Network Screening on Low-Volume Roads using Risk Factors

Kazi Tahsin Huda and Ahmed Al-Kaisy *

Posted Date: 28 December 2023

doi: 10.20944/preprints202312.2174.v1

Keywords: low-volume roads; crash prediction; Empirical Bayes; network screening; classification and regression tree; risk factors

Original Article

Network Screening on Low-Volume Roads Using Risk Factors

Kazi Tahsin Huda [1] and Ahmed Al-Kaisy [2]*

[1] Department of Engineering Technology and Construction Management, University of North Carolina, Charlotte, United States of America; khuda@charlotte.edu
[2] Department of Civil Engineering, Montana State University, Bozeman, Montana, United States of; alkaisy@montana.edu
* Correspondence: alkaisy@montana.edu

Abstract: This paper proposes a new method for network screening on rural low-volume roads. These roads are important as they provide critical access to agricultural land and tourist attractions. Most low-volume roads belong to the lowest functional class (local rural roads) and thus are built to lower design standards. The conventional hot spot network screening techniques may not be appropriate for low-volume roads due to the sporadic nature of crashes occurring on these roads. Contrarily sophisticated network screening approaches require extensive roadway and traffic data that are often unavailable to local agencies for lack of resources, and/or lack of technical expertise. This research attempts to address these obstacles in low-volume roads network screening which aims to identify candidate sites for safety improvements. The research used an extensive low-volume road sample from the state of Oregon and the Empirical Bayes expected number of crashes in developing the proposed models for network screening. The proposed models do not require exact measurement of roadway geometric features as all geometric variables were classified into categories that are easy to compile by local agencies. Further, the method could be used with and without traffic data, without compromising the effectiveness of the network screening process.

Keywords: low-volume roads; crash prediction; Empirical Bayes; network screening; classification and regression tree; risk factors

1. Introduction

Low-volume roads (LVRs) constitute a large proportion of the rural roadway net-work providing critical access to remote rural areas and tourist attractions. These roads are characterized by low traffic volumes, lower functional classification, and many of them are built to lower standards (e.g., narrow lanes and shoulders, non-forgiving roadsides, etc.). Moreover, average travel speeds on these roads are generally high (like other rural roadways) which largely explains the higher crash fatality rates on these roads compared to those in urban areas [1]. In the United States as well as in most other developed countries, safety is managed on the highway network using highway safety improvement programs. These programs, which aim to reduce the number and severity of crashes on a road network, are ongoing programs funded by the government and are usually managed by state transportation agencies. A critical step in these programs is network screening, where candidate sites for safety improvement are identified and ranked according to priority throughout the network for further consideration and analysis. A common rule used by most safety improvement programs is to maximize safety benefits at the selected sites to ensure optimum use of safety funds.

Historically, the most common approach to network screening is the use of crash history, i.e., crash frequencies, severities, and/or rates, in identifying sites for safety improvements. While this approach may work reasonably well for networks with high traffic exposure (e.g., urban areas and major rural routes), it may prove unsuitable for low-volume roads. Specifically, due to low traffic volumes, only a limited number of crashes occur during the analysis period, which tend to be scattered sporadically throughout the network. Therefore, it is unlikely that a site would consistently show a significant number of crashes to be ranked high on a ranking list using crash frequencies. On

 doi:10.20944/preprints202312.2174.v1

the other hand, many LVR sites with very low crash frequencies may still rank high when the crash and/or fatality rates are used in screening the network. Sites selected using the latter approach may not necessarily result in maximizing safety benefits at the network level. Therefore, a more sophisticated network screening approach, which considers factors other than crash history, such as roadway and traffic characteristics, is more effective in screening LVR networks. However, these methods (e.g., Empirical Bayes) often require detailed roadway, traffic, and crash data and use advanced statistical tools, which may present implementation challenges. Specifically, many LVRs are owned and operated by local agencies (townships, counties, tribal governments, etc.) that usually lack adequate resources and technical staff to adopt such sophisticated methods. Therefore, an ideal network screening method for LVRs would require little resources while being effective in identifying sites associated with high potential of crash reduction.

In this study, LVRs are defined as rural roads with Annual Average Daily Traffic equal to or less than 1000 vehicles per day (vpd). This definition is consistent with that used in a few other studies found in the literature [2–4].

2. Background

Network screening is a well-researched area, and several network screening methods were developed over time. However, only limited information is available about network screening for rural two-lane roadways, and more so for rural low-volume roads. Only a couple of methods found implementation in practice, while a few other methods were proposed in the literature without necessarily being implemented by highway agencies.

