INVESTIGATION OF DIFFERENCE BETWEEN NETWORK SCREENING RESULTS BASED ON MULTIVARIATE AND SIMPLE CRASH PREDICTION MODELS

Jiří Ambros
Centrum dopravního výzkumu, v.v.i.
Líšeňská 33a
636 00 Brno, Czech Republic
phone +420 541 641 362
fax +420 541 641 712
e-mail jiri.ambros@cdv.cz
(corresponding author)

Veronika Valentová
Centrum dopravního výzkumu, v.v.i.
Líšeňská 33a
636 00 Brno, Czech Republic
phone +420 541 641 355
e-mail veronika.valentova@cdv.cz

Zbyněk Janoška
Centrum dopravního výzkumu, v.v.i.
Líšeňská 33a
636 00 Brno, Czech Republic
phone +420 541 641 799
e-mail zbynek.janoska@cdv.cz

Submitted for presentation at the 94th Annual Meeting of the Transportation Research Board, January 11-15, 2015.

Total number of words: 4472 + 6 figures/tables × 250 = 5972

ABSTRACT

Identification of hazardous road locations (network screening) is the first step of road network safety management process. According to state-of-the-art knowledge it should be conducted using empirical Bayes technique which relies on a crash prediction model (safety performance function). However there is a dilemma in choice of a model: researchers strive for multivariate models, which demand area-wide databases of several variables; on the other hand practitioners require simple models, based only on fundamental variables, which are more easily available and able to be periodically updated. The research question was: "What is the difference between network screening results based on multivariate and simple crash prediction models?" For the purpose of investigation multivariate and simple crash prediction models were developed for regional road network of South Moravia (Czech Republic) and used in network screening. The results based on both models were compared and discussed – the conclusion is that results from network screening with simple model are generally comparable to the multivariate model.

1 INTRODUCTION

Identification of hazardous road locations (also called network screening) is defined as 'the process by which a road network is screened to identify sites that require safety investigation' (*1*). Having a long tradition in traffic engineering it is seen as the first step of road network safety management process (*2, 3, 4*). It is also one of the most frequent tasks; for example in a survey of which safety analyses are conducted in road agencies in the US and Canada (*1*), identification of hazardous road locations was selected by 100% of 32 interviewees.

The identification process has to be efficient, since resources are limited and should not be wasted on incorrectly identified sites, while not treating the ones which are truly unsafe and may not be identified (*1*). The process should also enable ranking the locations from most likely to least likely to realize a reduction in crash frequency with implementation of countermeasures (*3*). In order to fulfill these requirements the empirical Bayes (EB) approach has been proposed (*5*); network screening using the EB technique provides the best diagnostic performance and is therefore the recommended solution (*6, 7, 8, 9*). Use of the EB method should rely on a crash prediction model, which are also often called safety performance function (*10*). There are basically two types of models: simple and multivariate (*11, 1*). The difference is in the independent variables (covariates): while simple model involves only traffic volume and segment length, multivariate models use also other variables, usually geometric characteristics.

In practice network screening is in the responsibility of road agency; however developing a model is not an easy task (*12*) and it may be demanding in terms of data requirements and staff qualification. On the contrary a simple model, requiring basic data only, provides an easier solution and road agencies in several countries have been using them, for example Australia (*13*), Denmark (*14*), Finland and Lithuania (*15*) or the Netherlands (*16*). Simple models are also used as 'baseline' models in Highway Safety Manual (*3*) and SafetyAnalyst software (*17*); the predictive methodology, applied in these two sources, consists of two steps: (1) developing baseline models for nominal conditions, and (2) multiplying the 'baseline' models by crash modification factors (CMFs) to capture changes in geometric design and operational characteristics (deviations from nominal conditions).

Nevertheless a number of researchers have warned against using simple models use since they may introduce important omitted-variable bias (*18, 19, 20*). Therefore while researchers strive for multivariate models, which demand area-wide databases of several variables, practitioners require simple models, based only on fundamental variables, which are more easily available and able to be periodically updated. The situation was also described as a dichotomy between what is used in practice and what is used by frontline safety researchers (*20*). The state-of-the-art review of European practices even concluded that recent advances in statistical crash modeling are mostly irrelevant for practical use of crash models (*21*).

