Submitted for ITE Western District Student Paper Award, San Diego, California June 2017

Exploratory Analysis of Roadway Departure Crashes Contributing Factors Based on Classification and Regression Trees

Mohammadreza Hashemi[1*], **Adrian Ricardo Archilla**[2]
1-Ph.D. Student & Graduate Research Assistant, University of Hawaii at Manoa
2-Associate Professor, University of Hawaii at Manoa

ABSTRACT
According to the Federal Highway Administration (FHWA), Roadway Departure (RwD) crashes account for approximately 56% of highway fatalities in the United States. Therefore, investigation of RwD crashes contributing factors is an important step towards fulfilling the vision of zero deaths and serious injuries. Classification and Regression Trees (CART) is a non-parametric methodology for exploratory analysis, since it makes no presumptions about the relationship between dependent and independent variables. Using 4-year crash data from Island of Oahu, Hawaii, this paper employs a CART model to identify RwD crashes contributing factors such as roadway geometry, roadway inventory, and environmental characteristics. The results reveal that horizontal alignment, lighting condition, number of lanes, and speed limit are the most important factors in RwD crashes. The model shows that drivers are at greater risk of being involved in a RwD crash when driving on curves and straight roads with equal or less than 2 lanes with no daylight, respectively. In addition, the results indicate that a speed limit of more than 35 mph on straight roads is associated with a higher likelihood of RwD crashes since the drivers may lose control of their vehicles more easily. The model is developed with 70% of data and tested with the other 30%. The training model assigned the highest probability of the type of crash (RwD vs. non-RwD) to the actual type of crash in 72% of the cases. In practice, decision makers may consider associated contributing factors to prioritize the location and type of countermeasures to reduce RwD effectively.

1-INTRODUCTION
Highway fatalities and injuries in the United States continue to be a major national safety issue. While injuries are a concern, safety performance is typically expressed in terms of fatalities and fatal crashes because more extensive data are available for these crashes. According to the FHWA, although there has been a downward trend in both the fatal crash rates and number of deaths on highways, the fact that there are still 34,674 average annual fatalities (2007-2014) indicates that much work still needs to be done *(1)*. RwD crashes account for 56% of these fatalities *(1)*. Likewise, the statistics in the State of Hawaii also indicate that the majority of fatalities are related to RwD crashes *(2)*. The population of the island of Oahu was 953,207 in 2010 (approximately 72% of the resident population of the state of Hawaii). Oahu has 1,547.56 centerline miles of roads. The average annual total number of crashes from 2008 to 2011was 4,921. Although the percentage of RwD crashes is only about 25% of all crashes, 40% of fatalities are categorized as RwD crashes *(2)*.Therefore, it is necessary to investigate the RwD crashes contributing factors in detail.

Data mining is the process of applying different methods such as classification with the purpose of uncovering hidden patterns in large data sets. Decision trees are one of the classification algorithms available. In fact, it is also a predictive model. Decision trees are of two types: classification trees and regression trees (depending on whether the target variable is categorical or numerical). CART is a term used to cover both methods *(3)*. It is also one of the common data mining techniques used in many disciplines such as medicine, meteorological science, and engineering. Unlike parametric regression, CART is a non-parametric methodology that makes no presumption between dependent and independent variables. The CART methodology also has been applied in transportation safety problems. This model had acceptable prediction power in comparison to commonly used regression models *(4)*.

This research develops a CART model to analyze RwD using Oahu crash data from 2008 to 2011. The model uncovers the most important explanatory variables (contributing factors) that can distinguish

*Corresponding author. Email: hashemi@hawaii.edu

RwD crashes form non-RwD crashes. Also, it provides a classification tree to show graphically the circumstance in which the RwD crashes are more frequent.

