Accident Analysis and Prevention 43 (2011) 461–470

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Analyzing angle crashes at unsignalized intersections using machine learning techniques

Mohamed Abdel-Aty*, Kirolos Haleem

Department of Civil, Environmental & Construction Engineering, University of Central Florida, Orlando, FL 32816-2450, United States

ARTICLE INFO

Article history:
Received 5 November 2009
Received in revised form 5 September 2010
Accepted 3 October 2010

Keywords:
Data mining
Machine learning
Multivariate adaptive regression splines
MARS
Angle crash
Random forest
Unsignalized intersections
Crash prediction

ABSTRACT

A recently developed machine learning technique, multivariate adaptive regression splines (MARS), is introduced in this study to predict vehicles' angle crashes. MARS has a promising prediction power, and does not suffer from interpretation complexity. Negative Binomial (NB) and MARS models were fitted and compared using extensive data collected on unsignalized intersections in Florida. Two models were estimated for angle crash frequency at 3- and 4-legged unsignalized intersections. Treating crash frequency as a continuous response variable for fitting a MARS model was also examined by considering the natural logarithm of the crash frequency. Finally, combining MARS with another machine learning technique (random forest) was explored and discussed. The fitted NB angle crash models showed several significant factors that contribute to angle crash occurrence at unsignalized intersections such as, traffic volume on the major road, the upstream distance to the nearest signalized intersection, the distance between successive unsignalized intersections, median type on the major approach, percentage of trucks on the major approach, size of the intersection and the geographic location within the state. Based on the mean square prediction error (MSPE) assessment criterion, MARS outperformed the corresponding NB models. Also, using MARS for predicting continuous response variables yielded more favorable results than predicting discrete response variables. The generated MARS models showed the most promising results after screening the covariates using random forest. Based on the results of this study, MARS is recommended as an efficient technique for predicting crashes at unsignalized intersections (angle crashes in this study).

1. Introduction

Intersections are the connecting nodes for roadway networks. Few studies have addressed the safety of unsignalized intersections. An important reason is the inadequacy and difficulty to obtain data for analysis at these intersections, as well as the limited crash counts.

Statistical models are common tools for estimating safety performance functions of many transportation systems (e.g., Kulmala, 1995; Lord, 2000; Oh et al., 2003; Miaou and Lord, 2003; Caliendo et al., 2007) for identifying relationships between the crash occurrence and a set of explanatory covariates. Also, predicting crashes is another important application of safety performance functions that can help identify hazardous sites for further safety remedy. The most common probabilistic models used by safety analysts for modeling vehicle crashes are the traditional Poisson and Negative Binomial (NB). Since prediction is an essential objective of crash models, some studies that focused on developing models for mainly predicting vehicle crashes were Lord (2000), Xie et al. (2007) and Li et al. (2008). Researchers are always trying to introduce and develop statistical tools for effectively predicting crash occurrence.

This study aims at exploring the potential of applying a recently developed machine learning technique, the multivariate adaptive regression splines (MARS), for precise crash prediction. This was demonstrated through various applications of MARS via data collected at unsignalized intersections. Another objective is to explore the significant factors that contribute to angle crash occurrence at unsignalized intersections by utilizing a recently collected extensive dataset of 2475 unsignalized intersections.

2. Literature review

Using crash prediction models in safety studies can be found in previous literature (Hauer et al., 1988; Persaud and Dzbik, 1993; Sawalha and Sayed, 2006; Abdel-Aty and Radwan, 2000). NB models are widely used in predicting crash frequency at intersections (Harwood et al., 2000; Lord and Bonneson, 2006), since they can efficiently accommodate the over-dispersion characteristic often observed in crash data (Miaou and Lord, 2003; Park and Lord, 2007). For improving the NB framework, Anastasopoulos and Mannering

* Corresponding author. Tel.: +1 407 823 5657; fax: +1 407 823 3315.
E-mail address: mabdel@mail.ucf.edu (M. Abdel-Aty).

0001-4575/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2010.10.002

(2009) attempted to examine the random-parameters NB model, and found that it has the potential of providing a fuller understanding of the factors affecting crash frequency.

Recently, researchers have proposed new methods for modeling and predicting crashes that are comparable to NB and Poisson models. Examples of those methods are neural networks (Mussone et al., 1999; Abdelwahab and Abdel-Aty, 2002), Bayesian neural networks (Riviere et al., 2006; Xie et al., 2007) and support vector machine "SVM" (Li et al., 2008). However, neural networks models always suffer from their interpretation complexity, and sometimes they over-fit the data (Vogt and Bared, 1998). For this, Bayesian neural networks were introduced that can accommodate data over-fitting. As an example, Xie et al. (2007) applied the Bayesian neural networks in predicting crashes, and found that they are more efficient than NB models. Also, Li et al. (2008) applied a simpler technique than the Bayesian neural networks, which is SVM, to data collected on rural frontage roads in Texas. They fitted several models using different sample sizes, and compared the prediction performance of those models with the NB and Bayesian neural networks models. They found that SVM models are more efficient predictors than both NB and Bayesian neural networks models.

MARS is a multivariate non-parametric regression technique introduced by Friedman (1991). MARS is considered a nonparametric technique because it does not require priori assumption about the form of the relationship between the dependent and the independent variables. Moreover, it can reveal the required relationship in a piecewise regression function. This technique is effective while analyzing complex structures in the data such as nonlinearities and interactions; characteristics usually observed in crash data. Also, MARS is a regression-based technique that does not suffer from the "black-box" limitation, where the output is easily understood.

The application of the MARS technique can be found in previous studies (e.g., Put et al., 2004; Francis, 2003; Attoh-Okine et al., 2003). For example, Put et al. (2004) concluded that MARS has some advantages compared to the more traditionally complicated techniques such as neural networks.

The following studies address safety analysis of angle crashes at unsignalized intersections. In their study, Najm et al. (2001) showed that 41.6% of angle crashes occurred at signalized intersections, 36.3% at stop-signed intersections, and 22.1% at intersections with no controls or other control types. Studies done by Datta and Dutta (1990) and Datta (1991) concluded that the number of right-angle crashes decreased at an intersection when the traffic control was changed from a stop sign to a traffic signal.

From the aforementioned studies, it can be noted that MARS has promising advantage of accounting for nonlinearities in crash data analysis, and for improving prediction. This was one of the motivations of this study.

