Accident Analysis and Prevention 41 (2009) 98–107

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Exploring precrash maneuvers using classification trees and random forests

Rami Harb[a,*], Xuedong Yan[b], Essam Radwan[c], Xiaogang Su[d]

[a] Department of Civil and Environmental Engineering, University of Central Florida, Orlando, FL 32816-2450, United States
[b] Center for Transportation Research, University of Tennessee, 309 Conference Center Bldg, Knoxville, TN 37996-4133, United States
[c] Center for Advanced Transportation Systems Simulation, Department of Civil and Environmental Engineering, University of Central Florida, Orlando, FL 32816-2450, United States
[d] Department of Statistics and Actuarial Science, University of Central Florida, Orlando, FL 32816-2450, United States

ARTICLE INFO

Article history:
Received 24 February 2008
Received in revised form 26 August 2008
Accepted 29 September 2008

Keywords:
Accident avoidance
Rear-end accidents
Head-on accidents
Angle accident
Decision trees
Random forests

ABSTRACT

Taking evasive actions vis-à-vis critical traffic situations impending to motor vehicle crashes endows drivers an opportunity to avoid the crash occurrence or at least diminish its severity. This study explores the drivers, vehicles, and environments' characteristics associated with crash avoidance maneuvers (i.e., evasive actions or no evasive actions). Rear-end collisions, head-on collisions, and angle collisions are analyzed separately using decision trees and the significance of the variables on the binary response variable (evasive actions or no evasive actions) is determined. Moreover, the random forests method is employed to rank the importance of the drivers/vehicles/environments characteristics on crash avoidance maneuvers. According to the exploratory analyses' results, drivers' visibility obstruction, drivers' physical impairment, drivers' distraction are associated with crash avoidance maneuvers in all three types of accidents. Moreover, speed limit is associated with rear-end collisions' avoidance maneuvers and vehicle type is correlated with head-on collisions and angle collisions' avoidance maneuvers. It is recommended that future research investigates further the explored trends (e.g., physically impaired drivers, visibility obstruction) using driving simulators which may help in legislative initiatives and in-vehicle technology recommendations.

Published by Elsevier Ltd.

1. Introduction

Many safety experts do not refer to traffic collisions as accidents and consider that they could and should have been avoided (What Causes Accidents, 2007). In fact, under critical traffic conditions leading to motor vehicle accidents, drivers can either take or not take evasive actions to avoid the crashes. Taking evasive actions vis-à-vis critical traffic situations grants drivers an opportunity to avoid the crash occurrence or at least diminish its severity. According to Uc et al., avoiding a crash requires continuous monitoring of neighboring vehicles, and anticipating and adjusting to changes in their speeds and positions under the pressure of time, which rely on multiple cognitive abilities (Uc et al., 2006). According to General Estimate System (GES) data for years 2002, 2003, and 2004 (NHTSA, 2004), over 30% of the drivers involved in rear-end accidents, over 25% of the drivers involved in head-on accidents, over 20% of drivers involved in angle collisions, and over 5% of the drivers involved in sideswipe collisions had no evasive actions prior to the accident occurrence (see Fig. 1). The reader may refer to Section 2.1 for detailed description of the GES data set used in this study. The noteworthy number of drivers not performing corrective actions prior to crashes underscores the need to explore the factors associated with the crash avoidance maneuvers.

The National Highway Traffic Safety Administration (NHTSA) shows great interest in accident avoiding technologies such as lane departure warning and brake assist (Safety Regulators Shifting Focus to Accident Avoidance, 2006). As a result, many automobile manufacturers studied and included collision avoidance warning systems (CAWS) in new car models that are designed to notify drivers about potential hazards from roadway departure and other vehicles (Araki et al., 1996; Chen et al., 1997; Hirst and Graham, 1997; NHTSA, 2002; Tilin, 2002; Clement and Taylor, 2006; Maltz and Shinar, 2007; Sengupta et al., 2007). Moreover, in an effort to enhance traffic accident avoidance skills, several agencies offer driver training courses for accident avoidance. These classes are designed to improve driving skills and help drivers reduce human errors when faced with hazardous traffic situations. All soldiers, civilian employees, and contractor employees who drive army-owned or leased vehicles must complete the training course (Wheeled Accident Avoidance, 2006).

Several researchers studied the reliability of different collision avoidance systems. For instance, Bliss and Acton (2003) studied the unreliability of collision avoidance systems and their effect on

* Corresponding author.
E-mail addresses: Ramyz5@yahoo.com (R. Harb), yxd22222@yahoo.com (X. Yan), Aeradwan@mail.ucf.edu (E. Radwan), Xsu@mail.ucf.edu (X. Su).

0001-4575/$ – see front matter. Published by Elsevier Ltd.
doi:10.1016/j.aap.2008.09.009

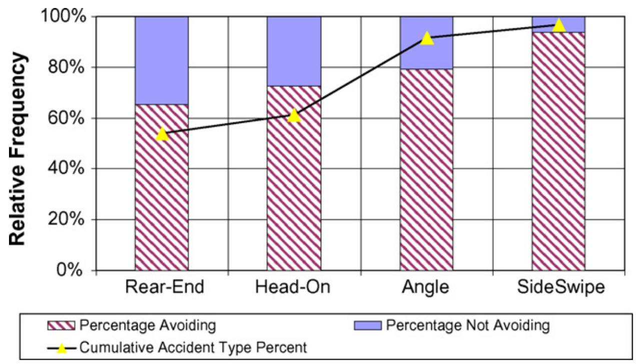

Fig. 1. Percentage of drivers trying to avoid accidents.

driving performance. It was concluded that reliable alarms generate more reaction from drivers (Bliss and Acton, 2003). Any mistrust of the alarm systems will result in reduced alarm response frequency (Bliss, 1993). Moreover, in the final report for the Automotive Collision Avoidance Systems (ACAS) Program, NHTSA states that drivers will likely ignore alarms that are not reliable (NHTSA, 2000). Therefore, it is crucial to explore the overall aspects and traits associated with the avoidance maneuvers to help develop efficient and reliable collision warning systems.

To study the accident avoidance maneuvers, many researchers employ driving simulators. Dangerous environments and hazardous traffic situations may be replicated in a safe driving simulator environment. For instance, Hancock and Ridder (2003) used a driving simulator to study the behavioral responses of drivers in the final seconds and milliseconds of induced hazardous situations that could lead to traffic accidents. Ikeda et al. (2001) studied the accident avoidance performance with respect to the age of the drivers using a driving simulator. Charlton (2007) used the University of Waikato-New Zealand driving simulator to examine the role of attentional, perceptual, and lane placement factors in drivers' behavior at horizontal curves. Ho et al. (2006) used the University of Oxford driving simulator to assess the effectiveness of vibrotactile warning signals in preventing front-to-rear-end collisions. Maltz and Shinar (2007) compared imperfect in-vehicle collision avoidance warning systems (IVCAWS) to a higher reliability collision avoidance warning system. Other studies using driving simulators entailed older drivers' avoidance capability in traffic accidents such as reduced vision field under complex work (Uno and Hiramatsu, 1995), effect of variable message signs on driver speed behavior on a section of expressway under adverse fog conditions (Kolisetty et al., 2006), deterioration in response for evasive maneuvers (Uno and Hiramatsu, 1998), deterioration of accident avoidance capability (Nishida, 1998), and rear-end accident avoidance resulting from horizontal visibility blockage (Harb et al., 2007a,b). Although the simulator studies provide a safe driving environment to test drivers' crash avoidance behaviors, these studies are generally limited by the sample size and the particular traffic scenario design. Therefore, an overall exploration analysis of the environments/drivers/vehicles factors associated with drivers' evasive maneuvers is needed, which could be a starting point for more in depth analyses using driving simulator or instrumented vehicles.