2-BACKGROUND

FHWA published a strategic plan for RwD crashes and defined a RwD crash as: a non-intersection crash in which a vehicle crosses an edge line, a centerline, or otherwise leaves the traveled way *(1)*. Therefore, RwD crashes include both run-off the road (ROR) and head-on crashes. ROR crashes involve vehicles that leave the travel lane and encroach onto the shoulder and beyond and hit one or more of any number of natural or artificial objects, such as bridge walls, poles, embankments, guardrails, parked vehicles, and trees *(5)*. Also, a head-on crash typically occurs when a vehicle crosses a centerline or a median and crashes into an approaching vehicle *(6)*.

Previous research studies focused mostly on specific types of RwD crashes (e.g., studying only fatal ROR crashes). For instance, Peng et al. investigated the relationship between single vehicle ROR crashes and the geometric characteristics of rural two lane roads by using negative binomial and multinomial logit models. They found that crash frequency and severity will increase when there is a decrease in lateral clearance or shoulder width *(7)*. Liu & Subramanian, using data pertaining only to fatal single-vehicle ROR crashes, identified roadway, driver, environment, and vehicle-related factors associated with fatal single-vehicle ROR crashes. Their results show that driver sleepiness, alcohol impairment, roadway alignment with curves, speeding, passenger car, rural roadway, high speed limit road, and adverse weather were significant factors related to the high risk of fatal single-vehicle run-off-road crashes *(8)*. Hashemi & Archilla investigated the most prevalent and the most probable circumstances of RwD crashes by using the Bayesian Statistical Approach. Roadway design, human factors, and environmental variables were included in their analysis as well *(2)*.

Parametric regression models have been frequently applied to safety data. Some of the common challenges in development regression models are linearity assumption in linear models and the difficulty to develop robust non-linear models, sensitivity to outliers, auto correlation and homoscedastic disturbance *(9)*. CART is one of the data mining techniques that has been commonly employed in different disciplines. There is no presumption between dependent variables and independent variables in CART. Many previous studies in safety analysis have used the CART methodology. Chang & Cheng compare a CART model and negative binomial regression model to establish the relationship between traffic accidents and highway geometric variables, traffic characteristics, and environmental factors. The prediction performance between the CART and the negative binomial regression models showed that CART is a good alternative method for analyzing freeway crash frequencies *(4)*. Pande & Abdel-Aty examined the freeway traffic parameters leading to lane-change related collisions by using classification trees for selecting the explanatory variables *(10)*. De O˜naa et al. found that decision trees are useful to extract decision rules form crash data. For instance, it was inferred that the crashes are more severe for women in low light conditions; therefore decision makers can prioritize actions based on a classification of crash severities *(11)*.

Generally, less information was found in the literature about RwD crashes as a single concept or developing a decision tree to investigate the influential factors affecting RwD crash occurrence.

3-DATA COLLECTION AND PROCESSING

For this study, the state of Hawaii motor vehicle accident reports are used to analyze the of RwD crashes in Oahu for four consecutive years (2008-2011). After detailed examination of a database created from police report forms, the actions identifying whether a crash is a RwD or non-RwD crash are identified.

This study considers the type of crash as a dependent variable. The type of crash is categorized as either a RwD crash or a non-RwD crash. All the variables were treated as categorical except for total number of lanes and speed limit which are ordinal. In cases that the independent variables were categorical, an attempt was made to group the data into fewer number of classes for each variable to facilitate the interpretation of the results. For instance, all the crashes occurring in zones with speed limits less than or equal to 25 mph (the maximum speed limit on residential roads in Hawaii) are placed in a same class. Table 1 provides detailed information regarding the variables.