3. Methodology

3.1. Multinomial adaptive regression splines model characteristics

According to Abraham et al. (2001), splines are defined as "an innovative mathematical process for complicated curve drawings and function approximation". To develop any spline, the X-axis representing the space of predictors is broken into number of regions. The boundary between successive regions is known as a knot (Abraham et al., 2001). While it is easy to draw a spline in two dimensions (using linear or quadratic polynomial regression models), manipulating the mathematics in higher dimensions can be accomplished using the "basis functions", which are the elements of fitting any MARS model.

According to Friedman (1991), the MARS method is a local regression method that uses a series of basis functions to model complex (such as nonlinear) relationships. The global MARS model is defined according to Put et al. (2004) as shown in Eq. (1):

$$\hat{y} = a_0 + \sum_{m=1}^{M} a_m B_m(x) \tag{1}$$

where $\hat{y}$ is the predicted response; a_0 is the coefficient of the constant basis function; $B_m(x)$ is the mth basis function, which can be a single spline function or an interaction of two (or more) spline functions; a_m is the coefficient of the mth basis function; and M is the number of basis functions included in the MARS model.

According to Put et al. (2004), there are three main steps to fit a MARS model. The first step is a constructive phase, in which basis functions are introduced in several regions of the predictors and are combined in a weighted sum to define the global MARS model (as indicated in Eq. (1)). This global model usually contains many basis functions, which can cause an over-fitting. The second step is the pruning phase, in which some basis functions of the over-fitting MARS model are deleted. In the third step, the optimal MARS model is selected from a sequence of smaller models.

As indicated in Put et al. (2004), the first step for describing the three MARS steps is created by continually adding basis functions to the model. Basis functions in MARS consist of a single spline function or a product (interaction) of two (or more) spline functions for different predictors (Put et al., 2004). Those basis functions are added in a "two-at-a-time" forward stepwise procedure, which selects the best pairs of spline functions in order to improve the model. Each pair consists of one left-sided and one right-sided truncated function defined by a given knot location, as shown in Eqs. (2) and (3), respectively.

$$[-(x-t)_+^q] = \begin{cases} (t-x)^q; & x < t \\ 0; & \text{otherwise} \end{cases} \tag{2}$$

$$[+(x-t)_+^q] = \begin{cases} (x-t)^q; & x > t \\ 0; & \text{otherwise} \end{cases} \tag{3}$$

From Put et al. (2004), it is to be noted that the search for the best predictor and knot location is performed in an iterative process. The predictor, as well as knot location having the most contribution to the model, are selected first. Also, at the end of each iteration, the introduction of an interaction is checked for model improvement.

The second step is the pruning step, where a "one-at-a-time" backward deletion procedure is applied in which the basis functions with the least contribution to the model are eliminated. This pruning is based on the generalized cross-validation (GCV) criterion (Friedman, 1991). The GCV criterion is used to find the overall best model from a sequence of fitted models, where a larger GCV value tends to produce a smaller model, and vice versa. The GCV criterion is estimated using Eq. (4).

$$GCV(M) = \frac{1}{N} \frac{\sum_{i=1}^{N} (y_i - \hat{y})^2}{(1 - C(M)/N)^2} \tag{4}$$

where N is the number of observations; y_i is the response for observation i; $\hat{y}$ is the predicted response for observation i; and $C(M)$ is a complexity penalty function, which is defined as shown in Eq. (5):

$$C(M) = M + dM = M(1 + d) \tag{5}$$

where M is the number of non-constant basis functions (i.e., all terms of Eq. (1) except for "a_0"); and d is a user-defined cost for each basis function optimization. According to Put et al. (2004), the higher the cost d is, the more basis functions will be eliminated.

Finally, the third step is used for selecting the optimal MARS model. This selection is based on an evaluation of the prediction characteristics of the different fitted MARS models.

3.2. Random forest technique

Since the random forest technique was attempted in this study in conjunction with MARS, a brief discussion about this technique is presented. Random forest is one of the most recent and promising machine learning techniques proposed by Breiman (2001). It is well known for selecting important variables from a set of variables. To select the important covariates, the R package provides the mean decrease Gini "IncNodePurity" diagram. This diagram shows the node purity value for every covariate (node) of a tree by means of the Gini index (Kuhn et al., 2008). A higher node purity value represents a higher variable importance. For more details about the random forest technique, readers can refer to Breiman (2001) and Harb et al. (2009).

For the NB framework, two relevant sources addressing the methodological approach of NB models can be found in Miaou (1994) and Poch and Mannering (1996).

4. Prediction performance assessment

To examine the significant prediction performance of the MARS technique, there were two main evaluation criteria used, the mean absolute deviance (MAD) and the mean square prediction error (MSPE). The MAD and MSPE criteria were also used in previous studies (Lord and Mahlawat, 2009; Jonsson et al., 2009; Li et al., 2008). Eqs. (6) and (7) show how to evaluate MAD and MSPE, respectively. In this study, the MAD and MSPE values were normalized by the average of the response variable. This was done because crash frequency has higher range; hence, error magnitude is relatively higher. However, considering the natural logarithm of crash frequency results in a smaller range; hence, error magnitude is relatively lower. By this, the comparison between the MARS models using discrete and the continuous responses holds. A better prediction performance of the model is obtained by having smaller values of MAD and MSPE:

$$MAD = \frac{1}{n * \bar{y}} \rightleftarrows \sum |y_i - \mu_i| \quad (6)$$

$$MSPE = \frac{1}{n * \bar{y}} \sum (y_i - \mu_i)^2 \quad (7)$$

where n is the sample size in the prediction dataset; y_i is the observed crash frequency for intersection i; μ_i is the predicted crash frequency for intersection i; and $\bar{y}$ is the average of the response variable.

5. MARS applications

There were three main applications performed in this study using the MARS technique. The first one dealt with a comparison between the fitted NB and the MARS models while treating the response in each of them as a discrete variable (crash frequency). For the scope of this study, the traditional NB framework was used, and the training dataset used for calibration was 70% of the total data, while the remaining 30% was used for prediction (validation). Thus, two NB angle crash frequency models were developed for 3- and 4-legged unsignalized intersections using a training dataset (1732 intersections) for four-year crash data from 2003 till 2006. This crash type was specifically selected, as it is considered one of the most frequent and severe crash types occurring at unsignalized intersections (Summersgill and Kennedy, 1996; Layfield, 1996; Agent, 1988). Afterwards, using the same significant predictors in each of the two models, two MARS models were fitted, and compared to the corresponding NB models. The prediction assessment criteria were performed on a test dataset (743 intersections) for the four-year crash data as well.

