Accident Analysis and Prevention 85 (2015) 45–57

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Safety performance of traffic phases and phase transitions in three phase traffic theory

Chengcheng Xu[a,b], Pan Liu[a,b,*], Wei Wang[a,b], Zhibin Li[a,b]

[a] Jiangsu Key Laboratory of Urban ITS, Southeast University, Si Pai Lou #2, Nanjing 210096, China
[b] Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Si Pai Lou #2, Nanjing 210096, China

ARTICLE INFO

Article history:
Received 9 December 2014
Received in revised form 12 June 2015
Accepted 24 August 2015
Available online 12 September 2015

Keywords:
Traffic flow
Three phase traffic theory
Crash likelihood
Freeway
Phase transition

ABSTRACT

Crash risk prediction models were developed to link safety to various phases and phase transitions defined by the three phase traffic theory. Results of the Bayesian conditional logit analysis showed that different traffic states differed distinctly with respect to safety performance. The random-parameter logit approach was utilized to account for the heterogeneity caused by unobserved factors. The Bayesian inference approach based on the Markov Chain Monte Carlo (MCMC) method was used for the estimation of the random-parameter logit model. The proposed approach increased the prediction performance of the crash risk models as compared with the conventional logit model. The three phase traffic theory can help us better understand the mechanism of crash occurrences in various traffic states. The contributing factors to crash likelihood can be well explained by the mechanism of phase transitions. We further discovered that the free flow state can be divided into two sub-phases on the basis of safety performance, including a true free flow state in which the interactions between vehicles are minor, and a platooned traffic state in which bunched vehicles travel in successions. The results of this study suggest that a safety perspective can be added to the three phase traffic theory. The results also suggest that the heterogeneity between different traffic states should be considered when estimating the risks of crash occurrences on freeways.

1. Introduction

With the widespread use of freeway traffic surveillance instruments, increased attention has been given to identifying the traffic flow conditions prior to crash occurrences to better understand to what extent freeway traffic flow affects safety. A number of traffic flow measures, such as the difference in detector occupancy between adjacent lanes, the standard deviation of vehicle speed and the difference in upstream and downstream speeds, have been identified as crash precursors (Golob and Recker, 2004; Golob et al., 2004; Abdel-Aty and Rajashekar, 2006; Hossain and Muromachi, 2011; Pande et al., 2011; Xu et al., 2013a,b). Various crash risk models have been developed to link the likelihood of crash occurrences on freeways to crash precursors. Such information is needed in dynamic traffic management systems (DTMS) on freeways to apply proactive control strategies to reduce crash likelihood (Hossain and Muromachi, 2011; Papageorgiou and Kotsialos, 2002; Allaby et al., 2007).

The crash risk models developed in previous studies generally did not take into consideration the varying crash mechanism across different traffic states on freeways. The models were developed to relate crash likelihood to freeway traffic flow parameters, no matter if the traffic is operating in a congested or free flow state. Such assumption ignores the fact that drivers behave differently across various macroscopic traffic flow states. More specifically, the mechanism of crash occurrences and the relationship between crash likelihood and crash precursors may be quite different across various traffic states.

In experimental and theoretical studies of traffic flow, freeway traffic was divided into various aggregate states to identify complex dynamical behavior (Hall et al., 1992; Kerner and Rehborn, 1996; Wu, 2002). So far it is still not sure how these traffic flow states can be linked to safety. In the authors' previous study, freeway traffic flow was divided into five states using traffic occupancy measured at four nearby loop detector stations. It was found that the effects of traffic flow on crash likelihood are different across various traffic states (Xu et al., 2012). In a more recent study the authors further evaluated the safety performance associated with freeway traffic flow operating at different levels of service (LOS). It was found that

* Corresponding author at: Jiangsu Key Laboratory of Urban ITS, Southeast University, Si Pai Lou #2, Nanjing 210096, China.
E-mail addresses: iamxcc1@163.com (C. Xu), pan_liu@hotmail.com (P. Liu), wangwei@seu.edu.cn (W. Wang), lizhibin-2002@163.com (Z. Li).

http://dx.doi.org/10.1016/j.aap.2015.08.018
0001-4575/

a safety level can be assigned to freeway LOS, and the contributing factors to crash likelihood are different across various LOS. A crash risk prediction model that considered the varying contributing factors to crash likelihood across various LOS was developed (Xu et al., 2014).

The study presented in this paper is a continuous effort following the authors' previous studies to understand how freeway traffic flow affects safety. More specifically, the present study aims to evaluate the safety performance associated with freeway traffic flow in the framework of the three phase traffic theory, which was proposed by Kerner in 1996 (Kerner and Rehborn, 1996; Kerner, 1998). The three-phase traffic theory is a prominent approach for modeling freeway traffic flow without a fundamental diagram. It was developed to predict and explain the empirical spatiotemporal features of traffic breakdown and the resulting traffic congestion. In the three-phase traffic theory, freeway traffic flow was classified into three phases: (1) free-flow phase (F), which is characterized by high vehicle speeds. There is a monotonously increasing and almost linear relationship between flow and density in free flow; (2) synchronized flow phase (S), in which the speeds in neighboring lanes have a tendency of synchronization, and (3) wide moving jams (J), which is characterized by very high densities and very low speeds, sometimes as low as zero (Kerner, 2004; Kerner and Klenov, 2009).

The phase transitions between traffic phases were used to explain the features of spatial-temporal congested patterns. The phase transition from free flow to synchronized flow (F→S) can be considered the onset of congestion. The traffic break down phenomenon can be explained by the F→S transition (Kerner, 2004). The phase transition F→S is more probable than phase transition from free flow to wide moving jams (F→J) in free flow at the same density (Kerner, 1998, 2002). The phase transition F→S at a bottleneck can be explained by a competition between over-acceleration and speed adaptation (Kerner and Klenov, 2003, 2009). The lane changing behavior exhibits dual roles in phase transitions. The lane changing behavior that causes a strong reduction in following vehicle speeds in the target lane is responsible for the phase transition F→S or S→J. In contrast, lane changing to a faster lane can lead to the phase transition S→F or J→S (Kerner and Klenov, 2009).

Although the three phase traffic flow theory was originally introduced to predict and explain the empirical spatiotemporal features of traffic breakdown and the resulting traffic congestion (Kerner, 2004), it also has potential to explain the safety performance that is associated with freeway traffic. For example, the abrupt speed change and small safety gaps in phase transitions may indicate hazardous traffic conditions under which crashes are more likely to occur. In past, significant efforts have been devoted to establishing a relationship between the behavioral characteristics of drivers and the mechanism of phase transitions. The transitions between three phases were studied to reveal the relationship between empirical spatiotemporal features of traffic flow and the traffic behaviors. Combined with the crash risk prediction models, the results of previous traffic flow studies can help us better understand the mechanism of crashes in each traffic flow state. In addition, the research will also provide quantitative results regarding the safety performance associated with different traffic phases and phase transitions. The research results can thus be considered a supplement to the previous studies on three phase traffic theory in which safety has not been fully considered.

2. Literature review

2.1. Freeway traffic flow and safety

Understanding to what extend freeway traffic flow affects safety has been an emerging topic for the past ten years. Using high resolution loop detector data, Golob and Recker classified traffic flow conditions prior to crashes into different traffic states with distinct crash characteristics. It was found that each of the traffic flow state has a unique profile in terms of the type of crashes that are most likely to occur. In general, the single-vehicle and property damage only crashes are more likely to occur in the traffic states with low density, while the multi-vehicle and injury crashes are more likely to occur in the traffic sates with high density (Golob and Recker, 2004).

A number of studies have developed crash risk prediction models to quantitatively evaluate the impacts of traffic flow conditions on crash likelihood (Abdel-Aty and Rajashekar, 2006; Ahmed and Abdel-Aty, 2012; Hossain and Muromachi, 2011, 2012; Pande et al., 2011; Xu et al., 2013a,b). The central idea was to estimate the relative risks of crash occurrences given traffic flow data before crashes and under normal traffic conditions. Matched case-control design and unmatched case-control design have been two dominant approaches for modeling the risks of crash occurrences. With the matched case-control design, the non-crash cases (controls) are matched with crash cases (cases) according to some confounding factors such as the time and the locations of crashes, while with the unmatched case-control design the control samples were randomly selected.

The purpose of using the case-control study design was to account for the impacts of confounding factors. Recent studies suggested that both matched and unmatched case-control study design control for the impacts of confounding variables. The major difference is that the matched case-control study account for the impacts of confounding factors at the stage of selecting controls; while the unmatched case-control study takes into account confounding factors at the stage of data analysis (Rothman and Greenland, 1998; Bruce et al., 2008).

In some recent studies, artificial intelligent (AI) models, such as the artificial neural networks (Pande et al., 2011), the genetic programming (Xu et al., 2013a), the Bayesian networks (Hossain and Muromachi, 2012), and the classification and regression tree model (Hossain and Muromachi, 2011) were used for estimating crash likelihood. As compared with traditional regression models, the AI models do not require parametric assumptions about the distribution of data and a well-defined functional form between crash probability and explanatory variables. However, the AI models are complex to estimate, and work as black boxes. As a result, they cannot be directly used to identify the relationships between crash likelihood and various traffic flow variables.

2.2. Heterogeneity in effects of variables

The impacts of unobserved heterogeneity should be carefully considered when crash risk models are developed. However, this issue has been largely ignored in previous studies. Theoretically, the available explanatory variables account for only part of the variance in crash likelihood. Various unobserved variables, such as lane closures, work zones, design features, and driver behaviors, can introduce heterogeneity and change the impacts of explanatory variables. Ignoring unobserved heterogeneity may lead to inconsistent and biased estimation results and wrong inferences concerning crash prevention strategies to follow (Washington et al., 2003).

In some early studies, the fixed or the random effects specification was used to account for the heterogeneity (Shankar et al., 1998; Derrig et al., 2002). In the fixed effects specification, a dummy or intercept variable is used for each individual to estimate the individual specific effects; while in the random effects specification, the individual specific effects are assumed to follow a specific distribution. Instead of estimating the intercept for each individual, only the mean and the variance of the assumed distribution are estimated in the random effects model.

Some recent studies suggested that unobserved heterogeneity pertain potentially to all the parameters in a model (Munger et al., 2012; Park et al., 2010; Xiong and Mannering, 2013). Accordingly, the discrete and continuous mixing distributions have been used for describing the unobserved heterogeneity. The discrete mixing distributions lead to finite mixture models, while the continuous mixing distributions lead to random parameters models. In the random parameters models, the regression parameters are assumed to follow some continuous distribution, and allowed to be different across observations (Munger et al., 2012).

