International Journal of Transportation Science and Technology 10 (2021) 235–244

Contents lists available at ScienceDirect

International Journal of Transportation Science and Technology

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

Rule-based safety prediction models for rural two-lane run-off-road crashes

Subasish Das [a,*], Xiaoduan Sun [b], Ming Sun [b]

[a] Texas A&M Transportation Institute, Bryan, TX 77807, USA
[b] University of Louisiana at Lafayette, Lafayette, LA 70504, USA

ARTICLE INFO

Article history:
Received 25 April 2020
Received in revised form 23 June 2020
Accepted 18 August 2020
Available online 27 August 2020

Keywords:
Run-off-road (ROR) crash
Roadway departure
Safety performance function
Rules based model
Cubist

ABSTRACT

During 2015–2017, the yearly average of fatalities caused by roadway departure (RwD) crashes was 19,233 in the U.S. Roadway departure crashes, or crashes in which a vehicle crosses an edge line, centerline, or otherwise leaves the traveled way, account for 52 percent of all traffic fatalities in the U.S. during this time. A majority of RwD crashes are run-off-road (ROR) crashes; these are crashes that result in a vehicle crossing the edge line on either side of the roadway. In this study, the research team analyzed seven years (2010–2016) of rural two-lane ROR crash data from Louisiana to better comprehend ROR crashes in refining safety predictions for rural two-lane highways. Statistical model (negative binomial model) and three separate machine learning models (random forest, support vector machine, and Cubist) were applied to determine the best fit models. Overall, Cubist is characterized by a better performance in estimating ROR crashes on rural two-lane roadways. The Cubist approach introduced rules-based safety performance functions (SPFs) for total and fatal and injury crashes. This approach will be beneficial for the safety practitioners in tackling localized issues in crash data with an emphasis on prediction accuracy.

1. Introduction

Robust safety prediction models facilitate the effective incorporation of roadway safety into transportation planning, design, and operation that can quantitatively predict roadway safety. Many safety prediction models were developed in previous studies. Many of these models were introduced in the first edition of the Highway Safety Manual by the American Association of State Highway Transportation Officials (AASHTO, 2010). The state of Louisiana established several goals to decrease the number of crashes and ultimately save lives as part of their 2017 Strategic Highway Safety Plan (SHSP) (LHSC, 2017). One of the emphasis areas presented in the SHSP is 'Infrastructure and Operations.' This focus area includes roadway departure (RwD) or run-of-road (ROR) crashes, intersection safety issues, and non-motorized user safety issues. This implementation plan identifies prospective safety countermeasures that can help Louisiana reduce ROR crashes on their journey to Destination Zero Deaths. During 2015–2017, there were yearly 19,233 fatalities resulted from RwD, which is 52 percent of all the traffic fatalities in the United States (HSRG, 2019). ROR crash is defined as a crash that occurs after a vehicle crosses an edge line on either side of the roadways. ROR crashes are a subset of RwD crashes when the vehicle crosses the

Peer review under responsibility of Tongji University and Tongji University Press.
* Corresponding author.
E-mail address: s-das@tti.tamu.edu (S. Das).

https://doi.org/10.1016/j.ijtst.2020.08.001
2046-0430/

centerline and edge line of the opposite direction. To effectively reduce the amount of RwD and ROR crashes and fatalities, the Federal Highway Administration (FHWA) used the Strategic Approach & Plan as a guide to implementing countermeasures that: 1) keep vehicles in designated lane, 2) provide safe recovery from departures, and 3) reduce crash severity.

In recent years, models for crash count and severity distribution for different types of roadways have become more advanced. In particular, new statistical models and machine learning algorithms have been utilized, such as data mining, multivariate statistical model, and inclusion of random parameters to empirical Bayesian and full-Bayesian hierarchical approaches. Regression models generally examine the mean effects of the factors and, consequently, disregard any subgroup or cluster effect. Any resulting interventions then often follow suit in failing to consider any subgroup effect. Non-parametric or machine learning regression techniques have recently appeared as a convenient alternative that maintains the efficiency of standard statistical models while considering the complex nature of traffic crash occurrence scenarios. Additionally, conventional safety analysis methods do not consider localized or sb-group effect in the dataset. In many cases, this issue is addressed either by using local calibration factor or subdivide the dataset with broad groups using either length or traffic volume thresholds. Rules-based machine learning models can address these issues.

It is important to note that machine learning models can determine complex and non-linear associations between a response variable and a wide range of exploratory variables without any prior knowledge of underlying processes with high prediction accuracy (McCabe et al., 2017). However, these models are not useful for practitioners due to a lack of interpretability. This paper shows the value of using rule-based modeling to identify subgroup effects without requiring assumptions on the subgroups using seven years (2010–2016) of ROR crash data on rural two-lane roadways in Louisiana. Rather than an uninterpretable machine learning model, rule-based multi-variate linear regression models provide better predictions with model explanations.

2. Literature review

A top priority of roadway safety engineers and planners is to improve roadway safety. One of the most prominent research areas within the diverse traffic safety research field is crash data analysis to assess the safety of a transportation facility (e.g., interstates, arterials, intersections). Crash prediction models can capture complex interactions in traffic safety data. They can also be used to make engineering judgments and analytical assumptions about a given crash occurrence. In 2010, Lord and Mannering (2010) conducted a comprehensive review of crash frequency studies and analyzed their limitations. Savolainen et al. (2011) conducted a similar systematic review of injury severity related studies in 2011. In 2014, Mannering and Bhat (2014) summarized analytic methods used in both crash frequency studies and injury severity studies and included suggestions for future research. Their findings suggested that the key approach of most studies was to investigate the relationship between many different variables and crash occurrence or severity.

