimization based approach, called Optimization and Prediction with hybrid Loss function (OPL), to fuse the linear regression and the results of seasonal autoregressive integrated moving average (SARIMA) model jointly. The OPL hybrid model takes advantage of the unique strengths of linear correlation in social media features and SARIMA model in time series prediction. Experiments on events nearby a subway station show that OPL reports the best forecasting performance compared with other state-of-the-art techniques. In addition, an ensemble model is developed to leverage the weighted results from OPL and support vector machine regression together. As a result, the prediction accuracy and robustness further increases.

1.1 Introduction

Passenger flow prediction is critical for planning, management and operations of public transit systems(Chen and Wei, 2011). The output from the prediction can benefit transit network design, route scheduling, and station crowd regulation operations (Hasan et al., 2013). The majority of the previous studies lie in forecasting day-to-day recurrent passenger flow (Leng et al., 2013; Sun et al., 2015; Sun et al., 2014; Wei and Chen, 2012). However, when it comes to non-recurrent events (e.g. sporting game, concert, running race, etc.), because of its irregularity and inconsistency, passenger flow prediction turns into a very challenging task. Very limited methods have been proposed in the literature.

For solving this problem, instead of revising existing methods, we intend to leverage a new kind of data -- social media. User-generated contents on social media strengthen linkage and interactions between users, meanwhile provide a large amount of information. The vast information is able to capture the public attention, which is one of the common traits of events.

However, social media data is much difficult to process compared with traditional relational data. There still exist several major challenges in handling social media data, which is unstructured, noisy, gigantic, and contains a variety of information. Take Twitter data for example. Only in 2014, we have collected over 29.7 million geo-tagged posts bounded in the New York City Area. At individual post level, a fundamental question of data mining arises: what it is talking about, and what event information it contains. Thus the first challenge (C1), within a transportation context, is how to identify transportation-related events that each post refers to. An individual geo-tagged post is able to provide social activity analysis at spatial-temporal aggregated level. Transportation authorities can leverage such information to identify hot spots and further indicate passenger flows in near future for public gathering. Therefore, the second challenge (C2) is how to develop a method to coordinate social media for forecasting passenger flow, especially under event occurrences.

This chapter aims to address challenges (C1) and (C2). More specifically, under event occurrences, we intend to extract event information from geo-tagged social media data, and leverage both historical transit data and real-time social media data to forecast future passenger flow at subway stations. The following questions will be investigated: (i) Can social media be used to identify public events in real life? (ii) How to build the prediction model by the features extracted from social media? To the best of our knowledge, there has not been considerable published research on the effects of passenger flow prediction with social media.

The section has the following structure. Section 1.2 summarizes related works about recent popular transportation prediction techniques and the uses of social media in transport applications. An overview of the data, including subway passenger flow and social media, is given in Section 1.3. Section 1.4 describes the setup of event detection approach. Section 1.5 presents a detailed analysis of the relationship between event passenger flow and social media. Section 1.6 presents the technical details of prediction modeling and experiments on real-world datasets. Finally, Section 1.7 provides concluding remarks.

1.2 Related works

There is a vast literature in short-term transportation forecasting (Vlahogianni et al., 2004). Generally, there are two groups of approaches receiving wide attention, namely, parametric and non-parametric techniques.

The common parametric techniques include autoregressive integrated moving average model (ARIMA), exponential smoothing (Williams et al., 1998), and historical average (Hobeika and Kim, 1994). Especially, ARIMA has been fully developed for various transportation prediction purposes, including traffic occupancy (Ahmed and Cook, 1979), travel time (Zhang and Rice, 2003) and traffic flow (Williams, 2001). Previous research (Lee and Fambro, 1999; Williams and Hoel, 2003) shows ARIMA performs well for stationary and non-event time series. With the rise of data mining and science, non-parametric techniques also have been widely adopted recently. Neural network (Tsai et al., 2009; Yasdi, 1999), support vector machine for regression (SVR) (Wu et al., 2004) and k-nearest neighbor (Guo et al., 2013) were used to build the traffic volume prediction model for the time-series data.

The passenger flow prediction belongs to the subcategory of short-term transportation prediction. Some researchers adopted both kinds of prediction techniques to forecast the passenger flow for railway (Gong, 2010; Tsai et al., 2009) (Jiang et al., 2014), bus stop (Gong et al., 2014; Jiang et al., 2014; ZHANG et al., 2011), and subway stations. Specifically for passenger flow prediction at subway stations, there are different prediction levels, respectively, at whole transit lines (Leng et al., 2013; Wei and Chen, 2012), at one station with passenger transfer flow (Sun et al., 2014), and at one station with entrance and transfer flow (Sun et al., 2015). All of them obtained a desirable predict result of typical commuting volumes. However, none of them adds consideration of atypical conditions.

Recently, more and more attempts have been made to implement The Internet and social media analysis in the domain of transportation. A huge group of people in the online community generates a tremendous amount of content. Chaniotaks and Antoniou (Chaniotakis and Antoniou, 2015) proposed a generic methodological framework for collecting and analyzing the data from social media. And other researchers took advantages of using crowdsourcing these resources to capture the incoming non-recurrent events

(Pereira et al., 2015a) and explained the causes of transport overcrowding (Pereira et al., 2015b). Studies are trying to exploit this area mainly fall into two applications, traffic detection, and traffic prediction, with supervised learning techniques.

In the application of traffic detection, Wanichayapong et al. (2011) used synthetic analysis to classify the traffic incident information into spatial categories from the social media data. Schulz et al. (2013) extracted features from part-of-speech tagging and words in Twitter posts and developed classifiers to detect car accident occurrences. They applied spatial and temporal filtering to locate the accidents. Daly et al. (2013) built a system called Dublin's Semantic Traffic Annotator and Reasoner to use natural language processing techniques to analyze social media contents in order to capture real-time traffic conditions. Mai and Hranac (2013) explored the time and location of the related Twitter posts after traffic incidents occurred. They found that the majority of tweets are posted within 5 hours and 25 miles for freeway incidents. Gal-Tzur et al. (2014) used the Twitter messages sent from transportation authorities to develop classifiers to identify the posts related to transportation information. Moreover, they presented a keyword-based hierarchical schema to categorize these posts. Chen et al. (2014) tried to detect traffic congestion and location solely based on social media data by using topic modeling and hinge-loss Markov random fields. D'Andrea et al. (2015) utilized Twitter data and developed a support vector machine model to recognize useful keywords from tweets and detect traffic events in the area of highway road network. Kumar et al. (2014) incorporated social media to detect road hazards by sentiment and language analysis. Most recently, Zhang et al. (2016) studied and revealed the characteristics of traffic flow surge near the tweet concentration, which is defined as a cluster of keywords for traffic related events.

For traffic prediction, He et al. (2013) proposed a long-term traffic prediction models with social media features for a freeway network in San Francisco Bay area. They found that there exists a negative correlation between social activity on the web and traffic activity on the roads. Ni et al. (2014) tried to forecast freeway traffic flows under special event conditions by taking into account information derived from social media. Lin et al. (2015) applied linear regression models for predicting the impact of inclement weather on freeway speed with the help of social media.

For subway and transit, Collins et al. (2013) used sentiment analysis of transit riders' short messages on social media to measure their satisfaction about transit. They found that the social media posts with the sharp increased negative sentiment indicated some transit incidents, like fire and delays.

Above studies show that there is great potential to use social media to locate right information for transportation applications. However, none of the previous studies explores the effectiveness of using social media for passenger flow prediction in public metro transit systems.

1.3 Dataset

This study expands the successful applications of social media data to predict passenger volume at a subway station. We focus our study on subway station "Mets – Willets Point" on Line 7 in New York City. The station is selected based on two main reasons. First, "Mets – Willets Point" is adjacent to not one but two stadiums, Citi Field and USTA Billie Jean King National Tennis Center (NTC). Citi Field is the home stadium of New York Mets Baseball team, and NTC hosts the annual US Open grand-slam tennis

tournament. Second, the sports events always obtain public attention. From our observation, there is a substantial volume of social media posts referring to the events.

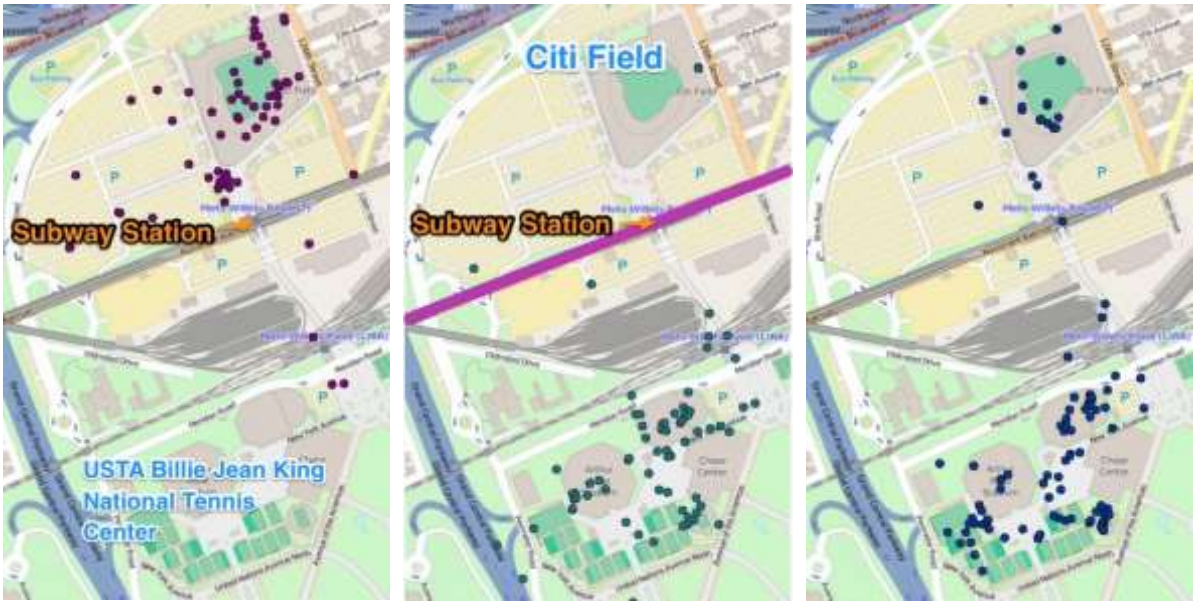
We collected the turnstile usage at "Mets – Willets Point" subway station from Metropolitan Transportation Authority (MTA). In order to cover various types of events, the time range is set from April 2014 to October 2014, in which various events occur nearby.

Table 1.1 Sample tweets before events

Event			**Sample Twitter Message**	
Type	*Start Time*	*Details*	*Create at*	*Text content*
Baseball game	2014-05-14 19:10	Mets vs. Yankee	2014-05-14 18:22:22	Checked in CITI field for the yankees vs mets game w yankees mets
Tennis games	2014-08-25 19:00	US Open 1st round	2014-08-25 17:49:46	I'm at 2014 usopen tennis championships in flushing ny
Baseball game + Tennis games	2014-08-28 19:00 (T) 19:10 (B)	US Open 2nd round & Mets vs. Braves	2014-08-28 18:29:10	love this place billy jean king national tennis centre us open

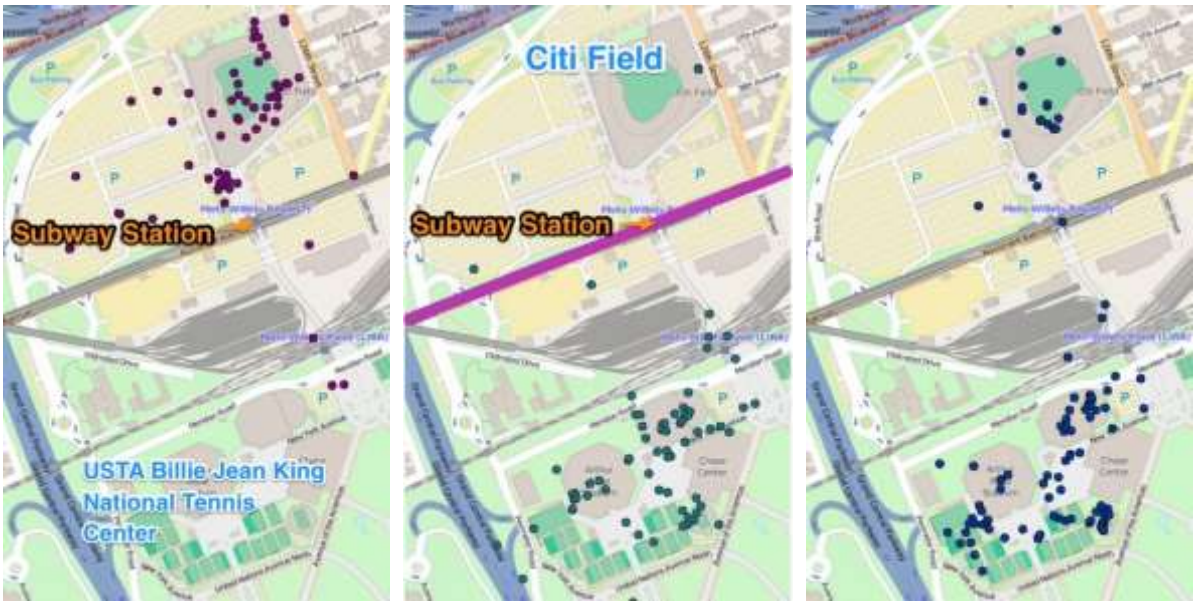
Turnstile devices record passengers passing each turnstile for either entry or exit, and it reports the aggregated number every four hours. In this chapter, we aggregate both entry and exit flows as total passenger flow, which is of transit agency's interest.

We collected Twitter data, known as tweets, as social media data. Twitter message is an online text post limited to 140 characters by Twitter users. Tweets were collected in the same temporal window through Twitter Streaming API with geo-location filter. The spatial bounding box was set to cover only the subway station and two stadiums. Because of the location filter, besides text content, username and timestamp, each tweet contains its geographic coordinate. Inside the post, users are able to prefix by a # symbol with words, which is called the Twitter hashtag. A hashtag provides unique tagging convention to facilitate tweets with certain topics, contexts or events. The aforementioned information from Twitter messages defines a tweet in this chapter.



(a) Baseball game (b) Tennis games (c) Baseball game + Tennis games

Fig.1. shows the locations of tweets sent two hours before different types of events start. As it can be seen, tweets were mostly sent from the stadium in which the event was held. Moreover, different events correspond to different social media activities, and to various levels of public attention. From social media data perspective, the characteristics of tweets, like time stamps, geolocations, text content, quantity ratios, etc., lead to such differences. Our objective is to find ways to measure these differences in social media data and leverage them into prediction models to forecast subway passenger flow.



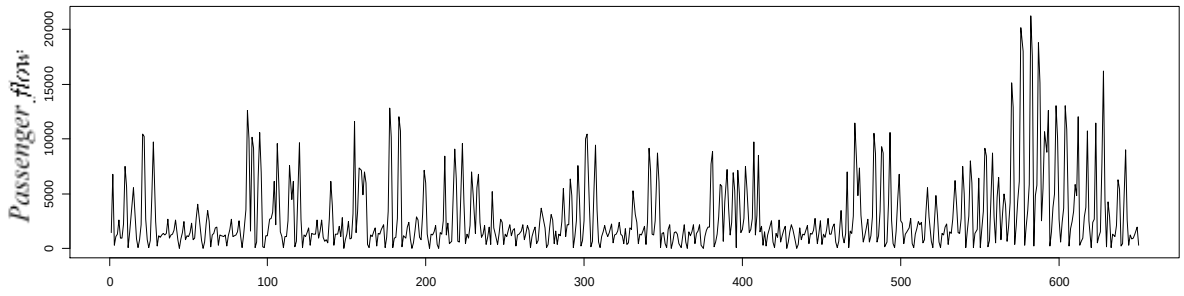
(a) Baseball game (b) Tennis games (c) Baseball game + Tennis games

Fig.1.2 Geographic distribution of tweets two hours before the events

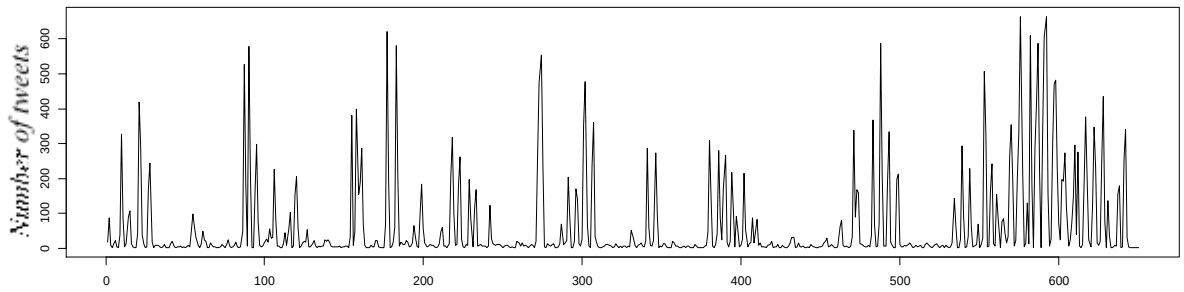
1.4 Hashtag-based Event Identification

The events held in stadiums were well attended. The attendance not only brings a high volume of passenger flows but also activities on Twitter, shown in Fig.1.3 As one can see, event scenarios generate large spikes of social media activity and passenger flow at the same time.

We assume that the complete schedule of all events is unknown for transit operators. The subway station Mets-Willets Point could coordinate transit passengers for two major sports events, US Open Tennis Championships and Major League Baseball for New York Mets. The former was held late August and early September over a two-week period, and the latter was held from April to September 2014. However, after



(a)



Time Periods from April 2014 to October 2014

(b)

Fig.1.3 Comparisons of passenger flow (a) and number of tweets (b)

initial examinations, we found that there were other events like concerts and speeches being held nearby as well. Therefore, we need to identify the events by social media data.

Instead of detecting the exact topic of the events (Cordeiro, 2012b; Ramage et al., 2010; Weerkamp and de Rijke, 2012), we would like to examine tweets within the area and probe whether there will exist events involving high social activities. To correctly identify the events, rather than using the complex machinery of latent variable topic models (e.g. Latent Dirichlet Allocation (Blei et al., 2003)), we employ the Twitter hashtags to measure social media activities and provide the context for them (Giridhar et al., 2014).

Hashtag extraction is the first step of the proposed event detection algorithm. We denote t as one of the time intervals, with $t = 1, \dots, T$, where T is the total number of four-hour intervals. HL_t is the list of hashtags during t. $HL_t = \{H_{t1}, \dots, H_{tj}, \dots, H_{tJ_t}\}$, where H_{tj} is the j^{th} hashtag and J_t is the total number of hashtags labeled by Twitter users during t.

Furthermore, let $\boldsymbol{M}^H \in \mathbb{R}^{T \times S}$ denote the hashtag matrix, where S is the number of hashtags. Its element $M^H_{t,s}$ corresponds to the occurrence of the s^{th} hashtag in the t^{th} time interval. In the hashtag matrix, all the hashtags over time intervals merge into the columns. Various words and phrases depict different aspects of social activities. In sum, the column names of hashtag matrix are the hashtags, the rows stand for the time intervals, and each entry in the matrix corresponds to the frequency of the hashtag.

Below are the steps of event detection by hashtags.

Algorithm 1: Hashtag-based Event Identification

Input: Tweets within the area

Output: Hashtag matrix $\boldsymbol{M}^H \in \mathbb{R}^{T \times S}$

1. Hashtags extraction

$HL_t = \{H_{t1}, \dots, H_{tj}, \dots, H_{tJ_t}\} \quad \forall t \in [1, T]$

2. Lexical analysis

$HL \equiv \bigcup_{t=1}^{T} HL_t$

Remove stop words, punctuation and duplicated strings from HL

3. Label all collected tweets by hashtag

$TW_{p,s} \equiv$ calculate the occurrence of s^{th} word in *HL* of p^{th} tweet

for p^{th} tweet $p = 1 \; to \; P$ do

for s^{th} word in *HL*

Append the $TW_{p,s}$ as a new column for p^{th} tweet

4. Build hashtag matrix ($\boldsymbol{M}^H \in \mathbb{R}^{T \times S}$)

Each row of $\boldsymbol{M}^H$ represents the vector of *HL*

$OC_t \in \mathbb{R}^S \equiv$ the occurrence of each element in *HL* for time interval t

for $t = 1 \; to \; T$ do

$OC_t = \sum_{p \in T} \sum_{s \in S} TW_{p,s}$

$\boldsymbol{M}^H_t = OC_t$

5. Peak detection

for $t = 1 \; to \; T$ do

Rank OC_t based on $\sum_S \; OC_{t,s}$ from the largest to the smallest.

for $s = 1 \; to \; S$ do

Sort $OC_{t,s}$ from largest to smallest.

Since there could be different hashtags for different time intervals, it is trivial to see that $\boldsymbol{M}^H$ is originally a sparse column-wise matrix, and each column corresponds to the frequency of hashtag in each time interval. By concatenating hashtag list HL_t over *t*, it converts $\boldsymbol{M}^H$ to a full storage matrix in order to sort the hashtag matrix row by row for peak detection afterward.

Moreover, instead of directly utilizing the occurrence of hashtags labeled by Twitter users, we extract the string vector of hashtags and use it to label the text content of each tweet. It will facilitate the approach to capture those tweets about a similar topic without hashtags.

Finally, we implement peak detection to extract most frequently occurring hashtags as event hashtags, representing social media activities with context. In Table 1.2, the top 3 frequently occurring hashtags are presented. Moreover, we use the sum of all occurring hashtags for each time interval to measure the social media activity. High-rank number of hashtags indicates that the corresponding time interval is under event occurrence.

Table 1.2 shows that the there are various detected events, including US Open, baseball games, music shows, running races, etc. In order to justify the method, we compare the detection results with the true home game schedule of New York Mets, which had long time range and a decent number of games. There were 81 game days during April 2014 to October 2014 for New York Mets. After eliminating the days with missing Twitter data, 65 game days remain. Since the objective of the event detection is to sense the positive events instead of non-events, we evaluate the identification results with *precision*, *recall* and F_1 score.

The proposed method achieves good performance in identifying those baseball events, i.e., the *precision* is 98.27%, *recall* 87.69% and F_1 score 0.9268.

