IJE TRANSACTIONS A: Basics Vol. 33, No. 7, (July 2020) 1375-1386

International Journal of Engineering

Journal Homepage: www.ije.ir

Single-vehicle Run-off-road Crash Prediction Model Associated with Pavement Characteristics

M. Akbari[a], G. Shafabakhsh*[a], M. R. Ahadi[b]

[a] *Faculty of Civil Engineering, Semnan University, Semnan, Iran*
[b] *Transportation Research Institute, Ministry of Roads and Urban Development, Tehran, Iran*

PAPER INFO

Paper history:
Received 12 March 2020
Received in revised form 30 April 2020
Accepted 12 June 2020

Keywords:
Highway Segmentation
Pavement Physical Characteristics
Crash Prediction Models
Run-off-road Accidents
Functional Forms
Nonlinear Negative Binomial Regression

ABSTRACT

This study aims to evaluate the impact of pavement physical characteristics on the frequency of single-vehicle run-off-road (ROR) crashes in two-lane separated rural highways. In order to achieve this goal and to introduce the most accurate crash prediction model (CPM), authors have tried to develop generalized linear models, including the Poisson regression (PR), negative binomial regression (NBR)), and non-linear negative binomial regression models. Besides exposure parameters, the examined pavement physical characteristics explanatory variables contain pavement condition index (PCI), international roughness index (IRI) and ride number (RN). The forward procedure was conducted by which the variables were added to the core model one by one. In the non-linear procedure and at each step, 39 functional forms were checked to see whether the new model gives better fitness than the core/previous model. Several measurements were taken to assess the fitness of the model. In addition, other measurements were employed to estimate an external model validation and an error structure. Results showed that in PR and NBR models, variables coefficients were not significant. Findings of the suggested nonlinear model confirmed that PCI, as an objective variable, follows the experts anticipation (i.e., better pavement manner associates with less ROR crashes). Finally, it should be noted that the roughness variable was insignificant at the assumed significance level, so it had no contribution to ROR crashes. The results imply that improving the pavement condition leads to a more probable decrease in the ROR crashes frequency.

***doi**: 10.5829/ije.2020.33.07a.25*

1. INTRODUCTION

Crash prediction models (CPMs) describe a mathematical relation between accident frequencies and defined explanatory variables. These models have been used to give an estimate of the expected crash frequency based on the characteristics of accident factors.

Among all types of accidents, the run-off-road (ROR) crash is one of the most severe crash occurrences. Many types of research have been done to increase the effectiveness and reliability of the existing safety plans related to ROR crashes. Some of which are related to roadside features [1-3].

However, there are several studies that include road infrastructure and geometry features [4-6].

Among all these studies, the influence of pavement condition and the riding quality aspects related to pavement physical characteristics have been neglected.

The main target of the recent study is to fill these research gaps by developing a CPM that incorporates the pavement condition and related riding quality measures to the frequency of ROR crash occurrence in two-lane separated rural highways. We will further aim to develop a way to formulate the relationship between ROR crash frequency and the following explanatory variables:

- Exposure variables (annual average daily traffic (AADT) and segment length),
- Pavement condition variable (pavement condition index (PCI)),
- Riding quality variables (international roughness index (IRI) and ride number (RN)).

*Corresponding Author Institutional Email: *ghshafabakhsh@semnan.ac.ir* (G. Shafabakhsh)

Please cite this article as: M. Akbari, G. Shafabakhsh, M. R. Ahadi, Single-vehicle Run-off-road Crash Prediction Model Associated with Pavement Characteristics, International Journal of Engineering (IJE), IJE TRANSACTIONS A: Basics Vol. 33, No. 7, (July 2020) 1375-1386

Two categories of models were used to achieve this goal, the generalized linear models (GLMs) and the nonlinear stochastic models. So, it will be recognized if the nonlinear model can give better results or if the GLM ones are significant.

1. 1. Literature Review The Primitive researches about pavement influences on accident rate go back to the 1970s. The first Pavement-Accident models were developed based on standard multiple linear-regression equations.

Later, Chan et al. [7] studied the effects of asphalt pavement conditions on traffic accidents. They developed twenty-one negative binomial (NB) models for different crash types. Those models examined the impact of the pavement characteristic variables, including rut depth (RD), international roughness index (IRI), and present serviceability index (PSI) on crash frequency rates. The primary outcome is that the RD models are insignificant for all types of accidents while the IRI and PSI parameters give a suitable fitness and execute well on the total crash data. The deficiency of their study might be related to conjunctions between single and multi-vehicle crash types in modeling.

Jiang et al. [8] correlated pavement management and traffic parameters to accident frequency occurrence. Results approved the significant impact of PSI and PDI on accident frequency. The frequency models have a direct relationship with the road roughness, i.e., less pavement roughness can be associated with less accident frequency.

Akbari et al. [6] studied the impact of pavement condition index (PCI) variable on the ROR crashes. They concluded that a unit increase in PCI variable reduces about 1.93% in frequency of ROR crashes.

It should be noted that the study on the impact of pavement characteristics on accident frequency (specially ROR crashes) are rare and this study carried out to fulfill this research gap.

2. DATA DESCRIPTION

The Semnan province has a critical situation not only in Iran (because of locating along with the internal East-West corridor), but also in Asian highways networks (as located in Asian Highway Route No. 1 (AH1) also known as the Silk Road).

The Semnan to Tehran link had some segments with the critical deteriorated pavement. High travel demand caused by heavy trucks are the reasons for such conditions that affected the pavement quality.

This Province does not have a good position in the ROR crash occurrence ranking. The presence of severe pavement distresses, and high ROR crashes are the main reasons for selecting this roadway. An executive report revealed that 42 percent of fatal crashes due to ROR crashes are related to Isfahan, Fars, Semnan and Kerman provinces and another document highlighted that 64.5 percent of all accidents which occurred in Semnan province were related to ROR crashes.

There are three types of data that are generally available for empirical analysis: time series, cross-section and pooled. Cross-sectional data, in statistics is a type of data collected by observing many subjects (such as road segments) at one point or period of time. The analysis might also have no regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences between selected subjects.

The cross-sectional sample provides a snapshot of those observed subjects, at that one point in time. Note that no one knows based on one cross-sectional sample if the dependent variable is increasing or decreasing; the model can only describe the current situation. Cross-section differs from time series, in which the same small-scale or aggregate entity is observed at various points in time. Another type of data, pooled data, combines both cross-sectional and time series data ideas and looks at how the subjects change over a time series.