The Highway Safety Manual (HSM) was published based on many years of highway safety research. It outlines tools and guidelines that can be used to perform quantitative safety analyses. The HSM also describes predictive methods that can be used to identify sites that have the potential for safety improvement. Part C of the recently published HSM contains crash prediction models for different types of facilities. The HSM recommends that users generate SPFs for their particular regions to improve the crash prediction process.

The HSM provides default SPFs for different roadway facilities. Many studies have developed SPFs for rural two-lane roadways (Wang et al., 2019; Wali et al., 2018; Li et al., 2017; Martz et al., 2017; Russo et al., 2016; Anarkooli et al., 2018; Cafiso et al., 2010; Hong et al., 2016; Garach et al., 2016). To use HSM models, there is a need for using local calibration factors. Some states may predict fewer crashes, while others may predict more than the default SPFs. These calibration factors account for the variations in traffic patterns, climate, topology, and other relevant factors. Numerous studies have developed state-specific calibration factors (Llopis-Castelló et al., 2019; Sun et al., 2018; Tarko et al., 2018; Mehta and Lou, 2013; Qin et al., 2014; Abdel-Rahim and Sipple, 2015). The EB method uses weighted average principle. It has been widely used in many safety studies and is recommended by the HSM (AASHTO, 2010; Pratt et al., 2018; Sun and Das, 2013a, 2013b; Sun et al., 2014; Das et al., 2018, 2013; Das, 2015; Zou et al., 2019; Das et al., 2020; Wu et al., 2018).

The literature review reveals that majority of the SPF studies ignored the consideration of local effects of the localized data. For example, rural two-lane roadway SPFs are mostly developed based on local data or calibration factors with some common sub-categories based on either length or AADT thresholds. Rules based modeling can mitigate this issue by introducing local clusters based on the properties of the used variables by keeping an emphasis on the prediction accuracy. It calls for demonstrating of using rule-based modeling to identify cluster or subgroup effect in crash data analysis. The current research effort contributes to the safety research by developing rules-based regression models to predict ROR crashes on rural two-lane roadways more accurately.

3. Research methodology

To compare the effectiveness of different estimation techniques for ROR crashes on rural two-lane highways, the project team analyzed the following three machine learning regression techniques: the cubist method, random forest algorithms, and support vector regression.

3.1. Cubist

The concepts of the Cubist framework were developed by (Quinlan, 1992, 1993, 1994). The Cubist framework places a multivariate linear model at each leaf, and then categorical decision trees are expanded to handle continuous classes with the M5 model. Because the measures are developed at each tree, outcomes from the Cubist framework are more precise than outcomes from a regression tree, which only contains a single value at each leaf. Another option for enhancing predictive power is to use similar training cases to estimate the value at a given set of points or measures. A presumption of Cubist is that it is a composite model that incorporates a model tree, reformulated as rules, with the instance-based method. Furthermore, composite models, integrating instances, and model trees are more accurate than model trees alone.

3.2. Random forest

The random forest algorithm (RF) is based on the random subspace method (Ho, 1998) and the bagging principle (Breiman, 2001). It, therefore, also depends on developing a collection of decision trees with random predictors. The critical byproducts of RF include the out of bag (OOB), variable importance measures, and error rate. The OOB value is also known as the misclassification rate, and as the number of trees increases, the value becomes smaller. The variable importance ranking is found by using classification accuracy and the Gini impurity. When a given variable is randomly changed, the importance ranking measures how much the mean squared error increases. If the prediction error shows no change when the variable is altered, then the importance measures will not change significantly. Similarly, the mean squared error (MSE) of the variable will only change slightly, indicating that the specified variable is not significant. If the MSE decreases significantly when the variable changes, then the variable is considered critical.

3.3. Support vector regression (SVR)

In 1963, Lerner and Vapnik introduced the Generalized Portrait algorithm. This method included a core algorithm used to develop Support Vector Machine (SVM), which are statistical learning theory algorithms. In 1974, Vapnik established the field of statistical learning theory (History of SVM, 2020). Reports state that Vapnik et al. presented the current form of the SVM on the basis of a separable bipartition problem at AT&T Bell Laboratories in 1992 (Smola and Schölkopf, 2004). The objective of SVM is to map the data x into a high-dimensional feature space F through a nonlinear mapping.

One of the data-driven trend recognition algorithms for function approximation and regression is Support Vector Regression (SVR). The SVR approach presumes the error approximation to the data with model generalization. Although there are various versions of the SVR, the traditional model, ϵ-SVR, is detailed in the study conducted by Smola and Schölkopf (2004). To obtain more information about SVR, readers should consult the Cornejo-Bueno et al. (2016) paper discussed in this study.

The ϵ-SVR method for aggression comprised of, given a set of training vectors $\mathbb{T} = \left\{\left(x_i, K_S^i\right), i = 1, \cdots, l\right\}$, where x_i stands for a vector of predictive variables and K_s^i is a measure of crash frequency count (specifies the facility type of rural roadways), training a model in Eq. (1):

$$\widehat{K}_s(x) = f(x) + b = w^T\phi(x) + b \tag{1}$$

where $\widehat{K}_s(x)$ is a prediction of K_s to reduce the risk function in Eq. (2):

$$Riskfunction(R) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} L\left(K_s^i, f(x_i)\right) \tag{2}$$

where the norm of w directs the model smoothness, $\phi(x)$ is a function of protection of the input space to the feature Hilbert space, b is a parameter of bias, and $L\left(K_s^i, f(x_i)\right)$ is the loss function. The steps can be used to develop an SVR model for a training dataset.