This paper studies the environments/drivers/vehicles factors associated with crash avoidance maneuvers based on the General Estimate System (GES) for years 2002, 2003, and 2004. The GES database records whether drivers took evasive actions or not prior to crashes in addition to the type of evasive actions (braking, steering, etc.). Rear-end collisions, head-on collisions, and angle collisions are analyzed individually to explore their environments, vehicles, and drivers' characteristics associated with the crash avoidance maneuver. It should be noted that sideswipe accidents were excluded from this analysis due to the small sample size. Previous studies successfully applied logistic regression to identify the statistical significance of independent variables on binary dependent variables (Hing et al., 2003; Stamatiadis and Deacon, 1995). Although logistic regression is a proper method to estimate the significance of independent variable on a dichotomous dependent variable, it makes it difficult to detect and interpret complex or high-order interactions among independent variables (Morgan and Sonquist, 1963; Su et al., 2008). Therefore, classification trees, nonparametric models, are utilized to analyze the accident avoidance maneuvers (evasive actions versus no evasive actions) for each accident type. It should be mentioned that tree methods handle interactions implicitly. In other words, one should not pick up a terminal tree node and trace up its ancestors in order to look for higher-order interactions. In addition, the random forests technique is employed to determine the independent variables' importance ranking for each accident type.

The remainder of this paper is organized as follows. Section 2 presents the crash database used in this study including the years of interest, the target variable, the data preparation, and the limitations of the data. In the same section, the classification trees and random forests techniques are briefly described. Section 3 illustrates the classification trees models and the variable importance rankings for rear-end collisions, head-on collisions, and angle collisions. Section 4 summarizes the findings of the analyses including discussions recommendations for future research.

2. Methodology

2.1. Crash database

The General Estimates System (GES) database for years 2002, 2003, and 2004 is used in this study. The GES database obtains its data from a nationally representative probability sample selected from the estimated 6.3 million police-reported crashes which occur annually. This database includes crashes that result in fatalities, injuries, and major property damage. The crash reports are chosen from areas that reflect the geography, roadway mileage, population, and traffic density of the United States. For more detailed information, refer to the GES Analytical User's Manual (NHTSA, 2004).

The GES database is a relational database consisting of three main files: accident, vehicle/driver, and person. Each file deals with a specific aspect of traffic crashes. These files may be linked as needed by the crash report case number and vehicle number. The accident file contains information on crash characteristics and environmental conditions at the time of the crash. The vehicle/driver file contains general information describing all vehicles and drivers involved in the crash. The person file contains general information describing all persons involved in the crash: drivers, passengers, pedestrians, pedal cyclists, and non-motorists.

Since 1992, five precrash variables have been added to the vehicle/driver file to identify: (1) what was the vehicle doing just prior to the critical pre-crash event (P_CRASH1), (2) what made the vehicle's situation critical (P_CRASH2), (3) what was the corrective action made, if any, to this critical situation (P_CRASH3), (4) what was the stability of the vehicle just prior to impact (P_CRASH4), (5) and what were the results of the vehicle's pre-crash stability coded in variable P_CRASH4 (P_CRASH5). To investigate the significant factors associated with drivers' crash avoidance actions; this study mainly focuses on the variables P_CRASH2 and P_CRASH3. P_CRASH2 was used to identify the vehicles/drivers that were under a critical traffic situation and might have chances to take corrective crash avoidance actions. Therefore, the data included in this study were restricted to the following critical traffic events that drivers encountered: other motor vehicle in lane (P_CRASH2 = 50–59), another vehicle encroaching into this vehicle's lane (P_CRASH2 = 60–78), pedestrian, pedacylist or other non-motorist (P_CRASH2 = 80–85), object or animal in roadway (P_CRASH2 = 87–92). It should be cautioned that the drivers who did not have the chance to perform crash avoidance actions such as vehicle loss of control (P_CRASH2 = 1–9), vehicle traveling over the lane line or the roadway edge (P_CRASH2 = 10–19) were excluded from this study. Further, P_CRASH3 was used to identify whether drivers took evasive actions or not. This study classified the drivers who encountered the critical traffic events (P_CRASH3) into two pre-crash groups: evasive actions group (P_CRASH3 = 1) given a value of 1 and no evasive actions group (P_CRASH3 = 2–98) given a value of 0. Therefore, comparing the distribution of the drivers/vehicles/environments between drivers taking evasive maneuvers and drivers not taking evasive maneuvers may identify the significant factors associated with the probability of drivers' taking evasive actions.

It should be mentioned that there is a large proportion of unknown "P_CRASH3" observations (P_CRASH3 = 99) in the GES data set used for this study. In 2002, 2003, and 2004 there were 68.11%, 70.21%, and 72.79% missing P_CRASH3 observations, respectively. A possible reason for the large percentage of unknown data is that generally, the onsite police officer may not specify driver's evasive actions without a site witness or a clear indication. The unknown cases were excluded from this research.

Similarly to all empirical crash databases, some limitation may exist in the General Estimate System (GES) database used in this study. These data come from police investigations using evidence from multiple sources of questionable consistency across a range of different police officers. This issue has been raised by many traffic safety experts; however, empirical traffic crash databases, including the GES database are widely used and often reliable for exploring crash characteristics. It is suggested that the same study be conducted using a different empirical database and that the results be compared.

2.2. Classification tree

A classification tree classifies observations by recursively partitioning the predictor space. Due to its nonparametric nature and easy interpretation, decision trees have received wide popularity from various fields, especially since the introduction of the Classification and Regression Trees (Breiman et al., 1984). The Classification and Regression Trees (CART) procedure consists of three steps: first, growing a large tree structure; then pruning it to obtain a sequence of nested subtrees; and finally selecting the best tree model from the subtree sequence via a validation method.

For this analysis, the above described GES data set consists of n independent and identically distributed observations (or drivers) $\{(y_i, \underset{\sim}{x}_i) : i = 1, \ldots, n\}$, where y_i denotes the ith binary response taking values of either 0 (drivers with no avoidance maneuver) or 1 (drivers with avoidance Maneuver) and $\underset{\sim}{x}_i = (x_{i1}, \ldots, x_{ip}) \in \Re^p$ is the p-dimensional input vector (drivers/vehicles/environments characteristics) for the ith driver. First, a single bisection of the whole data is performed, which is termed as the root node. A node virtually represents a data set in the hierarchical tree structure. If a node has descendents, it is termed as an internal node; otherwise it is termed as a terminal node or leaf. For a given node t, an impurity measure i_t is introduced to assess how drivers with no avoidance maneuvers and drivers with avoidance maneuvers are mixed together. A commonly used impurity measure is based on entropy as denoted in

$$i_t = -p_t \log(p_t) - (1 - p_t)\log(1 - p_t), \tag{1}$$

where p_t is the proportion of drivers with avoidance maneuvers in a specified node t. In Eq. (1), $i_t = 0$ when $p_t = 1$ or 0, which corresponds to the two ideally pure cases where the data set contains drivers with avoidance maneuver only or drivers with no avoidance maneuver only. The impurity measure, i_t, reaches its maximum when $p_t = 0.5$ that is when drivers with avoidance maneuvers and drivers with no avoidance maneuver are equally mixed together. Since our independent variables x_{ij} are categorical having each different levels in $C = \{c_1, \ldots, c_K\}$, an allowable split s of the data, induced by binary criteria: $x_{ij} \in B$ or $x_{ij} \notin B$, will be used for any subset $B \subset C$. A preferable bisection of the data will split the data into two purer halves. Suppose that node t is bisected into two child nodes: the left child node t_L; and the right child node t_R. It is intended that the total impurity in the two resultant child nodes, $i_{t_L} + i_{t_R}$, be greatly reduced when compared to the original node, i_t. Searching over all permissible splits, the best split s^* is the one that yields the greatest reduction of impurity, namely, $\Delta i(t) = i_t - (i_{t_L} + i_{t_R})$ is maximized by s^*. The root node is then split into two child nodes according to the best split. Subsequently, the child nodes will be further split into halves. Repeating this procedure will lead to a tree-structured model.