2.1. Network screening methods implemented in practice

Expectedly, methods that are reported to be used by highway agencies on rural lo-cal roads are generally simple and require a minimal amount of information for implementation. One such method utilizes the Federal Highway Administration (FHWA) Systemic Project Selection tool [5]. Risk factors are identified based on the relationship between roadway variables and crash occurrence. Locations having one or more of these risk factors are scored with "1" or an asterisk. After reviewing the locations, the risk factors are reassessed for their usefulness in the identification of safety improvement locations in the whole network. Any risk factor that is present in every location of the network is discarded. Finally, the locations are ranked based on the presence of risk factors. Sites with a higher number of risk factors indicate higher crash risk and therefore have higher priority. A few states and counties have reported the use of this tool. The “star” approach [6,7] in ranking sites for safety improvement consideration follows the same concept. If a risk factor of interest is present at a site, then a star is assigned. The higher the number of stars, the higher the risk at a site and the higher priority the site receives.

3.2. Methods proposed in literature

Several studies proposed network screening methods on rural roads with no re-ported implementation or adoption by highway agencies. Most of these methods use the relationship between risk factors or surrogate safety measures and crash numbers to predict the future number of crashes or level of risk throughout the network. Traffic or roadway characteristics that have a strong correlation with crash occurrence are known as risk factors and may be used as surrogate safety measures. A study [8] for rural roads in Wyoming has developed crash prediction models using the Negative Binomial Regression (NBR) and Poisson regression method. The models used crash rate, traffic volume, and speed and found that the high crash rates have a strong correlation with high volume and high speeds. The study also found that the NBR models had a better fit of the data. Another study [9] developed a full Bayes multivariate crash prediction model for rural two-lane roads of Pennsylvania. This study used traffic volume and segment length in pre-dicting crash frequencies. Al-Kaisy et al. [10] developed a risk index to identify at-risk locations for state-owned LVRs in Oregon. This method used three elements namely, geometric features, crash history, and

 doi:10.20944/preprints202312.2174.v1

traffic exposure. Suitable weights were assigned to each element and then the elements along with their weights were used to calculate a risk index for a site. The same author also proposed a scoring-based network screening approach [11,12] using relationships between crash occurrences and risk factors, for Montana. A study from New Zealand [13] have used speed to identify at-risk horizontal curves on rural highways. The study calculated the operating speeds on curves using an Austroads (Australian transportation agency) operating speed prediction model. Then the operating speeds were compared with the radius of the horizontal curves to assess the design limitation of the curves.

The more sophisticated Empirical Bayes (EB) method [14] was also reported as being evaluated on rural local roads. That study [15] used simulated data for low volume two-lane roads to measure the performance of different screening methods. The study concluded that the EB method outperformed other methods considered in this investigation. Superior performance of the EB method for LVRs was also reported by a US-based study [16] using empirical crash data. Another European study [17] used the difference between observed and predicted crashes over per unit length to find the accident potential for a site. Sites with positive accident potentials were considered as at-risk sites. Lastly, an Ethiopian study [18] used the difference between the EB expected crash numbers and the predicted crash numbers to identify and rank at-risk sites on a rural highway.

3. Motivation

The sporadic nature of the crash occurrences on LVRs makes the conventional network screening methods using crash history unsuitable for use on rural local road networks. Further, the couple of network screening methods that have been reported in the literature of being implemented by highway agencies are simplistic and may not be adequately effective in identifying sites deserving more consideration of safety treatments. Lastly, the more sophisticated network screening methods (e.g., EB method) usually require extensive data that is typically unavailable or inaccessible for rural low-volume roads. Therefore, the current study aims to develop a relatively robust method for network screening on rural LVRs that is practical in that it requires a minimal amount of information and technical expertise for implementation.

4. Study Data

State-owned LVR data from Oregon was used for developing this new method. The study sample consisted of around 850 miles of low volume roads from different regions in the state. The Oregon Department of Transportation (ODOT) online databases and video logs were used to identify and compile traffic, geometric and crash data. Data was collected for roadway segments that are 0.05 mile in length. This resolution was chosen so that roadside information could be extracted accurately. No major intersections were included in the dataset. Intersections along with one 0.05-mile segment before the intersection and another 0.05-mile segment after the intersection were skipped. All the segments were for two-lane paved rural roads with a posted speed limit of 55 mph.

Traffic volume and roadway data were collected for each of the 0.05-mile segments. Roadway data included information about lane width, shoulder width, horizontal curves, spiral curves, vertical grade, driveway density, side slope, and roadside fixed objects. The Oregon DOT video logs were used to collect driveway density, side slope rating, and fixed object rating data. Fixed objects and side slope data were collected categorically as it was not possible to collect their exact in-formation from video logs. Crash data from 2004 to 2013 were also collected for each of the 0.05-mile segments.

5. Methodology

As stated previously, the proposed new method should satisfy two requirements: 1) the method is effective in identifying sites that are more deserving of safety treatments, and 2) the method requires little information and technical skills that are accessible to local agencies. To satisfy these requirements, the method developed in this research incorporated the following two features:

- The EB expected number of crashes was selected as a basis for network screening in the proposed method. This is to ensure effective network screening given the favorable performance of the EB

 doi:10.20944/preprints202312.2174.v1

method for LVRs [15,16]. In addition, multiple studies [19–21] have also reported favorable performance of the EB method over other methods used for network screening.