Table 1.2 Sample events and their hastags

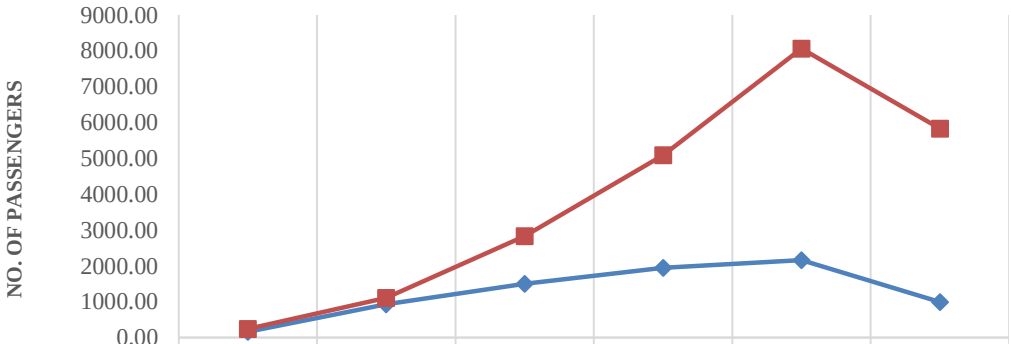
Date	**Hour**	**No. of EH**	**Top Hashtags**		
3/31	17 to 21	65	mets	openingday	ny
4/5	13 to 17	306	mets	reds	baseball
4/9	17 to 21	34	amaluna	cirquedusoleil	citifield
5/14	17 to 21	710	mets	yankees	subwayseries
5/31	9 to 13	85	happiest5k	queens	ny
6/7	17 to 21	75	digifestnyc	nyc	selfie
8/25	17 to 21	437	usopen	tennis	usopen2014
8/31	13 to 17	609	usopen	mets	tennis

Note that there are two reasons to use event hashtags instead of the quantity of tweets directly. First, there is a chance that high volume of tweets does not necessarily indicate event and attendance. In our observation, a conversation between users, commercial promotions or information dissemination could also

generate a high quantity of tweets. The proposed hashtag-based method is able to diminish the effects of these unrelated tweets. Second, the top event hashtags can describe what the event is about, though the hashtags might not be formal English words. It can be seen in ***Error! Reference source not found.***Table 1.2, different kind of events and baseball teams can be easily recognized by the top event hashtags.

1.5 Events Characteristics

Different events in stadiums bring different size of audience to the sites, in which the passenger flow at the subway station varies accordingly.



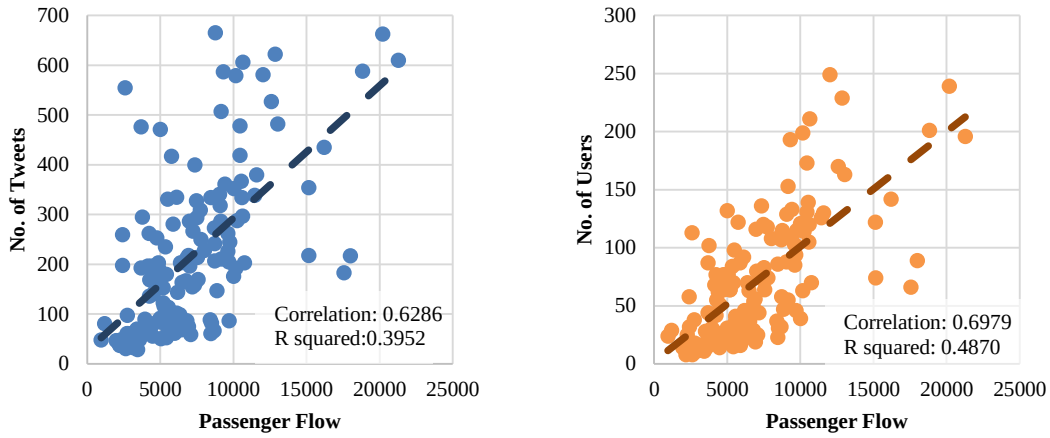
	5:00	9:00	13:00	17:00	21:00	24:00 + 1 hour
Non Event Volume	161.88	929.07	1500.75	1949.77	2166.36	997.72
Event Volume	241.76	1113.89	2828.61	5087.21	8066.04	5829.68
Ratio of Non-Event over Event Flow	66.96%	83.41%	53.06%	38.33%	26.86%	17.11%

Fig.1.4 Average event/nonevent daily passenger flow at Mets-Willies Point station

As shown in Fig., there are huge differences between event and ordinary transit traffic in quantity, more importantly in variation. This difference inevitably leads to the difficulty of transit prediction by traditional time series models (e.g. ARIMA).

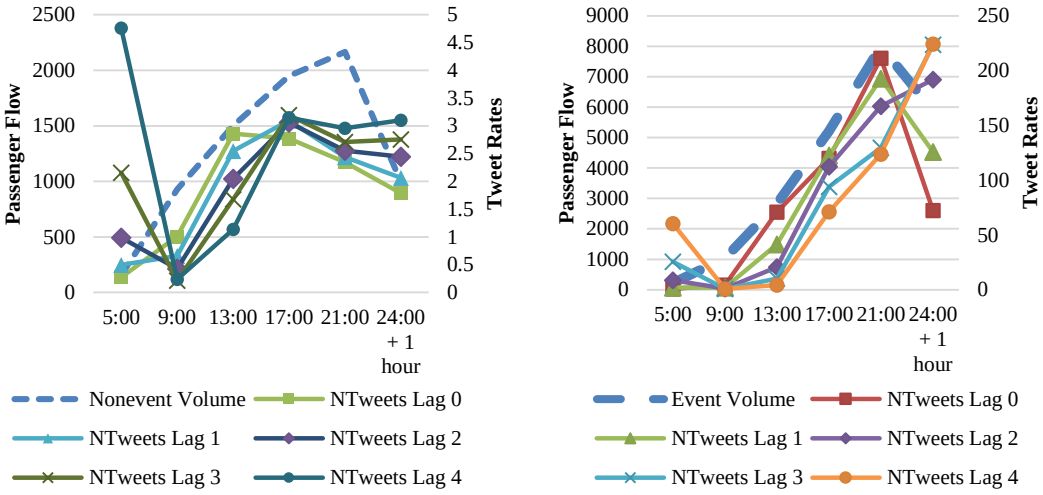
On the other hand, in Fig we plotted the number of tweets against passenger flow under event occurrences in (a), and the number of Twitter users against passenger flow in (b). As one can see, a linear trend is observed between tweet counts on passenger flow. The correlation coefficient is above 0.62 and adjusted R^2 value is above 0.39. The R^2 values indicate that the number of users is a more robust predictor. We reasonably believe that there exists a moderate positive correlation between tweet counts and event passenger flows. This result gives us the confidence to explore further the prediction modeling of social media on the event passenger flow.

Note that our study is restrained to the extent that the geo-tagged tweet is available. For some of the time periods, the amount of tweets is very small despite the time of day. In this case, event identification measures social media activities and automatically excludes these time periods from the correlation study and the following analysis.



(a) Number of Tweets V.S. Passenger flow (b) Number of Users V.S. Passenger flow

Fig.1.5 The correlation between tweet rates and passenger flow under events



(a) Nonevent (b) Event

Fig.1.6 Average Passenger flow V.S. Average Tweet Rates at Citi Field Station

1.6 Prediction modeling

In this section, we intend to investigate whether or not the content of social media will assist in forecasting event passenger flow. The first step is to identify the best time lags for the prediction models.

To measure the tweets quantifiable, we define two types of feature as tweets rates from social media data:

- *NTweets(t)*: Number of event-related tweets at time step *t*.
- *NUsers(t)*: Number of unique tweet users at time step *t*.

Because the record time interval of transit passenger flow is four hours, we also aggregate the tweets data

in four-hour intervals. If the predicted passenger flow is at time *t*, we shift tweet rates to earlier hours: *t-1,t-2,…t-L,* since prediction requires features ahead of passenger flow time. Based on the positive correlation of tweet rates and passenger flow in Fig, we construct a linear regression (LR) model, where passenger flow is the dependent variable, and tweet rates over different hours are independent variables.

The highest predictive correlation is achieved when the tweet rates are calculated based on one hour prior to event time range. We obtain an adjusted R^2 value of 0.616 in lag one-hour case. For comparison, the R^2 values in lag zero and two-hour cases are, respectively, 0.488 and 0.512. Also, shown in Fig.1.6, one can see that the curve of tweets rates with one hour lag fits best to the curve of event passenger flow, whereas for non-event passenger flow there are no obvious patterns between tweets rate and passenger flow. Based on such analysis, we will include tweet rates with one-hour lag into the base prediction model in the following analysis.

Next, we implement cross validation to compare the results of LR model and two popular prediction models: average prediction (AVG) and seasonal autoregressive integrated moving average (SARIMA). We generate an experiment with 100 runs of datasets from the event detection result, and each run takes inputs by randomly splitting the entire dataset into training (70%) and test (30%) sets.

The prediction performance is evaluated by two metrics, namely Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE).

In our experiment with 100 runs, the LR model with tweet rates improves the MAPE by 33.08% comparing with SARIMA (See Fig. for details). Notice that such good performance is achieved by the LR with two variables only. However, the LR model does not capture the relation between time steps, since the passenger flow data are time series in nature.

We conduct a comparison of R^2 values between two models: 1) the Tweets-based LR model and 2) the historical-flow-based SARMIA model. The experiment obtains adjusted R^2 value of 0.616 for the LR, 0.400 for the SARMIA, and 0.696 for combined features of both. As one can see, around 60% of the event passenger flow variance can be explained by the number of tweets variation. And around 40% of the variance comes from historical time-series flow data, which includes a large portion of day-to-day recurrent passenger flow and a small portion of the non-recurrent event flow. The combination of these two methods shows better R2 value since the LR provides event-related features while the SARIMA present the features related to time series and routine flow.

Inspired by the above experiment with two modeling methods, we propose a convex optimization based approach, called Optimization and Prediction with hybrid Loss function (OPL), to fuse the LR model and the SARIMA model in the objective function jointly. The OPL model aims to take advantage of unique strengths of line regression in social media features and SARIMA model in time series prediction.

The hypothesis of the proposed model is a parametric linear model, defined as:

$$h_w(x) = 1 + w_1x_1 + w_2x_2 + \cdots + w_nx_n \qquad x_0 = 1$$

Where x_i is i^{th} feature and its corresponding coefficient is w_i. In total, there is the experiment with m=100 runs. Each entry of the experiment is one of the four-hour intervals from the event detection result. Following our experiment design, we randomly split the m runs into training m_{train} (70%) and test m_{test}

(30%). The two tweet rates, *NTweets* and *NUsers*, with one-hour lag act as features in the model.

We construct the total loss function as:

$$J(w,\hat{y}) = \sum_{j}^{m_{train}} (y^{(j)} - h_w(x^{(j)}))^2 + \alpha \cdot \sum_{j}^{m_{test}} \left(\hat{y}^{(j)} - h_w(x^{(j)})\right)^2 + \beta \sum_{j}^{m_{test}} \left(\hat{y}^{(j)} - y^{*(j)}\right)^2 \quad 1.1$$

The idea behind the loss function is to combine the modeling of the predictions on both training and test data as well as the predictions from time series model. Equation 1.1 contains three main parts. The first component is the sum of least square for the training set, which is the same as linear regression. The second component incorporates the prediction part directly into the loss function in order to minimize the square error from test data. In addition, to fuse the results of SARIMA, we manage to add the sum of least square between OPL predicted $\hat{y}^{(j)}$ and SARIMA predicted $y^{*(j)}$ into Equation 1.1 as the third component. $y^{*(j)}$ plays the role of regularization to leverage the whole loss function. Since OPL only includes two independent variables, in the trail experiments, it shows that it is not necessary to equip L1 regularization to prevent overfitting. In sum, OPL adopts the moderately large correlated social media features, and incorporates the prediction results from conventional time series model.

To minimize Equation (1), we first vectorize all variables and coefficients:

$$W \in \mathbb{R}^n \qquad Y \in \mathbb{R}^{m_{train}}$$
$$X^{train} \in \mathbb{R}^{m_{train} \times n} \qquad \hat{Y} \in \mathbb{R}^{m_{test}}$$
$$X^{test} \in \mathbb{R}^{m_{test} \times n} \qquad Y^* \in \mathbb{R}^{m_{test}}$$

Then, the loss function is transformed into:

$$J(W,\hat{Y}) = tr(Y - X^{train} \times W^T) \times \left(Y - X^{train} \times W^T\right)^T) + \alpha \cdot tr(\hat{Y} - X^{test} \times W^T) \times \left(\hat{Y} - X^{test} \times W^T\right)^T) + \beta \cdot tr((\hat{Y} - Y^*) \times \left(\hat{Y} - Y^*\right)^T)$$

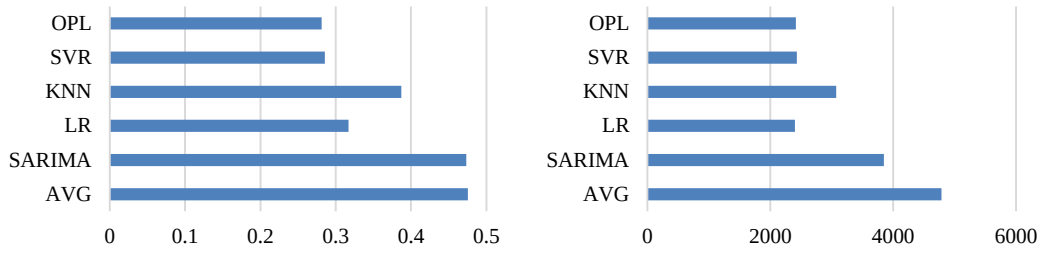
Take partial derivative of the above equation with respect to W and $\hat{Y}$, respectively and we get:

$$\nabla_W J(W,\hat{Y}) = [\left(X^{train}\right)^T \times X^{train} + \alpha \cdot (X^{test})^T \times X^{test}] \times W^T - \alpha \cdot (X^{test})^T \times \hat{Y}^T - (X^{train})^T \times Y^T = 0 \quad 1.2$$

$$\nabla_{\hat{Y}} J(W,\hat{Y}) = \alpha \cdot X^{test} \times W^T - (\alpha + \beta) \cdot \hat{Y}^T + \beta \cdot Y^{*T} = 0 \quad 1.3$$

Then, we use the gradient descent method to solve Equations 1.2 and 1.3 to find a local minimum of $\hat{Y}$. Given Equation 1.1, gradient descent starts with an initial set of $(W,\hat{Y})$ and iteratively moves toward a set of values that minimize the function. Each iteration takes a step in the negative direction of the function gradient. Because the Equation 1.1 is convex, the result of OPL shall be the global optimal values.

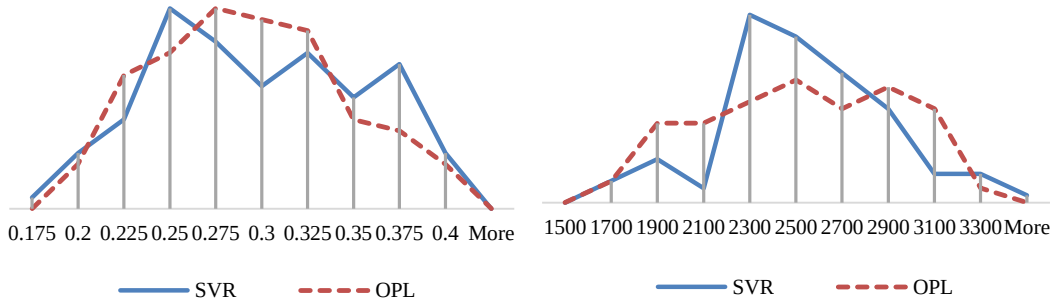
In order to benchmark our proposed method against existing popular prediction approaches, we introduce two nonparametric methods, including SVR and k-nearest neighbors (KNN). The prediction process utilizes cross-validation as well.



(a) MAPE (b) RMSE

Fig.1.7 Performance metrics of the prediction models

Fig.1.7 illustrates that the OPL yields better prediction accuracy than other methods. Compared with the LR, the OPL improves MAPE by 11.4%. Also, one can see that the SVR presents desirable prediction performance as well. The SVR and the OPL have different characteristics. The SVR is a nonparametric technique that considers tweet rates only. The OPL is a parametric method and incorporates the prediction results from conventional time series model. Further, a detailed comparison is conducted by another 100 randomly generated runs.



(a) MAPE (b) RMSE

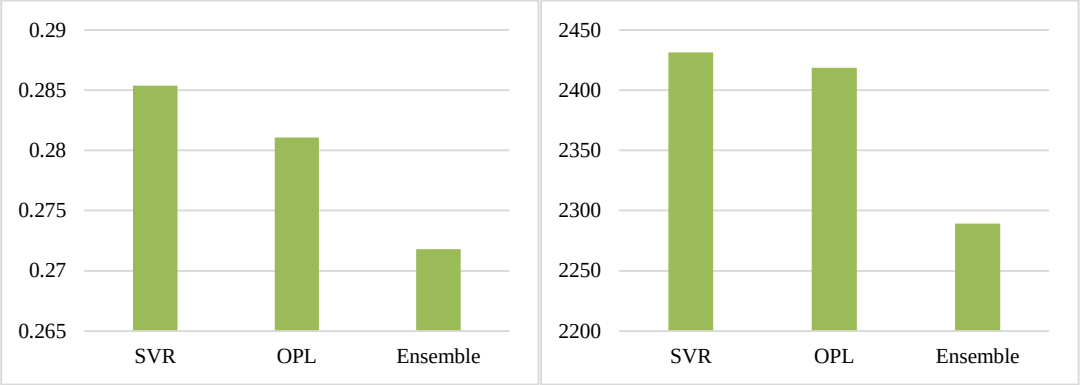
Fig.1.8 The distributions of test errors to compare the SVR and OPL

Fig.1.8 depicts the distributions of test errors for both SVR and OPL. While either method performs relatively well on its own, it shows the distributions are heterogeneous for both metrics, MAPE and RMSE. The heterogeneity of error distributions encourages us to combine the merits from both techniques. Inspired by the aggregation approach proposed by (Tan et al., 2009), we implement stacking -- an ensemble learning approach to merge the prediction results of the SVR and OPL.

$$\hat{Y} = P\left(X^{train}\middle|OPL\right)\cdot \underset{\hat{Y}}{\operatorname{argmin}} J\left(W,\hat{Y}|OPL\right) + P\left(X^{train}\middle|SVR\right) \cdot \underset{\hat{Y}}{\operatorname{argmin}} J\left(W,\hat{Y}|SVR\right) \quad 1.4$$

We estimate $\hat{Y}$ by Equation 1.4. The weighted probabilities come from normalized root mean square error

of training data. The output averages the argument of the minimum for both SVR and OPL.



(a) MAPE (b) RMSE

Fig.1.9 Improvement from ensemble learning from the OPL and SVR

As one can see from Fig.1.9, the ensemble approach yields better prediction accuracy than either OPL or SVR. It is worth mentioning that the improvement over the conventional SARIMA is more than 40%. Notice that tweet features are obtained from no-cost and real-time social media data. The results indicate the promising value of using social media for passenger flow prediction under event conditions.

1.7 Conclusions

In this chapter, we have addressed two important questions, in brief, whether social media data is able to signify public gathering events, and what techniques can be used to model the passenger flow prediction by the features extracted from social media.

First, we exploit social media to detect various events with hashtags. In order to capture events precisely, the hashtags from the Twitter users have been analyzed, tuned, adapted and applied with lexical processing techniques and peak detection. Our approach achieves good performance with precision 98.27% and recall 87.69% for the baseball games. It is a simple but efficient method to capture the events related to public gathering with high social media activity.

Second, we propose a convex optimization model called Optimization and Prediction with hybrid Loss function (OPL) to fuse the least squares of linear regression and the prediction results of SARIMA in the same objective function. The OPL hybrid model aims to take advantage of the unique strengths of line regression in social media features and SARIMA model in time series prediction. Among several popular prediction methods, OPL shows the best results in terms of MAPE and RMSE. In addition, by comparing the distribution of prediction errors of OPL with SVR, which is a popular nonparametric and nonlinear method, it is found that their performance shows heterogeneous error patterns. Therefore, an ensemble model is developed to leverage the weighted results from OPL and SVR jointly. As a result, the prediction accuracy and robustness further increases.

Overall, social media data show the capability in passenger flow prediction under event conditions. Social media offers a cost-effective way to obtain real-time traveler related data, and fills the gap between day-to-

day passenger flow volume and abruptly changing non-recurrent event volume. The positive correlation between passenger flow and social media activity plays a significant role as transit demand indicator in the public transit system.

In future, one could further explore the minimum percentage of social media use in an event that leads to a respectable accuracy, and how such minimum can be estimated in order to compute a trust index for the regression result.

2 An Exploratory Study on the Correlation between Twitter Concentration and Traffic Surge

Social media receives increasing attentions as a crowdsourced information source in traffic operations and management. The tweets, which are blogged and shared by the broad masses of people, may be associated with some major social activities. These tweets are called "Twitter concentrations" in this chapter. The public activities behind Twitter concentrations potentially pose more pressure on traffic network and cause traffic surge within a specified time and location. However, it still remains unknown how closely the Twitter concentration and traffic surge are correlated with each other. Our study fuses a set of tweets and traffic data collected during the whole year of 2014 in North Virginia Region, and mainly investigates the correlation between Twitter concentration and traffic surge in July. The results show the promise and effectiveness of our proposed methods and even provide insights in the causality of the non-recurrent traffic surge.

2.1 Introduction

Road traffic surge aggravates the jammed condition and worsens the level of service of road links. The consequences of traffic surge may vary, including traffic delay, fuel wasting, drivers' frustration, etc. Some of the traffic surges may be accounted by the recurrent features of traffic patterns such as time-of-day characteristics and weekday-weekend differences. This kind of traffic increase is predictable in most cases, and people living by usually get accustomed to it. Other kinds of traffic surge, which are more unpredictable and hazardous, may correlate with non-recurrent traffic patterns such as road accident, bad weather, malfunction of traffic signals, festival parades, etc. Fig.2.1 illustrates the differences between recurrent and non-recurrent traffic surge. The non-recurrent traffic patterns and corresponding traffic surge problems, which are caused by major social activities, are the main interests of this chapter.