Second, to determine the optimal tree size, Breiman et al. (1984) proposed the regression and classification trees (CART) algorithm. In their proposal, a large tree T_0 is first grown by applying very mild stopping rules. To identify the best subtree, which must be a subtree of T_0, Breiman et al. (1984) proposed a cost-complexity pruning algorithm to iteratively truncate weak links or internal nodes. The results are a nested sequence of optimally pruned subtrees.

Third, the best subtree is selected from this sequence via a validation method. One way to do so is to apply an independent validation sample. In this method, the whole data set is first randomly divided into two parts: the training (70% of the observations in this study) set and the validation set (30% of the observations in this study). A large initial tree is grown and pruned using the training test. Then, the validation set is run down the subtree sequence and the misclassification rate is recalculated for each subtree. The best subtree is the one that provides the smallest misclassification rate.

The tree-based methods provide outstanding exploration or description of the data and are widely used in many application fields, such as data mining, medical diagnosis and/or prognosis. Model-based statistical significance testing has been fundamental in traditional statistics; However, it becomes problematic when dealing with large data sets (e.g., GES data used in this research). This is because even practically negligible effects would become statistically significant as long as the sample size is large enough. This issue has been addressed by many authors including Friedman (Friedman, 1997; Hand et al., 2001). To circumvent this difficulty, researchers from data mining, machine learning, and pattern recognition have approached the problem from a different perspective. The main idea of their approach is to shift the central point of modeling to prediction accuracy via validation. Various methods including recursive partitioning have been developed and thriving in this regards (e.g., the test sample method, cross-validation, and bootstrap).

One of the many attractive features offered by decision trees is variable importance ranking. It reveals the "important" factors in classifying drivers with evasive actions and drivers with no evasive actions. However, the final tree structure may not expose the variables' importance ranking since it could be completely masked by another correlated input. There are many methods available for extracting variable importance information. The next section elab-

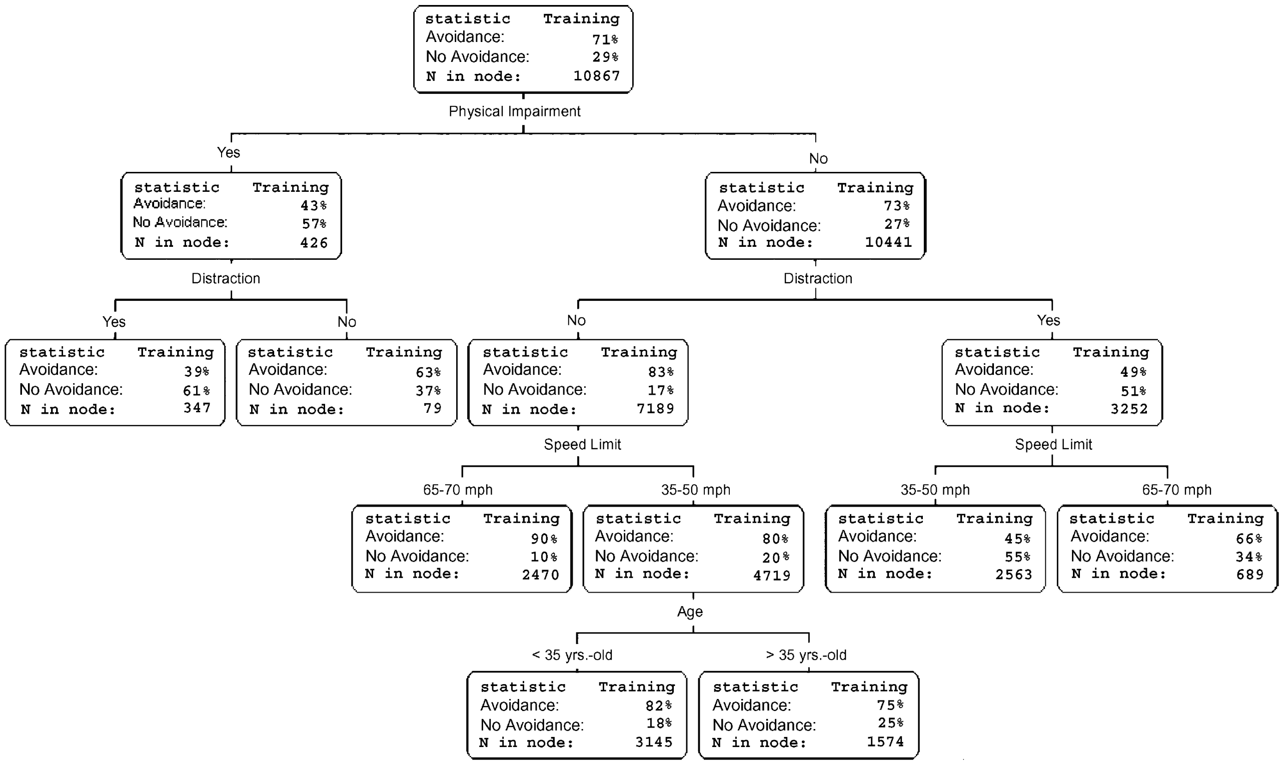

Fig. 2. Rear-end accidents' tree.

orates on the technique used in this study to determine the variable importance ranking.

2.3. Random forests

The procedure used by random forests (Breiman, 2001) is among the newest and most promising developments in extracting the variables' importance ranking. The term random forest comes from random decisions forests that was first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Brieman's bagging idea and Ho's "random subspace method" to construct a collection of decision trees with controlled variations (Ho, 1995).

In random forests, one grows a number of trees by bootstrapping the sample (i.e., randomly selecting n observations from the original data with replacement) and searching over only a randomly selected subset of inputs at each split. The results from these trees are then integrated in an appropriate way. To describe how to compute variable importance, let L_b denote the bth bootstrap sample and $L - L_b$ denote its out-of-bag sample containing all observations in the original data L that are not selected into L_b. After each tree T_b is constructed, send the out-of-bag sample $L - L_b$ down and record the misclassification rate r_b. By sending $L - L_b$ down to T_b one applies the tree structure T_b to the data $L - L_b$. For example, if we send a new data set $L - L_b$ down to the tree structure T_b shown in Fig. 2 of the manuscript, then that means we first break the data into two parts: drivers who have physical impairment go to the left child node and those who do not go to the right child node. Next, the two child nodes will be further split according to next criterion: whether there was a distraction. The splitting would end till it gets to a terminal node. Following this way, each observation will fall into one and only one of the terminal nodes in the tree structure. Among others, one common way of obtaining the predicted value (i.e., evasive actions or no evasive actions) in decision trees is via majority vote. Take the farthest left terminal node in Fig. 2 for example. Based on the current data, there are 347 subjects in this node, a majority (61%) of them being "no avoidance" and 39% being "avoidance". Thus the predicted value for all individuals in this node would be "no avoidance". And the misclassification rate for this node only would be 39%. The overall misclassification rate for the whole tree structure can be calculated in the same manner by integrating the prediction results from all its terminal nodes.

Then the values of the jth input in $L - L_b$ are randomly permuted and the out-of-bag sample is run down T_b again. Let r_b^j denote the resulting misclassification rate. This is done for every input ($j = 1, \ldots, p$) and the procedure is repeated for a total of B bootstrap samples. Finally, the variable importance is the average of the relative differences between r_b^j and r_b. The algorithm shown below summarizes the whole procedure for computation of variable importance via random forests.