The second application dealt with treating the response in the fitted MARS models as a continuous variable. This was considered while considering the natural logarithm of crash frequency, and the same training and test datasets were used. This application was proposed due to the high prediction capability of the MARS technique while dealing with continuous responses, as shown by Friedman (1991) and indicated in Kim (2000). Hence, the passage from fitting a MARS model with a discrete response to a MARS model with a continuous response was done in an organized manner for the main core of this study. This was done to assess using MARS with continuous responses for improving crash prediction, so that researchers would gain better understanding on the most efficient way to fit a predictive MARS model.

The third application dealt with combining MARS with the random forest technique for screening the variables before fitting a MARS model. Then, a comparison between the MARS models (with the covariates initially screened using random forest) and the MARS models (with the covariates initially screened from the NB model) was held.

6. Data collection and preparation

The analysis conducted in this study was performed on 2475 unsignalized intersections collected from six counties in the state of Florida. The county selection was based on its geographic location in Florida, so as to represent the Northern, Southern, Central, Eastern and Western parts of the state.

The Crash Analysis Reporting (CAR) system maintained by the FDOT (Florida Department of Transportation) was used to identify all the state roads (SRs) in those six counties. Then, the random selection method was used for choosing some state roads. Unsignalized intersections were then identified along these randomly selected SRs using "Google Earth" and "Video Log Viewer Application". This application is an advanced tool developed by FDOT, and has two important features, the "right view" and the "front view". The "right view" option provides the opportunity of identifying whether a stop sign and a stop line exist or not. The "front view" feature provides the opportunity of identifying the median type as well as the number of lanes per direction more clearly.

Geometric, traffic and control fields of the collected intersections were merged with the Roadway Characteristic Inventory database (RCI) for the 4 years (2003, 2004, 2005 and 2006). The RCI database – which is developed by FDOT – includes physical and administrative data, such as functional classification, pavement, shoulder and median data related to the roadway (the new web-based RCI application). The angle crash frequency for those identified unsignalized intersections was determined from the CAR database, and a finally merged dataset was created. Table 1 shows a summary statistics for angle crashes in the calibration (training) and validation (test) databases for both 3- and 4-legged intersections. It can be noticed that there is an over-dispersion in the datasets; hence, the use of the NB framework was appropriate.

A full description of the important variables used in the NB and MARS modeling procedures for 3- and 4-legged unsignalized intersections is shown in Table 2. From Table 2, regular unsignalized intersections are those intersections having distant stretches on the minor approaches, whereas access points include parking lots at plazas and malls, and driveways that are feeding to the major approach. Due to the unavailability of AADT on most minor roads, an important traffic covariate explored in this study is the surrogate measure for AADT on the minor approach, which is represented by the number of through lanes on this approach.

Table 1
Summary statistics for angle crashes in the training and test databases in 2003–2006.

	Three-legged training dataset in 4 years "2003–2006"	Four-legged training dataset in 4 years "2003–2006"	Three-legged test dataset in 4 years "2003–2006"	Four-legged test dataset in 4 years "2003–2006"
Number of intersections	1341	391	596	147
Total number of crashes	1197	1008	585	312
Mean crash frequency per intersection	0.892	2.578	0.981	2.122
Crash standard deviation per intersection	1.734	3.856	2.079	2.808

7. Results

7.1. Modeling angle crash frequency at 3- and 4-legged unsignalized intersections using the NB technique

Using SAS (2002), the NB angle crash frequency model for both 3- and 4-legged unsignalized intersections is shown in Table 3. This table includes the generalized *R*-square criterion as a goodness-of-fit statistic.

7.1.1. Three-legged model interpretation

From Table 3, there is a statistical significant increase in angle crashes with the increase in the logarithm of AADT (which inherently means an increase in traffic volume). As AADT relatively increases, vehicles coming from the minor approach find it difficult to cross the major road due to congestion; hence, angle crash risk might increase.

There is a reduction in angle crashes with the increase in the logarithm of the upstream distance to the nearest signalized intersection. This is expected since there is enough spacing for vehicles on the minor approach to cross the major road before the platoon of vehicles from the signalized intersection hinders the gap acceptance; and thus, angle crash risk decreases. In other words, angle crashes are lower when vehicles' distribution is scattered, i.e., as the distance increases, platoons generally dissolve (particularly when overtaking is allowed).

There is an increase in angle crashes with the increase in truck percentage on the major road. This is anticipated due to possible vision blockage caused by trucks; thus, angle crash risk could increase.

Compared to access points, regular unsignalized intersections have longer stretches on the minor approach; thus, angle crashes increase, and as shown in Table 3, the increase is statistically significant. Also, ramp junctions have high angle crashes due to traffic turbulence in merging areas.

The existence of one left turn lane on each major road direction significantly increases angle crashes, compared to no left turn lanes. This is due to a high possible conflict pattern between left turning vehicles from both minor and major approaches.

Compared to open medians, undivided medians have the least significant decrease in angle crashes due to the reduction in conflict points.

Compared to the eastern part of Florida (represented by Brevard County), the highest increase in angle crashes occurs in the western part (represented by Hillsborough County), followed by the northern part (represented by Leon County), then the southern part (represented by Miami-Dade County), and finally the central part (represented by Orange and Seminole Counties).

7.1.2. Four-legged model interpretation

From Table 3, as found in the 3-legged model, increasing the logarithm of AADT significantly increases angle crashes.

There is a significant increase in angle crashes with the increase in the logarithm of the spacing between successive unsignalized intersections. This result contradicts to the study done by Layfield (1996), who concluded that there were fewer right-angle crashes for a relatively large spacing between the minor approaches of urban unsignalized intersections.

Similar to the 3-legged model, compared to access points, regular unsignalized intersections as well as unsignalized intersections next to railroads experience a significant increase in angle crashes.

The existence of one left and right turn lane on each major road direction significantly increases angle crashes, compared to no left and right turn lanes, respectively. Once more, this is due to a high possible conflict pattern between vehicles crossing from both minor and major approaches.

Two-way left turn lanes as well as undivided medians on the major approach increase angle crashes, when compared to open medians, and the increase is statistically significant for undivided medians. This shows the hazardous effect of having two-way left turn lanes for 4-legged intersections. This conforms to the study done by Phillips (2004) who found that two-way left turn lanes experience more crashes than raised medians.

As the size of intersections increase, angle crashes increase. This is anticipated due to the higher angle crash risk maneuver at relatively bigger intersections. Increasing intersection size is mainly associated with increasing number of lanes, which means increasing total vehicular flow on each approach, which was found to increase angle crash frequency. Intersections with 3 total lanes on the minor approach have the only significant increase.