4. Data

The research team collected seven years (2010–2016) of crash data from the Louisiana Department of Transportation and Development (LADOTD). This data contains several databases separated by year, crash data, vehicle data, and geometric data. The ROR crash characteristics are defined in the 'harmful event' variables in the vehicle data. Fig. 1 illustrates the flowchart of data integration and analysis. Extracted ROR vehicle level data was later merged with crash and crash level geometric (known as DOTD table) data. LADOTD also maintains a roadway inventory database. The ROR crashes were then assigned to the roadway segments by using the control section, and logmile information (logmile from and logmile to). The total segment length of the rural two-lane roadways in Louisiana is 11,702 miles. These roadways experience 32,583 total, and 16,661 fatal and injury (known as KABC) ROR crashes in seven years. The average yearly ROR total and KABC crashes are 4655 and 2380, respectively.

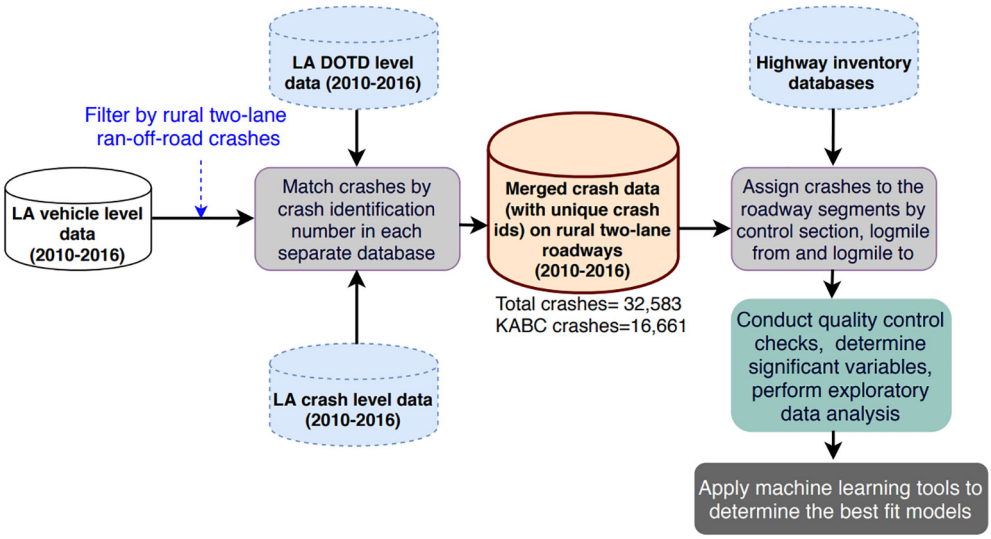

Fig. 1. Flowchart of data integration and analysis.

Table 1
ROR Crashes by Year and Severity Type.

Severity	2010	2011	2012	2013	2014	2015	2016	Total
Fatal (K)	148	127	148	116	142	128	120	929
Severe Injury (A)	59	50	63	40	54	47	44	357
Moderate Injury (B)	634	676	636	596	618	647	664	4471
Minor Injury (C)	1586	1637	1546	1494	1429	1608	1604	10,904
No Injury (O)	2178	2135	2182	2328	2278	2389	2432	15,922
Total	4605	4625	4575	4574	4521	4819	4864	32,583

Table 1 shows the number of ROR crashes for each year from 2010 to 2016. The total crashes for each year are distributed in the table based on their severity, and each level of severity is totaled for all seven years. No Injury (O) shows the highest frequency for all seven years with Minor Injury (C) following with the second-highest frequency. The severity with the lowest occurrence is Severe Injury (A) with a total of 357. It is interesting that severe ROR crashes are lower in frequency than fatal ROR crashes. From 2010–2016, similar pattern trends in each level of severity are shown. In 2016, the highest number of ROR crashes was recorded, with a total of 4864 crashes (5.6% increase from 2010).

Fig. 2 shows the distribution of ROR crashes on roads throughout Louisiana. The red color represents the areas with more ROR crashes, and the yellow colors identify the areas with less ROR crashes. The figure shows that ROR crashes are more heavily concentrated on roads in the southern part of the state compared to the northern part. This is likely because there is a larger rural roadway network in that region.

Table 2 describes the key variables for ROR crashes and shows the statistics of each variable. The traffic volumes are widely varied on these networks, with a mean value of 2254 and a standard deviation of 2398 vehicles per day (VPD).

5. Results

The final dataset contains segment level crash information from 4882 crashes. Model performance was analyzed and validated using a five-fold cross-validation procedure. For this purpose, the full dataset was separated into subsamples (n = 5) with an equal representation of data. Based on the four remaining subsamples, each sequential subsample was used to validate the trained model independently; a sample of 2000 segments was randomly selected to compare the model performance. The study aimed to lower the computation time for three different algorithmic techniques by using the sampling method. The cubist framework was initially constructed without considering composite or committee models. The standard statistical performance measures used to evaluate model performance include the coefficient of determination (R^2), Root Mean Square Error (RMSE), and mean absolute error (MAE). For example, RMSE is a measure of the dispersion of the residuals or the standard deviation of the residuals. RSME predicts the parameter values, the standard deviation of the error term with certain n degrees of freedom. The value of RMSE is expressed in Eq. **(3)**:

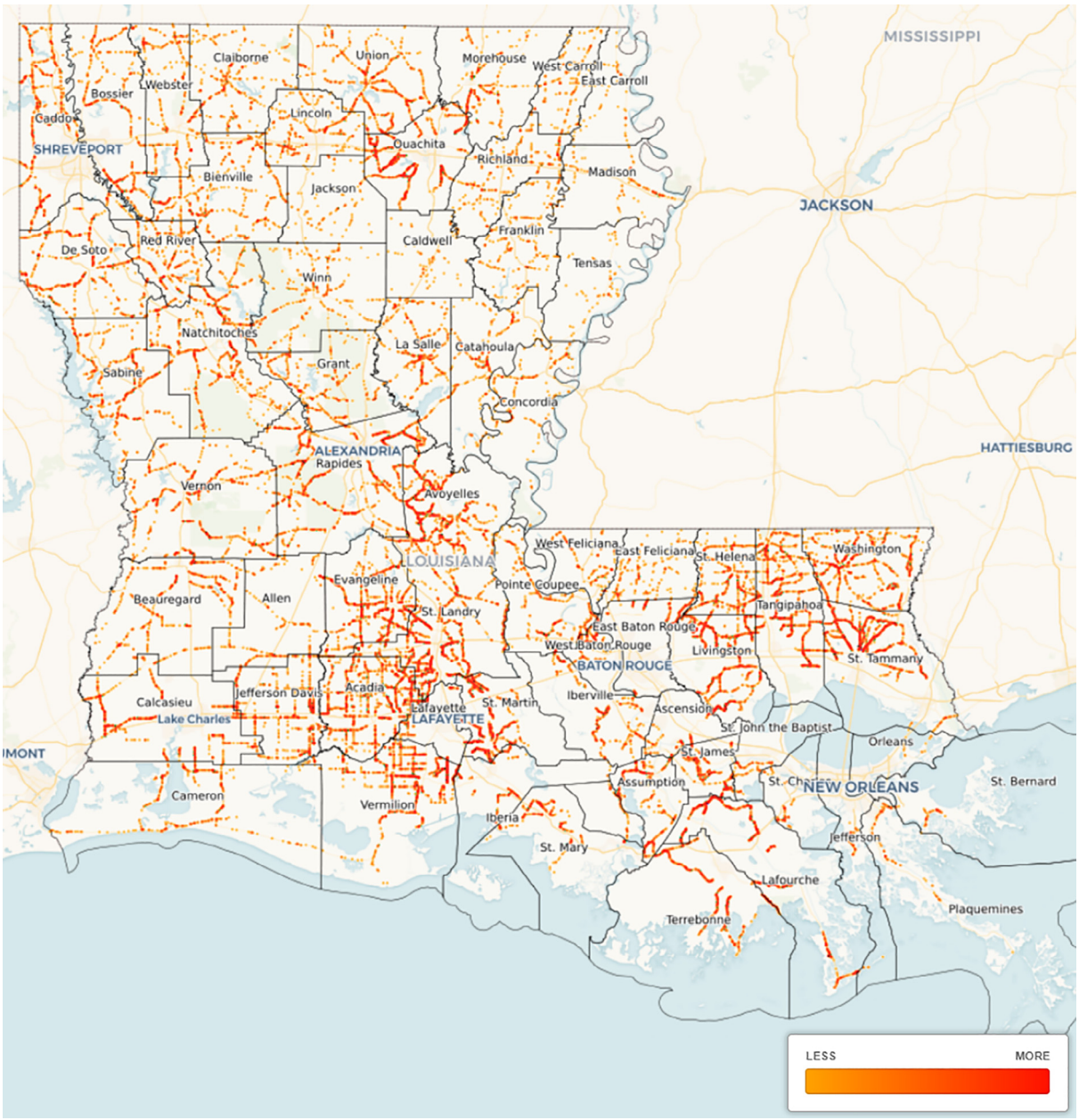

Fig. 2. Rural two-lane ROR crashes in Louisiana.

Table 2
Descriptive statistics of key variables.

Variable	Description	Mean	Std. Dev.	Min	Max	IQR[1]
Length	Segment Length (mi.)	2.40	2.47	0.01	17.40	3.27
Shou_W_R	Shoulder Width Right (ft.)	4.94	2.86	0.00	20.00	3.00
Shou_W_L	Shoulder Width Left (ft.)	4.96	3.10	0.00	89.00	3.00
Pave_Wid	Pavement Width (ft.)	22.09	2.37	9.00	64.00	4.00
AADT	Annual Average Daily Traffic (vehicle per day or vpd)	2254	2398	20	22660	2344
Tot	Total Crashes (2010–2016)	6.67	11.81	0	186	8
KABC	Fatal and Injury Crashes (2010–2016)	3.41	6.02	0	96	4

Note: [1]IQR = Interquartile Range.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \quad (3)$$

Two R packages (cubist and caret) were used in developing the models (Kuhn and Quinlan, 2018; Kuhn, 2018). The study also used negative-binomial (NB) statistical model to compare the model performance with the machine learning models. The performance of different models is illustrated by the RMSE and R^2 values listed in Table 3. Smaller RMSEs result in smaller standard errors and best-fitted models. The values show that the Cubist model yields the highest accuracy than the other two machine learning models. It can be argued that the performance of other two machine learning models are also similar to the Cubist model. It is important to note that RF and SVM are conventional black-box machine learning models with limited interpretation powers. Cubist can generate easily interpretable linear regression models based on the rules-fit. Thus, Cubist is selected as the suitable method for this analysis. The rest of the analysis was conducted using Cubist method.

Based on the performance statistics of all models, the Cubist model was later chosen to develop the rules-based SPFs. The Cubist model features a boosting-like scheme, known as a committee, where iterative model trees are generated in sequence. To manage the number of model trees, this study used the committee option. To change the predictions from the rule-based model, the cubist uses the nearest-neighbor feature. First, a model tree (with or without committees) is constructed. Then, Cubist can identify its nearest neighbors and establish the average of these training set points after a sample is anticipated by the model. Readers can consult Quinlan (1993) for the details of the adjustment determination procedure.