3. Data sources

This study used crash and traffic flow data collected from three freeway sections in the state of California, United States. More specifically, data were collected on a 22-mile segment on the I-880N freeway in 2010, a 20-mile segment on the I-880S freeway in 2008 and an 18-mile segment on I-5N freeway in 2009. The spacing between the loop detector stations along the selected three freeway sections ranges from 0.43 to 0.61 miles with an average of 0.5 miles. Both the crash and traffic flow data were obtained from the Highway Performance Measurement System (PeMS) maintained by the California Department of Transportation. In total, 1386 crash observations were identified on the three freeway sections, including 509 crash observations on the I-880N freeway segment, 458 crash observations on the I-880S freeway segment, and 419 crash observations on the I-5N freeway segment. The data from the I-880N freeway segment were used to identify the safety performance of various traffic states, and to develop real-time crash risk models. The data from the other two freeway segments were used to test for the validity of research findings.

The PeMS database provides 30-s raw loop detector data, including vehicle count, vehicle speed, and detector occupancy. The traffic data were collected from the loop detector station that is nearest to crashes (see Fig. 1). Previous studies suggested that to clearly identify traffic phase transitions approximately 15-min traffic flow data are needed (Neubert et al., 1999; Kerner, 2009; Kerner et al., 2002). The same procedure was followed in this study. For each crash, the authors collected 15-min traffic data ending at five minutes prior to crash occurrences. For example, if a crash occurred at 15:00 pm, the traffic data were extracted from 14:40 to 14:55 pm. The purpose of omitting the data from 14:55 to 15:00 pm was to compensate for any inaccuracies in the reported crash occurrence time (Golob and Recker, 2004). It also helps to identify hazardous traffic conditions ahead of crash occurrence time to make preemptive measures possible (Abdel-Aty and Rajashekar, 2006; Pande et al., 2011; Hossain and Muromachi, 2011; Xu et al., 2013a).

The present study is based on a matched case-controlled structure in which the crash cases are the traffic data before crash occurrences, while the non-crash cases, i.e., the controlled cases are the paired traffic data in crash free conditions. Previous studies suggested that the statistical power is negligible by using a control-to-case ratio beyond 4:1 (Ahrens and Pigeot, 2005; Rothman and Greenland, 1998). Thus, the control-to-case ratio of 4:1 was used in this study. For each crash case, the authors randomly selected four paired observations of the non-crash traffic data on the basis of three matching factors, including the time, the location, and the season. For example, the crash No.989 occurred at post-mile 26.84 at 13:15 pm on July 20, 2010. Traffic data taken at the nearest detector station from 12:55 pm to 13:10 pm on July 20, 2010 were included in the crash cases as an observation. Then the paired crash free traffic data taken at the same loop detector station during the same period on four randomly selected crash-free days in the same season were used as four observations in the non-crash cases.

The extracted 30-s raw detector data for crash and non-crash cases were aggregated into 5-min intervals and converted into the 25 traffic flow variables presented in Table 1. The average and standard deviation of traffic flow variables were aggregated across lanes. The autocorrelation and cross correlation were calculated on the basis of 30-s traffic data aggregated over lanes. Aggregating traffic data across lanes may lose important information about lane-by-lane variations. Thus, the authors also calculated the average absolute differences in 30-s traffic variables between adjacent lanes in fifteen minutes to take into account lane-by-lane variations. The average time headway was calculated as 3600 divided by average flow rate across lanes in vehicles per hour. The average distance headway was calculated as the product of average time headway and average vehicle speed.

Loop detectors sometimes suffer from intermittent hardware problems and other random errors, resulting in invalid traffic data. To obtain complete and accurate loop detector data, the records with missing values were discarded. In addition, traffic data were considered invalid or not usable under one or more of the following conditions: (1) the average speed was greater than 0 mile/h (mph) and the flow rate was 0 vehicle/h (vph); (2) the average occupancy was greater than 100%; (3) the occupancy was greater than 0% and the flow rate was 0 vph; (4) the average speed was greater than 100 mph; or (5) the flow rate was greater than 0 vph and the occupancy was 0%.

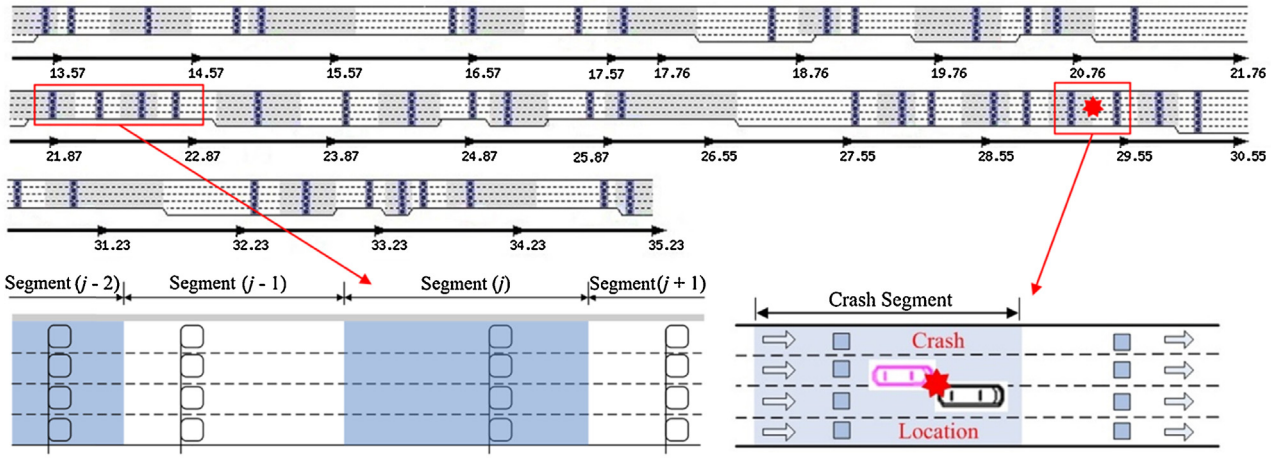

Fig. 1. Locations of loop detectors on the I-880N freeway and definition of crash segment.

Table 1
Description of variables.

Variable	Description
AvgCnt	Average vehicle count during 15-min period (veh/30 s)
AvgSpd	Average vehicle speed during 15-min period (mile/h)
AvgOcc	Average detector occupancy during 15-min period (%)
DevCnt	Std. dev. of vehicle count during 15-min period (veh/30 s)
DevSpd	Std. dev. of vehicle speed during 15-min period (mile/h)
DevOcc	Std. dev. of detector occupancy during 15-min period (%)
CovCnt	Coefficient of variation of count during 15-min period (veh/30 s)
CovSpd	Coefficient of variation of speed during 15-min period (mile/h)
CovOcc	Coefficient of variation of occupancy during 15-min period (%)
AcrCnt	First order autocorrelation of count during 15-min period
AcrSpd	First order autocorrelation of speed during 15-min period
AcrOcc	First order autocorrelation of occupancy during 15-min period
CrrCnt	Cross correlation of count between leftmost and rightmost lane during 15-min
CrrSpd	Cross correlation of speed between leftmost and rightmost lane during 15-min
CrrOcc	Cross correlation of occupancy between leftmost and rightmost lane during 15-min
DifCnt	The difference in 5-min average vehicle count between 5–10 min and 15–20 min time interval (veh/30 s)
DifSpd	The difference in 5-min average vehicle speed between 5–10 min and 15–20 min time interval (mile/h)
DifOcc	The difference in 5-min average detector occupancy between 5–10 min and 15–20 min time interval (%)
LDifCnt	Average absolute difference in vehicle counts between adjacent lanes (veh/30 s)
LDifSpd	Average absolute difference in vehicle speeds between adjacent lanes (mile/h)
LDifOcc	Average absolute difference in detector occupancies between adjacent lanes (%)
Avgtimehw	Average vehicle time headway during 15-min period (s)
Avgdisthw	Average vehicle distance headway during 15-min period (mile)
Devtimehw	Std. dev. of average vehicle time headway during 15-min period (s)
Devdisthw	Std. dev. of average vehicle distance headway during 15-min period (mile)

4. Methodology

The statistical methods used in this study are briefly discussed. A Bayesian conditional logit model was developed to estimate the relative safety performance associated with various traffic phases and phase transitions. The random forest technique was then used to explore the traffic flow variables that contribute to crash occurrences in various traffic phases and phase transitions. Finally, the random-parameter logit models were developed to estimate the crash likelihood with consideration of varying contributing factors across different traffic states.

4.1. Bayesian conditional logit model

The conditional logit model was used to quantitatively evaluate the relative safety performance of various traffic phases while controlling for the effects of other confounding variables, such as weather, geometric features, pavement conditions, etc. The model can be expressed as (c.f., Bruce et al., 2008):

$$y_{ijk} \sim \text{Bernoulli}(p_{ijk}) \tag{1}$$

$$P(y_{itk}) = \frac{1}{\left\{1 + \exp\left[-\alpha_i + \sum_{k=1}^{K} \beta_k x_{ijk}\right]\right\}} \tag{2}$$

where x_{ijk} represents the value of the kth unmatched variable for the case ($j=0$) or the jth control in the ith matched set. Thus, $\boldsymbol{X} = \{x_{ijk}\}$ includes all the cases and controls over all matched sets, where $i = 1, 2, \ldots, I$; $j = 0, 1, \ldots, J$; $k = 1, 2, \ldots, K$. I is the total number of matched sets; J is the number of controls in each matched set; and K is the number of explanatory variables. The α_i represents the effects of matching variables on crash likelihood for each matched set; β_k denotes the estimated coefficients for explanatory variables; and x_k represents the unmatched explanatory variables included in the model.

