J. Mod. Transport. (2019) 27(2):129–140
https://doi.org/10.1007/s40534-018-0178-6

www.springer.com/40534

Complementary parametric probit regression and nonparametric classification tree modeling approaches to analyze factors affecting severity of work zone weather-related crashes

Ali Ghasemzadeh[1] · **Mohamed M. Ahmed**[1]

Received: 12 February 2018 / Revised: 26 November 2018 / Accepted: 28 November 2018 / Published online: 21 December 2018

Abstract Identifying risk factors for road traffic injuries can be considered one of the main priorities of transportation agencies. More than 12,000 fatal work zone crashes were reported between 2000 and 2013. Despite recent efforts to improve work zone safety, the frequency and severity of work zone crashes are still a big concern for transportation agencies. Although many studies have been conducted on different work zone safety-related issues, there is a lack of studies that investigate the effect of adverse weather conditions on work zone crash severity. This paper utilizes probit–classification tree, a relatively recent and promising combination of machine learning technique and conventional parametric model, to identify factors affecting work zone crash severity in adverse weather conditions using 8 years of work zone weather-related crashes (2006–2013) in Washington State. The key strength of this technique lies in its capability to alleviate the shortcomings of both parametric and nonparametric models. The results showed that both presence of traffic control device and lighting conditions are significant interacting variables in the developed complementary crash severity model for work zone weather-related crashes. Therefore, transportation agencies and contractors need to invest more in lighting equipment and better traffic control strategies at work zones, specifically during adverse weather conditions.

✉ Ali Ghasemzadeh
aghasemz@uwyo.edu

Mohamed M. Ahmed
mahmed@uwyo.edu

[1] Department of Civil and Architectural Engineering, University of Wyoming, Laramie, WY 82071, USA

Keywords Adverse weather · Work zone · Safety · Crash characteristics · Probit model · Decision tree

1 Introduction

The widespread aging of the US roads and the increase in traffic demand have raised the need for more transportation maintenance projects; hence, affecting both safety and operations of roadways. Over 27% of 67,523 work zone crashes in 2013 involved injuries or fatalities. In fact, over 579 fatalities were reported due to work zone-related crashes in 2013 [1]. A number of studies have been conducted on work zone crashes. Results from these studies showed that crash rate and frequency are increased by the presence of work zones [2–6]. Weng et al. [7] used association rules to investigate factors that might have a significant effect on work zone crash casualties. They found that driving under the influence, the speed limit over 40 mph and work zones without traffic control devices have the most significant effects on work zone casualty risk. Debnath et al. [8] identified the most frequent hazards at work zones from road-worker perspective. They found that speeding vehicles are the common work zone hazard. Wang and Qin [9] investigated the severity of single-vehicle crashes at work zones. They concluded that speeding is one of the key factors affecting the severity of work zone crashes. Characteristics of work zone rear-end crashes were analyzed by Qi et al. [10], and they found that driving under the influence, lighting condition, the presence of pedestrians, and roadway defects have the highest effects on severity of work zone rear-end crashes.

It is worth mentioning that there is no agreement among different studies about the severity of work zone crashes. More specifically, some studies reported that work zone

crashes were significantly more severe than non-work zone crashes [11]. On the other hand, some studies claimed that work zone crashes were less severe than non-work zone crashes [12, 13]. There is another group of studies, which mentioned that there is no significant difference between the severity of work zone and non-work zone crashes [14, 15]. While these studies are important and useful to understand factors affecting work zone crashes, it is largely unknown whether the harmful effects of adverse weather conditions can exacerbate the severity of work zone crashes or not.

Prior experience from road safety studies suggested that adverse weather conditions significantly affect the operations and safety of roadways [16–20]. More than 7,400 people are killed, and over 673,000 people are injured due to weather-related crashes every year on the US roadways [21]. Qin et al. [22] found that heavy rain leads to a fewer number of crashes but more severe crashes. Yu et al. [23] found that crashes are extremely affected by weather conditions, especially those that occurred in mountainous freeways. Huang et al. [24] found that the severity of fog- and smoke-related crashes is higher compared with crashes in clear weather condition. Eisenberg et al. [25] analyzed effects of snowfalls on crashes injury severity. They found that snowy days have less fatal crashes than dry days; however, they have more injury and property damage only (PDO) crashes. Ahmed et al. [26] analyzed the interaction between roadway geometry and real-time weather and traffic data on mountainous freeways. They found that the probability of being involved in a crash could be doubled during the snowy seasons. Another study depicted that both injury and non-injury crash rates increase during the winter season; however, injury crashes are more severe during snowy winter season in comparison with other non-snow seasons [27].

Nevertheless, there is still a lack of studies examining the impact of weather conditions on work zone injury severity. This paper utilized data from the second Strategic Highway Research Program (SHRP2) Roadway Information Datasets (RID) to shed some lights on the different factors affecting work zone crash frequency and severity in different weather conditions.

The contributions provided in this paper are emphasized as follows. First, considering the fact that weather data extracted from weather stations are not that accurate, this study utilized weather information provided in police crash reports. It is worth mentioning that a previous study by the authors showed about 60% accuracy in identifying weather-related crashes using the weather data obtained from the weather stations [28]. Second, a new methodology, which is based on combining the traditional probit regression analysis and data mining decision tree technique, was utilized to overcome the limitations of each standalone modeling technique. Third, factors affecting the severity of work zone weather-related crashes were identified and discussed.

2 Data source

This study utilized 8 years (2006–2013) of crashes including 3,028 weather-related crashes (1,887 PDO and 1,141 fatal + injury) to determine factors affecting work zone weather-related crashes in Washington State. The crash data were extracted from the Strategic Highway Research Program 2 (SHRP2) Roadway Information Dataset (RID).

RID consists of roadway data collected from mobile data collection project, government, public, and private parties, in addition to supplemental crash history data which is used in this study. For more information about RID, see [29].

