The Viability of Using Automatic Vehicle Identification Data for Real-Time Crash Prediction

Mohamed M. Ahmed and Mohamed A. Abdel-Aty

***Abstract*—Real-time crash prediction research attempted the use of data from inductive loop detectors; however, no safety analysis has been carried out using traffic data from one of the most growing nonintrusive surveillance systems, i.e., the tag readers on toll roads known as automatic vehicle identification (AVI) systems. In this paper, for the first time, the identification of freeway locations with high crash potential has been examined using real-time speed data collected from AVI. Travel time and space mean speed data collected by AVI systems and crash data of a total of 78 mi on the expressway network in Orlando in 2008 were collected. Utilizing a random forest technique for significant variable selection and stratified matched case–control to account for the confounding effects of location, time, and season, the log odds of crash occurrence were calculated. The length of the AVI segment was found to be a crucial factor that affects the usefulness of the AVI traffic data. While the results showed that the likelihood of a crash is statistically related to speed data obtained from AVI segments within an average length of 1.5 mi and crashes can be classified with about 70% accuracy, all speed parameters obtained from AVI systems spaced at 3 mi or more apart were found to be statistically insignificant to identify crash-prone conditions. The findings of this study illustrate a promising real-time safety application for one of the most widely used and already present intelligent transportation systems, with many possible advances in the context of advanced traffic management.**

***Index Terms*—Automatic vehicle identification (AVI), freeway/expressway, intelligent transportation system (ITS), safety risk.**

I. Introduction

TRAFFIC detection technology is the main spine of any intelligent transportation system (ITS); there is a wider range of vehicle detection devices in use than ever before on highways, starting from the popular inductive loops and magnetometers to video and radar-based detectors. It is known that the history of the loop detector extends to 50 years ago when it was first developed in the 1960s, and the inductive loop detectors (ILDs) have become the most widely utilized sensors in traffic management systems. The ILD remained unchallenged for more than 30 years because of its simple design, until less intrusive detection options became technologically advanced enough to offer a relief from some of the inherent challenges of the loop detectors. The main problem of the loops is reliability, since loop detectors tend to fail due to the very hard environment of the pavement, the temperature variation, and the resulted shifts in the pavement, which can break the wires and render the loop detector nonfunctional. According to the Traffic Detector Handbook 2006 [1], the actual loop detector failure rates differ from agency to agency because of the large number of variables that contribute to the failure. This failure rate is found to be consistent with the failure rate literature for different states and varies between 24% and 29% at any given time. The secondary problem of the loop detectors is maintenance, since cutting into the pavement to repair the defective loops may shorten the lifetime of the pavement or result in pavement damage. Moreover, maintenance is sometimes limited or not possible on congested roadways.

During the last decade, new nonintrusive detection devices were deployed as alternatives to ILDs, such as video, microwave and laser radar, passive infrared, and ultrasonic and acoustic sensors. Nowadays, nonintrusive detection devices have improved in terms of accuracy, cost, and ease of use. Installation and maintenance are relatively easy than the loop detectors since the nonintrusive detection devices can be mounted above or alongside the roadway and hence enhance and increase reliability. While the inductive loops are expected to continue to function for several years, many transportation agencies seem to be shifting attention to nonintrusive alternatives.

Automatic vehicle identification (AVI) is among other systems, such as satellite positioning and mobile communications using Global System for Mobile communication/General Packet Radio Service, that contributed in the advancement of the electronic toll collection (ETC) systems by first introducing the dedicated ETC lanes, where the vehicles slow down into channeled toll lanes and, recently, where the express ETC lanes have operated at highway speeds, also known as open road tolling (ORT). ORT with ETC technology nowadays are widely utilized worldwide to automate the payment process, increase system throughput and reduce congestion, improve customer service, enhance safety, apply congestion pricing, increase toll revenues, and reduce environmental impacts. ETC systems are composed of AVI that determines the ownership of the vehicle to be charged to the corresponding customer, automatic vehicle classification to charge different fair rates to different

Manuscript received November 12, 2010; revised May 10, 2011 and September 11, 2011; accepted September 30, 2011. Date of publication October 26, 2011; date of current version May 30, 2012. The Associate Editor for this paper was L. Li.

The authors are with the Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816 USA (e-mail: mahmed@knights.ucf.edu; m.aty@ucf.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2011.2171052

vehicle types, and video enforcement systems to capture images of the violator and/or license plate that pass through the ETC lanes without a valid transponder. The structure of the ETC systems depends on the following two main factors: 1) the tolling system and 2) the number of access points on the freeway in case the travel time estimation is incorporated within an Advanced Traveler Information System (ATIS). It is worth mentioning that the spacing between access points is about 1 mi or less for urban freeways and can exceed 3 mi for rural freeways. Prior to ETC systems, there were three main tolling systems, namely, the "closed ticket system," the "closed barrier system," and the "open barrier system." The advent of the new ETC systems changed the way toll roads are designed and operated. ETC systems have the ability to easily support other value-added services on the same technology platform. These services might include but not limited to fleet and engine management systems, emergency response services, congestion pricing, pay-as-you-drive insurance services, and navigation capabilities. The aspect of tolling (distance based, flat rate, or congestion based) and the type of facility and access (freeway, expressway, or conventional road) play an important role in the structure and spacing of the tag readers.