For decades, the traffic surge problem, which potentially causes and even worsens traffic congestion, has been given much attention. The state-of-art studies attempt to unveil the correlation between traffic increase and other variables, and several explanations have been put forward, which are described as follows. Bando et al. (1995) found that there exists the congestion that is induced by a small perturbation without any specific origin such as a traffic accident or a traffic signal. Arnott et al. (2006) argued that cars cruising for parking add to traffic congestion. Duranton et al. (2011) showed no evidence that the provision of public transportation affects vehicle kilometers traveled (VKT). Further, they proved that increased provision of roads or public transit is unlikely to relieve congestion. Anderson (2013) concludes that the cessation of transit service may increase average highway delay by up to 47%. Other studies even showed that higher congestion through restraining capacity for additional travel appears to be associated with the decrease in regional employment growth rates (Sweet, 2014). State-of-art studies investigate under which conditions and activities the traffic operations are influenced, and the traffic congestion deteriorates. However, the same as many other traffic problems that are closely related to human activities, the answers to correlation studies of the traffic congestion problems may be quite diverse.

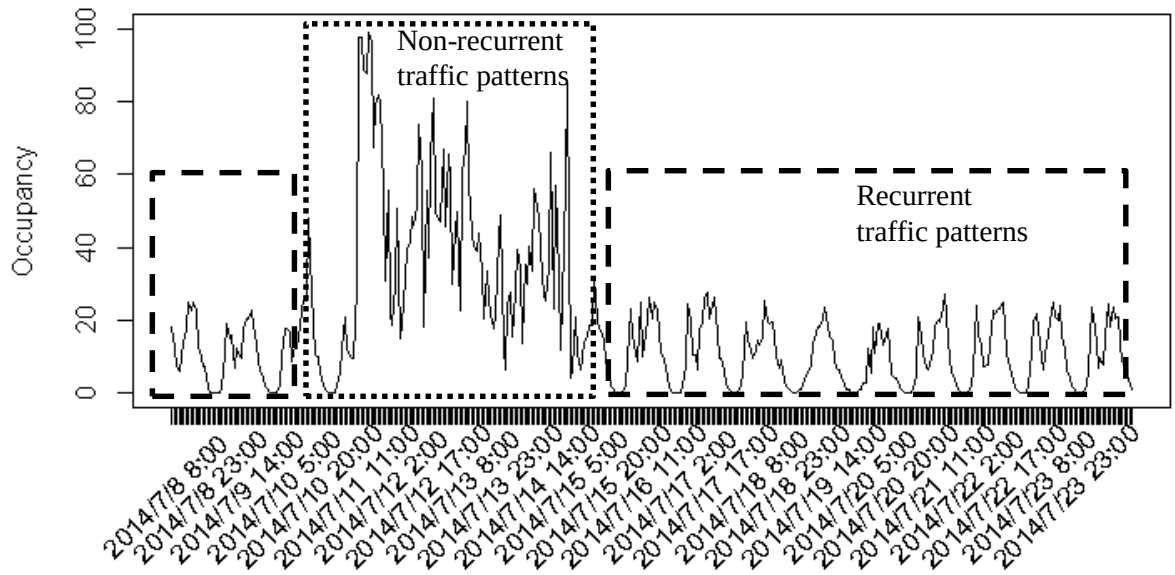


Fig.2.1 Recurrent traffic fluctuations of hourly traffic occupancy in one sample detector

Created in March 2006, Twitter founds a perfect stage of "We Media" and this makes possible the wide-range information retrieval about public activities from the massive majority of people. Over the past decades, social media has been validated useful to broadcast major events such as natural disasters (2013; Sakaki et al., 2010), bird flu (Aramaki et al., 2011), politic events (Shirky, 2011), etc. The problems are whether it can detect the traffic surge and whether there exists any possible correlation between the tweets and traffic surge. Usually, the Twitter contents that are related to specified events will occur disproportionally frequently over certain time and space, and it is possible to make a direct connection between tweets and events like ceremony opening (Balduini and Della Valle, 2012), celebrity death or festival parades (Schwarz, 2012). However, the non-recurrent traffic surge may be quite different from that of other events. This is because a variety of public activities possibly incurs more on-road traffic and cause non-recurrent traffic congestions. For example, on June 9th, 2013, an anomalous 10-mile traffic jam was detected on a major Southern California freeway. At the same time, the keywords "Obama" and "Impeach" occur much more frequently in the current tweets than the former ones (Giridhar et al., 2014). The contents that most people discuss and post on Twitter may imply a major social trend and bring more on-road traffic over specific time and space. In our study, we define Twitter concentrations as the tweets that involve a variety of traffic-related activities whose contents are widely created, consumed, distributed or shared. The goal of this chapter is to explore the correlation between Twitter concentrations and traffic surge.

There are mainly three challenges to be addressed: The first challenge lies in how to quantify the traffic surge that may result from major public events. The time-of-day traffic data collected by loop detectors inheritably contain detection errors, and the traffic occupancy may fluctuate over time-of-day. Thus, it may not be easy to reasonably interpret the traffic data from the observations of one or two days and also not easy to determine the traffic surge according to data records previously collected. Also, traditional studies focused mainly on several intersections (Teodorovic et al., 2001) or corridors (Lan et al., 2008; Schoenhof

and Helbing, 2007). In comparison, archiving, interpreting and summarizing the high-resolution traffic occupancy data in a large road network is quite challenging, considering the huge data size collected by large-scale fixed detectors. In our study, a clustering method is employed, and a detector-based probabilistic model to detect traffic surge is proposed.

The second challenge lies in the inheritable complexity and unstructured nature of tweet data: language ambiguity (Chen et al., 2014) and how to exact the traffic-related Twitter concentrations from the large collections of tweets is worth studying. We extract the Twitter concentrations by one or more keywords that make the tweets discriminatively different from that of others. The extracted tweets are further classified to label whether they are traffic-related or not. The prevailing methods of classifying tweets can be categorized into supervised and unsupervised techniques including Naïve Bayes classifier (Sankaranarayanan et al., 2009), online clustering (Phuvipadawat and Murata, 2010), support vector machine (Sakaki et al., 2010), hierarchical divisive clustering (Long et al., 2011), discrete wavelet analysis (Sakaki et al., 2010), continuous wavelet analysis (Cordeiro, 2012a), decision trees (Popescu and Pennacchiotti, 2010) etc. The performance of these methods is partially decided by the applications and data sources. In this chapter, we employed widely-accepted unsupervised and supervised learning techniques to classify the traffic-related tweets.

The third challenge is that public events that are reflected on the Twitter concentrations may exert different levels of influences on its surrounding traffic. Some of the activities may pose influence on more than one detectors, and some may be in effect within more than one hour. To properly interpret the traffic surge by the data collected from different locations and time periods, we aggregate the data by two different statistics measures: mean and 75th percentile. These values can be compared with the traffic-related Twitter concentrations to explore their correlations with traffic surge.

Our main contributions can be summarized as follows. First, a probability index is proposed to quantify the level of detector-based traffic surge in a large-scale road network that is detailed in Sections 2.3 and 2.4. Second, an effective detection method is proposed to extract, filter and classify the traffic-related Twitter concentrations from a total collection of tweet posts, and the method is introduced in Sections 2.5 and 2.6. Third, we develop a methodology to evaluate the correlation between a specified tweet post and its surrounding traffic. The details of the correlation studies are in Section 2.7. The chapter ends with some useful conclusions and thoughtful ideas in Section 2.8.

2.2 Data and Incentives

The study area, shown in Fig.2.2, is located in the vast road network of Northern Virginia (NOVA). The network that is more than 50 square kilometers consists of roads connected by more than 1200 signalized intersections. For each intersection, an average of 12 loop detectors is fixed on the approaches of the intersections. With these traffic detectors, the access to real-time traffic information in our study area is becoming routine as under growing pressure for improving traffic management (Leduc, 2008). The traffic occupancy data, which are usually employed as an index of a traffic jam, are collected at an interval of 15 minutes in July 2014. We only study the data collection in the daytime from 4:00 a.m. to 21:00 p.m.

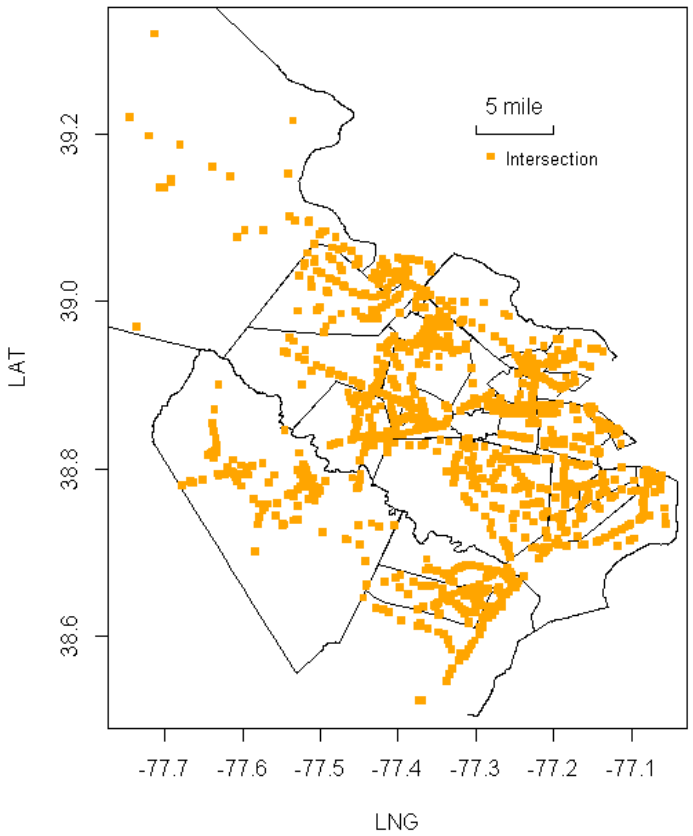


Fig.2.2 Study areas and the locations of intersections

Within NOVA area, we also collected the tweets through Twitter Streaming API with geo-location filter. More than 584,000 tweets were collected throughout 2014. We mainly investigate the tweets in July. To ensure that the tweets are collected from the general public, Twitter users from traffic authorities such as "I95VA", "MKA_NVA" have been removed from our study. Also, some of the tweets from media or press such as "nbcwashington" etc. were also excluded after our empirical examination. Our incentives are intrigued by some preliminary examinations of the tweets. For instance, in Feburary13 2014, there was a keyword surge of "capitalweather". People tweeted the delay caused by the "Biggest snow storm since Snowmageddon". Public service info feed "Metrobus Info" lively broadcasted the congestion at North Capitol & New York Ave in DC area. This was widely retweeted by other users. One can say that the public events reflected by Twitter concentrations potentially exert pressure on the road network and aggravate the traffic congestion.

2.3 Traffic clustering algorithm

Besides the large-scale sensor data, the traffic data inheritably possess time-of-day features such as AM peak and PM peak. Thus, the traffic surge should be justified by a time-of-day clustering method. It is worth mentioning that we do clustering on the data separately on weekdays and weekends because the traffic conditions may be quite different. The following algorithm works almost the same for weekdays and weekends, and we do not intentionally distinguish that.

Backed by this setup, we start from detector level. The traffic occupancy data of detectors are collected every 15 minutes, and we take the median of the traffic data collections in different hour period as the traffic signatures of the detectors. That is:

$\boldsymbol{O^d} = (O_1^d, O_2^d, \cdots, O_j^d, \cdots, O_N^d)$

where O_j^d is the occupancy median in the jth hour period in the detector d. There are in total N hour periods in one traffic signature $\boldsymbol{O^d}$. As our study hour period is from 4:00 a.m. to 21:00 p.m., N is set to be 18. The median value possibly eliminates the fluctuations of traffic data in different days and is less likely to be influenced by outliers than mean. Previous study argues that given the combination of direction, connectivity and locality of a road segment, one can distinctively determine the corresponding traffic signature (time-of-day features of speed) of a road segment with high probability (Banaei-Kashani et al., 2011). Enlightened by this idea, we also assume the following:

Assumption 1: there exist unchanged traffic signatures in a given detector. The time-of-day traffic occupancy over a certain hour period fall into a reasonable range, and those that are obviously higher from the feasible range are traffic surge.

The traffic signatures in more than 15000 traffic detectors constitute the raw database. To find the feasible occupancy range of each hour period, we employed the K-means algorithm with a principled way of finding the number of clusters and the cluster centers. This algorithm can partition the traffic signatures into finite groups of similar patterns and output the centers of clusters as well as the cluster IDs detectors belong to. The algorithm is shown as follows:

Algorithm 1: Traffic signature clustering

Input: The maximum number of clusters $\boldsymbol{\kappa}$ and the traffic signature matrix $\boldsymbol{O^d}$ (in this chapter, this matrix contains 15000 rows and 18 columns. Each row is the traffic signatures of detectors).

Output: Centers of clusters ($\boldsymbol{C^1}, \ldots, \boldsymbol{C^i}, \ldots, \boldsymbol{C^k}$);

The cluster IDs detectors belong to.

Assign the initial number of clusters k=2, initialize AIC= $+\infty$

Repeat

Implement K-means clustering algorithm with k clusters:

Pick randomly the cluster centers ($\boldsymbol{C^1}, \ldots, \boldsymbol{C^i}, \ldots, \boldsymbol{C^k}$);

Repeat

Cluster each traffic signature $\boldsymbol{O^d}$ to the nearest cluster center $\boldsymbol{C^i}$ with $\mathbf{min}(\boldsymbol{d}(\boldsymbol{O^{di}}, \boldsymbol{C^i}))$;

Replace $\boldsymbol{C^i}$ by $\boldsymbol{mean}(\boldsymbol{O^{di}})$;

Until none of the detectors switches clusters

Calculate the ratio of AIC change: diff(AIC)/AIC

Until diff(AIC)/AIC <= $\boldsymbol{\epsilon}$ or k= $\boldsymbol{\kappa}$

In the algorithm, $\boldsymbol{\epsilon}$ is a threshold value set to be 3% in this chapter. The Akaike information criterion (AIC) (Akaike, 1998) is employed to measure the relative quality of the clustering results.

$$AIC = \sum_{i}^{k} \sum_{d \in dom(i)} d(\boldsymbol{O}^{di}, \boldsymbol{C}^{i}) + K \cdot N$$

where $\boldsymbol{O}^{di}$ denotes the traffic signature of the dth detector that belongs to ith cluster. $\boldsymbol{C}^{i}$ is the center of the ith cluster. $d(\boldsymbol{O}^{di}, \boldsymbol{C}^{i})$ is the Euclidean distance between a traffic signature $\boldsymbol{O}^{di}$ and its cluster center $\boldsymbol{C}^{i}$. $dom(i)$ is the domain (collection) of all detector ID whose traffic signature belongs to ith cluster. k is the current number of clusters. N is the count of elements in a traffic signature which equals to 18 in our study. Theoretically, the smaller the AIC is, the better the clustering result should be. For computational efficiency, the algorithm stops when the increase of cluster number brings no more than 3% additional benefits. The AIC results are shown in Fig.2.3 When k=15, diff(AIC)/AIC goes lower than 3%.

There recommend two important criteria in selecting the number of clusters: First, the cardinality of small size clusters may decrease with the increasing of the number of clusters. The cardinality should not be too small because corresponding results from clusters of large cardinality produce more reliable cluster centers in the later study. Second, we may also use BIC or other statistics to measure the quality of clustering. Same as that of AIC, the results of other statistics do not also indicate an unconstraint large number of clusters.

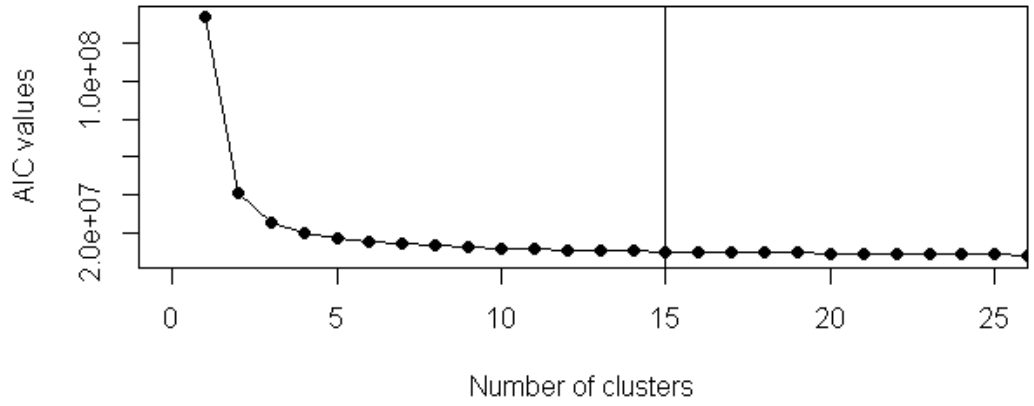


Fig.2.3 AIC values for different k

The clustered centers are shown in Fig.2.4. From the shape of our clustering results, it is not surprising that different cluster centers vary not only in shapes but also in scales. It reveals a clear time-of-day feature for each detector. This method can find the outliers in the traffic occupancy due to several of its advantages:

The method fully considers the time-of-day features of traffic patterns inherited in the traffic data.

The method is totally based on the field data which is in large scale. The aggregation of large-scale data may eliminate the possible noises from the results.

The method clusters the traffic occupancy only in July, which can diminish the effects of traffic operations in different months.

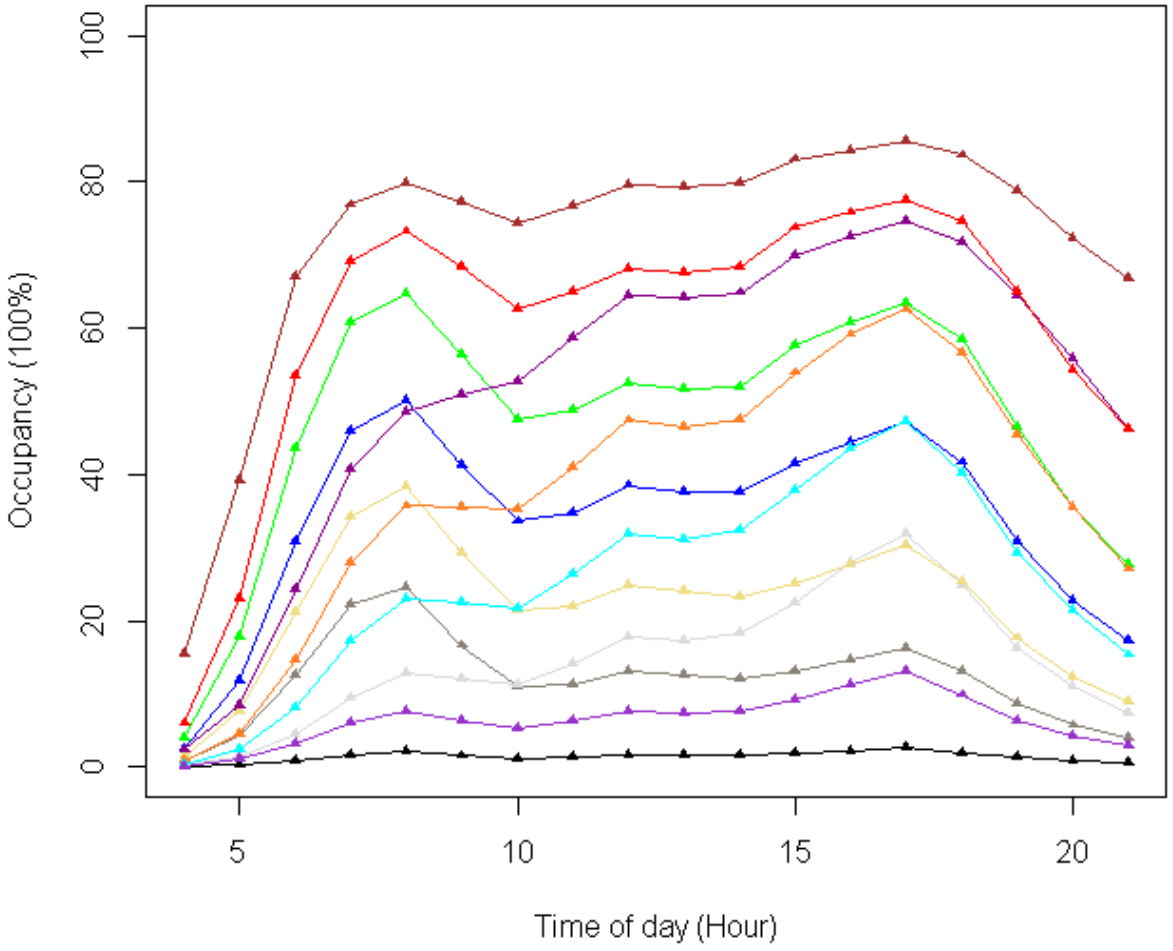


Fig.2.4 12 different clustered centers of traffic signatures

2.4 Traffic surge definition in detector level

For each cluster, the traffic occupancy over a specified hour period should be distributed around their cluster center. An outlier is far away from the cluster center, and its level of deviation from the center can be justified by calculating its probability. We empirically check the distributions of traffic occupancy in all hour periods in different clusters, and two of them are shown in Fig.2.5. After reviewing the empirical distributions, we can conclude they reasonably approximate them with a normal distribution.

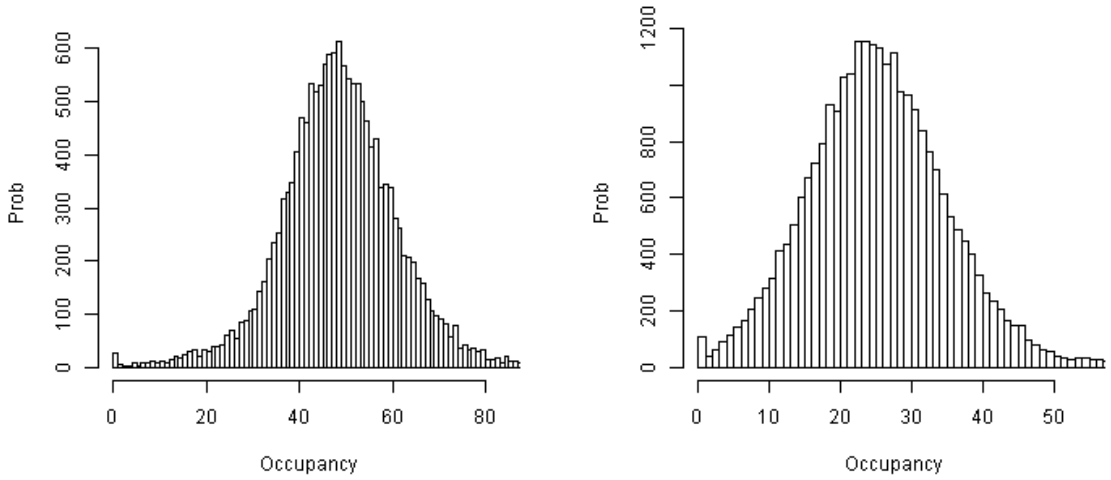


Fig.2.5 Distribution of traffic occupancy in two clusters at 7:00 p.m.