Similar to the 3-legged model, the highest increase in angle crashes occurs in the western part (represented by Hillsborough County), followed by the northern part (represented by Leon County), then the southern part (represented by Miami-Dade County) when compared to the eastern part (represented by Brevard County). The central part (represented by Orange and Seminole Counties) has no significant effect on angle crashes.

To show the result of the MARS model and different basis functions' coefficients, the MARS model for 4-legged angle crash frequency is presented in Table 4 as an example for illustration purposes.

Table 4 shows the different basis functions in the MARS model. From this table, it is noticed that there is an interaction term. Hence, the two variables forming the interaction term should be interpreted together. The interaction term is between Hillsborough County and unsignalized intersections with three total lanes on the minor approach. The equation representing this interaction term is:

"$-5.5343 * \text{Hills\_County} - 6.3123 * \text{Size\_Lanes\_3} + 7.4050 * \text{Hills\_County} * \text{Size\_Lanes\_3}$".

The interpretation for the formed equation is described as follows: for the case of Hillsborough (i.e., Hills_County = 1), the equation becomes:

"$(-6.3123 + 7.4050) * \text{Size\_Lanes\_3} - 5.5343$".

Table 2
Variables description for 3- and 4-legged unsignalized intersections.

Variable Description	Variable Levels for 3 Legs	Variable Levels for 4 Legs
Crash location in any of the 6 counties	Orange, Brevard, Hillsborough, Miami-Dade, Leon and Seminole	Orange, Brevard, Hillsborough, Miami-Dade, Leon and Seminole
Existence of stop sign on the minor approach	= 0; if no stop sign exists; = 1; if stop sign exists	= 0; if no stop sign exists; = 1; if only one stop sign exists on one of the minor approaches; = 2; if one stop sign exists on each minor approach
Existence of stop line on the minor approach	= 0; if no stop line exists; = 1; if stop line exists	= 0; if no stop line exists; = 1; if only one stop line exists on one of the minor approaches; = 2; if one stop line exists on each minor approach
Existence of crosswalk on the minor approach	= 0; if no crosswalk exists; = 1; if crosswalk exists	= 0; if no crosswalk exists; = 1; if only one crosswalk exists on one of the minor approaches; = 2; if one crosswalk exists on each minor approach
Existence of crosswalk on the major approach	= 0; if no crosswalk exists; = 1; if one crosswalk exists on one of the major approaches; = 2; if one crosswalk exists on each major approach	= 0; if no crosswalk exists; = 1; if one crosswalk exists on one of the major approaches; = 2; if one crosswalk exists on each major approach
Control type on the minor approach	= 1; if stop sign exists (1-way stop); = 3; if no control exists; = 5; if yield sign exists	= 2; if stop sign exists on each minor approach (2-way stop); = 3; if no control exists on both minor approaches; = 4; if stop sign exists on the first minor approach, and no control on the other
Size of the intersection[a]	= 1; for "1 × 2", "1 × 3" and "1 × 4" intersections; = 2; for "2 × 2" and "2 × 3" intersections; = 3; for "2x4", "2x5" and "2 × 6" intersections; = 4; for "2 × 7" and "2x8" intersections; = 5; for "3 × 2", "3 × 3", "3 × 4", "3 × 5", "3 × 6" and "3x8" intersections; = 6; for "4 × 2", "4 × 4", "4 × 6" and "4 × 8" intersections	= 2; for "2 × 2" and "2 × 3" intersections; = 3; for "2x4", "2x5" and "2 × 6" intersections; = 4; for "2 × 7" and "2x8" intersections; = 5; for "3 × 2", "3 × 3", "3 × 4", "3 × 5", "3 × 6" and "3 × 8" intersections; = 6; for "4 × 2", "4 × 4", "4 × 6" and "4 × 8" intersections
Type of unsignalized intersection	= 1; for access point (driveway) intersections; = 2; for ramp junctions; = 3; for regular intersections; = 4; for intersections close to railroad crossings[b]	= 1; for access point (driveway) intersections; = 3; for regular intersections; = 4; for intersections close to railroad crossings[b]
Number of right turn lanes on the major approach	= 0; if no right turn lane exists; = 1; if one right turn lane exists on only one direction; = 2; if one right turn lane exists on each direction[c]	= 0; if no right turn lane exists; = 1; if one right turn lane exists on only one direction; = 2; if one right turn lane exists on each direction
Number of left turn lanes on the major approach	= 0; if no left turn lane exists; = 1; if one left turn lane exists on only one direction; = 2; if one left turn lane exists on each direction[d]	= 0; if no left turn lane exists; = 1; if one left turn lane exists on only one direction; = 2; if one left turn lane exists on each direction
Number of left turn movements on the minor approach	= 0; if no left turn movement exists; = 1; if one left turn movement exists	= 0; if no left turn movement exists; = 1; if one left turn movement exists on one minor approach only; = 2; if one left turn movement exists on each minor approach
Land use at the intersection area	= 1; for rural area; = 2; for urban/suburban areas	= 1; for rural area; = 2; for urban/suburban areas
Median type on the major approach	= 1; for open median; = 2; for directional median; = 3; for closed median; = 4; for two-way left turn lane; = 5; for markings;= 6; for undivided median; = 7; for mixed median[e]	= 1; for open median; = 4; for two-way left turn lane; = 6; for undivided median
Median type on the minor approach	= 1; for undivided median, two-way left turn lane and markings; = 2; for any type of divided median	= 1; for undivided median, two-way left turn lane and markings; = 2; for any type of divided median
Skewness level	= 1; if skewness angle <= 75 degrees; = 2; if skewness angle > 75 degrees	= 1; if skewness angle <= 75 degrees; = 2; if skewness angle > 75 degrees
Posted speed limit on the major road	= 1; if posted speed limit < 45 mph; = 2; if posted speed limit >= 45 mph	= 1; if posted speed limit < 45 mph; = 2; if posted speed limit >= 45 mph

Table 2 (*Continued*)

Variable description	Variable levels for 3 legs	Variable levels for 4 legs
Number of through lanes on the minor approach[f]	= 1; if one through lane exists; = 2; if two through lanes exist; = 3; if more than two through lanes exist	= 2; if two through lanes exist; = 3; if more than two through lanes exist
Natural logarithm of the section annual average daily traffic on the major road; Natural logarithm of the upstream and downstream distances (in feet) to the nearest signalized intersection from the unsignalized intersection of interest; Left shoulder width near the median on the major road (in feet); Right shoulder width on the major road (in feet); Percentage of trucks on the major road; Natural logarithm of the distance between 2 successive unsignalized intersections[g]		

[a] The first number represents total number of approach lanes for the minor approach, and the second number represents total number of through lanes for the major approach.
[b] Railroad crossing can exist upstream or downstream the intersection of interest.
[c] One right turn lane on each major road direction for 3-legged unsignalized intersections: Two close unsignalized intersections, one on each side of the roadway, and each has one right turn lane. The extended right turn lane of the first is in the influence area of the second.
[d] One left turn lane on each major road direction for 3-legged unsignalized intersections: One of these left turn lanes is only used as U-turn.
[e] Mixed median is directional from one side, and closed from the other side (i.e., allows access from one side only).
[f] Surrogate measure for AADT on the minor approach.
[g] Continuous variables.