Central Florida's expressway system utilizes the AVI system for ETC and the provision of real-time information to motorists within the ATIS. This system estimates the segment travel time by monitoring the successive passage times of vehicles equipped with E-Pass, O-Pass, Sun-Pass, or electronic radio-frequency identification tags at expressway ORT plazas and exits. Data are gathered using AVI tag readers that are installed for the purpose of toll collection and additional tag readers installed solely for the purpose of estimating travel times. It is worth to mention that there are no specific guidelines

Commonly deployed ILDs measure time mean speed (TMS), whereas AVI systems measure space mean speed (SMS). TMS is defined as the arithmetic mean of the speed of vehicles passing a point during a given time interval. Hence, TMS only reflects the traffic condition at one specific point. On the other hand, SMS is the average speed of all vehicles occupying a given stretch of the road over some specified time period. Since not all vehicles are equipped with transponders, the accuracy of travel time estimation would depend on the percentage of vehicles that are equipped with transponders. The penetration of E-Pass users reached above 80% on Central Florida's expressway system. While traffic flow data collected from ILDs were a good safety measure in real-time proactive safety management, data collected from AVI have not been previously investigated in any safety-related study.

II. Background

Safety performance of a transportation facility can be assessed by crash data analysis as one of the most frequently used tools [2]. Crash performance functions were conventionally used to establish relationships among the traffic characteristics, roadway and environmental conditions, driver behavior, and crash occurrence. Although these models are useful to some extent, the aggregated nature of traffic parameters is not capable of identifying the real-time locations with a high probability of crashes.

On the other hand, real-time crash analysis captured the researchers' interest in the last decade since it has the capability of identifying crashes in real time and hence being more proactive in safety management rather being reactive. Madanat and Liu [3] used traffic flow and environmental conditions measured by surveillance sensors to estimate the incident likelihood for two types of incidents related to crashes and overheating vehicles. It was concluded that merging sections, visibility, and rain are the most significant factors affecting crash likelihood prediction. Loop detector data were used by Hughes and Council [4] to explore the relationship between freeway safety and peak period operations. They found that the variability in vehicle speeds was the most significant measure that affects crash occurrence, whereas macroscopic measures, such as average annual daily traffic and hourly volume, were poor measures in the analysis of safety. In addition, Feng [5] suggested that the reduction of speed variance may prevent crash occurrence.

Oh *et al.* [6] was the first to statistically link real-time traffic conditions and crashes. A Bayesian model was used with traffic data containing average and standard deviation of flow, occupancy, and speed for 10-s intervals. It was concluded that the 5-min standard deviation of speed contributes the most in differentiating between precrash and noncrash conditions. Although their sample size of 53 crashes is small, they showed the potential capability of establishing the statistical relationship. Moreover, the practical application of their finding is questionable since 5 min immediately before the crash is not an adequate time for any remedy actions.

Lee *et al.* [7] used the log-linear approach to model traffic conditions leading to crashes "precursor," and spatial dimension was added by using data from upstream and downstream detectors of crashes. Moreover, they used the speed profile captured by the detectors to estimate the actual crash time instead of using the reported crash time. They refined their analysis in a later study [8], and 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 average variation of speed across adjacent lanes was found to be insignificant.

A detailed study carried out by Golob and Recker [9] to analyze patterns in crash characteristics as a function of real-time traffic flow, nonliner canonical correlation analysis, and principal component analysis were used with three different sets of variables. The first set defined the lighting and weather condition; the second set defined crash the characteristics of collision type, location, and severity; and the third set consisted of real-time traffic flow variables. It was concluded that the median speed and the variation in speed between the left and interior lanes are related to the collision type. In addition, the inverse of the traffic volume has more influence than speed in determining the severity of the crash.

Matched case–control was used by Abdel-Aty *et al.* [10] to link real-time traffic flow variables collected by loop detectors and crash likelihood. Matched case–control was selected because it has the capability of eliminating the influence of location, time, and weather condition. They concluded that the average occupancy at the upstream station along with the coefficient of variation in speed at the downstream station, both during 5–10 min prior to the crash, were the most significant factors affecting crash likelihood prediction.

Abdel-Aty and Pemmanaboina [11] utilized principal component and logistic regression to estimate a weather index based on the rain readings at the weather station in the vicinity of the freeway. Using a matched case–control logit model, they were able to classify 58% of the crash cases using traffic loop data and the rain index.

Abdel-Aty and Pande [12] were able to capture 70% of the crashes using the Bayesian classifier-based methodology and the probabilistic neural network using different parameters of speed only. They found that the likelihood of a crash is significantly affected by the logarithms of the coefficient of variation in speed at the nearest crash station and two stations immediately preceding it in the upstream direction measured in the 5-min time slice of 10–15 min prior to the crash time.

Pande and Abdel-Aty [13] investigated lane-change-related crashes on a freeway using a classification tree procedure, and it was concluded that all sideswipe collisions and the angle crashes that occur on the inner lanes (leftmost and center lanes) of the freeway may be attributed to lane-changing maneuvers. The results also revealed that average speeds upstream and downstream of the crash location, the difference in occupancy on adjacent lanes, and the standard deviation of volumes and speed downstream of the crash location were the significant variables affecting crash occurrence.

Hourdakis *et al.* [14] developed an online crash-prone condition model using 110 live crashes, crash-related traffic events, and other contributing factors visualized from video traffic surveillance system (e.g., individual vehicle speeds and headways) over each lane in different places of the study area. They were able to detect 58% of the crashes successfully with a 6.8 false decision rate (where 6.8% of the crash cases were detected as noncrash cases).

Abdel-Aty *et al.* [15] used the random forest (RF) and multilayer perception neural network to test the transferability between different freeway corridors. Their model was successfully transferable from I-4 in Orlando to Dutch motorways.

Although a great effort has been performed in analyzing real-time data collected from ILDs in a safety framework, to the knowledge of the authors, no safety analysis has been carried out using traffic data from one of the most growing surveillance systems, i.e., the tag readers on toll roads (AVI). In this paper, for the first time, the identification of freeway locations with high real-time crash potential has been examined using real-time speed data collected from AVI systems. Stratified matched case–control logistic regression is used to classify the real-time traffic conditions measured by AVI into either leading or not leading to a crash. Matched case–control is used to control for the variability of different factors such as crash site, time, season, day of the week, etc. To select significant variables associated with the crash versus noncrash target variable, RF is utilized. RF has recently showed robustness in variable selection in transportation studies due to its stability over using a single decision tree [15], [16].

