Bayesian Updating Approach for Real-Time Safety Evaluation with Automatic Vehicle Identification Data

Mohamed M. Ahmed, Mohamed Abdel-Aty, and Rongjie Yu

Although numerous studies have attempted to use data from inductive loop and radar detectors in real-time crash prediction, safety analyses that have investigated the use of traffic data from an increasingly prevalent nonintrusive surveillance system have not included the tag readers on toll roads known as "automatic vehicle identification (AVI) systems." This paper (*a*) compares the prediction performance of a single generic model for all crashes and a specific model for rear-end crashes that used AVI data, (*b*) applies a Bayesian updating approach to generate full probability distributions for the coefficients, and (*c*) compares the estimation efficiency of the semiparametric Bayesian modeling with that of logistic regression with frequentist matched case control. A comparison of AVI data collected before all crashes and rear-end crashes with matched noncrash data revealed that rear-end crashes could be identified with a 72% accuracy, whereas the generic all-crash model achieved an accuracy of only 69% when different validation data sets were used. Moreover, the Bayesian updating approach increased the accuracy of both models by 3.5%.

Intelligent transportation systems rely heavily on detection systems to collect data that are essential to manage traffic, ease congestion, and provide motorists with travel time information. In the past decade, traffic safety studies showed that traffic safety could be incorporated into real-time traffic management systems as well as provide warnings of the increase in the risk situation to promote safety on freeways and expressways (*1–9*). These efforts have been devoted to link real-time traffic conditions to crash occurrence statistically. Most of this real-time crash prediction research attempted to use data collected from inductive loop detectors (ILDs) (*2–9*); however, traffic safety studies performed with data collected from automatic vehicle identification (AVI) systems are lacking (*10, 11*).

ILDs are the most commonly used sensors in traffic management systems and have helped with traffic operation for more than 50 years. Although ILDs suffer from many inherent problems, such as high failure rates and difficulty with maintenance, researchers in the traffic safety area have found that the data collected from loop detectors are useful for crash prediction in real time. According to the *Traffic Detector Handbook,* the actual loop detector failure rates differ from agency to agency because of the large number of variables that contribute to failures (*12*). This failure rate was found to be consistent with the failure rate from the literature for different states and to vary from 24% to 29% at any given time. New nonintrusive detection devices have become technologically advanced enough and sufficiently cost-effective that they may start replacing the commonly deployed intrusive detection devices. The new nonintrusive detection devices include video devices; microwave and laser radar; and passive infrared, ultrasonic, and acoustic sensors.

The central Florida expressway system uses an AVI system for nonstop toll collection as well as for the provision of real-time information to motorists within advanced traveler information systems. This system estimates the travel time on a segment by monitoring the successive passage times of vehicles equipped with E-Pass, O-Pass, or Sun-Pass electronic tags at expressway open-road toll plazas as well as at exit ramps. Data are gathered by AVI tag readers that are installed for the purpose of toll collection and additional tag readers installed solely for the purpose of travel time estimation.

The speed data collected from ILDs and AVI systems differ significantly. One main difference is that ILDs measure time mean speed, whereas AVIs measure space mean speed. Time mean speed is defined as the arithmetic mean of the speed of vehicles passing a point during a given time interval. Time mean speed therefore reflects the traffic condition at only one specific point. Space mean speed, however, is the average speed of all the vehicles occupying a given stretch of the road over some specified time period (several definitions of space mean speed exist, depending on how it is calculated; the definition provided in this paper is the best to describe the AVI's space mean speed). Because not all vehicles are equipped with AVI transponders, the accuracy of travel time estimation depends on the percentage of vehicles that are equipped with transponders. The penetration of E-Pass users is greater than 80% on central Florida's expressway system, which can provide reliable travel time estimates.

It is difficult to delineate from fundamental notions of time mean speed and space mean speed the measure of safety risk without detailed analyses, and a better understanding of these systems is essential in the safety context. Key questions are therefore whether AVI can be used to predict the risk of a crash in real time, what level of accuracy could be achieved for prediction of all crashes, and if that prediction performance can be improved when the single most frequent type of crash on freeways and expressways, the rear-end collision, is targeted (*13*). The impacts of these types of crashes on roadway operation are the most noticeable because most such crashes occur during congested periods (*14*).

Department of Civil, Environmental, and Construction Engineering, University of Central Florida, 4000 Central Florida Boulevard, Orlando, FL 32816-2450. Corresponding author: M. M. Ahmed, mahmed@knights.ucf.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2280, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 60–67.
DOI: 10.3141/2280-07

BACKGROUND

Real-time crash prediction drew researchers' attention in the past decade because it can help with proactive traffic management. Madanat and Liu estimated the likelihood of two types of incidents, crashes and overheated vehicles, using traffic flow and environmental conditions measured by surveillance sensors (*1*). They concluded that merging sections, visibility, and rain are the most significant factors affecting the likelihood of a crash. Hughes and Council used loop detector data to explore the relationship between freeway safety and operations during peak periods (*2*). They found that the variability in vehicle speeds was the most significant measure that affects crash occurrence, whereas macroscopic measures, such as annual average daily traffic and hourly volume, were poor measures in the analysis of safety.