Cubist models can be effectively used and applied with very few specifications of model parameters (tunable). In many cases, only a number of rules will require enhancement for the given data, which makes this method extremely appealing to learn complex conjunctions between the outcome variable or response and explanatory variables. Although 80% of the segment information was designated as the training data, the remaining information was used to test the modeling performance. The maximum number of committees was determined to be 20. Three different instances (instance number *3, 5, 7*) were selected for the final model development. Based on the initial explorations, committees with more than 20 did not start to significant advancements in model predictability. To lower computation time, the instances were narrowed to 7. Fig. 3 displays the RMSE values produced from various tuning or committee-instance scenarios. The best prediction was coordinated at 10 committees- 7 instances combination for training data. The combination was 15 committees-7 instances for the test data.

Table 4 lists the model performance measures for both training and test datasets. However, the performance on the training data set shows a slightly lower RMSE score. The measures indicate that the model performs well in both training and test data sets.

At each split of the tree, Cubist saves a linear model (after performing the feature selection) that is permitted to have terms for each variable used in the current split or any previous split. The final prediction is a function comprised of all the linear models from the initial node to the terminal node. The percentages shown as the attribute usage in Table 5 reflects all the models involved in prediction. Pavement width and shoulder width (left) are identified as the least significant predictor variables, consistent with the training and test data for both total and KABC crashes (see Table 5). Compared to the test data, the two most significant predictors (segment length and AADT) of the training data show higher prediction percentages. The other data evaluation parameters (average error, relative error, and correlation coefficient) in both training and test data models are not much deviated.

Rather than relying on an uninterpretable machine learning model, the reliance on rule-based models makes Cubist more equipped to manage model explanation. SPFs are equations used to estimate the average number of annual crashes at a location as a function of contributing factors. The estimated or predicted number of crashes (N) (for a project or a site) can be predicted by multiplying three main parts: base SPF ($C_{Predicted}$), CMFs, and a calibration factor, C, as shown in Eq. (4).

$$N = C_{Predicted} \times C \times \prod_{CMF} \quad (4)$$

Table 3
Model Performances Based on Sample Data (Total Crashes).

Models	Min	1st Q.	Median	Mean	3rd Q.	Max
RMSE						
NB	8.2349	8.9701	9.3457	9.9812	12.4535	16.7001
RF	6.1182	6.9958	7.5437	7.7028	8.0980	12.1033
SVR	6.3115	7.1807	7.5596	7.7765	8.1638	11.8817
Cubist	6.1650	6.9975	7.4935	7.6974	8.0623	12.3698
R^2						
NB	0.2702	0.3124	0.3345	0.3370	0.3550	0.3890
RF	0.4013	0.4772	0.5192	0.5193	0.5581	0.6145
SVR	0.4074	0.4735	0.5123	0.5085	0.5429	0.6028
Cubist	0.4171	0.4769	0.5252	0.5165	0.5524	0.6290

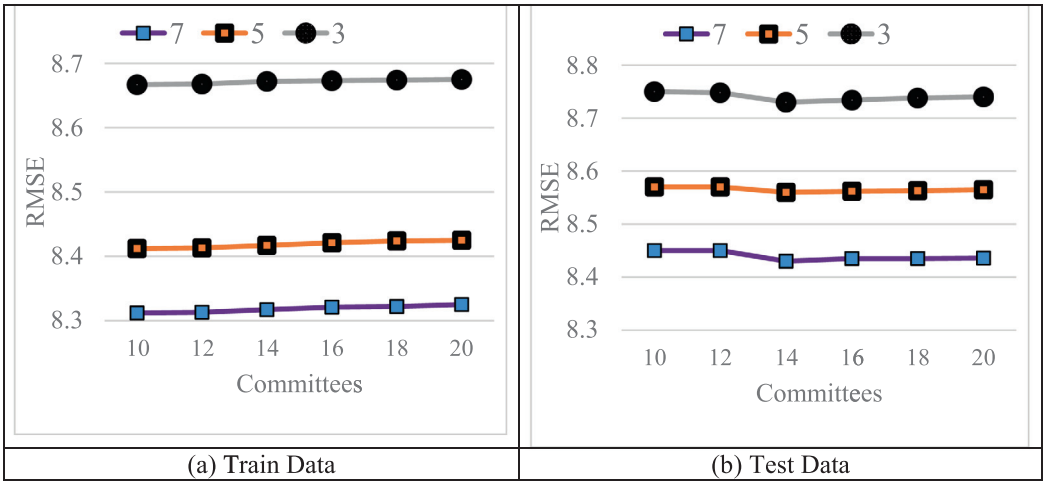

Fig. 3. Tuning parameters and RMSE values (total crashes).

Table 4
Performances of the Models on Train and Test Data.

Crash Type	Dataset	No. of Segments	Committees	RMSE	R^2	MAE
Total Crashes	**Train**	3905	20	8.3157	0.4806	4.2297
	Test	977	20	8.4126	0.4788	4.3234
KABC Crashes	**Train**	3905	20	4.1832	0.5162	2.2417
	Test	977	20	4.2615	0.5007	2.2446

Table 5
Attribute Use and Data Evaluation Parameters.