III. Description of a Roadway Network

A. General Description

The network that was studied is about 78 mi of freeways consisting of three toll roads in Orlando, FL, i.e., State Road (SR) 408, SR417, and SR528. SR408 is nearly 23 mi that extends from Florida's Turnpike in west Orlando to Challenger Parkway in the east. Traffic on SR408 is mostly commute traffic since it connects the east and the west of Central Florida and passes through the downtown area. SR417 and SR528 are 33 and 22 mi, respectively. SR417 connects Sanford to East Orlando with a high percentage of noncommuters travelling between the Orlando–Sanford International Airport, the Orlando International Airport, and the attraction areas; however it also includes many commuters from North Orlando. SR528 provides a crucial connection for residents and tourists between the attractions area, the Orlando International Airport, and the East Coast beaches and Cape Canaveral. As mentioned earlier, Central Florida's expressways are equipped with an AVI system for toll collection and travel time estimation. In this paper, Fig. 1 shows the expressway network and the AVI segments; the AVI segment tag readers are spaced according to toll plaza locations and locations of exits of interest to provide the travel time. Table I provides summary statistics of the AVI segments on each of the studied freeways: SR408 has 23 AVI segments on the eastbound and 24 on the westbound of average length of 0.9 mi; SR417 has 21 AVI segments on both directions, whereas SR528 has eight and nine AVI segments on the eastbound and westbound, respectively; SR528 has longer AVI segments that vary from 1.07 to 7.56 mi with an average length of approximately 3 mi.

IV. Data Description and Preparation

There were two sets of data used in the study, i.e., the expressway AVI archived data from SR408, SR417, and SR528 in Orlando and the corresponding crash data for year 2008. The Orlando–Orange County Expressway Authority (OOCEA) archives and maintains only the processed 1-min SMS and the estimated average travel time along the defined road segments. The unprocessed original time stamps of the tag readings are not available; these data are typically discarded after the travel time is processed due to privacy issues. The crash data were obtained from the road crash database maintained by the Florida Department of Transportation for year 2008.

The crashes have been assigned on each segment; three upstream segments and three downstream segments were identified to be considered in the preliminary analysis. The first upstream and downstream segments were named US1 and DS1, respectively. The subsequent upstream segments were named US2 and US3, respectively, whereas the subsequent segments in the downstream direction were named DS2 and DS3, respectively. The data structure is shown in Fig. 2.

AVI data corresponding to each crash case were extracted in the following process: for example, a crash occurred on February 7, 2008 (Thursday) at 2:00 P.M., SR408 eastbound, the crash segment G was identified using Geographic Information System (GIS) software, in addition to other six segments

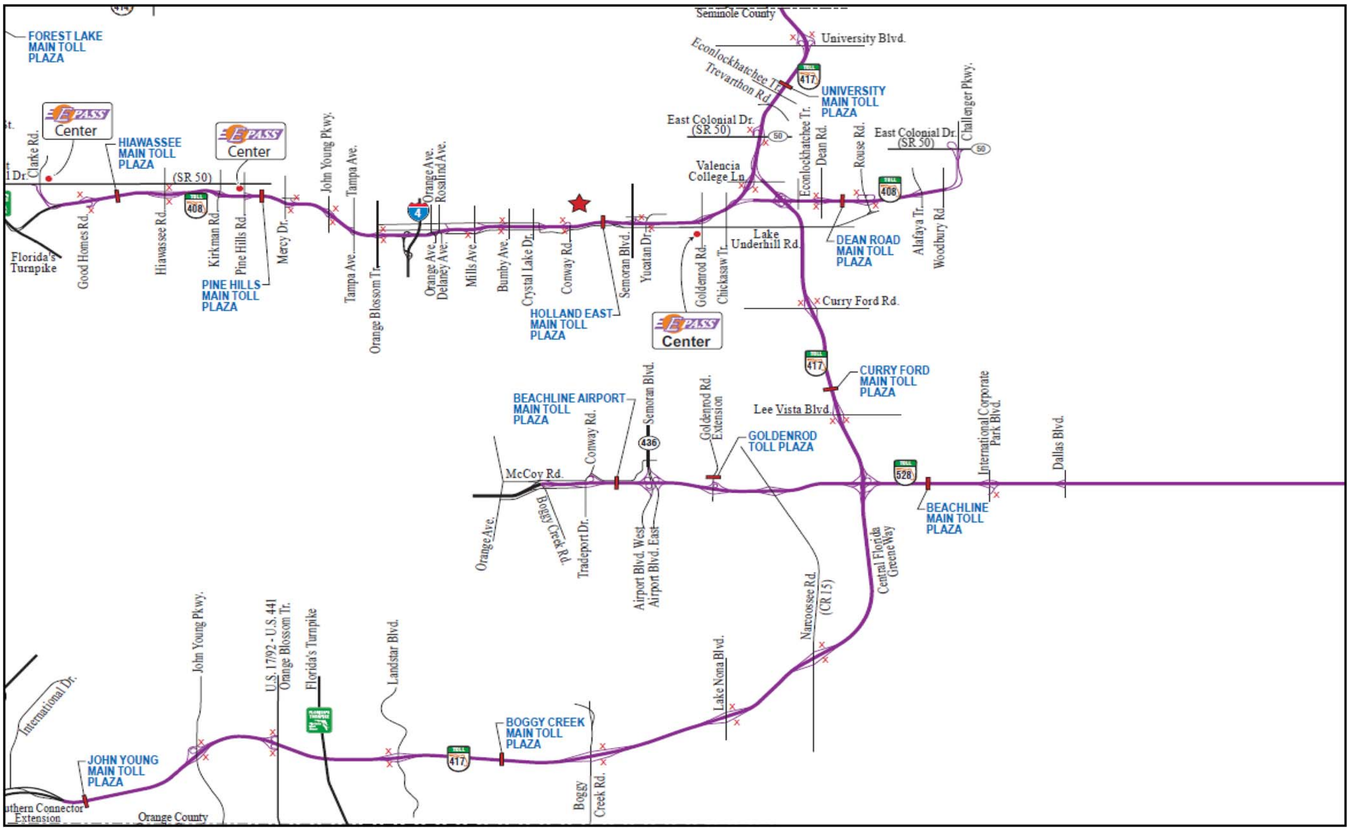

Fig. 1. Expressway network in Orlando. (Source: OOCEA System's Toll Facility Reference Manual).