Oh et al. were the first to show the potential ability to establish a statistical relationship to link real-time traffic conditions and crashes (*3*). They used a Bayesian model with traffic data containing the average and standard deviation flow, occupancy, and speed for 10-s intervals and concluded that the 5-min standard deviation of speed contributes the most to the differentiation of precrash and noncrash conditions. Lee et al. used the log-linear approach to model traffic conditions leading to crash precursors (*4*). They added a spatial dimension by using data from upstream and downstream detectors of crashes. They refined their analysis in a later study by considering the average variation of speed on each lane, the average variation of the difference in speed across adjacent lanes, and traffic density (*5*). The coefficient of temporal variation in speed was found to have a relatively longer-term effect on crash potential than density, whereas the effect of the average variation in speed across adjacent lanes was found to be insignificant.

Golob and Recker conducted a detailed study to analyze patterns in crash characteristics as a function of real-time traffic flow (*6*). Nonlinear canonical correlation analysis and principal component analysis were used with three different sets of variables. The first set defined lighting and weather conditions; the second set defined the crash characteristics of collision type, location, and severity; and the third set consisted of real-time traffic flow variables. They concluded that the median speed and the variation in speed between the left and interior lanes are related to the collision type. In addition, they found that the inverse of the traffic volume has more influence than the speed in determining the severity of a crash.

Abdel-Aty et al. used a matched case–control study to link real-time traffic flow variables collected by loop detectors and the likelihood of a crash (*7*). They concluded that the average occupancy at the upstream station along with the coefficient of variation in speed at the downstream station, both during the 5 to 10 min before the crash, were the most significant factors affecting the likelihood of a crash. Abdel-Aty and Pande were able to capture 70% of the crashes using a Bayesian classifier-based methodology, the probabilistic neural network, using different parameters of speed only (*8*). They found that the likelihood of a crash is significantly affected by the logarithm of the coefficient of variation of the speed at the nearest crash station and the two stations immediately preceding it in the upstream direction measured in the 5-min time slice 10 to 15 min before the crash. In a later study, Abdel-Aty and Pande (*9*) developed a strategy to identify real-time traffic conditions prone to result in rear-end crashes using freeway ILD data. They were able to achieve accuracy greater than that from the single generic model for all crashes, and their model was capable of identifying 75% of rear-end crashes 5 to 10 min before their occurrence with a reasonable false-alarm rate.

Although real-time crash prediction models that use data collected from ILDs have been described in the literature, no safety analyses that have been conducted with traffic data from tag readers on toll roads (AVI) have been found.

Ahmed and Abdel-Aty identified expressway locations with high crash potential using real-time speed data collected from an AVI system on 78 mi of the expressway network in Orlando, Florida (*10*). By use of the random forest technique for selection of significant variables and stratified matched case–control analysis to link the crash data to the space mean speed, the logarithm of the odds of a crash occurrence were calculated. It was concluded that none of the speed parameters obtained from AVI systems spaced, on average, 3 mi or more apart was able to identify crash-prone conditions in a statistically significant manner. The results suggested that the AVI data could be useful only if the segments of the AVI system were within 1.5 mi, on average. The results showed that the likelihood of a crash is statistically significantly related to speed data obtained from the AVI system, and the model achieved about 70% accuracy.

In a later study, Ahmed et al. used 3 years (2007 to 2009) of AVI system data collected from a 15-mi segment on I-70 in Colorado and real-time weather data collected from three weather stations on the roadway segment and concluded that data from an AVI system and real-time weather data provide good measures of the risk of a crash in real time (*11*). It was concluded that the 10-min average speed at the crash segment during the 5 to 15 min before the crash and the average visibility during the 1 h before the crash are the most significant factors affecting the likelihood of a crash on a freeway. The risk of a crash increased 6.5% for each unit decrease in the 10-min average speed, and it increased 37% for each unit decrease in the average visibility measured over the 1 h before the crash. The findings from these previous two studies suggest that the speed data collected from AVI systems can provide a good measure of the risk of a crash within advanced traffic management systems.

In this paper, a generic semiparametric Bayesian matched case–control model was calibrated for all crash types, and another model was calibrated for rear-end crashes. The study also investigated if knowledge about the covariates at the same location from previous years can provide a better fit and enhance the ability of the model to predict crashes more accurately. For this approach to be examined in real-life applications, data from 1 year (2007) were used to calibrate the model by the use of classical (frequentist) matched case–control logistic regression. The coefficient estimates were then used as priors in a Bayesian matched case–control analysis to update the coefficients with data from another year (2008) and data from a different year (2009) were used for validation.

Unlike other studies that have been limited by the availability of data and so carried out the sensitivity analysis with the same data that were used to calibrate the model, this study used a separate set of data for validation and scoring of the model.

DATA COLLECTION AND PREPARATION

The SR-417 expressway section under consideration and for which AVI data were available is 33 mi long. Central Florida's expressways are equipped with an AVI system for toll collection and travel time estimation; the 33-mi section contains 22 AVI tag readers in both directions, and the average spacing is 1.47 mi.

Two sets of data were used in the study: archived data from the AVI system for SR-417 in Orlando and the corresponding crash data for

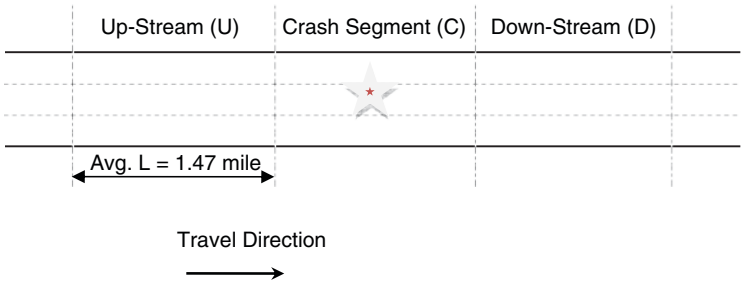

FIGURE 1 Scheme for segments in AVI system (avg. L = average length).