3 Methodology

In this study, factors affecting work zone weather-related crashes were identified using a complementary parametric and nonparametric crash severity model. In fact, both parametric and nonparametric models have their own advantages and disadvantages. For example, parametric models such as probit and logistic regression can provide the relationship between a response variable and predictors, and the results obtained from them are easy to interpret and understand [30]. However, the problem with these models is that there are many pre-assumptions in parametric models, which might negatively affect the accuracy of the results. Risk factors can also exhibit various exposure effects in different circumstances in parametric models (hidden effects problem). These shortcomings cannot be addressed using the common parametric models such as logistic and probit regression models [31]. One of the major solutions to address the hidden effects problem is to split the full sample data into several sub-datasets using the nonparametric classification tree method [31]. This method might be an effective way to see the effects of risk factors' hidden exposure in different circumstances. In addition, using nonparametric models is beneficial as these models have the ability to provide high prediction accuracy [31]. Therefore, this study utilized the probit–classification tree method, which is a complementary method utilizing both nonparametric (classification tree) and parametric (probit regression) models, to analyze the effect of different risk factors on work zone weather-related crashes. The dependent variable in the model is the crash severity levels, and the explanatory variables are the factors which influence

crash severity levels. In this study, the crash severity is defined as a binary variable, which has two levels including severe crashes (injury and fatal crashes) and non-severe crashes (PDO crashes).

3.1 Decision tree

A decision tree can be used for both continuous and nominal target variables. When a decision tree is used to predict a continuous target variable, it is called a regression tree, and when it is used to classify a nominal target variable, it is called a classification tree [32]. Two main components of decision trees are the "root node" and the "leaf node." The root node is the node located at the top of the tree and contains all the data, and the "leaf node" refers to the termination node and has the lowest impurity. More specifically, based on the independent variable (splitter) that creates the best homogeneity, the root node is divided into two child nodes. This procedure (partitioning the target variable recursively) will be continued until all the data in each node reach their highest homogeneity; then, tree growing will stop. This node, which does not have any branches, is called the "leaf node," and each path from the top of the tree (the root node) to each leaf (the terminal node) can be considered as a rule. It is worth mentioning that data in each child node are purer (more homogenous) than the data in the upper parent node [33]. In order to find possible splits among all variables, a splitting criterion (test) is performed. Splitting criterion is the main design component of a decision tree [34]. More specifically, in the decision tree learning algorithm, the splitting criterion's role is to measure the quality of each possible split among all variables. Two common splitting criteria that can be used to grow a decision tree are Chi-square and Gini reduction. In this study, Gini splitting criterion is used to select which variable and split pattern will be used to best split the node.

Gini impurity shows the level of data impurity. More specifically, it shows the incorrect classification probability of a randomly chosen record from the specific node in the subset. The procedure of selecting variables and split scheme that can be used to best split the parent node is as follows [35]:

1. Determine the node impurity: Considering the t as a parent node, the node impurity $i(t)$ can be calculated using the Gini index definition, which is provided in Eq. (1).

$$i(t) = 1 - \sum_{j}^{M} \left(\frac{n_j}{N}\right)^2, \tag{1}$$

where M represents the number of classes, n_j represents the number of class j elements, and N depicts all elements in the node.

If a node is homogeneous, then the value obtained from Eq. (1) would be minimal. However, the value would be higher for the less pure nodes.

2. Determine the impurity reduction (Δi): For all possible splits in the values for the variable x, the impurity reduction on the parent node t caused by a split s is calculated as follows:

$$\Delta i(x, s, t) = i(t) - \sum_{j=1}^{n_t} F_j i(t_j), \tag{2}$$

where n_t is the number of child nodes of the parent node t; F_j is the proportion of class j elements divided by all elements in the node (the proportion mentioned in the parenthesis in Eq. (1)). $\Delta i > 0$ indicates that elements in the child nodes are purer than elements in the parent node, and $\Delta i \leq 0$ denotes that elements in the child nodes are not purer than the parent node.

The best split which can be shown by s^* associated with the variable x can be identified by comprehensively searching all possible splits related to the variable x. More clearly, s^* causes the maximum impurity reduction.

3. Determine the global maximum impurity reduction: The previous step would be repeated, and the best splits for all variables would be determined. Among all the best splits, the split s^* associated with the variable x^* determines the global maximum impurity reduction.
4. Determine the leaf node: If $\Delta i(x^*, s^*, t) > 0$, then choose the variable x^* and split s^* to split the parent node. If $\Delta i(x^*, s^*, t) \leq 0$, then the parent node will be considered as the leaf node.
5. Stop the splitting: By satisfying one of these two criteria, the splitting would be stopped:
   (a) Any nodes cannot be further split ($\Delta i(x^*, s^*, t) \leq 0$), or the number of elements in the node is less than the preset minimum number.
   (b) The current tree depth reaches the preset maximum tree depth. Otherwise, go to Step 1.

For more information about the decision tree method, see [32].

3.2 Probit regression

Probit regression is a common statistical method that can estimate the coefficients of predictors [36]. This method was chosen to develop a separate probit regression for each group identified by the decision tree model. The forward

selection method was used to find the best fitting model. The advantage of using probit regression model is the ability to estimate marginal effects. Probit regression is accordingly implemented to each data group to find the coefficient of each explanatory variable.

Probit regression has the S-shape and can be used for dealing with binary response variables. Probit model (probit$[P(x)] = y$) is shown as Eq. (3) [36].

$$y = \alpha + \beta x, \tag{3}$$

where α is the probability of response when explanatory variables are the reference level (or when $x = 0$) and β represents the change in the probability per unit change in x, a parameter to be estimated.

The cumulative distribution function (CDF) $F(x)$ for a random variable x could be defined as $F(x) = P(X \leq x)$ in a way that x could be varied between $-\infty$ and ∞. By increasing the x, $F(x)$ increases from 0 to 1 because the probability of having success would be increased. Let us consider x as a continuous random variable and plot the cumulative distribution function as a function of x; the result looks like an S-shape when $\beta > 0$ [36].