Initialize all importance measures V_j's to zeros, for $j = 1, K, p$

Do $b = 1, K, B$
- Generate bootstrap sample L_b and obtain the out-of-bag sample $L - L_b$
- Based on L_b, grow a tree T_b by searching over m randomly selected inputs at each split
- Send $L - L_b$ down to T_b to compute the misclassification rate r_b
- For all predictors X_j, $j = 1, K, p$, do
  - ○ Permute the values of X_j in the out-of-bag sample $L - L_b$
  - ○ Send the permuted $L - L_b$ down to T_b to compute the misclassification rate r_b^j
  - ○ Update $V_j \leftarrow V_j + \frac{r_b^j - r_b}{r_b}$
- End do

End do

Average $V_j \leftarrow V_j/B$

Table 1
Input variables into tree models.

Variable	Categories	Tree 1 (rear-end) row % of each level		Tree 2 (head on) row % of each level		Tree 3 (angle) row % of each level	
		No avoidance	Avoidance	No avoidance	Avoidance	No avoidance	Avoidance
Age (years)	≤25	26.74	73.26	22.86	77.14	19.47	80.53
	26–35	26.64	73.36	20.70	79.30	22.98	77.02
	36–45	28.95	71.05	23.02	76.98	24.84	75.16
	46–55	31.04	68.96	19.70	80.30	27.63	72.37
	56–65	30.54	69.46	23.44	76.56	27.65	72.35
	66–75	35.68	64.32	45.65	54.35	32.79	67.21
	≥75	36.99	63.01	42.86	57.14	33.87	66.13
Body type	Passenger car	29.30	70.20	26.01	73.99	25.63	74.37
	LTV	28.32	71.18	21.49	78.51	23.34	76.66
	Trucks	23.02	76.98	8.66	91.34	18.32	81.68
Gender	Male	27.04	72.96	21.66	78.34	21.67	78.33
	Female	31.14	68.86	25.42	74.53	28.08	71.92
Alcohol involvement	Yes	35.96	64.04	22.42	77.58	23.96	76.04
	No	28.18	71.82	36.54	63.46	17.92	82.08
Physical impairment	Yes	56.76	43.24	22.20	77.80	23.67	76.33
	No	26.38	73.14	46.43	53.57	18.57	81.43
Speed (m/h)	10–20	36.74	63.26	31.58	68.42	35.95	64.05
	21–30	19.45	80.55	23.60	76.40	28.63	71.37
	31–40	14.69	85.31	15.49	84.51	19.41	80.59
	41–50	12.39	87.61	16.33	83.67	16.45	83.55
	51–60	9.67	90.33	13.25	86.75	8.43	91.57
	61–70	14.14	85.86	12.90	87.10	9.87	90.13
	>70	24.21	75.79	0.00	100.00	0.00	100.00
Visibility obstruction	No	25.59	74.41	24.72	75.23	23.89	75.11
	Yes	47.19	52.81	38.64	61.36	44.76	55.24
Distraction	No	15.32	84.68	26.19	73.31	24.87	75.13
	Yes	52.23	43.77	42.67	57.33	53.13	46.82
Number of lanes	One	39.73	60.27	7.14	92.86	29.41	70.59
	Two	28.09	71.91	20.93	79.07	22.29	77.71
	Three	24.69	75.31	28.47	71.53	24.35	75.85
Alignment	Curve	30.10	69.90	14.95	85.05	13.22	86.78
	Straight	31.90	68.10	27.48	72.52	29.13	70.87
Profile	Grade	22.74	77.26	12.41	87.59	1365	86.35
	Hillcrest	21.57	78.43	34.88	65.12	17.97	82.03
	Sag	30.00	70.00	66.67	33.33	20.00	80.00
	Straight	27.28	72.72	17.81	82.19	22.27	77.73
Lighting condition	Daylight	28.30	71.70	21.73	78.27	23.66	76.34
	Dark	29.07	70.93	25.57	74.43	24.80	75.20
Weather condition	No adverse condition	30.10	69.90	23.55	76.45	23.55	76.45
	Adverse condition	19.72	80.28	19.82	80.13	26.15	73.85

The Enterprise Miner developed by SAS (SAS Institute), which is another implementation of CART, is used in this study to develop the classification trees. Since Enterprise Miner does not offer the random forest method for variable importance ranking, the R software (The R-Project) is used to implement this procedure.

3. Tree models and analyses

Based on the univariate analysis (contingency tables) for each accident type, 13 factors show significant association at 0.05 significance level, with the likelihood of drivers taking evasive actions to avoid crashes (P_CRASH3). Table 1 shows the variables whose chi-square statistic's *P*-values are smaller than 0.05. Some of the variables found insignificant include: drug involvement, trafficway (rural versus urban location), highway type (divided versus undivided), road surface condition, speed limit, vehicle condition, road system identifier (state, local, expressway, etc.). Table 1 illustrates the distribution of these variables for each accident type and demonstrates that drivers' physical impairment ("blackout", "drowsiness", "sleepiness", "fatigued", "sleepiness", "requires cane or crutches", "paraplegic or restricted to wheelchair", "impaired due to previous injury", "deaf"), drivers' distraction ("by other occupants", "by moving object in vehicle", "while using phone", "using internal devices in vehicle", "distracted by outside person or object", "eating or drinking", "smoking"), drivers' visibility obstruction, speed, and vehicle type ("passenger car and light truck vehicle (LTV)" or "trucks and large trucks") may affect the accident avoidance maneuver significantly.

Three binary classification trees were developed in this study. The target variable in all three trees is P_CRASH3 which indicates whether the *i*th driver took evasive actions or not prior to the crash occurrence. Each node in the trees contains the percentage of the drivers with evasive maneuvers, the percentage of the drivers with no evasive maneuvers, and the number of observations. The input variables in each tree are listed in Table 1. The comparison between the drivers who took evasive actions and the drivers who did not take evasive actions may unveil the characteristics of the latter to help develop appropriate countermeasures that would increase the likelihood of evasive maneuvers.

3.1. Tree model #1: rear-end collisions

In tree model #1 (rear-end collisions), the input data, summarized in Table 1, comprises 10,867 observations (or drivers) from which 3151 drivers did not perform evasive maneuvers and 7716 drivers implemented evasive maneuvers. The resulting misclassification rate of this decision tree was 0.256. For the best tree size, shown in Fig. 2, the number of terminal nodes is determined to be 7.

As shown by the tree diagram of Fig. 2, drivers' physical impairment was used to create the first split and generate two internal child nodes. This indicates that drivers' physical impairment may have the most significant effect on evasive maneuvers in rear-end collisions. According to the first split, physically impaired drivers may be less likely to perform evasive actions compared to physically unimpaired drivers (43% with evasive maneuvers versus 73% with evasive maneuvers) when they are faced with potential rear-end collisions. The second level of this tree model demonstrates that distracted drivers may be more prone to fail to take evasive actions. In fact, distracted drivers with physical impairment may be less likely to have evasive maneuvers compared to undistracted drivers with physical impairment (39% versus 63%). Moreover, distracted drivers with no physical impairment may be less likely to have evasive maneuver compared to undistracted drivers with no physical impairment (43% versus 83%). The third level of the rear-end collisions classification tree shows that drivers may be more likely to implement evasive maneuvers at higher speed limits of 65 and 75 m/h compared to lower speed limits ranging between 35 and 50 m/h. In fact, distracted drivers with no physical impairment are more likely to perform evasive maneuvers under speed limits of 65 and 70 m/h compared to speed limits ranging between 35 and 50 m/h (66% versus 45%). On the other hand, undistracted drivers with no physical impairment may also be more likely to perform evasive maneuvers under speed limit of 65 and 70 m/h compared to speed limits ranging between 35 and 50 m/h (90% versus 80%). The fourth level of this tree reveals that undistracted and unimpaired drivers younger than 35 years old could be more likely to take evasive actions compared to drivers older than 35 years old (82% versus 72%) at speed limits ranging between 35 and 50 m/h. It should be noted that the age demarcation at 35 years old was implemented by the algorithm of CART since resulted in the lowest node misclassification rate. Interested researchers may investigate the sensitivity of the model at different age demarcations by manually aggregating over and under that age demarcation and looking at the percentage avoidance and the percentage no avoidance. However, the misclassification rate of the node will increase and the resulting misclassification rate of the entire model increases.