- The proposed method employs classified variables that can easily be compiled by local agencies using staff with limited technical skills.

5.1. Overview of the EB Method

The EB method employs the predicted number of crashes and the observed number of crashes in estimating the expected number of crashes at a specific site. This estimation is developed using Eq. (1):

$$N_{expected} = w * N_{predicted} + (1 - w) * N_{observed} \quad (1)$$

This equation assigns different weights (contributions) to the predicted and observed numbers of crashes in estimating the expected number of crashes. The predicted number of crashes is found using a predictive mathematical model, also known as safety performance function (SPF). While many SPFs are proposed in the literature for various types of highway facilities, this study used the SPF proposed by the Highway Safety Manual (HSM) for rural two-lane highways, that is calibrated for the state of Oregon. The predictive HSM method predicts the number of crashes at a specific site in a two-step process. First, the method utilizes the SPF mathematical model in predicting the number of crashes under base conditions (e.g., for two-lane highway segments, inputs are the AADT and segment length). Then, the method accounts for roadway and roadside characteristics (sometimes referred to as risk factors) that deviate from the base conditions using crash modification factors (CMFs). The observed number of crashes is the number of reported crashes at the specific site for the same analysis period.

5.2. Classification of roadway variables

The more sophisticated network screening methods use exact values for various roadway characteristics (sometimes referred to as risk factors) in applying the screening procedure. While the use of exact values may improve the accuracy of the screening process, it requires access to extensive databases or on-site detailed measurements which are typically beyond the resources available for small local agencies. Hence, implementing such sophisticated methods might not be feasible for most local agencies. One possible solution to this problem is to use a classified format instead of the exact values for roadway characteristics that can easily be acquired by local agencies. This approach would classify any roadway variable or risk factor (e.g., grade) into a limited number of categories for use in the prospective screening procedures. Therefore, the issue that had to be addressed by this research is how to set the cut-off values between the classes or categories of classified variables. To avoid any subjectivity in the process of setting cut-off values, it was decided to use the classification and regression tree (CART) data mining technique. The CART analysis classifies an independent variable based on the effect it has on a dependent variable. It can be used to determine the relative importance of different variables for identifying homogeneous groups within the data set. The analysis recursively partitions observations in a matched data set, consisting of a categorical or continuous dependent (response) variable and one or more independent (explanatory) variables, into progressively smaller groups [22]. Each partition is a binary split. During each recursion, splits for each independent (explanatory) variable are examined and the split that maximizes the homogeneity of the two resulting groups with respect to the dependent (re-sponse) variable is chosen. In this study, the dependent variable was the expected crash numbers (from the HSM EB method), and the independent variables were the roadway characteristics or risk factors. When the CART analysis failed to provide suitable cut-off values for classifying the risk factors, then the identified trends of crash occurrences with different risk factors were used for determining suitable cut-off values for classification.

The main purpose of classifying the data was to ease the efforts of data collection. However, collecting exact data for a couple of variables is comparatively easier than others and therefore these variables were intentionally not classified. The first variable is the driveway density. The driveway

 doi:10.20944/preprints202312.2174.v1

density for a segment can easily be obtained from updated maps (for example Google Earth, Google Maps) of the segments. The second variable is traffic volume as AADT has increasingly been used by highway agencies for many purposes and as such highway agencies usually measure or estimate AADT for most of their roadway networks. Further, it was deemed impractical to classify volume data due to the difficulty in selecting traffic volume class without prior knowledge about traffic volume. The side slope and roadside fixed object data were excluded from the CART analysis because these variables were already collected in a classified format using the roadway video logs accessible through the Oregon DOT website.

5.3. Development of Regression Model

Upon classifying roadway (or risk) factors, models were developed to predict the level of risk (or safety) of roadway segments using the study sample. Two models were developed for this purpose. The response (dependent) variable in both the models was the EB expected number of crashes, which is a function of the HSM predicted number of crashes and observed number of crashes. The explanatory (in-dependent) variables included classified roadway factors and traffic exposure (AADT) for the first model and only classified roadway factors for the second model. This was done so that the network screening could be performed with or without the use of AADT data, as many local agencies lack resources to acquire traffic volume data on their respective networks.

6. Study Results

This section presents the results of the different analyses that were carried out to develop the new network screening method. The proposed method is in the form of regression models that use roadway characteristics and traffic exposure in assessing the level of risk or safety for roadway segments.

6.1. Level of Risk

The proposed models use the EB expected number of crashes as an indicator of the level of risk associated with any roadway segment that is part of the network. This measure incorporates roadway physical characteristics, traffic exposure, and observed crashes in assessing the level of risk, and thus is more comprehensive than using either the HSM predicted number of crashes alone or the observed number of crashes alone. This measure serves as the response or dependent variable in the proposed models.

6.2. Model Variables

This section discusses the explanatory or independent variables used in developing network screening predictive models. Table 1 shows the explanatory variables for the proposed models. Some of the cut-offs were rounded to the nearest whole number for practical purposes.