The best definition for the relation of CDF, normal distribution and probit model provided by Agresti [36] is "*when F is the CDF of a normal distribution, the model provided in Eq.* (3) *is equivalent to the probit model*." This statement is clearly shown in Eqs. (4), (5), and (6) as the normal density function provided in Eq. (6) and the CDF of normal distribution depicted in Eq. (5) [36].

$$P(x) = F(x). \tag{4}$$

As mentioned, assuming the F follows a normal cumulative distribution:

$$F(x) = \Phi(x) = \int_{-\infty}^{x} \phi(z)\mathrm{d}z, \tag{5}$$

where $\phi(z)$ is the normal density function and defined as in Eq. (6).

$$\phi(z) = \frac{\pi \exp\left(\frac{-(z-\mu)^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}}, \quad z \sim N(0,1), \quad \mu = -\frac{\alpha}{\beta}, \quad \sigma = \frac{1}{|\beta|}, \tag{6}$$

where z follows the normal distribution with the mean of 0 and standard deviation equal to 1.

For more information about probit regression, see [36].

3.3 Variables description

Table 1 shows the selected variables for developing the work zone crash severity model in adverse weather conditions. The dependent variable is crash severity with two levels: severe crashes (fatal and injury crashes) and non-severe crashes (PDO crashes). Explanatory variables can be considered as driver behavior, roadway conditions, environmental conditions, and crash characteristics.

4 Results and discussion

The dataset was divided into three sub-datasets (training, validation, and testing). Among them, 60% of the data were assigned to the training sub-dataset, 30% to the validation sub-dataset, and 10% to the testing sub-dataset. More clearly, 3023 observations were used and divided among mentioned sub-datasets (i.e., 1813 observations were assigned to the training sub-dataset, 906 observations were assigned to the validation sub-dataset, and 302 observations were assigned to the testing sub-dataset). In addition, the maximum tree depth of 10, which is suitable considering the number of variables, was selected. The minimum number of data per node was selected as 40 so that all the crash severity classes have enough data for the subsequent probit analysis in each leaf node.

The Gini reduction test found any node at level four at the significance level of 0.05 to be insignificant based on the obtained results. Hence, the tree stops growing at this level, and the tree comprising four leaf nodes is selected. Optimal decision tree size ensures mitigating the overfitting issue. Two criteria were considered: First, the misclassification rate for the training and validation datasets should be similar to some extent; and second, the misclassification rate and average squared error should be as low as possible. Misclassification rates for the training and validation were obtained as 0.32 and 0.35, respectively. The average square errors obtained were 0.21 and 0.22 for the training and validation datasets, respectively. Therefore, both criteria were satisfied based on the obtained results.

Figure 1 shows the selected decision tree structure for the probit–classification tree model. As can be seen, collision type, the presence of a traffic control device, and lighting conditions were found to be significant interacting variables. Based on the classification shown in Fig. 1, the training dataset was divided into four groups representing four leaf nodes, including

Group (1): collision type = 3 and 4 (sideswipe and other crashes);
Group (2): collision type = 1 and 2 (rear-end and angle crashes) and traffic control = 2 (none);
Group (3): collision type = 1 and 2 (rear-end and angle crashes) and traffic control = 1 (stop/signal/yield sign/other) and lighting condition = 2 (other);

Table 1 Variables descriptions

Variable	Description	Type	Definition	Assigned code
Crash severity level	Severity of a crash	Binary	PDO	1
			Fatal + injury	2
Speed	Maximum posted speed	Binary	Below 50 mph	1
			Above 50 mph	2
Gender	Driver's gender	Categorical	Male	1
			Female	2
Age	Driver's age	Categorical	Young < 25	1
			Middle (25–64)	2
			Old > 64	3
Land use	If the crash occurred in an urban area	Binary	Urban	1
			Rural	2
Vehicle involved	Number of motor vehicles involved	Categorical	Greater than or equal to three	1
			Less than three	2
Collision type	Type of crash	Categorical	Rear-end	1
			Angle	2
			Sideswipe	3
			Other	4
Driving under the influence	If the driver was under the influence of alcohol or drugs	Binary	Sober	1
			Drunk	2
Roadway characteristic	If there is a curve at the crash location	Binary	No	1
			Yes	2
Lighting condition	If the crash occurred during the daylight	Binary	Yes	1
			No	2
Weather	Type of severe weather condition	Categorical	Rain	1
			Snow	2
			Fog	3
			Other	4
Vehicle action before crash	Vehicle last action before the crash	Categorical	Stopped	1
			Making left or right turn	2
			Changing lane	3
			Slowing	4
			Other (going straight, etc.)	5
Roadway type	Roadway type in crash location	Categorical	One way	1
			Two-way divided	2
			Two-way undivided	3
			Other	4
Traffic control devices	Presence of traffic control devices at crash location	Binary	Stop/signal/yield sign/other	1
			None	2
Vehicle age	Vehicle age	Categorical	Less than 5 years	1
			Between 5 and 10 years	2
			Greater than 10 years	3
Vehicle type	Whether the vehicle is considered as a big vehicle, a small vehicle or a motorcycle	Categorical	Big vehicles (bus, truck, pickup, etc.)	1
			Small vehicles (sedan, coupe, etc.)	2
			Motorcycle	3

Table 1 continued

Variable	Description	Type	Definition	Assigned code
Airbag	Whether the vehicle is equipped with airbag or not	Categorical	Equipped	1
			Not airbag equipped	2
Construction type	Different construction types at crash location	Categorical	Construction	1
			Maintenance	2
			Utility	3
			Other	4
Crash location	Whether the crash is within work zone or in external traffic backup caused by work zone	Binary	Within work zone	1
			In external traffic backup	2
Surface condition	Surface condition at crash location	Binary	Dry	1
			Other (wet, snow, etc.)	2