Table 3
Angle crash frequency model at 3- and 4-legged unsignalized intersections.

Variable description	Three-legged model		Four-legged model	
	Estimate[a]	*P*-Value	Estimate[a]	*P*-Value
Intercept	−7.1703 (1.3369)	<0.0001	−9.0650 (1.6736)	<0.0001
Natural logarithm of AADT on the major road	0.6741 (0.1120)	<0.0001	0.7151 (0.1662)	<0.0001
Natural logarithm of the upstream distance to the nearest signalized intersection	−0.0878 (0.0493)	0.0747	N/S[b]	
Natural logarithm of the distance between 2 successive unsignalized intersections	N/S		0.1200 (0.0604)	0.0471
Percentage of trucks on the major road	0.0272 (0.0168)	0.1049	N/S	
Unsignalized intersections close to railroad crossings	0.4368 (0.5317)	0.4114	1.0322 (0.3608)	0.0042
Regular unsignalized intersections	0.4069 (0.1193)	0.0007	0.4959 (0.1341)	0.0002
Unsignalized ramp junctions	0.5238 (0.3137)	0.0949	N/A[d]	
Access point unsignalized intersections (Driveways)	–[c]		–[c]	
One left turn lane exists on each major road direction	0.3495 (0.1754)	0.0463	0.4647 (0.2067)	0.0246
One left turn lane exists on only one major road direction	0.1642 (0.1324)	0.2149	0.6440 (0.2420)	0.0078
No left turn lane exists on the major approach	–[c]		–[c]	
One right turn lane exists on each major road direction	N/S		0.5842 (0.2678)	0.0292
One right turn lane exists on only one major road direction	N/S		0.0869 (0.2149)	0.6860
No right turn lane exists on the major approach	N/S		–[c]	
One left turn exists on any of the minor approaches	−0.6274 (0.2112)	0.0030	N/S	
No left turn lane exists on the minor approach	–[c]		N/S	
Mixed median exists on the major approach	−0.7215 (0.2795)	0.0099	N/A	
Undivided median exists on the major approach	−0.4342 (0.1504)	0.0039	0.3488 (0.2144)	0.1038
Marking exists on the major approach	−0.3797 (0.3128)	0.2248	N/A	
Two-way left turn lane exists on the major approach	−0.3779 (0.1891)	0.0457	0.0059 (0.1828)	0.9743
Closed median exists on the major approach	−0.5805 (0.2529)	0.0217	N/A	
Directional median exists on the major approach	−0.6773 (0.2874)	0.0184	N/A	
Open median exists on the major approach	–[c]		–[c]	
Posted speed limit on major road >= 45 mph	0.2201 (0.1156)	0.0568	N/S	
Posted speed limit on major road < 45 mph	–[c]		N/S	
"4 × 2", "4 × 4", "4 × 6" and "4 × 8" intersections	N/S		0.0443 (0.5968)	0.9408
"3 × 2", "3 × 3", "3 × 4", "3 × 5", "3 × 6" and "3 × 8" intersections	N/S		0.9531 (0.3527)	0.0069
"2 × 7" and "2 × 8" intersections	N/S		0.8813 (0.7924)	0.2660
"2 × 4", "2 × 5" and "2 × 6" intersections	N/S		0.2661 (0.2806)	0.3430
"2 × 2" and "2 × 3" intersections	N/S		–[c]	
Dummy variable for Seminole County	0.1889 (0.2394)	0.4302	−0.0427 (0.2795)	0.8786
Dummy variable for Orange County	0.6930 (0.1911)	0.0003	0.0604 (0.2669)	0.8211
Dummy variable for Miami-Dade County	0.7522 (0.2104)	0.0004	1.0695 (0.2575)	<0.0001
Dummy variable for Leon County	0.8489 (0.1985)	<0.0001	0.5336 (0.2786)	0.0555
Dummy variable for Hillsborough County	1.0528 (0.1988)	<0.0001	1.1046 (0.2304)	<0.0001
Dummy variable for Brevard County	–[c]		–[c]	
Dispersion	1.1442 (0.1113)		0.8379 (0.1043)	
Generalized *R*-square[e]	0.19		0.31	

[a] Standard error in parentheses.
[b] N/S means not significant.
[c] Base case.
[d] N/A means not applicable.
[e] Generalized *R*-square = 1 − (residual deviance/null deviance): the residual deviance is equivalent to the residual sum of squares in linear regression, and the null deviance is equivalent to the total sum of squares (Zuur et al., 2007).

Table 4
Angle crash frequency model at 4-legged unsignalized intersections using MARS.

Basis function	Basis function description	Estimate[a]	*P*-Value
Intercept	Intercept	2.1314 (5.3912)	0.6928
Log_AADT	Natural logarithm of AADT on the major road	0.6831 (0.5134)	0.1840
Hills_County	Hillsborough County	−5.5343 (1.9559)	0.0049
Orange_County	Orange County	−1.4406 (0.4560)	0.0017
Size_Lanes_3	"3 × 2", "3 × 3", "3 × 4", "3 × 5", "3 × 6" and "3 × 8" intersections	−6.3123 (2.2146)	0.0046
Acc_Point	Access points	−1.3737 (0.3382)	<0.0001
Hills_County[a] Size_Lanes_3	An interaction term	7.4050 (1.8259)	<0.0001
$(\text{Log\_AADT} - 10.389)_+$	A truncated power basis function for "Log_AADT" at "10.389"	6.4480 (1.3054)	<0.0001
$(\text{Log\_AADT} - 11.112)_+$	A truncated power basis function for "Log_AADT" at "11.112"	−25.5042 (7.3651)	0.0005
	Generalized *R*-square	0.52	

[a] Standard error in parentheses.

