Improving Safety Performance Function Model Fit Using Exploratory Regression Techniques

Riana Tanzen[1], Eric Green, Reginald Souleyrette, Arnold Stromberg

Department of Civil Engineering, University of Kentucky, 161 Oliver H. Raymond Building, Lexington, KY 40506, USA, riana.tanzen@uky.edu (Author 1)
Kentucky Transportation Center/ Department of Civil Engineering, University of Kentucky, 140A Oliver H. Raymond Building, Lexington, KY 40506, USA, eric.green@uky.edu (Author 2)
Department of Civil Engineering, University of Kentucky, 161 Oliver H. Raymond Building, Lexington, KY 40506, USA, souleyrette@uky.edu (Author 3)
Department of Statistics, University of Kentucky, 305C Multidisciplinary Science Building, 725 Rose St, Lexington, KY 40506, USA, astro11@email.uky.edu (Author 4)

1. Introduction

Safety Performance Functions (SPFs) are the preferred safety analysis for Kentucky's network screening process. When developing state-specific SPFs, there is a tradeoff between the quality of the SPFs and the coverage of the roadway network. The quality of the model can be evaluated by a variety of goodness-of-fit measures. Coverage of the roadway network is important since any portion of the network not used in model development will likely require Adjustment Factors (AF). Unfortunately, AFs, are not available for all roadway and intersection configurations.

This paper explores the tradeoffs between model quality and network coverage. A variety of model forms are tested and compared to two extreme conditions: 1) a generic model with 100% network coverage and, 2) a specific model based on homogeneous base conditions (lane width, shoulder width, curvature) producing the best quality model. Models were developed using ranges of roadway geometrics to increase the roadway coverage from the specific model. For example, instead of specifying 9-feet lanes, the model may include segments with 9-to-13-feet lanes. Additionally, roadway geometrics were added to the model form as variables to increase network coverage. Instead of specifying as a base condition, lane width can be added to the model as an additional term so that all values of lane widths can be represented in the model. The idea here is that narrower lanes are likely to be correlated to more crashes in contrast to the generic model where lane width does not affect the model.

2. Methodology

2.1 Data Preparation

The Kentucky Transportation Cabinet (KYTC) maintains the roadway data for all state-maintained roads in the roadway centerline network and Highway Information System (HIS). For this study, the dataset for rural two-lane roads with several attributes including traffic volume, lane width, shoulder width, and horizontal curves was extracted. The roadways were segmented into homogenous sections where traffic volume and other geometric attributes remain constant.

For this study, five years (2013-2017) of crash data has been used that was collected from the Kentucky State Police (KSP) maintained database. The crashes (classified by severity using the KABCO scale[2]) were assigned to corresponding segments. Crashes of all severities were summarized into total crashes and had been linked to the homogenous segments of rural two-lane roads.

2.2 Development of Safety Performance Functions

SPFs are crash prediction models based on statistical regression modeling of historical crash data and they correlate crashes to segment length and traffic volume (See Equation 1). The methodology outlined by the *Highway Safety Manual* (HSM) recommends specifying "base conditions" for each SPF and may include roadway features like lane width, shoulder width, horizontal curves, etc (*5*). Crash

[1] Corresponding author. Tel.: +859-475-2151;
E-mail address: riana.tanzen@uky.edu

[2] K= fatal, A = incapacitating injury, B = non-incapacitating injury, C = possible injury, and O = no injury/property damage only (PDO)

Modification Factors (CMF) are used to account for any difference between the base condition and any segment's geometric attribute and adjust the crash prediction. In Kentucky, CMFs are referred to as Adjustment Factors (AF) when used for this purpose (*1*). The study used an open-source automation tool named "SPF-R[3]", a script written in Rstudio developed by the Kentucky Transportation Center to develop SPFs.

$$SPF\ Predicted\ Crashes = e^{a} * L * AADT^{b} * AF_1 * AF_2 * \ldots. \tag{1}$$

Where, L= Length of a segment (miles); AADT = Average Annual Daily Traffic; a = Regression parameter for intercept; b = Regression parameter for AADT; AF = Adjustment Factors.