Table 1. Explanatory Variables of the proposed models.

Risk Factors	Approximate Ranges of Variables	Categories	Terms	Statistics/Frequencies
Lane Width (LW)	LW < 11 ft	1	Narrower	Frequency: 1,964
	LW ≥ 11 ft	2	Wider	Frequency: 14,550
Shoulder Width (SW)	SW < 2 ft	1	Narrower	Frequency: 5,508
	SW ≥ 2 ft	2	Wider	Frequency: 11,006
Degree of Horizontal Curvature (DC)	DC = 0°	0	Straight	Frequency: 10,038
	DC < 9°	1	Mild	Frequency: 5,322
	9° ≤ DC < 28°	2	Moderate	Frequency: 615
	DC ≥ 28°	3	Sharp	Frequency: 539
Grade (G)	G < 4 %	0	Mild	Frequency: 12,593

 doi:10.20944/preprints202312.2174.v1

	G ≥ 4 %	1	Steep	Frequency: 3,921
Driveway Density (DD) (driveways per mile)	Exact Number			Minimum: 0 First Quartile: 0 Median: 0 Mean: 4.479 Third Quartile: 0 Maximum: 100
Side Slope (SS)	Steep	1	Steep	Frequency: 3,363
	Moderate	2	Moderate	Frequency: 10,871
	Flat	3	Flat	Frequency: 2,280
Fixed Objects (FO)	Many	1	Many	Frequency: 10,700
	Some	2	Some	Frequency: 4,458
	Few	3	Few	Frequency: 1,356
Volume (V)	Exact Volume			Minimum: 60 First Quartile: 260 Median: 430 Mean: 496.5 Third Quartile: 620 Maximum: 2500
Dependent Variable: EB Expected Number of Crashes				Minimum: 0.0083 First Quartile: 0.0376 Median: 0.0561 Mean: 0.0994 Third Quartile: 0.0821 Maximum: 5.1193

As shown in this table, traffic volume (AADT) and driveway density use exact values. The AADT is usually measured or estimated for highway segments as it is used by highway agencies for different purposes. Classifying traffic volume was deemed inappropriate for the difficulty in selecting the appropriate class without a prior knowledge of traffic volume. It was also decided to use the exact number of driveways per mile as this information is easily accessible using aerial photography available through web mapping applications such as Google Maps. All other variables shown in this table are in a classified format. Roadside variables, side slope, and fixed objects, were already compiled in classified format when the information was extracted from video logs, and therefore no classification was needed. The CART analysis discussed earlier was used to classify the other variables, namely: lane width, shoulder width, degree of curvature, and grade. The open-source statistical software R was used for running the CART analyses.

The CART results for the degree of curvature are shown in Figure 1. It can be seen from the output that 86 percent of the data comprises segments that have a degree of curvature less than 8.9 degrees. This 86 percent corresponds to the lowest average expected number of crashes of 0.067. The next branch shows that 10 percent of the data comprises segments with degrees of curvature equal to or greater than 8.9 degrees but less than 28 degrees. This group corresponds to the average expected number of crashes of 0.14. The third branch of the output shows segments with degrees of curvature equal to or greater than 28 degrees. This group constitutes 3 percent of the data and corresponds to an average expected number of crashes of 0.86. Based on this output, four classes were selected for horizontal curvature: straight segments with degrees of curvature of zero, mild curves with degree of curvature greater than 0 but less than 8.9 degrees, moderate curves with degrees of curvature greater than or equal to 8.9 degrees but less than 28 degrees, and sharp curves with degrees of curvature greater than or equal to 28 degrees. The CART result regarding shoulder width is shown in Figure 2. The output shows that the segments with shoulder widths equal to or greater than 1.8 feet comprise 63 percent of the dataset and correspond to an average expected crash number of 0.068. The remaining 37 percent of the dataset (for shoulder widths of less than 1.8 feet) corresponds to a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2023 doi:10.20944/preprints202312.2174.v1

higher average expected crash number of 0.099. Therefore, 1.8 feet was chosen as the cut-off for categorizing the shoulder widths.