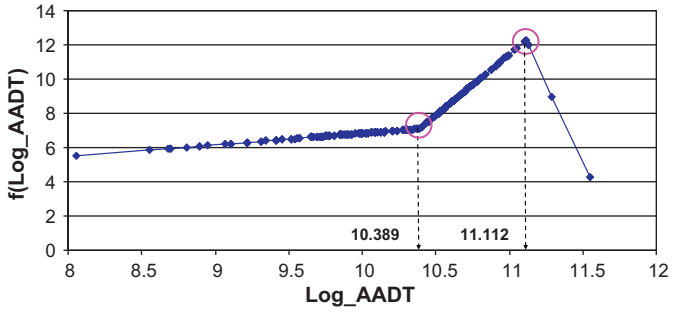

Fig. 1. Plot of the Basis function for "Log_AADT".

The formed equation can be simplified as "1.0927 * Size_Lanes_3 – 5.5343". Thus, the individual coefficient of "Size_Lanes_3" is "1.0927". This means that, in Hillsborough County, the angle crash frequency increases for intersections with three total lanes on the minor approach, when compared to other intersection sizes used in the analysis.

Also, from Table 4, it is noted that there is a nonlinear performance for the continuous variables "Log_AADT", as shown in its truncated basis function at "10.389" and "11.112". In order to understand the nonlinear function of "Log_AADT", a plot for its basis function is shown in Fig. 1. The basis function "*f*(Log_AADT)" according to the fitted MARS model is:

$$\text{"0.6831} * \text{Log\_AADT} + 6.4480 * (\text{Log\_AADT} - 10.389)_+ - 25.5042 * (\text{Log\_AADT} - 11.112)_+\text{"}.$$

As previously shown in Eq. (3), the term "$(\text{Log\_AADT} - 10.389)_+$" equals "Log_AADT – 10.389" when Log_AADT > 10.389, and zero, otherwise. The same also applies for "$(\text{Log\_AADT} - 11.112)_+$". By this, the plot in Fig. 1 can be formed, where the basis function "*f*(Log_AADT)" is plotted against all the values of "Log_AADT". From this figure, it can be noticed that there are two knots, "10.389 and 11.112", when there is a sudden break in the straight line. This demonstrates the nonlinear performance of the variable "Log_AADT" with angle crash frequency.

7.2. Comparing MARS and NB models

For the first application of MARS in this study, a comparison between the two fitted MARS models and the corresponding NB models, while treating the response in each as a discrete one (i.e., crash frequency), is shown in Table 5. The R package was utilized to estimate the MARS models via the library "polspline". It is worth mentioning that all the three steps to fit a MARS model (as previously mentioned in Section 3) were automatically done in R, and R finally gives an output for the final "best" fitted MARS model. The MARS models were generated using the default *GCV* value "3" in R. The *GCV* criterion is mainly used in the pruning step; hence, over-fitting is less likely to occur. From this table, it is noticed that the MSPE values for MARS in the 3- and 4-legged models are lower than the corresponding NB models. As for the MAD values, they are lower for the NB models. However, there is still a good potential in applying the MARS technique. The generalized *R*-square is much higher for the MARS models.

7.3. Examining fitting MARS model with continuous response

To examine the higher prediction capability of MARS while dealing with continuous responses (Friedman, 1991), the two MARS models using the same important NB covariates were fitted while considering the natural logarithm of crash frequency. A default GCV value of "3" was used. The assessment criteria for the generated MARS models are shown in Table 6 (middle portion).

By comparing the MAD and MSPE values from this table with those from the previously fitted MARS models in Table 5, it is noticed that the MAD and MSPE values shown in the middle portion of Table 6 are much lower; hence, higher prediction capability. Also, the generalized *R*-square values in Table 6 are higher than those in Table 5.

7.4. Using MARS in conjunction with random forest

Since the MARS technique showed promising prediction performance, especially while dealing with continuous responses, an additional effort to test screening all possible covariates before

Table 5
Comparison between the fitted MARS and the NB models in terms of prediction and fitting.

	Angle three-legged model		Angle four-legged model	
	MARS	NB	MARS	NB
Prediction				
MAD[a]	1.27	1.07	1.08	0.85
MSPE[a]	3.08	3.96	2.95	3.30
Fitting				
Generalized *R*-square	0.39	0.19	0.52	0.31

[a] MAD and MSPE values are normalized by the average of the response variable.

Table 6
Prediction and fitting performance of different MARS models using a continuous response formulation.

	MARS models using significant covariates from the NB model shown in Table 3		MARS models using screened covariates from random forest	
	Angle three-legged model MARS[a]	Angle four-legged model MARS[a]	Angle three-legged model MARS[a]	Angle four-legged model MARS[a]
Prediction				
MAD[b]	1.01	0.69	0.99	0.69
MSPE[b]	0.74	0.61	0.74	0.58
Fitting				
Generalized *R*-square	0.47	0.67	0.47	0.65

[a] Response is the natural logarithm of crash frequency.
[b] MAD and MSPE values are normalized by the average of the response variable.

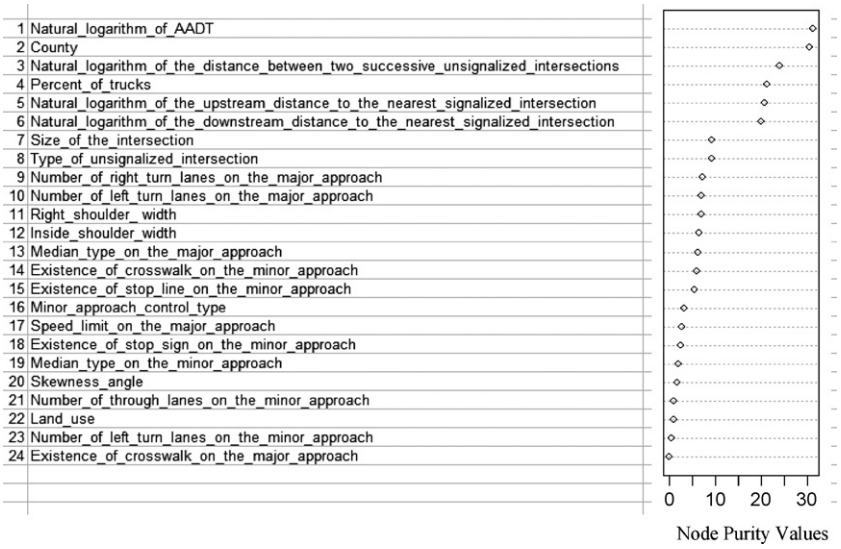

Fig. 2. Variable importance ranking using node purity measure.

fitting a MARS model was explored. This leads to utilizing the random forest technique (Breiman, 2001) before fitting a MARS model for variable screening and ranking important covariates. Using the R package, all possible covariates in the two attempted models were screened via the library "random Forest". The random forest technique was performed with 50 trees grown in the three and four-legged training datasets, and it was noticed that the attempted number of trees "50" was sufficient enough to obtain stabilized results.

