ARTICLE IN PRESS

Transportation Research Part C xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways

Qi Shi *, Mohamed Abdel-Aty [1]

Department of Civil, Environmental and Construction Engineering, University of Central Florida, Engineering II-215, Orlando, FL 32816, United States

ARTICLE INFO

Article history:
Received 28 May 2014
Received in revised form 17 January 2015
Accepted 23 February 2015
Available online xxxx

Keywords:
Big Data
Real-time
Congestion
Safety
Urban expressway

ABSTRACT

The advent of Big Data era has transformed the outlook of numerous fields in science and engineering. The transportation arena also has great expectations of taking the advantage of Big Data enabled by the popularization of Intelligent Transportation Systems (ITS). In this study, the viability of a proactive real-time traffic monitoring strategy evaluating operation and safety simultaneously was explored. The objective is to improve the system performance of urban expressways by reducing congestion and crash risk. In particular, Microwave Vehicle Detection System (MVDS) deployed on an expressway network in Orlando was utilized to achieve the objectives. The system consisting of 275 detectors covers 75 miles of the expressway network, with average spacing less than 1 mile. Comprehensive traffic flow parameters per lane are continuously archived on one-minute interval basis. The scale of the network, dense deployment of detection system, richness of information and continuous collection turn MVDS as the ideal source of Big Data. It was found that congestion on urban expressways was highly localized and time-specific. As expected, the morning and evening peak hours were the most congested time periods. The results of congestion evaluation encouraged real-time safety analysis to unveil the effects of traffic dynamics on crash occurrence. Data mining (random forest) and Bayesian inference techniques were implemented in real-time crash prediction models. The identified effects, both indirect (peak hour, higher volume and lower speed upstream of crash locations) and direct (higher congestion index downstream to crash locations) congestion indicators confirmed the significant impact of congestion on rear-end crash likelihood. As a response, reliability analysis was introduced to determine the appropriate time to trigger safety warnings according to the congestion intensity. Findings of this paper demonstrate the importance to jointly monitor and improve traffic operation and safety. The Big Data generated by the ITS systems is worth further exploration to bring all their full potential for more proactive traffic management.

1. Introduction

In an age of data explosion, almost every aspect of social activities is impacted by the abundance of information. The information, characterized by alarming volume, velocity and variety, is often referred to as "Big Data" (Beyer and Laney, 2012). As one fundamental elements of human life, transportation also confronts the promises and challenges brought about

* Corresponding author. Tel.: +1 (407) 823 0300.
E-mail addresses: shiqi@knights.ucf.edu (Q. Shi), m.aty@ucf.edu (M. Abdel-Aty).
[1] Tel.: +1 (407) 823 1374; fax: +1 (407) 823 3315.

http://dx.doi.org/10.1016/j.trc.2015.02.022
0968-090X/

by the Big Data era. Big Data in transportation arena, enabled by the rapid popularization of Intelligent Transportation System (ITS) in the past few decades, is often collected continuously from different sources over vast geographic scale. Huge in size and rich in information, the seemingly disorganized data could considerably enhance experts' understanding of their system. In addition, proactive traffic management for better system performance is made possible due to the real-time nature of the Big Data in transportation.

Operation efficiency and traffic safety have long been deemed as priorities among highway system performance measurement. While efficiency could be evaluated in terms of traffic congestion, safety is studied through crash analysis. Extensive works have been conducted to identify contributing factors and remedies of traffic congestion and crashes respectively. These studies lead to gathering consensus that operation and safety have played as two sides of a coin, ameliorating either would have a positive effect on the other. With the advancement of Big Data, monitoring and improvement of both operation and safety proactively in real-time have become an urgent call.

This study focuses on Central Florida Expressway Authority (CFX)'s system in Central Florida area. The system of interest consists of three expressways that are located in the densely populated urban area. The toll expressways communicate downtown area, airport and other attraction areas in Orlando, serving both commuters and tourists. Multiple ITS systems are equipped on the system for electronic toll collection and traveler information. In 2013, the authority introduced Microwave Vehicle Detection System (MVDS) to monitor traffic conditions across different sections of the expressways. A total of 275 MVDS detectors are densely allocated along the 75-mile expressway network, with average spacing less than 1 mile. Comprehensive traffic flow parameters are archived on one-minute interval basis without interruption. As a result, the large geographic scale of deployment and continuous data collection provide a full view of network performance and serve as the source of Big Data. In this paper, real-time operation and safety analyses based on MVDS data are carried out in the hope to shed some light on the Big Data applications in real-time traffic operation and safety monitoring and improvement.

2. Background

Effective strategies to improve traffic operation and safety simultaneously require profound understanding about their features and relationship. In the age of information, these objectives could be efficiently realized through Big Data applications. Traffic congestion can be viewed as a product of the interaction between demand and capacity. Periodic high demand at specific bottlenecks during peak hours can result in recurrent congestion while incidents, especially crashes, reducing roadway capacity temporarily lead to non-recurrent congestion. To catch this dynamic process, Big Data generated from the ITS detection system could be leveraged to develop congestion measurement in real-time. In the meantime, crash occurrence is often regarded as random events affected by human behavior, roadway design, traffic flow and weather conditions. Big Data applications also introduce new perspectives in safety analysis. Thanks to the advantages brought by Big Data, researchers are able to restore the traffic condition for each crash case and draw general conclusions using individual crash data. As a result, Big Data applications in the current work will focus on developing congestion measurement and uncovering the relationship between safety and congestion, both in real-time.

2.1. Big Data in transportation arena

Big Data in the transportation arena comes not only from a single source but many. Currently, the most widely used data sources are traffic surveillance systems. According to Martin et al. (2003), the state-of-art detection technologies fall into three categories: in-roadway detectors, over-roadway detectors and off-roadway technologies. One of the most representative and widely used in-roadway detectors in real-time crash analysis is the inductive loop detector. The loop detectors have been implemented since the early stage of automatic traffic surveillance thus they are applications of a relatively mature technology. However, they have several drawbacks such as disruption of traffic for installation and repair, and high failure rates in certain conditions (Martin et al., 2003) such as poor road surface conditions and adverse weather. An over-roadway sensor is one that is mounted above the roadway itself or alongside the roadway, offset from the nearest traffic lane by some distance. Existing over-roadway sensors range from video image processors to more up-to-date microwave radar sensors. Compared with the in-roadway sensors, the over-roadway sensors have the significant advantage that they minimize the disruption of traffic during installation and maintenance. Probe vehicle is off-roadway detection technology that is developing fast. Compared with the other two types of sensors mentioned above, probe vehicles also require in-vehicle devices in addition to fixed infrastructure. Current probe vehicle technologies include Global Positioning System (GPS), cellular phones, Bluetooth, Ground-Based Radio Navigation, Automatic Vehicle Identification (AVI) and Automatic Vehicle Location (AVL) (Turner et al., 1998; Martin et al., 2003). With sufficient probe vehicles, they could also provide real-time traffic information at individual vehicle level. Nevertheless, since only part of the vehicles are equipped with in-vehicle devices, some traffic indicators might not be complete or accurate.

Other data sources such as demographic data, weather reporting system, geometric characteristics, and crash data are also extensively used in traffic operation, safety management and research. In recent decades with mobile devices, social media data also become a promising data source. To make the most of these data, efficient data integration and fusion have to be carried out. In real-time traffic safety analysis for example, according to crash locations and times, real-time traffic,

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

weather and geometric data from nearby detection locations, weather stations and geometric information database could all be integrated together for analysis. In the future with more data available, the data integration would play an important role in Big Data applications in the transportation arena.

The benefits of Big Data technologies include direct and indirect applications. Direct applications could be congestion reduction, incident prediction, and travel time estimation. Indirect applications are carried out through enhancement of traffic modeling in the model development, calibration and validation processes. Traffic simulation could also be greatly improved based on the real data collected from the field.

2.2. Congestion measurement

Traditionally, volume-to-capacity (V/C) ratios and level of service (LOS) are implemented by transportation authorities as indicators of congestion intensity (Grant et al., 2011). Nevertheless, traffic demand can vary substantially in both temporal and spatial dimensions. Roadway capacity can also be reduced by incidents as discussed above. In such cases, V/C ratios and LOS lack the capability to capture the variability of congestion. Big Data from ITS facilities, on the other hand, provides professionals with much more detailed insights of congestion; they can reflect the overall performance of the whole system, zoom into specific locations or time periods, and observe the changes of congestion intensity in time and space. Above all, they can monitor congestion in real-time. By doing so, quick, precise and effective response is made possible.