This dilemma was an inspiration for this paper. Its research question was: "What is the difference between network screening results based on multivariate and simple crash prediction models?" For the purpose of investigation crash prediction models were developed for regional road network of South Moravia (Czech Republic) and used in network screening. The results based on simple and multivariate models were compared in order to see the potential differences.

2 DATA AND MODELING

Segmentation

The studied network consists of road sections (excluding intersections) of single carriageway two-lane paved rural roads, also known as ‘secondary roads’ (*22*). Their total length of approx. 995 km was divided into homogeneous segments with respect to following variables: annual average daily traffic (AADT), speed limit reduction, road category, number of lanes, paved shoulder. A change of any of these variables marked the end of a segment and beginning of another one. Figure 1 illustrates the principle of this segmentation, following the work of Cafiso et al. (*23*).

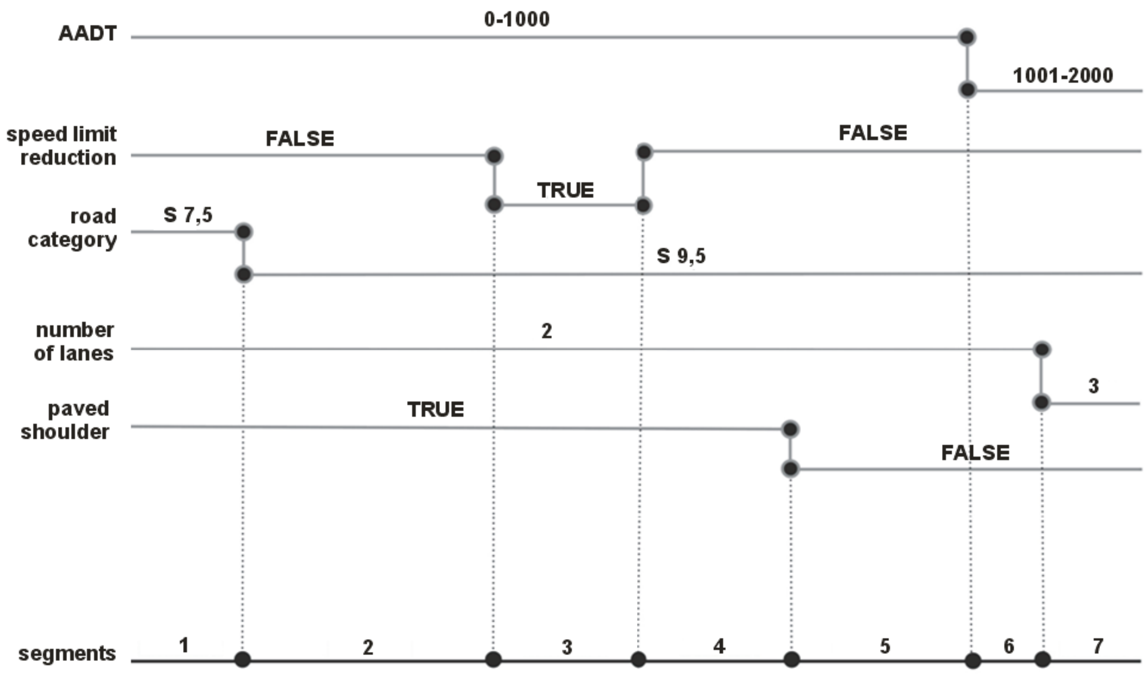

FIGURE 1 Principle of division into homogeneous segments.

In order to obtain segment lengths which will be practical for follow-up safety inspections, segments longer than 500 m were divided into 250 m parts. After this segmentation most of the segments (78%) were 250 m long, 17% were longer and 5% were shorter. Their total number was 3,764.

Variables

These segments were assigned specific values of response variable (crash frequency) and explanatory variables (exposure data, road and traffic characteristics, context and environment variables) which represent safety-related features. Recorded crash frequency data for the period 2009 – 2012 were obtained from the Czech Traffic Police. They include all injury crashes, i.e. with slight, severe or fatal personal consequences. There were 1,030 crashes in total, ranging between 0 and 12 crashes within a segment. Further explanatory variables were added:

– Annual average daily traffic (AADT) and percentage of heavy goods vehicles (HGV) data were used to represent the crash exposure. These data were acquired from Czech Road and Motorway Directorate, based on the results of national traffic census 2010.

– The segment length; its generation was described in the previous paragraph.