Some variables like geometric design and roadside features usually keep their characteristics for years, but some variables such as pavement characteristics, gradually deteriorate every year. The main restriction of data gathering for this study is to survey pavement condition characteristics, especially for a continuous bases in several years. Analyzing the pavement condition index (PCI) for a certain road segments during a continuous years is not only too expensive but also time consuming. So in this study, the cross-sectional data of one calendar year were considered to describe the current situation.

Although a high correlation between some of the variables exists, none of them were eliminated. Significance and goodness-of-fit measurements were used to consider the model stability instead of considering their correlations [9].

In multiple regression analysis, the nature and significance of the relations between the explanatory and independent variables are often of particular interest. Multicollinearity is a common problem when estimating linear or generalized linear models, including logistic regression and Cox regression. When the predictor variables are correlated among themselves, multicollinearity among them is said to exist which leading to unreliable and unstable estimates of regression coefficients. Most data analysts know that multicollinearity is not a good thing. But many do not realize that there are several situations in which multicollinearity can be safely ignored. In this situation, the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data [10]. Multicollinearity does not reduce the predictive power or

reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors.

That is, a multivariate regression model with collinear predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.

As it discussed in previous paragraphs, this issue appears in linear or generalized linear models. These types of models are a part of our study but the main part of our modelling is refer to nonlinear negative binomial regression. So this issue is not a matter to our models comparison study.

Table 1 shows the statistical details of contributed variables. As discussed previously, the cross-sectional data of one Iranian calendar year were considered to describe the current situation of examined roadway.

2. 1. Accident Data The single-vehicle ROR crash is applied to those events in which an errant vehicle is encroaching the roadway. This encroachment may lead to three types of outcomes: non-strike incidence, crash with non-fixed objects at roadside, and collisions involving roadside features and fixed-objects. All of these accidents are considered in this study.

The accident data were collected from FAVA department of Rahvar Police Office. During the study period (i.e., 20 March 2011 to 20 March 2012) among all 373 single- and multi-vehicle crashes, there were 166 single-vehicle ROR crash occurrences in the Semnan to Tehran roadway.

Shafabakhsh et al. [11] studied the spatial analysis of frequency of rural accidents and concluded that the accidents are considerable within 30 km from Garmsar.

2. 2. Exposure The lack of exposure parameters leads definitely to uncertainty in any CPM outputs. Researchers have emphasized that AADT is one of the most critical exposure variables for any crash prediction model one of the many exposure units is a million vehicle-kilometer of travel per year. The exposure function for the i^{th} segment is given by Eq. (1). The scale of traffic data variable can be estimated as AADT/1000 [12, 13].

$$Exposure_i = \frac{AADT_i \times 365 \times Length_i}{10^6} \quad (1)$$

Homogenous segments give better results and have more significant variables. Furthermore, short road sections have undesirable impacts on the linear regression models outputs. Therefore, long segments are more suitable and recommendable.

The distance between Semnan and Tehran is 150 kilometer, which is a two-lane separated rural highway. So, the link is separated into 55 homogeneous segments based on its pavement condition.

2. 3. Pavement Surface Characteristics Data

2. 3. 1. Pavement Condition Index (PCI) The PCI is a numerical indicator that exemplifies the present condition of a pavement by considering the observed distresses. The PCI for the roadway is estimated from the collected visual survey distress data and is based on the standard practice of roads PCI surveys, ASTM D6433-11 [14].

In this study, the pavement data were collected by a car equiped to a spcial instrument that continuously takes photographs while passing the lanes. Captured photos cover a 2*3.8 square meters area with a 0.5-meter longitudinal coverage to fit consecutive margins. The 100-meter-longitudinal merged photos are processed using Road Mapper® Software, and then the procedure of PCI calculating will be performed (see Figure 1). After assessing the PCI of each 100-meter units, the PCI of defined homogenous segments is calculated by taking the average of PCIs from involved 100-meter units. The summary of defined segments PCI is given in Table 1.

2. 3. 2. International Roughness Index (IRI) The international roughness index (IRI) is a mathematical transformation of a longitudinal profile that represents pavement roughness results in vehicle vibration. This model relates the movements of a simulated quarter-car to a single longitudinal wheel path profile at a constant speed of 80 km/h. This method is certificated in ASTM E1926-08 [15]. The IRI value is zero for newly constructed roadway, and it grows up to 12 for deteriorated ones.

TABLE 1. Details of selected variables for each homogenous segment in the study

Variables and Abbreviations		Units	Min.	Max.	Mean	Std. Dev.
Exposure	Annual Average Daily Traffic (AADT)	veh/day	9570	10380	9830	380
	Segment Length	km	0.7	6.7	2.74	1.53
Pavement Condition Index (PCI)		%	12.5	99.38	69.35	30.4
International Roughness Index (IRI)		m/km	1.39	9.47	3.67	2.29
Riding Number (RN)		1 to 5	0.72	3.79	2.56	1.05
Run-Off-Road (ROR) Crashes Frequency		No.	0.0	6.0	1.16	1.26

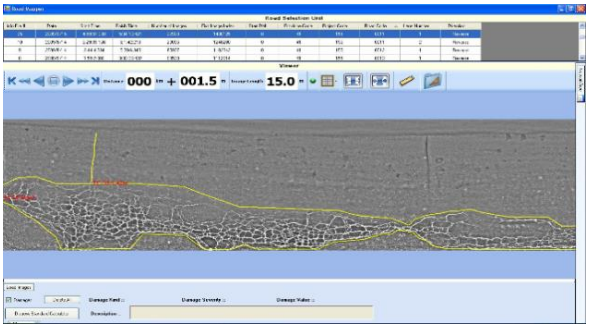

Figure 1. A view of Road Mapper Software interface and typical measurement of pavement distresses

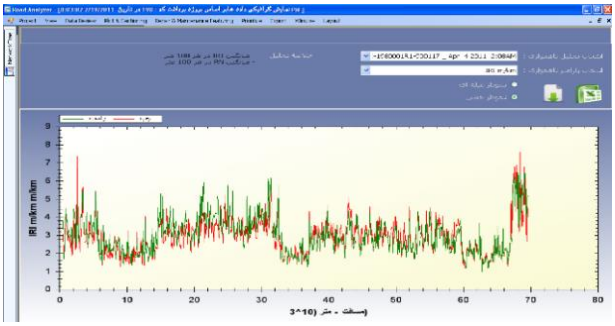

Figure 2. A The view of Road Analyzer Software interface and typical illustration of IRI oscillation

2. 3. 3. Ride Number (RN) The ride number (RN) is defined as the ride-ability index of a pavement that ranges from 0 to 5 for roadways indicator of impassable to perfect conditions, respectively. The RN depends on the physical properties of specific kinds of measurement tools and is based on the nature of the longitudinal profile. Therefore, the RN is a time-independent parameter.