3 years (2007 through 2009). The Orlando–Orange County Expressway Authority archives and maintains only the processed 1-min space mean speed and the estimated average travel time along the defined roadway segments. The unprocessed original time stamps of the tag readings are not available; these data are typically discarded for privacy reasons after the travel time is processed. The crash data were obtained from the Crash Analysis Resource maintained by the Florida Department of Transportation for the same years.

In a previous study, it was found that the occurrence of a crash was mostly related to the crash segment: one segment in the upstream direction and another segment in the downstream direction. Therefore, these segments were considered in the data extraction process and modeling parts of the study (*10*). The crashes were assigned to each segment; upstream and downstream segments were identified to extract their corresponding data from the AVI system. The upstream, crash, and downstream segments were named U, C, and D, respectively. The AVI system segment scheme is illustrated in Figure 1.

Data from the AVI system corresponding to each crash case were extracted in the following process. If a crash occurred on February 7, 2008 (Thursday), at 2:00 p.m. on SR-417 eastbound, the crash segment (C) was identified by the use of geographic information system software, and data were obtained for two other segments (one in the upstream direction and one in the downstream direction) for the period from 1:30 to 2:00 p.m. (30 min). Data for four noncrash cases for the same season (to control for weather conditions), location, and time for different Thursdays were also extracted. Data for the crash and the noncrash cases were extracted only when no crashes were observed within 1 h of the original crash at the same AVI system segment. Four crashes occurred within the crash segment a few minutes after the first crash, but data for these crashes were not considered because all speed parameters would have been affected by the first crash event.

The extracted 1-min speed data were aggregated to different aggregation levels of 2, 3, 5, and 10 min to investigate the aggregation level that provides the best accuracy in the modeling part of the study. The 5-min aggregation level was found to provide a better statistical fit [a smaller deviance information criterion (DIC)] and relatively higher classification accuracy. The 30-min speed data were divided into six time slices, in which Time Slice 1 represents the period between the crash and 5 min before the crash and Time Slice 6 represents the interval between 25 and 30 min before the crash. The data for Time Slice 1 were discarded in the analysis because 5 min would not provide enough time for a successful intervention to reduce the risk of a crash in a proactive safety management strategy. Moreover, the actual crash might not be precisely known. Golob and Recker discarded the 2.5 min of traffic data immediately preceding the reported time of each crash to avoid the uncertainty over the actual time of the crash (*6*). In general, with the proliferation of mobile phones and closed-circuit television cameras on expressways, a crash is usually almost immediately identified.

Average speeds, standard deviations of the speed, and the logarithms of the coefficient of variation of the speed (the standard deviation of the speed divided by the average speed) were calculated over the 5-min time intervals. The measure notations take the general form $xy\_z\beta$, where xy takes the value of AV, SD, or CV, for average, standard deviation, and coefficient of variation of speed, respectively; z represents AVI system segments and takes the value of U, C, or D, for upstream, crash, and downstream segment, respectively; and β takes a value of from 2 to 6, which refers to the time slices.

Unlike ILD data, which are known to suffer from high percentages of missing observations or bad readings, AVI system data are missing less than 5% observations and have no unreasonable values of speeds. The missing speed data were imputed by preservation of the distribution of the original data, and then the coefficient of variation was calculated. The final data set had a total of 45 variables consisting of three speed parameters for each of the three AVI segments at five time intervals (time slices).

Although crashes involving driving under the influence of alcohol or drugs and distraction-related crashes were less than 2% of total crashes, data for those crashes were excluded from the crash data set to examine only the effect of short-term turbulence in traffic speed. Hence, the analysis presented in this study is based on 447 total crashes, 171 of which were rear-end crashes.

METHODOLOGY

Matched Crash–Noncrash Analysis

The study design used a matched case–control methodology, which is a robust way of examining the crash precursors accounting for confounding factors such as time of crash, seasonal effect, and location, including all related geometric characteristics. Case–control studies are expected to provide more accurate results, as they eliminate confounding factors by matching (*15*). For each selected crash case, a random selection of m controls (noncrash cases) was chosen to account for the matching factors of location, time of day, day of week, and season (Orlando has two distinct weather seasons, and matched noncrash cases were taken from the same season for each crash case).

Although the matched case–control methodology can handle the confounding factors, other confounding factors, such as the behav-

ior of individual drivers, are not considered because the matching is for location and time variables only. Different *m:*1 ratios (ratio of the sample size of noncrash cases to crash cases) were examined, and *m* equal to 4 was found to give an odds ratio of relatively higher precision (lower standard error). Previous studies showed that negligible power is gained through addition of controls beyond 3-to-1 matching (*16*). Finally, the matched set (stratum) was formed of $m(4) + 1$ observations.

The modeling is estimated under the conditional likelihood principle of statistical theory, which accounts for within-stratum differences between crash and noncrash speed parameters. Use of the conditional likelihood eliminates the parameters associated with the covariates used for matching (e.g., crash time and location).