Real-time congestion measurement often defines congestion based on travel time or speed. The selection of a specific congestion measure depends on the available ITS detection systems on the managed roadways. Many agencies have introduced the Automatic Vehicle Identification (AVI) systems to keep track of vehicles at different AVI locations and calculate the travel time. Travel Time Index (TTI) is widely used to measure the extra time taken during peak hours compared with that of non-peak hours (Schrank et al., 2012). Additional time-based congestion measures include hours of congestion, Planning Time Index (PTI) and delay, etc. Other prevalent ITS detection systems such as loop and radar detector systems record the spot speed information. Speed-based congestion measures are developed for these systems. Washington State DOT (WSDOT) defines congestion based on the ratio between real-time speed detected by loop detectors and the posted speed limit (Hammond, 2012). Dias et al. (2009) proposed congestion index (CI) in their study defined as

$$\begin{aligned}\text{Congestion index (CI)} &= \frac{\text{free flow speed} - \text{actual speed}}{\text{free flow speed}} \quad \text{when CI} > 0;\\ &= 0 \text{ when CI} \leqslant 0\end{aligned} \tag{1}$$

Compared with choosing a fixed speed as congestion threshold, CI is a more flexible and consistent term to reflect the congestion intensity since posted speed limit can vary across roadway sections. The system of interest in this study deploys numerous MVDS detectors along the expressways to continuously monitor the traffic conditions at their installed locations. Consequently, congestion index in Eq (1) is adopted as the congestion measure for the MVDS traffic data.

2.3. Real-time crash prediction

Existing traffic safety analysis can be broken down into two broad categories: aggregate analysis targeting the crash frequency and disaggregate analysis studying each crash case. The effects of congestion on safety have been evaluated in both modeling frameworks. Crash frequency analyses are more traditional and feasible before the age of Big Data since they use aggregated information and require less data collection effort. Baruya (1998) in his study found crash frequency was negatively associated with speed which is a result of congestion. Wang et al. (2009) developed time-based congestion index and stated that congestion had no impact on the crash frequency based on their spatial analysis. The effects of congestion in the above analyses were inconclusive. This could partly be interpreted as congestion could be location and time specific: averaging congestion intensity over time or space would counteract the true effect of congestion on crash occurrence.

The fast development of ITS detection system, however, boosts new methodologies to cope with the above issues. As Big Data is becoming more available, it is possible to retrieve real-time traffic information for each crash case. The benefits are self-evident; by evaluating traffic safety in real-time, the information to be used is more accurate, more relevant. Real-time crash prediction belongs to disaggregate analysis which classifies crash and non-crash based on real-time traffic information prior to crashes. Two general approaches exist in real-time safety evaluation, statistical methods and data mining methods. Data process for these approaches is relatively more complicated, requiring detailed information during short time interval (e.g., 5/10 minutes) which is normally unavailable in aggregate analysis. Statistical methods such as simple/matched-case logistic regression (Abdel-Aty et al., 2004) and Bayesian statistics (Abdel-Aty et al., 2012) present the effects of candidate variables in a more interpretable way. Data mining based methods could be neural networks (Pande and Abdel-Aty, 2006), random forests (Abdel-Aty et al., 2008), and support vector machines (Yu and Abdel-Aty, 2013), etc. Data mining methods are applauded for their high prediction accuracy but criticized for the black-box-like process. Nowadays, many researchers begin taking advantages of both approaches by utilizing data mining technique for variable selection and statistical technique for an interpretable variable effects evaluation.

Currently real-time crash analysis emphasizing on congestion has not been adequately explored. Some existing research include Christoforou et al. (2011) and Hossain and Muromachi (2012). Nevertheless, considering the analytical methods

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

enabled by Big Data, the interrelation of congestion and traffic safety, and the vision to improve them together, this topic worth thorough investigation. In this study, we adopt real-time safety evaluation incorporating congestion. In the hope to propose an integrated improvement strategy for congestion and safety, we further introduce the First Order Reliability Method (FORM) widely used in structural reliability analysis to determine when it is appropriate to trigger the warning to the expressway system.

2.4. Rear-end crash

It has been widely accepted that the same traffic state could impose distinct impact on traffic safety regarding to crash type and severity. Congestion's effect on traffic safety, to be properly analyzed, should also follow the same reasoning. So far, the relationship between congestion and crash severity gained relatively more research attention (Shefer and Rietveld, 1997; Quddus et al., 2009; Wang et al., 2009). Crash types under congestion, on the other hand, haven't been adequately addressed.

Common sense informs us that on freeways/expressways single vehicle crashes are more likely to occur with driving errors under free flow condition; sideswipe crashes might be caused by inappropriate lane-change behaviors and speed variation between lanes; and rear-end crashes often involve multiple vehicles with the leading vehicle suddenly decelerating. Under congestion, vehicles approaching the queue end have to slow down in advance otherwise the likelihood of rear-end crashes could be increased considerably. Lee et al. (2006) stated that rear-end crashes took a large proportion of total freeway crashes. Abdel-Aty et al. (2007) further confirmed that rear-end crashes are highly related with congestion, particularly if speed variation and average occupancy are elevated. Golob and Recker's (2003) research conclusions were in line with the above study, associating rear-end collisions with high variations in relatively low speed. Christoforou et al. (2011) implemented multivariate probit model and indicated that the rear-end crashes were more probable under congestion while sideswipes more probable under "intermediate" density traffic regimes. In all, to achieve a more conclusive statement about the relationship between crash occurrence and crash mitigation under congestion, the crash types should be taken into consideration.

By synthesizing the discussion above, this study underlines the Big Data applications in congestion and safety analyses and improvement. Speed-based congestion index is applied to continuously monitor the traffic congestion in spatial–temporal dimensions. The relationship between congestion and rear-end crashes is explored within real-time modeling framework and FORM analysis. Based on the analysis results, strategies established on MVDS traffic data are proposed to improve operation and safety in real-time on the urban expressways.

3. Data preparation

The expressway system of interest in this study that consists of SR 408, SR 417 and SR 528 is partly operated by CFX as shown in Fig. 1. SR 408 travels through downtown Orlando and accommodate more commuter traffic. SR 417 links southern and northern Orange County. SR 528 connects the Orlando international airport and the coast resort area. SR 408 and SR 528 are both connected with SR 417. On the 75-mile system, a total of 275 MVDS detectors have been installed. Table 1 shows the current deployment of the MVDS system on each expressway. From the table, it can be seen that the system is well covered with the average distance between adjacent detectors less than 1 mile.

The MVDS system could monitor traffic flow conditions at the lane level on roadway segments, ramps, toll plaza cash lanes and express lanes. In total, the three expressways generate about 1.5 million number of traffic readings each day. Hence from a volume point of view, the amount of data archived is huge and still grows rapidly. Currently, the data are collected continuously at a one-minute interval. According to the need of traffic authority, they could be archived at even shorter time interval such as 20 or 30 s. Currently on the examined system, the authority has implemented real-time traffic data to provide drivers with travel time estimation. From a velocity point of view, the speed of data collection enables traffic operators to manage their system and react in real-time. The collected data are in structured numeric format and contain detailed traffic parameters. The MVDS system could detect spot speed, volume, occupancy and simple vehicle classification on each lane. Based on these variables, other indicators such as average and standard deviation of them could also be calculated. In addition to the basic traffic flow variables, several indexes have been calculated to represent certain traffic flow characteristics. For example, Oh et al. (2006) calculated RCRI (rear-end crash collision risk index) with 5-minute level data. Lee et al. (2006) proposed overall average flow ratio (OAFR) to represent the total number of lane changes in all lanes. Dias et al. (2009) introduced congestion index to represent the congestion conditions on the roadway. In traffic safety analysis, these ITS data are typically merged or combined with roadway geometric characteristics, weather conditions and driver information depending on their availability. Thus from a variety point of view, rich information could be extracted from MVDS systems and the need of data merging and combination added the complexity in data processing. Each of the three standards (volume, velocity and variety) indicates that the MVDS traffic data should be treated and exploited as Big Data. The MVDS data have been collected since July 2013. At the time of this research, eight-month traffic data have been collected.