Table 1 Description of Variables

Variable	Type	Description
Crash Type	Categorical	Target Variable: Non-RwD (0), RwD (1)
Area Type	Categorical	Rural (1), Urban (2)
Road Owner	Categorical	County (1), State (2)
Weather	Categorical	Clear (1), Cloudy (2), Rain (3), Other (4)
Lighting Conditions	Categorical	Daylight (1), Dawn/Dusk (2), Spot/Continuous Lighting (3), Dark (4)
Land Use	Categorical	School (1), Business (2), Residential (3), Industrial (4), Recreational (5), Farm/Fields (6), No Development (7), Other (8)
Road Classification	Categorical	Two-way undivided road (1), Two-way divided road (2), Other (3)
Horizontal Alignment	Categorical	Straight (1), Curve (2)
Vertical Alignment	Categorical	Level (1), Downhill or Uphill (2), Hill Crest or Sag (3)
Pavement Surface Type	Categorical	Concrete (1), Asphalt (2), Gravel or Dirt (3), Other (4)
Surface Wetness Condition	Categorical	Dry (1), Wet (2), Other (3)
Number of Lanes	Ordinal	Lanes ≤ 2 (1), 3≤Lanes ≤4 (2), 5 ≤Lanes (3)
Speed Limit	Ordinal	Speed≤25mph (1), 25mph<Speed≤35mph (2),35mph <Speed (3)

4-METHODOLOGY

CART constructs a model by repeatedly dividing a subset of data (called the parent node; the first parent node is called root node) into two descendants (child nodes) by a measure of impurity. The main purpose of partitioning the data is to minimize the impurity for all child nodes (a pure child node is called terminal node or leaf). Figure 2 presents schematic sketch of CART.

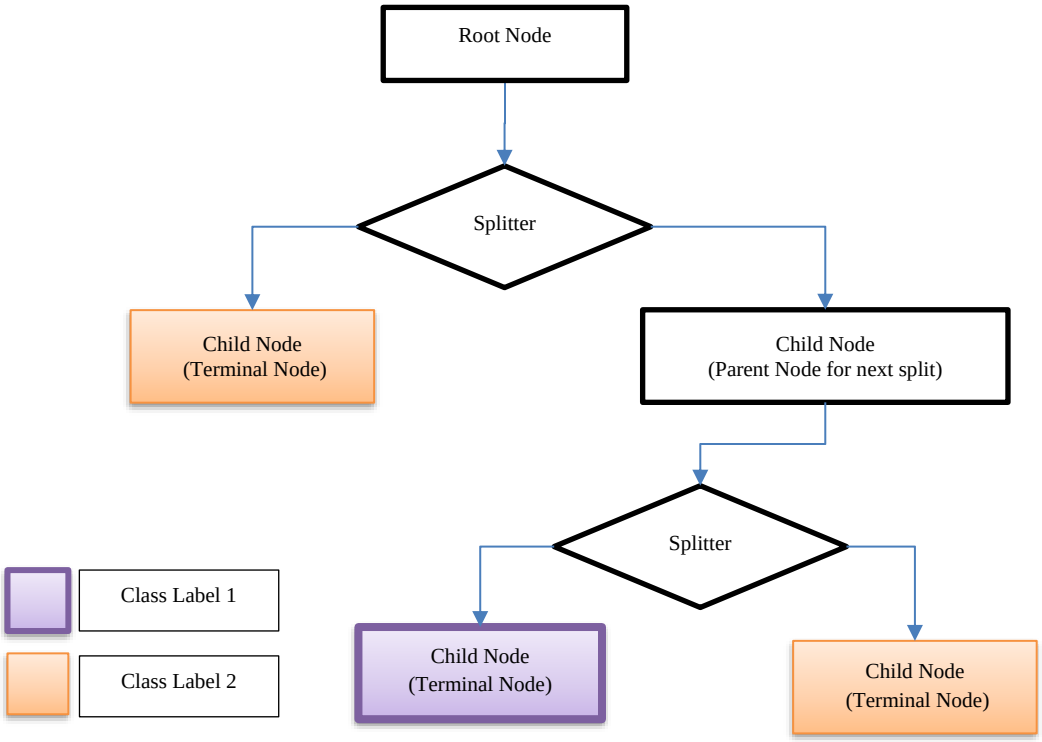

Figure 2 CART Methodology

Following the notation in *(3)*, a CART model can be summarized into three steps:

1) Establishing a set of binary questions to split each parent node.