TABLE I
SUMMARY STATISTICS FOR AVI SEGMENTS

Freeway			Automated Vehicle Identification Segments				
State Road ID		Length (mile)	Number of AVI Segments	Length (mile)			
				Min.	Mean	Max.	S.D.
SR408	EB	23	23	0.15	0.92	2.31	0.56
	WB		24	0.14	0.88	2.28	0.55
SR417	NB	33	21	0.21	1.49	2.98	0.75
	SB		21	0.25	1.46	2.87	0.70
SR528	EB	22	8	1.27	2.96	7.56	2.24
	WB		9	1.07	2.80	7.56	2.20

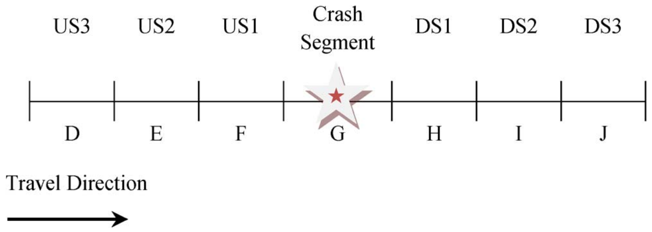

Fig. 2. AVI segment scheme.

(three in the upstream and three in the downstream directions) from 1:30 P.M. to 2:00 P.M. (30 min). Five randomly noncrash cases were also determined for the same location and time for different Thursdays, where no crashes were observed within 1 h of the original crash time.

The extracted 1-min speed data were aggregated to different aggregation levels of 2, 3, 5, and 10 min to investigate the best aggregation level that will give better accuracy in the modeling part. The 5-min aggregation level was found to be the best aggregation level. The 30-min speed data were divided into six time slices: time slice 1 represents the period between the crash time and 5 min prior to the crash time until time slice 6, which represents the interval between 25 and 30 min prior to the crash occurrence. Time slice 1 was discarded in the analysis since it will not provide enough time for successful intervention to reduce crash risk in a proactive safety management strategy. Moreover, the actual crash time might not be precisely known. Golob and Recker [17] discarded the 2.5 min of traffic data immediately preceding each crash reported time to avoid uncertainty of the actual crash time. In general, with the proliferation of mobile phones and closed-circuit television cameras on freeways, crash time is almost usually immediately identified.

In the modeling part, letters were assigned to each segment in accordance with the crash location to define the location of the crash segment with respect to the upstream/downstream segments. The assigned letters are D, E, F, G, H, I, and J, with G being the segment that the crash occurred on, segments F, E, and D are (in order) the closest segments to the crash segment in the upstream direction, whereas segments H, I, and J are (in order) the closest segments to the crash segment in the downstream direction, as shown in Fig. 2.

The average speeds, standard deviations of the speed, and logarithm of the coefficient of variation in speed were calculated over the 5-min time intervals. The nomenclature takes the following form: $XYS\_Z\beta$. XY takes the value of AV, SD, or CV for average, standard deviation, or coefficient of variation,

TABLE II
NUMBER OF CRASHES ON FREEWAY CORRIDORS

State Road ID		Number of crash cases	Number of non-crash cases
SR408	EB	180	720
	WB	160	640
	Both Directions	340	1360
	Total	1700	
SR417	NB	96	384
	SB	69	276
	Both Directions	165	660
	Total	825	
SR528	EB	82	328
	WB	83	332
	Both Directions	165	660
	Total	825	
Sub Total		670	2680
Total Observation		3350	

respectively. S stands for speed. Z represents AVI segments and takes values of D to J, whereas β takes the values from 2 to 6, which refer to the time slices.

Unlike ILD data, which are known to suffer from a high percentage of missing observations or bad reading, AVI data have less than 5% missing observations with no unreasonable values of speeds. The missing data for the speed were imputed by preserving the distribution of the original data, and then, the coefficient of variation was calculated. The final data set had a total of 105 variables consisting of three speed parameters for each of the seven AVI segments at five time intervals (time slices).

To examine the effect of short-term turbulence of traffic speed only, crashes involving driving under the influence of alcohol or drugs and distraction-related crashes were excluded from the crash data set. A total number of 670 crashes were considered in the analysis. Table II provides the number of crash/noncrash cases used in the study for the studied freeways.

V. METHODOLOGY

A. RF and Important Variable Selection

RF is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. The method combines Breiman's "bagging" idea and the random selection of features, as independently introduced by Ho [18] and Amit and Geman [19], to construct a collection of decision trees with controlled variation. RF has the capability of handling thousands of variables without deletion or deterioration of accuracy. Using ensembles of predictors for classification has proved to give more accurate results than using a single predictor. Moreover, RF has an advantage over the traditional classification trees of obtaining unbiased error estimates with no need for a separate cross-validation test data set. When a particular tree is grown from a bootstrap sample, one third of the training cases are left out and not used in the growing of the tree, and the left-out cases are called out-of-bag (OOB) data [20]. Abdel-Aty *et al.* [15] and Harb *et al.* [16] showed that RF may be used as a robust data mining technique to determine important variables in the transportation field.