During the studied time period, a total of 581 crashes occurred on the three expressways of which 243 are rear-end crashes (Table 2). SR 408 has both the highest crash count and rear-end crash rate among the three expressways. This phenomenon is credited to the large traffic volume and commuter traffic on SR 408 compared with the other two expressways.

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

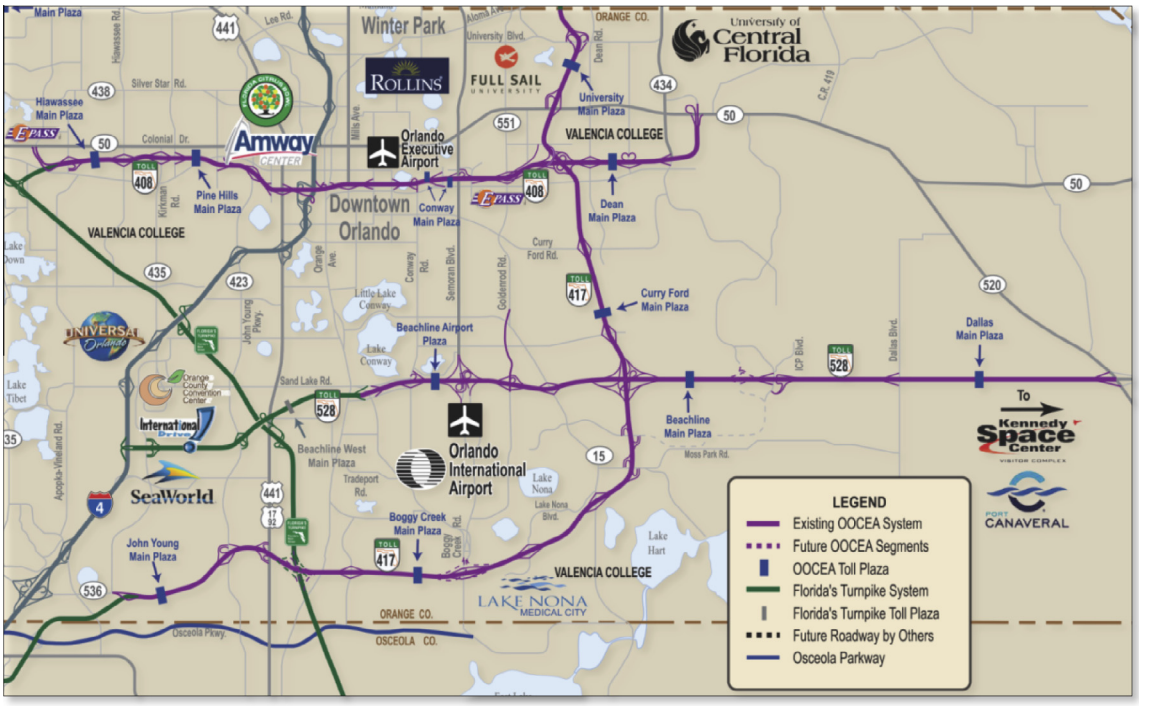

Fig. 1. Expressway system operated by CFX (CFX, 2014).

Table 1
MVDS deployment on CFX system.

Route	Length (mi)	Direction	Mainline detectors	Distance between adjacent detectors			
				Mean	Std Dev	Min	Max
SR 408	21.4	EB	55	0.38	0.18	0.10	1.00
		WB	55	0.39	0.18	0.10	1.00
SR 417	31.5	NB	55	0.58	0.28	0.20	1.30
		SB	55	0.58	0.28	0.20	1.20
SR 528	22.4	EB	26	0.84	0.79	0.10	3.00
		WB	29	0.84	0.82	0.10	3.10

Table 2
Crash occurrence on expressways.

Expressway	Rear-end crashes		
	No	Yes	Total
SR 408	133(45.86%)	157(54.14%)	290
SR 417	106(76.26%)	33(23.74%)	139
SR 528	99(65.13%)	53(34.87%)	152
Total	338(58.18%)	243(41.82%)	581

For each crash, traffic data of 5–10 minutes prior to the crash from two upstream and two downstream MVDS detectors (Fig. 2) closest to the crash location are collected and aggregated.

For the 243 rear-end crashes, 962 non-crash cases are matched. In previous matched case-control studies, control cases were selected for the same location at the same time and same weekday but in different weeks (Abdel-Aty et al., 2004; Ahmed and Abdel-Aty, 2012; Xu et al., 2012). The motivation behind this design is to control the variability caused by the time of day, season and roadway geometric characteristics. On the urban expressways, congestion tends to be recurrent especially when large commuter traffic is expected. Therefore following the traditional procedure by extracting traffic parameters at the same time as a crash case for non-crash cases might conceal the true effect of congestion on safety which is exactly what we are interested in. Taking this into account, we matched non-crash cases by extracting traffic conditions at

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

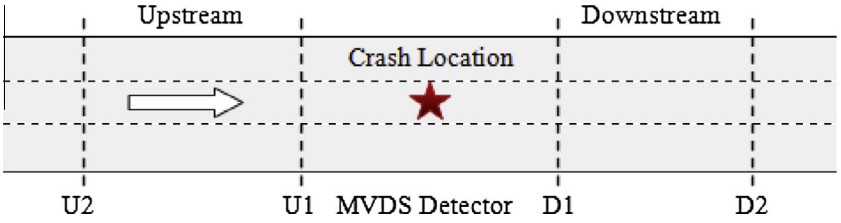

Fig. 2. Crash location and MVDS detector assignment.

the same location three hours and six hours before and after crash during which time no crash was observed, making the ratio between study and control group approximately 1:4. The total 1205 observations are then divided into training (70%) and validation (30%) data sets. When making the data partition, we made sure that the crash case and its matched non-crash cases are assigned to the same data set. For each MVDS detector selected (xx: u1, u2, d1, d2), logarithm of volume (log_xx_vol), truck percentage (xx_trkpct), average, standard deviation and logarithm of coefficient of variation of the speed (xx_avgspd, xx_stdspd, and log_xx_cvspd), speed difference between inner and outer lanes (xx_spddiff), number of lanes at the detection location (xx_lanes) and congestion index (xx_ci) were extracted.

Besides traffic parameters, the spatial–temporal related properties of crashes were also prepared. Whether the crash occurred during peak hours (peak) defined as 7:00–9:00 am and 17:00–19:00 pm were set up as binary variable. The posted speed limit (maxspeed), the horizontal curvature and existence of auxiliary lanes near ramps at the crash locations were all incorporated from FDOT Roadway Characteristics Inventory (RCI) database.

4. Methodology

4.1. Real-time congestion monitoring

The speed-based congestion index is adopted in this study to measure the congestion intensity on both spatial and temporal scales. The free flow speed is the 85th percentile speed at the detection location. According to Eq. (1), CI is a continuous variable with the range between 0 and 1. The increase in CI value indicates higher congestion level. The CI has been aggregated into 5-minute intervals at each detection location. Filled contour plots are created to visualize the congestion distribution.

4.2. Random forest

Random forest is an ensemble classifier using many decision tree models to vote for the most popular class (Breiman, 2001). A single decision tree suffers from high variance or bias. In contrast, random forest offers unbiased estimates of the classification error as trees are added to the forest. Also, strong law of large numbers guarantees random forest is robust against over-fitting.

One of the basic practices of random forest in real-time traffic safety evaluation is to estimate variable importance (Abdel-Aty and Haleem, 2011; Yu and Abdel-Aty, 2014). The random forest algorithm estimates the importance of a variable by looking at how much prediction error increases (or decrease in accuracy) when OOB (out-of-bag) data for that single variable is permuted (Liaw and Wiener, 2002). Another measure is the total decrease in node impurities denoted by Gini coefficient from splitting on the variable, averaged over all trees (Breiman, 2006). The former measure has the drawback that it overestimates the variable importance of highly correlated variables (Strobl et al., 2008). The latter, on the other hand, does not perform fairly with predictors of many categories (Strobl et al., 2007). In this study, both measures along with Pearson's correlation test were employed to select the important variables for real-time safety evaluation.