3.2. Tree model #2: head-on collisions

In tree model #2 (head-on collisions), there are 1105 observations (drivers) from which 254 drivers with evasive actions and 851 drivers with no evasive actions. The resulting misclassification rate was 0.221. For best tree size, shown in Fig. 3, the number of terminal nodes is determined to be 5.

As shown by Fig. 3, drivers' visibility obstruction was used to create the first split of the data. This indicates that drivers' visibility obstruction can have the most significant effect on evasive

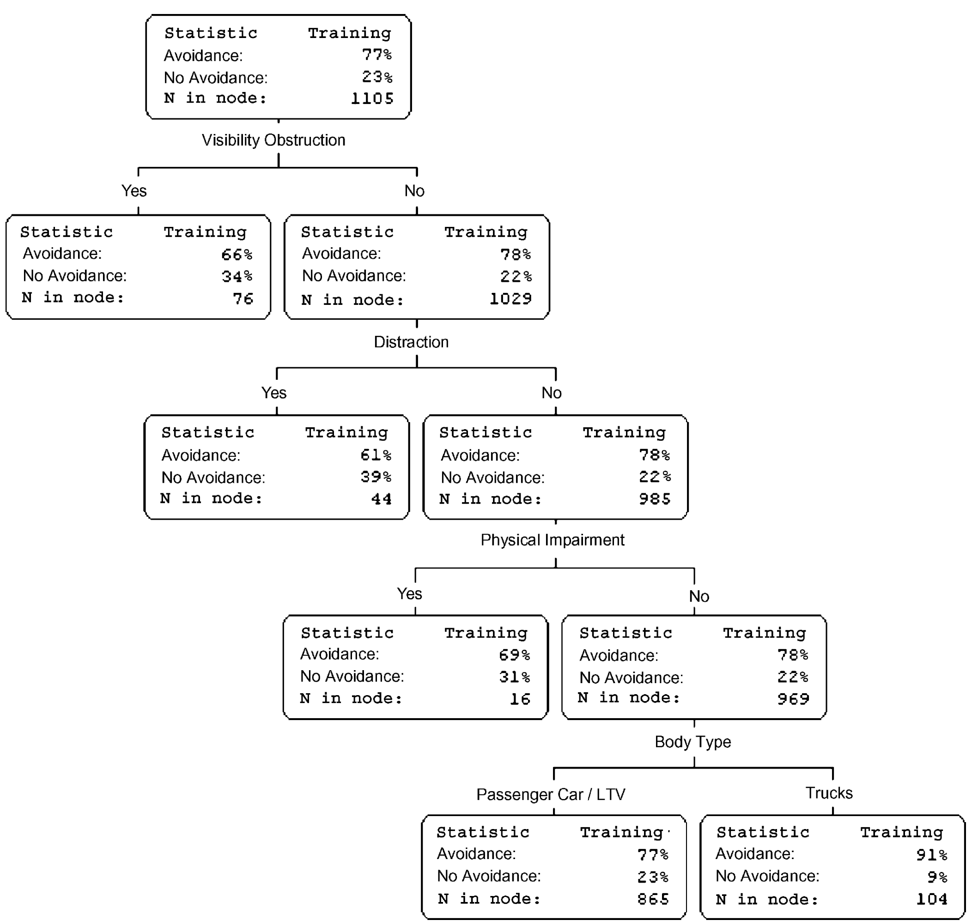

Fig. 3. Head-on accidents' tree.

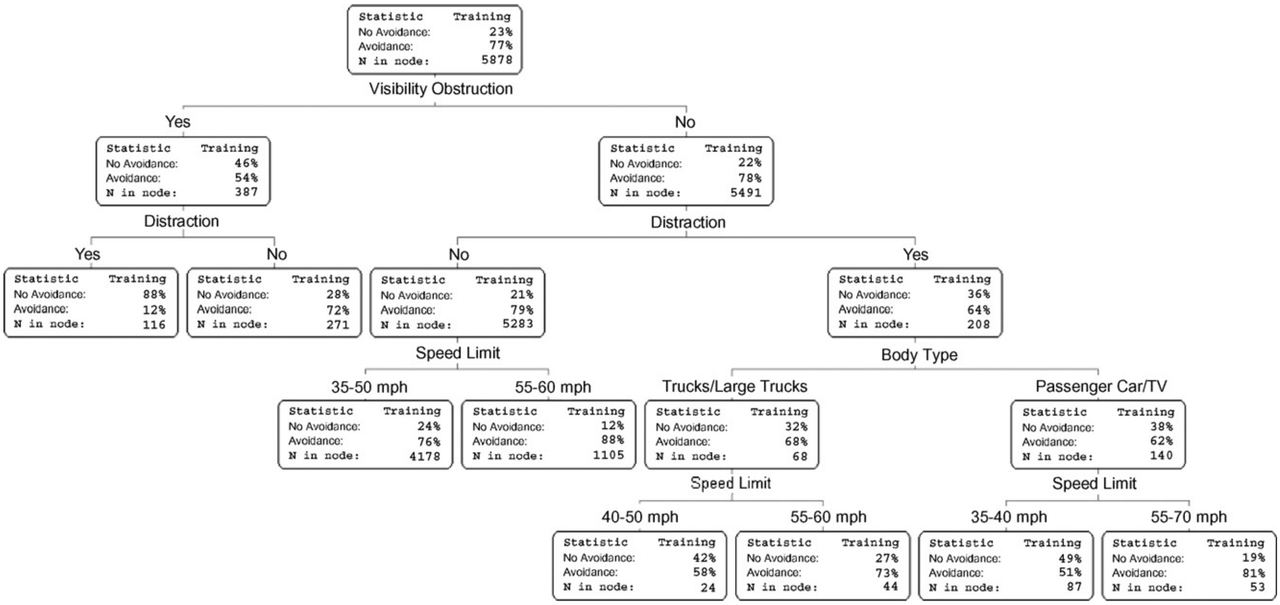

Fig. 4. Angle accidents' tree.

maneuvers in head-on collisions. The first level of this decision tree denotes that drivers experiencing visibility obstruction may be less likely to perform evasive maneuvers compared to drivers with clear visibility (66% versus 78%). The second level of the head-on collisions tree demonstrates that distracted drivers may be less likely to take evasive actions compared to undistracted drivers (61% versus 78%). The third level shows that physically impaired drivers are less likely to perform evasive maneuvers compared to physically unimpaired drivers (69% with versus 78%). The fourth level of this decision tree illustrates that truck drivers may be more likely to take evasive actions compared to non-truck (passenger cars/LTV) drivers (91% versus 77%).

3.3. Tree model #3: angle collisions

In tree model #3 (angle accidents), there are 5878 observations (drivers) from which 1352 drivers did not take evasive actions and 4526 drivers took evasive actions. The resulting misclassification rate was 0.226. For the best tree size, shown in Fig. 6, the number of terminal nodes is determined to be 8.