– Road and traffic characteristics data were also obtained from a Czech Road and Motorway Directorate database, reflecting the state in July 2010. Density of intersections with minor rural roads and density of roadside facilities were computed as frequencies divided by segment lengths. Some variables were not well represented across their range of values: regarding number of lanes, 95% cases were two-lane segments; regarding presence of speed limit reduction, 99% cases were without reduction, i.e. with 90 kph speed limit applied. These two variables were thus removed.

– Another variable was related to the level of quality of road pavement. It is defined according to collected data about cracks and potholes in five classes, from 1 (excellent) to 5 (wrecking). Data collected by Czech company PavEx Consulting in 2011 were used (*24*), where sections were classified according to the worst quality level presence; however only the list containing sections with levels 4 (unsatisfactory) and 5 (wrecking) was available to authors. Therefore for further analyses each segment was assigned the ratio of length with the pavement quality level 4 or 5; the value is between 0 and 1 (0 means quality better than 4 or 5; 1 means total length in these quality levels).

– Some of context and environment variables were also acquired from the Czech Road and Motorway Directorate database, as of July 2010. The data for other ones were collected additionally: these were average curvature change rate (CCR), as a traditional alignment consistency indicator, and forest environment, which is linked to wet surface or game crashes (*25*). CCR was computed as sums of angles between vertices divided by the sum of their lengths (*26*). The information about continuous forest around the road segment was collected manually from on-line maps, which utilize the data of Czech Environmental Information Agency (CENIA).

All variables and their characteristics are summarized in Table 1.

TABLE 1 Overview of data with description and descriptive statistics of variables

Data type	Abbr.	Description	Data type and unit	Descriptive statistics (min / max / mean / SD or frequencies)
Crash data	R	4-year frequency of reported injury crashes	count	0 / 12 / 0.27 / 0.69
Exposure data	$AADT$	Annual average daily traffic	continuous [vehicle per day]	91 / 18,498 / 2,459.37 / 2,229.86
	HGV	HGV percentage	continuous	0.06 / 0.50 / 0.18 / 0.06
	L	Segment length	continuous [metres]	51.00 / 499.88 / 264.29 / 64.03
Road and traffic data	CAT	Road category	binary (0 = 7.5 m; 1 = 9.5 or 11.5 m wide)	0: 3,156; 1: 608
	SH	Paved shoulder	binary (0 = FALSE; 1 = TRUE)	0: 3,333; 1: 431
	PAV	Pavement quality	continuous	0 / 1 / 0.49 / 0.49
Context and environment data	CCR	Average curvature change rate	continuous [gon per km]	0.0 / 1,498.18 / 98.03 / 134.27
	INT	Density of intersections with minor rural roads	continuous [number per km]	0.0 / 16.90 / 1.16 / 2.40
	FAC	Density of roadside facilities	continuous [number per km]	0.00 / 52.00 / 2.58 / 5.76
	FOR	Forest environment	binary (0 = FALSE; 1 = TRUE)	0: 2,977; 1: 787

Modeling

Prior to the modelling, variables inter-correlation was investigated. The highest significant correlation was found between road category and paved shoulder (almost 0.67). Since the difference in lane width for each road category is minimal (category 7.5 is by 0.25 m narrower than 9.5 and 11.5), the connection may be therefore caused by the relation between road category and paved shoulder, which is governed by Czech standards: wider roads tend to be equipped with paved shoulder. Since the amount of intercorrelation was considerable and given the fact that road category is a derived variable, it was removed from the data set.

A negative binomial regression modeling was used to develop the models. This type is probably the most frequently used model in crash-frequency modelling (*18*). Details of models development may be found elsewehere (e.g. *27*). The required prediction model form was as follows:

$$P_i = \beta_0 \cdot AADT_i^{\beta_1} \cdot \exp(\textstyle\sum_{i=2}^{n} \beta_i x_i) \qquad (1)$$

where β_i are coefficients to be estimated in modelling and x_i are explanatory variables.

SPSS procedure GENLIN was used for the modelling. Only variables with 95% statistically significant influence were kept in the model. The *simple* model included two explanatory variables: AADT and segment length. The *multivariate* model was built by choosing the explanatory variables function forms and adding them into a model (or removing from the model) in steps. Goodness-of-fit was checked by three criteria: maximal decrease of Akaike information criterion (AIC), maximal decrease of overdispersion parameter value and the most ideal shape of cumulative residuals graph (*28*).