4.3. Bayesian logit model

To predict real-time crash likelihood, logistic regression models under Bayesian framework were evaluated. Logistic models and their extensions have been widely used in real-time safety studies with data from different sources, such as loop detectors (Abdel-Aty et al., 2004), microwave radar sensors (Yu et al., 2013), Automatic Vehicle Detection (AVI) data (Abdel-Aty et al., 2012; Ahmed and Abdel-Aty, 2012) and vehicle trajectory data (Oh and Kim, 2010). Geometric characteristics and weather data have also been proven to be useful in logistic models (Abdel-Aty et al., 2012; Yu et al., 2013; Xu et al., 2013). Therefore this statistical method is able to handle information from different sources. With the fast development in Big Data, it is expected that new data sources could be incorporated in this modeling frame in the future. The target variable is the binary indicator of crash occurrence, with probability p for crash case ($y = 1$) and $1 - p$ for non-crash case ($y = 0$). Three types of logistic models were constructed and their performances compared: (1) the matched case-control logit model; (2)

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

fixed effects logit model; (3) random parameters logit model differentiating peak and non-peak hours. The specifications of the logit models are illustrated below:

$$y_i \sim Binomial(p_i, 1) \tag{2}$$

For model type (1),

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{X_i\beta} + \varepsilon_{j[i]} \tag{3}$$

For model type (2),

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \boldsymbol{X_i\beta} \tag{4}$$

For model type (3),

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_{0[t]} + \boldsymbol{X_i\beta_t} \tag{5}$$

where β_0 is constant term, $\boldsymbol{\beta}$ is the vector of explanatory variable coefficients. $\beta_{0[t]}$ and $\boldsymbol{\beta_t}$ bear the same meaning except that they are for peak ($t = 1$) and non-peak ($t = 0$) traffic hours. $\varepsilon_{j[i]}$ denotes the contribution to the logit of all terms constant within the j^{th} group (Hosmer and Lemeshow, 2004). In Bayesian inference, prior distributions for parameters have to be justified first. Here non-informative priors were assigned. For $\beta_0/\beta_{0[t]}$ and each element in $\boldsymbol{\beta}/\boldsymbol{\beta_t}$, they were assigned to follow $normal(0, 10^6)$. ε_j has normal distribution $normal(0, 1/\tau)$ where $\tau \sim gamma(0.001, 0.001)$.

The models were calibrated in the software WinBUGS. Three chains were simulated with the first 5000 iterations fed as burn-in out of the 15,000 iterations. Parameter convergence was assured by checking if the trace plots of the three chains appear to be overlapping one another (Spiegelhalter et al., 2003). The Deviance Information Criterion (DIC) was used as a Bayesian measure of model complexity and fit. Smaller DIC indicates better model. Bayesian Credible Interval (BCI) was used for parameter estimation. If the 95% BCI does not contain 0, then the effect of the variable is significant.

4.4. First Order Reliability analysis

Reliability and risk analysis is essential concern in structural engineering. The combined actions of the elements of the structure will lead to system failure when specific conditions are met. As a result, reliability analysis is used to distinguish safe and unsafe conditions. In traffic safety, we can also interpret each crash as a failure of the expressway system. Similarly, we can derive the combination of traffic parameters indicating potential hazards on the system.

Reliability is denoted by the probability of limit state function (LSF) $g(\boldsymbol{X}) > 0$ where $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$. When $g(\boldsymbol{X})$ is less than zero, it means that the system is in failure region. Then reliability could be numerically expressed as

$$R = P\{g(\boldsymbol{X}) > 0\} = \int_{g(x)>0} f_x(\boldsymbol{X})dx \tag{6}$$

where $f_x(\boldsymbol{X})$ is the joint distribution of $\boldsymbol{X}$. In the logit model of real-time crash prediction, each case is corresponded with p, if p is higher than a cutoff point (c), then the case will be classified as crash, otherwise non-crash. Based on inverse logit function, we can apply the reliability analysis in real-time safety evaluation as follows:

$$p_i = \frac{\exp(\beta_0 + \boldsymbol{X_i\beta})}{1 + \exp(\beta_0 + \boldsymbol{X_i\beta})} < c \tag{7}$$

$$g(\boldsymbol{X}) = -\boldsymbol{X_i\beta} - \beta_0 + \log\left(\frac{c}{1-c}\right) > 0 \tag{8}$$

To get the joint distribution $f_x(\boldsymbol{X})$, the distribution of each significant variable and the correlations between them were determined first. SAS procedure SEVERITY was used to test candidate distributions and −2 Log Likelihood chosen as the comparison criterion. Then the joint probability density function (PDF) is calculated and transformed from the variables' original $\boldsymbol{X}$ space to the standard normal $\boldsymbol{U}$ space using Nataf transformation (Li et al., 2008) as illustrated with two variables in Fig. 3.

When the transformed joint PDF $f_x(\boldsymbol{U})$ with the corresponding LSF $g(\boldsymbol{U})$ are available, First Order Reliability Method (FORM) is applied to calculate the critical point. The term "first order" means that the method approximates the LSF by taking the first order Taylor expansion. The critical point is achieved by maximization of the transformed joint PDF at the limit state $g(\boldsymbol{U}) = 0$, which means it has the highest probability to fail. Then it is transformed back into $\boldsymbol{X}$ space. A more detailed description of FORM can be found in Yu et al. (2013).

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

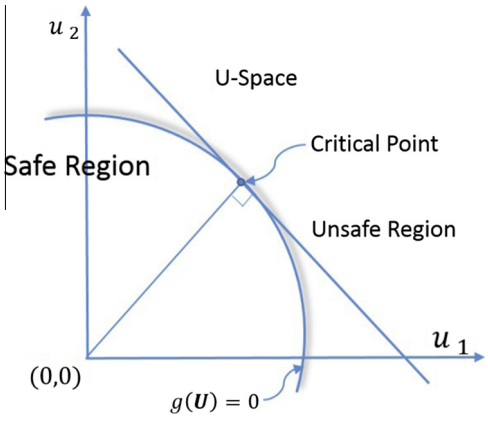

Fig. 3. Reliability expressed in U space (Nowak and Collins, 2012).

5. Congestion monitoring and safety modeling results

5.1. Congestion evaluation

Of the eight-month period, the most recent month (February, 2014) was selected to reflect the congestion conditions on the three expressways. As mentioned, the CI values were aggregated at five-minute intervals for each station. To achieve more stable conclusions about congestion segments and time duration, they were averaged by the weekdays in February. Currently, CFX applies TTI as their congestion measurement. TTI of 1.25 and 2.0 are defined as thresholds for moderate and high congestion. The ratio between actual travel speed and free flow speed given the two TTI congestion thresholds are namely 4:5 and 2:1, which are equivalent to CI of 0.2 and 0.5 respectively. As a result, CI values of 0.2 and 0.5 are set up as the moderate and high congestion thresholds. From Fig. 4a–f below, it can be concluded that the level of congestion is highly localized and time specific. For the same expressway, morning and evening peak hours are identified on multiple detecting points. For the same location, the congestion levels vary significantly across different time of day. Therefore, to achieve more accurate congestion detection, continuous monitoring is necessary. For traffic safety studies, the use of real-time congestion measurement should also be encouraged to reveal the true effect of congestion on crash occurrence.

5.2. Variable selection

The random forest model for variable selection was constructed in R-based data mining tool Rattle (Williams, 2011). In the model specification, 4 variables were randomly sampled at each split. As suggested by Breiman (2006), the number of trees to grow in random forest should not be too small, to ensure that every observation gets predicted at least a few times, we grew 500 trees. The variable importance rankings only listed the most important 20 variables out of the total 37 candidate variables; other candidate variables deemed unimportant by the algorithm were omitted. Fig. 5 depicts that while there are minor differences between the variable rankings by the two importance ranking methods, the most important types of variables are basically the same. Logarithm of volume, peak hour, average speed and congestion index are the crucial variables. However, the correlation between variables should be investigated before identifying the variables to be incorporated in the final model. To cope with the issue, Pearson's correlation test in Table 3 and a simple logistic regression were run to keep the significant variables but controlling for correlations.

Synthesizing the results from random forest, correlation test and preliminary logistic regression, four variables were selected for the real-time crash prediction model: the peak hour indicator, logarithmic volume and average speed at U2 station and the CI at D1 station. Descriptive statistics of these variables are provided in Table 4.

5.3. Real-time logit models

Three types of real-time logit models based on the training data were tested and compared. All the four variables involved in the final model stage appear to be significant at 95% confidence interval (Table 5). Peak hour is proved to significantly increase crash likelihood. Logarithmic volume at U2 station is positively related with crash occurrence while the average speed at the same location is negatively associated with rear-end crash. The congestion index at D1 station is also found to contribute to crashes. While congestion index is a direct measure of congestion intensity, higher traffic volume and lower speed are also regarded as indirect indicators of congestion. In this study, the effects of traffic parameters at both upstream and downstream locations all converged to the same statement: rear-end crashes are significantly affected by traffic congestion on urban expressways.