2) Defining a splitting rule. For this, an impurity function $\phi(s,t)$ is used to examine the goodness of split s for selecting the best splitter in node t. The impurity measure $i(t)$ of node t for a problem with j classes is:

$$i(t) = \phi(\, p(1|t), p(2|t) \ldots, p(j|t)\,) \quad (1)$$

where $p(j|t)$ is the proportion of cases in class j to all cases at node t. The splitter s at node t divides the data into two child nodes with proportion of p_R and p_L; then the change in impurity measure by splitter s at node t is:

$$\Delta i(s,t) = i(t) - p_R i(t_R) - p_L i(t_L) \quad (2)$$

The model tries to maximize $\Delta i(s,t)$ in node t; therefore, this leads to the minimization of the overall impurity of the model in a recursive process.

The most common criterion to measure the impurity is Gini index *(3)*. Defining $\pi(j)$ as the prior probability for category j, N as the total number of observations, and N_j as the total number of observation in category j, the Gini index, a particular functional form for Equation 1, is defined as:

$$i_{Gini}(t) = \sum_{i \neq j} p(i|t)\, p(j|t) \quad (3)$$

$$P(j|t) = \frac{p(j,t)}{p(t)} \quad (4)$$

$$p(j,t) = \pi(j)\frac{N_j(t)}{N_j} \quad (5)$$

$$p(t) = \sum_j p(j,t) \quad (6)$$

$$\sum_j p(j|t) = 1 \quad (7)$$

$$\pi(j) = \frac{N_j}{N} \quad (8)$$

3) Defining some criteria to declare a node as a terminal node. The model continues partitioning until the impurity becomes zero in each terminal node. Since a classification tree with too many branches and terminal nodes is impractical, some criteria are needed to stop the splitting process. It can be done either by defining a threshold as a minimum amount of changes in the impurity measure at node t to perform splitting, or a maximum value for the depth of the tree. Finally, it is common practice to assign a category label to each terminal node according to the category with highest probability. However, this practice is contradictory with a probabilistic analysis.

The CART methodology can apply other criteria to improve the classification tree structure, such as surrogate, pruning, and misclassification cost. A more detailed explanation of the methodology is provided in Breiman et al *(3)*.

4-1-Variable Importance

In a CART model, it is possible that the effect of a variable V_1 is masked by another variable V_2. This means that if the model eliminates V_1, then V_2 can be added to the new model and make it almost as good as the first model. Therefore, the importance of a variable is not measured only by its use as a splitter in the optimal tree. In fact, the importance of each variable is the summation of changes in impurity measure by using that variable as a splitter $\tilde{s}_{V_m}$, for all nodes $t \in T$ (T is total nodes). Consequently, the measure of importance of a variable V_m in a CART model is *(3)*:

$$VIM(V_m) = \sum_{t \in T} \frac{N_t}{N}\, \Delta i(\tilde{s}_{Vm}, t) \quad (9)$$

4-2-Model Assessment

A CART model is usually evaluated by accuracy of prediction. Accuracy is a percentage of cases classified correctly and it has been used traditionally to evaluate quality of a classification model.

Table 2 Model Assessment Table

		Predicted	
		True	False
Observation	True	True Positives (TP)	False Negatives (FN)
	False	False Positives (FP)	True Negatives (TN)

$$Sensitivity = True\ Positive\ Rate: TPR = \frac{TP}{(TP+FN)} \quad (10)$$

$$1 - specificity = False\ Positive\ Rate: FPR = \frac{FP}{(FP+TN)} \quad (11)$$

Moreover, the accuracy of prediction in a model is:

$$Accuracy = \frac{(TP+TN)}{(P+N)} \quad (13)$$

5-RESULTS

5-1-Variable Importance

Table 3 presents the normalized variable importance in the model. The horizontal alignment is the most important explanatory variable in the model. This means that the proclivity of a vehicle to leave the road depends on the horizontal alignment. Lighting condition is second important variables. Lighting conditions play a role for drivers to see clearly and keep the vehicle on the road. The number of lanes is third most important variable. Speed limit is the next important variable. This is reasonable since it is more difficult to control the car or perform an evasive maneuver at higher speeds. Other variables have normalized importance less than 20%. In fact, the model shows that road classification, road owner, and weather are not significant enough to distinguish RwD crashes from non RwD crashes. For instance, from this model, it can be inferred that rainy weather is not a good variable to separate RwD from non-RwD crashes. Although it might be true that there are more RwD crashes in rainy days, it is also possible that total number of non RwD crashes increases in a similar proportion.