Fig. 2 shows the purity values for every covariate in the 4-legged training dataset. The highest variable importance ranking is the natural logarithm of AADT, followed by the county location, then the natural logarithm of the distance between two unsignalized intersections, etc., until ending up with the existence of crosswalk on the major approach. The resulted variable importance ranking demonstrates the significant effect of the spatial covariates on angle crashes, with the distance between successive unsignalized intersections being the most significant. To screen the covariates, a cut-off purity value of "10" was used. This leads to selecting seven covariates (labeled from "1" till "7" in Fig. 2). Four out of seven variables were previously found significant in the four-legged NB model (logarithm of AADT, county, logarithm of the distance between successive unsignalized intersections and intersection size). Those seven covariates were then fitted using MARS, with the response being the natural logarithm of crash frequency, as it revealed the most promising prediction capability.

To assess whether there is an improvement over the two generated MARS models using the important variables from the NB model, the same evaluation criteria were used, as shown in Table 6 (right most portion). Comparing the MAD and MSPE values in Table 6 (middle and right most portions), it is noticed that there is a reduction (even if it is small) in the MAD and MSPE values for the two MARS models using screened covariates from random forest. The resulted generalized R-square values are relatively high; hence,

Table 7
MARS model at 4-legged unsignalized intersections after screening the variables using random forest.

Basis function	Basis function description	Estimate[a]	*P*-Value
Intercept	Intercept	−2.9252 (0.8759)	0.0009
Log_AADT	Natural logarithm of AADT on the major road	0.2376 (0.0852)	0.0055
Hills_County	Hillsborough County	0.5529 (0.0922)	<0.0001
Miami_County	Miami-Dade County	0.5362 (0.1031)	<0.0001
$(\text{Log\_AADT} - 11.112)_+$	A truncated power basis function for "Log_AADT" at "11.112"	−8.3871 (1.9002)	<0.0001
$(\text{Log\_AADT} - 10.778)_+$	A truncated power basis function for "Log_AADT" at "10.778"	2.6198 (0.6390)	<0.0001
	Generalized *R*-square	0.65	

[a] Standard error in parentheses.

showing better model fit. This demonstrates that using MARS after screening the variables using random forest is quite promising.

The final fitted MARS model using the seven selected covariates at 4-legged unsignalized intersections is presented in Table 7, where the response is the logarithm of angle crash frequency. From this table, it is noticed that the positive coefficient for the logarithm of AADT concurs with that deduced from Table 3. Also, there is a nonlinear performance for the continuous variable "Log_AADT" with the logarithm of angle crashes, as shown in its truncated basis function at "10.778" and "11.112".

8. Conclusions

This study investigated multiple applications of the machine learning technique "MARS" for analyzing angle crashes, which is capable of yielding high prediction accuracy. Also, exploring the significant factors contributing to angle crash occurrence at unsignalized intersections was another objective.

The fitted NB angle crash models showed several important variables affecting safety at unsignalized intersections. These include traffic volume on the major road, the upstream distance to the nearest signalized intersection, the distance between successive unsignalized intersections, median type on the major approach, percentage of trucks on the major approach, size of the intersection and the geographic location within the state. These variables are considered different than what was significantly deduced from predicting rear-end crashes at unsignalized intersections (Haleem et al., 2010).

While comparing the fitted MARS and NB models using a discrete response, it was concluded that both MARS and NB models yielded efficient prediction performance. Treating crashes as continuous response while fitting MARS models using the significant variables from the NB model was investigated. This was done by considering the natural logarithm of crash frequency. It was concluded that the fitted MARS models yielded better prediction performance than MARS models with the discrete response.

Finally, fitting MARS models after screening variables using the random forest technique was attempted. It was concluded that applying MARS in conjunction with the random forest technique showed slightly better results than fitting MARS model using the important variables from the NB model.

The findings of this study show that the MARS technique is a promising approach for predicting crashes at unsignalized intersections if prediction is the sole objective (particularly angle crashes, as deduced from this study). Hence, for achieving the most promising prediction accuracy, important variables should be initially selected using random forest before fitting a MARS model. Still, NB regression models are recommended as a valuable tool for understanding those geometric and traffic factors affecting safety at unsignalized intersections, as they are easy to explain.

This study provides some useful safety applications. For example, the application of MARS in before-after studies could be applied by fitting MARS models before and after applying the required safety countermeasure at certain sites. Afterwards, the predicted crashes (before and after) at those sites would be compared. This procedure should account for the regression-to-the-mean effect, and not erroneously overestimate the effect of this safety countermeasure. To account for the regression-to-the-mean effect, a long-term crash history should be used (for example, five years before applying the safety countermeasure, and five years after its application). Moreover, the empirical Bayes method should be used by accounting for the sites of interest, as well as the reference sites to estimate the expected crashes. The expected crashes at the sites of interest in the after period had the countermeasure not been implemented are estimated from two sources; the crash history of the sites of interest, as well as the predicted crashes at the reference sites using the MARS technique as a safety performance function.

Acknowledgement

The authors wish to thank the Florida Department of Transportation for funding this research.