Thus, we can justify the severity of traffic increase based on the normal distribution. We define the traffic surge probability as probability $\emptyset\left(Z=\frac{o_j^{di}-c_j^i}{\sigma_j^i}\right)$ during the jth hour period in detector d and cluster i. Here C_j^i is the cluster center in Cluster i during hour j; σ_j^i is the standard deviation of traffic occupancy in cluster i. The closer the value is to 1, the greater the traffic surge should be.

2.5 Twitter concentration extraction and filtering

We can automatically extract Twitter concentrations that have the following features:

They are related to real events that many people witness and are willing to share their observations or experience via Twitter.

They potentially have connections with on-road traffic-related activities and may involve some kind of traffic movement.

First, we extract the Twitter concentrations from the set of tweets from January 2014 to December 2014. In most of the cases, keywords contained in the tweet can differentiate it from other tweets. Our algorithm first splits the tweet texts into separate word characters that form a large word database. In the database, we first search for the keywords that frequently appear in each day. Then, for all the frequent words, we select the words that frequently appear in one day but not so frequently or even vanish in other days.

Algorithm 2: Twitter concentration extraction

Input: Tweets collections throughout 2014

Output: Keywords of each day and tweets that contain the keywords

For each day

Decompose the tweets into vectors of words on that day

Count frequency $\boldsymbol{f_{w(d)}}$ of each word $\boldsymbol{w}$ on day $\boldsymbol{d}$;

Pick the words that satisfy $\boldsymbol{f_{w(d)} \geq \epsilon \sum_w f_{w(d)}}$

For all selected $\boldsymbol{w(d)}$

Select the words $\boldsymbol{w(d)}$ as the keywords $\boldsymbol{k(d)}$ that satisfy

$$\boldsymbol{count\left[w(d) \cap dom\left(w(d)\right)\right] \leq \tau}$$

For all $\boldsymbol{k(d)}$

Extract the tweets that contain $\boldsymbol{k(d)}$ in July

We set the frequency threshold for the keywords by a ratio parameter $\boldsymbol{\epsilon}$.. $count\left[w(d) \cap dom\left(w(d)\right)\right]$ counts the frequency of $w(d)$ in the domain of all $w(d)$, i.e. $dom\left(w(d)\right)$ and whose frequency is no larger than $\boldsymbol{\tau}$ is selected. For the selection of $\boldsymbol{\epsilon}$, if we increase the value of ϵ, it may miss some important keywords due to the large datasets. If we lower ϵ, it may incur more computations. Our experiences show that there are not so much difference between 10% and values lower than 10%. The value of τ should not be too large because longer periods of events will diminish the enthusiasm of the people and these events may not be a reflection of Twitter concentration. One may increase τ if their data covers more than one year because there may exist yearly events. In our study, we set $\boldsymbol{\epsilon}$ and $\boldsymbol{\tau}$ to be 10% and 3 respectively. By comparing the frequent words in different days, the stop-words such as "is", "and", "us", etc. can be eliminated and the remaining frequent words are the keywords that may indicate a kind of social activity. Table 2.1 shows some keywords of the day in July and some possible related social events.

Table 2.1 6 keywords and related social events in July 2014

Date	Keyword	Social events
7/1/2014	waffles	Waffle House restaurant just tweeted the most American waffle breakfast possible
7/3/2014	louis	Louis Zamperini, an American war survivor in World War II, died
7/4/2014	freedom	Celebration activity such as Freedom Fest 2014 fireworks
7/18/2014	fotosdeprinceroyce	Prince Royce concerts
7/16/2014	wjmc2014	Washington Journalism and Media Conference
7/26/2014	silver	A new metro line: silverline, is opened

Table 2.2 Transportation lexicon

Accidents	Carpooling	Drive	Junc	Passenger	Seatbelts	Trains
Arrival	Carriage	Driver	Junction	Passengers	Shuttle	Transit
Arrivals	Cars	Drivers	Junctions	Passing	Sidewalks	Transport
Arrive	Collision	Drives	Kilometer	Pedestrian	Speed	Transports

Arriving	Combustion	Driving	Kilometers	Pedestrians	Speeding	Trasport
Auto	Commute	Drop	Lane	Periods	Speedlimits	Travel
Automobile	Commuter	Eastbound	Licence	Petrol	Speeds	Travelcar
Automobiles	Commuters	Eastern	Line	Pickup	Standstill	Travelcard
Automotive	Commuting	Exhaust	Lines	Priced	Steer	Travelcards
Baggage	Congested	Exit	Link	Queues	Steering	Traveline
Bicycle	Congestion	Flows	Metering	Rd	Stops	Travelled
Bicycled	Connect	Freeflow	Motor	Ride	Taxi	Traveller
Bicycles	Connection	Freeway	Motorbike	Rider	Taxicabs	Travellers
Bicyclists	Connections	Freight	Motorbikes	Riders	Taxiing	Travellers
Bike	Crossing	Heading	Motorcycle	Riding	Taxis	Travelling
Bikeability	Crossings	Heathrow	Motorcycles	Road	Taxiway	Travelling
Biker	Crossroad	Highspeedrail	Motorcyclist	Roadmap	Taxiways	Travels
Bikes	Cyclist	Highway	Motorcyclists	Roads	Terminal	Trip
Bikesafe	Cyclists	Highways	Motoring	Roadsafety	Tolled	Trips
Bikesharing	Delay	Incident	Motorist	Roadside	Tolling	Truck
Biking	Delays	Incidents	Motorists	Roadways	Tolls	Trucks
Biofuel	Departing	Int0change	Motorization	Roadworks	Tour	Trunk
Biofuels	Departure	Interchange	Motorization	Roundabout	Tourists	Tunnels
Brake	Departureboard	Interchanges	Motorized	Route	Tow	Turn
Brakes	Departures	Intercity	Motorized	Routes	Towing	Turning
Braking	Destination	Intermodal	Motorpoint	Routing	Townbound	Uncongested
Breaks	Destinations	Intersection	Motorway	Runway	Track	Van
Bus	Direction	Intersections	Motorways	Runways	Tracking	Vans

Buses	Directional	Interurban	Navigating	Rush	Tracks	Vehicles
Busstop	Distance	Journey	Navigation	Safety	Traffic	Vehicular
Car	Districts	Journeyplanner	Parkers	Scooter	Trail	Wait
Cargo	Dock	Journeys	Parking	Scooters	Trailers	

Second, we filter the tweets extracted from the first algorithm to decide whether they are traffic-related. Compared with tweet classification, this algorithm is an unsupervised method that roughly estimates whether the tweets are traffic-related and works to shrink the size of the tweet collection. We hypothesize that the individuals describe their events by event-related words, and each traffic-related tweet should have one or more traffic-related words. A transportation lexicon is shown in Table 2.2 which is referred to (Gal-Tzur et al., 2014). We made some revisions by excluding some words that are related to air, water, railway traffic, etc. The tweets that contain at least one term in the lexicon are reserved otherwise discarded. We finally extracted 1179 candidate traffic-related tweets.

2.6 Twitter concentration classification and labeling

Twitter concentration classification is a supervised learning method that calculates the correlation between the tweets and traffic. We employed the logistic regression model, which is first introduced in 1958 (Freedman, 2009), as our learning model to train and test the tweets. We first train the model and use the model to label the candidate tweets obtained in Section 2.5 is traffic-related. The model is as follows:

$$\boldsymbol{F}(\boldsymbol{X}) = \frac{1}{1 + \mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{X}+\boldsymbol{\beta_0}}}$$

where $\boldsymbol{X} = (X_1, X_2, \dots X_i \dots X_m)^T$. X_j represents the vector of ith feature and there are m features in total. $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots \beta_i \dots \beta_m)$ is the vector of coefficients of each feature. $\boldsymbol{\beta_0} = (\beta_0 \dots \beta_0)$ is the vector of intercepts. $\boldsymbol{F}(\boldsymbol{X}) = (F(X_1), F(X_2), \dots F(X_i) \dots F(X_m))$ is the vector of probability values of the dependent variable.

The classification method proceeds in the following steps:

First, we pick randomly 2000 tweets that contain one or more words in the Transportation Lexicon in Table 2.2 from the tweet collections of the whole year. We manually label them to judge whether they are traffic-related. The labeled results are taken as the ground truths as well as the dependent variables $\boldsymbol{F}(\boldsymbol{X})$.

Second, each tweet is further decomposed into separate word characters that are called "tokens" in our study. The tokens can be English character, number or even Latinized letters and are taken as the candidates of independent variables. There are more than 6000 tokens in total.

Third, we conduct a stop-word filtering on the candidate feature words. The stop-word filtering is a prevailing method in page analyzer and article analyzer in preprocessing of natural language (Rajaraman et al., 2012). It can rule out the tokens that have no apparent linguistic meanings or significant event indications including articles, conjunctions, prepositions, pronouns, etc. The stop-word list we used referred to (Ranks-NL, 2015).

Fourth, we include those tokens that may correlate with the labels. The correlation benchmark we choose is phi coefficient (Cramér, 1999), which is widely accepted as a measure of association between two binary variables. The coefficient (usually denoted as ϕ) between two variables x and y is calculated as:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{11} + n_{10})(n_{01} + n_{00})(n_{11} + n_{01})(n_{10} + n_{00})}}$$

When $x = 1$, n_{11} and n_{10} are the counts separately for $y = 1$ and $y = 2$; when $x = 2$, n_{01} and n_{00} are the counts separately for $y = 1$ and $y = 2$. Those tokens whose correlation coefficient ϕ are higher than 0.05 are selected. These tokens totaling 71 are taken as our covariate features $\boldsymbol{X}$.

Fifth, we estimate the coefficients of the variables in the regression model by maximum likelihood estimation (MLE). This likelihood estimation can be realized by an iterative process such as Newton's method (Ryaben'kii and Tsynkov, 2006) and the estimation of both coefficient values and significance are detailed in (Cohen et al., 2013). To increase the accuracy of the predicted model, we implement 5-fold cross validation (Geisser, 1993), which is a popular model validation method. Cross-validation can give insight on how the model will generalize to an independent dataset. Directed by this method, the dataset is randomly partitioned into 5 folds. The classification model is trained on 4 folds, and the remaining fold is used for testing the trained model. This procedure is repeated 5 times and each fold is used exactly once as a test data set. We finally obtained an overall estimation by averaging 5 test results. The accuracy of the model is 0.76.

Finally, the prediction model obtained in the previous step is employed to test the candidate traffic-related tweets obtained in Section 2.5. In our study, we take $F(X_i)$ as the traffic accident probability of ith tweet data. The results show that of all 1179 tweets from the first classifier, 164 tweets may correlate with the traffic with $F(X_i) > 0.5$.

2.7 Correlations between Twitter concentrations and traffic operations

For each tweet, we mainly study the traffic related information within certain spatial and temporal ranges. The temporal ranges are set to be before and after one hour when the tweet is blogged. The spatial ranges are set to be 100m around where a tweet is blogged. It is worth noting that:

Public activities related to Twitter concentrations may happen either before or after when the tweet is blogged. So does the traffic surge.

As the geographic impact of public activities may vary, the traffic surge may exist in one or even more intersections nearby.

Thus, influenced by public activities related to the Twitter concentrations, there are mainly two different traffic surges: traffic surge in part of the detectors or over shorter time periods; traffic surge in most detectors and over long time periods. The first kind of traffic surge can be justified by the 75th percentile value of traffic surge probability:

$$q_{traffic} = Q3\left(\left\{\emptyset\left(Z = \frac{O_j^{di} - C_j^i}{\sigma_j^i}\right), d \in dom(d) \cap j \in dom(j)\right\}\right)$$

Where j is the hour period, d is the detector ID and i is the cluster ID. $dom(d)$ is the domain of all the detectors within the geo-scale of the tweets and $dom(j)$ is the domain of all time periods within the time-

scale of the tweets. $Q3()$ is the operator of 75th percentile. As this kind of traffic surge is dramatic in only part of the detectors while relatively mild in other detectors, value of $q_{traffic}$ corresponding to a tweet should be relatively high to justify a traffic surge.

The second kind of traffic surge can be justified by averaged traffic surge probability:

$$p_{traffic} = \frac{1}{NUM} \sum_{j \in dom(j)} \sum_{d \in dom(d)} \emptyset\left(Z = \frac{O_j^{di} - C_j^i}{\sigma_j^i}\right)$$

Where NUM is the total number of traffic occupancy data related to a tweet.

For a traffic-related Twitter concentration, its correlation to traffic surge can be justified by a threshold value of either $p_{traffic}$ or $q_{traffic}$. Here is our assumption on the detection of traffic surge from Twitter concentrations:

Assumption 2: For a traffic-related Twitter concentration, its correlation to traffic surge can be justified by either $q_{traffic} \geq q_{traffic}^0$ or $p_{traffic} \geq p_{traffic}^0$, where $q_{traffic}^0$ and $p_{traffic}^0$ are two parameters.

Given Assumption 2, the public events of Twitter concentrations can impact the surrounding traffic, and this impact can be justified by these two traffic surge probabilities. Two important findings are worth mentioning:

Empirical results show that the impact of different threshold values of $q_{traffic}$ on the result is low and even negligible.

If we set the threshold of $q_{traffic}$ to be 0.8, the percentage values of traffic-justified Twitter concentration events may change with $p_{traffic}$ as shown in Fig.2.6. Given $q_{traffic}^0$=0.8 and $p_{traffic}^0$=0.5, 127 out of 164 Twitter concentrations (77.4%) can be justified by traffic surge.

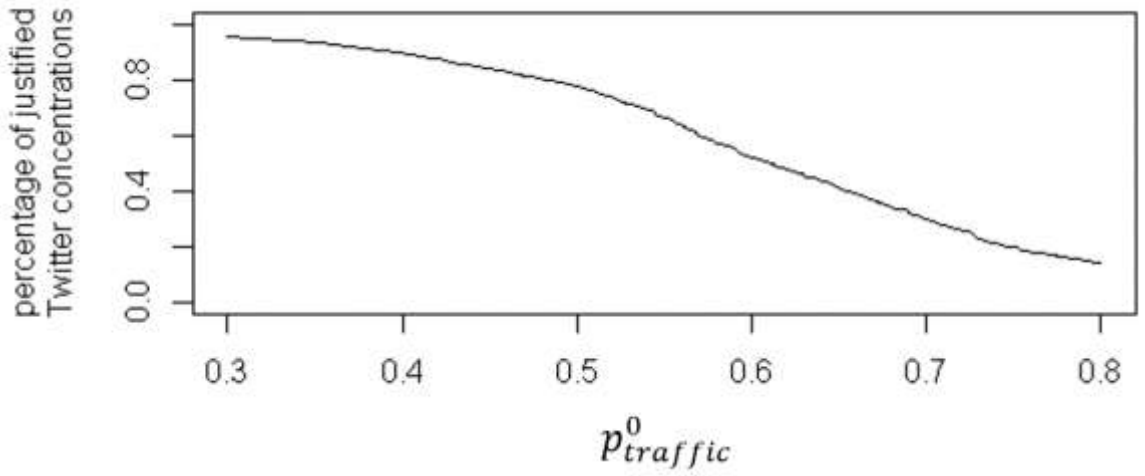


Fig.2.6 Percentage of traffic-justified Twitter concentrations under different threshold $p_{traffic}^0$

Different threshold values $p_{traffic}^0$ may influence the final results. It is obvious that a higher $p_{traffic}^0$, indicating a more serious traffic surge condition, may correspond to a public event that arouses more Twitter concentrations. It will be of great use in future study to further explore and quantify the severity levels of traffic surge using Twitter concentration.

Table 2.3 presents two Twitter concentrations and its corresponding public events. Fig.2.7 illustrates the time-of-day fluctuations of traffic surge probabilities of these two Twitter concentrations. One can see that the overall traffic surge probabilities in Fig.2.7(a) and 7(b) are above 0.5 for Twitter concentration (1). As a comparison for (2), traffic surge probabilities in one detector are high (see Fig.2.7(d)), but low in the other (see Fig.2.7(c)). This figure characterizes the influence levels of different Twitter concentrations in different geographic scales. In Table 2.3, keywords "silverline" and "4thofjuly2014" can justify the correlation between tweets and major public events. The Twitter concentrations indicate the occurrence of traffic-related activities that result from the public events. The results prove the potentials of Twitter concentrations in detecting the traffic surge. One can see that without knowing the type of public events in advance, detecting the traffic-related Twitter concentrations assists in interpreting the causality of traffic surge and provides insights for better decision-making in urban traffic management.

Table 2.3 The keywords, Twitter concentrations and public events corresponding to Fig.7

	Keyword	Twitter concentration	Public events
(1)	silverline	waiting at Wiehle to ride silverline	Silverline metro opened on July 26
(2)	4thofjuly2014	waiting for my friend to get here so we can roll out easternshore, 4thofjuly2014	Celebration activities on Independence Day

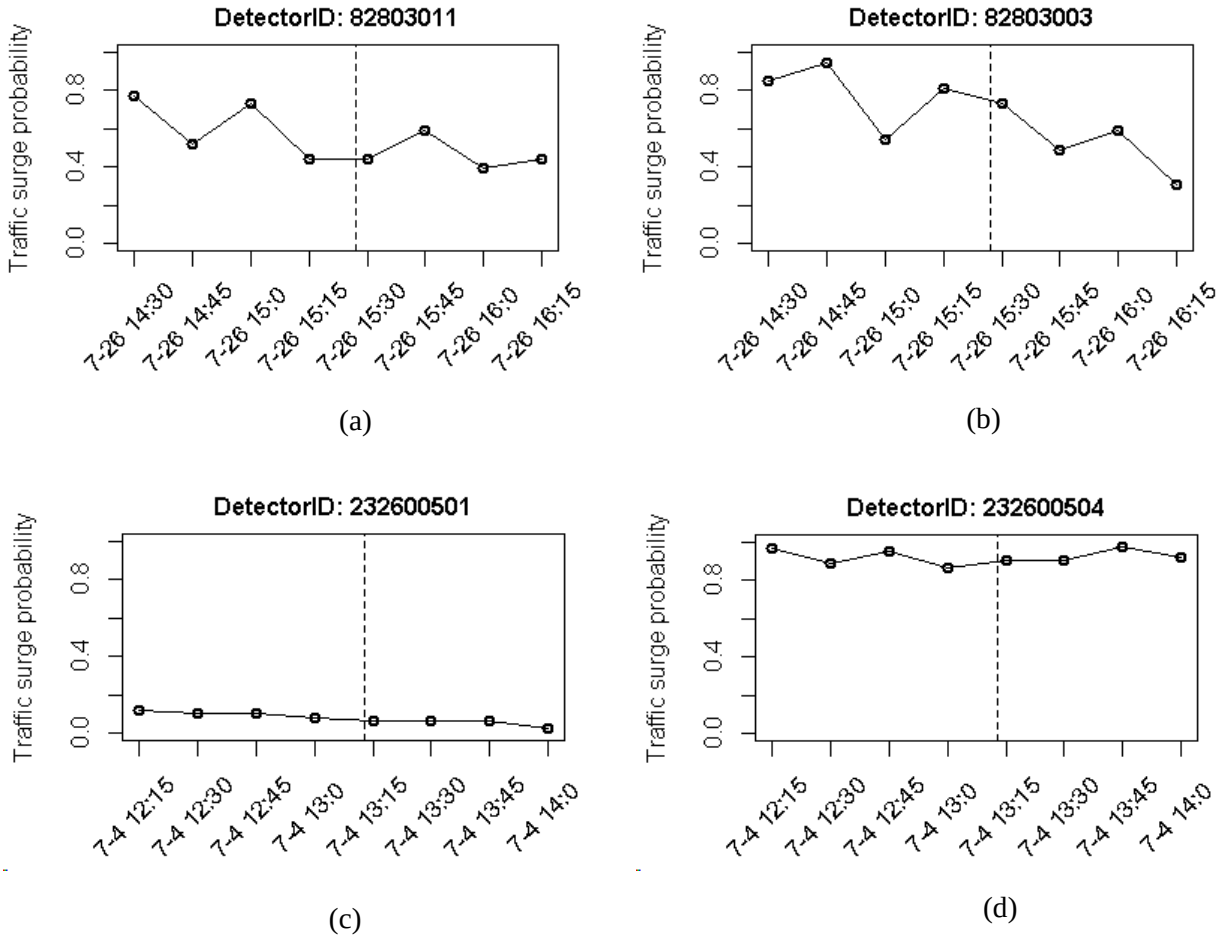


Figure 2.7 The traffic surge probabilities at different time-of-day that are related to Twitter concentration (1) and (2), listed in Table 2.3. Figure (a) and (b) represent two detectors associated with Twitter concentration (1), Figure (c) and (d) another two detectors with Twitter concentration (2). The dash line indicates when the tweet is blogged.

2.8 Conclusions and discussions

This chapter investigates the correlation between Twitter concentrations and the traffic surge on July 2014. The results prove the potentials of using tweets to detect the traffic surge within a given scale of space and time. First, the traffic occupancy over a certain period may follow a normal distribution, and this feature is fully exploited to derive the probability that quantifies the traffic surge. Second, the correlation between Twitter concentrations and traffic surge indicate that the major social activities that are related to traffic could possibly deteriorate the nearby traffic congestions. Our experiments show that 77.4% of traffic-related Tweeter concentrations can be justified by local traffic surge.

Following these findings, one may further the study by tackling the limitations of the current approach. In our study, the tweets are collected through Twitter Streaming API with geo-location filter and cannot possibly cover all the traffic surge of the whole region. This may be due to the limited volume of geo-tagged tweets. Also, the classification method employed in this chapter may be limited by the size of the training datasets, and the precision of the results may increase by incorporating more tweets.

The potential applications of our study are also promising: First, the traffic surge detection algorithm is built on the "big data" analysis of previous data collections. It can precisely unveil the traffic patterns in a large road networks and even identify the anomaly traffic conditions. Second, this study can help traffic operators understand the cause of traffic surge and improve short-term prediction of traffic congestion (especially non-recurrent congestion) on roadways in the future. Third, the Twitter concentration can broadcast the traffic-related events in a much more timely and quickly manner than traditional broadcasting media. Monitoring the social media data may deliver useful traffic event information, including traffic accident, traffic jam, road construction, etc. It will also be interesting to analyze the spatial-temporal correlations between traffic patterns and Twitter concentrations in future research.