Matched case–control studies are based on the classical prospective logistic regression model, with binary outcome *y* (case–control status), covariate (*x*), and stratum level *N.* Suppose that *N* stratum has one crash and *m* noncrash cases are in stratum *j,* where *j* is equal to 1, 2, 3, . . . , *N.* The term $p_j(x_{ij})$ is the probability that the *i*th observation in the *j*th stratum is a crash, where the vector of *k* speed parameters $x_1, x_2, \ldots, x_k$ can be denoted $x_{ij} = (x_{1ij}, x_{2ij}, \ldots, x_{kij})$, where $i = 0, 1, 2, \ldots, m$ and where $j = 1, 2, \ldots, N$. This crash probability may be modeled by the following linear logistic regression model described in a study by Abdel-Aty et al. (*7*):

$$\text{logit}\left\{p_j\left(x_{ij}\right)\right\} = \alpha_j + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} \tag{1}$$

where β is a coefficient. The logistic regression model for matched case–control studies differs from that for unmatched studies in that it allows the intercept to vary among the matched units of cases and controls. The intercept α summarizes the effect of variables used to form strata on the crash probability, and it is different for the different strata.

To account for the stratification in the analysis, a conditional likelihood is constructed. The crash probabilities cannot be estimated by matched case–control logistic regression by the use of Equation 1, however, because the conditional likelihood function $L(\beta)$ is independent of the intercept terms $\alpha_1, \alpha_2, \ldots, \alpha_N$; therefore, the effects of matching variables cannot be estimated. This conditional likelihood function is expressed as follows:

$$L(\beta) = \prod_{j=1}^{N}\left[1 + \sum_{i=1}^{m}\exp\left\{\sum_{u=1}^{k}\beta_u\left(x_{uij} - x_{u0j}\right)\right\}\right]^{-1} \tag{2}$$

where *u* denotes the set of crash cases just before the *i*th ordered event time. However, the values of the β parameters that maximize the conditional likelihood function given by Equation 2 are also the estimates of the β coefficients in Equation 1. These estimates are log odds ratios and may be used to approximate the relative risk of a crash.

In this analysis, the PHREG procedure in SAS software (Version 9.2) is used. PHREG provides the hazard ratio, which is another term for relative risk used in SAS. In addition, a prediction model can be developed by use of the log odds ratios under this matched crash–noncrash analysis. This model can be demonstrated by consideration of two observation vectors, $x_{1j} = (x_{11j}, x_{21j}, x_{31j}, \ldots, x_{k1j})$ and $x_{2j} = (x_{12j}, x_{22j}, x_{32j}, \ldots, x_{k2j})$ from the *j*th strata on the *k* speed parameters. By use of Equation 1, the log odds ratio of the occurrence of a crash because of speed parameter vector x_{1j} relative to traffic speed vector x_{2j} has the following form:

$$\log\left\{\frac{\dfrac{p\left(x_{1j}\right)}{\left[1-p\left(x_{1j}\right)\right]}}{\dfrac{p\left(x_{2j}\right)}{\left[1-p\left(x_{2j}\right)\right]}}\right\} = \beta_1\left(x_{11j} - x_{12j}\right) + \beta_2\left(x_{21j} - x_{22j}\right) + \cdots + \beta_k\left(x_{k1j} - x_{k2j}\right) \tag{3}$$

The right-hand side of Equation 3 is independent of α_j and can be calculated by use of the estimated β coefficients. Thus, the relative log odds ratio described above (left-hand side of Equation 3) may be used to predict crashes by replacement of x_{2j} with the vector of values of the traffic flow variables in the *j*th stratum of noncrash cases. One may use the simple average of all noncrash observations within the stratum for each variable. Let $\bar{x}_{2j} = (\bar{x}_{12j}, \bar{x}_{22j}, \bar{x}_{32j}, \ldots, \bar{x}_{k2j})$ denote the vector of mean values of noncrash cases of the *k* variables within the *j*th stratum. Then, the log odds ratio of crash cases relative to noncrash cases may be approximated by the following equation:

$$\log\left\{\frac{\dfrac{p\left(x_{1j}\right)}{\left[1-p\left(x_{1j}\right)\right]}}{\dfrac{p\left(\bar{x}_{2j}\right)}{\left[1-p\left(\bar{x}_{2j}\right)\right]}}\right\} = \beta_1\left(x_{11j} - \bar{x}_{12j}\right) + \beta_2\left(x_{21j} - \bar{x}_{22j}\right) + \cdots + \beta_p\left(x_{k1j} - \bar{x}_{k2j}\right) \tag{4}$$

Therefore, the log odds ratio can be used to predict crashes by establishment of a threshold value that attains the desirable crash classification accuracy.

Bayesian Updating Approach

This study uses the Bayesian semiparametric Cox proportional hazards model (PHM) to explain the relationship between an event (crash) occurring at a given time and a set of risk factors in a matched case–control design and to control mainly for the confounding factors of time, location, and season. The Cox PHM is commonly used for survival analysis; an important distinction in survival analysis is how the time dependence in the event process (the baseline hazard in the absence of any covariate effects) is parameterized. Cox's semiparametric model assumes a parametric form for the effects of the covariates, but it allows an unspecified form for the baseline hazard. Therefore, the Cox PHM can be used regardless of whether the survival time is discrete or continuous. The Cox PHM is performed with the SAS program (Bayes PROC PHREG) (*17*) by formation of a stratum for each matched set, a dummy variable for the survival time is created in the data set such that all the crash cases in a matched set have the same event time value, and the corresponding noncrash cases (controls) are censored at later times.