As shown in Fig. 4, drivers' visibility obstruction was used to create the first split of the data. This indicates that drivers' visibility obstruction has the most significant effect on taking evasive actions in angle collisions. Drivers suffering form visibility obstruction may be less likely to implement evasive actions compared to striking drivers with clear visibility (54% with evasive actions versus 78% with no evasive actions). The second level of the angle collisions tree shows that distracted drivers are less likely to take evasive maneuvers compared to the undistracted drivers (12% with evasive actions versus 72% with evasive actions for distracted drivers and 64% with evasive actions versus 79% with evasive actions for undistracted drivers). The third level of this tree shows that undistracted drivers with clear visibility may be more likely to implement evasive maneuvers at speed limits of 55 and 60 m/h compared to speed limits ranging between 35 and 50 m/h (88% with evasive actions versus 76% with evasive actions). Distracted truck and large truck drivers may be more likely to perform evasive maneuvers compared to distracted passenger car or light truck drivers (68% with avoidance maneuver versus 62% with avoidance maneuver). The fourth level of this tree shows that both distracted and undistracted drivers, driving any type of vehicles (trucks, large trucks, passenger cars, or LTVs) may be more prone to implement evasive actions under higher speed limits (for trucks and large trucks, 73% with evasive actions under speed limits of 55 and 60 m/h versus 58% with evasive actions under speed limits varying between 55 and 70 m/h. For passenger cars and LTVs, 81% with evasive actions under speed limits of 35 and 40 m/h versus 51% with evasive actions under speed limits ranging between 55 and 70 m/h).

3.4. Variables importance ranking

As mentioned in Section 2, variables' importance ranks the imperative factors in classifying drivers with evasive actions and drivers with no evasive actions. One tree structure (or the final tree structure) may not unveil the variables' importance ranking since it could be completely masked by another correlated input. Therefore, the random forests technique is employed where $B = 100$ trees were grown for each accident type. The variables' importances (V_j) were recorded for each variable for each tree T_b. Then, the average importance of each variable (for each accident type) is determined (V_j/B) and plotted in Fig. 5.

Fig. 5 summarizes the variables' importance ranking on accident evasive maneuvers for all three accident types. Drivers' evasive maneuvers are affected similarly in angle collisions and head-on collisions. In fact, drivers' visibility obstruction, drivers' distraction, drivers' physical impairment, vehicle type, and drivers' alcohol involvement have similar importance on the evasive maneuvers for head-on collisions and angle collisions. As for rear-end collisions, speed limit, number of lanes, and drivers' age have greater effect on accident evasive maneuvers compared to head-on collisions and angle collisions. This suggests that rear-end collisions' evasive maneuvers should be analyzed separately from angle collisions' evasive maneuvers and head-on collisions' evasive maneuvers. Drivers' distraction is ranked the second highest in significance on the evasive maneuvers in all three accident types. This fact warrants further attention to the sources of distraction and their possible countermeasures. The most important factor that affects head-on collisions and angle collisions' evasive maneuvers is drivers' visibility obstruction. Drivers' restricted visibility may deprive them from detecting a traffic hazards which may increase the probabil-

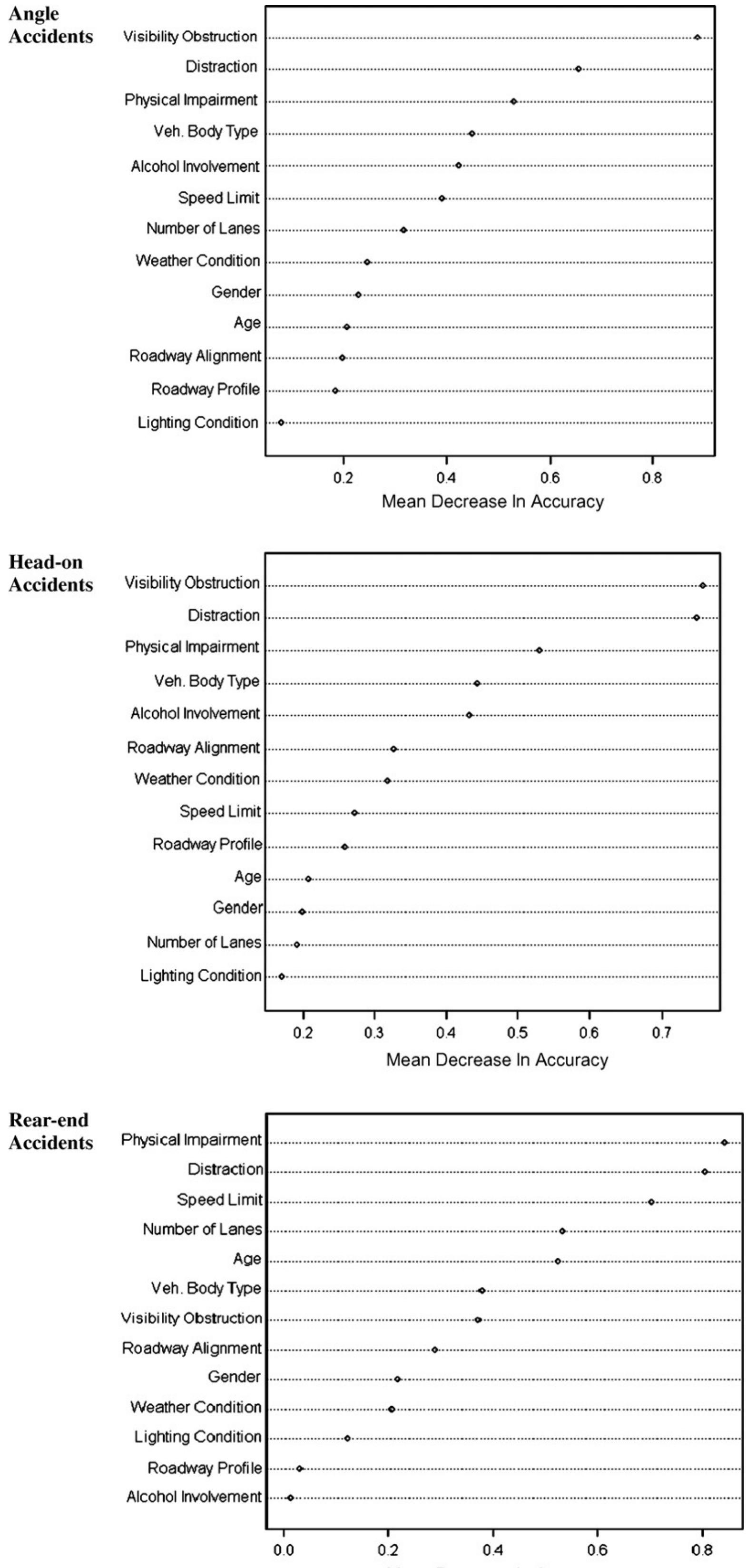

Fig. 5. Variables' importance ranking.

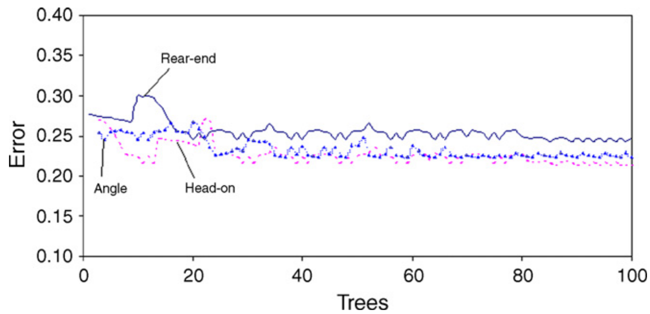

Fig. 6. Out-of-bag error rate with different number of trees.

ity of no evasive actions. Moreover, drivers' physical impairment ranks third in importance on the evasive maneuvers for head-on collisions and angle collisions and first for rear-end collisions. This finding emphasizes the necessity for clinical studies on different types of physical impairments and their effect on accident evasive actions.

Since it is essential to determine whether the number of trees is sufficiently large to obtain stable results, a sensitivity analysis showing the performance of the random forests' algorithm with the gradual increase of the number of trees is conducted. The out-of-bag (OOB) error rate, as recommended by the R package used in this study, is the selected measure of performance used in our analyses. Fig. 6 displays the (OOB) error rate for rear-end collisions, head-on collisions, and angle collisions at each iteration (from $B = 1$ to 100 trees). As shown by Fig. 6, for all three accident types, at $B = 80$ trees and more, the OOB error rates relatively stabilize indicating that the selected number of trees ($B = 100$) was sufficient for the random forests' algorithm to yield and reach relatively stable results.