In this study, the longitudinal profile of the roadway is picked up by an instrumented vehicle. After that, the Road Analyzer® software is used to analyze the collected raw data. The variables of IRI and RN will then be calculated based on the ASTM E1926-08 (see Figure 2) [15], together with ASTM E1489-08 [16]. The summary of calculated IRI and RN values are given in Table 1.

3. METHODOLOGY AND MODELS FRAMEWORK

The methodological issues and processes of crash prediction models (CPMs) should be discussed as the following procedures:

- Collecting of explanatory variables' data,
- Offering appropriate functional/model forms,
- Running suggested models,
- Evaluating Goodness-Of-Fit,
- Estimating models' error.

3. 1. Generalized Linear Regression Models Framework (GLMs) GLMs supply a class of fixed-effect regression models for dependent variables such as crash counts [17].

3. 1. 1. Poisson Regression Model Form This model develops a model of accident frequency variable Y, which follows a Poisson distribution with a mean value λ. Equation (2) describes the probability function of such models.

$$P(Y_i = y_i) = f_{Y_i}(y_i;\lambda_i) = \frac{e^{-\lambda_i}.\lambda_i^{y_i}}{y_i!} \quad (2)$$

where y_i is the number of ROR crashes on the i^{th} road segment, $P(y_i)$ is the probability of y_i count of ROR occurrences on the i^{th} road segment and λ_i is the expected value of y_i.

The model is developed through a linear predictor η_i to convert the nonlinear function $E(y_i)$ into a generalized linear one that is given in Equation (3).

$$\eta_i = g(\lambda_i) = \log(\lambda_i) = \beta_0 + \sum_{i=1}^{k} \beta_i.x_i \quad (3)$$

3. 1. 2. Negative Binomial Model Form Since accident data usually have different variances and mean values, the negative binomial (NB) regression models are recommended. Although the forms of generalized linear predictor and logarithm link function for NB and Poisson regression models are similar, they have differences shown below [18]:

- The expected value (y_i) conforms to the NB distribution.
- The error term is added to the NB model.

The expected value ($E(y_i)$) for the NB regression model, is given in Equation (4). The variance function of the NB model reassigns as Equation (5) [19].

$$E(y_i) = \lambda_i = \exp(\beta_i.X_i + \varepsilon_i) \quad (4)$$

$$Var(y_i) = \lambda_i + \lambda_i^2/\phi \quad (5)$$

The variables coefficients, β_i, and dispersion parameters, κ, are obtained by using the maximum likelihood estimation (MLE) method. The GENMOD procedure in SAS represents one of the suitable tools for this purpose. The λ_i function for both the Poisson and NB regression models are redefined by adding the exposure parameter, and therefore, the η_i function would be rewritten as Equation (6).

$$\eta_i = \log(Exposure) + \beta_0 + \sum_{i=1}^{k} \beta_i.x_i \quad (6)$$

3. 1. 3. Decision Criteria to Select GLM Type The dispersion parameter, σ_d, is estimated based on the Poisson error structure (Equation (7)). The DOF is described as the subtraction of the number of observations from that of the model variables. The Pearson χ^2 is defined in Equation (8).

$$\sigma_d = \frac{Pearson\chi^2}{\text{degree of freedom (DOF)}} \tag{7}$$

$$Pearson\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{E}(Y_i))^2}{Var(Y_i)} \tag{8}$$

where y_i is the number of occurred accidents at each segment, $\hat{E}(y_i)$ is the number of expected accidents at the segment *i*, and $Var(Y_i)$ is the variance of the dependent variable.

3. 2. Nonlinear Negative Binomial Regression Model Form The employed forward procedure to develop nonlinear CPM is shown in Figure 3. The structure of nonlinear CPM was given in Equation (9).

$$E(y) = f(Exposure) \times g(PCI) \times h(IRI) \times l(RN) \tag{9}$$

where $E(y)$ is the expected ROR crash frequency, and *f*, *g*, *h*, and *I* are the proposed functional forms for each independent variable.

The NLMIXED procedure in SAS maximizes an approximation of the log-likelihood integrated over the random effects. The log-likelihood functions for the negative binomial distribution, which must be maximized, are defined in Equations (10) and (11) [9]:

$$L(y, \mu, \kappa) = \sum_i l_i \tag{10}$$

$$l_i = y_i.\log(\kappa.\mu_i) - (y_i + 1/\kappa).\log(1 + \kappa.\mu_i) + \log\left(\frac{\Gamma(y_i + 1/\kappa)}{\Gamma(y_i + 1).\Gamma(1/\kappa)}\right) \tag{11}$$

where L is the log-likelihood function, l_i is an individual contribution to the log-likelihood, y_i is the response, μ_i is an estimated mean, and κ is the dispersion parameter.

3. 3. Functional Form Selection and Goodness-of-Fit Statistics Two groups of functional forms should be investigated to determine the goodness-of-fit (GOF) for the developed nonlinear CPM. The first one is the core functional form, and the second group includes the expanded functional forms. The recommended criteria to estimate the models' goodness-of-fit by NLMIXED procedure in SAS are $-2LL$, AIC, AICC, and BIC. All of which benefit the same rule; smaller values indicate better data fitness [20].

By estimating the likelihood ratio test (LRT) (Equation (12)) and the scaled deviance values, deciding on whether the core model is improved by adding new independent variables will depend on how close these values are to $\chi^2_{0.1,(df_2 - df_1)}$ at the 90% confidence level. The LRT and scaled deviance measures are used to test the null hypothesis which remarks all coefficients of the added variables are zero (i.e. $H_0 : \beta_{p+1} = 0$) [4, 13].

$$LRT = -2\log\left(\frac{\text{Likelihood of core model}}{\text{Likelihood of alternative model}}\right) \tag{12}$$

3. 4. Model Error Estimates The error statistics are practically beneficial to investigate how well the model fits the data. The Mean Absolute Error (MAE) criterion provides a measure of the average misprediction of the model.

In Equation (13), a value close to zero implies that the CPM predicts well the observed data [21-24].

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |O_i - E_i| \tag{13}$$

where n is the number of observations, O_i, is the measure of performance observed from the field data (i.e., the observed ROR crashes) and E_i is the measure of performance, which was estimated by the CPMs (i.e., estimated ROR crash by proposed CPMs).