	Total Crashes		**KABC Crashes**	
	Train	**Test**	**Train**	**Test**
Attribute Use				
Length	97%	87%	99%	93%
AADT	100%	88%	98%	91%
Shou_W_R	50%	28%	47%	60%
Pave_Wid	29%	5%	20%	38%
Shou_W_L	39%	3%	19%	39%
Data Evaluation				
Average Error	4.8	4.7	2.6	2.6
Relative Error	0.67	0.62	0.68	0.68
Correlation coefficient	0.63	0.67	0.65	0.58

It is worth to mention that the current study is limited to prediction only. One can use empirical Bayes (EB) method with the SPFs from the Cubist framework. Table 6 lists the generated rules and linear SPFs developed by each rule for total crashes. It is important to note that the sum of the number of cases in train data are not same as the total number of cases for train or test data. Several rules can consider the same segment if it is in the criteria of the generated rules. These findings are in line with the findings of most of the SPF literature. In most of the rules, segment length and AADT have positive signs. Shoulder width (both right and left) usually contains a negative sign. However, for some rules, the signs of shoulder width are positive which requires further investigation. A closer look at the rules show that presence of both shoulder widths (right and left both) and larger threshold of shoulder widths are associated with these positive signs. As this study is limited to ROR crashes only, information on both shoulder widths are not directly associated with the ROR crash counts. Table 7 lists the generated rules and linear SPFs developed by each rule for KABC crashes.

One important feature of Cubist is its ability to develop rules-based regression models. Model interpretability is another important feature of Cubist. Fig. 4 shows how the model fits the training data for significant variables such as AADT and segment length. The dotted points indicate the total observed crashes by length and AADT values. The blue line denotes the fitted values (from Cubist prediction models), and the red line denotes the number of committees. The trend of the blue line suggests that the sub-group affects length in terms of the predicted values in the generated rules.

Table 6
SPFs Developed from Rules Using Training Data Model (Total Crashes).

Rules	Cases	Mean (range)	Esti. Error	SPF by Rule
Training Data				
Rule 1: Length <= 0.878	1495	0.9 (0, 19)	0.9	$Crash_i = -0.7 + 2.12\ Length + 0.00021 AADT$
Rule 2: 0.878 < Length <=2.154	801	4.3 (0, 51)	3.0	$Crash_i = -3.1 + 2.05\ Length - 0.83 Shou\_W\_R + 0.72 Shou\_W\_L + 0.00063 AADT + 0.09 Pave\_Wid$
Rule 3: Length > 2.154 and AADT <= 920	624	5.6 (0, 42)	3.6	$Crash_i = -6.9 + 0.00729\ AADT - 1.91 Shou\_W\_L + 1.6 Shou\_W\_R + 0.88\ Length + 0.21 Pave\_Wid$
Rule 4: Length > 2.154 and 920 < AADT <=1934	441	14.7 (0, 85)	8.1	$Crash_i = -2 + 0.00635\ AADT + 1.93\ Length - 0.68\ Shou\_W\_L$
Rule 5: Length > 6.498 and AADT <= 1934	227	15.0 (0, 85)	8.4	$Crash_i = -8.3 + 0.01916\ AADT + 0.7\ Length - 0.37\ Shou\_W\_L$
Rule 6: Length > 2.154 and AADT > 920	985	18.0 (0, 159)	10.0	$Crash_i = -9.7 + 0.00188\ AADT + 1.71\ Length - 0.47\ Shou\_W\_R + 0.55\ Pave_{Wid} - 0.22\ Shou\_W\_L$
Rule 7: Length > 6.498 and Shou_W_L <=5 and 920 < AADT <= 1440	48	25.3 (3, 85)	15.3	$Crash_i = -122.4 + 6.44\ Pave\_Wid$
Rule 8: Length > 6.498 and Shou_W_L <=5 and 1440 < AADT <= 1934	21	30.1 (4, 61)	16.6	$Crash_i = -179.5 + 0.12117\ AADT + 0.43\ Pave\_Wid$
Rule 9: Length > 4.33 and Shou_W_L <= 5 and AADT > 1934	115	31.0 (2, 116)	15.9	$Crash_i = 5 + 6.71\ Length - 4.5 Shou\_W\_L + 0.00065\ AADT$
Rule 10: Length > 4.33 and 5 < Shou_W_L <=7 and AADT > 1934	28	37.6 (4, 159)	32.1	$Crash_i = -55.6 + 81.96\ Shou\_W\_L + 13.22\ Length + 0.00028\ AADT$

Table 7
SPFs Developed from Rules Using Training Data Model (KABC Crashes).

Rules from Train Data	Cases	Mean (range)	Esti. Error	SPF by Rule
Training Data				
Rule 1: Length <= 1.662	2054	0.8 (0, 15)	0.8	$FI_i = -0.5 + 1.42\ Length + 0.00013\ AADT$
Rule 2: Length > 1.662 and AADT <= 954	747	2.7 (0, 24)	1.9	$FI_i = -2.7 + 0.00433\ AADT + 0.41\ Length - 0.08\ Shou\_W\_R + 0.06\ Pave\_Wid$
Rule 3: Length > 1.662 and AADT > 954	1104	8.7 (0, 96)	4.9	$FI_i = -0.2 + 1.64\ Length + 0.00109\ AADT - 0.44\ Shou\_W\_R$
Rule 4: Length > 4.32 and 954 < AADT <= 4140	368	10.8 (0, 57)	5.9	$FI_i = 2 + 0.00289\ AADT - 1.28\ Shou_{W_L} + 1.19\ Length$
Rule 5: 4.32 < Length <= 6.932 and Pave_Wid > 23 and AADT > 4140	41	16.4 (3, 58)	9.3	$FI_i = -15.9 + 24.56\ Pave_Wid + 5.51\ Length + 0.0017\ AADT - 1.18\ Shou\_W\_L$
Rule 6: Length > 4.32 and Pave_Wid <= 23 and AADT > 4140	21	20.4 (3, 53)	12.5	$FI_i = 12.5 + 0.002\ AADT$
Rule 7: Length > 6.932 and Pave_Wid > 23 and AADT > 4140	21	21.7 (0, 96)	13.5	$FI_i = -12.5 + 67.02\ Pave_{Wid} + 0.00006\ AADT + 0.03\ Length - 0.02\ Shou\_W\_L$