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

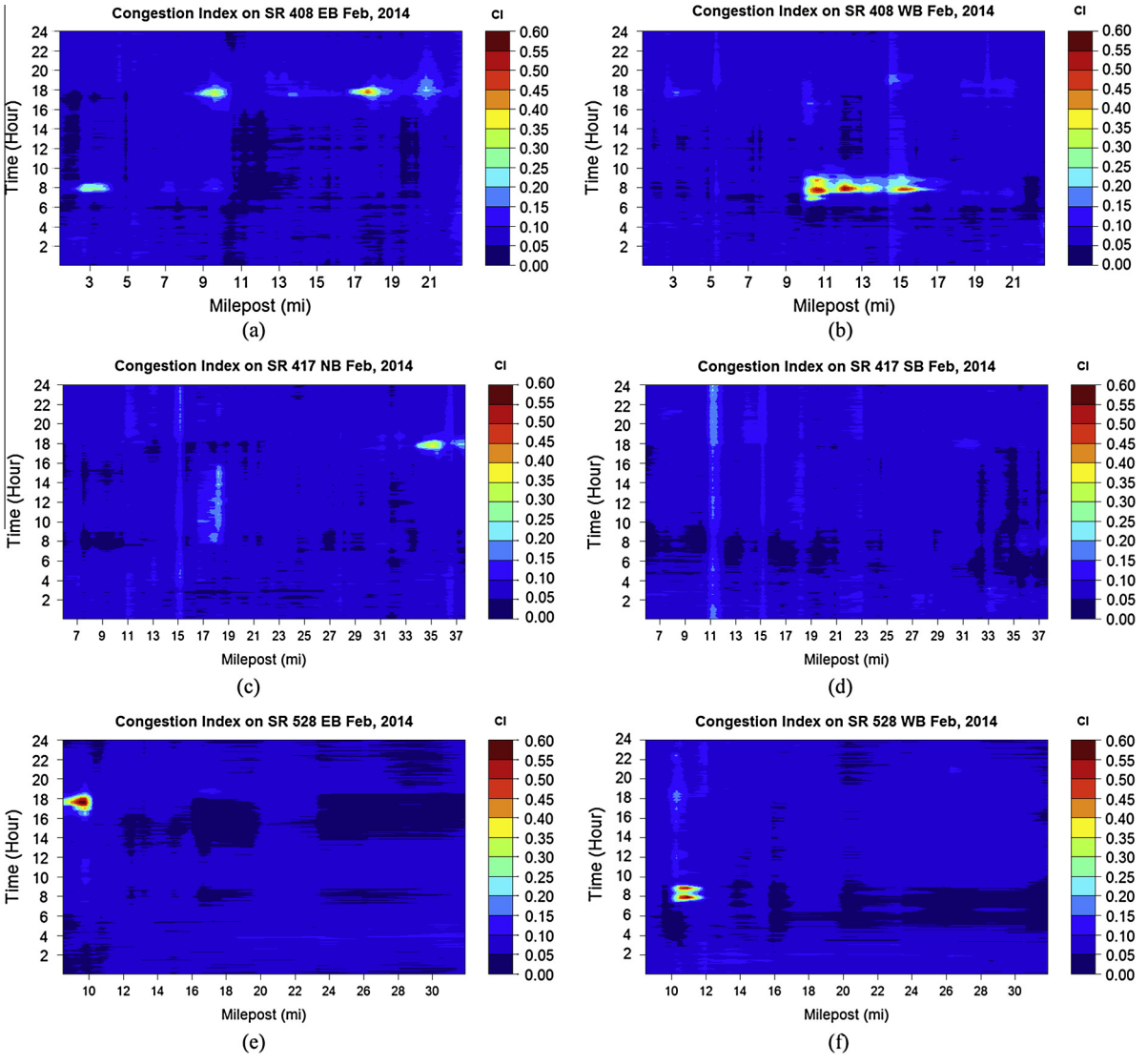

Fig. 4. Congestion evaluation for the expressway system (a)–(f).

As for the model selection, both model performance and prediction power were compared. The DIC values for the three models are comparable with random parameter model having slightly lower DIC. To evaluate the prediction power of the model, an appropriate cutoff point that classifies the membership (crash/non-crash) of each observation has to be determined first. Since the cutoff point is a probability value, theoretically it ranges from 0 to 1. The selection of a cutoff point that is too small would cause high sensitivity (crashes correctly identified) and low specificity (non-crashes correctly identified). On the other hand, a cutoff point that is too large would cause low sensitivity and high specificity. In these cases, the cutoff point is biased towards either sensitivity or specificity, thus not optimized. As a response, a graphical method is to select the optimal point where sensitivity and specificity curves cross (Hosmer and Lemeshow, 2004). Only in this case, the cutoff point is neither biased towards sensitivity nor specificity. The optimal cutoff point in this study is 0.14 as illustrated in Fig. 6. The prediction outcome, sensitivity, specificity, overall accuracy rate and area under the ROC (AUC) of both training and validation data sets are calculated for each model (Table 6). The performances of the three models based on training data are similar. In the validation data set, the random parameter outperforms others in specificity and overall accuracy rate.

Based on the model fitting and prediction power of the training and validation data, we prefer the random parameter model as the input for reliability analysis. Another practical reason we choose the random parameter model is that FORM analysis could not deal with categorical variables.

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

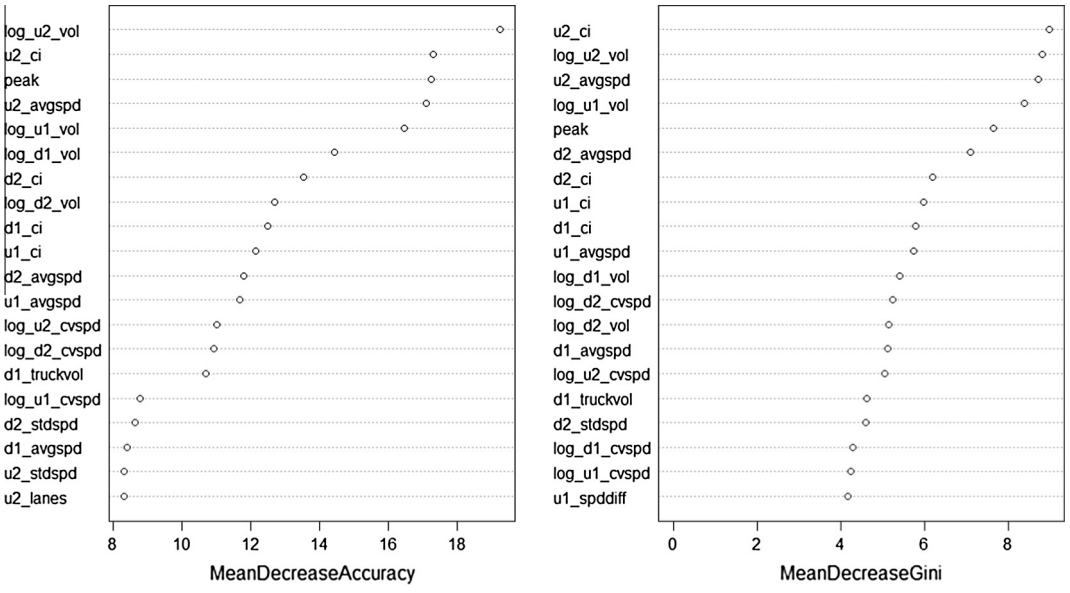

Fig. 5. Variable importance based on random forest.

Table 3
Pearson's correlation test for variables in the final model.

Pearson's correlation	peak	log_u2_vol	u2_avgspd	d1_ci
peak	1.0000	0.3437	−0.2767	0.3289
log_u2_vol	0.3437	1.0000	−0.2426	0.1331
u2_avgspd	−0.2767	−0.2426	1.0000	−0.3705
d1_ci	0.3289	0.1331	−0.3705	1.0000

Table 4
Summary of descriptive statistics.