The classical Cox semiparametric model estimates the coefficients of parameters solely on the basis of the information from the observed data, whereas the Bayesian Cox semiparametric model makes use of the combined information of the prior as well as the observed data to estimate the coefficients of the parameters. In the Bayesian framework, the data are used to update beliefs about the behavior of the parameter to assess its distributional properties as well as possible. PROC PHREG with the Bayes option generates a Markov chain that contains the approximate posterior distribution of samples by use of Gibbs sampling and the adaptive rejection

sampling algorithm (*18, 19*). DIC, a Bayesian generalization of the Akaike information criterion, is used to measure the complexity and fit of the model (*20*). In addition, a sensitivity analysis is conducted to measure the accuracy of each of the estimated models by use of a different set of data from 2009 for validation.

RESULTS AND DISCUSSION

Model Estimation and Diagnostics for All Crashes Versus Rear-End Crashes

As mentioned earlier, a frequentist matched case–control model was estimated for all crashes that occurred on the expressway section in 2007. The data set comprises 690 observations (138 crash cases and 552 noncrashes as a control). With prior knowledge of the likely range of values of the parameters from 2007, informative priors were specified for parameters for all crashes that occurred in 2008 (165 crashes and 660 noncrashes) to avoid use of the same data in the updating process. Use of noninformative priors in the Bayesian estimation resulted in the same estimate obtained with the frequentist model.

In the Bayesian update, one chain of 20,000 iterations was set up in SAS on the basis of the convergence speed and the magnitude of the data set. Before inferences are drawn from the posterior sample, the trace, autocorrelation, and density plots should be examined for each parameter to be content that the underlying Markov chain has converged. According to Brooks and Gelman convergence diagnostics, the trace, autocorrelation, and density plots for the two significant parameters shown in Figure 2 suggest that the mixing in the chain is acceptable with no correlation (*21*). After the convergence is ensured, the first 2,000 samples were discarded as adaptation and burn-in.

A univariate analysis was first conducted to check the significance of each variable. Different automatic search techniques (stepwise, forward, and backward) were attempted to identify significant variables in the multivariate analysis. These procedures were implemented to identify which terms were still statistically significant in the presence of other factors. Because variables not significant at the .05 level may still be associated with the response after adjustment for other covariates, any variable with a p-value of <.25 in the univariate analysis results was considered eligible to enter the multivariate model (*17*). The three search techniques agreed that two variables are significantly associated with crash occurrence.

Table 1 shows the estimates of the means and standard deviations of the beta coefficients, credible intervals, and hazard ratios for the all-crashes model; two variables were found to be significant: SD_C2 and AV_D2. The standard deviation of speed of the crash segment at Time Slice 2 (5 to 10 min before the crash) has a positive beta coefficient, whereas the average speed of the adjacent downstream segment at Time Slice 2 has a negative beta coefficient. These values mean that a high level of variation in the speed at the crash segment and a decrease in the average speed at the downstream segment may increase the risk of a crash at this location. A decrease in speed downstream might represent buildup of a queue.

The hazard ratio is the exponent of the beta coefficient and represents an estimate of the expected change in the risk ratio of a crash versus noncrash per unit change in the corresponding factor. The hazard ratio of 1.14 means that the risk of a crash increases by 13% for each unit increase in SD_C2. The hazard ratio is multiplicative in nature for the continuous variables, which means that a two-unit increase in SD_C2 changes the risk by 1.14^2, or 1.28 (a 28% increase).

With the same methodological updating approach described earlier, a Bayesian matched case–control model was estimated only for rear-end crashes that occurred in 2008 by the use of informative priors from the frequentist model that was estimated with data only for rear-end crashes that occurred in 2007. The data set for 2007 had 280 observations (56 rear-end crash cases and 224 noncrashes as a control), whereas the data set for 2008 used to update the model coefficients had 305 observations (61 rear-end crashes and 244 noncrashes). The convergence was similarly assessed by the use of plots for trace, autocorrelation, and density; and the model converged reasonably.

Table 2 shows the means and standard deviations of the beta coefficient estimates, credible intervals, and hazard ratios. SD_C2 and AV_D2 were found to be significant. However, for the standard deviation of speed at the crash segment at Time Slice 2, the hazard ratio for the rear-end-crash model increased by more than twice the hazard ratio for the all-crash model; and for the average speed of the downstream segment at Time Slice 2, the hazard ratio decreased by about 20%. This result may indicate that an increase in the variation of the speed at any given segment coupled with a decrease in the average speed at the downstream segment may result in a rear-end crash more than any other type of crash.

One limitation in the current AVI archiving system, however, is that the system does not record the percentage of lane changes per segment. This percentage can be calculated by development of an algorithm to compare the unique tag identifier for each individual vehicle at the beginning and end of each segment. Moreover, the algorithm can process the original raw AVI system data in a way that provides space mean speed by lane. In that way, a better picture of not only the longitudinal variation in speed at the AVI system segment but also the variation across the lanes can be comprehended. The availability of detailed lane speed data may also help to identify other types of crashes, such as sideswipe and angle crashes.

Use of the informative prior slightly enhanced the model fit; the DIC decreased from 652.371 to 647.695 for the all-crashes model and from 111.278 to 106.097 for rear-end crashes.