Table 3 Importance of the Variables

Independent Variable	Normalized Importance	Independent Variable	Normalized Importance
Horizontal Alignment	100.0%	Pavement Surface Type	15.7%
Lighting Conditions	86.1%	Surface Wetness Condition	10.9%
Number of Lanes	70.7%	Area Type	10.6%
Speed Limit	54.2%	Weather	5.0%
Vertical Alignment	20.9%	Road Owner	1.8%
Land Use	18.3%	Road Classification	1.0%

5-2-CART Model

Figure 3 depicts the CART model developed with the training data. The training model was developed with 70% of the data (selected randomly from the full data set), and the test model was built with the other 30%. Horizontal alignment is the first splitter to divide the crash data. It separates those crashes occurring on curves (node 2) and on straight roads (node 1). Node 2 shows that the probability of a crash being a RwD crash on curves are about 2 times the value for straight roads. Since about 25% of all RwD crashes occur on curves, the model has considered it the most important characteristic of RwD crashes. Notice that the model stopped splitting at node 2 because it could not find another variable on curves to partition further

the crash data into more homogeneous nodes. Therefore, RwD crashes on curves are the first terminal node that should be considered for safety evaluation.

For those crashes on straight roads, the model split node 1 by the total number of lanes into crashes on roads with 2 lanes or less (node 3) and crashes on roads with more than 2 lanes (node 4). The model shows that the probability of a crash being a RwD crash on roads with 2 lanes or less is more than two times the value for roads with more than two lanes. The model split node 3 into those crashes with daylight condition (node 5) and no daylight condition (node 6). The results show the probability of a crash being a RwD crash with no daylight condition is higher and considers node 6 a terminal node. Then model split the crashes in node 5 according to whether the speed is less than or equal to 35mph (node 9) or more than 35mph (node10). For crashes that occur on a straight road, with equal or less than 2 lanes, and daylight condition, the probability of being categorized as a RwD crash is about 2 times higher on roads with a speed limit greater than 35 mph than on roads with a speed limit of less than or equal to 35 mph. The interpretation of the results for the right side of node 1 is similar. There are couple of additional points worth mentioning. First, the speed limit is found to be better splitter than lighting condition for roads with more than 2 lanes. Second, the model has used vertical alignment as a splitter to minimize the impurity of node 12, which shows that crashes in steep grades have higher probability of being RwD crashes.