3 Traffic Accident Detection with Both Traffic and Social Media Data

Social media receives increasing attentions as crowdsourced information for traffic operations and management. One recent trending study is to use social media to detect on-site traffic accidents. However, it remains unknown how effective the social media based detection methods is as compared with traditional loop detector based method. In this chapter, we first explore the features of keywords and their association rules inherent in the accident-related tweets and explore the potentials of tweets in accident detection. Combining the traffic flow and occupancy data, our prediction results show that tweets can sometimes respond to traffic accident much more quickly than traditional methods and can even find some un-documented accidents which make up for the deficiencies of VDOT records. Also, the limitations and disadvantages are also discussed which provide insights in utilizing social media data to assist accurate on-site traffic accident detection.

3.1 Introduction

The traffic accident is considered as one of the most important urban problems worldwide and may break down the traffic flow and disturb the traffic operations. Major traffic accidents can sometimes cause irreparable damages, injuries, and even fatalities. National Highway Traffic Safety Administration (NHTSA), which publishes yearly reports on traffic safety facts, states that since 1988 more than 5,000,000 car crashes occur in the States each year and about 30% of them bring fatalities and injuries (NHTSA, 2015). After years of research, it has been widely accepted that significant reductions of accident impact can be achieved through effective detection methods and corresponding management strategies. Accurate monitoring of traffic and effective detection of traffic accidents are critical to modern transportation management.

Due to the fact that major traffic accidents potentially interrupt the traffic flow, traditional attempts in traffic accident detection focus mainly on monitoring fluctuations and changes of one or more traffic-related metrics such as the traffic flow, occupancy, speed, etc. Some methods leverage the time-of-day characteristics and geographic features to identify the anomalies that may indicate a traffic accident. For example, Teng et al. (2003) features from traffic measurements in incident conditions are significantly different from those in normal conditions. Payne et al. (1978) used freeway traffic flow data for the detection of accidents and other lane blockage incidents that temporarily disrupt traffic flow; Tsai et al. (1979) applied a pattern-recognition approach to improve incident-detection algorithms; Sethi et al.(1995) separate incidents by locations and achieved a better detection rates by measuring the speed difference and speed ratio; Samant et al. (2000) developed an effective traffic incident detection algorithm to extract incident-related features from traffic patterns. Jin et al. (2009) proposed an incident decision-making algorithm to detect traffic incidents on the basis of traffic flow-occupancy relationships. With years of dedications within the field, the detection methods and algorithms based on the detector-based data are becoming mature. The algorithm includes various regression analysis (Sethi et al., 1995; Yuan and Cheu, 2003), Artificial Neural Network (Khan and Ritchie, 1998), Bayesian-based Network (Abdulhai and Ritchie, 1999; Zhang and Taylor, 2006), Time series algorithms (Teng and Qi, 2003; Willsky et al., 1980), etc. It is worth mentioning that besides measurements from loop detectors, the probe vehicle data also proves to be a reliable data source and can be included in the fixed detector and probe vehicle

algorithms in studies like (Sethi et al., 1995). Similarly, Amin et al. (2012) proposed to utilize the capability of a GPS receiver to monitor the speed of a vehicle and detect accident based on monitored speed; Park et al. (2015) estimated incident impacts and incident detection by using probe vehicle techniques, etc.

Despite the adaptabilities of these studies, the improvement in the accuracy of detection with only traffic data still meets certain challenges. First, most of the previous research, which utilized the field data to detect the traffic accidents, build on the implicit assumption that the data is reliable. However, requiring real-time data from traffic detectors is very expensive in maintenance and operations. Detector failures or data errors are perennial problems in traffic operations. For example, Illinois Department of Transportation (IDOT) in Chicago reported that no more than 5 percent of their loops (detectors) are inoperative at any given time (Kell et al., 1990). The percentage is not low enough as compared to the rate of a traffic accident. The problem of malfunctioned sensors cause even more troubles in incident detection in large regions, say, an area with more than 10,000 signalized intersections. Second, the uncertainty nature of traffic patterns and non-recurrent social activities may undermine the potential of traffic metrics in justifying the traffic accidents. Besides traffic accidents, daily traffic operations may suffer breakdowns by other factors such as parades, road constructions, running races, etc. Thus, the metrics including the traffic flow and occupancy inherently perform as an indirect support for traffic accidents instead of a direct proof. To address these challenges, there are efforts in applying clustering or classification methodologies such as K-means (Münz et al., 2007) on large data collections to diminish the errors. Other tendencies lie in incorporating more facts that relate to the real-time interaction of accidents such as probe vehicle trajectory.

Different from data sources from loop detectors or on-road vehicles, Twitter, the microblogging service that has received increasing attentions in recent years, has been gradually accepted as a direct user-contributed information source in event detection. Twitter creates an online environment where content is created, consumed, promoted, distributed, discovered or shared for purposes that are primarily related to communities and social activities, rather than functional task-oriented objectives (Gal-Tzur et al., 2014). Thus, in Twitter each tweeter acts as a data source of "We Media" and it is possible to retrieve the wide-range information from the broad masses of people in a timely manner. What is more, as more users tweet on mobile devices than on PC (Protalinski, 2012), the corresponding time and location information along with Twitter will be of great use in the detection and broadcasting of social events including earthquake (Sakaki et al., 2010), bird flu (Aramaki et al., 2011), politic events (Shirky, 2011), etc. The location effectiveness and timeliness features of Twitter can even find side-proof in a previous study which uses the GPS-enabled smartphones (White et al., 2011). As tweeter are able to describe what is happening on the scene site rather than a post-event recall and their tweeting locations may be quite near the scene site, the tweet contents are usually the priority in most of the studies through automatic detection of words in tweets especially the those that occur disproportionally frequently at the current time (Giridhar et al., 2014). There are also similar trials in the transportation fields. For example, Mai et al. (2013) compared incident records with Twitter messages and proved the potentials for information from Twitter to add context to other traffic measurements as a supplemental data source. Schulz et al. (2013) used microblogs to detect the small scale incidents; Gal-Tzur et al. (2014) conducted a corridor study on the correlation

between tweet and traffic jam. Gu et al. (2016) combined the data sources from Twitter, incident records, Here, etc. and employed the Naïve-Bayes classification to detect five major incident types; D'Andrea et al. (2015) compared accuracies and precisions of different regression models including Naïve-Bayes, Support Vector Machine, Neural Network, Decision Tree in detecting traffic incidents from Twitter stream. Most of these studies focus on methodologies of automatic information extraction from tweet contents through the state-of-the-art techniques of natural language processing (NLP) and this is usually the most difficult part in applying the tweet information for traffic-related purposes.

The challenges of using tweets to detect traffic accidents are also obvious. There are two major challenges to be addressed before the use of tweets in traffic accident detection. First, as compared to events that arouse enormous public concerns such as key basketball games, extreme weathers or traditional festivals, the influence of traffic accidents are comparably a "midget". From our observation, tweets related to traffic accidents are thus in small quantity. What's more, most of them are confined to a small area and limited to a relatively short time interval and some researchers call them small-scale events (Schulz et al., 2013). Thus, the effectiveness and limitations of tweets in detecting small-scaled events, especially the features of timeliness, accuracy, etc., should be explored and discussed. Second, the challenge in tweets lies in its inherent complexity and unstructured nature of data: language ambiguity (Chen et al., 2014). The common methods in detecting the traffic-related events include support vector machine (D'Andrea et al., 2015; Schulz et al., 2013), natural language processing (Li et al., 2012; Wanichayapong et al., 2011), etc. which explore the semantic features in the keywords. However, as the context of tweet is limited to 140 words and the tweet contents try to be concise, keyword detection is sometimes not sufficient for accurate automatic language processing. For example, "internet traffic is slow" and "internet shows traffic is slow" may deliver totally different information. To address above challenges, the association rules in the tweet contents should be explored and implemented in the traffic accident detection. Third, also due to the word limitation, some tweet contents which do not give enough descriptions to the incident types. Even some of incidents may come from their suppositions. The traffic incidents in (Gu et al., 2016) include car crash, construction work, bad weather, traffic congestion, etc. but all these types of incidents may also incur the congestion. The tweet users may even tweet there should be a traffic accident when they are delayed on road. As shown in Table 3.1, not all tweet users prefer to give a clear description of the traffic accidents. In our study, we only label the accident-related tweets which have explicit indications of traffic accidents and . However, the credibility of these tweets still needs further verification.

Table 3.1 Tweet samples describing the general traffic information, general traffic incident and road accident

General information	"I am waiting at the silver line, exciting"
	"Always hate the signals ahead of the hip-hop, making me sick"
General incident	"standstill for 1 hour, there must be accidents in front"
	"this is typical NOVA traffic, what a bad day"

Traffic accident	"major accident next to the sunoco near the parkway a car got flipped over"
	"the worst car accident possible just happened in front of me"

Instead of studying all traffic-related incidents, our interests lie in specifying the traffic accident including "collision", "disabled vehicle" and "vehicle on fire". Under this purpose, we employed two major supervised learning models: support vector machines (SVMs) (Cortes and Vapnik, 1995) and the supervised latent Dirichlet allocation (sLDA). From the view of SVMs, a tweet post can be disintegrated into a bag of words and those traffic-related words such as "accident", "crash", etc. can be taken as important features in the model. Besides single word features, the correlation features between these traffic-related words are also important because people sometimes describe a topic with word groups and these word groups ususally have more specific indications than single words. Thus, the association rules between words can possibly increase the prediction precisions of the model and should also be fully considered. From the view of sLDA, a tweet post can be disintegrated into a bag of topics. The proportations of those topics can be approximated by the Dirichlet distribution and even be inferred from word distriubtion in each topic in tweets. According to (Mcauliffe and Blei, 2008), the parameters of the topic and word distributions can be inferred by Expectation-Maximization (E-M) algorithm based on the labelled word documents. These words associated within a topic should be further examined. Our contributions can be summarized as: First, in addition to separately analyze the words, we reveal the association rules between words in each Twitter post and include the association features in our SVMs model for a more accurate accident detection; Second, we explore the possibility to increase detection accuracy and precision by combining the traffic-related metrics and tweet information. The role of traffic features in improving the accident detection is discussed further. Third, we compare results of traditional learning models and topic models. The drawbacks of topic models in detecting the specific traffic-related event: traffic accident is revealed. Fourth, by comparing the prediction results of several models with the ground truth from the traffic management log, we found that the tweets can possibly supplement the current accident detection records. The advantages and disadvantages of accident detection based on tweets are also discussed.

The rest of the section is organized in the following steps: Section 3.2 introduces the study area and the raw data sources. Section 3.3 details the model we use. Both the individual and paired token features are extracted and regression results with different features are presented and discussed. Section 3.4 compares our regression results with a supervised topic modelling method. Section 3.5 details the process of extracting the traffic-related information and explore the possibility of traffic information to improve the prediction. Section 3.6 compares the accident-related tweets with the ground truth to reveal the pros and cons of accident detection based on tweets. In Section 3.7, we conclude this chapter with a few empirical findings and generalizations together with some thoughtful discussions.

3.2 Data description

3.2.1 Raw data

The study area, shown in Fig.3.1 (a), is located in the vast road network of Northern Virginia (NOVA). With 2.8 million residents (about a third of the state), NOVA is the most populous region of Virginia and

the Washington D.C. Metropolitan Area. It has long been known for its heavy traffic (Cervero, 1994). The road network is a 50 square-kilometer (31 miles) area with more than 1,200 signalized intersections. In our study, we mainly include three categories of data:

The tweet data were collected through Twitter Streaming API with geo-location filter. Filtering by the coordinates, we extracted tweets posted only from NOVA region. There are more than 584,000 tweets from January 2014 to December 2014. Each tweet posts are coupled with specific date, time and location information. The tweets are the reflection of what people are interested at the specific time and location. Thus, they can justify the traffic accident if the text content has a clear expression of it. The location information is the paired latitude and longitude where the tweets are posted. The resolution of the location can be as high as 100 meters. Automatic extraction of accident-related tweets can be of great use in traffic management. The effectiveness of the detection is the major topic in our study.

The traffic data are collected by loop detectors equipped at the approach of the intersection. The detectors amount to nearly 15,000 in NOVA. These loop detectors keep recording the traffic flow and occupancy at an interval of 15 minutes. With these traffic detectors, the access to real-time traffic information in our study area is becoming routine as under growing pressure for improving traffic management (Leduc, 2008).

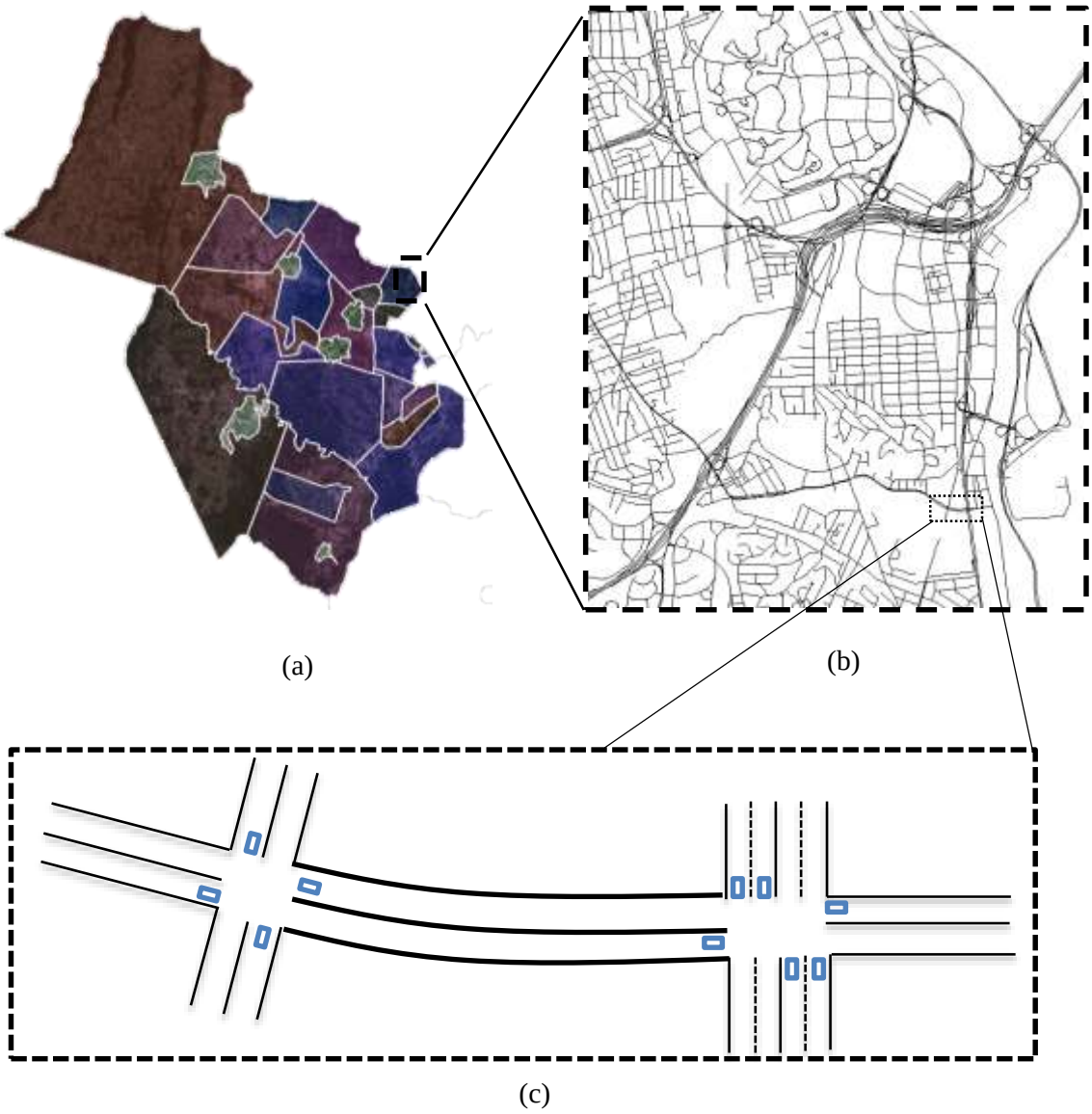


Fig.3.1 (a) Geographic districts of the study area, (b) road network map of one sample region and (c) locations of the detectors on approaches of the intersections

We even refer to the traffic management log maintained by Virginia Department of Transportation (VDOT). The traffic management log is an accident database recording the historical accidents in NOVA in the past few years. There are about 52,496 accidents happen in our study area throughout Year 2014. Each accident database is paired with detailed information of latitude, longitude, date, time and corresponding incident description. Such data are taken as the ground truth in our classification model to reveal the coverage and effectiveness of tweets in accident detection.

3.2.2 Tweet data preprocessing

We preprocess the raw data of tweet data to constitute the database that can be used for further analysis. The first step is to extract the candidate tweets that possibly describe the on-site traffic accidents. Usually, these candidate tweets should contain one or more keywords such as “accident” or “crash” that are

accident-related and we can assume that people describe traffic accidents by accident-related words. However, there has been no consensus on such a vocabulary of the accident-related words. Thus, we turn to the traditional news media and collect about 100 articles of news that broadcast the traffic accident. In all these articles, we select the words that appear the most frequently. The frequency of word is the times that a specific word appears in these articles. Except the common words such as "I", "is", etc. and those that reflect specific geographic and event features, we found that most of the articles mention the words with frequency higher than 20% as shown in Table 3.2.

Table 3.2 Accident-related words

"accident", "incident", "crash", "collision", "head on", "damage", "pile up", "rear end", "rear-end", "sideswipe", "lost control", "rolled over", "roll over", "tailgating", "police"

The second step is to extract the candidate tweets based on the accident-related words. We can apply the filter based on keywords to obtain the accident-related tweets. As compared to traditional media, social media blogs are broadcasted by the crowds and are without the editorial review. Some of the words may be grammatically correct. Thus, to ensure both the accuracy and sample size, certain rules must be followed:

Include the words that are relevant to accidents but apparently misspelled or personally modified including "acident", "incdent", etc.

Include other variations of accident-related words such as the word pairs that have a hyphen in word pair such as roll-over, etc.

Exclude the words related to transportation authority or news media.

Finally, we obtained more than 3500 candidate tweets. These candidate tweets can later be used to train the accident detection model.

3.3 Classification by SVMs

3.3.1 Process

In this section, we employ the supervised learning model SVMs (Karatzoglou et al., 2005) and the process of inferring functions of these models. The supervised learning consists of two major components: labelling and modelling. Labelling refers to the manual labelling process on the candidate tweets and the manual label is the categorical value assigned to each tweet. In our study, the two-class manual label is employed deciding whether the tweets are accident-related or not. After labelling, more than 400 tweets are taken as accident related. From the rest of tweets, we select non-accident-related tweets of the same size of accident-related tweets and combine them to constitute a tweet database. These tweets are symbolized as $\boldsymbol{T} = \{T_1, T_2, \dots, T_i, \dots, T_M\}$ and ith tweet is T_i. The corresponding label for T_i is L_i.

The output of the model is the manual label while the input is the features extracted from the tweet and the traffic information. The input features are one of major concerns in this chapter and will be fully detailed. SVMs can employ different kernel functions to keep the computational load reasonable. In our study, we employ the linear kernel to train and predict the models.

In the process of model training, we further implement 5-fold cross validation (Geisser, 1993) to increase the accuracy of the predicted model. Cross-validation can give insight on how the model will generalize to an independent dataset. Directed by this method, the dataset is randomly partitioned into 5 folds. The classification model is trained on 4 folds, and the remaining fold is used for testing the trained model. This procedure is repeated 5 times and each fold is used exactly once as a test data. We finally obtained an overall estimation by averaging 5 test results.

3.3.2 Token filtering and stemming

To fetch the proper features, each tweet is further decomposed into components. These components may including words, characters, numbers or even Latinized symbols which are collectively called "token". There are more than 10000 tokens from all eligible tweets. We can assume that some of the tokens may have no explicit meanings while some other tokens can potentially convey one or more instantaneous ideas and feelings of the tweeters. Part of them will be selected as the features of the regression models after necessary filtering and stemming. The steps can be illustrated as Fig.3.2.

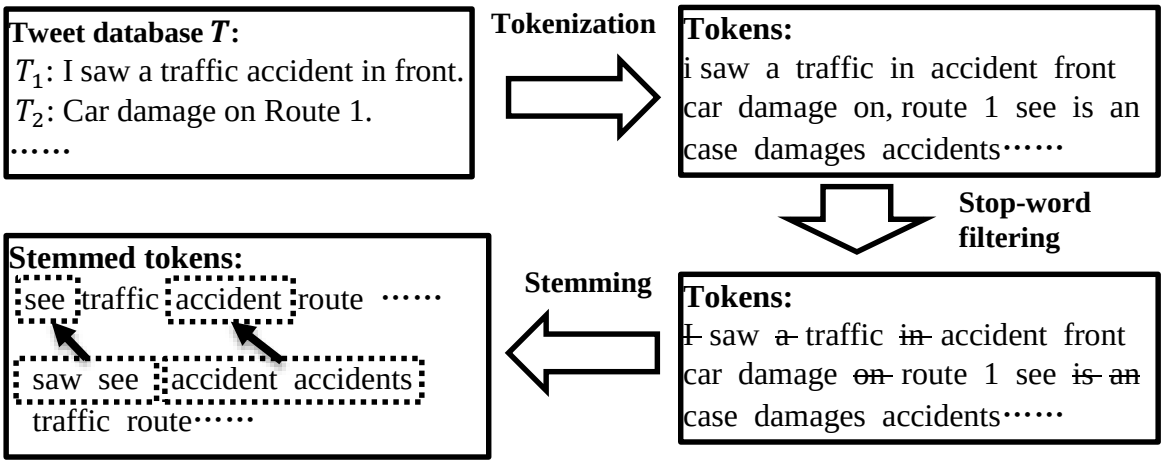


Fig.3.2 Steps of token filtering and stemming

First, the punctuation marks convey almost no meanings and should be discarded and all other words should be converted into lower case. Meanwhile, some of the words or characters that have no apparent linguistic meanings or significant event indications should be filtered out before the processing. These words are referred as stop-words. Stop-word filtering is a prevailing method in page analyzer and article analyzer in preprocessing of natural language (Rajaraman et al., 2012). The stop-word list we used refer to (Ranks-NL, 2015).