The basis of the RF algorithm is first to choose the number of trees to grow and the number of m variables that would be selected to split each node to produce stable results and a minimum OOB error rate. The OOB error rate depends on two main components, namely, the correlation between any two trees in the forest and the strength of each individual tree in the forest. The correlation between any two trees in the forest increases the error rate, whereas increasing the strength of the individual trees decreases the forest error rate. Reducing m reduces both the correlation and the strength, and increasing it increases both. Somewhere in between is an optimal range of m that can be found using the OOB. Alternatively, a default value of the number of the candidate variables that will be randomly selected at each split m can be used for classification $m = (p)^{1/2}$, where p is the total number of variables. RF monitors the error rate for observations left out of the bootstrap sample OOB for each grown tree on a bootstrap sample. Fig. 3 shows the OOB error rate against different tree numbers; it is noted that 1000 trees are enough to achieve a constant minimum error rate and hence produce stable estimates.

Using the package "randomforest" [21] in the "R Software" [22], the RF model was estimated; using $m = 6$ variables that were randomly sampled as candidates at each split, the OOB error rate was found to be a minimum of 0.183 and 65.24% of variance explained by the model. Important variable selection based on the mean decreases Gini "IncNodePurity" as the node purity value increases the importance of the variable increase [23].

Examining RF with each data set for the three roadway corridors, most of the important variables were related to the segment that the crash occurred on, first upstream and downstream segments for SR408 and SR417. While SR528 did not return any reasonable results, SR408 and SR417 showed similar results in variable selection. Therefore, the combined data were considered in the final run. Fig. 3 shows the important variables from the RF produced for the combined data of SR408 and SR417 in both directions. The logarithm of the coefficient of variation in speed at crash segment G at time slice 2 from 5 to 10 min before the crash time (log_CVS_G2), average speed on downstream segment H in time slice 2 (AVS_H2), and the standard deviation of speed of the upstream segment between 5 and 10 min before the crash (SDS_F2) were found to be the most important variables according to Node Purity.

Hence, only variables related to the crash segment and the nearest upstream and downstream segments were included in the matched case–control modeling procedure.

B. Matched Crash–Noncrash Analysis

The study design utilized a matched case–control methodology, which is a simple and 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 [24]. For each selected crash case,

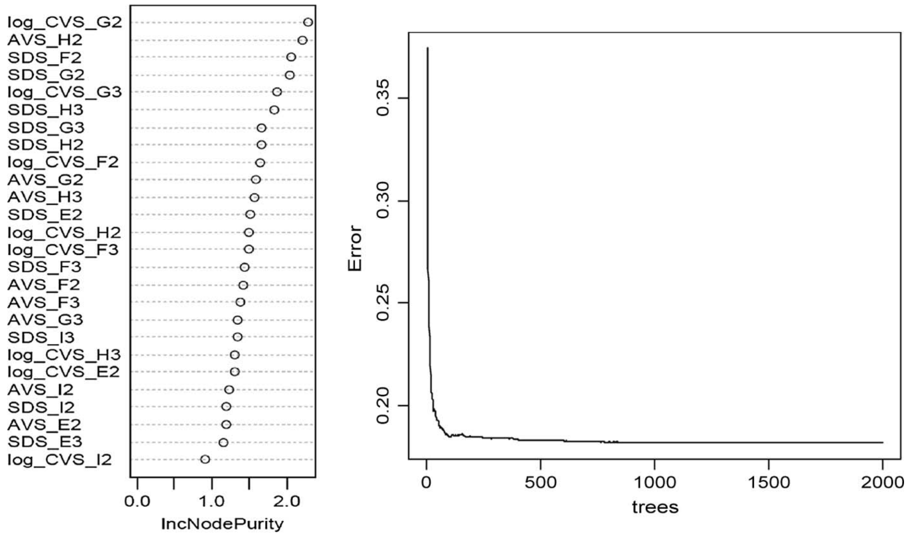

Fig. 3. Variable importance and OOB error.

randomly selected m controls (noncrash cases) were selected on account of matching factors of location, time of day, day of week, and season (Orlando has two distinct weather seasons, and matched noncrash cases are taken from the same season for each crash case). Although matched case–control can handle the confounding factors, other confounding factors such as individual drivers' behavior is not considered since the matching is for location and time variables only. Different $m:1$ ratios have been examined; $m = 4$ was found to give slightly better results. Previous studies show that negligible power is gained through adding controls beyond 3-to-1 matching [23]. Finally, the matched set (stratum) was formed of $m\ (4) + 1$ observations. Modeling is performed under the conditional likelihood principle of statistical theory accounting 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 there are N strata with one crash and m noncrash cases in stratum j, where $j = 1, 2, 3, \ldots, N$. $p_j\ (x_{ij})$ is the probability that the ith observation in the jth stratum is a crash, where the vector of k speed parameters $x_1, x_2, \ldots, x_k$ can be noted by $x_{ij} = (x_{1ij}, x_{2ij}, \ldots, x_{kij})$, $i = 0, 1, 2, \ldots, m$ and $j = 1, 2, \ldots, N$. This crash probability may be modeled by the following linear logistic regression model, as described in a study by Abdel-Aty *et al.* [10]:

$$\text{Logit}\{P_j(X_{ij})\} = \alpha_j + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_k X_{kij}. \quad (1)$$

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

To account for stratification in the analysis, a conditional likelihood is constructed. It should be noted that the crash probabilities cannot be estimated using (1) since conditional likelihood function $L(\beta)$ is independent of intercept terms $\alpha_1, \alpha_2, \ldots, \alpha_N$, and hence, the effects of matching variables cannot be estimated. This conditional likelihood function is expressed as follows:

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

However, the values of β parameters that maximize the conditional likelihood function given by (2) are also the estimates of the β coefficient in (1). These estimates are log odds ratio and may be used to approximate the relative risk of a crash.