Mainly, Root Mean Squared Error (RMSE) is the root of Mean Squared Prediction Error (MSPE), which is a standard indicator of the models error. The equation of RMSE is given by Equation (14). Results that are closer to zero represent better data fitness [25-27].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (O_i - E_i)^2} \tag{14}$$

4. MODELING RESULTS AND DISCUSSION

The procedure of developing models is a well-defined forward procedure. Then by considering a proper GOF measure for each functional form, the statistical significance of a new model could be examined. There are three steps for estimating the variables coefficients and their dispersion values. At each step, these two essential notes must be considered:

- The new model contains exposure variables and an added variable that has the best GOF.
- In nonlinear CPMs, that is necessary to control the suggested 39 functional forms for all the new variables.

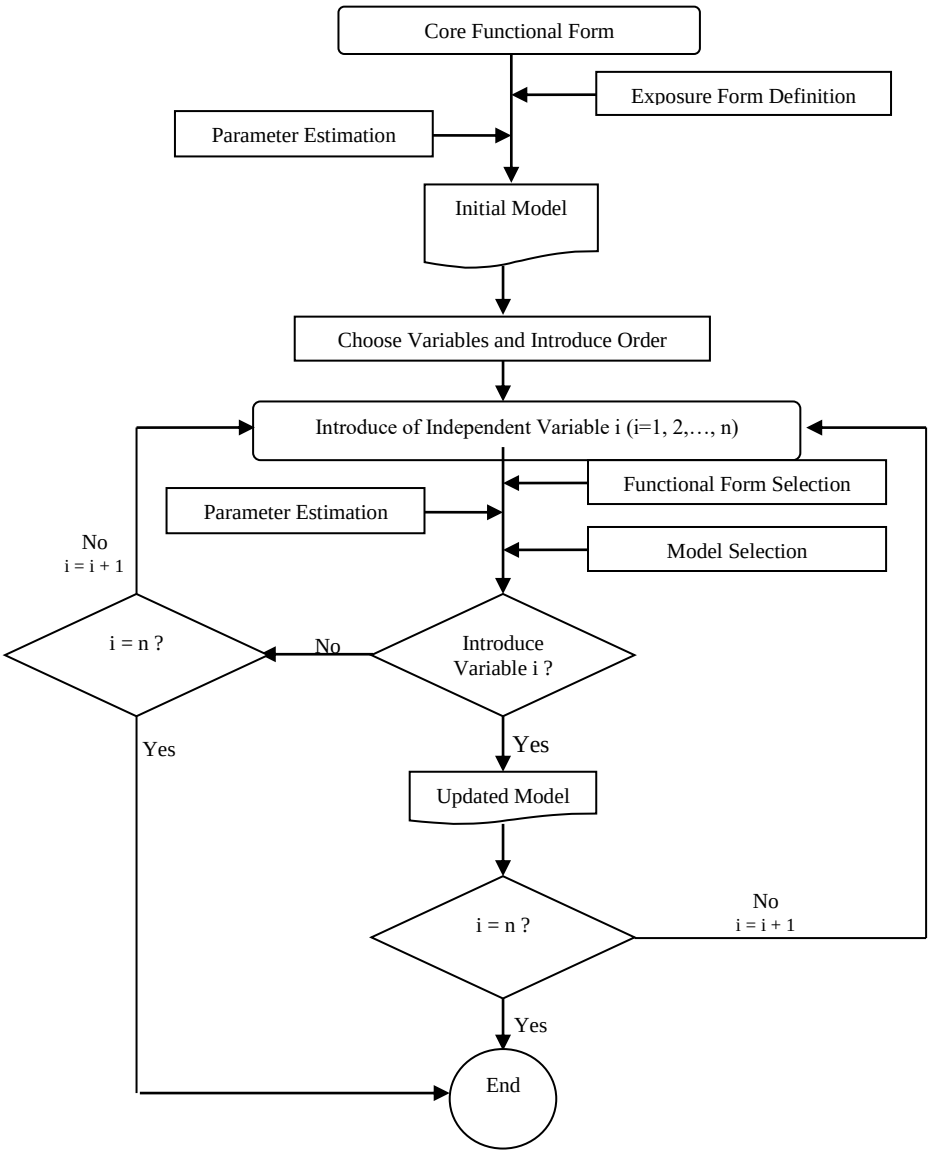

Figure 3. The nonlinear modelling procedure overview

4. 1. Parameter Estimates

4. 1. 1. Generalized Linear CPMs Based on Equation (6), and concerning both gathered ROR crashes and independent variables data, the Poisson regression (PR) and NB regression (NBR) models are developed (see Table 2).

The variables coefficient indicates the effect of a variable change on the accident frequency value. A change in a given variable by an amount of one percent corresponds to a change in accident frequency by the value of $(\exp(\beta_i)-1)\times 100\%$ changes. The p-values column in Table 2 show that all variables are not statistically significant at the level of 90%. In this regard, the necessity of using nonlinear forms of regression models can be revealed.

The criteria for assessing GLMs' GOF are the scaled deviance ratio and the σ_d value. The obtained DOF is 51. In Table 3, the scaled deviance ratio for PR and NBR are estimated to be equal to 1.6575 and 1.0327, respectively. Further, the σ_d values for PR and NBR are estimated to be equal to 2.7346 and 1.6287, respectively. Comparing these values confirms that the NB regression model is well-fitted to the data much more than the Poisson model. This result reflects the effect of dispersion parameter consideration in the NBR model, which is estimated to be equal to 0.5114.

4. 1. 2. Nonlinear Negative Binomial CPM **First step:** The first step in modeling the procedure is the formation of the core model which only contains the

exposure variables. The core model performs as a reference model and provides the base structure for introducing an alternative model which includes a newly added variable. The remained explanatory variables are added continuously to update the core model. By using the NLMIXED procedure in SAS, the coefficient of exposure variable converges to the value of 135.39 after 15 iterations.

The calculated $-2LL$ is equal to 172.5 and the BIC criterion is 180.5; these values will be the references to decide how the newly introduced independent variables can improve the CPM. The dispersion parameter value of the model is 0.5579.

Second step: PCI, IRI, and RN are separately added on to the core model, and 39 candidate functions are chosen as link equations between ROR crashes and independent variables. Rather than intrinsic functions, their coefficients, GOF criteria and dispersion values are embedded in Table 4 which examined the PCI alternative link equations.

As shown in Table 4, the 13th functional form has the lowest GOF values among the other functions which means that it has the best fitness for the study data. The β_{1-1} value is the coefficient of exposure variable which should be reassigned for each function.

These functional forms and mentioned procedures are implemented for IRI and RN variables. The results of the second step are listed in Table 5. In this table, the model name $FF_{2\text{-}1}$ is the abbreviation of the suggested functional form for the first variable's model (i.e., PCI's model) in the second step.