	Description	Mean	Std Dev	Min	Max
crash	Rear-end crash: non-crash = 0; crash = 1	0.201	0.401	0.000	1.000
peak	Peak hours: nonpeak = 0; peak = 1	0.161	0.368	0.000	1.000
log_u2_vol	Log volume of U2 station	4.611	1.118	0.693	6.762
u2_avgspd	Average speed of U2 station	62.138	8.735	2.500	98.000
d1_ci	Congestion index of D1 station	0.058	0.103	0.000	0.909

Table 5
Variable effects and model comparison.

	Random effect		Fixed effect		Random parameter	
	Mean	95% BCI	Mean	95% BCI	Mean	95% BCI
Intercept	–	–	−1.505	(−3.487, 0.541)	−1.031[1] −3.315[2]	(−4.176, 2.589) (−5.268, −0.092)
peak	1.905	(1.435, 2.419)	1.857	(1.373, 2.363)	– –	– –
log_u2_vol	0.275	(0.116, 0.435)	0.382	(0.171, 0.596)	0.338[1] 0.823[2]	(0.117, 0.554) (0.291, 1.374)
u2_avgspd	−0.057	(−0.070, −0.045)	−0.042	(−0.066, −0.016)	−0.032[1] −0.048[2]	(−0.059, −0.002) (−0.087, −0.017)
d1_ci	6.053	(3.253, 9.546)	6.809	(3.658, 10.920)	7.288[1] 6.190[2]	(3.428, 12.160) (6.200, 10.630)
Model estimation						
$\overline{D}$	634.211		632.975		629.562	
p_D	4.879		5.171		6.652	
DIC	639.090		638.146		636.214	

[1] non-peak hours; [2] peak hours.

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

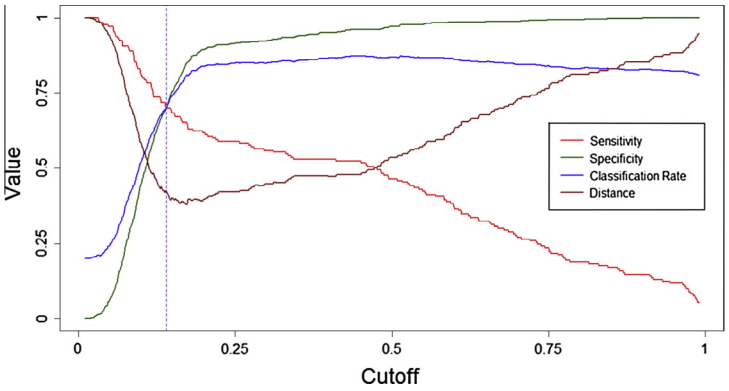

Fig. 6. Cutoff point determination.

Table 6
Model classification results.

	Random effect		Fixed effect		Random parameter	
Training data						
Observed						
Predicted	0	1	0	1	0	1
0	493	182	474	201	491	184
1	52	118	50	120	52	118
Sensitivity	0.706		0.706		0.694	
Specificity	0.702		0.702		0.727	
Accuracy	0.703		0.703		0.721	
AUC	0.774		0.779		0.781	
Validation data						
Observed						
Predicted	0	1	0	1	0	1
0	205	82	202	85	210	77
1	23	50	23	50	23	50
Sensitivity	0.685		0.685		0.685	
Specificity	0.714		0.704		0.732	
Accuracy	0.708		0.700		0.722	
AUC	0.755		0.755		0.755	

5.4. Reliability analysis

In reliability analysis, since we are more interested in the congestion's effects on crashes, we focus on the peak hour conditions during which time period congestion is more predictable. As discussed earlier, when the combined effects of individual elements in a system reach certain state, the system has high probability of failure and reliability analysis could determine the critical point. As the application of reliability analysis in traffic safety evaluation, the critical point indicates that the crash contributing factors at this combination would most likely cause potential unsafe conditions on the expressway system. When such state is reached, it can be understood as that unsafe condition emerges. Safety interventions will then be triggered. Since the critical point is derived based on real-time traffic data 10–5 minutes prior to crashes, when such conditions are encountered, it could be treated as crash prone conditions. Traffic authority could proactively warn the motorists through the use of Dynamic Message Signs (DMS). On the basis of random parameter logistic regression results, we defined the LSF for peak hour as

$$g(\boldsymbol{X}) = -0.823\log(\text{u2\_vol}) + 0.048(\text{u2\_avgspd}) - 6.190(\text{d1\_ci}) + 1.50 \tag{9}$$

Then the distribution for each variable was fitted as displayed in Table 7. Five types of candidate distributions were tested. For each distribution fitting, convergence and goodness of fit were recorded. The smaller −2 Log Likelihood value, the better the candidate distribution fits. The log_u2_vol follows normal distribution. Weibull distribution has the lowest −2 Log Likelihood values for u2_avgspd and d1_ci.

Given the distribution of each variable, the FORM analysis was conducted using OpenSees software (Mazzoni et al., 2006). In FORM analysis, it is the whole expressway system instead of individual locations that is under inspection. As input for the FORM model, basic statistics (i.e. mean and standard deviation) and the correlation matrix of the three variables during peak hours are required. The results of the FORM critical point are shown in Table 8.

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

Table 7
Distribution fitting and selection.

Distribution	log_u2_vol			u2_avgspd			d1_ci		
	Converged	−2 LL	Selected	Converged	−2 LL	Selected	Converged	−2 LL	Selected
Normal	Yes	496	Yes	Yes	1663	No	Yes	−76	No
Lognormal	Yes	1053	No	Yes	1757	No	Yes	−443	No
Exponential	Yes	1098	No	Yes	2058	No	Yes	−389	No
Weibull	Yes	555	No	Yes	1645	Yes	Yes	−458	Yes
Gamma	Yes	769	No	Yes	1719	No	Yes	−452	No

−2 LL: −2 Log Likelihood.

Table 8
FORM analysis for peak hour.

	Basic statistics		Pearson correlations			FORM critical point
	Mean	Std. Dev	log_u2_vol	u2_avgspd	d1_ci	
log_u2_vol	5.007	0.561	1	−0.366	0.259	5.17
u2_avgspd	67.448	4.785	−0.366	1	−0.371	67.0
d1_ci	0.071	0.034	0.259	−0.371	1	0.075

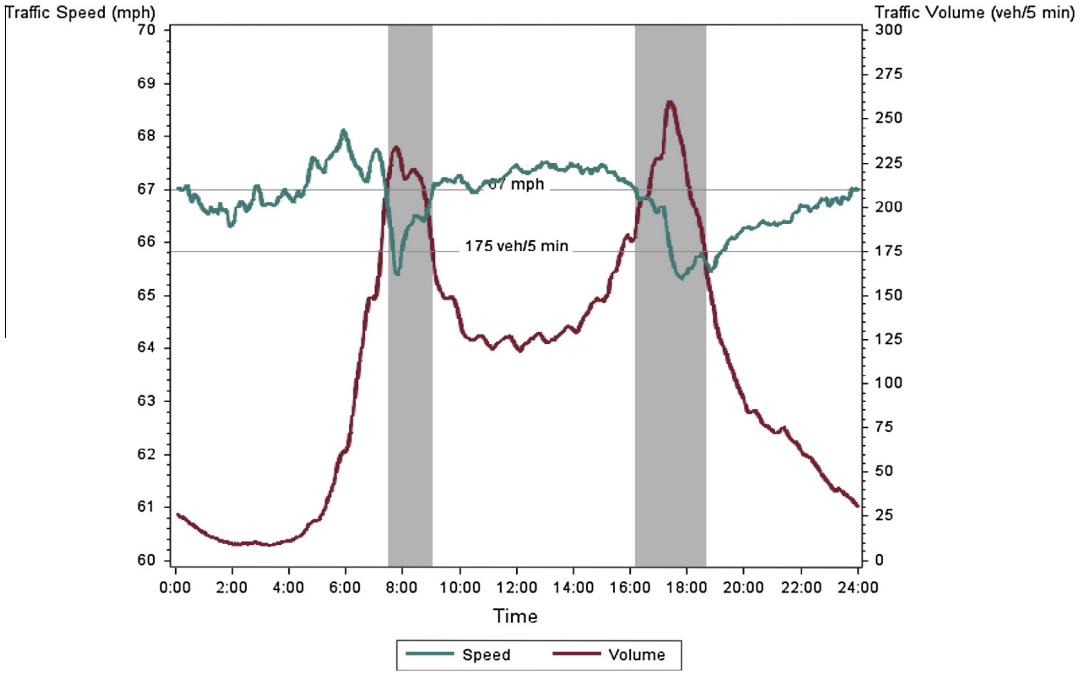

Fig. 7. System speed and 5-minute volume profiles.