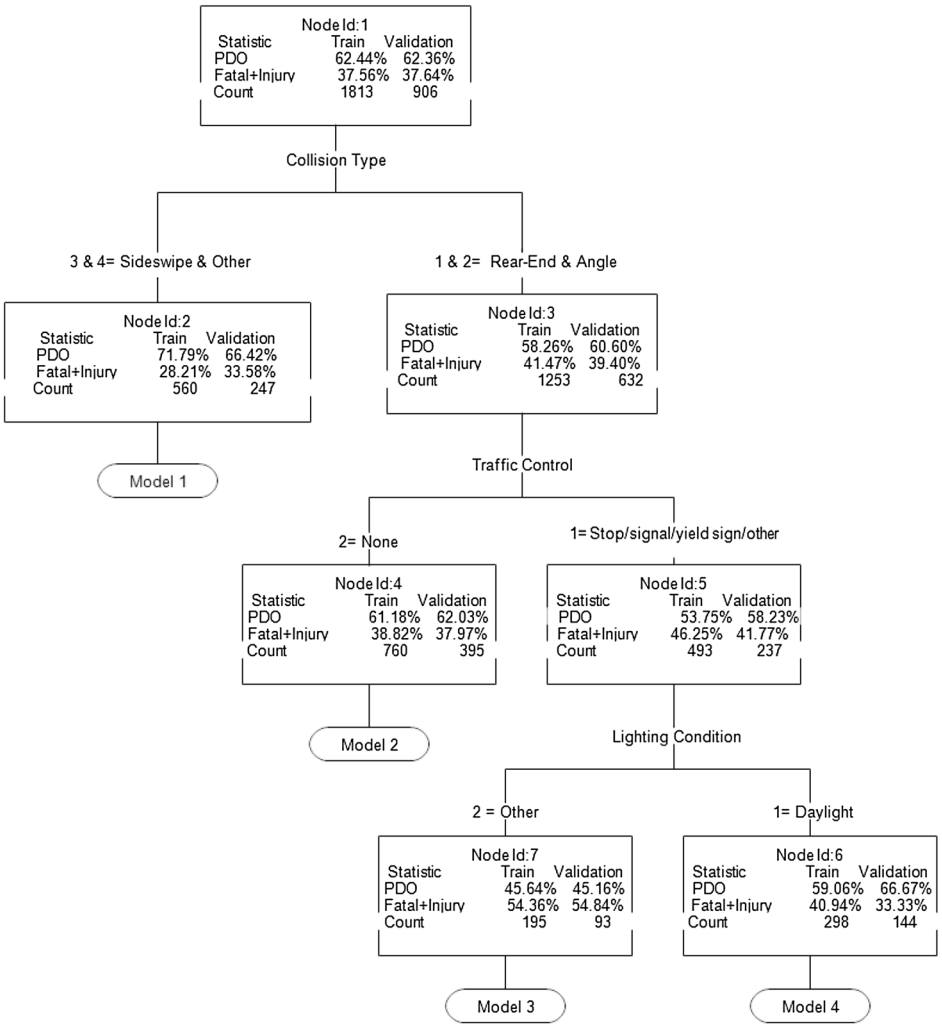

Fig. 1 Decision tree structure for the logistic-decision tree model

Group (4): collision type = 1 and 2 (rear-end and angle crashes) and traffic control = 1 (stop/signal/yield sign/other) and lighting condition = 1 (daylight).

For each above-mentioned group, a separate probit regression model was developed, and the marginal effect of risk factors for work zone weather-related crashes was calculated. For instance, in case of Group (2), 19 variables are available for developing a probit model, since the lighting condition is equal to 2, which means that this variable is fixed in this model. It is worth mentioning that for Group 2, collision type still has two categories (rear-end and angle crashes), and therefore, this variable is not fixed and will be used for further analysis in the probit model.

Table 2 shows the results obtained from developing a separate probit model for each classified group using the estimated decision tree algorithm. In fact, four models were estimated for work zone weather-related crashes using the probit–classification tree procedure. To confirm the suitability and fitness of the models, three statistics including Akaike information criterion (AIC) statistic, the Schwarz criterion (SC) statistic, and the − 2 log-likelihood statistic were used, and the results are given in Table 3.

As can be seen in model 1, the marginal effects of risk factors on the crash severity of a driver who is involved in collision types 3 and 4 (sideswipe and other) in work zone weather-related crashes can be identified. Speed limit is one of the factors found to be significant in model 1. In fact, a higher speed limit dramatically increases the severity of work zone weather-related crashes. More specifically, drivers in work zone weather-related crashes occurred in a segment with less than 55 mph posted speed limit were 13% less likely to be involved in severe crashes (the marginal effect is − 0.13 in Table 2) in comparison with the reference level (drivers who were driving in a work zone segment with a posted speed limit greater than 55 mph). In fact, driving over the speed limit could be risky because it provides insufficient time for

Table 2 Estimation of probit classification tree regression for work zone weather-related crash severity

Model	Variable		Coefficient	SE	Wald Chi-square	*P* value	Marginal effect	
							d*y*/d*x*	SE
Model 1	Intercept		0.5876	0.2132	7.5971	0.0058	–	–
	Speed limit	1	− 0.3769	0.0972	15.0471	0.0001	− 0.1353255	0.0313671
	Vehicle age	1	− 0.3709	0.1120	10.9751	0.0009	− 0.1333592	0.0309113
	Vehicle age	2	− 0.0637	0.1046	0.3715	0.5422	− 0.0262309	0.0060800
	Sobriety	1	− 0.5067	0.1833	7.6390	0.0057	− 0.1323963	0.0306881
	Gender	1	− 0.1876	0.0928	4.0818	0.0433	− 0.0548812	0.0127209
	Roadway characteristic	1	0.2864	0.0987	8.4145	0.0037	− 0.0665939	0.0154358
	Collision type	3	− 0.2600	0.0948	7.5265	0.0061	− 0.0746425	0.0173014
Model 2	Intercept		− 1.0630	0.5961	3.1806	0.0745	–	–
	Roadway type	2	1.1801	0.2017	34.2453	< 0.0001	0.3710220	0.0917906
	Speed limit	1	− 0.2757	0.1092	6.3771	0.0116	− 3.6055224	0.0182795
	Vehicle type	1	− 0.1790	0.0776	5.3202	0.0211	− 2.5847597	0.6394681
	Crash location	1	− 0.2710	0.0982	7.6225	0.0058	− 0.1153495	0.0285374
	Roadway characteristic	1	− 0.2302	0.1041	4.8938	0.0270	− 0.0930068	0.0230098
	Surface condition	2	0.8450	0.2979	8.0437	0.0046	0.2948777	0.0729526
	Collision type	1	− 0.4907	0.2117	5.3725	0.0205	− 0.1845932	0.0456683
Model 3	Intercept		6.7347	334.7	0.0004	0.9839	–	–
	Sobriety	1	− 1.3562	0.3125	18.8339	< 0.0001	− 0.4300327	0.1684546
	Airbag	1	− 0.4780	0.2352	4.1319	0.0421	− 0.1320891	0.0517426
	Vehicle involved	2	− 0.6947	0.1659	17.5375	< 0.0001	− 0.2196478	0.0860416
Model 4	Intercept		− 0.7614	0.3911	3.7887	0.0516	–	–
	Vehicle action before crash	1	0.2731	0.1345	4.1238	0.0423	0.0793240	0.0205579
	Vehicle action before crash	2	− 1.0430	0.5113	4.1618	0.0413	− 0.3807359	0.0986730
	Speed limit	1	− 0.3002	0.1405	4.5622	0.0327	− 0.1242374	0.0321978
	Weather condition	1	0.9319	0.3227	8.3385	0.0039	0.1333412	0.0345572
	Land use	1	0.3460	0.1459	5.6216	0.0177	0.1054947	0.0273404