In this analysis, procedure PHREG in SAS 9.2 is utilized. PHREG provides the hazard ratio, which is another term for relative risks used in SAS. In addition, a prediction model can be developed using the log odds ratios under this matched crash–noncrash analysis. This can be demonstrated by considering 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 jth strata on the k speed parameters. Using (1), the log odds ratio of crash occurrence due to speed parameters vector x_{1j} relative to traffic speed vector x_{2j} will have the following form:

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

The right-hand side of (3) is independent of α_j and can be calculated using estimated β coefficients. Thus, the aforementioned relative log odds ratio [left-hand side of (3)] may be utilized for predicting crashes by replacing X_{2j} with the vector of values of the traffic flow variables in the jth stratum of noncrash cases. One may use the simple average of all

TABLE III
OVERALL MODEL ESTIMATES AND FIT STATISTICS

Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Log_CVS_G2	1	0.21018	0.08901	5.5763	0.0182	1.234

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	3255.499	3249.915
AIC	3255.499	3251.915
SBC	3255.499	3256.253

TABLE IV
SR408 MODEL ESTIMATES AND FIT STATISTICS

Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Log_CVS_G2	1	0.27305	0.11513	5.6254	0.0177	1.314

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1536.143	1530.482
AIC	1536.143	1532.482
SBC	1536.143	1536.310

noncrash observations within the stratum for each variable. Let $\overline{x}_{2j} = (\overline{x}_{12j}, \overline{x}_{22j}, \overline{x}_{32j}, \ldots, \overline{x}_{k2j})$ denote the vector of mean values of noncrash cases of k variables within the jth stratum. Then, the log odds ratio of a crash relative to noncrash cases may be approximated by the following equation:

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

Hence, the log odds ratio can be used for predicting crashes by establishing a threshold value that attains the desirable crash classification accuracy.

As mentioned earlier, important variables were found to be related to the crash segment and two adjacent segments in the upstream and downstream directions at time slices 2 and 3 according to the results obtained in RF. These 18 variables only of AVS, SDS, and CVS were considered for further analysis using matched case–control.

VI. RESULTS AND DISCUSSION

In the preliminary analysis, a model was built for the combined data sets for all freeway sections. A univariate analysis was conducted first to check the significance of each variable. Different automatic search techniques of stepwise, forward, and backward were attempted to identify significant variables in multivariate analysis. These procedures were implemented to identify which terms were still statistically significant in the presence of other factors. Since variables not significant at 0.05 may be still associated with the response after adjusting for other covariates, any variable with $P < 0.25$ in the univariate results were considered eligible to enter into the multivariate model. There was an agreement between the three search techniques that the log of the coefficient of variation in speed of the crash segment at time slice 2 (Log_CVS_G2) is the only significant variable. This variable has a positive β coefficient, which means that the odds of a crash increase as the variation in speed increases and the average speed decreases at the segment of the crash at 5–10 min before the crash occurrence. Table III shows the hazard ratio for the significant variable. The hazard ratio is the exponent of the β coefficient, and it represents an estimate of the expected change in the risk ratio of having a crash versus noncrash per unit change in the corresponding factor. The hazard ratio of 1.234 means that the risk for a crash increases 1.234 times for each unit increase in Log_CVS_G2. It should be noted that the hazard ratio is multiplicative in nature for the continuous variables: this means that a two-unit increase in Log_CVS_G2 changes the risk by 1.234^2 or 1.52.

Since the combined data sets were collected from different populations, it was worth investigating each of the three freeway corridors separately. Therefore, other models were developed for each of the three freeways individually; univariate and multivariate analyses using automatic search techniques have been conducted.

All speed parameters related to SR528 were found to be statistically insignificant. It is worth mentioning that using toll tag readers to estimate travel times introduces a delay in generating observed travel times: for example, if a travel time of T minutes is observed, then that travel time applies to a vehicle that entered the segment T minutes ago. The length of the AVI segment plays a significant role in the SMS estimation: for example, if a number of vehicles entered a segment of 1-mi length, then it should be expected to have them exit the segment within 1 min in a normal traffic condition given that the speed is 60 mi/h. On the other hand, if the length of the AVI segment is 7 mi, then the estimated travel time applies to vehicles that entered the segment 7 min ago. Moreover, during times of rapid change in the segment travel time, this delay on long segments can reduce the usefulness of AVI data since the estimated measures will not be able to capture the variation in the SMS. In particular, this delay may mean that toll tag readers along long segments are ineffective tools for incident prediction.

The final model for SR408 resulted in one significant variable, i.e., LogCVS_G2 (log of the coefficient of variation in speed) from segment G (crash segment) at time slice 2 (5–10 min before the crash), as shown in Table IV. The variable has a positive β coefficient, which means that the odds of a crash increase as the variation of speed increases at the crash segment. This could be also explained that on average

TABLE V
SR417 Model Estimates and Fit Statistics

Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
SDS_G2	1	0.12163	0.05649	4.6357	0.0313	1.129
AVS_H2		-0.05683	0.02336	5.9183	0.0150	0.945

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	654.827	643.355
AIC	654.827	647.355
SBC	654.827	653.295

TABLE VI
Classification Results for SR408 and SR417

		SR408		
Frequency Percent Row % Column %		Predicted		Total
		0	1	
Actual	0	728 42.82 53.53 86.98	632 37.18 46.47 73.23	1360 80.00
	1	109 6.41 32.06 13.02	231 13.59 67.94 26.77	340 20.00
Total		837 49.24	863 50.76	1700 100.00

		SR417		
Frequency Percent Row % Column %		Predicted		Total
		0	1	
Actual	0	362 43.88 54.85 87.65	298 36.12 45.15 72.33	660 80.00
	1	51 6.18 30.91 12.35	114 13.82 69.09 27.67	165 20.00
Total		413 50.06	412 49.94	825 100.00

of 1-mi AVI segment, the increase in the standard deviation coupled with decrease in the average speed 5–10 min before the crash (since the coefficient of variation in speed includes the standard deviation as the nominator and the average speed as the denominator) may increase the likelihood of crash occurrence. This indicates an increase in the turbulence of traffic. The hazard ratio is found to be 1.314, which means that the crash risk increases 1.314 times for each unit increase in Log_CVS_G2. Moreover, the hazard ratio increased from 1.234 in the overall model to 1.314. This indicates that the risk for a crash increased by 8% for each unit increase in Log_CVS_G2 when SR528 and SR417 data sets were excluded from the model.