The p-values of coefficients emphasize that the PCI and RN parameters are significant at the assumed level of 90% but the IRI variable does not satisfy this criterion. Although the $g(PCI)$ and $g(RN)$ functions improved the core model, the PCI and RN variables will remain for the next step. The comparison of GOF criteria shows that the LRT values for all models are equal to 0.5,

TABLE 2. Analysis of parameter estimates about GLMs criteria for assessing GLMs' GOF

	Estimate		Standard Error		Wald 95% Confidence Limits				Chi-Square		Pr > ChiSq	
	PR	NBR	PR	NBR	PR		NBR		PR	NBR	PR	NBR
Intercept	3.4120	3.5032	2.0822	2.6779	-0.669	7.4931	-1.7455	8.7518	2.69	1.71	0.1013	0.1908
PCI	-0.0196	-0.019	0.0116	0.0137	-0.042	0.0030	-0.046	0.0078	2.88	1.95	0.0897	0.1630
IRI	0.145	0.1389	0.233	0.2986	-0.311	0.6016	-0.446	0.7242	0.39	0.22	0.5336	0.6418
RN	0.8469	0.8572	0.6978	0.8748	-0.520	2.2145	-0.8573	2.5717	1.47	0.96	0.2249	0.3271
Dispersion	-	0.5114	-	0.2492	-	-	0.0231	0.9998	-	-	-	-

TABLE 3. Parameters for assessing GLMs' GOF

	Value		Value/DF	
	PR	NBR	PR	NBR
Deviance	84.5309	52.6702	1.6575	1.0327
Pearson Chi-Square	139.4632	83.0612	2.7346	1.6287
Log Likelihood	-64.6943	-60.5226	-	-

TABLE 4. Examined functional forms for PCI independent variable and estimated GOF criteria of the second step

No.	Functional Forms for h(PCI)	Variables' coefficients				Goodness-of-fit criteria				κ Value
		β_{1-1}	β_{2-1}	β_{2-2}	β_{2-3}	-2LL	AIC	AICC	BIC	
1	$1+\beta_{2-1}.\log(x)$	156.66	-0.0331			172.5	178.5	179.0	184.5	0.5552
2	$1+\beta_{2-1}.x$	148.51	-0.0013			172.4	178.4	178.9	184.5	0.5519
3	$\beta_{2-1}+\beta_{2-2}.x$	6.0946	24.368	-0.0313		172.4	180.4	181.2	188.5	0.5519
4	$x^{\beta_{2-1}}$	156.10	-0.0348			172.5	178.5	179.0	184.5	0.5554
5	$\beta_{2-1}.x^{\beta_{2-2}}$	12.4941	12.494	-0.0348		172.5	180.5	181.3	188.5	0.5554
6	$x+\beta_{2-1}.x^2$	5.8231	-0.008			172.2	178.2	178.7	184.2	0.5453

7	$\beta_{2-1}.x + \beta_{2-2}.x^2$	1.6451	3.5398	-0.0285		172.2	180.2	181.0	188.2	0.5453
8	$1 + \beta_{2-1}.x + \beta_{2-2}.x^2$	1.7490	3.2892	-0.0265		172.2	180.2	181.0	188.2	0.5450
9	$\beta_{2-1} + \beta_{2-2}.x + \beta_{2-3}.x^2$	2.9694	18.9458	1.2102	-0.01	172.0	182.0	183.3	192.1	0.5412
10	$x^2 + \beta_{2-1}.x^3$	5.84E-7	-1862.7			228.4	234.4	234.9	240.5	2.4266
11	$\beta_{2-1}.x^2 + \beta_{2-2}.x^3$	-0.1718	-0.9051	0.0085		176.3	184.3	185.1	192.4	0.6445
12	$1 + \beta_{2-1}.x^2 + \beta_{2-2}.x^3$	-2.3E-8	-3.3E6	22577		188.0	196.0	196.8	204.0	0.9728
13	$x.e^{\beta_{2-1}.x}$	8.4118	-0.0193			172.0	178.0	178.5	184.0	0.5428
14	$\beta_{2-1}.x.e^{\beta_{2-2}.x}$	2.9003	2.9003	-0.0193		172.0	180.0	180.8	188.0	0.5428
15	$1 + \beta_{2-1}.x.e^{\beta_{2-2}.x}$	135.39	0.8451	-9.7402		172.5	180.5	181.3	188.5	0.5579
16	$\beta_{2-1}.x+e^{\beta_{2-2}.x}$	1.5034	1.5319	-5.5338		182.5	190.5	191.3	198.5	0.8479
17	$x^{\beta_{2-1}}.e^{\beta_{2-2}.x}$	4.4560	1.2209	-0.0232		172.0	180.0	180.8	188.0	0.5422
18	$e^{\beta_{2-1}.x}$	148.13	-0.0013			172.4	178.4	178.9	184.5	0.5522
19	$\beta_{2-1}.e^{\beta_{2-2}.x}$	12.1707	12.1707	-0.0013		172.4	180.4	181.2	188.5	0.5522
20	$1/(1 + \beta_{2-1}.x^9)$	Err	Err	-	-	Err	Err	Err	Err	Err
21	$1/(1 + e^{-(\beta_{2-1}+\beta_{2-2}.x)})$	135.39	1.0000	1.001		172.5	180.5	181.3	188.5	0.5579
22	$1/(1 + e^{-\beta_{2-1}.x})$	135.39	1.0007			172.5	178.5	179.0	184.5	0.5579
23	$\beta_{2-1}.x/\sqrt{1+x^2}$	11.6373	11.6373			172.5	178.5	179.0	184.5	0.5579
24	$\beta_{2-1}.x/\sqrt{1+\beta_{2-2}.x^2}$	4.0430	4.0430	0.0141		172.5	180.5	181.3	188.5	0.5602
25	$\beta_{2-1}.e^{-\beta_{2-2}.e^{-\beta_{2-3}.x}}$	11.6357	11.6357	1.0000	1.000	172.5	182.5	183.7	192.5	0.5579
26	$(1-\beta_{2-1}.e^{-2.x})/(1+\beta_{2-1}.e^{-2.x})$	135.39	1.0000	1.0000		172.5	180.5	181.3	188.5	0.5579
27	$2.\beta_{2-1}.e^{-x}/(1+\beta_{2-1}.e^{-2.x})$	1.22E7	1.22E7	1.0000		5114	5123	5123	5131	1.1E7
28	$\beta_{2-1}.x/(1+\beta_{2-2}.x)$	16.0773	16.0773	1.8891		172.5	180.5	181.3	188.5	0.5586
29	$1/(1+\beta_{2-1}.x)$	147.74	0.00135			172.4	178.4	178.9	184.5	0.5525
30	$1/(1+\beta_{2-1}.x^2)$	146.07	0.00001			172.4	178.4	178.9	184.4	0.5499
31	$\beta_{2-1}/(1+\beta_{2-2}.x^2)$	12.0861	12.0861	0.00001		172.4	180.4	181.2	188.4	0.5500
32	$1/(1+\beta_{2-1}.x^3)$	145.90	1.62E-7			172.4	178.4	178.8	184.4	0.5480
33	$1/(1+\beta_{2-1}.x^4)$	145.78	1.75E-9			172.4	178.4	178.8	184.4	0.5468
34	$1/(1+\beta_{2-1}.x^5)$	145.48	1.8E-11			172.3	178.3	178.8	184.4	0.5463
35	$1/(1+\beta_{2-1}.x^{\beta_{2-2}})$	135.39	-0.9308	-6.6334		172.5	180.5	181.3	188.5	0.5579
36	$1/(1+x^{\beta_{2-1}})$	135.39	-8.4585			172.5	178.5	179.0	184.5	0.5579
37	$1/(1+\beta_{2-1}.x^6)$	144.97	1.8E-13			172.4	178.4	178.8	184.4	0.5464
38	$1/(1+\beta_{2-1}.x^7)$	144.31	1.7E-15			172.4	178.4	178.8	184.4	0.5469
39	$1/(1+\beta_{2-1}.x^8)$	140.66	1.1E-17			172.4	178.4	178.9	184.4	0.5514