References

Abdel-Aty, M., Radwan, E., 2000. Modeling traffic accident occurrence and involvement. Accident Analysis and Prevention 32 (5), 633–642.
Abdelwahab, H., Abdel-Aty, M., 2002. Artificial Neural Networks and Logit Models for Traffic Safety Analysis of Toll Plazas. Transportation Research Record: Journal of the Transportation Research Board, No. 1784. TRB, National Research Council, Washington, DC, pp. 115–125.
Abraham, A., Steinberg, D., Philip, N., 2001. Rainfall forecasting using soft computing models and multivariate adaptive regression splines. IEEE SMC Transactions: Special Issue on Fusion of Soft Computing and Hard Computing in Industrial Applications.
Agent, K., 1988. Traffic Control and Accidents at Rural High-Speed Intersections. Transportation Research Record No. 1160. Transportation Research Board, Washington, DC, pp. 14–21.
Anastasopoulos, P., Mannering, F., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41, 153–159.
Attoh-Okine, N., Mensah, S., Nawaiseh, M., 2003. A new technique for using multivariate adaptive regression splines (MARS) in pavement roughness prediction. Proceedings of the Institution of Civil Engineers 156 (1), 51–55.
Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.
Caliendo, C., Guida, M., Parisi, A., 2007. A crash-prediction models for multilane roads. Accident Analysis and Prevention 39 (4), 657–670.
Crash Analysis Reporting System. http://tlhost01.dot.state.fl.us/bluezone/FDOT_Session/default.htm (accessed 20.01.08).
Datta, K., 1991. Head-On, Left-Turn Accidents at Intersections with Newly Installed Traffic Signals. Transportation Research Record No. 1318. TRB, The National Academies, Washington, DC, pp. 58–63.
Datta, K., Dutta, U., 1990. Traffic Signal Installation and Accident Experience. ITE Journal, Washington, DC, pp. 39–42.
Francis, L., 2003. Martial chronicles is MARS better than neural networks. Casualty Actuarial Society Forum (Winter), 27–54.
Friedman, J., 1991. Multivariate adaptive regression splines. Ann Stat 19, 1–141.
Google Earth. http://earth.google.com/ (accessed 20.01.08).
Haleem, K., Abdel-Aty, M., Santos, J., 2010. Multiple applications of the multivariate adaptive regression splines technique in predicting rear-end crashes at unsignalized intersections. Journal of the Transportation Research Board, in press.
Harb, R., Yan, X., Radwan, E., Su, X., 2009. Exploring precrash maneuvers using classification trees and random forests. Accident Analysis and Prevention 41 (1), 98–107.
Harwood, D., Council, F., Hauer, E., Hughes, W., Vogt, A., 2000. Prediction of the Expected Safety Performance of Rural Two-lane Highways. Federal Highway Administration, Final Report, FHWA-RD-99-207.
Hauer, E., Ng, J.C.N., Lovell, J., 1988. Estimation of Safety at Signalized Intersections. Transportation Research Record No. 1185, pp. 48–61.
Jonsson, T., Lyon, C., Ivan, J., Washington, S., Schalkwyk, I., Lord, D., 2009. Investigating differences in the performance of safety performance functions estimated for total crash count and for crash count by crash type. Paper Presented at the Transportation Research Board 88th annual meeting, Washington, DC.
Kim, J., 2000. MARS modeling for ordinal categorical response data: a case study. The Korean Communications in Statistics 7, 711–720.
Kuhn, S., Egert, B., Neumann, S., Steinbeck, C., 2008. Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. BMC Bioinformatics 9 (400).
Kulmala, R., 1995. Safety at Rural Three- and Four-Arm Junctions: Development and Applications of Accident Prediction Models. VTT Publications 233. Technical Research Centre of Finland, Espoo.
Layfield, R., 1996. Accidents at Urban Priority Crossroads and Staggered Junctions. TRL Report 185. Transport Research Laboratory, Crowthorne, UK, p. 120.
Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using Support Vector Machine models. Accident Analysis and Prevention 40 (4), 1611–1618.
Lord, D., 2000. The prediction of accidents on digital networks: characteristics and issues related to the application of accident prediction models. Ph.D. Dissertation. Department of Civil Engineering, University of Toronto, Toronto, Ontario.
Lord, D., Bonneson, J., 2006. Development of Accident Modification Factors for Rural Frontage Road Segments in Texas. Zachry Department of Civil Engineering, Texas A&M University, College Station, TX.
Lord, D., Mahlawat, M., 2009. Examining the application of aggregated and disaggregated Poisson-gamma models subjected to low sample mean bias. Paper Presented at the Transportation Research Board 88th Annual Meeting, No. 1290, Washington, DC.
Miaou, S., 1994. The relationship between truck accidents and geometric design of road section: Poisson versus Negative Binomial regression. Accident Analysis and Prevention 26 (4), 471–482.

Miaou, S., Lord, D., 2003. Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes versus Empirical Bayes. Transportation Research Record, No. 1840, pp. 31–40.
Mussone, L., Ferrari, A., Oneta, M., 1999. An analysis of urban collisions using an artificial intelligence model. Accident Analysis and Prevention 31 (6), 705–718.
Najm, W., Smith, J., Smith, D., 2001. Analysis of Crossing Path Crashes (Report No. DOT HS 809 423). National Highway Traffic Safety Administration, Washington, DC.
Oh, J., Lyon, C., Washington, S., Persaud, B., Bared, J., 2003. Validation of the FHWA Crash Models for Rural Intersections: Lessons Learned. Transportation Research Record No. 1840, pp. 41–49.
Park, E., Lord, D., 2007. Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. Transportation Research Record No. 2019, pp. 1–7.
Persaud, B., Dzbik, L., 1993. Accident Prediction Models for Freeways. Transportation Research Record, No. 1401, pp. 55–60.
Phillips, S., 2004. Empirical collision model for four-lane median divided and five-lane with TWLTL segments. MS Thesis. North Carolina State University, Raleigh, NC.
Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection-accident frequency. Journal of Transportation Engineering 122 (2), 105–113.
Put, R., Xu, Q., Massart, D., Heyden, Y., 2004. Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies. Journal of Chromatography A 1055, 11–19.
Riviere, C., Lauret, P., Ramsamy, J., Page, Y., 2006. A Bayesian neural network approach to estimating the energy equivalent speed. Accident Analysis and Prevention 38 (2), 248–259.
Roadway Characteristic Inventory. http://webapp01.dot.state.fl.us/Login/default.asp (accessed 10.02.08).
SAS Institute Inc., 2002. Version 9 of the SAS System for Windows. Cary, NC.
Sawalha, Z., Sayed, T., 2006. Traffic accidents modeling: some statistical issues. Canadian Journal of Civil Engineering 33 (9), 1115–1124.
Summersgill, I., Kennedy, J., 1996. Accidents at Three-Arm Priority functions on Urban Single Carriageway Roads. TRL Report 184. Transport Research Laboratory, Crowthorne, UK, p. 74.
The New Web-based RCI Application. http://webservices.camsys.com/trbcomm/docs/presfla2005.htm (accessed 10.07.08).
Video Log Viewer Application. http://webapp01.dot.state.fl.us/videolog/ (accessed 10.02.08).
Vogt, A., Bared, J., 1998. Accident Models for Two-Lane Rural Roads: Segments and Intersections. Publication FHWA-RD-98-133. FHWA, U.S. Department of Transportation.
Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural network models: an empirical analysis. Accident Analysis and Prevention 39 (5), 922–933.
Zuur, A., Ieno, E., Smith, G., 2007. Analyzing Ecological Data. Statistics for Biology and Health, New York.