Second, some of the words have different writing expressions due to the grammatical reasons but convey almost the same meanings such as "accidents" and "accident". The token stemming is necessary to reduce these inflected (or sometimes derived) words to their word stem, base or root form. In this study, we employ the Porter stemming algorithm (Porter, 1980) for the token stemming and each token is grouped into the proper stemmed token.

After token filtering and stemming, each tweet T_i can be summarized several stemmed tokens. Of all the tweets $\boldsymbol{T}$, there are more than 3000 stemmed tokens symbolized as $\{t_1, t_2, \ldots \ldots t_j\}$. The stemmed tokens are the features for each tweet T_i and each tweet has different token features. If the tweet contains a stemmed token, the corresponding token features are labeled as 1 otherwise 0. Thus, the token features and the tweets $\boldsymbol{T}$ form our binary database $\boldsymbol{D_S}$ and it will be used for the feature selection.

Token filtering and stemming is the initial and necessary step for regression analysis in different models.

3.3.3 Classification with individual word

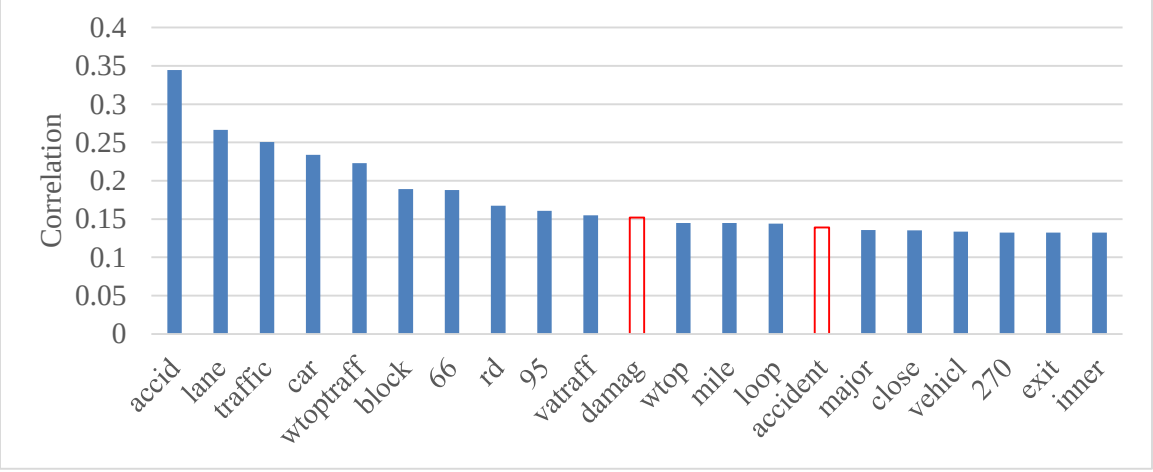
This section describes the steps of selecting features from individual token in the database $\boldsymbol{D_S}$ and the corresponding results. We focus on correlation between the individual token and our manual label. The correlation benchmark we choose is phi coefficient (Cramér, 1999), which is widely accepted as a measure of association between two binary variables. The coefficient (usually denoted as ϕ) between two variables x and y is calculated as:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1*}n_{0*}n_{*0}n_{*1}}} \quad (1)$$

Where all notations are defined in the following table:

	$y = 1$	$y = 0$	Total
$x = 1$	n_{11}	n_{10}	n_{1*}
$x = 0$	n_{01}	n_{00}	n_{0*}
Total	n_{*1}	n_{*0}	n

Those tokens whose $|\phi|$ is higher than 0.1 are selected. Following this rule, 27 tokens are selected and some of them are shown in Fig.3.3.



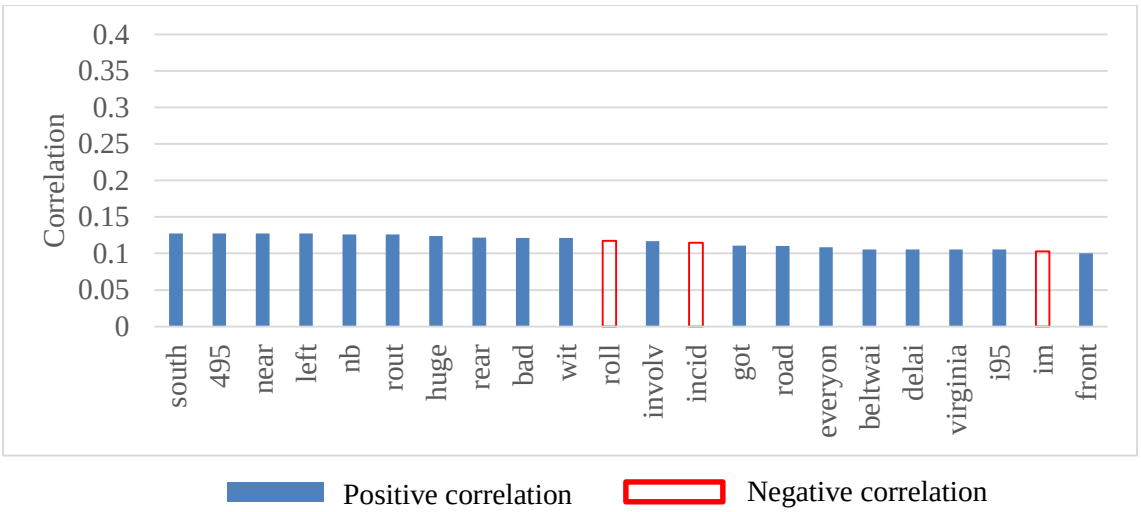


Fig.3.3 Correlations between the manual label and the individual stemmed tokens

From Fig.3.3, the stemmed tokens may be different from their original words in which "accid" refers to "accident"; "accident" does to "accidently"; "incid" does to "incident". Some of the tokens may be accounted by the geographic uniqueness such as "66", "95", and "495" which indicates the route number, and this means tweeter prefers to report the traffic accidents with route name; some may be the topic-related words including "traffic", "accident", etc.; other words such as "damage" or "accidently" are too general in our daily lives and thus lose the uniqueness in describing the traffic accident.

With selected individual tokens as the input of the regression model, one can simply compare the results by token features. To evaluate the achieved results in different models, we employed statistical metrics: accuracy and precision:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} / \frac{TN}{TN + FN} \tag{3}$$

Where

	$P_{Model} > 0.5$	$P_{Model} \leq 0.5$
$Groundtruth = 1$	TP	FP
$Groundtruth = 0$	FN	TN

One can set different correlation coefficient ϕ to determine the number of the token features in the regression model. Theoretically, more token features may increase the accuracy and precision of the regression results but may increase the computational time, model complexity or even cause over fitting as shown in Fig.3.4. When we set the ϕ as 0.2, there is only 4 qualified tokens. With the decreasing of individual tokens, the number of individual tokens boom as expected while the accuracy of the prediction will increase in a comparably much slower speed. When ϕ is equal to 0.15, we can obtain an accuracy of 0.784 with 11 individual tokens. With decreasing of ϕ, more tokens involved will not change the results

significantly. One can see that with less than 15 words, the tweets can be classified with an accuracy around 0.78 which means there is room for the model to improve.

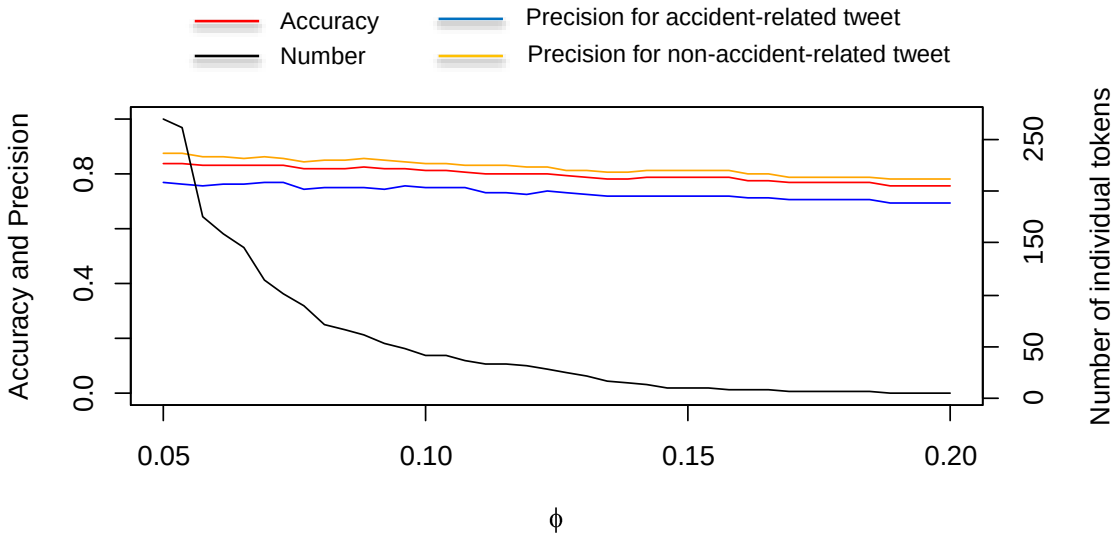


Fig.3.4 Regression results with individual tokens under different values of correlation coefficient ϕ

3.3.4 Classification with paired words

Features from individual token may sometimes not be sufficient to automatically classify the tweets because these may overlook the interconnections between words and sometimes the associations between words can have much more significant indications than single ones. For example, in a tweet post, the occurrence of word "car" conditioned by "accident" may increase the accident-related probability. Conversely, the occurrence of token "car" conditioned by "maintenance" or "repair" may undermine the likelihood of accident-related tweets.

In this section, we select the features from paired words by studying the association rules between the manual label and the stemmed tokens in the binary database $\boldsymbol{D_S}$. The association rules can be unveiled by the Apriori algorithm (Agrawal and Srikant, 1994; Hahsler et al., 2007). Apriori algorithm can find the regularities in large-scale binary data by two major probabilities: support and confidence.

We label all stemmed tokens in the database as $\boldsymbol{t} = \{t_1, t_2, \dots, t_j, \dots, t_N\}$. Given a stemmed token t_j, support of t_j is the proportion of tweets which contains t_j in the database.

$$supp(t_j) = \frac{sizeof(\{T_i, t_j \subseteq T_i\})}{sizeof(\{T_i\})} \tag{4}$$

Where t_j is the jth token; T_i is the ith tweet. Setting a threshold of $supp(t_j)$, we can filter out a limited number of qualified t_j. Similar to the support of each individual token, we can even calculate the support of paired tokens $supp(t_{j_1} \cap t_{j_2} \cap \dots \dots \cap t_{j_m})$:

$$supp(t_{j_1} \cap t_{j_2} \cap \dots \dots \cap t_{j_m}) = \frac{sizeof(\{T_i, t_{j_1} \cap t_{j_2} \cap \dots \dots \cap t_{j_K} \subseteq T_i\})}{sizeof(\{T_i\})} \tag{5}$$

Where $j_1 \neq j_2 \neq \cdots \neq j_K$. The paired tokens can be the combination of any two or more individual tokens. The concurrent tokens in one tweet post are quite common such as "traffic accident", "severe injury", etc. One can see that support deals mainly with the frequencies of one or more tokens. As our tweet database are filtered according to several different keywords, the word combinations of accident-related tweets may be also quite different. Thus, support of paired tokens can possibly capture different concurrent tokens that can possibly be used as the features in the model. But not all of them may be qualified as the features in the model. Besides support, the association rule between manual label and the paired tokens can be further revealed by confidence calculated as:

$$conf\left(L_i \Rightarrow t_{j_1} \cap t_{j_2} \cap \ldots\ldots \cap t_{j_m}\right) = \frac{supp\left(L_i \cap t_{j_1} \cap t_{j_2} \cap \ldots\ldots \cap t_{j_m}\right)}{supp\left(t_{j_1} \cap t_{j_2} \cap \ldots\ldots \cap t_{j_m}\right)} \quad (6)$$

In the confidence calculation, we focus more on paired tokens that are related to traffic accident which means L_i is equal to 1 in Equation (6). The maximum size of a paired token feature is theoretically equal to the total counts of tokens in $\boldsymbol{t}$, but due to the limited size of the tweet posts, larger size will be of no use. Also, if one increase the size of the paired tokens, the computational time will dramatically increase bringing almost no benefit. Our initial examinations show that almost no association rule exists in tweets when size of paired tokens is larger than 7.

In most of the previous study, setting support and confidence is sometimes mandatory. The setting of support can be a small value which can include as many as paired tokens for feature selection. The setting of confidence, as compared, usually influence the results significantly and different values should be further studied in the classification see its impact. We conducted an empirical studies to see how the token features can reveal the language of customs of tweeter in describing traffic accident. When support is equal to 0.01 and confidence is equal to 0.1, our results show that most paired tokens contain "accident". Other paired tokens can be traced in the findings as shown in Fig.3.3.

Table 3.3 Paired tokens by Apriori algorithm

accid	vatraff	accid	close	accid	got	
accid	mile	accid	road	accid	im	
accid	wtop	lane	block	accid	car	
accid	major	accid	block	accid	lane	block
accid	left	lane	wtoptraff	accid	lane	wtoptraff
accid	near	accid	wtoptraff	accid	car	got
accid	95	accid	bad			
accid	66	accid	lane			
accid	rd	accid	traffic			
accid	involv	car	got			

Same as individual tokens in Section 3.3.3, the paired token features in the database are equal to 1 if the tweet contains the corresponding paired tokens and 0 otherwise. We made analysis by incorporating paired token features into the regression model. When selecting individual tokens, the ϕ is set to be 0.15; when selecting paired tokens, support is set to be 0.01 and this value is set to involve as many as possible the paired tokens to compare the regression results; the confidence value is manually changed from 0.1 to 1.

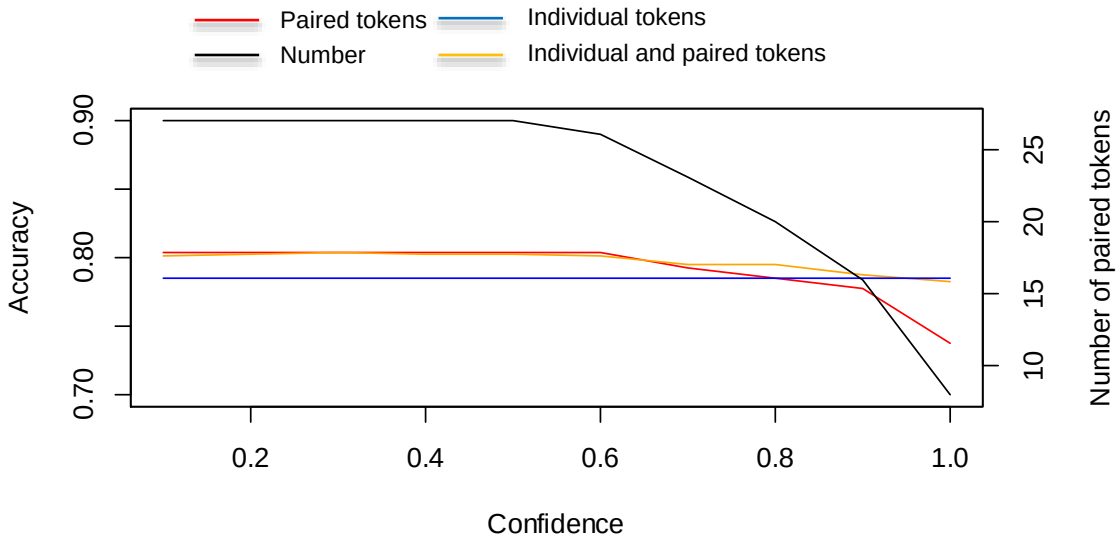


Fig.3.5 Regression results with paired tokens under different values of confidence where support is equal to 0.01

Our major findings are that incorporating the paired token features can improve the accuracy by 2~3%. By comparing the results of different confidence values, one can see that high value of support will not obtain enough paired tokens to improve the accuracy of the prediction. With confidence value decreasing, more paired tokens are involved in the model as expected but only a few of them can improve the accuracy and precision. Our data shows that when confidence is set to be 0.6, around 25 paired token features can obtain an accuracy of 0.808. Recall that around 100 individual token features can obtain an accuracy around 0.8, paired tokens are more efficient in using less words to obtain the same accuracy.

It is worth mentioning that our study only focus on the association rule between tokens and the manual label not that between all tokens. Thus, one can call that a supervised association rule mining instead. One can be satisfied with the regression results because it can obtain the accuracy with much less features involved in a more efficient way. There may be two major explanations for this:

As compared to the detailed accident reports or web news, the tweets are limited to 140 words and the word counts may be even less in practice. Tweet users are more accustomed to word (token) pairs to detail the incidents which can be both concise and clear.

The association rules of tweets needs less computing time than that of long contexts. According to our study, the required computing time for results in Table 3.3 takes about 17 min with computers (i7 3720QM, 32G RAM). Short text length of tweets facilitates the association rule mining.

3.4 Comparing with the classification by sLDA

3.4.1 Process

In this section, we employ the supervised Latent Dirichlet allocation (sLDA) (Mochihashi, 2009) and compare the results between SVMs and topic analysis in classifying the traffic accidents from tweets. As compared to the SVMs, topic analysis assumes that a topic is a probability distribution over a group of words (tokens) which describe a semantic theme and the features of a document can be divided into several different topics instead of different words (tokens). Thus, sLDA is capable of reducing the dimensionality of the words. As compared most of the topic models including Latent Dirichlet allocation (LDA) which are unsupervised, sLDA can infer latent topics predictive of the response on the basis of a manual label. The major differences between unsupervised and supervised topic models is the techniques to reduce dimensionality. The advantages of sLDA have been proved in several studies. However, the effectiveness of tweets is under question mainly in first, compared to the data sources like film reviews (Boyd-Graber and Resnik, 2010), image (Rasiwasia and Vasconcelos, 2013), etc., tweets have less words and may not generate reliable topics; second, unlike topics like Named Entity (Xu et al., 2009), sentiment (Lin et al., 2012), etc., traffic-related topics are comparably less general.

According to (Mcauliffe and Blei, 2008), each tweet post and label arises from the following generative process:

Draw topic proportions $\theta|\alpha \sim Dir(\alpha)$;

For each word

(a) Draw topic assignment $z_n|\theta \sim Mult(\theta)$;

(b) Draw topic assignment $w_n|z_n, \beta_{1:K} \sim Mult(\beta_{z_n})$;

Draw response variable $y|z_{1:N}, \eta, \sigma^2 \sim Mult(\eta^T \bar{z}, \sigma^2)$.

Where $Dir(\alpha)$ is the Dirichlet distribution; $Mult(\theta)$ is the multinomial distribution; z_n is the topic of the word w_n (token); β_{z_n} is the multinomial distribution parameter for z_n; $\bar{z} = (1/N)\sum_{n=1}^{N} z_n$. We follow the generative process and E-M procedure in (Mcauliffe and Blei, 2008) to infer the unknown parameters in the topic and word distributions. We implement 5-fold cross validation in the process of model training as in Section 3.3 and this will not be detailed.

3.4.2 Comparisons of classification results

In sLDA, there are only two latent topics (label): accident-related and non-accident-related. For the tweet post, the words are stemmed and tokenized and stop-words are filtered as discussed in Section 3.3. The results are shown in Table 3.4 and Fig.3.6. The topic words are tokenized before classification as discussed in Section 3.3.2.

Fig.3.6 compares the regression results between sLDA and SVMs with paired token features. In SVMs, support and confidence are set to be 0.01 and 0.6 respectively. The results of sLDA are slightly lower than SVMs. The precision is better than that in (Gu et al., 2016) mainly because there are 5 different categories (labels) in that study. However, the precision for non-accident-related is too low meaning that a number of non-accident tweets are predicted wrongly by sLDA.

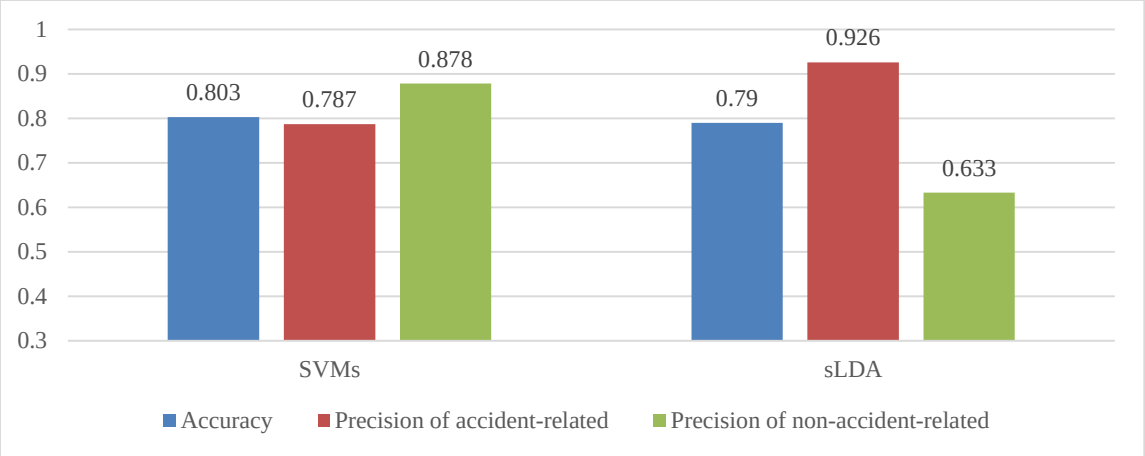


Fig.3.6 Comparisons of accuracy and precision between SVMs and sLDA

The first 30 topic tokens generated by sLDA are shown in Table 3.3. In the table, Topic 2 refers to the tokens that are accident-related while topic 1 is the reverse. By comparing with the results of individual tokens in Fig.3.3, the positive correlated tokens are bolded and blue-colored while the negative correlated tokens are shaded and red-colored. Those tokens in topic 2 list contains most of the tokens that have a positive relationship with the manual label while some in topic 1 are those with a negative relationship. One can admit that to some extent, sLDA properly classify the tweets that are accident-related but the precision may be greatly influenced by the tokens that have no specific meanings.