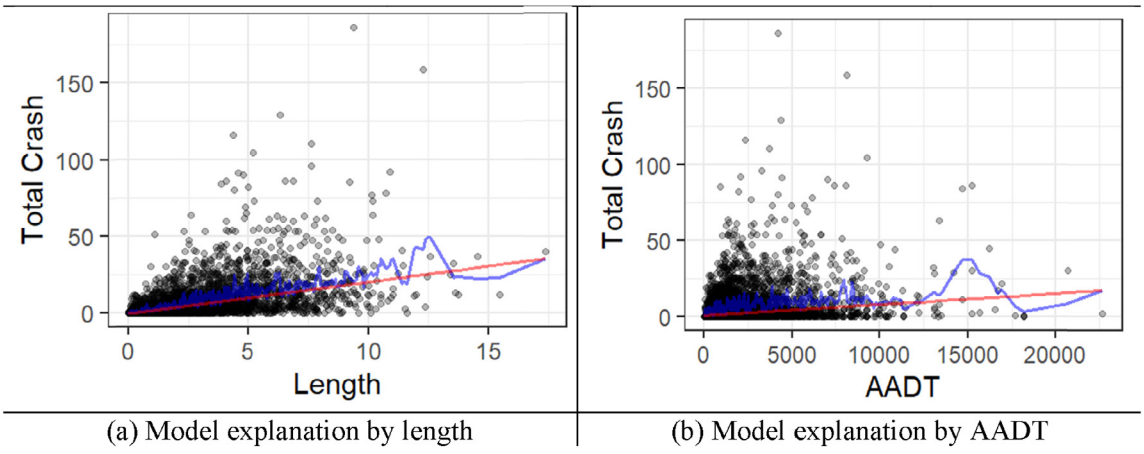

Fig. 4. Interpretation from the developed models based on training data (total crashes).

6. Conclusions

Safety improvement on rural two-lane roadways is very important. Nearly 70 percent of the Louisiana's state-maintained roadways are rural two-lane roadways. A large proportion of ROR crashes happened on these roadways. Understanding association between the key factors and crash occurrences can allow LADOTD to implement suitable countermeasures to reduce these crashes. This study applied statistical model (NB) and three machine learning models to determine the best-fit modeling techniques. With Cubist showing higher prediction accuracies compared with the other two models, the research team developed rules-based regression models for ROR total and KABC crashes on rural two-lane roadways.

This study shows that data-driven prediction algorithms such as Cubist are more robust compared to the statistical models for better prediction accuracies. Additionally, no hidden assumption is required for machine learning models. The current study can be considered as a starting point in introducing the significance of rules-based regression models to the existing SPFs. The SPFs and model interpretation visualizations developed from the models are beneficial for the safety practitioners and policymakers for easy interpretation and decision making to improve safety on rural two-lane roadways.

The current study is not without limitations. First, the current study is limited to total and KABC crashes only. There is a need for developing KA and KAB crashes, which is not done in this study. Another limitation is the usage of police reported crash data and human errors in data compilation. Limitations of the current study offer directions for future research in this domain.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors declare that the contents of this article have not been published previously. All the authors have contributed to the work described, read, and approved the contents for publication in this journal. All the authors have been certified by their respective organizations for human subject research.