To account for the selection bias introduced by the matched case-control design, a conditional likelihood needs to be developed. The conditional probability that the first vector of the explanatory variables $\boldsymbol{x}_{i0}$ in the ith matched set corresponds to the case, conditional on $\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ij}$ being the vectors of explanatory variables in the ith matched set, is givens as:

$$P_i^c = \frac{\exp\left(\sum_{k=1}^{K} \beta_k x_{i0k}\right)}{\exp\left(\sum_{k=1}^{K} \beta_k x_{i0k}\right) + \sum_{j=1}^{J} \exp\left(\sum_{k=1}^{K} \beta_k x_{ijk}\right)} \tag{3}$$

Thus, the likelihood function of the conditional logit can be written as:

$$\begin{aligned} f(Y|\boldsymbol{\beta}) &= \prod_{i=1}^{I} f(y_{i0} = 1|\boldsymbol{\beta}) \\ &= \prod_{i=1}^{I} P_i^c = \exp\left\{\sum_{i=1}^{I}\sum_{k=1}^{K}(\beta_k x_{i0k}) \right. \\ &\quad \left. - \sum_{i=1}^{I} \log\left[\sum_{j=0}^{J} \exp\left(\sum_{k=1}^{K} \beta_k x_{ijk}\right)\right]\right\} \end{aligned} \tag{4}$$

This study used the Bayesian inference approach based on Markov Chain Monte Carlo (MCMC) method for the specification of the conditional logit model. In contrast to the point estimations in the traditional maximum likelihood estimation (MLE) approach, the Bayesian modeling technique considers and characterizes all unknown parameters as random variables under a prior distribution. The estimates of the mean, standard deviation, and quartiles of the coefficients can be determined by the posterior distribution. Based on the Bayes' theorem, the posterior distribution of parameters can be estimated as:

$$f(\boldsymbol{\beta}|\boldsymbol{Y}) = \frac{f(\boldsymbol{Y}, \boldsymbol{\beta})}{f(\boldsymbol{Y})} = \frac{f(\boldsymbol{Y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})}{\int f(\boldsymbol{Y}, \boldsymbol{\beta}) d\boldsymbol{\beta}} \propto f(\boldsymbol{Y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}) \tag{5}$$

where $f(\boldsymbol{\beta}|\boldsymbol{Y})$ denotes the posterior joint distribution of parameters $\boldsymbol{\beta}$ conditional upon dataset $\boldsymbol{Y}$. $f(\boldsymbol{Y}, \boldsymbol{\beta})$ represents the joint probability distribution of dataset $\boldsymbol{Y}$ and model parameters $\boldsymbol{\beta}$. $f(\boldsymbol{Y}|\boldsymbol{\beta})$ is the likelihood conditional on model parameters $\boldsymbol{\beta}$. The function $\pi(\boldsymbol{\beta})$ represents the prior distribution of model parameters $\boldsymbol{\beta}$. The non-informative prior distributions were used for the model parameters, which can be specified as:

$$\boldsymbol{\beta} \sim \text{Normal}(\mathbf{0}_K, 10^6 \mathbf{I}_K) \tag{6}$$

where $\mathbf{0}_K$ is a $K \times 1$ vector of zeros; $\mathbf{I}_K$ is a $K \times K$ identity matrices. Based on the specification of the prior distributions for the model parameters $\boldsymbol{\beta}$, the posterior joint distribution $f(\boldsymbol{\beta}|\boldsymbol{Y})$ can be written as:

$$f(\boldsymbol{\beta}|\boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}) = \prod_{i=1}^{I} f(y_{i0}=1|\boldsymbol{\beta}) \times \prod_{k=1}^{K} N\left(\beta_k|u_k, \Sigma_k\right)$$
$$\propto \exp\left\{\sum_{i=1}^{I}\sum_{k=1}^{K}(\beta_k x_{i0k}) - \sum_{i=1}^{N}\log\left[\sum_{j=0}^{J}\exp\left(\sum_{k=1}^{K}\beta_k x_{ijk}\right)\right] - \frac{1}{2}\sum_{k=1}^{K}\frac{(\beta_k)^2}{10^6}\right\} \quad (7)$$

To generate realizations from the posterior joint distribution of the model parameters, the Markov Chain Monte Carlo (MCMC) method was followed to draw parameters sequentially from Eq. (7). Since the conditional distributions in Eq. (7) is nonstandard, the Metropolis-Hasting sampling approach (Hastings, 1970) was used to generate random draws. The inference was made based on the remaining draws after discarding the draws during the burn-in period.

4.2. Random forest

The RF technique was introduced by Breiman in 2001 (Breiman, 2001). It has been considered one of the most efficient methods for evaluating the importance of variables (Harb et al., 2009). Compared with traditional variable selection methods, such as the classification trees and stepwise regression, RF has the capability of handling multi-collinearity problem of candidate variables. It has demonstrated high capability in obtaining unbiased and stable results without a need for a separate cross-validation test data set. Moreover, RF runs efficiently on high dimensional datasets even with thousands of variables.

The RF technique was used to identify the traffic flow variables contributing to the risks of crash occurrences for each traffic state. RF is a machine learning method that consists of an ensemble of randomized classification and regression trees (Breiman, 2001). A predetermined number of classification and regression trees are generated randomly and aggregated to give one single prediction. The RF model chooses the classification with the most votes from all the trees in the forest. During the training procedure, each classification and regression tree is developed through bootstrap sampling by which n samples are selected randomly with replacement from the original training dataset. When building a classification and regression tree, the best split at each node is searched from a randomly selected subset of the whole predictors.

The random forest method uses two measures to evaluate the importance of each variable, including a measure based on the Gini index and the out-of-bag (OOB) error rate. The measure based on the Gini index was used in this study to select the traffic flow variables that contribute to the crash likelihood for different traffic flow phases. The decrease in the Gini index at each node was calculated for the variable that was used to make the split. Then, the importance measure for this variable was given as the average decrease in the Gini index over all trees in the forest. The Gini index for node t in a single tree can be calculated as:

$$G(t) = 1 - \sum_{i=1}^{m} p^2(i|t) \quad (8)$$

where $G(t)$ denotes the Gini index for the node t; i represents the number of classes; and $p(i|t)$ represents the estimated class probabilities.

4.3. Bayesian random parameter logit model

This study focuses on estimating the crash likelihood for each traffic state using traffic flow data only. Other variables, such as weather conditions, geometric features, and driver behaviors were controlled for using the matched case-control study design. However, additional heterogeneity that is caused by observed factors such as drivers' behavior can still be introduced, resulting in biased estimates of crash risk models. Previous studies found that the effects of traffic flow variables on crash likelihood are different across various weather conditions and geometric features (Hossain and Muromachi, 2011; Xu et al., 2013b). It is thus important to consider the varying effects of traffic flow variables on crash likelihood across observations.

To account for the unobserved heterogeneity in the dataset, the Bayesian random parameter logit model was used. The random parameter logit model allows some or all parameters to vary across observations. In this study, the Bayesian inference approach based on MCMC method was applied to estimate the random parameter logit model. As compared with the traditional maximum likelihood method, the Bayesian inference approach has three major advantages. First, the Bayesian inference approach offers a flexible structure to update the posterior distribution when new information is available, which is important for improving the temporal and spatial transferability of a crash risk prediction model. Second, a non-informative prior can prohibit or reduce the possibility of parameters toward near non-identified regions, which can cause the classical methods difficult to convergence (Balcombe et al., 2009). Finally, the Bayesian inference approach can produce the marginal likelihood that can be used for model comparison and testing in a way that cannot be achieved by the classical estimation methods.

The random parameter logit model was developed for each traffic state to account for the within-group variation. As the traffic state (group membership) of each observation is known, separate crash risk prediction models were developed for different traffic states instead of using the mixture modeling approach to account for the between-group variation. Developing separate models can account for the observed heterogeneity across different traffic states, i.e., the varying impacts of traffic variables on crash likelihood across different traffic states. More importantly, developing separate crash risk models is also expected to capture the varying contributing factors to crash likelihood under different traffic states.

The random parameter logit model can be expressed as:

$$z_m = \beta_{0,m} + \boldsymbol{\beta}_{(1),m}\boldsymbol{x}_{(1),m} + \boldsymbol{\beta}_{(2)}\boldsymbol{x}_{(2),m} + \varepsilon_m \quad \varepsilon_m \overset{iid}{\sim} \text{logistic}(0, 1)$$
$$y_m = \begin{cases} 0, & \text{if } z_m \le 0 \\ 1, & \text{if } z_m > 0 \end{cases} \quad (9)$$

where y_m represents the crash indicator, if a crash occurred, $y_m = 1$; otherwise, $y_m = 0$ ($m = 1, 2, \ldots, M$). For each crash indicator y_m, $\boldsymbol{x}_{(1),m} = [x_{(1),1,m}, x_{(1),2,m}, \ldots, x_{(1),K(1),m}]$ is a $1 \times K(1)$ vector of contributing factors that can be denoted by $x_{(1),k,m}$ ($k = 1, 2, 3, \ldots, K(1)$). The vector $\boldsymbol{\beta}_{(1),m} = [\beta_{(1),1,m}, \beta_{(1),2,m}, \ldots, \beta_{(1),K(1),m}]^T$ is the parameter vector for the contributing factor vector $\boldsymbol{x}_{(1),m}$. These parameters were assumed to vary across observations. The $\boldsymbol{x}_{(2),m} = [x_{(2),1,m}, x_{(2),2,m}, \ldots, x_{(2),K(2),m}]$ is a $1 \times K(2)$ vector of other contributing factors that can be denoted as $x_{(2),k,m}$ ($k = 1, 2, 3, \ldots, K(2)$). The vector $\boldsymbol{\beta}_{(2)} = [\beta_{(2),1}, \beta_{(2),2}, \ldots, \beta_{(2),K(2)}]^T$ is the parameter vector for the contributing factor vector $\boldsymbol{x}_{(2),m}$. Each parameter in the $\boldsymbol{\beta}_{(2)}$ was assumed to be fixed across observations. The constant $\beta_{0,m}$ is also assumed to vary across observations.