Table 3.4 Lists of top 30 topic words (tokens) of sLDA models on tweets

Topic 1	like im damag incid accident dont shit can tailgat will u lol roll time peopl dai hi make cant hit wa onli love thing still control life call lost someon
Topic 2	car accid traffic lane got bad wtoptraff block befor road rd close todai va involv almost 3 66 polic 2 95 two near left okai bu major 1 front

3.5 Improvements of classification by traffic-related information

In principle, the fusion of multi-source data provides significant advantages over single source data (Hall and Llinas, 1997) and the integrations of association features inherent in the tweet contents and other data sources are expected to produce more synthetic and informative results. As the traffic accident potentially influence the road traffic operations, the abnormal patterns of traffic-related information are potential features. It is also a viable method of monitoring the traffic operations in traditional studies (Coifman et al., 1998; Oh et al., 2001). Two major problems exist: first, the impact of traffic accident to its surround areas is unknown both in time and geographic scale; second, the traffic patterns are difficult to identify given the large volumes of historical data. Here we employ a systematic method to extract the traffic-related features.

3.5.1 Traffic pattern identification

The recurrent traffic pattern of each detector can be unveiled by studying the historical traffic volume and occupancy data. For each detector, we evenly divide the traffic occupancy into N separate groups. For each traffic occupancy group, we take the median of the corresponding traffic flow values as the traffic

signature. We use the median because it is less affected by outliers than mean. The traffic signature of a detector d is defined as the vector of these traffic flow values. That is $\boldsymbol{F^d} = (F_1^d, F_2^d, \cdots, F_o^d, \cdots, F_N^d)$. Where F_o^d is the median value of traffic flow given a range of occupancy o in detector d. One can see that for each detector, the traffic pattern is a vector of N traffic flow values. If there is no traffic flow record over a certain occupancy, we employed the linear interpolation of traffic flow median of adjacent occupancies. We can finally obtain the traffic signatures of more than 15,000 detectors in over 1,250 signalized intersections. In this chapter, N is set as 50.

According to the thorough study of the fundamental diagram (Jin and Ran, 2009), it is widely accepted that there exists a relationship between the traffic flow and occupancy (or density). However, the hypothesis of this relationship is diverse (e.g. triangle, parabola, trapezoid, broken-line, etc.). In our study, we do not make assumptions about this relationship between $\boldsymbol{F^d}$ and its corresponding occupancy. Instead, we assume the unchanged nature of the relationship:

Assumption 1: there exists an unchanged traffic signature in a given location. The traffic flow corresponding to a certain occupancy interval will mostly fall into a reasonable range, and those that deviate from the feasible range are traffic outliers.

To validate the assumption, we employ the K-means algorithm without pre-defining the clustering centers and the number of clusters to reveal the relationship. K-means clustering algorithm can partition the traffic signatures into finite groups of similar patterns. The inputs are the traffic signatures of all detectors, and the outputs are the collection of cluster centers and the cluster IDs that detectors belong to. We employ Akaike information criterion (AIC) (Akaike, 1998) to find the proper number of clusters. AIC measures the relative quality of the clustering results, shown in Equation (7).

$$AIC = \sum_{i}^{k} \sum_{d \in dom(i)} d(\boldsymbol{F}^{di}, \boldsymbol{C}^i) + k \cdot N \tag{7}$$

Where $\boldsymbol{F}^{di}$ denotes the traffic signature of the dth detector that belongs to ith cluster. $\boldsymbol{C}^i$ is clustering center of the ith cluster. $d(\boldsymbol{F}^{di}, \boldsymbol{C}^i)$ is the Euclidean distance between traffic signature $\boldsymbol{F}^{di}$ and its clustering center $\boldsymbol{C}^i$. $dom(i)$ is the domain (collection) of all detectors in ith cluster. k is the current number of clusters. N is the count of elements in a traffic signature, which equals to 50 in our study.

Our algorithm starts with the lower bound of the number of clusters and iterates the K-means clustering by increasing the cluster number. We calculate the AIC difference between the current iteration and the previous one. The iteration ends until the AIC difference is less than ϵ. The algorithm is as follows:

Algorithm:

Input: The maximum number of clusters $\boldsymbol{K}$, and traffic signature $\boldsymbol{F^d}$ for all detectors. (in this study, our data has more than 15,000 rows and 50 columns. Each row represents the traffic signature of a detector.)

Output: Centers of clusters $(\boldsymbol{C^1}, \ldots, \boldsymbol{C^i}, \ldots, \boldsymbol{C^k})$;

Cluster IDs detectors belong to.

Assign the initial number of clusters k=2, initialize AIC= +∞

Repeat

Implement K-means clustering algorithm with k clusters:

Pick randomly the cluster centers ($\boldsymbol{C^1}, \ldots, \boldsymbol{C^i}, \ldots, \boldsymbol{C^k}$);

Repeat

Cluster each traffic signature $\boldsymbol{F}$ to the nearest cluster center $\boldsymbol{C^i}$ with $\mathbf{min}(\boldsymbol{d}(\boldsymbol{F^{di}}, \boldsymbol{C^i}))$;

Replace $\boldsymbol{C^i}$ by $\boldsymbol{mean}(\boldsymbol{F^{di}})$;

Until none of the detectors **switch** clusters;

Calculate the AIC difference between each cycle;

Until AIC difference <= $\boldsymbol{\epsilon}$ or k= $\boldsymbol{\kappa}$

The AIC values will theoretically decrease with the increase of k. In this chapter, we set ϵ as 3%. When k=15, the change in AIC goes lower than 3%, as shown in Fig.3.7.

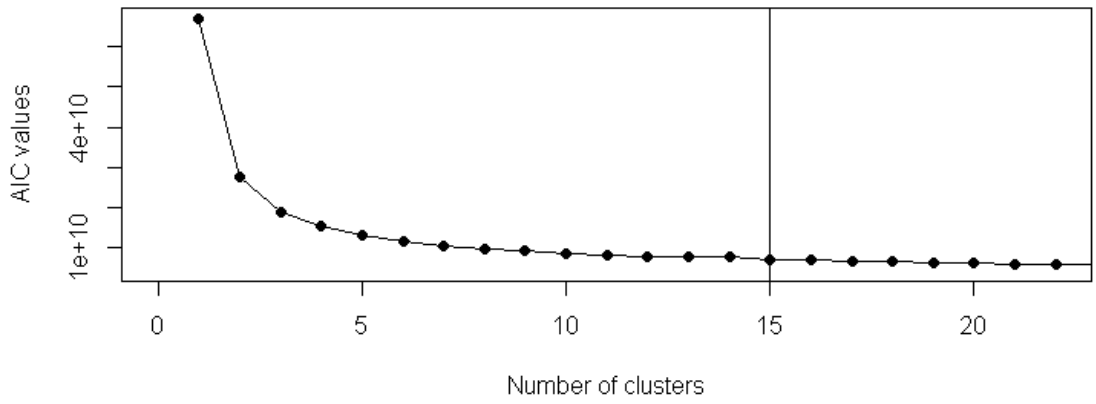


Fig.3.7 AIC values for different number of clusters

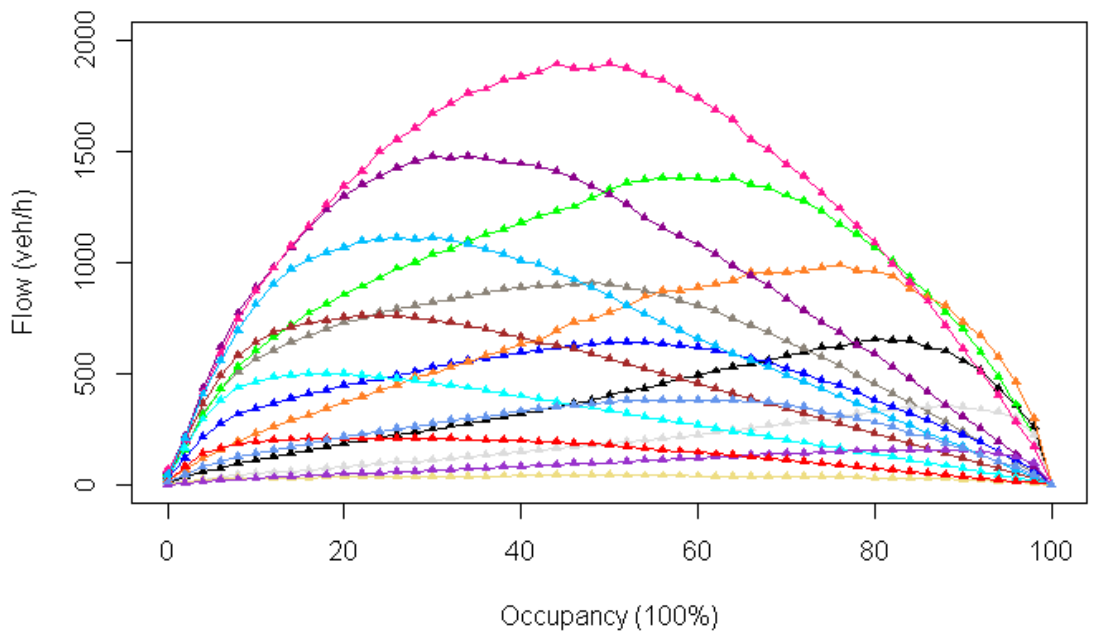


Fig.3.8 15 different clustered centers of traffic signatures

Thus, we finally cluster nearly 15,000 detectors into 15 different groups. The centers of clusters are shown in Fig.3.8. From the shape of our clustering results, it is not surprising that the relationships between traffic flow and occupancy differ greatly from each other. Unlike a predefined relationship, this method has certain advantages:

The method is totally driven by the analysis of large-scale data. The aggregation analysis of large-scale data can lead to reduced noise in the results.

The method clusters the traffic signatures with similar traffic patterns and potentially identifies the location of detectors that hold similar characteristics in the road network.

The method excludes the influences of daily differences or time-of-day differences inherited in the traffic data.

3.5.2 Abnormal pattern identification and traffic-related features

The output cluster centers represent the relationship between traffic flow and occupancy. One can intuitively figure out the possible traffic outliers by comparing the clustered center to the original data as shown in Fig.3.9.

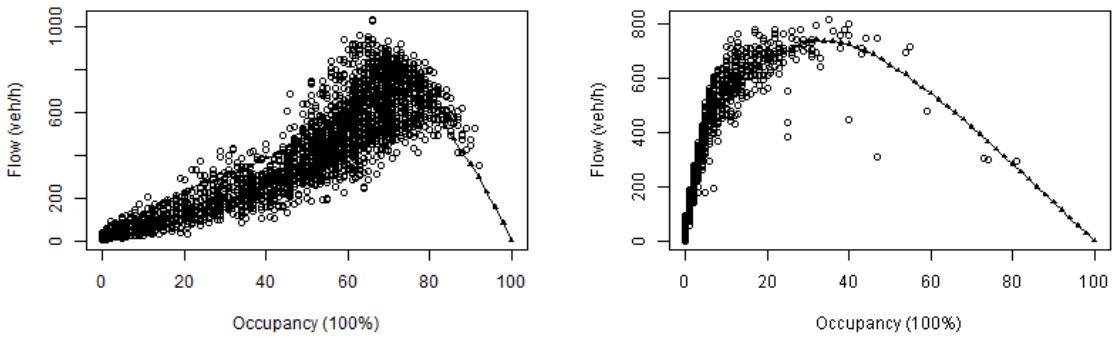
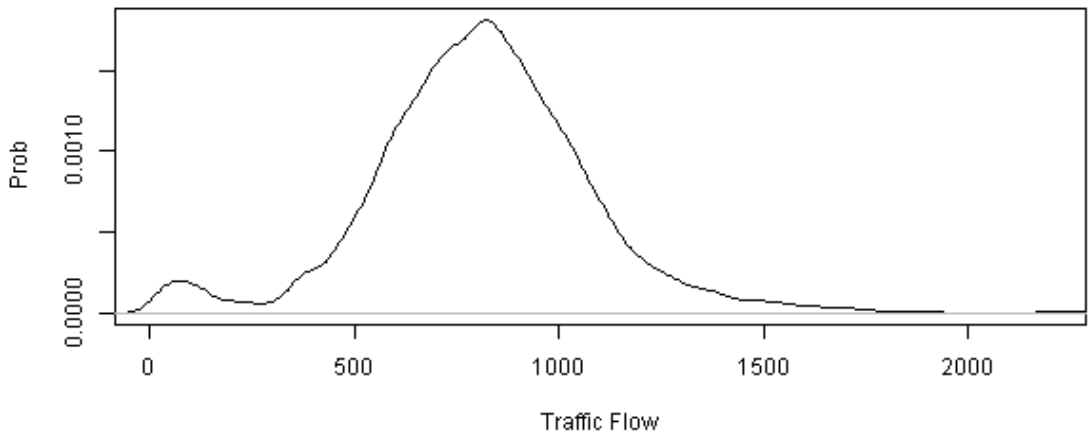
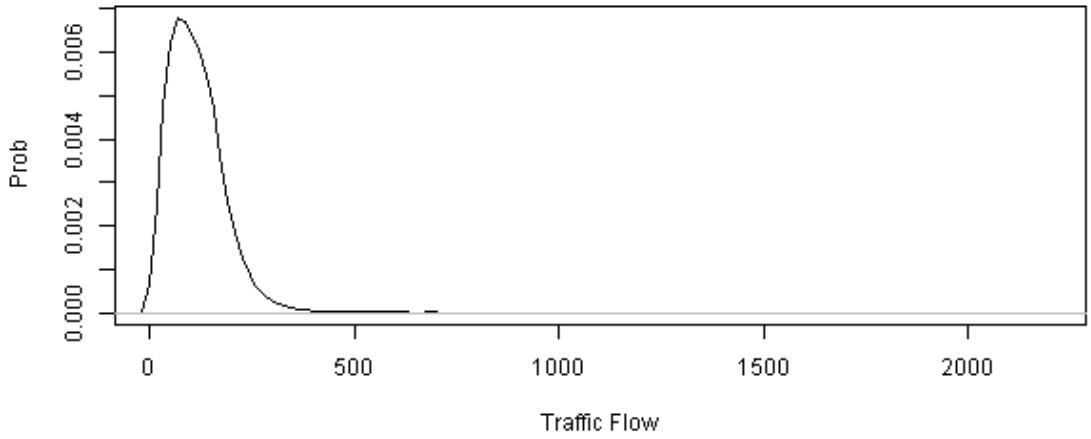


Fig.3.9 Comparisons between clustered centers and the original traffic flow and occupancy data in two sample detectors

For each cluster, the traffic flows over a specified occupancy interval are distributed around their cluster centers. Further, the outliers can be quantified by a probabilistic method that measures its deviation degree. Our empirical examinations show that the distributions of the traffic flow in a particular cluster and occupancy interval follows a Gaussian distribution shown in Fig.3.10. The traffic outliers can be intuitively identified in the distribution tail.



(a)



(b)

Fig.3.10 The traffic flow distribution over a range of occupancy in Cluster (a) Number 1 and (b) Number 5. In both cases, the occupancy interval is set as [20%, 22%].

Thus for each detector, the abnormal degree of traffic-related data can be quantified by the cumulative probability of the distribution.

$$P^{dt} = \Phi\left(\left|\frac{\mathcal{F}_o^{dt} - C_o^i}{\sigma_o^i}\right|\right) \tag{8}$$

Where P^{dt} is the probability for detector d over time period t. i indicates the ith cluster of d; $\mathcal{F}_o^{dt}$ is the traffic flow data over traffic occupancy interval o; σ_o^i and C_o^i is the standard deviation and center of traffic flow in Cluster i over occupancy interval o. P^{dt} quantifies abnormal probability for the deviation of traffic data from its cluster centers. The larger P^{dt} is, the worse the traffic operations should be and the more likely the traffic is influenced by traffic accident. This probability can be employed as the traffic-related feature in our model.

In the process of extracting traffic-related features of a tweet, we mainly study the traffic related information within certain spatial and temporal ranges. The temporal ranges are set to be before and after one hour when the tweet is blogged. The spatial ranges are set to be 100m around where a tweet is blogged. For each tweet, the corresponding abnormal probabilities will be further aggregated based on two major considerations: From the geographic perspective, as the geographic impact of the traffic accident may vary, the increase of traffic probabilities may happen either in all places around the accident site or just only in partial places. From the temporal perspective, as the traffic accident may happen either before or after when the tweet is blogged, the increase of abnormal probabilities may happen either over the whole time period or just only a certain time span. According to these considerations, two features are then generated for our regression model for each tweet:

$$p_{traffic} = \frac{1}{NUM} \sum_{t \in dom(t)} \sum_{d \in dom(d)} P^{dt} \quad (9)$$

$$q_{traffic} = Q3(\{P^{dt}, d \in dom(d) \cap t \in dom(t)\}) \quad (10)$$

Where t is the hour period; d is the detector ID and i is the cluster ID; $dom(d)$ is the domain of all the detectors within the geo-scale of the tweets and $dom(j)$ is the domain of all time periods within the time-scale of the tweets; $Q3()$ is the operator of 75th percentile; NUM is the total number of traffic data related to a tweet. It is worth mentioning that both $p_{traffic}$ and $q_{traffic}$ are discretized before putting into the regression model.

3.5.3 Classification results with traffic-related information

We make a simple comparison between the prediction results with and without traffic-related features. In SVMs, support and confidence are set to be 0.01 and 0.6 respectively. The results shown in Fig.3.11 indicate that traffic-related features will not improve the prediction results.

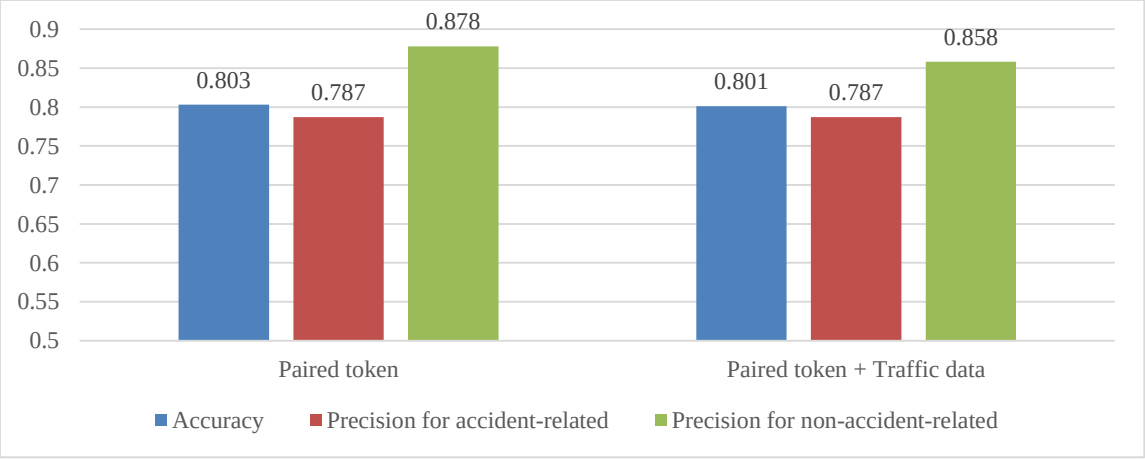


Fig.3.11 Comparisons of accuracy and precision with and without traffic features

One possible explanations for this may be that some traffic accidents may not influence the traffic operations, or the locations of the tweets do match that of the accident sites. This needs to be verified by comparing the accident-related tweets with the traffic management log.

3.6 Comparison with ground truth

3.6.1 Coverage of accident-related tweets

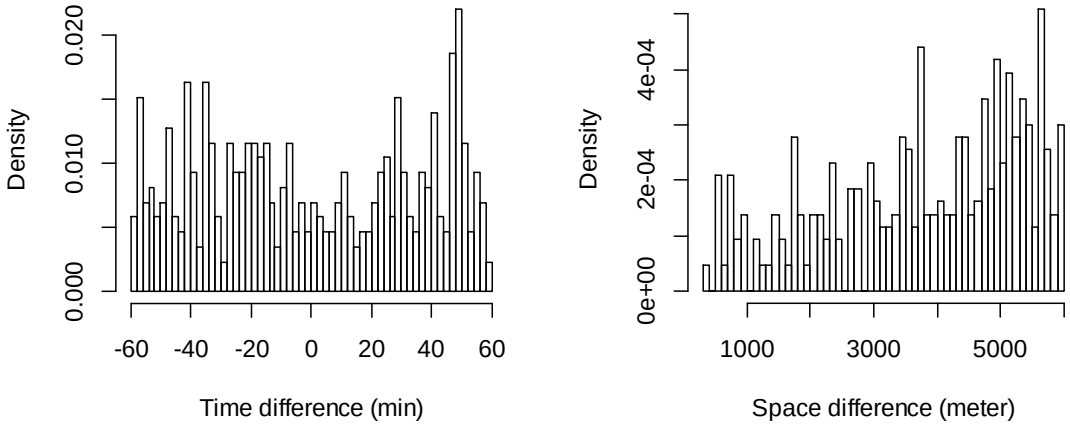
Even though the accident-related tweets in our study amounts to 422, it is still a very small number as compared to 52,496 accident records obtained from VDOT. It is worth mentioning that we only consider the geo-tagged tweets which take no more than 5% of all tweets posted online. One advantage of geo-tagged tweets to detect traffic accidents is that they can provide clear longitude and latitude information of where the tweets are posted. For those tweets without specific latitude and longitude messages, we can possibly infer their locations according to their tweet messages (Ikawa et al., 2012). However, it is obvious that not all non-geo-tagged tweets can provide enough location messages. Given the low coverage of geo-tagged tweets and the disadvantages of non-geo-tagged tweets, one can see that online tweets in NOVA are still unlikely to cover all the traffic accidents with high probability.

Thus, the tweets are more probable to be a viable supplement rather than a replacement to the existing detection method. This is mainly because the they are relatively small-scaled incidents (Schulz et al., 2013) and seldom arouse public attentions. The influence of them may not be as high as that of earthquake or festival parades, not all travelers are willing to leave a corresponding messages online. Also, when passing by the site of traffic accident, most of the drivers cannot tweet about it just for their own traffic safety.

3.6.2 Features of time and space differences between tweets and accidents

As the traffic-related information may not improve the prediction results, the traffic conditions around where the tweets are posted may not be significantly influenced. Possible explanations may be found by examining the time and space differences between the accident-related tweets and corresponding accidents. We compare accident-related tweets with the traffic management log from Virginia Department of Transportation (VDOT). Given an accident-related tweet, we extract the accident records from the log which is close to the tweet locations over a certain time window. The time window is the set 1 hour before and after the tweet time. The comparison results are insightful in studying the potentials of tweets in traffic accident detection.