Table V provides the estimates and fit statistics for the model for SR417; two variables came out to be significant, i.e., SDS_G2 and AVS_H2. The standard deviation of speed of the crash segment at time slice 2 has a positive β coefficient, whereas the average speed of the adjacent downstream segment at time slice 2 has a negative β coefficient. This means that a high variation in speed at the crash segment with a decrease in the average speed in the downstream segment may increase the risk of having a crash at this location. A decrease in speed downstream might represent a queue buildup.

The results from both models suggest that the real-time crash prediction models are not transferable from one road to another due to the differences in the driver population and the structure of the AVI system; it is noteworthy that both roads have different types of road users, as stated before in the data description part. However, transferability might be possible for roadways with similar AVI system spacing and population. These findings were depicted by Pande *et al.* [25], although the data they used were collected from very similar loop detector structures in Central Florida (I-4 and I-95). They found that it might not be advisable to use the same model for two freeways with different driver populations or travel patterns.

To implement the estimated model in real-time application, sensitivity analysis is conducted. Table VI shows the sensitivity and specificity for the final models. Sensitivity is the proportion of crashes that are correctly identified as crashes, whereas specificity is the proportion of noncrashes that are correctly identified as noncrashes by the model [26]. Sensitivity and specificity can be calculated using the odds ratio given by (4). For example, the mean of two variables SDS_G2 and AVS_H2 of all four noncrash cases for the SR417 model was calculated within each matched set. The estimated vector of these noncrash means replaced the vector in (4) for the jth matched set. The odds ratio can be estimated by utilizing the β coefficients from the model in (4), where the vector is the actual observation in the data set. Sensitivity values were found to be 67.94% and 69.09%, whereas the two models achieved specificity values of 53.53% and 54.85% for SR408 and SR417, respectively, at a threshold equal to 1. Classification accuracy is considered good for all crash types, and accuracy would be expected to increase when evaluating specific crash types [27].

Both models have relatively high false-positive rates at a threshold of 1 (about 46% were classified as crashes incorrectly), whereas the false-negative rates were low (about 32% of crashes were classified as noncrashes). Different classification accuracy can be obtained by changing the threshold depending on the management strategy. The threshold should be carefully chosen in the real-world application; a large number of false alarms might affect the drivers' compliance with the system and hence reduce the effectiveness of the system. Nevertheless, advanced traffic management (ATM) objectives of reducing turbulence to improve operation can still be achieved even with a high percentage of false alarms. ITS strategies, such as variable speed limits, could be introduced without the drivers' knowledge of false alarms or not.

VII. Conclusion and Recommendations

While the most common application of AVI is ETC and travel time estimation, there is a promising traffic safety application in the context of ATM. This paper has implemented for the first time data collected from AVI in a real-time traffic safety analysis. AVI data were found to be promising in providing a measure of crash risk in real time. The operation-based management of expressways can benefit from the collected AVI traffic data not only to ease the congestion and enhance the operation but also to provide warnings of an increased risk situation on the crash risk measures identified in this study to increase safety on freeways and expressways.

Travel time and SMS data were collected from tag readers (AVI) of a total of 78 mi on the Central Florida expressway network in Orlando in 2008. Historical crash data were collected for the same period and study road sections. Utilizing RF for significant variable selection and stratified matched case–control to account for the confounding effects of location and time, the log odds of crash occurrence may be obtained, and hence, a proactive safety management system may be incorporated with existing ATIS.

The estimated speed collected from the AVI systems is different from that collected from ILDs, AVI systems measure the average speed of all vehicles occupying a given stretch of the road over some specified time period. Therefore, the AVI segment length plays an important role in estimating the SMS that will be used in any traffic safety management strategy. On one hand, the results suggest that the AVI data could only be useful if the AVI segments are within 1.5 mi on average; on the other hand, it has been found that the model is not easily transferable from one road to another unless the AVI structure and driver population are similar. The coefficient of variation in speed at the crash segment during 5–10 min prior to the time of the crash is found to be the most significant factor affecting the crash likelihood on a freeway with tag readers spaced 1 mi on average and mostly commute drivers, whereas the standard deviation of the speed at the crash segment and the average speed at the adjacent downstream segment were found to be the most significant factors on another freeway section with AVI segments length of an average of 1.5 mi with mixed types of road users.

All speed parameters obtained from AVI systems spaced on average at 3 mi apart were found to be statistically insignificant to identify crash-prone conditions. Although this paper has shown that AVI segments within 1.5 mi may be useful in real-time crash analysis, further investigation is needed to determine the exact cutoff and threshold values of the appropriate length of the AVI segment to be used as a guideline in ITS applications.

Acknowledgment

The authors would like to thank OOCEA and PBS&J for providing the AVI data that were used in this paper. All opinions and results are solely those of the authors.

References

[1] L. A. Klein, D. Gibson, and M. K. Mills, "Traffic detector handbook: Third edition—Volume II," Federal Highway Admin. (FHWA), Washington, DC, Rep. FHWA-HRT-06-108, 2006.

[2] M. Abdel-Aty and A. Pande, "Crash data analysis: Collective vs. individual crash level approach," *J. Safety Res.*, vol. 38, no. 5, pp. 581–587, 2007.

[3] S. Madanat and P. Liu, "A prototype system for real-time incident likelihood prediction," Transp. Res. Board, Nat. Res. Council, Washington, DC, IDEA Project Final Rep. (ITS-2), 1995.