In the model development, influence of three variables was not significant and they were removed (HGV percentage, roadside facility density and intersection density). Final models parameters are listed in Table 2.

TABLE 2 Parameters $\boldsymbol{\beta_i}$ of simple and multivariate models, their Akaike information criteria (AIC) and overdispersion parameters

Model	Inter-cept	ln($AADT$)	L	CCR	FOR (if FALSE if TRUE)	PAV	SH (if FALSE if TRUE)	AIC	Over-dispersion parameter
Simple	-7.978	0.776	0.003	–	–	–	–	4671	1.236
Multivariate	-9.032	0.895	0.002	0.002	-0.309 0	-0.173	0.370 0	4599	1.062

Using the coefficient values model equations are formed as follows:

– Simple model:

$$P_i = \exp(-7.978) \cdot AADT_i^{0.776} \cdot \exp(0.003 \cdot L_i) \qquad (2)$$

– Multivariate model:

$$P_i = \exp(-9.032) \cdot AADT_i^{0.895} \cdot \exp(0.002 \cdot L_i) \cdot \exp(0.002 \cdot CCR_i) \cdot \begin{Bmatrix} \exp(-0.309)\ if\ FOR_i = FALSE \\ 1\ if\ FOR_i = TRUE \end{Bmatrix} \cdot$$
$$e^{-0.173 \cdot FAC} \cdot \begin{Bmatrix} \exp(0.370)\ if\ SH_i = FALSE \\ 1\ if\ SH_i = TRUE \end{Bmatrix} \qquad (3)$$

Based on the coefficient signs it is obvious that:

- Crash frequency increase is associated with traffic volume, segment length, curvature change rate, presence of forest and paved shoulder. All these influences are logical and consistent with general literature (*16, 25*).
- On the other hand crash frequency decrease is associated with the share of unsatisfactory or wrecking pavement condition. While this may sound contradictory, similar results were also found in other studies (*25, 29*). It is also possible that crash frequency in such conditions is influenced by other variables not controlled for.

Network screening

Models were used to obtain predicted crash frequency (P) for each segment (i). Empirical Bayes estimate of expected crash frequency (EB) was then calculated, using predicted crash frequency, recorded crash frequency and length-dependent overdispersion parameter (*5, 30*). Finally potential for safety improvement (PSI) was obtained as a difference between predicted crash frequency and EB estimate (*31*).

$$EB_i = w_i \cdot P_i + (1 - w_i) \cdot R_i \quad (4)$$

$$w_i = \frac{k_i}{k_i + P_i} \quad (5)$$

$$k_i = k \cdot L_i \quad (6)$$

$$PSI_i = EB_i - P_i \quad (7)$$

where

EB_i	EB estimate
w_i	weight
P_i	predicted crash frequency
R_i	recorded crash frequency
k_i	overdispersion parameter
L_i	segment length
PSI_i	potential for safety improvement

Values of PSI for both simple and multivariate models were used for network screening. After their descending sorting, two lists of segment numbers were developed. Following previous studies (e.g. *7*), 1%, 2.5% and 5% upper tails were further used in order to investigate the differences.

3 COMPARISON AND RESULTS

The results consisted of two ranked lists of segments: one from network screening based on multivariate model, the second from network screening with simple model. Both sets had variants for 1%, 2.5% and 5% upper tails. In order to discuss the potential differences between them, three comparisons were made. Details of used statistical tests may be found in various in statistical textbooks (e.g. *32*).

Firstly the lists of segment numbers were compared using Spearman's rank correlation coefficient. It compares the amount of difference between two rankings of the same variable (segment numbers in this case). The numbers that did not match had to be excluded. Table 3 reports the results;

the coefficients for all the upper tails were above 0.9 (statistically significant at the 0.01 level, two-tailed).

TABLE 3 Results of comparison of segment numbers

Upper tail	# total segments	# matched segments	Spearman's rank correlation coefficient
1%	38	35	0.97
2.5%	94	78	0.96
5%	188	185	0.89

Secondly equality of statistical distributions of PSI values was tested. According to Shapiro-Wilk test of normality the samples are not normally distributed. Using Mann-Whitney U test, which is appropriate for other than normally distributed data, distributions were then found to be equal across both simple and multivariate results, with exception of the list based on 5% upper tail. The tests were conducted at significance level 0.05. Table 4 shows the results of null hypotheses testing.