Classification Accuracy for All-Crash Model Versus Rear-End-Crash Model

Sensitivity analyses were conducted to implement the estimated models in a real-time application. Table 3 shows the sensitivities and the specificities for the final models. Sensitivity is the proportion of crashes that are correctly identified as crashes by the model, whereas specificity is the proportion of noncrashes that are correctly identified as noncrashes by the model (*22*). The sensitivity and the specificity can be calculated by use of the odds ratio given by Equation 4. For example, the means of the two variables SD_C2 (which is the standard deviation of the speed of the crash segment at Time Slice 2, which is 5 to 10 min before the crash) and AV_D2 (which is the average speed of the downstream segment at Time Slice 2) were calculated for all four noncrash cases within each matched set. The estimated vector of these means for the noncrash cases replaced the vector in Equation 4 for the jth matched set. The odds ratio can be estimated by use of the beta coefficients from the model updated with the 2008 data set in Equation 4, in which the vector is the actual observation in the 2009 data set for all crashes and rear-end crashes.

The sensitivities were found to be 69.44% and 72.22% for all crashes and rear-end crashes, respectively, by use of the Bayesian matched case–control model with noninformative priors; and they increased to 72.92% and 75.93%, respectively, by use of the

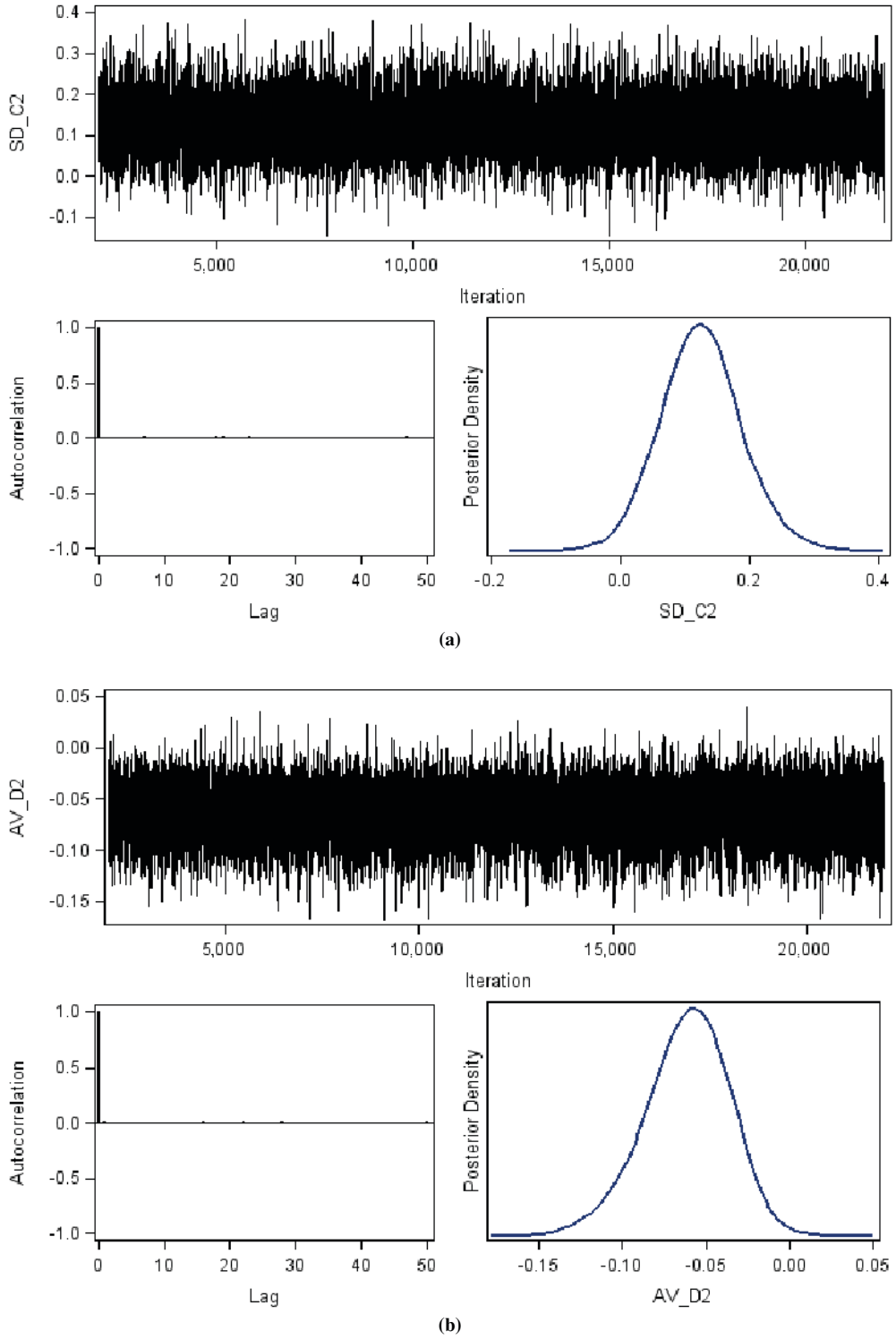

FIGURE 2 Diagnostics plots for the all-crash model: (*a*) SD_C2 and (*b*) AV_D2.