and they have the same priority to be selected for the next step. By considering the values $\chi^2_{0.1,1}$, these new models do not satisfy the significance requirement. It is because the LRT values are less than 2.71. Nevertheless, this small improvement rejects the null hypothesis and in the next step by adding other remained variables, CPMs will fit quite well to the data. It is essential to mention that the user must carefully interpret LRTs when the values fall under the null hypothesis boundary.

Third step: At this stage, the second variable is added to previous models. It should be noted that the adding order of these variables changes the results because of the effect of functional form selection at each stage. By considering the GOF criteria of the given models in Table 6, it appears that the $FF_{3\text{-}1}$ model has the best fit among the others.

The models name $FF_{3\text{-}1}$ is the abbreviation of the suggested functional form for the new variable (i.e., RN) at the third step. The coefficients' p-value emphasizes that the coefficients of the $FF_{3\text{-}1}$ model are significant at the level of around 90%, but the coefficients of the $FF_{3\text{-}2}$ model do not satisfy this criterion. The LRT value for the $FF_{3\text{-}1}$ model is 2.1 (concerning the $FF_{2\text{-}1}$ model) and compared to $\chi^2_{0.1,1}$ distribution with one DOF (i.e., the value of 2.71), the new model significantly improves the prior model. The final form of nonlinear CPM is defined as Equation (15).

$$ROR - Accident_{nonlinear\ CPM} = 6.4521 \times Exposure \times (PCI \times e^{-0.03331.PCI}) \times (e^{0.4735.RN}) \quad (15)$$

4. 2. Safety Effects of Independent Variables

All GOF criteria and p-values of the suggested non-linear model (i.e., Model $FF_{3\text{-}1}$ which is represented in Equation (15) reflect that all variables have a statistically significant effect on the ROR accident occurrence. The independent variables coefficients of the model clarify how these explanatory parameters are associated with the total frequency of single-vehicle ROR crashes. The positive coefficient reflects the direct relationship between the accident frequency and the independent variable. On the contrary, it has a reverse effect if the coefficient is negative. Based on this point of view, the CPM's variables will be discussed below.

4. 2. 1. PCI Variable The negative sign for PCI's coefficient predicts that segments with higher PCI values will probably have lower ROR crash rates. The exponential coefficient for PCI is equal to 0.9672 indicates that the ROR crash rate is reduced by a factor of 0.967. The percentage change in accident rate associated with 1% increase in the PCI is -3.31%. The 1.01 multiplier refers to the form of Model $FF_{3\text{-}1}$; in other words, it shows a 1% increase in PCI in addition to its exponential coefficient increment.

TABLE 5. Estimated coefficients and GOF criteria for independent variables of the second step

Model name	Variable	Functional Forms for g(.)	Variables' coefficients		Goodness-of-fit criteria				κ (p-value)
			β_{1-1} (p-value)	β_{2-1} (p-value)	-2LL	AIC	AICC	BIC	
$FF_{2\text{-}1}$	g(PCI)	$PCI.e^{\beta_{2-1}.PCI}$	8.4118 (0.0195)	-0.0193 (0.001)	172.0	178.0	178.5	184.0	0.5428 (0.0397)
$FF_{2\text{-}2}$	g(IRI)	$1/(1+\beta_{2-1}.IRI^4)$	143.06 (<.0001)	0.000082 (0.5877)	172.0	178.0	178.5	184.1	0.5573 (0.0353)
$FF_{2\text{-}3}$	g(RN)	$RN.e^{\beta_{2-1}.RN}$	161.99 (0.0330)	-0.3956 (0.0197)	172.0	178.0	178.4	184.0	0.5546 (0.0363)

TABLE 6. Estimated coefficients and GOF criteria for independent variables of the third step (p-values are given in parenthesis)

Model name	Variables' order	Functional Forms for h(.)	Variables' coefficients			Goodness-of-fit criteria				κ (p-value)
			β_{1-1} (p-value)	β_{2-1} (p-value)	β_{3-1} (p-value)	-2LL	AIC	AICC	BIC	
$FF_{3\text{-}1}$	g(PCI)×h(RN)	$e^{\beta_{3-1}.RN}$	6.4521 (0.032)	-0.033 (0.005)	0.4735 (0.110)	169.9	177.9	178.7	185.9	0.5115 (0.04)
$FF_{3\text{-}2}$	g(RN)×h(PCI)	$1+\beta_{3-1}.PCI$	140.51 (0.053)	-0.063 (0.839)	-0.0069 (0.011)	170.4	178.4	179.2	186.4	0.5154 (0.04)

This outcome is logical and compatible with experts expectations and drivers experiments. Distresses force driver to change his direction rapidly, so this fast reaction leads to careless steering most of the time and a ROR crash is consequently expectable.

4. 2. 2. RN Variable The functional form of Model FF3-1 clarifies that as the riding quality grows up, the ROR crashes increases. Referring to Equation (15), the RN's exponential coefficient is equal to 1.6056 and an increase in the RN value by a unit will result in a 60.56% increase in the ROR accident rate. The RN's coefficient sign is positive; that means if the RN value increases, higher accident rates will be expected. As mentioned in section 2.3.3., RN is a subjective riding quality criterion. This definition confirms that the riders anticipations and reactions have an important role in the ROR accident occurrence. The higher RN values allow the rider to drive monotonously and inattentively; by considering such behaviors, the results of the model will be acceptable and consistent with what has been expected.