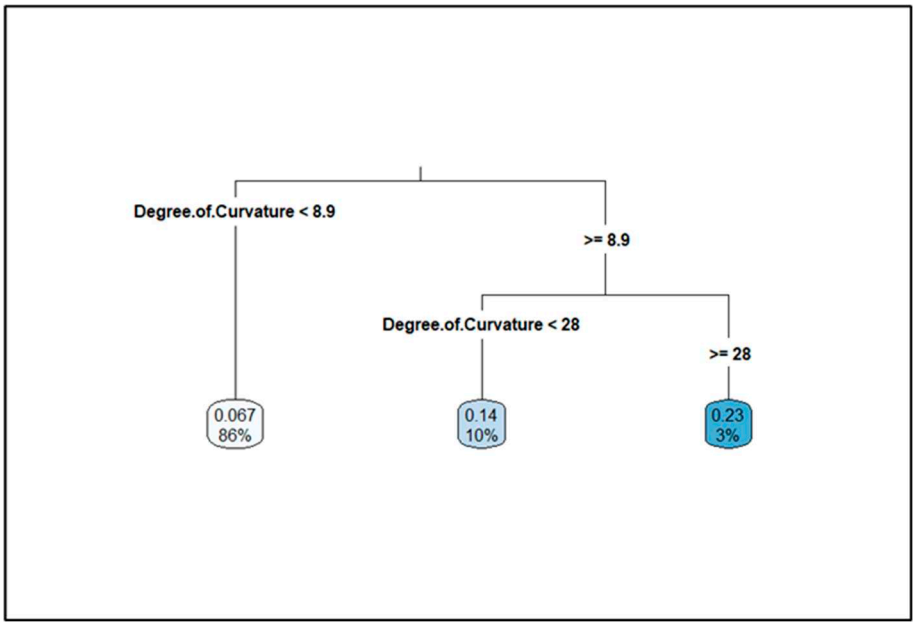

Figure 1. CART output for degree of horizontal curvature.

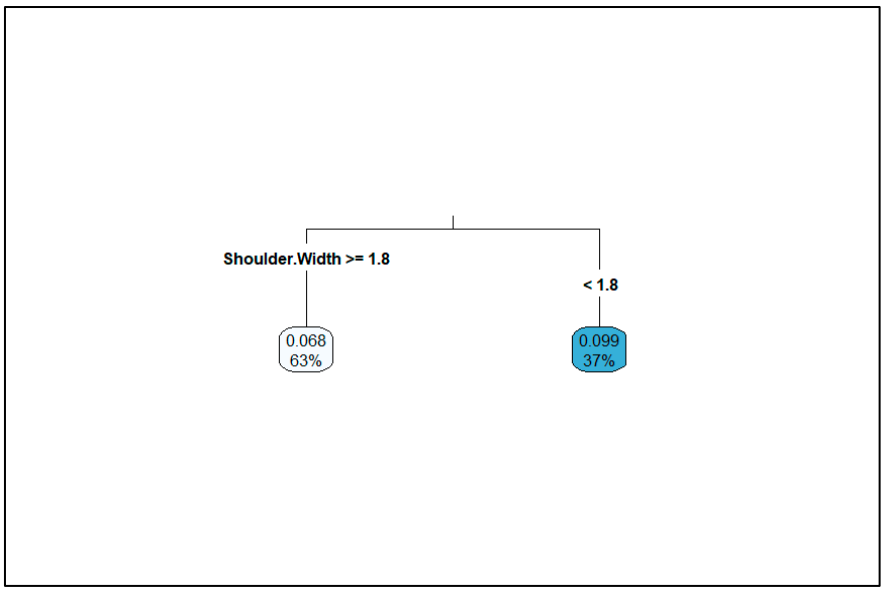

Figure 2. CART output for shoulder width.

The CART results regarding lane width and grade were trivial as the analysis lumped all the data set in one class, thus providing no utility in classifying the variables. For these roadway characteristics, descriptive statistics, and engineering judgment were used to help in classifying the variables. For lane width, two classes were created: wider lanes with lane width of 11 feet or wider, and narrower lanes with any lane width less than 11 feet. Figure 3 shows the observed crashes per mile for the two lane-width classes used in the proposed models. For grade, the observed crashes per mile were plotted as a function of grade as shown in the bar chart in Figure 4. It is clear from this figure that the observed crashes notably increase when grade exceeds 4%. Therefore, two classes were used for classifying this variable: mild grade with grade percentage less than 4%, and steep grade with grade percentage equal or greater than 4%.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2023 doi:10.20944/preprints202312.2174.v1

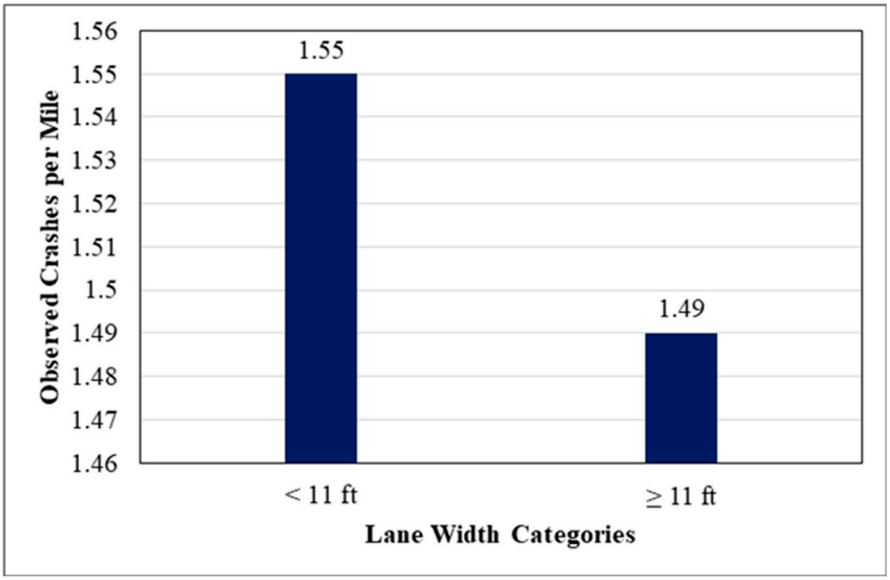

Figure 3. Crash rates for lane width categories.