Bayesian updating approach with specified informative priors from 2007. Both models had reasonable rates of false-positive results. At a threshold value of unity, 42.01% of all crashes and 45.83% of rear-end crashes were incorrectly classified. Different rates of false-positive results can be obtained when the threshold is changed on the basis of the management strategy. The threshold should be chosen carefully in real-world applications; large numbers of false alarms might affect drivers' compliance with the system and hence reduce its effectiveness. Nevertheless, the objective of advanced traffic management systems to reduce turbulence to improve operations can still be achieved even with a high percentage of false alarms. Strategies that are part of intelligent transportation system,

TABLE 1 SR-417 Model Estimates and Hazard Ratios for All Crashes, 2008

	β Coefficient				Hazard Ratio			
			Credible Interval		β Coefficient		Credible Interval	
Parameter	Mean	Standard Deviation	2.5%	97.5%	Mean	Standard Deviation	2.5%	97.5%
SD_C2	0.1256	0.0639	0.00312	0.2562	1.1362	0.0729	1.0031	1.2920
AV_D2	–0.0614	0.0257	–0.1167	–0.0153	0.9408	0.0241	0.8899	0.9848

NOTE: Fitness statistics: DIC = 647.695 (a smaller DIC is better); pD (effective number of parameters) = 2.149.

TABLE 2 SR-417 Model Estimates and Hazard Ratio for Rear-End Crashes, 2008

	β Coefficient				Hazard Ratio			
			Credible Interval		β Coefficient		Credible Interval	
Parameter	Mean	Standard Deviation	2.5%	97.5%	Mean	Standard Deviation	2.5%	97.5%
SD_C2	0.9151	0.3852	0.1986	1.7065	2.6949	1.1318	1.2197	5.5096
AV_D2	–0.2627	0.1520	–0.6147	–0.0313	0.7776	0.1124	0.5408	0.9692

NOTE: Fitness statistics: DIC = 106.097 (a smaller DIC is better); pD = 1.611.

such as variable speed limits, could be introduced so that drivers would have no knowledge of the occurrence of a false alarm.

CONCLUSIONS

Although traffic flow data collected from ILDs are useful for real-time proactive safety management, no studies have attempted to use data from AVI systems for real-time assessments of safety risks. Data from AVI systems were found to provide a measure of the risk of a crash in real time comparable to that obtained with data from ILDs. The operation-based management of expressways can benefit from the data collected not only for estimation of tolls collected and travel times but also to provide warnings of situations of increased risk.

Few studies have predicted crashes by type using real-time traffic data collected on freeways and expressways. In contrast, this study explicitly classified and compared a generic model for all types of crashes with a model for a specific crash type (rear-end crashes) using data collected from tag reader (AVI) systems on expressways.

The paper presents a Bayesian updating framework to identify real-time traffic conditions prone to cause crashes by the use of expressway AVI data. Three years of crash data and the corresponding AVI data on SR-417 in Orlando were used, and a classical (frequentist) matched case–control model was estimated with data from 2007. With prior knowledge of the likely range of values of the parameters from 2007 on the same expressway corridor, informative priors were specified for the parameters in a semiparametric Bayesian matched case–control framework to avoid use of the same data in the updating process. This approach was applied to all crashes and then to rear-end crashes. By comparison of the data from the AVI system for the period preceding all crash types and rear-end crashes with data for noncrashes, the hazard ratio for the standard deviation of the speed of the crash segment in the 5 to 10 min before the crash for the rear-end crash model was found to increase by more than twice the hazard ratio for the overall crash model. The hazard ratio for the average speed of the downstream segment in the 5 to 10 min before the crash decreased. This may indicate that the increase in the variation of the speed at any given segment coupled with a decrease in the average speed in the downstream segment may result in a rear-end crash more than any other type of crash.

The classification accuracy for the model of rear-end crashes is greater than that achieved by the generic all-crashes model: 72.22% of the rear-end crashes were correctly identified, whereas the generic all-crashes model correctly identified only 69.44% of crashes. Moreover, the proposed Bayesian updating approach showed a better fit in the form of relatively lower DIC values by the use of informative priors. The accuracy of both models also increased to 75.93% and 72.92% for rear-end and all crashes, respectively.

The proposed methodology leads to an estimate of risk much more efficiently than ordinary logistic regression with frequentist matched case control. Use of the Bayesian updating approach is strongly recommended as a robust technique to reduce uncertainty in the parameters and increase the accuracy of the model fit.

Although the AVI system can provide the percentage of lane changes per segment by comparison of the unique tag identifier for each vehicle at the beginning and end of the segment as well as provide space mean speed for each lane to estimate the variation in speed across lanes, the algorithm and the archiving system of the AVI system in their current forms do not report this information, and therefore, expressway authorities are encouraged to update their archiving systems.

This paper suggests that in their current form, data from AVI systems can provide an acceptable real-time assessment of the safety risk for all crash types in general and for rear-end crashes in particular. Furthermore, with minor modifications to how tag readers are structured and how the data from the AVI system are processed and

TABLE 3 Classification Results

Variable	Predicted: 0 (Noncrash)	Predicted: 1 (Crash)	Total
All Crashes			
Actual			
0 (noncrash)			
Frequency	334	242	576
Percentage	46.39	33.61	80.00
Row (%)	57.99[a]	42.01[b]	na
Column (%)	89.54	69.74	na
1 (crash)			
Frequency	39	105	144
Percentage	5.42	14.58	20.00
Row (%)	27.08[c]	72.92[d]	na
Column (%)	10.46	30.26	na
Total			
Frequency	373	347	720
Percentage	51.81	48.19	100.00
Rear-End Crashes			
Actual			
0 (noncrash)			
Frequency	117	99	216
Percentage	43.33	36.67	80.00
Row (%)	54.17[a]	45.83[b]	na
Column (%)	90.00	70.71	na
1 (crash)			
Frequency	13	41	54
Percentage	4.81	15.19	20.00
Row (%)	24.07[c]	75.93[d]	na
Column (%)	10.00	29.29	na
Total			
Frequency	130	140	270
Percentage	48.15	51.85	100.00

NOTE: na = not applicable.
[a]Specificity.
[b]False-positive rate.
[c]False-negative rate.
[d]Sensitivity.

archived, it will be possible to enhance the predictive accuracy and extend the proposed methodology to other crash types.