The random parameter $\boldsymbol{\beta}_{(1),m}$ is assumed to be normally distributed as $\boldsymbol{\beta}_{(1),m} \sim N(\boldsymbol{u}_{(1)}, \boldsymbol{\Sigma}_{(1)})$ with $\boldsymbol{u}_{(1)} = [u_{(1),1}, u_{(1),2}, \ldots, u_{(1),K(1)}]^T$

and $\boldsymbol{\Sigma}_{(1)} = \mathrm{diag}[\Sigma_{(1),1}, \Sigma_{(1),2}, \ldots, \Sigma_{(1),K(1)}]$. The full data likelihood of the random parameter logit model is given as:

$$f(Y|\boldsymbol{\Theta}) = \prod_{m=1}^{M} f\left(y_m|\beta_{0,m}, \boldsymbol{\beta}_{(1),m}, \boldsymbol{\beta}_{(2)}\right) = \prod_{m=1}^{M}\left[\frac{e^{\beta_{0,m}+\boldsymbol{\beta}_{(1),m}\boldsymbol{x}_{(1),m}+\boldsymbol{\beta}_{(2)}\boldsymbol{x}_{(2),m}}}{1+e^{\beta_{0,m}+\boldsymbol{\beta}_{(1),m}\boldsymbol{x}_{(1),m}+\boldsymbol{\beta}_{(2)}\boldsymbol{x}_{(2),m}}}\right]^{y_m} \times \left[\frac{1}{1+e^{\beta_{0,m}+\boldsymbol{\beta}_{(1),m}\boldsymbol{x}_{(1),m}+\boldsymbol{\beta}_{(2)}\boldsymbol{x}_{(2),m}}}\right]^{1-y_m} \quad (10)$$

where $\boldsymbol{\Theta}$ represents the vector of all parameters, including the random parameter vector $\boldsymbol{\beta}_{(1)}$, the fixed parameter vector $\boldsymbol{\beta}_{(2)}$, the random constant $\boldsymbol{\beta}_0$, the random parameters mean vector $\boldsymbol{\mu}_{(1)}$, the random parameters variance vector $\boldsymbol{\Sigma}_{(1)}$, the random constant mean μ_0, and the random constant variance Σ_0. Thus,

$$\Theta = \left[\boldsymbol{\beta}_0, \boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)}, \mu_0, \sum\nolimits_0, \boldsymbol{\mu}_{(1)}, \sum\nolimits_{(1)}\right] \quad (11)$$

The prior distribution $\pi(\boldsymbol{\Theta})$ reflects the prior knowledge of the model parameters $\boldsymbol{\Theta}$. The non-informative priors that are diffuse with large variance are used to 'let the data speak for themselves', so that the inferences are unaffected by the information external to the current data (Gelman et al., 2004). The prior distributions for all parameters $\boldsymbol{\Theta}$ are specified as:

$$\begin{aligned}&\beta_{0,m}\sim N\left(\mu_0, \sum\nolimits_0\right), \quad \boldsymbol{\beta}_{(1),m}\sim N\left(\boldsymbol{\mu}_{(1)}, \sum\nolimits_{(1)}\right), \quad \boldsymbol{\beta}_{(2)}\sim N\left(\bar{\boldsymbol{\mu}}_{(2)}, \bar{\sum}_{(2)}\right)\\ &\mu_0\sim N(\bar{a}_0, \bar{b}_0), \quad \sum\nolimits_0\sim IG(\bar{c}_0, \bar{d}_0)\\ &\mu_{(1),k}\sim N(\bar{a}_{(1),k}, \bar{b}_{(1),k}), \quad \sum\nolimits_{(1),k}\sim IG(\bar{c}_{(1),k}, \bar{d}_{(1),k}), \quad k=1,2,\ldots,K(1)\end{aligned} \quad (12)$$

where N represents the normal distribution; and IG represents the inverse gamma distribution. The priors of the fixed parameters vector and the random parameters mean vector follow normal distributions, and the random parameters variances follow inverse gamma distributions. The parameters with over lines denote the hyper-parameters that are set as:

$$\begin{aligned}&\bar{\boldsymbol{\mu}}_{(2)} = \mathbf{0}_{K(2)}, \quad \bar{\sum}_{(2)} = 10^6\mathbf{I}_{K(2)}\\ &\bar{a}_0 = \bar{a}_{(1),k} = 0, \quad \bar{b}_0 = \bar{b}_{(1),k} = 10^6, \quad k=1,2,\ldots,K(1)\\ &\bar{c}_0 = \bar{c}_{(1),k} = 0.001, \quad \bar{d}_0 = \bar{d}_{(1),k} = 0.001, \quad k=1,2,\ldots,K(1)\end{aligned} \quad (13)$$

where $\mathbf{0}_{K(2)}$ is a $K(2)\times 1$ vector of zeros; $\mathbf{I}_{K(2)}$ is a $K(2)\times K(2)$ identity matrices. Based on the specification of the prior distributions for the model parameters $\boldsymbol{\Theta}$, the posterior joint distribution $f(\boldsymbol{\Theta}|\boldsymbol{Y})$ can be written as:

$$\begin{aligned}f(\boldsymbol{\Theta}|\boldsymbol{Y}) &\propto f(\boldsymbol{Y}|\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})\\ &= \prod_{m=1}^{M} f(y_m|\beta_{0,m}, \beta_{(1),m}, \beta_{(2)}) \times \prod_{m=1}^{M} N\left(\beta_{0,m}|\mu_0, \sum\nolimits_0\right)\\ &\times \prod_{k=1}^{K(1)}\prod_{m=1}^{M} N\left(\beta_{(1),k,m}|\mu_{(1),k}, \sum\nolimits_{(1),k}\right)\\ &\times \prod_{k=1}^{K(2)} N\left(\beta_{(2),k}|\bar{\mu}_{(2),k}, \bar{\sum}_{(2),k}\right) \times N(\mu_0|\bar{a}_0, \bar{b}_0) \times IG\left(\Sigma_0|\bar{c}_0, \bar{d}_0\right)\\ &\times \prod_{k=1}^{K(1)} N(\mu_{(1),k}|\bar{a}_{(1),k}, \bar{b}_{(1),k}) \times \prod_{k=1}^{K(1)} IG\left(\Sigma_{(1),k}|\bar{c}_{(1),k}, \bar{d}_{(1),k}\right)\end{aligned} \quad (14)$$

The conditional posterior distributions of all components of vector $\boldsymbol{\Theta}$ are proportional to the joint distribution $f(\boldsymbol{Y}, \boldsymbol{\Theta})$, specified by Eq. (14). Before implementing the MCMC sampling algorithm, the conditional posterior distribution for each parameter needs to be derived. The authors only presented how to derive the conditional posterior distributions for the parameters $\boldsymbol{\mu}_{(1)}$, $\boldsymbol{\Sigma}_{(1)}$, and $\boldsymbol{\beta}_{(1)}$. The other parameters can be derived using the similar methods. The authors compute the conditional posterior for the means of the random parameters $\boldsymbol{\mu}_{(1)}$ as follows:

$$\begin{aligned}&f(\mu_{(1),k}|\boldsymbol{Y}, \boldsymbol{\Theta}\backslash\mu_{(1),k})\\ &\propto \prod_{m=1}^{M} N\left(\beta_{(1),k,m}|\mu_{(1),k}, \sum\nolimits_{(1),k}\right) \times N(\mu_{(1),k}|\bar{a}_{(1),k}, \bar{b}_{(1),k})\\ &\propto \exp\left\{\left(-\frac{1}{2}\right)\left[\left(u_{(1),k} - \frac{\sum_{(1),k}\bar{a}_{(1),k} + \bar{b}_{(1),k}\sum_{m=1}^{M}\beta_{(1),k,m}}{\sum_{(1),k} + M\bar{b}_{(1),k}}\right)^2 \Big/ \frac{\sum_{(1),k}\bar{b}_{(1),k}}{\sum_{(1),k} + M\bar{b}_{(1),k}}\right]\right\}\\ &\propto N\left(\frac{\sum_{(1),k}\bar{a}_{(1),k} + \bar{b}_{(1),k}\sum_{m=1}^{M}\beta_{(1),k,m}}{\sum_{(1),k} + M\bar{b}_{(1),k}}, \frac{\sum_{(1),k}\bar{b}_{(1),k}}{\sum_{(1),k} + M\bar{b}_{(1),k}}\right)\end{aligned} \quad (15)$$

The conditional posterior for the covariance of the random parameters $\boldsymbol{\Sigma}_{(1)}$ is obtained as follows:

$$\begin{aligned}f\left(\sum\nolimits_{(1),k}|\boldsymbol{Y}, \boldsymbol{\Theta}\backslash\sum\nolimits_{(1),k}\right) &\propto \prod_{m=1}^{M} N\left(\beta_{(1),k,m}|\mu_{(1),k}, \sum\nolimits_{(1),k}\right)\\ &\times IG\left(\sum\nolimits_{(1),k}|\bar{c}_{(1),k}, \bar{d}_{(1),k}\right) \propto \frac{1}{\left(\sum_{(1),k}\right)^{M/2+\bar{c}_{(1),k}+1}}\\ &\times \exp\left[-\frac{\sum_{m=1}^{M}(\beta_{(1),k,m} - \mu_{(1),k})^2 + 2\bar{d}_{(1),k}}{2\sum_{(1),k}}\right]\\ &\propto IG\left(\frac{M}{2} + \bar{c}_{(1),k}, \frac{\sum_{m=1}^{M}(\beta_{(1),k,m} - \mu_{(1),k})^2 + 2\bar{d}_{(1),k}}{2}\right)\end{aligned} \quad (16)$$

The conditional posterior for the parameter vector $\boldsymbol{\beta}_{(1)}$ is obtained as follows:

$$\begin{aligned}f(\boldsymbol{\beta}_{(1)}|\boldsymbol{Y}, \boldsymbol{\Theta}\backslash\boldsymbol{\beta}_{(1)}) &\propto \prod_{m=1}^{M} f(y_m|\beta_{0,m}, \boldsymbol{\beta}_{(1),m}, \boldsymbol{\beta}_{(2)})\\ &\times \prod_{k=1}^{K(1)}\prod_{m=1}^{M} N\left(\beta_{(1),k,m}|\mu_{(1),k}, \sum\nolimits_{(1),k}\right)\\ &\propto \exp\left\{\sum_{m=1}^{M}\beta_{0,m}y_m + \sum_{m=1}^{M}\boldsymbol{\beta}_{(1),m}\boldsymbol{x}_{(1),m}y_m + \sum_{m=1}^{M}\boldsymbol{\beta}_{(2)}\boldsymbol{x}_{(2),m}y_m\right.\\ &- \sum_{m=1}^{M}\log[1 + e^{\beta_{0,m}+\boldsymbol{\beta}_{(1),m}\boldsymbol{x}_{(1),m}+\boldsymbol{\beta}_{(2)}\boldsymbol{x}_{(2),m}}]\\ &\left.- \frac{1}{2}\sum_{k=1}^{K(1)}\sum_{m=1}^{M}\frac{(\beta_{(1),k,m} - \mu_{(1),k})^2}{\Sigma_{(1),k}}\right\}\end{aligned} \quad (17)$$

Once all the full conditional posterior distributions are derived, the following MCMC sampling algorithm can be used to simulate the posterior joint parameter distribution:

(i) Generate μ_0 from $f(\mu_0|\boldsymbol{Y}, \boldsymbol{\Theta}\backslash\mu_0)$:
(ii) Generate $\mu_{(1),k}$ from $f(\mu_{(1),k}|\boldsymbol{Y}, \boldsymbol{\Theta}\backslash\mu_{(1),k})$ $(k = 1, 2, \ldots, K(1)$:
(iii) Generate Σ_0 from $f(\Sigma_0|\boldsymbol{Y}, \boldsymbol{\Theta}\backslash\Sigma_0)$:
(iv) Generate $\Sigma_{(1),k}$ from $f(\Sigma_{(1),k}|\boldsymbol{Y}, \boldsymbol{\Theta}\backslash\Sigma_{(1),k})$ $(k = 1,2, \ldots, K(1))$:
(v) Generate $\boldsymbol{\beta}_0$ from $f(\boldsymbol{\beta}_0|\boldsymbol{Y}, \boldsymbol{\Theta}\backslash\boldsymbol{\beta}_0)$:
(vi) Generate $\boldsymbol{\beta}_{(1)}$ from $f(\boldsymbol{\beta}_{(1)}|\boldsymbol{Y}, \boldsymbol{\Theta}\backslash\boldsymbol{\beta}_{(1)})$:
(vii) Generate $\boldsymbol{\beta}_{(2)}$ from $f(\boldsymbol{\beta}_{(2)}|\boldsymbol{Y}, \boldsymbol{\Theta}\backslash\boldsymbol{\beta}_{(2)})$:

The models were specified by repeating steps (i)–(vii) until the predetermined maximum iteration times of the MCMC algorithm were reached. The first four steps were performed through the Gibbs sampling approach (Geman and Geman, 1984). The conditional distributions in the last three steps (v)–(vii) are nonstandard. Thus the Metropolis-Hasting sampling approach was used to generate random draws (Hastings, 1970). The inference was then made based on the remaining draws after discarding the draws during the burn-in period.

The test of significance for the standard deviation is the most commonly used method to determine whether a random parameter should be used. However, this method cannot be used in this study, because the variance that follows an inverse-gamma distribution is always positive. The Bayes factor analysis was then used to determine whether a random parameter should be used. The Bayes factor of model M1 and M2 is defined as the ratio of the marginal likelihood of these two models. Previous studies suggested that the fitness of M1 is significantly better than that of M2 when the Bayes factor is greater than three (Kass and Raftery, 1995). We calculated the Bayes factors of the models with and without a random parameter for each of the variables. A random parameter is used if the Bayes factor is greater than three.

5. Results of data analyses

5.1. Identification of traffic states

According to the three-phase traffic theory, freeway traffic can be classified into three phases, including free flow (F), synchronized flow (S), and wide moving jams (J), as well as four transitional states, including the transition from free flow to synchronized flow (F→S), the transition from synchronized flow to free flow (S→F), the transition from synchronized flow to wide moving jams (S→J), and the transition from wide moving jams to synchronized flow (J→S). The traffic states can be identified given the traffic flow characteristics measured from loop detector stations.

The free-flow phase is characterized by high vehicle speeds and low traffic density. Free flow can be easily distinguished from congested flow using the time series plot of speed and occupancy. Fig. 2 illustrates a time series of 30-s aggregates of the speed and occupancy at a loop detector station. The transitional state from free to the synchronized flow and the wide moving jams, can be identified by a sudden change in both speed and occupancy. Theoretically, freeway traffic may be transferred from free flow to either synchronized flow or wide moving jams. However, previous studies already demonstrated that the wide moving jams generally do not emerge with the free flow phase (Kerner, 2004). That is, the transitional state F→J cannot be easily observed. Thus, the transitional state F→J was not considered in the present study.

In the three phase traffic theory, the synchronized flow and wide moving jams are two phases of congested flow. In the synchronized state, the measurements of flow rate and density exhibit an irregular pattern in the fundamental diagram. The pattern cannot be described by the functional relationship defined in the classical traffic flow theory. The correlation between density and flow rate was then used to identify the synchronized flow, wide moving jams, and the transitional states between them (Neubert et al., 1999; Knospe et al., 2002). Fig. 3 illustrates the correlation between density and flow rate in congested flow during randomly selected two hours. The synchronized flow was characterized by weak correlations between flow rate and density with a correlation parameter lower than 0.2. Accordingly, the wide moving jams were identified by strong correlation between density and flow rate with a correlation parameter greater than 0.5.

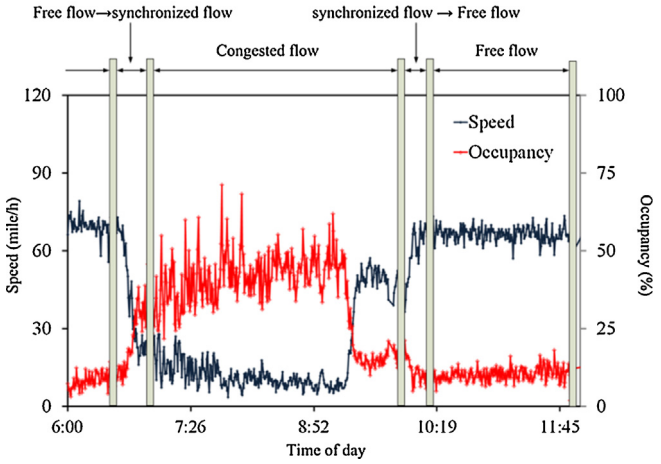

Fig. 2. Time series plot of 30-s aggregates of speed and occupancy.

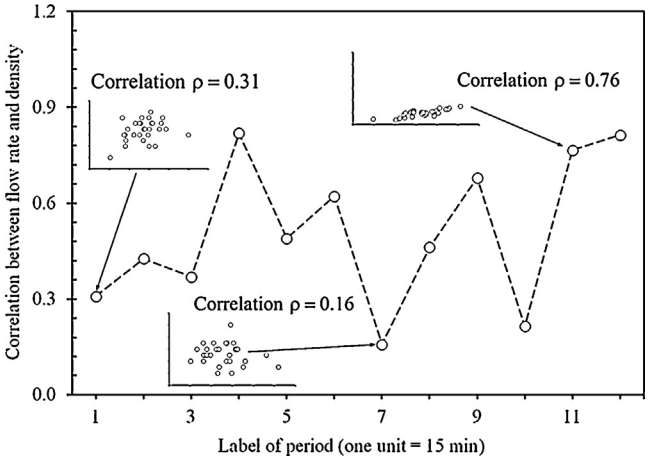

Fig. 3. Correlation between density and flow rate in congested flow.

The transitions between synchronized flow and wide moving jams were characterized by intermediate correlation values between 0.2 and 0.5 (Neubert et al., 1999; Knospe et al., 2002). In this condition the direction of phase transition was determined by the change in speed. More specifically, the reduction in speed over time was considered an indicator for the transitional state from synchronized flow to wide moving jams (S→J), and vice versa. Note that the identification of traffic states followed the procedures suggested by previous studies (Neubert et al., 1999; Knospe et al., 2002; Kerner, 2004). On the basis of these rules, the traffic states for each crash and non-crash case in the three freeway datasets were identified. The distributions of crash and non-crash cases across traffic states are summarized in Table 2.

5.2. Safety performance of seven traffic states

A Bayesian conditional logit model was developed using both case and control samples from the I-880N freeway to evaluate the relative safety performance of each traffic state. The seven traffic states were incorporated in the model with six indicator variables

Table 2
Distribution of crash and non-crash cases for different traffic states.

Traffic state	Non-crash		Crash		Total	
	Frequency	Percentage (%)	Frequency	Percentage (%)	Frequency	Percentage (%)
Free Flow (F)	3998	72.1	669	48.3	4667	67.3
F→S	229	4.1	164	11.8	393	5.7
S→F	188	3.4	57	4.1	245	3.5
Synchronized Flow (S)	408	7.4	156	11.3	564	8.1
S→J	240	4.3	136	9.8	376	5.4
J→S	227	4.1	96	6.9	323	4.7
Wide moving jams (J)	254	4.6	108	7.8	362	5.2
Total	5544	100.0	1386	100.0	6930	100.0

in which the free-flow state was considered the reference level. At this stage the model did not include other explanatory variables such as occupancy and the standard deviation of speed. It was assumed that these parameters were highly correlated with traffic state.

Three parallel MCMC chains were constructed for Bayesian inference. Each MCMC chain consisted of 10,000 iterations, including an initial "burn-in" period of 2000 iterations. The estimations of each parameter from the MLE method were considered initial values. The initial values for multiple MCMC chains were dispersed throughout the 90% confidence intervals of the estimated parameters from the MLE. The convergence of the posterior distribution samples was checked by the visual inspection of the trace plots, posterior density plots, and autocorrelation function plots. In addition, the Gelman-Rubin potential scale reduction (PSR) was also checked. If the PSR was lower than 1.1, the multiple chains were considered converged (Gelman et al., 2004). The estimation results of the Bayesian conditional logit model are given in Table 3. The 95% credible interval for each parameter in Table 3 indicates that the traffic states significantly affect the risks of crash occurrences. The odds ratio for each variable was used to quantify the safety performance of each traffic state.

The model specification results suggest that the safety performance differs significantly across various traffic states. The odds ratios of all indicator variables are significantly greater than one, indicating that the free flow state has the best safety performance in terms of the lowest crash likelihood. The transitional states F→S and S→J have the highest crash likelihood, followed by the wide moving jams. The risks of crash occurrence associated with the transitional state S→F is higher than those in the free flow state, but lower than those in synchronized flow. Similarly, the crash likelihood in the transitional state J→S is greater than those in synchronized flow, but lower than those associated with wide moving jams. The relative safety performance associated with each traffic state is illustrated in Fig. 4.

The wide moving jams are characterized by very high traffic density and moving structures that are bonded by two fronts where vehicle speed changes greatly (Kerner, 2004; Knospe et al., 2002). The very high traffic density leaves less space for taking crash avoidance maneuvers prior to crash occurrence. Besides, the stop-and-go inside the jams can introduce turbulent traffic conditions that may lead to crashes. As a result, the crash likelihood in wide moving jams are higher than those in free flow and synchronized flow. The synchronized flow is characterized by the intermediate traffic density and synchronized speed between vehicles. Accordingly, the crash likelihood associated with synchronized flow are greater than those associated with the free flow phase, but lower than those associated with the wide moving jams.