SE standard error

Table 3 Pure probit regression

Parameter		Estimate	SE	Wald Chi-square	*P* value	Marginal effect	
						dx/dy	SE
Intercept		− 0.4316	0.1834	5.5401	0.0186	–	–
Vehicle action before crash	1	0.1439	0.0650	4.9053	0.0268	0.0613226	0.0088342
Vehicle action before crash	2	− 0.5228	0.1567	11.1332	0.0008	− 0.1802264	0.0259636
Roadway type	1	− 0.4321	0.1938	4.9698	0.0258	− 0.1553926	0.0223860
Roadway type	2	0.2769	0.0990	7.8209	0.0052	0.0938238	0.0135164
Roadway type	3	0.3120	0.1003	9.6716	0.0019	0.1096382	0.0157946
Speed limit	1	− 0.1901	0.0627	9.1990	0.0024	− 0.0718147	0.0103457
Traffic control	2	− 0.1597	0.0534	8.9583	0.0028	0.0551023	0.0079381
Sobriety	2	0.6802	0.1291	27.7685	< 0.0001	− 0.2363750	0.0340524
Gender	1	− 0.1469	0.0487	9.1061	0.0025	− 0.0533711	0.0076887
Roadway characteristic	2	0.2216	0.0645	11.7957	0.0006	− 0.0803298	0.0115724
Surface condition	2	0.2705	0.1374	3.8766	0.0490	0.0793566	0.0114322
Land use	1	− 0.1587	0.0666	5.6741	0.0172	0.0563574	0.0081189
Vehicle involved	1	− 0.1547	0.0490	9.9761	0.0016	− 0.0548034	0.0078950
Collision type	1	0.2220	0.0673	10.8795	0.0010	0.0742384	0.0106949
Collision type	2	0.5739	0.1374	17.4388	< 0.0001	0.1945923	0.0132321
Collision type	3	− 0.2636	0.0969	7.4078	0.0065	− 0.0918510	0.0132321

suitable response to control and handle unexpected situations, and this might be exacerbated during adverse weather conditions such as heavy rain or fog.

The findings indicated that among those drivers who had sideswipe and other types of crashes (all crashes except rear-end and angle crashes), female drivers had a higher casualty risk in comparison with male drivers. Two factors have been identified for the higher risk of being injured associated with the female drivers at work zone crashes in the literature, i.e., risk-taking behavior, which is higher in younger female drivers, and increased number of female drivers (exposure effect) [31, 37]. Based on the identified marginal effects for model 1, male drivers were 5% less likely to be involved in severe crashes in comparison with female drivers.

Vehicle age is another factor that turned out to be a significant predictor in model 1. Results showed that older vehicles were severely impacted in work zone weather-related crashes. This result confirms previous studies, which depicted the impact of vehicle age on driver injury severity [38, 39]. More specifically, vehicles less than 5 years old were 13% less likely to be involved in severe work zone weather-related crashes in comparison with those over 10 years old. It was also found that 5–10-year-old vehicles were 2% less involved in severe work zone weather-related crashes in comparison with older vehicles.

Driving under the influence is another factor that came out to be significant in model 1. Particularly, sober drivers were 13% less likely to be involved in severe work zone weather-related crashes in comparison with drivers under the influence, which supports previous studies [40, 41].

Roadway characteristics (presence of a curve) were also found to have a significant effect in this model. More specifically, drivers who were involved in a work zone weather-related crash at non-curved sections were 6% less likely to be involved in severe crashes in comparison with drivers who were involved in a work zone weather-related crash at a curve. The presence of curves at work zone segments could adversely affect the sight distance and handling of the vehicle in critical situations. The adverse weather might exacerbate this negative effect. Risky overtaking might be another reason for severe crashes at work zones on road curves.

The model also estimated the effect of collision type (sideswipe crashes) on the severity of crashes. More clearly, drivers who were involved in sideswipe crashes were 7% less likely to be involved in severe crashes in comparison with other collision types. It is worth mentioning that most of the sideswipe crashes were not severe in comparison with head-on crashes at work zones, which will be discussed in model 2 interpretation.

Model 2 indicates the results of the probit model for those drivers who were involved in rear-end and angle crashes, in locations with no traffic control devices. Particularly, model 2 showed that route type is a contributing factor in crash severity model. Drivers who were involved in work zone weather-related crashes in rural or secondary routes were 37% more likely to be involved in severe

crashes in comparison with drivers who were driving on interstate highways. This could be because most of the rural roads are narrow two-way two-lane undivided roads which do not have enough shoulder width. Even secondary roads do not have the same quality as interstate roadways, which might intensify the severity of crashes in rural and secondary roads.

Vehicle type was found to be a significant factor in model 2. Drivers who were driving passenger cars and involved in work zone weather-related crashes were 2.5 times less likely to be involved in severe crashes in comparison with those driving other vehicles.