2.2.1 Generic SPF

The "generic" SPF included segments with all the roadway features of interest (lane width, shoulder width, and horizontal curve classes) without specifying any base conditions. However, the generic model could show omitted variable bias as influential explanatory variables might have been excluded from the model. There are two ways to address omitted variable bias: filtering the database by specifying base conditions and/or including additional independent variable(s) in the model. Nonetheless, each method comes with its limitations. Although the application of filters is a convenient way, it requires adjustment factors to adjust crash prediction for the segments that may be different from the base conditions. Therefore, the unavailability of adjustment factors can limit the application of the SPF. On the other hand, including too many parameters in the regression model might lead to overfitting of SPFs and it may not be possible to include all the relevant explanatory variables that could have potential impacts on safety (*6*).

2.2.2 SPFs with Specific Base Conditions

SPFs are preferably developed using specified base conditions. In this study, multiple iterations were performed with various sets of base conditions, and the most reliable SPF was chosen based on goodness-of-fit measures i.e. CURE plots, modified R^2, CURE deviation percentage (CDP), maximum absolute CURE deviation (MACD), and inverse overdispersion[4]. This model will be referred to as the "specific" model. The base conditions used for this model are : Lane Width = 9 feet; Shoulder Width = 3 feet; Curve Class = A.

2.2.3 SPFs with Ranges of Base Conditions

Though SPFs can be improved by specifying base conditions, in absence of appropriate AFs, they cannot be used in the subsequent steps. In this study, a series of models have been developed using a range of values for the attribute. For each variable, the ranges were expanded around the particular value used for the "specific" model development. This ensures the inclusion of more segments used in the SPF and fewer segments requiring AFs. Similar to the "specific" model, several goodness-of-fit measures were used to evaluate the models. Table 1 summarizes the base conditions for the best five SPFs.

Table 1: Base conditions (ranges) used to develop SPFs

Models	*Base Conditions*		
	Lane Width	*Shoulder Width*	*Curve*
R1	9	0-3	A, B
R2	9	3-6	A, B
R3	9-13	3	A, B
R4	8-10	3	A, B
R5	9-12	3	A

2.2.4 SPFs with Additional Explanatory Variables

Another way of mitigating omitted variable bias is to include additional variables along with segment length and traffic volume in the model. For example, instead of filtering the dataset with a lane width of 9 feet, lane width could be included in the model as an explanatory variable. Therefore, AFs are no longer required for lane width and the sample size of the dataset is increased for model development. Table 2 summarizes the models developed including a variety of configurations of the

[3] http://github.com/irkgreen/SPF-R

[4] For modified R2 and inverse overdispersion, higher values are preferred. For CDP and MACD lower values are preferred.

variables where LW, SW and CUDEG are lane width, shoulder width, and degree of vertical curvature respectively.

Table 2: Additional variables and model forms of SPFs

Model	*Variable added*	*Model form*
V1	LW	$Y = L * e^{a} AADT^{b1} * e^{LW*b2}$
V2	SW	$Y = L * e^{a} AADT^{b1} * e^{SW*b2}$
V3	Roadway Width (LW+SW)	$Y = L * e^{a} AADT^{b1} * e^{(LW+SW)*b2}$
V4	LW, SW	$Y = L * e^{a} AADT^{b1} * e^{LW*b2} * e^{SW*b3}$
V5	LW, SW, LW*SW (Interaction term)	$Y = L * e^{a} AADT^{b1} * e^{LW*b2} * e^{SW*b3} * e^{LW*SW*b4}$
V6	CUDEG	$Y = L * e^{a} * AADT^{b1} * (2 * CUDEG)^{b_2} * (\frac{CUDEG}{5730 * L})^{b3}$
V7	LW, SW, CUDEG	$Y = L * e^{a} * AADT^{b1} * e^{LW*b2} * e^{SW*b3} * (2 * CUDEG)^{b_4} * (\frac{CUDEG}{5730*L})^{b5}$

2.3 Empirical Bayes Estimate

The Empirical Bayes (EB) method estimates the expected crashes using a mathematical combination of the observed and predicted crash frequencies. For each segment, Equations 2 and 3 were used to calculate the EB estimates for each model where $\boldsymbol{\theta}$ (Theta) is the inverse overdispersion parameter.

$$\textbf{EB Expected Crashes} = weight * SPF\ predicted\ crashes + (1 - weight) * observed\ crashes \quad (2)$$

$$\boldsymbol{weight} = \frac{1}{1 + \frac{\left(\frac{SPF\ Crashes}{Segment\ Length}\right)}{\theta}} \quad (3)$$