The transitional state F→S is usually caused by two traffic behaviors occurring within a random local disturbance in free flow in which the speed is suddenly reduced: (1) the speed reduction of preceding vehicle forces the following vehicles to decelerate abruptly (see Fig. 4(d)); and (2) the following vehicle changes lane and forces the vehicle on the adjacent lane to decelerate (see Fig. 4(e)). Considering the fact that the transitional state F→S usually occurs in free-flow with relatively high speed, both traffic behaviors are quite dangerous. The instability and dangerous traffic behaviors cause the crash likelihood associated with the transitional state F→S to be higher than those associated with synchronized flow (see Table 3). The transitional state S→J is caused by considerable speed perturbation in synchronized flow with high traffic density. The abrupt reduction in speed in the transitional state S→J causes the crash likelihood to be larger than those associated with the wide moving jams.

In some cases a vehicle changes to a faster lane, and the lane-change behavior does not result in significant speed reduction. In this condition the lane-change behavior may result in the phase transition from synchronized flow to free flow, or from wide moving jams to synchronized flow (Kerner and Klenov, 2009). Such lane-change behaviors may cause the dissolution of the disturbances that potentially lead to turbulent traffic conditions in synchronized flow or wide moving jams. The crash likelihood associated with the transitional state S→F and J→S are thus lower than those associated with synchronized flow and wide moving jams, respectively.

Table 3
Estimation results of conditional logit model.

Variables	Parameters Estimates				Odds ratios			
	Mean	S.D.	2.5%	97.5%	Mean	S.D.	2.5%	97.5%
Free flow (F)[a]	–	–	–	–	–	–	–	–
F→S	2.221	0.228	1.758	2.657	9.460	2.256	5.803	14.260
S→F	1.238	0.305	0.576	1.797	3.613	1.112	1.780	6.032
Synchronized flow (S)	1.846	0.223	1.411	2.243	6.489	1.410	4.102	9.426
S→J	2.251	0.238	1.722	2.763	9.768	2.338	5.598	15.840
J→S	1.968	0.262	1.460	2.472	7.406	1.926	4.305	11.840
Wide moving jams (J)	2.111	0.258	1.639	2.618	8.535	2.238	5.148	13.710

[a] Free flow (F) is the reference level.

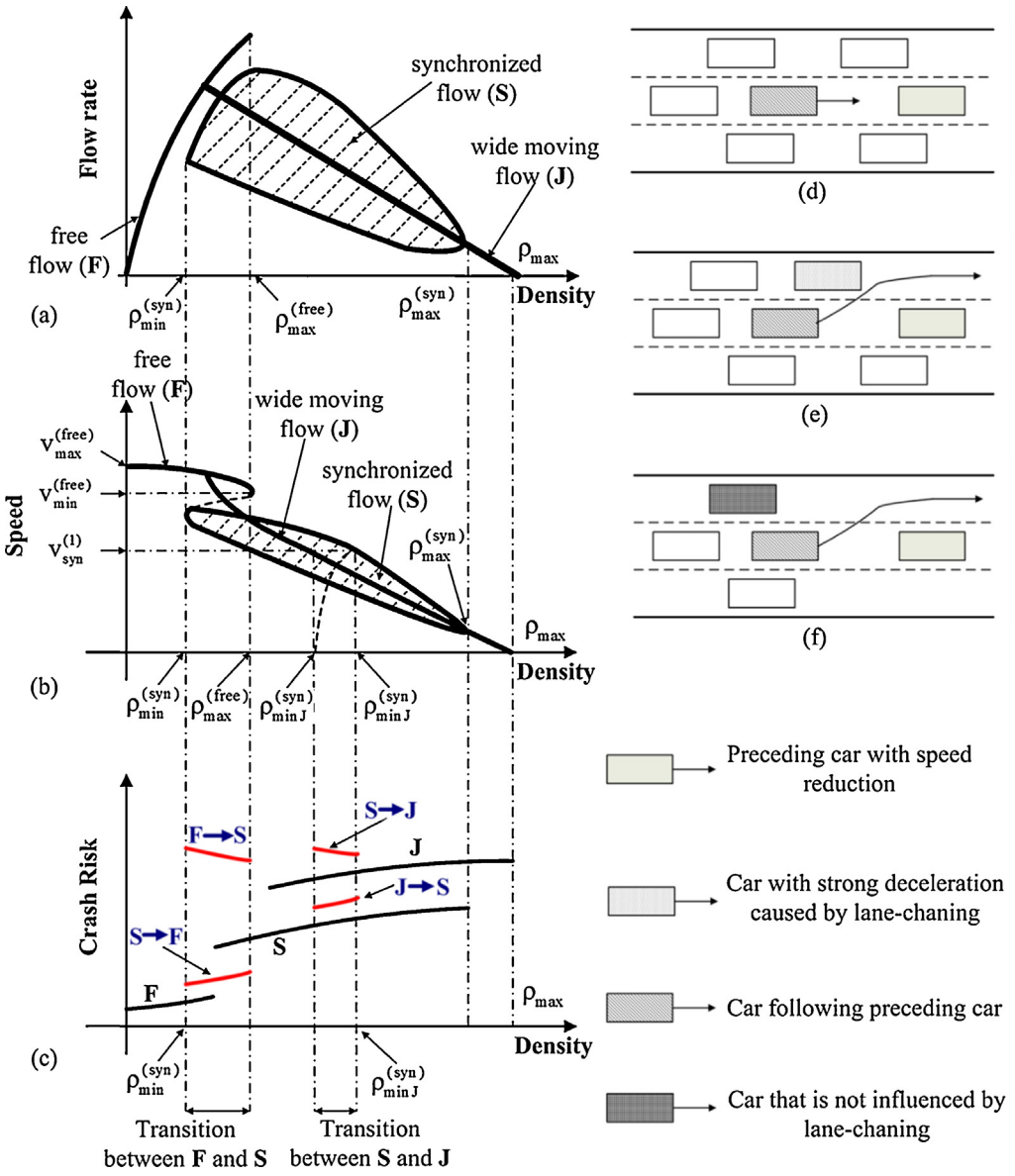

Fig. 4. Safety performance of traffic flow states: (a) Free flow, synchronized flow, and wide moving jams in flow rate and density plane (Kerner, 2004). (b) Traffic states in speed-density plane (Kerner, 2004). (c) Crash likelihood associated with each traffic state. (d) Speed adaptation effect that causes phase transition. (e) Lane changing behavior that causes phase transition F→S or S→J. (f) Lane changing behavior that causes phase transition S→F or J→S.

5.3. Contributing factors to crash likelihood

The RF technique was used for identifying the most important variables that contribute to crash likelihood for various traffic states. The RF analyses were repeated for 100 times using the samples obtained from the I-880N freeway. The average OOB error rates for different numbers of trees were calculated. It was found that with 500 trees a constant minimum error rate was achieved. The number of trees in the forest was then set to be 500. The traffic flow variables given in Table 1 were considered for the RF analyses. The RF analyses were conducted for each traffic state for fifty times. The average normalized variable importance was then calculated. The average normalized variable importance for the top six variables for various traffic states is given in Fig. 5. As expected, the traffic flow variables contributing to crash likelihood are quite different across traffic states.

The difference in traffic flow variables over time and the first order autocorrelation of traffic flow variables are the main factors that contribute to the crash likelihood in free flow (see Fig. 5(a)). The finding is interesting because the autocorrelation of traffic flow variables on short time scales was an indicator for the presence of platoons in traffic flow (Neubert et al., 1999; Knospe et al., 2002). In the three phase traffic theory, free flow represents the traffic flow state before breakdown. The results imply that the free flow state can be further divided into two sub-phases in terms of their safety performance, including a true free flow state in which the interactions between vehicles are minor, and a platooned traffic state in which bunched vehicles travel in successions. In platooned traffic drivers have limited space to take actions to avoid a collision, resulting in higher crash likelihood than those in the true free flow state. This finding is consistent with the classification of traffic flow proposed by Wu in 2002 that the traffic flow in the fluid region can be further divided into a free and a convoy state (Wu, 2002).

For wide moving jams, the difference in detector occupancy between adjacent lanes, the variation in average space

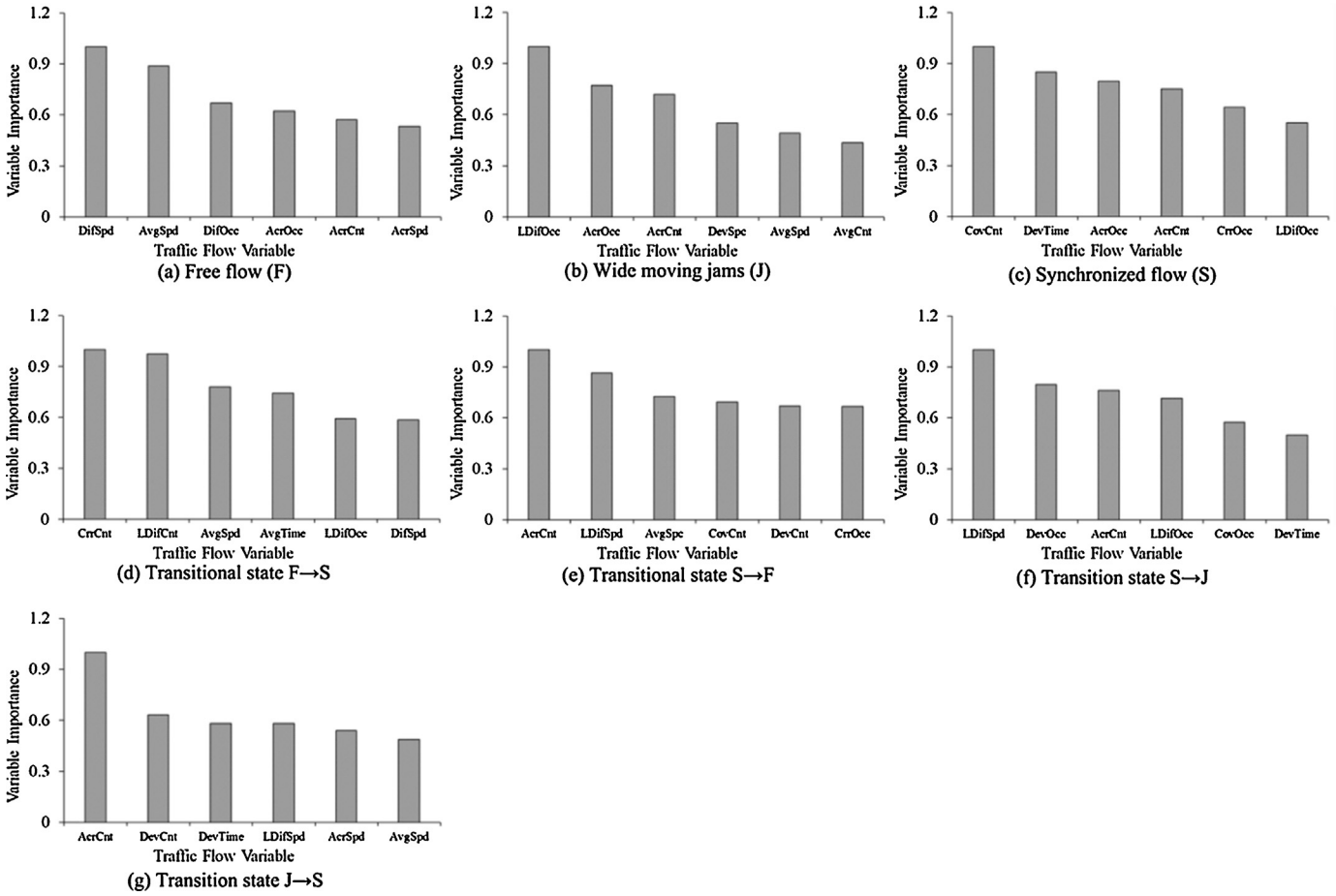

Fig. 5. Normalized variable importance for traffic states.