ACKNOWLEDGMENTS

The authors thank the Orlando–Orange County Expressway Authority and Atkins for providing the AVI data that were used in this study.

REFERENCES

1. Madanat, S., and P.-C. Liu. *A Prototype System for Real-Time Incident Likelihood Prediction.* IDEA project final report (ITS-2). TRB, National Research Council, Washington, D.C., 1995.
2. Hughes, R., and F. Council. On Establishing Relationship(s) Between Freeway Safety and Peak Period Operations: Performance Measurement and Methodological Considerations. Presented at 78th Annual Meeting of the Transportation Research Board, Washington, D.C., 1999.
3. Oh, C., J.-S. Oh, S.-G. Ritchie, and M.-S. Chang. Real-Time Estimation of Freeway Accident Likelihood. Presented at 80th Annual Meeting of the Transportation Research Board, Washington, D.C., 2001.
4. Lee, C., F. Saccomanno, and B. Hellinga. Analysis of Crash Precursors on Instrumented Freeways. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1784,* Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 1–8.
5. Lee, C., B. Hellinga, and F. Saccomanno. Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840,* Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 67–77.
6. Golob, T., and W. Recker. *Relationships Among Urban Freeway Accidents, Traffic Flow, Weather and Lighting Conditions.* California PATH working paper UCB-ITS-PWP-2001-19. Institute of Transportation Studies, University of California, Berkeley, 2001.
7. Abdel-Aty, M., N. Uddin, A. Pande, M. F. Abdalla, and L. Hsia. Predicting Freeway Crashes from Loop Detector Data by Matched Case–Control Logistic Regression. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1897,* Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 88–95.
8. Abdel-Aty, M., and A. Pande. Identifying Crash Propensity Using Specific Traffic Speed Conditions. *Journal of Safety Research,* Vol. 36, No. 1, 2005, pp. 97–108.
9. Abdel-Aty, M., and A. Pande. Comprehensive Analysis of the Relationship Between Real-Time Traffic Surveillance Data and Rear-End Crashes on Freeways. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1953,* Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 31–40.
10. Ahmed, M., and M. Abdel-Aty. The Viability of Using Automatic Vehicle Identification Data for Real-Time Crash Prediction. *IEEE Transactions on Intelligent Transportation Systems,* Vol. 13, No. 2, 2012, pp. 459–468.
11. Ahmed, M., R. Yu, and M. Abdel-Aty. Safety Application of Automatic Vehicle Identification and Real-Time Weather Data on Freeways. Presented at 18th ITS World Congress, 2011.
12. *Traffic Detector Handbook,* Vol. II, 3rd ed. Publication FHWA-HRT-06-108. FHWA, U.S. Department of Transportation, 2006.
13. *Analyses of Rear-End Crashes and Near-Crashes in the 100-Car Naturalistic Driving Study to Support Rear-Signaling Countermeasure Development.* Report DOT HS 810 846. NHTSA, U.S. Department of Transportation, 2007.
14. Abdel-Aty, M., N. Uddin, and A. Pande. Split Models for Predicting Multivehicle Crashes During High-Speed and Low-Speed Operating Conditions on Freeways. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1908,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 51–58.
15. Breslow, N., and N. Day. *Statistical Methods in Cancer Research,* Vol. I. *The Analysis of Case Control Studies.* IARC Scientific Publication No. 32. International Agency for Research on Cancer, Geneva, 1980.
16. S. Kuhn, B. Egert, S. Neumann, and C. Steinbeck. Building Blocks for Automated Elucidation of Metabolites: Machine Learning Methods for NMR Prediction. *BMC Bioinformatics,* Vol. 9, No. 1, 2008, p. 400.
17. *SAS/STAT 9.2 User's Guide,* 2nd ed. SAS Institute Inc., Cary, N.C., 2009. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#titlepage.htm. Accessed April 2011.
18. Gilks, W., N. Best, and K. Tan. Adaptive Rejection Metropolis Sampling Within Gibbs Sampling. *Applied Statistics,* Vol. 44, No. 4, 1995, pp. 455–472.
19. Gilks, W., and P. Wild. Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics,* Vol. 41, No. 2, 1992, pp. 337–348.
20. Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and V. D. Linde. Bayesian Measures of Model Complexity and Fit (with Discussion). *Journal of the Royal Statistical Society B,* Vol. 64, No. 4, 2003, pp. 583–616.
21. Brooks, S. P., and A. Gelman. Alternative Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics,* Vol. 7, 1998, pp. 434–455.
22. Agresti, A. *Categorical Data Analysis,* 2nd ed. John Wiley & Sons, Inc., Hoboken, N.J., 2002.

All opinions and results are those of the authors.

The Safety Data, Analysis, and Evaluation Committee peer-reviewed this paper.