2.4 Cross-Validation

This study used the train-test split method for cross validation by splitting the dataset into two parts: the training set and the testing set. 75 percent of data was assigned to the training set using a random number generating function, and the rest was used as the testing set. The training dataset was used to develop the "generic", "specific", and two best models (based on goodness of fit measures) from Table 1 and Table 2. Root mean square error (RMSE) is chosen as the validation metric and was calculated for the testing dataset in two ways:

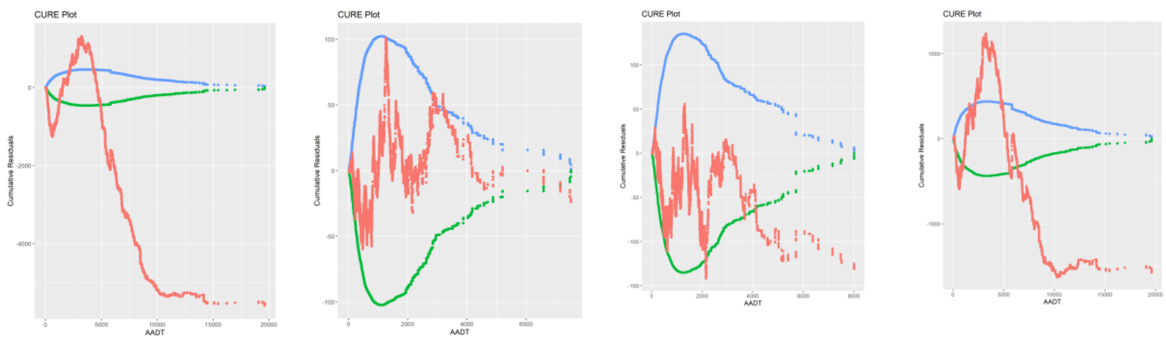
- **RMSE$_1$:** RMSE$_1$ calculates the errors by comparing the predicted crashes to the observed crashes. However, the randomness of crash data and its regression-to-the-mean bias might affect the RMSE.
- **RMSE$_2$:** RMSE$_2$ estimates the errors by comparing the predicted crashes to the EB estimate which is a function of the observed crashes and accounts for the regression-to-the-mean bias by pulling the crash counts to the mean (*14*).

3 Results and Discussions

3.1 Comparison of Goodness-of-Fit Measures

In this study, the CURE plots were used to visually evaluate the model fit and detect omitted variable bias. From visual inspection, the CURE plot from model R2 seemed to be the most significant among the models developed using ranges of base conditions (R1-R5). Although the CURE plot from model V7 had shown significant areas outside the bounds of the plot, it is the best among the seven models developed by adding model parameters (V1-V7). Figure 1 shows the CURE plots for the

"Generic" model, "Specific model", models R2 and V7. Other GOF measures including modified R^2, CDP, MACD, and Theta are summarized in Table 3.



(i) (ii) (iii) (iv)

Figure 1: CURE plots of (i)"Generic" model; (ii) "Specific" model; (iii) Model R2; (iv) Model V7

Table 3: Goodness-of-fit compared for all SPFs

Models	Modified R^2	CDP	MACD	Theta
Generic	**0.26**	**86.0**	**5582.9**	**1.163**
Specific	**0.65**	**0.6**	**101.0**	**2.230**
R1	0.59	4.5	112.9	1.950
R2	**0.60**	**2.0**	**141.9**	**2.094**
R3	0.35	6.0	635.8	1.607
R4	0.55	12.0	254.7	1.873
R5	0.52	37.8	297.1	1.800
V1	0.30	76.6	4583.8	1.206
V2	0.35	64.7	3148.8	1.279
V3	0.35	65.1	3177.5	1.278
V4	0.35	64.9	3136.5	1.281
V5	0.35	65.6	3128.4	1.284
V6	0.29	90.0	3687.5	1.239
V7	**0.37**	**63.4**	**1628.1**	**1.358**

The CURE plot resulting from the "generic" SPF (Figure 1 (i)) shows that 86% of the data have breached the bands. The model also has a low modified R^2 value and theta and high MACD (see Table 3) indicating an unreliable model. On the other hand, the CURE plot of the "Specific" model has the best CURE plot (see Figure 1) and the model possesses the highest values of modified R^2 and theta and the lowest values of CDP and MACD among all 14 models.

From Table 3 it is seen models R1-R5 also have significant GOF measures indicating the reliability of the models. Among them, model R2 developed performed the best. However, no improvement is observed when ranged base conditions were used compared to the "Specific" model and this is consistent with the CURE plots. Based on the GOFs, model V7 is better than the other six (V1-V6). This model appeared to portray an improvement in model predictions over the "Generic" model.