According to the critical point, when the system (the average condition over the whole network) congestion index reaches 0.075, logarithmic volume on a cross-section exceeds 5.17 (equivalent to volume of 175) and speed drops below 67 mph, there would be high probability of a crash on the expressway. Based on the profiles of these three parameters along time, the time period when all of them reaches the critical point is 7:30 to 9:00 in the morning and 17:00 to 18:40 in the evening illustrated by the shaded area in Figs. 7 and 8. The two figures visually indicate that the rear-end crashes are more likely to occur during morning and evening peak hours. In addition, the distribution of rear-end crashes in Fig. 9 on the studied expressway system also proves the conclusion from FORM analysis.

However, when the three system parameters reach the critical point, it is still desirable to know which segment has especially higher probability to have a crash. While it is impossible to conduct the FORM analysis at each detection location in this paper, a simplified method is proposed here by checking the volume, speed, and CI for each crash cases. During peak hours, the average CI for crash cases is 0.2 (Table 9), the same as the moderate congestion threshold with speed at 52

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

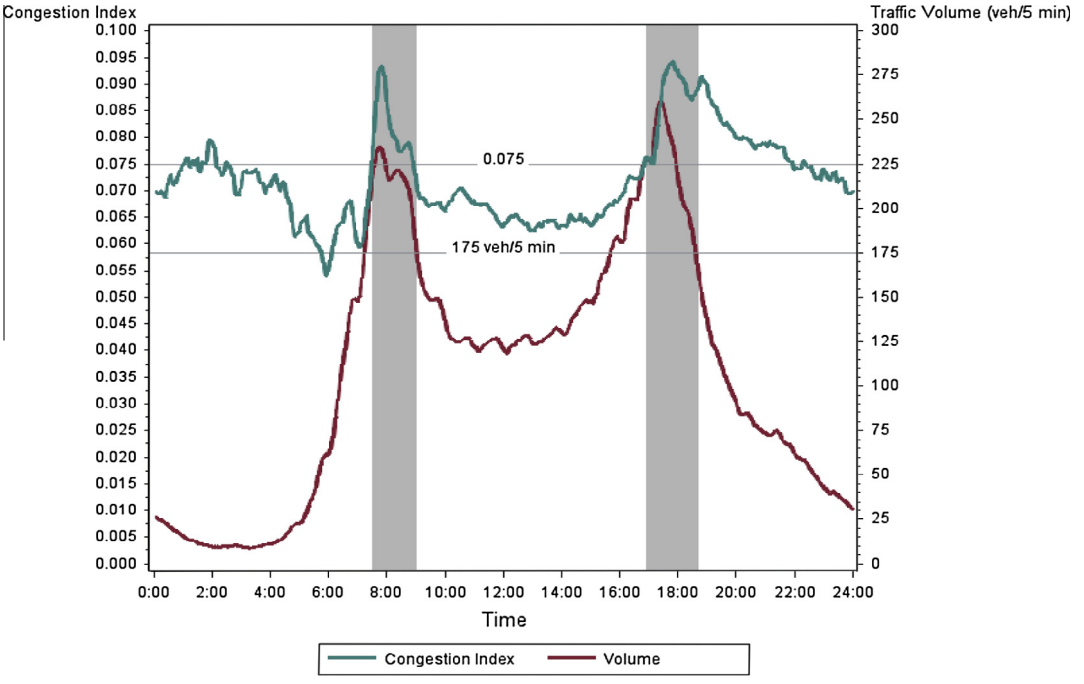

Fig. 8. System congestion index and 5-minute volume profiles.

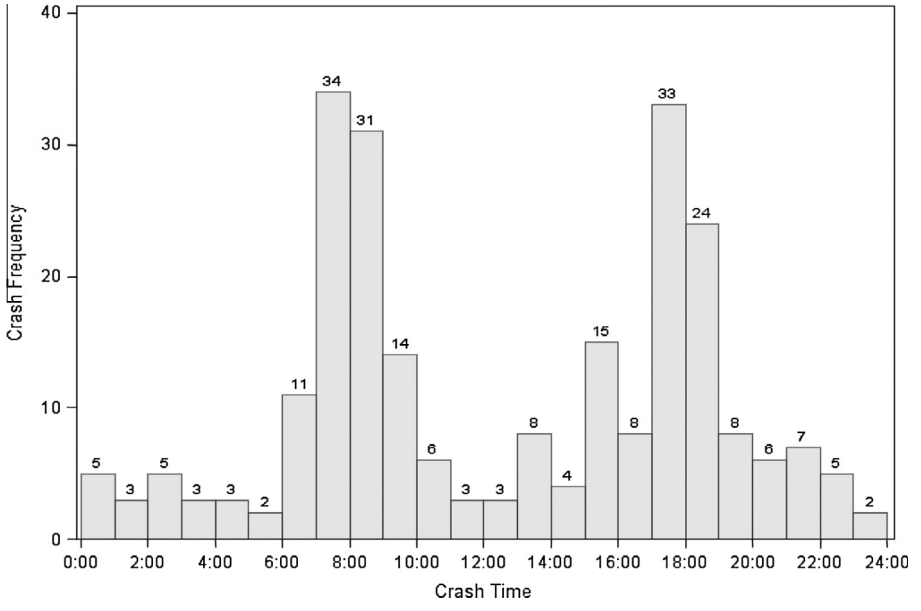

Fig. 9. Time distribution of rear-end crashes on expressway network.

mph and 5-minute traffic volume at 360. Therefore when individual detection locations encounter such traffic conditions, these specific locations could be deemed as crash-prone locations. Combining the conclusions of FORM analysis for the whole system and individual crash cases, real-time safety monitoring could be carried out using CI, volume and speed together as safety indicators: when the system has 5-minute volume and CI above 175 and 0.075 respectively and speed below 67 mph, the three indicators at each detector location will be examined; if CI at individual detection locations exceeds 0.20 with 5-minute volume greater than 360 and speed below 52 mph, safety countermeasures should be triggered at the upstream segment.

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

Table 9
Congestion index for crash cases.

peak	Variable	Crash count	Mean	Std Dev
0 (non-peak)	u2_avgspd	120	60.3210	12.0269
	u2_volume		160.0667	120.7305
	d1_ci		0.0683	0.1310
1 (peak)	u2_avgspd	123	52.7322	15.7393
	u2_volume		360.1789	184.5100
	d1_ci		0.2000	0.2349

Table 10
Real-time monitoring of congestion and traffic safety.

	Individual detector		
	Other	Speed <52 mph & 5-minute volume >360 and CI >= 0.2	
System	Other	Safe and no congestion	Congestion warning
	speed <67 mph & 5-minute volume >175 & CI >= 0.075	Prepare for safety warning	Congestion and safety warning

6. Conclusions

The rapid development of ITS systems in the past few decades has catalyzed the implementation of Big Data in the transportation arena. To harness the power of Big Data for better traffic system performance, it is vital to take full advantage of its real-time nature. In this study, the viability of monitoring and improving traffic operation and safety on urban expressways in Central Florida using real-time Microwave Vehicle Detection System (MVDS) data is researched. From perspectives of volume, velocity and variety, the MVDS should be regarded as a main source of Big Data. The detection system archives spot speed, volume, lane occupancy and volume by vehicle type per lane on minute basis. Congestion detection and the real-time safety analysis were developed for three expressways based on these data.

Traditional congestion measures lack the ability to capture the variability of congestion. Real-time congestion measurement based on Big Data is therefore more desirable to identify the congestion pattern in both the temporal and spatial dimensions. Congestion index was introduced to measure congestion intensity and visualized via filled contour plot. It was found that congestion on the urban expressways is highly time and location specific. Recurrent congestion during morning and evening peak hours are observed for specific locations. Faced with the large traffic demand during peak hours, the traffic authorities could not always expand the system capacity as a solution. Currently, DMS have been widely applied on the CFX system for travel time estimation. However, they could also be used for congestion warning. Information of congestion locations and potential delay would leave drivers enough time to adjust their speed and raise their awareness of surrounding traffic. If smoother traffic flow is achieved, it is expected that congestion will be alleviated. In extreme cases of total shutdown of the expressway, traffic could diverge at nearby ramps to avoid deterioration of congestion by DMS suggestions. As a conclusion, application of Big Data for better operation should emphasize real-time monitoring of traffic condition and a quick response based on the retrieved data.