TABLE 4 Results of comparison of PSI distributions

Upper tail	Model	Shapiro-Wilk test of normality (H_0: data is normally distributed)	Independent samples Mann-Whitney U test (H_0: the distributions are the same)
1%	Simple	H_0 rejected ($p < 10^{-3}$)	H_0 retained (p = 0.19)
	Multivariate	H_0 rejected ($p < 10^{-3}$)	
2.5%	Simple	H_0 rejected ($p < 10^{-3}$)	H_0 retained (p = 0.08)
	Multivariate	H_0 rejected ($p < 10^{-3}$)	
5%	Simple	H_0 rejected ($p < 10^{-3}$)	H_0 rejected (p = 0.01)
	Multivariate	H_0 rejected ($p < 10^{-3}$)	

The third test compares the results in terms of percentage of segments identified in both lists. The list from multivariate model was taken as the 'base' and differences were sought in the list from simple model. The results are reported in Table 5 two forms:

– In terms of count, i.e. how many segments would be unidentified due to using simple model instead of multivariate.

– In terms of PSI, i.e. how much of total PSI would the unidentified segments contain.

TABLE 5 Results of comparison of identified segments

Upper tail	# total segments	# missing segments	% not identified in both lists	
			count	PSI
1%	38	3	8%	5%
2.5%	94	16	17%	10%
5%	188	3	2%	1%

4 DISCUSSION AND CONCLUSIONS

Two models (simple and multivariate) were used for screening of the same road network. From final ranked lists of road segments 1%, 2.5% and 5% upper tails with their PSI values were selected for comparison. The results may be linked with following questions:

- *Are identified segments the same?* The lists of segment numbers were compared; the rank correlation coefficients for all the upper tails were significant and at least 0.9.

- *Are identified segments equally safe?* According to statistical test the distributions of PSI values in both lists are equal for 1% and 2.5% upper tail, but not for 5% upper tail. Thus with exception of the latter list, there is no significant difference between the lists based on simple or multivariate model.

- *How many segments were 'lost' due to using simple model only?* Between 2% and 17% of segments (depending on size of upper tail) were unidentified due to using simple model instead of multivariate.

- *How unsafe were the unidentified segments?* The unidentified segments contained between 1% and 10% of total PSI (depending on size of upper tail).

It is therefore obvious there are some differences between the results based on network screening with simple or multivariate model. Nevertheless the most critical segments (in 1% and 2.5% upper tail lists) are likely to be identified by both simple and multivariate model. The important finding is that the segments, which were 'lost' due to using simple model instead of multivariate, did not contain more than 10% of total PSI.

Similar agreement was reached by Florida researchers (*33*) who compared performance of network screening of flow-only and full models from SafetyAnalyst: based on t-test of PSI distributions they found no significant difference in the means between the two sets of PSIs. On the other hand a Texas study (*34*) showed that the baseline models combined with CMFs may produce much larger variance compared to the full models; the authors concluded that the full model should be preferred.

From a wider perspective there are two general orientations in crash modeling; Hauer (*35*) visualizes them as two different flashlights aimed at left or right side of the model.

$$predicted\ value = f(traits\ and\ parameters) \tag{8}$$

- Focus on the left side of equation (prediction) is the *focus on applications*. The user sees a model as the tool for generating estimates of expected crash frequency.
- Focus on the right side of equation (function of traits and parameters) is the *focus on research*. Researcher focuses on the value of the unknown parameters and on the function which links the traits and parameters; he/she is interested in understanding how changing the various traits will affect the predicted value in order to predict the safety effect of design choices and interventions. In this perspective the model represents the current understanding cause and effect and is thought to be a source of crash modification factors.

This distinction was also used by Persaud (*1*) who divided the models into two classes: crash prediction models and crash causation models. While crash causation models should be related to factors that *explain* crash causation, crash prediction models suffice with variables *associated* with crashes, for which data are more easily available. Data availability is often the most limiting factor in choice of explanatory variables (*36*); another option is choosing the variables representing risk factors which are amenable to change (*20*).

Nevertheless it is obvious that the perspective, adopted in the paper, is focused on network screening only, which is one of more pragmatic application of crash modeling, focusing on the application only. Such models should not be used in causation perspective nor for producing crash modification factors.