According to our examination, of more than 400 labeled accident tweets, there are about 300 of them can be traced to an accident record by VDOT. It is also possible that one tweet correspond to more than one accident records and there is not additional information for us to specify the exact one. In this case, we choose the accident record that is the nearest to the tweet location. The time when the tweets are posted can be either earlier or later than the starting time of the traffic accident records as shown in Fig.3.12(a). Suppose the starting time in the traffic management log is the time when the police arrives the accident site, nearly one third of the accident-related tweets are posted earlier than the traffic accident. This coincides with the findings in (D'Andrea et al., 2015) that tweets detect traffic accidents more than 1 hour earlier than traditional media. If so, detecting the accident-related tweets online can sometimes significantly reduce the response time.



(a) (b)

Fig.3.12 (a) Time and (b) space difference between the accident-related tweets and the accident records by VDOT.

As shown in Fig.3.12(b), the shortcomings are also obvious that the space differences are sometimes too large and it is hard for traffic operators to pinpoint the accident site solely with the latitude and longitude information of geo-tagged tweets. This also explains the reasons why traffic-related information cannot improve the prediction accuracy because the travelers usually tweet where they are far away from the accident sites. This increase the difficulties in the real-time detection of traffic accidents and these shortcomings can possibly be overcome by hinting the tweet contents. It is worth mentioning that unlike (D'Andrea et al., 2015), our study only focus on the tweets with a clear expression of traffic accidents. The tweets with side proof of accidents such as traffic jam or delay is not included.

3.6.3 Features of unrecorded tweet accidents

The comparison between tweets and accident records also reveal that some accident-related tweets express explicit meanings about the traffic accidents but cannot be traced by VDOT accident records. After our examinations, more than one third of these tweets are from the media channels such as “wtop”, “wtoptraffic”, etc. The locations for these tweets may not provide useful locations for accident detection. Other tweets may possibly be accounted by several reasons.

First, compared with the traffic management log maintained by VDOT, it is entirely possible that the tweets can capture the unexpected small events happened in our daily life. These events may include those “mild” accidents that do not incur the attention of traffic police and thus may not be included in the management log. The consequences of these events such as the road lanes blocking or cars slowing down may not last long and the corresponding affairs may come with a proper handling. If so, the unrecorded tweets may act as a supplement of the current accident detection system. The tweets can be:

- “woooo got rear ended on i495 going to md great way to start a monday morning”
- “holy shit i just crashed my dads car”

Second, other reasons possibly exist: some of the accident-related tweets may be posted too far away from the accident site; some tweet users retweet about an accident instead of seeing in person; some tweeters may misjudge the situations and their inferences are from the jammed conditions of the roadways. In sum, some tweets may just be alarms that are not entirely true. For example:

- "sooo the car just said attention there is a car accident 12 miles ahead wtf kin of car does that"
- "major vehicle accident southbound i95 near lorton va traffic dmv"

After comparing with the ground truth, it can be concluded that the tweets labeled by our model can possibly identify the existence of potential traffic accidents. This identifications may be faster than the traditional methods. The locations of the traffic accidents may not be just exactly the latitude and longitude where the tweets are posted and the traffic operators should incorporate more information sources pinpoint the locations. In sum, it is entirely possible to increase the efficiency of traffic accident detection by monitoring the geo-tagged tweets.

3.7 Conclusions and discussions

In this chapter, we employ the SVMs to detect the traffic accident from tweets. The prediction results are compared with that of sLDA and generate three important features: single token, paired token and traffic-related data to achieve a more accurate and effective on-site traffic accident detection. Our findings can be summarized as follows:

First, we thoroughly investigate the tweet contents related to traffic accidents. We found token features: single tokens and paired tokens that may correlate with the traffic accident labels. Our results show that paired tokens can possibly capture the association rules inherent in the accident-related tweets and increase the accuracy of the traffic accident detection.

Second, we unveil the relationships between traffic flow and occupancy based on the fundamental diagram using large-scale data and point out that these relationships vary different locations. We employ the K-means clustering algorithm to cluster the detectors into different patterns of fundamental diagrams. The traffic flows over a certain range of occupancy in a given cluster are observed to follow a Gaussian distribution. The derived traffic-related information may provide limited improvement for accident prediction.

Third, the comparison between the prediction results and the traffic management log maintained by VDOT provides insights in the studying the accident-related tweets: First, sometimes the tweet reflection on the traffic accident is much faster than the traditional methods and detecting the accident-related tweets online can sometimes significantly reduce the response time. Second, tweets can sometimes capture those "mild" accidents that do not incur the attention of traffic police and this indicates possibility of tweets making up for the deficiencies of traffic management log. Third, some accident-related tweets, include those posted by traditional media, are more probable to be a post-event recall rather than an expressions of instantaneous feelings. These tweets cannot give an exact location of the accident site and precise location detection should involve more data sources.

Finally, it is concluded that integrating social media data into the traffic-related study opens up a wide range of possibilities for research in on-site traffic accident detection. The results show that social media data are very noisy and even unreliable, so solely relying on social media data is still not a perfect option.

Further studies can focus on the data fusion of different data sources to better realize the purposes of other research such as traffic jam detection, traffic emergency evacuation, etc. The spatial-temporal features of traffic data are also worth studying for regional traffic operations. Note that our tweet data and traffic data are labeled by both time and locations. It would be an interesting extension to detect traffic event with non-geotagged tweets.

Reference

Abdulhai, B., Ritchie, S.G., 1999. Enhancing the universality and transferability of freeway incident detection using a Bayesian-based neural network. *Transportation Research Part C: Emerging Technologies* 7(5), 261-280.

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules, *Proc. 20th int. conf. very large data bases, VLDB*, pp. 487-499.

Ahmed, M.S., Cook, A.R., 1979. *Analysis of freeway traffic time-series data by using Box-Jenkins techniques*.

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle, *Selected Papers of Hirotugu Akaike*. Springer, pp. 199-213.

Amin, M.S., Jalil, J., Reaz, M., 2012. Accident detection and reporting system using GPS, GPRS and GSM technology, *Informatics, Electronics & Vision (ICIEV), 2012 International Conference on*. IEEE, pp. 640-643.

Anderson, M.L., 2013. Subways, strikes, and slowdowns: The impacts of public transit on traffic congestion. National Bureau of Economic Research.

Aramaki, E., Maskawa, S., Morita, M., 2011. Twitter catches the flu: detecting influenza epidemics using Twitter, *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 1568-1576.

Arnott, R., Inci, E., 2006. An integrated model of downtown parking and traffic congestion. *Journal of Urban Economics* 60(3), 418-442.

Balduini, M., Della Valle, E., 2012. Tracking Movements and Attention of Crowds in Real Time Analysing Social Streams–The case of the Open Ceremony of London 2012. *Semantic Web Challenge at ISWC*.

Banaei-Kashani, F., Shahabi, C., Pan, B., 2011. Discovering patterns in traffic sensor data, *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on GeoStreaming*. ACM, pp. 10-16.

Bando, M., Hasebe, K., Nakayama, A., Shibata, A., Sugiyama, Y., 1995. Dynamical model of traffic congestion and numerical simulation. *Physical Review E* 51(2), 1035.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3, 993-1022.

Boyd-Graber, J., Resnik, P., 2010. Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 45-55.

Cervero, R., 1994. Rail transit and joint development: Land market impacts in Washington, DC and Atlanta. *Journal of the American Planning Association* 60(1), 83-94.

Chaniotakis, E., Antoniou, C., 2015. Use of Geotagged Social Media in Urban Settings: Empirical Evidence on Its Potential from Twitter, *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*. IEEE, pp. 214-219.

Chen, M.-C., Wei, Y., 2011. Exploring time variants for short-term passenger flow. *Journal of Transport Geography* 19(4), 488-498.

Chen, P.-T., Chen, F., Qian, Z., 2014. Road traffic congestion monitoring in social media with hinge-loss Markov random fields, *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, pp. 80-89.

Cohen, J., Cohen, P., West, S.G., Aiken, L.S., 2013. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Coifman, B., Beymer, D., McLauchlan, P., Malik, J., 1998. A real-time computer vision system for vehicle tracking and traffic surveillance. *Transportation Research Part C: Emerging Technologies* 6(4), 271-288.

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), 2.

Cordeiro, M., 2012a. Twitter event detection: Combining wavelet analysis and topic inference summarization, *Doctoral Symposium on Informatics Engineering, DSIE*, pp. 11-16.

Cordeiro, M., 2012b. Twitter event detection: Combining wavelet analysis and topic inference summarization, *Doctoral Symposium on Informatics Engineering, DSIE*.

Cortes, C., Vapnik, V., 1995. Support-vector networks. *Machine learning* 20(3), 273-297.

Cramér, H., 1999. *Mathematical methods of statistics*. Princeton university press.

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.

Daly, E.M., Lecue, F., Bicer, V., 2013. Westland row why so slow?: fusing social media and linked data sources for understanding real-time traffic conditions, *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, pp. 203-212.
Duranton, G., Turner, M.A., 2011. The fundamental law of road congestion: Evidence from US cities. *The American Economic Review*, 2616-2652.
Freedman, D.A., 2009. *Statistical models: theory and practice*. cambridge university press.
Gal-Tzur, A., Grant-Muller, S.M., Kuflik, T., Minkov, E., Nocera, S., Shoor, I., 2014. The potential of social media in delivering transport policy goals. *Transport Policy* 32, 115-123.
Geisser, S., 1993. *Predictive inference*. CRC Press.
Giridhar, P., Amin, M.T., Abdelzaher, T., Kaplan, L.M., George, J., Ganti, R., 2014. Clarisense: Clarifying sensor anomalies using social network feeds, *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*. IEEE, pp. 395-400.
Gong, M., Fei, X., Wang, Z., Qiu, Y., 2014. Sequential framework for short-term passenger flow prediction at bus stop. *Transportation Research Record: Journal of the Transportation Research Board*(2417), 58-66.
Gong, W., 2010. ARMA-GRNN for passenger demand forecasting, *Natural Computation (ICNC), 2010 Sixth International Conference on*. IEEE, pp. 1577-1581.
Gu, Y., Qian, Z.S., Chen, F., 2016. From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies* 67, 321-342.
Guo, F., Krishnan, R., Polak, J., 2013. A computationally efficient two-stage method for short-term traffic prediction on urban roads. *Transportation planning and technology* 36(1), 62-75.
Hahsler, M., Grün, B., Hornik, K., 2007. Introduction to arules–mining association rules and frequent item sets. *SIGKDD Explor*.
Hall, D.L., Llinas, J., 1997. An introduction to multisensor data fusion. *Proceedings of the IEEE* 85(1), 6-23.
Hasan, S., Schneider, C.M., Ukkusuri, S.V., González, M.C., 2013. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics* 151(1-2), 304-318.
He, J., Shen, W., Divakaruni, P., Wynter, L., Lawrence, R., 2013. Improving Traffic Prediction with Tweet Semantics, *IJCAI*.
Hobeika, A.G., Kim, C.K., 1994. Traffic-flow-prediction systems based on upstream traffic, *Vehicle Navigation and Information Systems Conference, 1994. Proceedings., 1994*. IEEE, pp. 345-350.
Ikawa, Y., Enoki, M., Tatsubori, M., 2012. Location inference using microblog messages, *Proceedings of the 21st international conference companion on World Wide Web*. ACM, pp. 687-690.
Jiang, X., Zhang, L., Chen, X.M., 2014. Short-term forecasting of high-speed rail demand: A hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in China. *Transportation Research Part C: Emerging Technologies* 44, 110-127.
Jin, J., Ran, B., 2009. Automatic freeway incident detection based on fundamental diagrams of traffic flow. *Transportation Research Record: Journal of the Transportation Research Board*(2099), 65-75.
Karatzoglou, A., Meyer, D., Hornik, K., 2005. Support vector machines in R.
Kell, J.H., Fullerton, I.J., Mills, M.K., 1990. Traffic detector handbook.
Khan, S.I., Ritchie, S.G., 1998. Statistical and neural classifiers to detect traffic operational problems on urban arterials. *Transportation Research Part C: Emerging Technologies* 6(5), 291-314.
Kumar, A., Jiang, M., Fang, Y., 2014. Where not to go?: detecting road hazards using twitter, *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, pp. 1223-1226.
Lan, L.W., Sheu, J.-B., Huang, Y.-S., 2008. Investigation of temporal freeway traffic patterns in reconstructed state spaces. *Transportation Research Part C: Emerging Technologies* 16(1), 116-136.
Leduc, G., 2008. Road traffic data: Collection methods and applications. *Working Papers on Energy, Transport and Climate Change* 1, 55.
Lee, S., Fambro, D., 1999. Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. *Transportation Research Record: Journal of the Transportation Research Board*(1678), 179-188.
Leng, B., Zeng, J., Xiong, Z., Lv, W., Wan, Y., 2013. Probability tree based passenger flow prediction and its application to the Beijing subway system. *Frontiers of Computer Science* 7(2), 195-203.
Li, R., Lei, K.H., Khadiwala, R., Chang, K.C.-C., 2012. Tedas: A twitter-based event detection and analysis system, *Data engineering (icde), 2012 ieee 28th international conference on*. IEEE, pp. 1273-1276.
Lin, C., He, Y., Everson, R., Rüger, S., 2012. Weakly supervised joint sentiment-topic detection from text. *Knowledge and Data Engineering, IEEE Transactions on* 24(6), 1134-1145.

Lin, L., Ni, M., He, Q., Gao, J., Sadek, A.W., Director, T.I.T.I., 2015. Modeling the Impacts of Inclement Weather on Freeway Traffic Speed: An Exploratory Study Utilizing Social Media Data, *Transportation Research Board 94th Annual Meeting*.

Long, R., Wang, H., Chen, Y., Jin, O., Yu, Y., 2011. Towards effective event detection, tracking and summarization on microblog data, *Web-Age Information Management*. Springer, pp. 652-663.

Mai, E., Hranac, R., 2013. Twitter interactions as a data source for transportation incidents, *Proc. Transportation Research Board 92nd Ann. Meeting*.

Mcauliffe, J.D., Blei, D.M., 2008. Supervised topic models, *Advances in neural information processing systems*, pp. 121-128.

Mochihashi, D., 2009. LDA, a latent dirichlet allocation package. *ATR Spoken Language Communication Research Laboratories*.

Münz, G., Li, S., Carle, G., 2007. Traffic anomaly detection using k-means clustering, *GI/ITG Workshop MMBnet*.

NHTSA, N.H.T.S.A., 2015. 2013 Traffic Safety Facts FARS/GES Annual Report.

Ni, M., He, Q., Gao, J., 2014. Using social media to predict traffic flow under special event conditions, *The 93rd Annual Meeting of Transportation Research Board*.

Oh, C., Oh, J.-S., Ritchie, S., Chang, M., 2001. Real-time estimation of freeway accident likelihood, *80th Annual Meeting of the Transportation Research Board, Washington, DC*.

Park, H., Haghani, A., 2015. Real-time prediction of secondary incident occurrences using vehicle probe data. *Transportation Research Part C: Emerging Technologies*.

Payne, H.J., Tignor, S.C., 1978. Freeway incident-detection algorithms based on decision trees with states. *Transportation Research Record*(682).

Pereira, F.C., Rodrigues, F., Ben-Akiva, M., 2015a. Using data from the web to predict public transport arrivals under special events scenarios. *Journal of Intelligent Transportation Systems* 19(3), 273-288.

Pereira, F.C., Rodrigues, F., Polisciuc, E., Ben-Akiva, M., 2015b. Why so many people? Explaining Nonhabitual Transport Overcrowding With Internet Data. *Intelligent Transportation Systems, IEEE Transactions on* 16(3), 1370-1379.

Phuvipadawat, S., Murata, T., 2010. Breaking news detection and tracking in Twitter, *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*. IEEE, pp. 120-123.

Popescu, A.-M., Pennacchiotti, M., 2010. Detecting controversial events from twitter, *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, pp. 1873-1876.

Porter, M.F., 1980. An algorithm for suffix stripping. *Program* 14(3), 130-137.

Protalinski, E., 2012. More Twitter users chose to tweet from a mobile device rather than a PC in 2012, study says. The Next Web.

Purohit, H., Hampton, A., Bhatt, S., Shalin, V.L., Sheth, A., Flach, J., 2013. An Information Filtering and Management Model for Twitter Traffic to Assist Crises Response Coordination. *Special Issue on Crisis Informatics and Collaboration*.

Rajaraman, A., Ullman, J.D., Ullman, J.D., Ullman, J.D., 2012. *Mining of massive datasets*. Cambridge University Press Cambridge.

Ramage, D., Dumais, S.T., Liebling, D.J., 2010. Characterizing Microblogs with Topic Models. *ICWSM* 10, 1-1.

Ranks-NL, 2015. Default English stopwords list.

Rasiwasia, N., Vasconcelos, N., 2013. Latent dirichlet allocation models for image classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35(11), 2665-2679.

Ryaben'kii, V.S., Tsynkov, S.V., 2006. *A theoretical introduction to numerical analysis*. CRC Press.

Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake shakes Twitter users: real-time event detection by social sensors, *Proceedings of the 19th international conference on World wide web*. ACM, pp. 851-860.

Samant, A., Adeli, H., 2000. Feature extraction for traffic incident detection using wavelet transform and linear discriminant analysis. *Computer-Aided Civil and Infrastructure Engineering* 15(4), 241-250.

Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J., 2009. Twitterstand: news in tweets, *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*. ACM, pp. 42-51.

Schoenhof, M., Helbing, D., 2007. Empirical features of congested traffic states and their implications for traffic modeling. *Transportation Science* 41(2), 135-166.

Schulz, A., Ristoski, P., Paulheim, H., 2013. I see a car crash: Real-time detection of small scale incidents in microblogs, *The Semantic Web: ESWC 2013 Satellite Events*. Springer, pp. 22-33.

Schwarz, A., 2012. How publics use social media to respond to blame games in crisis communication: The Love Parade tragedy in Duisburg 2010. *Public Relations Review* 38(3), 430-437.
Sethi, V., Bhandari, N., Koppelman, F.S., Schofer, J.L., 1995. Arterial incident detection using fixed detector and probe vehicle data. *Transportation Research Part C: Emerging Technologies* 3(2), 99-112.
Shirky, C., 2011. The political power of social media. *Foreign affairs* 90(1), 28-41.
Sun, Y., Leng, B., Guan, W., 2015. A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system. *Neurocomputing* 166, 109-121.
Sun, Y., Zhang, G., Yin, H., 2014. Passenger Flow Prediction of Subway Transfer Stations Based on Nonparametric Regression Model. *Discrete Dynamics in Nature and Society* 2014.
Sweet, M., 2014. Traffic congestion's economic impacts: evidence from US metropolitan regions. *Urban Studies* 51(10), 2088-2110.
Tan, M.-C., Wong, S.C., Xu, J.-M., Guan, Z.-R., Zhang, P., 2009. An Aggregation Approach to Short-Term Traffic Flow Prediction. *Ieee Transactions on Intelligent Transportation Systems* 10(1), 60-69.
Teng, H., Qi, Y., 2003. Application of wavelet technique to freeway incident detection. *Transportation Research Part C: Emerging Technologies* 11(3), 289-308.
Teodorovic, D., Lucic, P., Popovic, J., Kikuchi, S., Stanic, B., 2001. Intelligent isolated intersection, *Fuzzy Systems, 2001. The 10th IEEE International Conference on*. IEEE, pp. 276-279.
Tsai, J., Case, E., 1979. Development of freeway incident-detection algorithms by using pattern-recognition techniques. *Transportation Research Record* 722, 113-116.
Tsai, T.-H., Lee, C.-K., Wei, C.-H., 2009. Neural network based temporal feature models for short-term railway passenger demand forecasting. *Expert Systems with Applications* 36(2), 3728-3736.
Vlahogianni, E.I., Golias, J.C., Karlaftis, M.G., 2004. Short-term traffic forecasting: Overview of objectives and methods. *Transport reviews* 24(5), 533-557.
Wanichayapong, N., Pruthipunyaskul, W., Pattara-Atikom, W., Chaovalit, P., 2011. Social-based traffic information extraction and classification, *ITS Telecommunications (ITST), 2011 11th International Conference on*. IEEE, pp. 107-112.
Weerkamp, W., de Rijke, M., 2012. Credibility-inspired ranking for blog post retrieval. *Information Retrieval* 15(3-4), 243-277.
Wei, Y., Chen, M.-C., 2012. Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C: Emerging Technologies* 21(1), 148-162.
White, J., Thompson, C., Turner, H., Dougherty, B., Schmidt, D.C., 2011. Wreckwatch: Automatic traffic accident detection and notification with smartphones. *Mobile Networks and Applications* 16(3), 285-303.
Williams, B., 2001. Multivariate vehicular traffic flow prediction: evaluation of ARIMAX modeling. *Transportation Research Record: Journal of the Transportation Research Board*(1776), 194-200.
Williams, B., Durvasula, P., Brown, D., 1998. Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models. *Transportation Research Record: Journal of the Transportation Research Board*(1644), 132-141.
Williams, B.M., Hoel, L.A., 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of transportation engineering* 129(6), 664-672.
Willsky, A.S., Chow, E.Y., Gershwin, S., Greene, C.S., Houpt, P.K., Kurkjian, A.L., 1980. Dynamic model-based techniques for the detection of incidents on freeways. *Automatic Control, IEEE Transactions on* 25(3), 347-360.
Wu, C.-H., Ho, J.-M., Lee, D.-T., 2004. Travel-time prediction with support vector regression. *Intelligent Transportation Systems, IEEE Transactions on* 5(4), 276-281.
Xu, G., Yang, S.-H., Li, H., 2009. Named entity mining from click-through data using weakly supervised latent dirichlet allocation, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1365-1374.
Yasdi, R., 1999. Prediction of road traffic using a neural network approach. *Neural computing & applications* 8(2), 135-142.
Yuan, F., Cheu, R.L., 2003. Incident detection using support vector machines. *Transportation Research Part C: Emerging Technologies* 11(3), 309-328.
ZHANG, C.-h., SONG, R., SUN, Y., 2011. Kalman filter-based short-term passenger flow forecasting on bus stop. *Journal of transportation systems engineering and information technology* 4, 025.
Zhang, K., Taylor, M.A., 2006. Effective arterial road incident detection: a Bayesian network based algorithm. *Transportation Research Part C: Emerging Technologies* 14(6), 403-417.

Zhang, X., Rice, J.A., 2003. Short-term travel time prediction. *Transportation Research Part C: Emerging Technologies* 11(3), 187-210.
Zhang, Z., Ni, M., He, Q., Gao, J., Gou, J., Li, X., 2016. An Exploratory Study on the Correlation between Twitter Concentration and Traffic Surge 2. *To appear in Transportation Research Record* 35, 36.