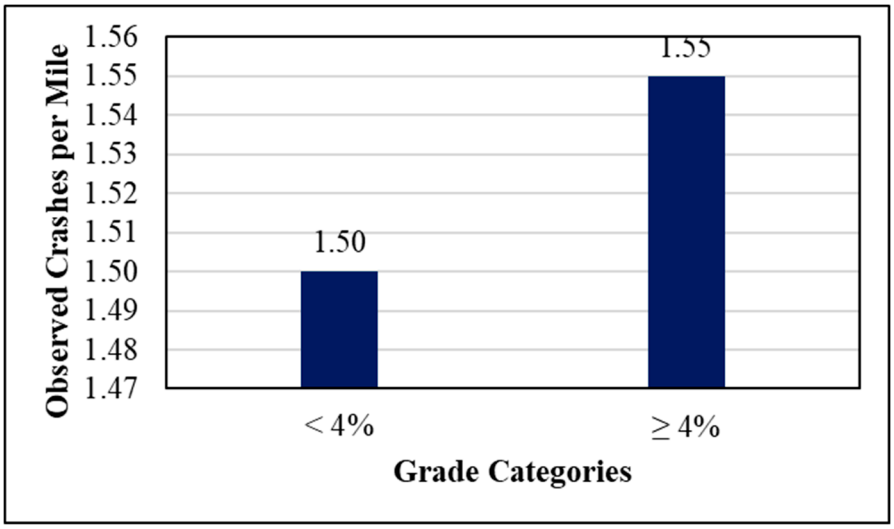

Figure 4. Crash rates for vertical grade categories.

6.3. Proposed Models

Multivariate ordinary least square linear regression analysis was used in developing the proposed models for low-volume road network screening. First, the data was split randomly into two parts. The first part consisting of 80 percent of the data (about 680 miles) was used to develop or train the model while the second part with the remaining 20 percent (about 170 miles) was used for testing the accuracy of the model.

The open-source statistical software R was used for analysis. Two models were developed: one with traffic exposure and another without traffic exposure information. The first model, which incorporates traffic volume as an explanatory variable, is shown in Eq. 2.

$$\ln(Exp) = -0.88\,LW - 0.34\,SW + 0.016\,DD + 0.001\,V + 0.24\,DC - 0.31\,SS - 0.21\,FO \quad (2)$$

Where, Exp = EB expected number of crashes, LW = Lane width, SW = Shoulder width, DD = Driveway density, V = Exposure (AADT), DC = Degree of curvature, SS = Side slope and FO = Fixed objects.

This model has a coefficient of determination (adjusted R-squared) of 0.9150. This indicates that the model can explain about 92 percent of the variability of the EB expected number of crashes, which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2023 doi:10.20944/preprints202312.2174.v1

is relatively high given the classified format used for most of the variables in this model. All variable coefficients were found significant at the 95% confidence level. All variable coefficients exhibited logical relationships with the explanatory variable, i.e., the expected number of crashes, given the fundamentals of safety science. Important information about the regression output for the first model is shown in Table 2.

Table 2. Regression output for the first model.

Variables	Coefficients	p-value	Significance at 95 percent confidence level
Lane Width (LW)	-0.88	$< 2 \times 10^{-6}$	✓
Shoulder Width (SW)	-0.34	$< 2 \times 10^{-6}$	✓
Driveway Density (DD)	0.016	$< 2 \times 10^{-6}$	✓
Volume (V)	0.001	$< 2 \times 10^{-6}$	✓
Degree of Curvature (DC)	0.24	$< 2 \times 10^{-6}$	✓
Side Slope (SS)	-0.31	$< 2 \times 10^{-6}$	✓
Fixed Objects (FO)	-0.21	$< 2 \times 10^{-6}$	✓
Adjusted R-squared = 0.915			

To account for situations where traffic volume data is not available to local agencies, regression analysis was conducted using the same data set while excluding the AADT from the analysis. The model is shown in Eq. 3 below.

$$\ln(Exp) = -0.53\,LW - 0.46\,SW + 0.02\,DD + 0.27\,DC - 0.28\,SS - 0.25\,FO \quad (3)$$

Where, Exp = EB expected number of crashes, LW = Lane width, SW = Shoulder width, DD = Driveway density, DC = Degree of curvature, SS = Side slope and FO = Fixed objects.