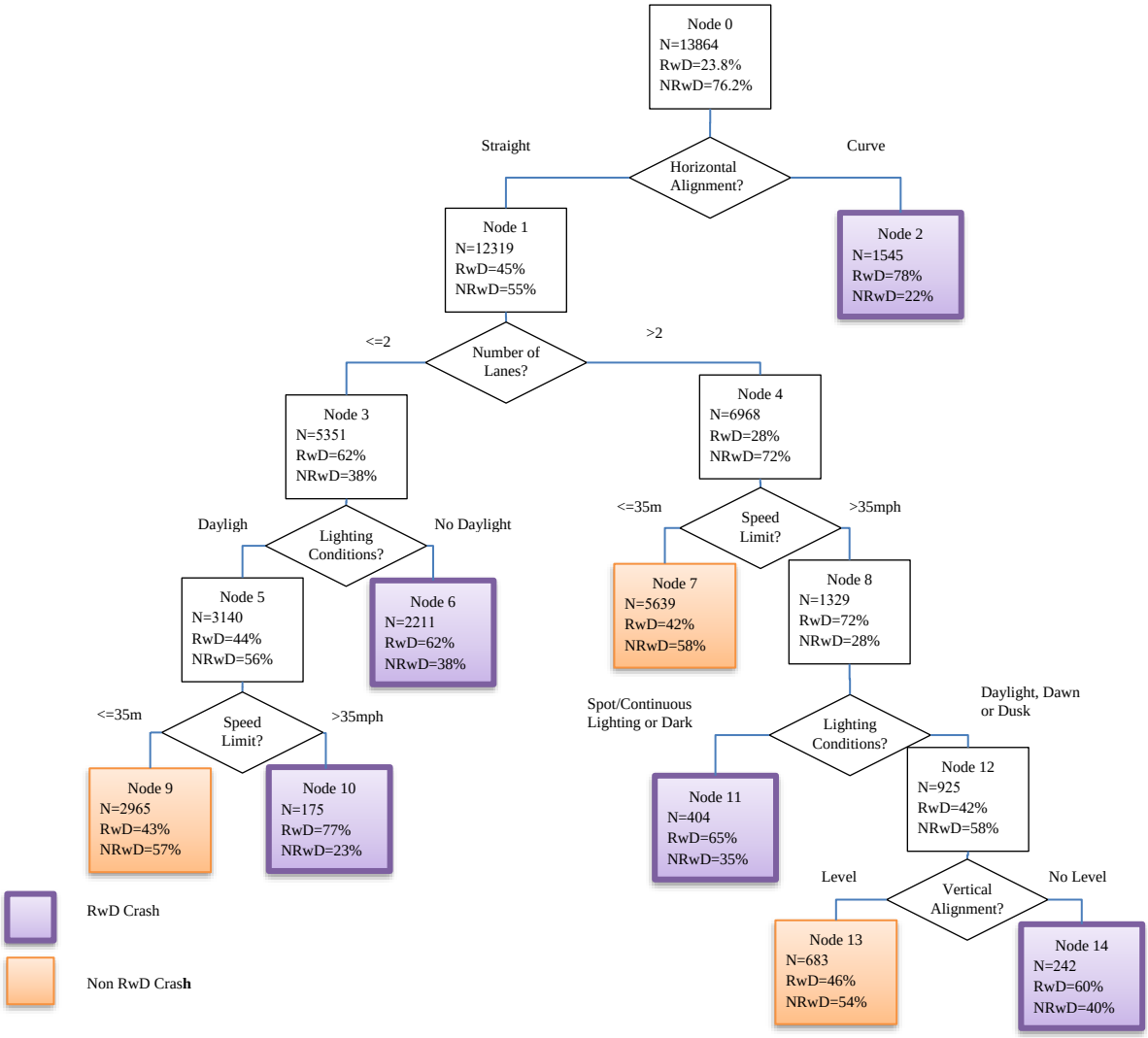

Figure 3 CART Model

5-3-Model Assessment

A common issue with classification trees is the unbalance of information in the dataset. If a majority of the dataset belongs to a specific category, the classification is bias towards that majority category; therefore, the goodness of fit for the classification of that class inflated. To solve this issue, Breiman et al*(3)* and Steinberg and Colla *(12)* suggests the consideration of equal prior probabilities in model. Although, the overall goodness of fit for the classification tree may decrease, the goodness of fit for classifying the minority category (in this case RwD crashes) increases significantly, which is a point of interest.

Table 4 presents the results of the prediction in both the training model and the test model. The results show that 75.8% of non RwD crashes and 61.4% of RwD crashes are correctly classified. Moreover, the accuracy for training, which is the total number of correctly predicted crashes to total number of crashes is 72.4%.

Table 4 Model Assessment for the CART Model

Sample	Observed	Predicted		
		Non RwD	**RwD**	**Percent Correct**
Training	Non RwD	8014	2555	75.8%
	RwD	1273	2022	61.4%
	Overall Percentage	67.0%	33.0%	72.4%
Test	Non RwD	3365	1084	75.6%
	RwD	562	810	59.0%
	Overall Percentage	67.5%	32.5%	71.7%

6-DISCUSSION & CONCLUSIONS

Classification is a data mining technique useful in both finding accurate classifier and predicting the structure of a phenomenon. The CART methodology has some advantages: it is non-parametric, it selects variables automatically, and it has the ability of using any combination of discrete variable. However, it also presents some challenges: the decision tree configuration is unstable since it might change by adding or removing a variable, it does not provide robust statistical significance indicators for each variable such as p-values in regression analysis, and the splitting process is done by only one variable at a time. Therefore, a modeler should consider these points before choosing CART. Nevertheless, despite these issues, it provides a good starting point for further analysis of large datasets.