[4] R. Hughes and F. Council, "On establishing relationship(s) between freeway safety and peak period operations: Performance measurement and methodological considerations," presented at the 78th Annu. Meeting Transportation Research Board, Washington, DC, 1999.

[5] C. Feng, "Synthesis of studies on speed and safety," presented at the 80th Annu. Meeting Transportation Research Board, Washington, DC, 2001.

[6] C. Oh, J. Oh, S. Ritchie, and M. Change, "Real time estimation of freeway accident likelihood," presented at the 80th Annu. Meeting Transportation Research Board, Washington, DC, 2001.

[7] C. Lee, F. Saccomanno, and B. Hellinga, "Analysis of crash precursors on instrumented freeways," presented at the 81st Annual Meeting Transp. Res. Board, Washington, DC, 2002.

[8] C. Lee, F. Saccomanno, and B. Hellinga, *Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic*. Washington, DC: Nat. Res. Council, 2003.

[9] T. Golob and W. Recker, "Relationships among urban freeway accidents, traffic flow, weather and lighting conditions," Inst. Transp. Stud., Univ. California, Berkeley, CA, California PATH Working Paper UCB-ITS-PWP-2001-19, 2001.

[10] M. Abdel-Aty, N. Uddin, F. Abdalla, A. Pande, and L. Hsia, "Predicting freeway crashes based on loop detector data using matched case–control logistic regression," *Transp. Res. Rec.*, vol. 1897, pp. 88–95, 2004.

[11] M. Abdel-Aty and R. Pemmanaboina, "Calibrating a real-time crash-prediction model using archived weather and ITS traffic data," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 2, pp. 167–174, Jun. 2006.

[12] M. Abdel-Aty and A. Pande, "Identifying crash propensity using specific traffic speed conditions," *J. Safety Res.*, vol. 36, no. 1, pp. 97–108, 2005.

[13] A. Pande and M. Abdel-Aty, "Assessment of freeway traffic parameters leading to lane-change related collisions," *Accid. Anal. Prev.*, vol. 38, no. 5, pp. 936–948, Sep. 2006.

[14] J. Hourdakis, V. Garg, P. Michalopoulos, and G. Davis, "Real-time detection of crash-prone conditions at freeway high-crash locations," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 1968, pp. 83–91, 2006.

[15] M. Abdel-Aty, A. Pande, A. Das, and W. Knibbe, "Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems," *Transp. Res. Rec.*, vol. 2083, pp. 153–161, 2008.

[16] R. Harb, X. Yan, E. Radwan, and X. Su, "Crash avoidance analysis using classification trees and random forest," presented at the 87th Annu. Meeting Transportation Research Board, Washington, DC, 2008.

[17] T. Golob and W. Recker, "A method for relating type of crash to traffic flow characteristics on urban freeways," *Transp. Res.—Part A, Policy Pract.*, vol. 38, pp. 53–80, 2004.

[18] T. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[19] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural Comput.*, vol. 9, no. 7, pp. 1545–1588, Oct. 1997.

[20] L. Breiman, "Some infinity theory for predictor ensembles," Statist. Dept., Univ. California, Berkeley, CA, Tech. Rep. 579, 2000.

[21] RandomForest: Breiman and Cutler's random forests for classification and regression. [Online]. Available: http://cran.r-project.org/web/packages/randomForest/

[22] R Development Core Team, R: A language and environment for statistical computing, Vienna, Austria: R Found. Stat. Comput., 2011. [Online]. Available: http://www.R-project.org/

[23] 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, p. 400, 2008.

[24] N. Breslow and N. Day, *Statistical Methods in Cancer Research*. Geneva, Switzerland: IARC, 1980.

[25] A. Pande, A. Das, M. Abdel-Aty, and H. Hassan, "Real-time crash risk estimation: Are all freeways created equal?" presented at the 90th Annu. Meeting Transp. Res. Board, Washington, DC, 2011.

[26] A. Agresti, *Categorical Data Analysis*, 2nd. New York: Wiley, 2002.

[27] A. Pande and M. Abdel-Aty, "Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways," presented at the TRB Annu. Meeting, Washington, DC, 2006, CD-ROM, Paper 06-0016.

Mohamed M. Ahmed received the B.Sc. degree in civil engineering from Al-Azhar University, Cairo, Egypt, in 2001 and the M.Sc. degree in transportation engineering from the University of Central Florida (UCF), Orlando, in 2009. He is currently working toward the Ph.D. degree with UCF.

He is currently a Research and Teaching Associate with UCF. His research interests include traffic safety analysis, intelligent transportation systems, and statistical and data mining applications in transportation engineering.

Mohamed A. Abdel-Aty received the B.Sc. and M.Sc. degrees in civil engineering from Alexandria University, Alexandria, Egypt, in 1985 and 1991, respectively, and the Ph.D. degree in transportation engineering from the University of California, Davis, in 1995.

He is a Professor of transportation engineering with the University of Central Florida (UCF). He is also the Program Director of Safety and Operation with the Center for Advanced Transportation Systems Simulation, UCF. His main expertise and interest is in the areas of traffic safety, travel demand analysis, and intelligent transportation systems. He is a leading traffic safety expert at both national and international levels. He has published more than 260 papers (140 in journals). His research in the real-time prediction of traffic crashes on freeways is recognized worldwide. He and his research team have moved freeway management from reactive incident detection to proactive incident prediction.

Dr. Abdel-Aty is a Registered Professional Engineer in the State of Florida. He is an Associate Editor of *Accident Analysis and Prevention*, the premier journal in safety. He is a member of the Editorial Board of the *Journal of Intelligent Transportation Systems*. He was the recipient of UCF's Distinguished Researcher of the Year Award in 2003 and UCF's Outstanding Graduate Teacher Award in 2007.