headway, and the autocorrelation of traffic variables are the main contributing factors to crash likelihood. The variability of average space headway is indicative of the frequency of accelerating and braking in traffic flow, which is highly correlated with crash likelihood. Besides, the imbalance of traffic between lanes may encourage drivers to change lanes more frequently, resulting in increased crash likelihood when density is high. In synchronized flow, the difference in occupancy between adjacent lanes, the variation of vehicle count, and the autocorrelation of traffic flow variables are the main factors for crash likelihood. According to previous studies, there is a tendency to the synchronization of vehicle speeds on each lane and across different lanes in synchronized traffic. The large variability in vehicle count and occupancy between adjacent lanes is indicative of the emergence of narrow jams in which vehicles accelerate and brake frequently. Besides, the autocorrelation of traffic data on short time scale suggests the presence of platooned traffic (Neubert et al., 1999; Knospe et al., 2002).

For the transitional state F→S, the major contributing factors to crash likelihood include the difference in vehicle count and detector occupancy between adjacent lanes, the average time headway and the difference in speed over time. Recall that the phase transition from free flow to synchronized flow is highly correlated with speed adaption and lane changes. The abrupt speed reduction of the preceding vehicle may pose the following vehicles to increased risks of read-end collisions if the time headways between them are small. In addition, the speed fluctuation and imbalanced traffic between lanes may encourage drivers to change lanes more frequently, resulting in increased crash likelihood in dense traffic. For the transitional state S→F, crash likelihood is mainly influenced by the difference in vehicle speed between adjacent lanes, the average space headway, and the variations in vehicle count. The transition from synchronized flow to free flow is mainly caused by the lane-change behaviors which are usually correlated with the difference in speed between adjacent lanes. The risks for sideswipe crashes may arise during lane changes when the space headways between vehicles are small.

The contributing factors to crash likelihood in the transitional state S→J include the difference in vehicle speed and the detector occupancy between adjacent lanes, and the variation of detector occupancy. As mentioned before, the imbalanced traffic between lanes generates more lane-change behaviors, resulting in increased risks of side-swipe crashes when the traffic is unstable. For the transitional state J→S, the risks of crash occurrence are mainly determined by the difference in vehicle speed between adjacent lanes, and the variation of vehicle count. The contributing factors for the crash likelihood in the transitional state J→S are very similar to those in the state S→F. In fact, the mechanism for phase transition is also similar for these two transitional states. That is, both transitional states are caused by the lane-change behaviors which are usually correlated with the difference in speed between adjacent lanes.

5.4. Crash risk prediction models

Crash risk prediction models were developed to quantitatively evaluate the impacts of traffic flow variables on crash likelihood

using the data from the I-880N freeway. A crash risk prediction model was developed for each traffic state to account for the observed heterogeneity across different traffic states. The traffic flow data account for only part of the variance in crash likelihood. To account for the unobserved heterogeneity caused by unobserved influence factors, the Bayesian random parameter logit model was used.

The estimated parameters of the conventional logit model for each traffic state by MLE method were used as initial values. Three parallel MCMC chains that consist of 10,000 iterations were constructed for Bayesian inference. The first 2000 iterations for each chain were discarded as the burn-in period. The PSR of 1.1 was used as the convergence criterion for each random parameter logit model (Gelman et al., 2004). Table 4 presents the estimation results of the Bayesian random parameter logit models for various traffic states. As expected, the contributing factors to crash likelihood for each traffic state in the crash risk prediction model are reasonably similar to those in the RF analyses. The contributing factors to the crash likelihood in various traffic states are quite different, and the same traffic flow variable has distinct effects on crash likelihood across different traffic states.

For comparison, the conventional logit models in which all parameters were fixed across observations were also developed. We compared the crash prediction accuracy of these two models at different false alarm rates (see Fig. 6). With the use of random parameter models, the prediction accuracy on average increases 13% over the conventional logit models. The area under the ROC curve (AUC) was also used for evaluating the prediction performance of the models (Bradley, 1997; Weiss and Provost, 2003; Xu et al., 2013a,b). The AUC values for the random parameters model are about 10% greater than those for the fixed-parameters models. It again confirms the finding that the proposed random parameter logit models are superior to the conventional logit models in predicting crash likelihood.

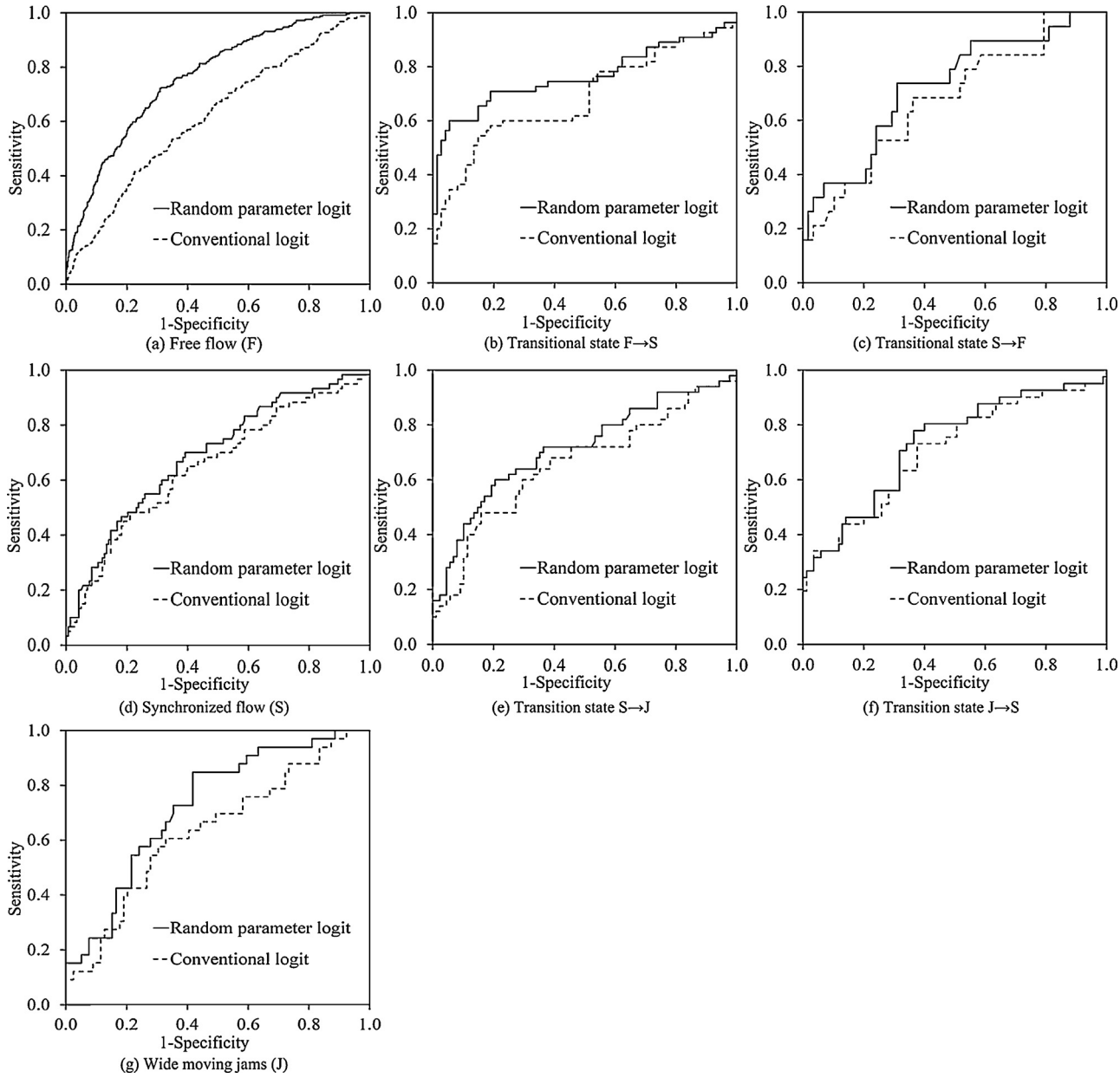

Fig. 6. The ROC curves of the random parameter and conventional logit models.

Table 4
Estimation results of crash risk models for various traffic states using I-880N freeway dataset.