4. Discussions and conclusions

The main objective of this study is to unveil the attributes of drivers, vehicles, and environments associated with crash evasive actions. The analysis showed that close to 30% of the drivers do not take evasive action vis-à-vis critical traffic conditions impending to motor vehicle crashes. Therefore, if one can explore the factors associated with the "no evasive actions", one can help develop more reliable collision warning systems and mitigate the number and the severity of the crashes.

In this study the GES database for years 2002, 2003, and 2004 is employed. The target binary variable is P_CRASH3 which identifies whether a driver performed evasive actions prior to the accident occurrence or not. Decisions trees along with random forests, a novel technique in traffic safety studies, are utilized to analyze P_CRASH3 for rear-end collisions, head-on collisions, and angle collisions. The random forests technique ranked the importance of the environments/vehicles/drivers characteristics on accident evasive maneuvers while eliminating potential effect of correlated inputs.

Drivers' distraction displayed the second highest negative effect on taking evasive actions in the three accident types. In fact, distractions such as engaging in cell phone conversations or other in-vehicle tasks while driving may result in failure to detect or delayed recognition of critical traffic events or discrete stimuli and drivers would be more likely to miss external traffic events (Alm and Nilson, 1994; Hancock and Ridder, 2003; McCartt et al., 2006; McKnight and McKnight, 1993; Strayer and Johnston, 2001; Stutts et al., 2003). Previous studies attempted to quantify different levels of distraction and modeled the distraction levels of different in-vehicle devices (Horrey and Wickens, 2006; Reed et al., 2008; Salvucci, 2005). However, at this point they do not have the ability to predict whether any in-vehicle device or activity produce an acceptable level of distraction (Reed et al., 2008). This study did not discriminate among the distraction sources and their individual effects on evasive maneuvers due to the limited sample size. A further investigation is suggested to compare the effects of different distraction causes on drivers' evasive maneuvers which may help in legislative initiatives and design recommendations.

Drivers younger than 35 years old may be more likely to take evasive actions compared to drivers older than 35 years old in rear-end collisions. This may be due to the fact that drivers' reaction time increases with age (Welford and Birren, 1965; Lings, 1991). The age factor underscores the need for ITS technologies to assist older drivers recognize the traffic hazard.

Truck drivers and large truck drivers may be more likely to perform evasive actions compared the passenger car drivers and LTV drivers. This may be due to the fact that truck drivers benefit from professional training programs including avoidance learning which plays a crucial role in the development and maintenance of safe driving (Fuller, 1992; Harrison, 2005).

Drivers under higher speed limits may be more likely to take evasive actions compared to lower speed limits. A study by Limpert and Gamero (1974) indicated that as speed increases, the number of drivers that attempt to avoid crashes through steering wheel inputs increases. This finding is also consistent with traffic flow literature (May, 1990) that drivers tend to be more sensitive (or alert) at higher speed. This fact may be due to the high alertness of drivers on higher speed limit facilities.

Drivers' visibility obstruction has the highest effect on head-on collisions and angle collisions' evasive maneuvers. If the drivers' sight is restricted they may not perceive a hazardous situation, which would deteriorate the proper space cushion that provides drivers enough reaction time. The drivers' visibility may be blocked by stationary objects or vehicles (Yan and Radwan, 2007), by moving vehicles on the road (Harb et al., 2007a,b), or may be the result of poor geometric design or bad weather. Previous studies showed that visibility obstruction may increase drivers' reaction times leaving them insufficient time to act upon a critical traffic event (Harb et al., 2007a,b). The interaction between drivers' distraction and drivers' visibility obstruction is emphasized in angle collisions and head-on collisions' evasive maneuvers in this study. Distracted drivers suffering visibility obstruction are significantly less likely to take evasive actions.

Drivers' physical impairment has the highest negative effect on evasive maneuvers in rear-end collisions and the third highest negative effect on evasive maneuvers in angle collisions and head-on collisions. Previous studies correlated drivers' physical impairment with precarious driving which can explain the negative effect on evasive actions. In fact, drowsiness, fatigue, sleepiness, impairment of visual perception, impairment of higher cognitive functions and of volition are anathema to safe driving (Brown, 1998; Lamond and Dawson, 1999; Thomas et al., 1998). This study did not distinguish between the different types of physical impairments ("blackout", "drowsiness", sleepiness"," fatigued", "requires cane or crutches", "paraplegic or restricted to wheelchair", "impaired due to previous injury", "deaf") recorded by the GES database due to the limited sample size. This significant negative effect of drivers' physical impairment on evasive maneuvers emphasizes the need of those drivers for additional attention such as special training courses or/and customized ITS technologies for different types of impairments.

This traffic crash database analysis provides an insight on the overall trends and significance of the drivers/vehicles/environments characteristics on evasive maneuvers vis-à-vis critical traffic conditions. It is suggested that future research take into consideration the trends explored in this study (especially drivers' physical impairment, drivers' visibility obstruction, and drivers' distraction) for more in depth analyses using driving simulators or instrumented vehicles. Different scenarios corresponding to the results from this study can be designed in a safe environment and human factors' characteristics can be collected and analyzed. Potential countermeasures can be tested (e.g., in-vehicle technology) which may help in legislative initiatives, design recommendations, and in-vehicle technology.

The authors acknowledge some limitations of this study. First, as mentioned in Section 3, and as a typical crash database, the GES data come from police investigations using evidence from multiple sources of questionable consistency across a range of different police officers. The authors recommend that similar analyses be conducted using a different data source. Second, due to the limited sample size, the authors did not discriminate between the different types of evasive actions (e.g., braking, steering), the different types of physical impairment, the different types of visibility obstruction, and the different types of distraction. It is recommended that future research disaggregate these variables if the sample is large enough or conduct these studies using driving simulators. Finally, crash databases lack of human factors characteristics (e.g., reaction time, deceleration rates) which may be important factors in analyzing drivers' evasive maneuvers. Therefore, the authors recommend further investigation using driving simulators.

References

Alm, H., Nilson, L., 1994. Changes in driver behavior as a function of hands free mobile phones—a simulator study. Accid. Anal. Prev. 26, 441–451.

Araki, H., Yamada, K., Hiroshima, Y., Toshio, I., 1996. Development of a rear-end collisions avoidance system. In: Proceedings of the 1996 IEEE Intelligent Vehicle Symposium, Tokyo, Japan, pp. 652–657.

Bliss, J.P., 1993. The cry-wolf phenomenon and its effect on alarm response. Unpublished Doctoral Dissertation. University of Central Florida, Orlando.

Bliss, J.P., Acton, S.A., 2003. Alarm mistrust in automobiles: how collision alarm reliability affects driving. Appl. Ergon. 34, 499–509.

Brieman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth, Inc, Pacific Grove, CA.

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

Brown, I.D., 1998. Driver fatigue. Hum. Factors 36, 298–314.

Charlton, S.G., 2007. The role of attention in horizontal curves: a comparison of advance warning, delineation, and road marking treatments. Accid. Anal. Prev. 39 (5), 873–885.

Chen, C., Quinn, R.D., Ritzmann, R.E., 1997. A crash avoidance system based on the cockroach escape response circuit. In: Proceedings on the 1997 IEEE International Conference on Robotics and Automation, Albuquerque, USA, pp. 2007–2012.

Clement, S., Taylor, M., 2006. The simple platoon advancement model of ITS technologies applied to vehicle control at signalized intersections. J. Adv. Transport. 40 (1), 1–21.