References

AASHTO, 2010. Highway Safety Manual, first ed. Washington D.C..
Abdel-Rahim, A., Sipple, M.C., 2015. National Institute for Advanced Transportation Technology, Idaho Transportation Department, and Federal Highway Administration. Calibration and Development of Safety Performance Functions for Rural Highway Facilities in Idaho.
Anarkooli, A.J., Persaud, B., Hosseinpour, M., Saleem, T., 2018. Comparison of Univariate and Two-Stage Approaches for Estimating Crash Frequency by Severity—Case Study for Horizontal Curves on Two-Lane Rural Roads. Accident Analysis & Prevention, 129, 382–389.
Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.
Cafiso, S., Di Graziano, A., Di Silvestro, G., La Cava, G., Persaud, B., 2010. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. Accid. Anal. Prev. 42 (4), 1072–1079.
Cornejo-Bueno, L., Borge, J.N., Alexandre, E., Hessner, K., Salcedo-Sanz, S., 2016. Accurate estimation of significant wave height with Support Vector Regression algorithms and marine radar images. Coast. Eng. 114, 233–243.
Das, S., 2015. Effectiveness of Inexpensive Crash Countermeasures to Improve Traffic Safety. Doctoral Dissertation. University of Louisiana at Lafayette.
Das, S., Le, M., Pratt, M., Morgan, C., 2020. Safety effectiveness of truck lane restrictions: a case study on Texas urban corridors. Int. J. Urban Sci. 24 (1), 35–49.
Das, S., Sun, X., Dixon, K., Rahman, M., 2018. Safety effectiveness of roadway conversion with a two way left turn lane. J. Traffic Trans. Eng. (Eng. Ed.) 5 (4), 309–317.
Das, S., Sun, X., He, Y., Wang, F., Leboeuf, C., 2013. Investigating the Safety Impact of Raised Pavement Markers on Freeways in Louisiana. Int. J. Eng. Res. Innov. 5 (2), 74–80.
Garach, L., de Oña, J., López, G., Baena, L., 2016. Development of Safety Performance Functions for Spanish Two-Lane Rural Highways on Flat Terrain. Accid. Anal. Prev. 95, 250–265.
Highway Safety Research Group. Louisiana Crash Reports. http://datareports.lsu.edu/ (accessed April 16, 2020).
History of SVM. <http://www.svms.org/history.html> (accessed April 16, 2020).
Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20 (8), 832–844.
Hong, J., Hariharan, B., Shankar, V., Venkataraman, N., Huang, S., Kezhiyur, A.J., Milton, J.C., Van Schalkwyk, I., 2016. Random Parameter Framework for Two-Lane Rural Roadways: Findings from Washington State.
Kuhn, M., Quinlan, R., 2018. Cubist: Rule- And Instance-Based Regression. Modeling. R package version (2), 2.
Kuhn, M., 2018. caret: Classification and Regression Training. R package version 6.0-81.
Li, L., Gayah, V.V., Donnell, E.T., 2017. Development of regionalized SPFs for two-lane rural roads in Pennsylvania. Accid. Anal. Prev. 108, 343–353.
Llopis-Castelló, D., Findley, D.J., Camacho-Torregrosa, F.J., García, A., 2019. Calibration of inertial consistency models on North Carolina two-lane rural roads. Accid. Anal. Prev. 127, 236–245.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A: Policy Practice 44, 291–305.
Louisiana Highway Safety Commission (LHSC), 2017. Louisiana Highway Safety Plan: Federal Fiscal Year 2018. Baton Rouge, LA.
Mannering, F., Bhat, C., 2014. Analytic methods in accident research: methodological frontier and future directions. Analytic Methods Accident Res. 1, 1–22.
Martz, P., Bill, A.R., Khan, G., Noyce, D.A., 2017. Safety Performance Function for Undivided Rural Two-Lane Roadways Using Regression Tree Analysis.
McCabe, M.F., Rodell, M., Alsdorf, D.E., Miralles, D.G., Uijlenhoet, R., Wagner, W., Lucieer, A., Houborg, R., Verhoest, N.E.C., Franz, T.E., Shi, J., Gao, H., Wood, E. F., 2017. The future of Earth observation in hydrology Hydrol. Earth Syst. Sci. 21, 3879–3914.
Mehta, G., Lou, Y., 2013. Calibration and development of safety performance functions for Alabama: two-lane, two-way rural roads and four-lane divided highways. Trans. Res. Rec.: J. Trans. Res. Board 2398, 75–82.

Pratt, M., Geedipally, S., Wilson, B., Das, S., Brewer, M., Lord, D., 2018. Pavement Safety-Based Guidelines for Horizontal Curve Safety. Report No. FHWA/TX-18/0-6932-R1.
Qin, X., Zhi, C., Vachal, K., 2014. Calibration of Highway Safety Manual Predictive Methods for Rural Local Roads.
Quinlan, J., 1993. Combining instance-based and model-based learning. In: Proceedings of the Tenth International Conference on Machine Learning, pp. 236–343.
Quinlan, J., 1994. Improved use of continuous attributes in C4. 5. J. Artif. Intell. Res. 4, 77–90.
Quinlan, J., 1992. Learning with continuous classes. Proc. fifth Int. Conf. Artif. Intell. 92, 343–348.
Russo, F., Busiello, M., Dell'Acqua, G., 2016. Safety performance functions for crash severity on undivided rural roads. Accid. Anal. Prev. 93, 75–91.
Savolainen, P., Mannering, F., Lord, D., Quddus, M., 2011. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accid. Anal. Prev. 43, 1666–1676.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics Computing 14 (3), 199–222.
Sun, C., Edara, P., Brown, H., Berry, J., Claros, B., Yu, X., 2018. Midwest Transportation Center, Missouri Department of Transportation, C. University of Missouri, A. Iowa State University, and Office of the Assistant Secretary for Research and Technology. Missouri Highway Safety Manual Recalibration.
Sun, X., Das, S., 2013a. A Comprehensive Study on Pavement Edge Line Implementation. Report No. FHWA/LA.13/508.
Sun, X., Das, S., 2013b. Developing Louisiana Crash Reduction Factors. Report No. FHWA/LA.12/506.
Sun, X., Das, S., Zhang, Z., Wang, F., Leboeuf, C., 2014. Investigating Safety Impact of Edgelines on Narrow, Rural Two-Lane Highways by Empirical Bayes Method. Trans. Res. Rec.: J. Trans. Res. Board 2433, 121–128.
Tarko, A.P., Romero, M., Hall, T., Sultana, A., 2018. Purdue University, Indiana Department of Transportation, and Federal Highway Administration. Updating the Crash Modification Factors and Calibrating the IHSDM for Indiana.
Wali, B., Khattak, A.J., Waters, J., Chimba, D., Li, X., 2018. Development of safety performance functions: incorporating unobserved heterogeneity and functional form analysis. Trans. Res. Record: J. Trans. Res. Board 2672 (30), 9–20.
Wang, K., Zhao, S., Jackson, E., 2019. Functional forms of the negative binomial models in safety performance functions for rural two-lane intersections. Accid. Anal. Prev. 124, 193–201.
Wu, L., Geedipally, S., Pike, A., 2018. Safety Evaluation of Alternative Audible Lane Departure Warning Treatments in Reducing Traffic Crashes: An Empirical Bayes Observational Before–After Study. Trans. Res. Record: J. Trans. Res. Board, 2672(21), 30–40.
Zou, Y., Ash, J., Park, B., Lord, D., Wu, L., 2019. Empirical Bayes estimates of finite mixture of negative binomial regression models and its application to highway safety. J. Appl. Statistics 45 (9), 1652–1669.