4. 3. Models' Error Structure As mentioned before, to justify the external validation and the overall accuracy of CPM, It is necessary to investigate the error statistics. The MAE value equals to 1.09 and confirms the acceptability of model fitness on the data. Furthermore, the RMSE value is equal to 1.5 which approves the accuracy of suggested nonlinear negative binomial CPM.

5. CONCLUSION

This study aimed to investigate the extent of the pavement physical characteristics and riding quality effects on the occurrence of single-vehicle ROR crashes on two-lane separated rural roadways. For this purpose, three types of databases were obtained, including AADT, physical pavement features (including length, PCI, IRI and RN of segments) and ROR crash data. By obtaining the raw data and processing them, 55 homogenous segments with a total length of 150.5 km were categorized to develop reliable CPM. The examined CPMs were classified into two different models; the generalized linear regression models (GLMs) and the nonlinear multivariate negative binomial regression model. These regression models were developed to estimate the variables coefficients and other GOF criteria.

Literature reviews showed that researchers agree with developing CPMs which fits well on the data with small counts such as ROR crashes. The results of developed GLMs imply that the models variables are not significant at the assumed level. In order to achieve this goal non-linear forms of negative binomial multivariate regression models were investigated. The proposed model indicated that the explanatory variables are complicated enough to use a non-linear model for rural roads accidents.

Based on GOF criteria, the proposed non-linear model (i.e., Equation (15)) statistically provides significant improvements in the model fit. The p-values of variables coefficients show that the PCI and RN variables are significant but IRI could not satisfy this criterion in modeling. Also, the given values of error estimate measures imply that Model $FF_{3\text{-}1}$ provides quite reliable predictions of ROR crash occurrence related to pavement physical characteristics.

The sign and coefficient of variables reasonably follow the expected outcomes. The proposed model shows that an increment in the PCI (which is an objective riding quality criterion) will cause a drop in the ROR crash frequency. Based on the suggested functional form of Eq. (15), a unit improvement in PCI (as a pavement manner criterion) corresponds to a 3.31 percent reduction in single-vehicle ROR crashes. This effectiveness emphasizes the importance of pavement management and maintenance programs. Therefore, the road safety authorities are responsible for periodically controlling the pavement condition and following the scheduled maintenance and repairing programs.

The RN's coefficient represents a different attitude; however, such interpretations become acceptable if we consider the concept of subjective riding quality that refers to the drivers anticipation. The variables coefficient shows that a unit increase in RN value associates with a 60.56 percent increase in the ROR crash rate. The RN is calculated from longitudinal profile measurements and is used to estimate subjective ride quality. Hence, that is not used to measure any pavement roughness or distress. This variable defines the comfort level of riding and relates to the nature of the longitudinal roadway profile. Highway engineers usually prefer to construct infrastructures that have high grades of RN. However, the results of this study reveal that these high grades most likely associate with higher single-vehicle ROR crashes. That is since the drivers seem to not expect a dangerous situation and they usually follow their steadily riding, which might result in driving weakness and/or drowsiness.

6. REFERENCES

1. Fitzpatrick, C.D., Harrington, C.P., Knodler Jr. M.A., Romoser, M.R.E., "The influence of clear zone size and roadside vegetation on driver behavior", ***Journal of Safety Research,*** Vol. 49, (2014), 97-104. doi: 10.1016/j.jsr.2014.03.006
2. Carrigan, C.E., Ray, M.H., "A new approach to run-off-road crash prediction", Proceeding of the 96th Annual Meeting Compendium of TRB, Washington D.C., United States, (2017). https://www.roadsafellc.com/linked/carrigan17h.pdf
3. Torre, F.L., Tanzi, N., Yannis, G., Dragomanovits, A., Richter, T., Ruhl, S., Karathodorou, N., Graham, D., "Accident prediction in

European countries: Development of a practical evaluation tool", Proceeding of the 7th Transport Research Arena (TRA), Vienna, Austria, (2018).