Variable	Mean	2.5%	97.5%
Free flow (F)			
Constant	−2.17	−2.42	−1.98
Std. dev.	0.38	0.15	0.70
AcrOcc	0.66	0.53	0.85
Std. dev.	0.18	0.11	0.27
AcrSpd	0.87	0.37	1.45
DifSpd	0.05	0.02	0.08
Std. dev.	0.07	0.03	0.14
Transitional state from free flow to synchronized flow (F→S)			
Constant	−1.75	−2.11	−1.36
Std. dev.	0.19	0.10	0.55
DifSpd	0.05	0.01	0.08
Std. dev.	0.02	0.01	0.11
LDifOcc	0.10	0.03	0.18
Transitional state from synchronized flow to free flow (S→F)			
Constant	−3.60	−5.04	−1.97
Std. dev.	0.20	0.10	0.60
AcrCnt	2.48	2.01	2.83
Std. dev.	0.05	0.03	0.13
LDifSpd	0.15	0.04	0.25
Std. dev.	0.03	0.01	0.15
Synchronized flow (S)			
Constant	−1.64	−1.94	−1.36
Std. dev.	0.18	0.10	0.53
CrrOcc	−1.04	−1.27	−0.75
Std. dev.	0.06	0.03	0.18
DevTime	2.74	2.58	3.20
Std. dev.	0.05	0.01	0.43
Transitional state from synchronized flow to wide moving jams (S→J)			
Constant	−2.39	−2.80	−1.90
Std. dev.	0.15	0.10	0.29
LDifSpd	0.07	0.03	0.12
Std. dev.	0.02	0.01	0.06
DevTime	2.22	1.19	3.06
Std. dev.	0.19	0.10	0.72
Transitional state from wide moving jams to synchronized flow (J→S)			
Constant	−3.43	−4.11	−2.97
Std. dev.	0.21	0.10	0.70
DevCnt	1.31	1.02	1.86
Std. dev.	0.05	0.03	0.13
DevTime	3.82	3.55	3.98
Std. dev.	0.05	0.03	0.10
Wide moving jams (J)			
Constant	−2.38	−2.62	−2.13
Std. dev.	0.06	0.03	0.15
AcrCnt	1.14	0.62	1.60
Std. dev.	0.05	0.03	0.12
LDifOcc	0.12	0.08	0.16
Std. dev.	0.02	0.01	0.07

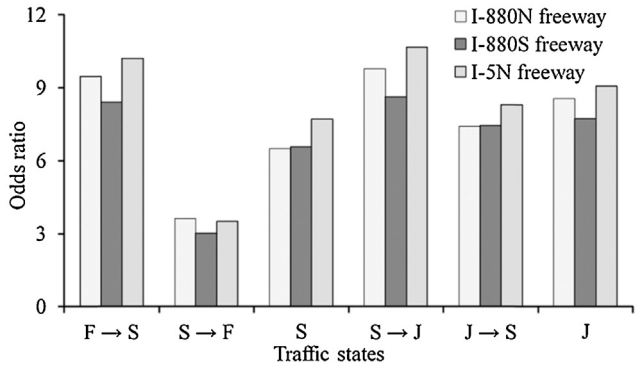

Fig. 7. The odds ratios of the conditional logit models using different datasets.

6. Conclusions and discussion

Various crash risk prediction models have been developed to establish a statistical relationship between the risks of crash occurrences and freeway traffic flow parameters. Combing the crash risk prediction technique and the three phase traffic theory we can better understand to what extend freeway traffic flow affects safety. With the proposed crash risk prediction models, a safety level can be assigned to the traffic phases and the transitional states in the three phase traffic theory. The free flow state has the best relative safety performance, followed by the transitional state S→F, synchronized flow, transitional state J→S, and wide moving jams. The transitional state F→S and S→J are the most hazardous traffic states in which crash likelihood are 8.46 and 8.77 times higher than those in the free flow state. To test for the validity of this finding, Bayesian conditional logit models were developed using data collected from the data collected from the I-880S and the I-5N freeway sections. The estimation results are very similar to those obtained using the data collected from the I-880N freeway section (see Fig. 7).

Crash risk prediction models were then developed to quantitatively evaluate the impacts of traffic flow variables on crash likelihood. The Bayesian random parameter logit model was utilized to account for the heterogeneity caused by unobserved factors. The proposed approach increased the prediction performance of the crash risk models as compared with the conventional logit model. More importantly, with the three phase traffic theory, the contributing factors to crash likelihood can be well explained by the mechanism of phase transitions. We believe that it is a significant advantage over traditional crash risk models which only focuses on the statistical relationship. The present study can be considered a supplement to the three phase traffic theory, and has potential to bridge previous studies in the areas of traffic safety and traffic flow.

Several issues need to be addressed in future studies. First, due to the differences in traffic conditions, driver behavior, and layout of loop detectors, the developed crash risk models may not be able to be directly transferred to another freeway. Additional efforts are needed to examine the transferability of the developed model. Proper methods, such as the Bayesian updating technique, can be used to improve the model transferability. Second, the matched case-control design was adopted in the present study to control for the influence factors. In the future, the unmatched case-control design can be considered to evaluate the impacts on crash likelihood of other variables than traffic flow variables, such as weather and geometric characteristics, and their interactions with the traffic states. Finally, future studies can incorporate the analyses of crash severity and collision type, which can promote a better understanding of the safety performance of different traffic states in the three-phase traffic theory.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Project #: 51322810). The authors would like to thank PeMS and Caltrans for providing the crash and traffic data that were used in this paper.

References

Abdel-Aty, M., Rajashekar, P., 2006. Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data. IEEE Trans. Intell. Transport. Syst. 7 (2), 167–174.
Ahrens, W., Pigeot, I., 2005. Handbook of Epidemiology. Springer.
Allaby, P., Hellinga, B., Bullock, M., 2007. Variable speed limits: safety and operational impacts of a candidate control strategy for freeway applications. IEEE Trans. Intell. Transport. Syst. 8 (4), 671–680.

Ahmed, M., Abdel-Aty, M., 2012. The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Trans. Intell. Transport. Syst. 13 (2), 459–468.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30, 1145–1159.
Balcombe, K., Chalak, A., Fraser, I., 2009. Model selection for the mixed logit with Bayesian estimation. J. Environ. Econ. Manag. 57, 226–237.
Bruce, N., Pope, D., Stanistreet, D., 2008. Quantitative Methods for Health Research: A Practical Interactive Guide to Epidemiology and Statistics. John Wiley & Sons Ltd.
Derrig, R.A., Segui-Gomez, M., Abtahi, A., Liu, L.L., 2002. The effect of population safety belt usage rates on motor vehicle-related fatalities. Accid. Anal. Prev. 34 (1), 101–110.
Golob, T., Recker, W., 2004. A method for relating type of crash to traffic flow characteristics on urban freeways. Transport. Res. A 38 (1), 53–80.
Golob, T., Recker, W., Alvarez, V., 2004. Freeway safety as a function of traffic flow. Accid. Anal. Prev. 36 (6), 933–946.
Gelman, A., Carlin, J., Stern, H., Rubin, D., 2004. Bayesian Data Analysis, second ed. Chapman and Hall, London.
Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6, 721–741.
Hall, F., Hurdle, V., Banks, J., 1992. Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. Transport. Res. Rec. 1365, 12–18.
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. 2012, 373–381.
Hossain, M., Muromachi, Y., 2011. Understanding crash mechanism and selecting appropriate interventions for real-time hazard mitigation on urban expressways. Transport. Res. Rec. 2213, 53–62.
Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57 (1), 97–109.
Harb, R., Yan, X., Radwan, E., Su, X., 2009. Exploring precrash maneuvers using classification trees and random forests. Accid. Anal. Prev. 41 (1), 98–107.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90 (430), 773–795.
Kerner, B.S., 1998. Experimental features of self-organization in traffic flow. Phys. Rev. Lett. 81, 3797–3800.
Kerner, B.S., 2004. The Physics of Traffic. Springer, Heidelberg.
Kerner, B.S., 2002. Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. Phys. Rev. E 65, 1–30.
Kerner, B.S., Klenov, S.L., 2003. Microscopic theory of spatial-temporal congested traffic patterns at highway bottlenecks. Phys. Rev. E 68, 1–20.
Kerner, B.S., Klenov, S.L., 2009. Phase transitions in traffic flow on multilane roads. Phys. Rev. E 80, 1–20.
Kerner, B.S., Rehborn, H., 1996. Experimental properties of complexity in traffic flow. Phys. Rev. E 53 (5), R4275–R4278.
Kerner, B., 2009. Introduction to Modern Traffic Flow Theory and Control. Springer, Heidelberg.
Kerner, B., Klenov, S., Wolf, D., 2002. Cellular automata approach to three-phase traffic theory. J. Phys. A: Math. Gen. 35, 9971–10013.
Knospe, W., Santen, L., Schadschneider, A., Schreckenberg, M., 2002. Single-vehicle data of highway traffic: microscopic description of traffic phases. Phys. Rev. E 65, 1–16.
Munger, D., L'Ecuyer, P., Bastin, F., Cirillo, C., Tuffin, B., 2012. Estimation of the mixed logit likelihood function by randomized quasi-Monte Carlo. Transport. Res. B: Methodol. 46 (2), 305–320.
Neubert, L., Santen, L., Schadschneider, A., Schreckenberg, M., 1999. Single-vehicle data of highway traffic: a statistical analysis. Phys. Rev. E 60 (6), 6480–6490.
Park, B., Zhang, Y., Lord, D., 2010. Bayesian mixture modeling approach to account for heterogeneity in speed data. Transport. Res. B 44 (5), 662–673.
Pande, A., Das, A., Abdel-Aty, M., Hassan, H., 2011. Real-time crash risk estimation are all freeways created equal? Transport. Res. Rec. 2237, 60–66.
Papageorgiou, M., Kotsialos, A., 2002. Freeway ramp metering: an overview. IEEE Trans. Intell. Transport. Syst. 3 (4), 271–281.
Rothman, K.J., Greenland, S., 1998. Modern Epidemiology, second ed. Lippincott Williams and Wilkins, Philadelphia.
Shankar, V., Albin, R., Milton, J., Mannering, F.L., 1998. Evaluating median cross-over likelihoods with clustered accident counts: an empirical inquiry using the random effects negative binomial model. Transport. Res. Rec. 1635, 44–48.
Weiss, G.M., Provost, F., 2003. Learning when training data are costly: the effect of class distribution on tree induction. J. Artif. Intell. Res. 19, 315–354.
Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2003. Statistical and Econometric Methods for Transportation Data Analysis. Chapman & Hall/CRC.
Wu, N., 2002. A new approach for modeling of fundamental diagrams. Transport. Res. A: Policy Pract. 36 (10), 867–884.
Xiong, Y., Mannering, F.L., 2013. The heterogeneous effects of guardian supervision on adolescent driver-injury severities: a finite-mixture random-parameters approach. Transport. Res. B 49, 39–54.
Xu, C., Wang, W., Liu, P., 2013a. A genetic programming model for real-time crash prediction on freeways. IEEE Trans. Intell. Transport. Syst. 14 (2), 574–586.
Xu, C., Wang, W., Liu, P., 2013b. Identifying crash-prone traffic conditions under different weather on freeways. J. Safety Res. 46, 135–144.
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., Liu, P., Wang, W., Li, Z., 2014. Identification of freeway crash-prone traffic conditions for traffic flow at different levels of service. Transport. Res. A: Policy Pract. 69, 58–70.