Crash location was found to be another significant factor in this model. The results showed that drivers who were involved in crashes within a work zone area were 11% less likely to be involved in severe crashes in comparison with drivers who were involved in crashes in external traffic backup caused by work zone. This finding is supported by other studies which noted lane closure might increase the traffic conflicts, as a road work mostly requires closing a part of the roadway [42]. This situation might be worse in adverse weather conditions as the low visibility for drivers could prevent them to perform appropriate maneuver in critical circumstances.

The model also showed the significant effect of surface conditions on the severity of crashes. Drivers who were involved in work zone crashes were 29% more likely to be involved in severe crashes if the surface is wet in comparison with a dry surface condition. The marginal effect of rear-end crashes (collision type = 1) is given in Table 2, which indicated that drivers who were involved in rear-end crashes (most common crashes in work zones) were 18% less likely to be involved in severe crashes in comparison with angle crashes during adverse weather conditions.

Model 3 specifies the results of the probit model for those drivers who were involved in rear-end and angle crashes, with the presence of traffic control devices and in the absence of daytime lighting conditions. Similar to model 1, sobriety came out to be significant, which means that the crash severity is affected by this factor. The marginal effect of a risk factor on crash severity showed that drivers equipped with airbags were 13% less likely to be involved in severe crashes, which is consistent with a previous study [31].

In addition, results showed that crashes with fewer vehicles involved were 21% less likely to be severe. This could be because of the severity of chain crashes. Asking drivers to keep the safe following distance and avoid tailgating might be helpful. Installing dynamic message signs at work zones could be beneficial to provide advisory messages for drivers.

Finally, model 4 reveals the results of the probit model for those drivers who were involved in rear-end and angle crashes, with the presence of traffic control devices and in daytime lighting conditions. Vehicle action before crash, speed limit, weather conditions, and land use came out to be significant factors affecting crash severity model. Indeed, vehicles making a left or right turn before a crash were 38% less likely to be involved in severe crashes in comparison with vehicles going straight (category 5 in Table 1).

It was also found that drivers who were driving on roads with posted speed limits less than 55 mph were 12% less likely to be involved in severe crashes. In addition, drivers who were involved in rainy weather condition crashes were 13% more likely to be involved in severe crashes. Finally, drivers who were driving in a rural area were 10% more likely to be involved in severe work zone weather-related crashes in comparison with those driving on urban roadways. This might be due to the fact that risky behaviors such as not wearing seatbelts are more common among rural drivers [43].

As mentioned earlier, another conventional probit model with all variables was developed to compare the results with complementary probit–classification tree models. All the 20 variables used for developing probit–classification tree were considered in developing the conventional probit model as well. Forward stepwise selection method was used to find the best model fit. The results of the probit regression model are given in Table 3.

As can be seen, 11 variables came out to be significant in the conventional probit model. However, some important contributing factors including vehicle age, vehicle type, crash location, airbag, number of vehicles involved, lighting, and weather conditions are not represented in this model. The conventional probit model fit statistics are provided in Table 4.

Receiver operating characteristic (ROC) curve was used to visualize models' performance. The predictive capability of the probit–classification tree models and conventional probit (base model) is shown in Fig. 2 using ROC curves. ROC curve shows the models' performance by plotting sensitivity versus 1-specificity. More clearly, if there are no relevant information and bias results, the ROC curve would be closer to the diagonal line. On the other hand, the best performed model would be close to the upper left corner.

Figure 2 shows that for training dataset which contains most of the data, all developed probit–classification tree models performed better than the conventional probit model. It is also worth mentioning that even though small portions of dataset were assigned to validation and testing datasets (in comparison with training dataset), still using probit–classification tree method showed a better performance in comparison with a conventional probit model.

Table 4 Model fit statistics for four models obtained from probit–classification tree

	Model 1		Model 2		Model 3		Model 4		Conventional probit	
	Intercept only	Intercept and covariates	Intercept only	Intercept and covariates	Intercept only	Intercept and covariates	Intercept only	Intercept and covariates	Intercept only	Intercept and covariates
Model fit statistics										
AIC	1119.4	1081.9	1714.4	1604.7	446.5	371.1	665.1	631.5	4003.28	3844.92
SC	1124.3	1120.5	1719.6	1682.0	450.3	408.8	669.3	682.1	4009.30	3965.19
− 2 Log-likelihood	1117.4	1065.9	1712.4	1574.7	444.5	351.1	663.1	607.5	4001.28	3804.92