How congestion affects the crash occurrence has been discussed in some existing literature. In aggregate safety analysis, the issue related to averaging congestion intensity might be the cause of the insignificant effects of congestion found in many crash frequency studies. In case of this study, it was verified that the congestion was highly localized and time specific. As a result, to gain better understanding whether congestion leads to more crashes, it was deemed better to be evaluated under real-time modeling framework. Big Data enables the restoration of traffic for each crash case. Moreover, rear-end crashes were selected as the target since their connection with congestion could be more straightforward. Both data mining and Bayesian statistics techniques were adopted to identify the leading contributing factors to crashes in real-time. The results concluded that peak hour, higher volume and lower speed at upstream locations, and high congestion index (CI) at downstream detection point significantly increased crash likelihood. Thus, direct (CI) and indirect (volume, speed) congestion indicators all support the assumption that congestion has an impact on rear-end crashes.

Different from previous real-time traffic safety studies, the current work takes one step further by incorporating reliability analysis to determine the conditions when it is appropriate to trigger safety warnings on the expressway. First-Order Reliability Method (FORM) model was constructed based on the real-time crash prediction model and the critical point of system CI, volume and speed was calculated. When the system reaches the critical point, it does not necessarily mean equal risk for each section. Accordingly, the values of the above three parameters for each crash case was investigated. It was found that the average CI for peak hour crashes was equal to the congestion threshold, which suggested that when congestion is detected at a specific location, both congestion and safety warnings should be sent to motorists.

As a final effort of this study, we propose combined real-time monitoring of congestion and safety on urban expressways through the MVDS system. A strategy taking both system and local congestion into account is summarized in Table 10. The proposed framework is to highlight the association between congestion and safety. On urban expressways, improving either one would be beneficial to the other.

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022

Despite that this study conducted relatively comprehensive analyses on both congestion and safety, there are still plenty of room for further improvement: (1) the current work focuses on rear-end crashes only. However, Big Data applications could be easily extended to other crash types and severities. (2) Multiple factors other than congestion could lead to crashes in real world. In this study, only traffic dynamics were included in the real-time models. However, to fully realize the power of Big Data, more data sources should be utilized. Real-time weather condition could be an important factor for expressway operation and safety as well. Future work should also make use of Big Data from other sources.

Acknowledgments

The authors wish to thank the Central Florida Expressway Authority (CFX) for funding this research and providing the data. Partial funding was provided by the Southeastern Transportation Center UTC consortium. The authors wish to thank Charles R. Lattimer, PE, PMP Sr. Project Manager, Intelligent Transportation Systems at ATKINS for his suggestions.

References

Abdel-Aty, M., Haleem, K., 2011. Analyzing angle crashes at unsignalized intersections using machine learning techniques. Accid. Anal. Prev. 43 (1), 461–470.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F.M., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transp. Res. Rec.: J. Transp. Res. Board 1897 (1), 88–95.

Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., Cunningham, R., Dhindsa, A., Dilmore, J., 2007. Linking Crash Patterns to ITS-related Archived Data: Phase II, Volume I: Real-time Crash Risk Assessment Models. Florida Department of Transportation, BD-550-5.

Abdel-Aty, M., Pande, A., Das, A., Knibbe, W.J., 2008. Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. Transp. Res. Rec.: J. Transp. Res. Board 2083 (1), 153–161.

Abdel-Aty, M.A., Hassan, H.M., Ahmed, M., Al-Ghamdi, A.S., 2012. Real-time prediction of visibility related crashes. Transp. Res. Part C: Emerg. Technol. 24, 288–298.

Ahmed, M.M., Abdel-Aty, M.A., 2012. The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Trans. Intell. Transp. Syst. 13 (2), 459–468.

Baruya, A., 1998. Speed–accident relationships on European roads. In: Proceedings of the 9th International Conference on Road Safety in Europe.

Beyer, M.A., Laney, D., 2012. The Importance of 'Big Data': A Definition. Gartner, Stamford, CT.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Breiman, L., 2006. Randomforest: Breiman and Cutler's Random Forests for Classification and Regression. R Package Version. <http://stat-www.berkeley.edu/users/breiman/RandomForests>.

CFX, 2014. Expressway Map. Central Florida Expressway Authority.

Christoforou, Z., Cohen, S., Karlaftis, M.G., 2011. Identifying crash type propensity using real-time traffic data on freeways. J. Saf. Res. 42 (1), 43–50.

Dias, C., Miska, M., Kuwahara, M., Warita, H., 2009. Relationship between congestion and traffic accidents on expressways: an investigation with Bayesian belief networks. In: Proceedings of 40th Annual Meeting of Infrastructure Planning (JSCE), Japan.

Golob, T.F., Recker, W.W., 2003. Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. J. Transp. Eng. 129 (4), 342–353.

Grant, M., Bowen, B., Day, M., Winick, R., Bauer, J., Chavis, A., 2011. Congestion Management Process: A Guidebook. United States Department of Transportation-Federal Highway Administration, Washington, DC.

Hammond, P.J., 2012. WSDOT 2012 Congestion Report. WSDOT's Comprehensive Annual Analysis of State Highway System Performance, 11th ed.

Hosmer Jr., D.W., Lemeshow, S., 2004. Applied Logistic Regression. John Wiley & Sons.

Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accid. Anal. Prev. 45, 373–381.

Lee, C., Abdel-Aty, M., Hsia, L., 2006. Potential real-time indicators of sideswipe crashes on freeways. Transp. Res. Rec.: J. Transp. Res. Board 1953 (1), 41–49.

Li, H., Lü, Z., Yuan, X., 2008. Nataf transformation based point estimate method. Chin. Sci. Bull. 53 (17), 2586–2592.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22.

Martin, P.T., Feng, Y., Wang, X., 2003. Detector Technology Evaluation. Mountain-Plains Consortium.

Mazzoni, S., Mckenna, F., Scott, M.H., Fenves, G.L., 2006. Opensees Command Language Manual. Pacific Earthquake Engineering Research (PEER) Center.

Nowak, A.S., Collins, K.R., 2012. Reliability of Structures. CRC Press.

Oh, C., Kim, T., 2010. Estimation of rear-end crash potential using vehicle trajectory data. Accid. Anal. Prev. 42 (6), 1888–1893. http://dx.doi.org/10.1016/j.aap.2010.05.009.

Oh, C., Park, S., Ritchie, S.G., 2006. A method for identifying rear-end collision risks using inductive loop detectors. Accid. Anal. Prev. 38 (2), 295–301.

Pande, A., Abdel-Aty, M., 2006. Assessment of freeway traffic parameters leading to lane-change related collisions. Accid. Anal. Prev. 38 (5), 936–948.

Quddus, M.A., Wang, C., Ison, S.G., 2009. Road traffic congestion and crash severity: econometric analysis using ordered response models. J. Transp. Eng. 136 (5), 424–435.

Schrank, D., Eisele, B., Lomax, T., 2012. TTI's 2012 Urban Mobility Report. Texas A&M Transportation Institute.

Shefer, D., Rietveld, P., 1997. Congestion and safety on highways: towards an analytical model. Urban Stud. 34 (4), 679–692.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., 2003. Winbugs User Manual. MRC Biostatistics Unit, Cambridge.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinf. 8 (1), 25.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinf. 9 (1), 307.

Turner, S.M., Eisele, W.L., Benz, R.J., Holdener, D.J., 1998. Travel Time Data Collection Handbook.

Wang, C., Quddus, M.A., Ison, S.G., 2009. Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. Accid. Anal. Prev. 41 (4), 798–808.

Williams, G., 2011. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Springer Science & Business Media.

Xu, C., Liu, P., Wang, W., Li, Z., 2012. Evaluation of the impacts of traffic states on crash risks on freeways. Accid. Anal. Prev. 47, 162–171.

Xu, C., Tarko, A.P., Wang, W., Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accid. Anal. Prev. 57, 30–39. http://dx.doi.org/10.1016/j.aap.2013.03.035.

Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.

Yu, R., Abdel-Aty, M., 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Saf. Sci. 63, 50–56.

Yu, R., Shi, Q., Abdel-Aty, M., 2013. Feasibility of incorporating reliability analysis in traffic safety investigation. In: Proceedings of the Transportation Research Board 92nd Annual Meeting.

Please cite this article in press as: Shi, Q., Abdel-Aty, M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.02.022