This model is very similar to the previous model in format, i.e., the logarithmic format for the dependent variable and the linear format for all independent variables. The coefficient of determination for this model is not much different from that of the previous model (0.9057). This indicates that dropping the AADT from the second model did not much affect the predictive ability of the model as the R-squared value did not change but slightly. All the coefficients for independent variables were found significant at the 95% confidence level. Again, all coefficients exhibit a logical relationship between the response variable and explanatory variables. Important information about the regression output for the no-volume model is shown in Table 3.

Table 3. Regression output for the no-volume model.

Variables	Coefficients	p-value	Significance at 95 percent confidence level
Lane Width (LW)	-0.53	$< 2 \times 10^{-6}$	✓
Shoulder Width (SW)	-0.46	$< 2 \times 10^{-6}$	✓
Driveway Density (DD)	0.02	$< 2 \times 10^{-6}$	✓
Degree of Curvature (DC)	0.27	$< 2 \times 10^{-6}$	✓
Side Slope (SS)	-0.28	$< 2 \times 10^{-6}$	✓
Fixed Objects (FO)	-0.25	$< 2 \times 10^{-6}$	✓
Adjusted R-squared = 0.9057			

6.4. Validation Results

The proposed models were validated using a different data set to make sure that models perform reasonably well on other low-volume roads that are part of the network. The testing (validation) data contained 20 percent (about 170 miles) of the original dataset and was selected randomly. The expected crash numbers for the testing data were predicted using both the proposed models. To examine the predictive ability of the models, the mean squared error (MSE) and the root mean

 doi:10.20944/preprints202312.2174.v1

squared error (RMSE) were used. The MSE is the mean of squared residuals, and the residuals are the difference between actual and predicted expected crashes. The RMSE is calculated as per Equation 4.

$$RMSE = \sqrt{\frac{\sum(e - \hat{e})^2}{n}} \quad (4)$$

Where, e = actual expected crashes, $\hat{e}$ = predicted expected crashes, and n = number of data points.

For a more comprehensive validation, the mean, and the total of the actual and the predicted expected crashes were also calculated. All these parameters were also calculated for the training dataset and are provided in Table 4.

Table 3. Regression output for the no-volume model.

	Training Data		Testing Data	
	Mean	**Total**	**Mean**	**Total**
Actual Expected Crashes	0.0875	1219	0.0827	288
Model Estimate	0.0859	1198	0.0863	301
No Vol. Model Estimate	0.0759	1059	0.0764	266
MSE	0.0372		0.0270	
MSE (no volume)	0.0324		0.0255	
RMSE	0.1928		0.1642	
RMSE (no volume)	0.1800		0.1596	

Considering the mean and the total expected crashes shown in Table 4, it is observed for the first model that the difference between the actual and predicted expected crashes are less than 5 percent for both the training and the testing datasets (1.8 and 1.7 percent for the training data; 4.3 and 4.5 percent for the testing data). These differences are considered particularly small given that all geometric variables are classified into a small number of categories. For the second model, the difference is about 13 percent (13.25 and 13.12 percent for the training; 7.62 and 7.64 percent for the testing datasets), which can be expected for a model that mostly uses categorical variables and lacks traffic exposure information. The small percentage differences between the mean and the total "actual" and "predicted" expected number of crashes suggest accurate prediction by the two models. Further, both the MSEs and the RMSEs are relatively small values which support the strong predictive ability of the models and are consistent with the high R-squared values discussed earlier. Furthermore, the testing data MSEs are observed to be smaller than those of the training data. Usually, higher testing data MSEs indicate an overfitted model. Comparing the RMSEs with the mean, however, shows that the variability of the predicted expected number of crashes is quite large. Again, much of this variability can be attributed to the use of categorical risk factors for predicting the expected number of crashes.

7. Summary and Concluding Remarks

This paper presented a new method for network screening for local and low-volume roads. The method involves the use of regression models for predicting the expected number of crashes using roadway characteristics with or without the use of traffic exposure data. A key advantage of this method is that it does not require exact information or access to extensive databases as it employs classified roadway variables that can easily be compiled by staff of local agencies. Simple and basic tools like Google maps can provide useful information for use as input to the proposed models. Another advantage of the proposed method is that it does not require extensive technical expertise for implementation. The prediction models consist of simple mathematical equations that can easily be understood and implemented. This would allow smaller local agencies (counties, townships, etc.) with limited technical expertise and lack of access to extensive databases to implement these models.

One potential limitation in the proposed method is that the regression models developed may be state- or region-specific, and therefore the use of these models outside of Oregon would ideally require the use of local data in developing state-specific or region-specific models using the same approach followed in this research.

The two proposed models have R-squared values of greater than 0.90 which suggests a very strong predictive ability and accurate predictions. The proposed models were also validated using a separate dataset and the results were deemed favorable overall. Further, the high R-squared values for the two models and the closeness of their numerical values (0.915 and 0.906) suggest that the contribution of crash history and traffic exposure in the estimation of the HSM EB expected number of crashes is not tangible on low-volume roads. This is consistent with the findings from a recent study [16] which found that HSM EB method relies more on risk factors to estimate the EB expected number of crashes under low-volume conditions.