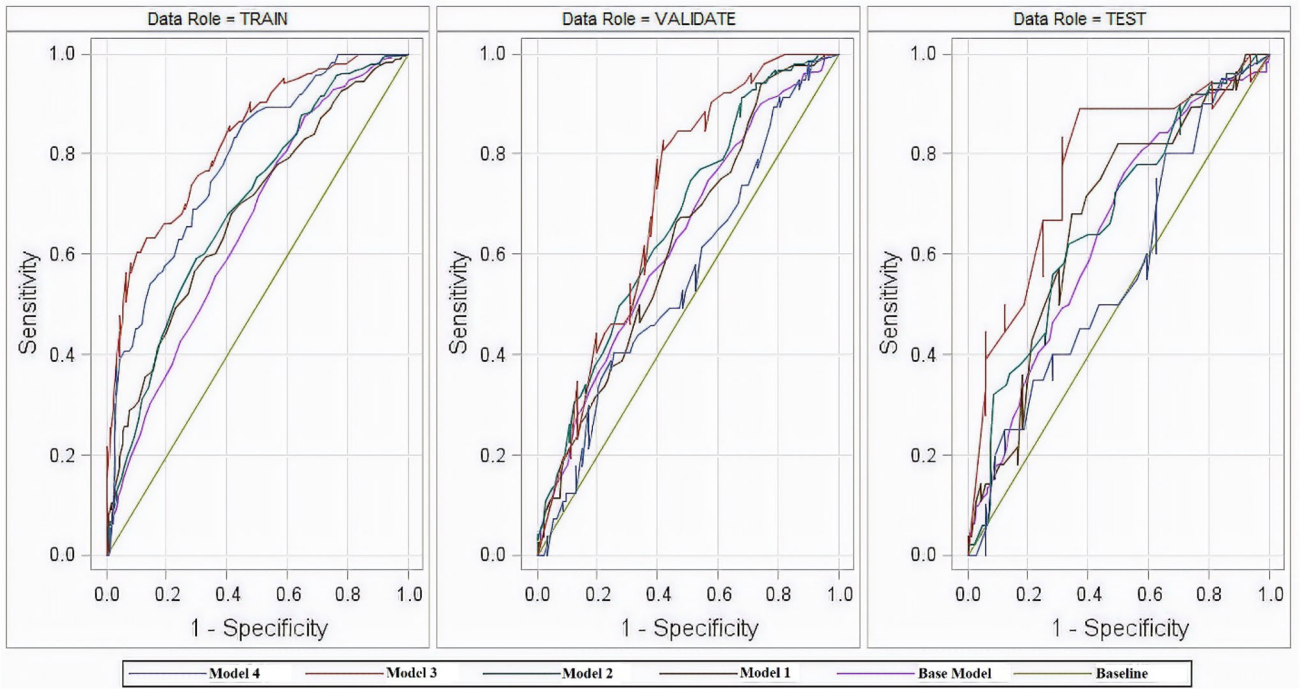

Fig. 2 Comparision among all models using ROC chart

5 Conclusions

This study explored the effects of vehicle, driver behavior, and environmental factors on work zone weather-related crashes severity using a method that combines both data mining technique and conventional parametric modeling. In addition, the obtained results from both conventional probit model and proposed probit–classification tree model were compared to better understand the advantages of the proposed model using the 8 years of work zone weather-related crashes.

The complementary methodology utilized in this study has the advantage of being used in safety analysis, as this model can compensate for the weak points of parametric and nonparametric models.

In comparison with the probit–classification tree model, some important contributing factors, such as vehicle age, vehicle type, crash location, airbag, number of the vehicles involved, lighting, and weather conditions, were excluded in the conventional probit model, which shows the capability of the developed model in identifying risk factors affecting work zone weather-related crashes.

The most interesting finding of this study is that both the presence of traffic control device and lighting conditions were found to be significant interacting variables in the developed complementary crash severity model for work zone weather-related crashes. This finding shows that more attention should be paid to the effect of weather conditions in different stages of work zone projects by transportation agencies and contractors. It is highly recommended to transportation agencies and contractors to invest in

adequate lighting equipment for nighttime work, which is the only option in most urban areas. In 2013, the American Traffic Safety Services Association developed the first and the most comprehensive set of nighttime lighting guidelines for work zones called "Nighttime Lighting Guidelines for Work Zones: A guide for developing a lighting plan for nighttime work zones" [44]. The manual provides a simple procedure for designing a nighttime lighting system for work zones that can be easily adopted by engineers, designers, and contractors without prior experience in illumination. However, more studies are needed to clarify and characterize driver behavior in different lighting, visibility, and weather conditions.

This study found that the presence of a traffic control device is a significant contributing factor to crash injury severity. Improving the temporary traffic control (TTC), including but not limited to the portable variable speed limit (VSL), to adjust speed limits during adverse weather conditions can be beneficial in this regard. Portable changeable message signs (PCMS) are also highly recommended to warn drivers about work zones to reduce the severity of work zone crashes.

Acknowledgements This work was conducted under the second Strategic Highway Research Program (SHRP2), which is administrated by the Transportation Research Board (TRB) of the National Academies of Sciences, Engineering, and Medicine, and it was sponsored by the Federal Highway Administration (FHWA) in cooperation with the American Association of State Highway and Transportation Officials (AASHTO).

References

1. Ullman GL, Pratt M, Fontaine MD et al (2018) Estimating the safety effects of work zone characteristics and countermeasures: a guidebook. Natl Coop Highw Res Progr Rep. https://doi.org/10.17226/25007
2. Ghasemzadeh A, Ahmed MM (2017) A tree-based ordered probit approach to identify factors affecting work zone weather-related crashes severity in North Carolina using the highway safety information system dataset. In: 96th Transportation research board annual meeting, Washington, D.C.
3. Ghasemzadeh A, Ahmed MM (2016) Crash characteristics and injury severity at work zones considering adverse weather conditions in Washington using SHRP 2 roadway information database. In: 95th Transportation research board annual meeting, Washington, D.C.
4. Khattak AJ, Khattak AJ, Council FM (2002) Effects of work zone presence on injury and non-injury crashes. Accid Anal Prev 34:19–29. https://doi.org/10.1016/S0001-4575(00)00099-3
5. Schrock SA, Ullman GL, Cothron AS, Cothron AS (2004) An analysis of fatal work zone crashes in Texas. Texas Transportation Institute, Texas
6. Akepati SR, Dissanayake S (2011) Characteristics and contributory factors of work zone crashes. In: Transportation Research Board 90th annual meeting
7. Weng J, Zhu J-Z, Yan X, Liu Z (2016) Investigation of work zone crash casualty patterns using association rules. Accid Anal Prev 92:43–52
8. Debnath AK, Blackman R, Haworth N (2015) A comparison of self-nominated and actual speeds in work zones. Transp Res Part F Traffic Psychol Behav 35:213–222
9. Wang K, Qin X (2014) Use of structural equation modeling to measure severity of single-vehicle crashes. Transp Res Rec J Transp Res Board 2432:17–25
10. Qi Y, Srinivasan R, Teng H, Baker R (2013) Analysis of the frequency and severity of rear-end crashes in work zones. Traffic Inj Prev 14:61–72
11. Garber N, Zhao M (2002) Distribution and characteristics of crashes at different work zone locations in Virginia. Transp Res Rec 1794:19–25
12. Garber NJ, Woo T-SH (1990) Accident characteristics at construction and maintenance zones in urban areas. No. VTRC 90-R12. Virginia Transportation Research Council
13. Hargroves BT (1981) Vehicle crashes in highway work zones. J Transp Eng 107:525–539
14. Chambless J, Chadiali AM, Lindly JK, McFadden J (2002) Multistate work zone crash characteristics. ITE J 72(5):46–50
15. Hall JW, Lorenz VM (1989) Characteristics of construction zone crashes. In: Transportation research record: journal of the transportation research board. TRB, National Research Council, Washington, D.C, pp 20–27
16. Ghasemzadeh A, Ahmed MM (2018) Utilizing naturalistic driving data for in-depth analysis of driver lane-keeping behavior in rain: non-parametric MARS and parametric logistic regression modeling approaches. Transp Res Part C Emerg Technol 90:379–392. https://doi.org/10.1016/j.trc.2018.03.018
17. Hammit BE, Ghasemzadeh A, James RM et al (2018) Evaluation of weather-related freeway car-following behavior using the SHRP2 naturalistic driving study database. Transp Res Part F Traffic Psychol Behav 59:244–259
18. Ghasemzadeh A, Ahmed MM (2017) Drivers' lane-keeping ability in heavy rain: preliminary investigation using SHRP 2 naturalistic driving study data. Transp Res Rec J Transp Res Board. https://doi.org/10.3141/2663-13
19. Ghasemzadeh A, Hammit BE, Ahmed MM, Young RK (2018) Parametric ordinal logistic regression and non-parametric decision tree approaches for assessing the impact of weather conditions on driver speed selection using naturalistic driving data. Transp Res Rec. https://doi.org/10.1177/0361198118758035
20. Ahmed MM, Ghasemzadeh A (2018) The impacts of heavy rain on speed and headway behaviors: an investigation using the SHRP2 naturalistic driving study data. Transp Res Part C Emerg Technol 91:371–384. https://doi.org/10.1016/j.trc.2018.04.012
21. Colyar J, Zhang L, Halkias J (2003) Identifying and assessing key weather-related parameters and their impact on traffic operations using simulation. In: Institute of Transportation Engineers 2003 Annual Meeting and Exhibit (held in conjunction with ITE District 6 Annual Meeting) Institute of Transportation Engineers
22. Qin X, Noyce D, Lee C, Kinar J (2006) Snowstorm event-based crash analysis. Transp Res Rec J Transp Res Board 1948:135–141