4. Theofilatos, A., Yannis, G., "A review of the effect of traffic and weather characteristics on road safety", ***Accident Analysis and Prevention,*** Vol. 72, (2014), 244-256. doi: 10.1016/j.aap.2014.06.017
5. Ambros, J., Valentová, V., Sedoník, J., "Developing updatable crash prediction model for network screening: Case study of Czech two-lane rural road segments", ***Journal of the Transportation Research Board,*** Vol. 2583, (2016), 1-7. doi: 10.3141/2583-01
6. Akbari, M., Shafabakhsh, Gh., Ahadi, M.R., "Evaluating the safety effects of pavement condition index (PCI) on frequency of run-off-road accidents", ***Journal of Transportation Infrastructure Engineering,*** Vol. 1, No. 3, (2015), 47-61. doi: 10.22075/jtie.2015.316
7. Chan, C.Y., Huang, B., Yan, X., Richards, S., "Investigating effects of asphalt pavement conditions on traffic accidents in Tennessee based on the pavement management system (PMS)", ***Journal of Advanced Transportation,*** Vol. 44, No. 3, (2010), 150-161. doi: 10.1002/atr.129
8. Jiang, X., Huang, B., Zaretzki, R.L., Richards, S., Yan, X., "Estimating safety effects of pavement management factors utilizing Bayesian random effect models", ***Traffic Injury Prevention,*** Vol. 14, No. 7, (2013), 766-775. doi: 10.1080/15389588.2012.756582
9. Jafari, R., Hummer, J.E., "Safety effects of access points near signalized intersections", Proceeding of the 92nd Annual Meeting Compendium of TRB, Washington D.C., United States, (2013).
10. Ashuri, A., Amiri, A., "Drift change point estimation in the rate and dependence parameters of autocorrelated poisson count processes using MLE approach: An application to IP counts data", ***International Journal of Engineering, Transactions A: Basics,*** Vol. 28, No. 7, (2015), 1021-1030. doi: 10.5829/idosi.ije.2015.28.07a.08
11. Shafabakhsh, Gh., Famili, A., Akbari, M., "Spatial analysis of data frequency and severity of rural accidents", ***Transportation Letters,*** Published Online: 08 Mar 2016, (2016). doi: 10.1080/19427867.2016.1138605
12. Roque, C., Cardoso, J.L., "Investigating the relationship between run-off-the-road crash frequency and traffic flow through different functional forms", ***Accident Analysis and Prevention,*** Vol. 63, (2014), 121-132. doi: 10.1016/j.aap.2013.10.034
13. van Petegema, J.W.H.(J.H.), Wegman, F., "Analyzing road design risk factors for run-off-road crashes in the Netherlands with crash prediction models", ***Journal of Safety Research,*** Vol. 49, (2014), 121-127. doi: 10.1016/j.jsr.2014.03.003
14. ASTM D6433-11, "Standard practice for roads and parking lots pavement condition index (PCI) surveys", American Standard, (2011). doi: 10.1520/D6433
15. ASTM E1926-08, "Standard practice for computing international roughness index (IRI) of roads from longitudinal profile measurements", American Standard, (2015). doi: 10.1520/E1926
16. ASTM E1489-08, "Standard practice for computing ride number (RN) of roads from longitudinal profile measurements made by an inertial profile measuring device", American Standard, (2013). doi. 10.1520/E1489
17. Basu, S., Saha, P., "Regression models of highway traffic crashes: A review of recent research and future research needs", ***Procedia Engineering,*** Vol. 187, (2017), 59-66. doi: 10.1016/j.proeng.2017.04.350
18. Wood, A.G., Mountain, L.J., Connors, R.D., Maher, M.J., Ropkins, K., "Updating outdated predictive accident models", ***Accident Analysis and Prevention,*** Vol. 55, (2013), 54-66. doi: 10.1016/j.aap.2013.02.028
19. Ye, Z., Zhang, Y., Lord, D., "Goodness-of-fit testing for accident models with low means", ***Accident Analysis and Prevention,*** Vol. 61, (2013), 78-86. doi: 10.1016/j.aap.2012.11.007
20. Hosseinpour, M., Yahaya, A.S., Sadullah, A.F., Ismail, N., Ghadiri, S.M.R., "Evaluating the effects of road geometry, environment, and traffic volume on rollover crashes", ***Transport,*** Vol. 31, No. 2, (2016), 221-232. doi: 10.3846/16484142.2016.1193046
21. Sharifi, Y., Hosseinpour, M., "A predictive model based ANN for compressive strength assessment of the mortars containing metakaolin", ***Journal of Soft Computing in Civil Engineering,*** Vol. 4, No. 2, (2020), 1-11. doi: 10.22115/scce.2020.214444.1157
22. Naderpour, H., Rezazadeh, D., Fakharian, P., Rafiean, A.H., Kalantari, S.M., "A new proposed approach for moment capacity estimation of ferrocement members using group method of data handling", ***Engineering Science and Technology: An International Journal,*** Vol 23, No. 2, (2020), 382-391. doi: 10.1016/j.jestch.2019.05.013
23. Ghannadiasl, A., Rezaei Dolaghb, H. "Sensitivity analysis of vibration response of railway structures to velocity of moving load and various depth of elastic foundation", ***International Journal of Engineering, Transactions C: Aspects,*** Vol 33, No. 3, (2020), 401-409. doi: 10.5829/ije.2020.33.03c.04
24. Ghazvinian, H., Karami, H., Farzin, S., Mousavi, S. "Effect of MDF-cover for water reservoir evaporation reduction, experimental, and soft computing approaches", ***Journal of Soft Computing in Civil Engineering,*** Vol 4, No. 1, (2020), 98-110. doi: 10.22115/scce.2020.213617.1156
25. Naderpour, H., Rafiean, A.H., Fakharian, P. "Compressive strength prediction of environmentally friendly concrete using artificial neural networks", ***Journal of Building Engineering,*** Vol 16, (2018), 213-219. doi: 10.1016/j.jobe.2018.01.007
26. Ghasemi, S., Bahrami, H., Akbari, M. "Classification of seismic vulnerability based on machine learning techniques for RC frames", ***Journal of Soft Computing in Civil Engineering,*** Vol. 4, No. 2 (2020), 13-21. doi: 10.22115/scce.2020.223322.1186
27. Lashkenari, M.S., KhazaiePoul, A., Ghasemi, S., Ghorbani, M., "Adaptive neuro-fuzzy inference system prediction of Zn metal ions adsorption by γ-Fe2o3/Polyrhodanine nanocomposite in a fixed bed column", ***International Journal of Engineering, Transactions A: Basics,*** Vol. 31, No. 10, (2015), 1617-1623. doi: 10.5829/ije.2018.31.10a.02

Persian Abstract

چکیده

هدف از این مطالعه، ارزیابی تاثیر مشخصات فیزیکی رویه راه روی فراوانی تصادفات خروج از جاده در جاده‌های دوخطه مجزا می‌باشد. به همین منظور و برای ارائه یک مدل پیش‌بینی دقیق، نویسندگان سعی کردند مدل‌های خطی تعمیم‌یافته‌ای را ارائه دهند که شامل رگرسیون پواسون، رگرسیون دوجمله‌ای منفی و رگرسیون دوجمله‌ای منفی غیرخطی می‌شود. علاوه بر پارامترهای درمعرض‌بودن، برخی متغیرهای توصیفی مربوط به خصوصیات فیزیکی رویه راه مانند شاخص وضعیت رویه راه، شاخص ناهمواری بین‌المللی راه و عدد سواری نیز در مدلسازی لحاظ شدند. برای مدلسازی، از فرآیند پیش‌رونده بهره گرفته شده است که در آن، متغیرها به ترتیب به مدل اولیه (هسته اصلی) افزوده می‌شوند. در فرآیند مدلسازی غیرخطی و در هر مرحله، ۳۹ فرم ساختاری کنترل شدند تا مشخص شود که مدل جدید آیا برازش بهتری نسبت به مدل اولیه یا مدل قبلی داشته است یا خیر. ابزارهای متعددی برای تخمین نیکویی برازش مدل مورد آزمون قرار گرفتند. همچنین، از ابزارهای دیگری نیز برای تخمین اعتبار بیرونی و ساختار خطای مدل‌ها بهره گرفته شده است. نتایج نشان دادند که در مدل‌های رگرسیون پواسونی و رگرسیون دوجمله منفی، ضرایب متغیرها معنی‌دار نبودند. یافته‌های مدل غیرخطی پیشنهادی تایید کردند که متغیر شاخص وضعیت روسازی به عنوان یک متغیر عینی، از انتظارات متخصصان تبعیت می‌کند (به عبارتی، وضعیت رویه بهتر با تصادفات خروج از جاده کمتر همبستگی دارد). نهایتاً، شایان ذکر است که متغیر ناهمواری در سطح معنی‌داری مفروض، معنی‌دار نبود و بنابراین، سهمی در تصادفات خروج از جاده ندارد. نتایج تاکید دارند که بهبود وضعیت روسازی منجر به کاهش احتمالی بیشتری در فراوانی تصادفات خروج از جاده می‌شود.