RwD crashes account for half highway fatalities; therefore, this research developed a CART model to explore it. The results showed that the CART model had acceptable performance in identifying influential variables and distinguishing the circumstances in which RwD crashes are more probable to occur. The results indicate that a crash is more likely to be a RwD crash on curves and on straight segments with 2 lanes or less, daylight condition, and a speed limit greater than 35 mph. The results are consistent with field observations in Oahu and the spatial distribution of RwD crashes. In fact, the regions with higher percentage of RwD crashes have roads with the characteristics that increase probability of RwD crash occurrence.

The CART diagram decision tree can help decision makers to prioritize the safety projects. For instance, this study showed that from 5 different scenarios (RwD crash terminal nodes), two of them are more important. The decision tree reveals that approximately a quarter of crashes occur on curves and another quarter of crashes occur on straight roads with equal or less than 2 lanes with no-daylight condition (note that this situation, however, does not correspond to a node with the one of the highest probabilities of a crash being a RwD crash). Therefore, implementation of low cost countermeasures for curves such as chevron signs, curve signs, and improving the lighting conditions on the second group roads can mitigate RwD crashes more effectively.

ACKNOWLEDGEMENTS
The financial support of the State of Hawaii Department of Transportation, Safety Division, Oahu District is greatly appreciated and acknowledged.
The contents of this paper reflect the views of the authors, who are responsible for the facts and accuracy of the facts presented herein. The contents do not necessarily reflect the official views or policies of the State of Hawaii, Department of Transportation or the Federal Highway Administration. This report does not constitute a standard, specification or regulation.

REFERENCES
1. Federal Highway Administration. Roadway Departure (RwD) Strategic Plan. http://safety.fhwa.dot.gov/roadway_dept/.
2. Hashemi, M., and A. R. Archilla. Potential Factors Affecting Roadway Departure Crashes in Oahu, Hawaii. 2016.
3. Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.
4. Chang, L. Y., and W. C. Chen. Data mining of tree-based models to analyze freeway accident frequency. *Journal of Safety Research*, Vol. 36, No. 4, 2005, pp. 365–375.
5. Neuman, T. R., R. Pfefer, K. L. Slack, K. K. Hardy, F. Council, H. McGee, L. Prothe, and K. Eccles. *A guide for addressing run-off-road collisions*. 2003.
6. Neuman, T., R. Pfefer, K. L. Slack, H. McGee, L. Prothe, K. Eccles, and F. Council. *NCHRP REPORT 500: Guidance for Implementation of the AASHTO Strategic Highway Safety Plan.Volume 4: A Guide for Addressing Head-On Collisions*. 2003.
7. Peng, Y., S. R. Geedipally, and D. Lord. Investigating the Effect of Roadside features on Single-Vehicle Roadway Departure Crashes on Rural Two-Lane Roads. 2012.
8. Liu, C., and R. Subramanian. *Factors related to fatal single-vehicle run-off-road crashes*. 2009.
9. Washington, S. P., M. G. Karlaftis, and F. L. Mannering. *Statistical and econometric methods for transportation data analysis*. CRC press, 2010.
10. Pande, A., and M. Abdel-Aty. Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention*, Vol. 38, No. 5, 2006, pp. 936–948.
11. de Oña, J., G. López, and J. Abellán. Extracting decision rules from police accident reports through decision trees. *Accident Analysis & Prevention*, Vol. 50, 2013, pp. 1151–1160.
12. Steinberg, D., and P. Colla. CART: tree-structured non-parametric data analysis. *San Diego, CA: Salford Systems*, 1995.