Acknowledgement: The authors would like to thank the Western Transportation Institute's Small Ur-ban, Rural, and Tribal Center on Mobility (SURTCOM) for their financial assistance to this research project.

Conflict of Interest: The authors declare no conflicts of interest regarding the publication of this paper.

References

1. National Highway Traffic Safety Administration (NHTSA). *Rural Urban Comparison of Traffic Fatalities*, 2018, Washington D.C.
2. Ewan, L., Al-Kaisy, A. and Hossain, F. Safety Effects of Road Geometry and Roadside Features on Low-Volume Roads in Oregon. *Transportation Research Record: Journal of the Transportation Research Board*, 2018, 2580, 47–55.
3. Federal Highway Administration. (FHWA). *Manual on Uniform Traffic Control Devices*, 2009, Washington, D.C. Accessed at http://mutcd.fhwa.dot.gov/. Accessed March 28, 2019.
4. Gross, F., Eccles, K., and Nabors, D. Low-Volume Roads and Road Safety Audits – Lessons Learned. *Transportation Research Record: Journal of the Transportation Research Board*, 2011, 2213, 37-45.
5. Federal Highway Administration. (FHWA). *Systemic Safety Project Selection Tool*, 2013, Washington, D.C. http://safety.fhwa.dot.gov/systemic/. Accessed March 25, 2019.
6. Minnesota Department of Transportation (MnDOT). *Minnesota's Systemic Approach Integrates Safety Performance into Investment Decisions for Local Roads*, 2014. https://www.fhwa.dot.gov/innovation/everydaycounts/edc_4/pdf/case_study_mn_oct2014.pdf. Accessed March 25, 2018.
7. North Dakota Department of Transportation. *Local Road Safety Program*, N.d. https://www.dot.nd.gov/divisions/safety/docs/LSRP/VIEW_WesternRegion_Agency_BurkeCounty.pdf. Accessed March 25, 2018.
8. Zhong. C., Sisiopiku, V. P., Ksaibati, K. and Zhong, T. Crash Prediction on Rural Roads. *3rd International Conference on Road Safety and Simulation*, 2011, September 14-16, Purdue University, Indianapolis.
9. Aguero-Valverdea, J., Wub, K. K., and Donnell, E. T. A multivariate spatial crash frequency model for identifying sites with promise based on crash types. *Accident Analysis and Prevention*, 2016, 87, 8–16.
10. Al-Kaisy, A., Veneziano, D., Ewan, L. and Hossain, F. *Risk Factors Associated with High Potential for Serious Crashes: Final Report*, 2015, Report No. FHWA-OR-RD-16-05.
11. Al-Kaisy, A. and Huda, K.T. *Developing a Methodology for Implementing Safety Improvements on Low-Volume Roads in Montana*, 2021, Final Report. Report No. FHWA/MT-21-004/9679-699.
12. Al-Kaisy, A. and Raza, S. A Novel Network Screening Methodology for Rural Low-Volume Roads. *Journal of Transportation Technologies*, 2023, 13, 4, 599-614.
13. Harris, D., Durdin, P., Brodieb, C., Tate, F. and Gardener, R. A road safety risk prediction methodology for low volume rural roads. *Australasian Road Safety Conference, Gold Coast*, 2015, Queensland, 14-16 October .
14. American Association of State Highway and Transportation Officials (AASHTO). *Highway Safety Manual*, 2010, First Edition. Washington D.C.
15. Cafiso, S., Di Silvestro, G., Persaud, B. and Ara Begum, M. Revisiting variability of dispersion parameter of safety performance for two-lane rural roads. *Transportation Research Record*, 2010, 2148(1), 38-46.

16. Al-Kaisy, A. and Huda, K. T. Empirical Bayes application on low-volume roads: Oregon case study. *Journal of Safety Research*, 2022, 80, 226-234.
17. Šenk, P., Ambros, J., Pokorný, P. and Striegler, R. Use of Accident Prediction Models in Identifying Hazardous Road Locations. *Transaction on Transport Sciences*, 2012, 5(4), 223-232.
18. Abebe, M. T. and Belayneh, M. Z. Identifying and Ranking Dangerous Road Segments a Case of Hawassa-Shashemene-Bulbula Two-Lane Two-Way Rural Highway, Ethiopia. *Journal of Transportation Technologies*, 2018, 8, 151-174.
19. Elvik, R. The predictive validity of Empirical Bayes estimates of road safety. *Accident Analysis and Prevention*, 2008, 40(6), 1964-1969.
20. Montella, A. A comparative analysis of hotspot identification methods. *Accident Analysis and Prevention*, 2010, 42, 571-581.
21. Manepalli, U. R. R. and Bham, G. H. An evaluation of performance measures for hotspot identification. *Journal of Transportation Safety and Security*, 2016, 8(4), 327-345.
22. De'ath, G. and Fabricius, K. E. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 2000, 81(11), 3178-3192.