Friedman, J., 1997. Data Mining and Statistics: What's the Connection? In: Keynote Address, 29th Symposium on the Interface: Computing Science and Statistics.

Fuller, R., 1992. Learned riskiness. Irish J. Psychol. 13, 250–257.

Hancock, P.A., Ridder, S.N., 2003. Behavioural accident avoidance science: understanding response in collision incipient conditions. J. Ergon. Hum. Factors 46 (12), 1111–1135.

Hand, D., Mannila, H., Smyth, P., 2001. Principles of Data Mining. MIT Press, pp. 93–144.

Harb, R., Radwan, E., Yan, X., Abdel-Aty, M., 2007a. Contribution of light truck vehicles to rear-end collisions. Accid. Anal. Prev. 35 (5), 1026–1036.

Harb, R., Radwan, E., Yan, X., 2007b. Larger size vehicles (LSV) contribution to red light running, based on a driving simulator experiment. Transport. Res. Part F 10, 229–241.

Harrison, W.A., 2005. A demonstration of avoidance learning in turning decisions at intersections. Transport. Res. Part F: Traffic Psychol. Behav. 8, 341–354.

Hing, J.Y.C., Stamatiadis, N., Aultman-Haull, L., 2003. Evaluating the impact of passengers on the safety of older drivers. J. Safety Res. 34, 343–351.

Hirst, S., Graham, R., 1997. The format and presentation of collisions warnings. In: Ian, N. (Ed.), Ergonomics and Safety of Intelligent Driver Interfaces. Lawrence Erlbaum Associates, Mahwah, NJ.

Ho, C., Reed, N., Spence, C., 2006. Assessing the effectiveness of "intuitive" vibrotactile warning signals in preventing front-to-rear-end collisions in a driving simulator. Accid. Anal. Prev. 38 (5), 988–996.

Ho, Tin Kam, 1995. Random decision forest. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada.

Horrey, W.J., Wickens, C.D., 2006. Examining the impact of cell phone conversations on driving using meta-analytic techniques. Hum. Factors 48 (1), 196–205.

Ikeda, A., Katou, T., Kinoshita, M., Aihara, M., 2001. Study of driver simulator considerations on difference in accident avoidance performance due to age. JSAE Rev. 23, 219–222.

Kolisetty, V.G.B., Iris, T., Asakura, Y., Kuroda, K., 2006. Effect of variable message signs on driver speed behavior on a section of expressway under adverse fog conditions—a driving simulator approach. J. Adv. Transport. 40 (1), 47–74.

Lamond, N., Dawson, D., 1999. Quantifying the performance impairment associated with fatigue. J. Sleep Res. 8, 255–262.

Limpert, R., Gamero, F.E., 1974. The accident avoidance potential of the motor vehicle: accident data, vehicle handling and safety standards. In: Proceedings of the 3rd International Congress on Automotive Safety, vol. 11, Washington, DC.

Lings, S., 1991. Assessing driving capability: a method for individual testing. Appl. Ergon. 22, 75–84.

Maltz, M., Shinar, D., 2007. Imperfect in-vehicle collision avoidance warning systems can aid distracted drivers. Transport. Res. Part F 10 (4), 345–357.

McCartt, A.T., Hellinga, L.A., Bratiman, K.A., 2006. Cell phones and driving: review of research. Traffic Injury Prev. 7, 89–106.

McKnight, A.J., McKnight, A.S., 1993. The effect of cellular phone use upon driver attention. Accid. Anal. Prev. 25, 259–265.

May, A.D., 1990. Traffic Flow Fundamentals. Prentice Hall, Englewood Cliffs, New Jersey.

Morgan, J.N., Sonquist, J.A., 1963. Problems in the analysis of survey data, and a proposal. J. Am. Stat. Assoc. 58, 415–434.

National Highway Traffic Safety Administration, 2000. Automotive Collision Avoidance Systems (ACAS) Program. Final Report No. DOT HS 809 080. US Department of Transportation, Washington, DC.

National Highway Traffic Safety Administration, 2002. Automotive collision avoidance system field operational test. Final Report No. DOT HS 809 462. US Department of Transportation, Washington, DC.

National Highway Traffic Safety Administration, 2004. GES Analytical User's Manual 1988–2004. US Department of Transportation.

Nishida, Y., 1998. Driving characteristics of the elderly. In: Proceedings of the JSAE (in Japanese with English summary), no. 9832521.

Reed, J.J., Trick, L.M., Matthews, M., 2008. Testing assumptions implicit in the use of the 15-second rule as an early predictor of whether an in-vehicle device produces unacceptable levels of distraction. Accid. Anal. Prevent. 40 (2), 628–634.

Safety Regulators Shifting Focus to Accident Avoidance, 2004. Access date September 2006. http://www.consumeraffairs.com/news04/2005/nhtsa_active_safety.html.

Salvucci, D.D., 2005. A multitasking general executive for compound continuous tasks. Cogn. Sci. 29, 457–492.

SAS Institute, 2004. Cary, NC.

Sengupta, R., Rezaei, S., Shladover, S.E., Cody, D., Dickey, S., Krishnan, H., 2007. Cooperative collision warning systems: concept definition and experimental implementation. J. Intell. Transport. Syst. 11 (3), 143–155.

Stamatiadis, N., Deacon, J.A., 1995. Trends in highway safety: effects of an aging population on accident propensity. Accid. Anal. Prevent. 27, 443–459.

Strayer, D.L., Johnston, W.A., 2001. Driven to distraction: dual-task studies of simulated driving and conversing on a cellular phone. Psychol. Sci. 12 (6), 462–466.

Stutts, J., Feaganes, J., Rodgman, E., Hamlett, C., Meadows, T., Reinfurt, D., Gish, K., Mercadante, M., Staplin, L., 2003. Distractions In Everyday Driving (Final Report). AAA Foundation for Traffic Safety in Washington. The University of North Carolina Highway Safety Research Center.

Su, X., Zhou, T., Yan, X., Fan, J., Yang, S., 2008. Interaction trees with censored survival data. Int. J. Biostat. 4 (1), Article 2. The R project for statistical computing. http://www.r-project.org.

Thomas, M., Throne, D., Sing, N., Redmond, T., Balkin, T., Wensensten, N., Russo, M., Welsh, A., Rowland, L., Johnson, D., Aladdin, R., Cephus, R., Hall, S., Belenky, G., 1998. The relationship between driving accidents and microsleep during cumulative sleep deprivation. J. Sleep Res. 7 (Suppl. 2), 275.

Tilin, A. 2002. You are about to crash. Wired Magazine, Issue 10.04. Http://www.wired.com/wired/archive/10.04/.

Uc, E.Y., Rizzo, M., Anderson, S.W., Shi, Q., Dawson, J.D., 2006. Unsafe rear-end collision avoidance in Alzheimer's disease. J. Neurol. Sci. 251, 35–43.

Uno, H., Hiramatsu, k., 1995. Aged driver's performances in combined operation task. Trans. JSAE (in Japanese with English summary) 26 (2), 62–66.

Uno, H., Hiramatsu, K., 1998. Older driver's avoidance capabilities in an emergent traffic situation. In: Proceedings of JSAE (In Japanese with English summary), no. 20005000. 2000.

Welford, A., Birren, J., 1965. Behavior, Aging, and the Nervous System: Biological Determinants of Speed of Behavior and Its Changes with Age. Thomas, Springfield, IL.

What Causes Car Accidents? Smart Motorist. Access date March 2007. http://www.smartmotorist.com/acc/acc.htm.

Wheeled Accident Avoidance. Access date September 2006. http://www.per.hqusareur.army.mil/services/safetydivision/new_page_19.htm.

Yan, X., Radwan, A.E., 2007. Effect of restricted sight distances on driver behaviors during unprotected left-turn phase at signalized intersections. Trans. Res. Part F 10, 330–334.