23. Yu R, Xiong Y, Abdel-Aty M (2015) A correlated random parameter approach to investigate the effects of weather conditions on crash risk for a mountainous freeway. Transp Res part C Emerg Technol 50:68–77
24. Huang H, Abdel-Aty M (2010) Multilevel data and Bayesian analysis in traffic safety. Accid Anal Prev 42:1556–1565
25. Eisenberg D, Warner KE (2005) Effects of snowfalls on motor vehicle collisions, injuries, and fatalities. Am J Public Health 95:120–124
26. Ahmed M, Abdel-Aty M, Yu R (2012) Assessment of interaction of crash occurrence, mountainous freeway geometry, real-time weather, and traffic data. Transp Res Rec J Transp Res Board 2280:51–59
27. Khattak A, Knapp K (2001) Interstate highway crash injuries during winter snow and nonsnow events. Transp Res Rec J Transp Res Board 1746:30–36
28. Ahmed M, Abdel-Aty M, Lee J, Yu R (2014) Real-time assessment of fog-related crashes using airport weather data: a feasibility analysis. Accid Anal Prev 72:309–317. https://doi.org/10.1016/j.aap.2014.07.004
29. Smadi O, Hawkins N, Hans Z, Bektas BA, Knickerbocker S, Nlenanya I, Souleyrette R, Hallmark S (2015) Naturalistic driving study: development of the roadway information database. Federal Highway Administration. SHRP 2 Report No. S2-S04A-RW-1
30. Abdel-Aty M, Abdelwahab H (2004) Modeling rear-end collisions including the role of driver's visibility and light truck vehicles using a nested logit structure. Accid Anal Prev 36:447–456
31. Weng J, Meng Q, Wang DZW (2013) Tree-based logistic regression approach for work Zone Casualty Risk Assessment. Risk Anal 33:493–504
32. Osei-Bryson K-M, Ngwenyama O (2014) Advances in research methods for information systems research. Springer, Berlin
33. Weng J, Meng Q (2011) Decision tree-based model for estimation of work zone capacity. Transp Res Rec J Transp Res Board 2257:40–50
34. Kovalerchuk B, Vityaev E (2000) Data mining in finance: advances in relational and hybrid methods. Springer, Berlin
35. Weng J, Meng Q (2012) Effects of environment, vehicle and driver characteristics on risky driving behavior at work zones. Saf Sci 50:1034–1042
36. Agresti A (2007) An introduction to categorical data analysis, 2nd edn. Wiley, New York
37. Kostyniuk LP, Molnar LJ, Eby DW (2000) Are women taking more risks while driving? A look at Michigan drivers. In: Women's travel issues second national conference
38. National Highway Traffic Safety Administration (2013) How vehicle age and model year relate to driver injury severity in fatal crashes. US Dep Transp
39. Glassbrenner D (2012) An analysis of recent improvements to vehicle safety. NHTSA Tech Rep (Report No DOT HS 811 572)
40. DeJong W, Hingson R (1998) Strategies to reduce driving under the influence of alcohol. Annu Rev Public Health 19:359–378
41. Keall MD, Frith WJ, Patterson TL (2004) The influence of alcohol, age and number of passengers on the night-time risk of driver fatal injury in New Zealand. Accid Anal Prev 36:49–61
42. Weng J, Meng Q, Yan X (2014) Analysis of work zone rear-end crash risk for different vehicle-following patterns. Accid Anal Prev 72:449–457. https://doi.org/10.1016/j.aap.2014.08.003
43. Hao W, Daniel J (2016) Driver injury severity related to inclement weather at highway–rail grade crossings in the United States. Traffic Inj Prev 17:31–38
44. American Traffic Safety Services Association (2013) Nighttime lighting guidelines for work zones. https://www.workzonesafety.org/files/documents/training/fhwa_wz_grant/night_lighting_guide.pdf