3.2 Cross Validation

The resulting RMSE values are presented in Table 4. Among the four models, the "specific" model has the best predictive ability and this is consistent with the results from the previous section. Nonetheless, the unavailability of AFs leads to the use of a "Generic" model. However, Table 4 shows that model V7 has better predictive ability compared to the "Generic" model.

Table 4: Cross Validation using RMSE

Models	*$RMSE_1$*	*$RMSE_2$*
Generic	1.27	1.13
Specific	0.94	0.62
R2	1.1	0.96
V7	1.15	1.00

4 Conclusions

This paper aimed to evaluate a tradeoff between the quality of the SPFs and the coverage of the roadway network. According to the HSM and current practices of several agencies, specifying base conditions for model development ensures the quality of the SPF. However, while maintaining the quality of SPF is the top priority, coverage of the roadway network is important too since any portion of the network not used in model development will likely require to be adjusted. Unfortunately, adjustment factors are not available for all roadway characteristics. This can limit the application of even a high-quality SPF and lead to the implementation of a generic SPF. Although a generic model covers the entire network, the quality has to be highly compromised.

This research demonstrates two approaches to model SPFs that tried to balance the quality of the model and roadway network coverage to achieve the maximum benefit. One approach recommended is to broaden the range of base conditions to incorporate as much network as possible. Another method suggested integrating independent variables into the model. Although the SPFs with base conditions (specific or ranges) show very prominent predictive power and model quality, they are still dependent on the availability of adjustment factors. On the other hand, the model with explanatory variables (lane width, shoulder width, and curvature) has stood out by portraying improvement in model fit and predictions compared to the generic model and is independent of any need for adjustments. However, the models described in the paper are developed using crashes of all severities indiscriminately. The next step of this research will integrate crash severity into the model development. Additionally, future research can be employed by exploring some other variables including vertical curve, speed limit, etc.

References

1. Tanzen, R. *The Relationship between Roadway Homogeneity and Network Coverage for Network Screening*. University of Kentucky, 2020.
2. Resende, P. T. V., and R. F. Benekohal. Effects of Roadway Section Length on Accident Modeling. 1997.
3. Miaou, S.-P. The Relationship between Truck Accidents and Geometric Design of Road Sections: Poisson versus Negative Binomial Regressions. *Accident Analysis & Prevention*, Vol. 26, No. 4, 1994, pp. 471–482.
4. Hauer, E., and J. Bamfo. Two Tools for Finding What Function Links the Dependent Variable to the Explanatory Variables. Presented at the proceedings of the ICTCT 1997 Conference, Lund, Sweden, 1997.
5. AASHTHO. *Highway Safety Manual, First Edition*. American Association of State Highway and Transportation Officials, Washington, D.C., 2010.
6. Srinivasan, R., and K. Bauer. *Safety Performance Function Development Guide: Developing Jurisdiction-Specific SPFs*. Publication FHWA-SA-14-005. Federal Highway Administration Office of Safety, 2013, p. 47.
7. Srinivasan, R., D. Carter, and K. Bauer. *Safety Performance Function Decision Guide: SPF Calibration versus SPF Development*. Publication FHWA-SA-14-004. FHWA, 2013.
8. Bauer, K. M., and D. W. Harwood. Safety Effects of Horizontal Curve and Grade Combinations on Rural Two-Lane Highways. *Transportation Research Record*, Vol. 2398, No. 1, 2013, pp. 37–49. https://doi.org/10.3141/2398-05.
9. Lyon, C., B. Persaud, and F. Gross. The Calibrator: An SPF Calibration and Assessment Tool User Guide. 2016, p. 98.
10. Green, E. *Segmentation Strategies for Road Safety Analysis*. Dissertation. University of Kentucky, 2018.
11. Kuhn, M., and K. Johnson. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press, 2019.
12. Varoquaux, G. Cross-Validation Failure: Small Sample Sizes Lead to Large Error Bars. *Neuroimage*, Vol. 180, 2018, pp. 68–77.
13. Hariharan, B. *Functional Form and Heterogeneity on Safety Performance Function Estimation*. Dissertation. The Pennsylvania State University, 2015.
14. Hauer, E., J. Kononov, B. Allery, and M. S. Griffith. Screening the Road Network for Sites with Promise. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1784, No. 1, 2002, pp. 27–32. https://doi.org/10.3141/1784-04.