Having these limitations in mind, the reported investigation showed the case when simple models may be used in network screening without causing a significant bias. For the authority of regional road network of South Moravia (Czech Republic) it may be recommended to rely on network screening with simple crash prediction models. Compared to multivariate modeling it will reduce their demands on time, qualified staff and data requirements.

ACKNOWLEDGMENTS

The authors are grateful to Salvatore Cafiso, Rune Elvik and Bhagwant Persaud for valuable consultations with crash prediction modelling. The study was conducted with support of the Czech Ministry of Interior's research project No. VG20112015013 "Identification and treatment of high risk road spots and sections" (IDEKO).

REFERENCES

1. Persaud, B. N. *Statistical Methods in Highway Safety Analysis: A Synthesis of Highway Practice*. NCHRP Synthesis 295. Transportation Research Board of the National Academies, Washington, D.C., 2001.
2. Hauer, E., J. Kononov, B. Allery, and M. S. Griffith. Screening the Road Network for Sites with Promise. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 1784*, Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 27-32.
3. *Highway Safety Manual.* First Edition. American Association of State Highway and Transportation Officials, Washington, D.C., 2010.
4. Montella, A. A comparative analysis of hotspot identification methods. *Accident Analysis and Prevention*, Vol. 42, 2010, pp. 571-581.
5. Hauer, E., D. W. Harwood, F. M. Council, and M. S. Griffith. Estimating Safety by the Empirical Bayes Method: A Tutorial. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 1784*, Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 126-131.
6. Cheng, W., and S. Washington. Experimental evaluation of hotspot identification methods. *Accident Analysis and Prevention*, Vol. 37, 2005, pp. 870-881.
7. Cheng, W., and S. Washington. New Criteria for Evaluating Methods of Identifying Hot Spots. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 2083,* Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 76-85.
8. Elvik, R. Comparative Analysis of Techniques for Identifying Locations of Hazardous Roads. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 2083,* Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 72-75.
9. Lim, I.-K., and Y.-J. Kweon. Identifying High-Crash-Risk Intersections: Comparison of Traditional Methods with the Empirical Bayes – Safety Performance Function Method. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 2364*, Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 44-50.

10. Elvik, R. A survey of operational definitions of hazardous road locations in some European countries. *Accident Analysis and Prevention*, Vol. 40, 2008, pp. 1830-1835.
11. Hauer, E. *Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Pergamon Press, Oxford, 1997.
12. Eenink, R., M. Reurings, R. Elvik, J. Cardoso, S. Wichert, and C. Stefan. *Accident Prediction Models and Road Safety Impact Assessment: recommendations for using these tools*. SWOV, 2008.
13. Jurewicz, C., and B. Thompson. Crash risk estimation and assessment tool. In *Australasian Road Safety Research, Policing and Education Conference*. Australian Transport Council, National Convention Centre, Canberra, 2010.
14. Greibe, P. Accident prediction models for urban roads. *Accident Analysis and Prevention*, Vol. 35, 2003, pp. 273-285.
15. Peltola, H., R. Rajamäki, and J. Luoma. A tool for safety evaluations of road improvements. *Accident Analysis and Prevention*, Vol. 60, 2013, pp. 277-288.
16. Reurings, M., and T. Janssen. *Accident prediction models for urban and rural carriageways (based on data from The Hague region)*. Report No. R-2006-14. SWOV Institute for Road Safety Research, Leidschendam, 2007.
17. Harwood, D. W., D. J. Torbic, K. R. Richard, and M. M. Meyer. *SafetyAnalyst: Software Tools for Safety Management of Specific Highway Sites*. Report No. FHWA-HRT-10-063. Federal Highway Administration, McLean, 2010.
18. Lord, D., and F. Mannering. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A*, Vol. 44, 2010, pp. 291-305.
19. Mitra, S., and S. Washington. On the significance of omitted variables in intersection crash modeling. *Accident Analysis and Prevention*, Vol. 49, 2012, pp. 439-448.
20. Mannering, F. L., and C. R. Bhat. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, Vol. 1, 2014, pp. 1-22.
21. Elvik, R. *Assessment and applicability of road safety management evaluation tools: Current practice and state-of-the-art in Europe*. TØI report 1113/2010. Institute of Transport Economics, Oslo, 2010.
22. Gatti, G., C. Polidori, I. Galvez, K. Mallschützke, R. Jorna, M. van de Leur, M. Dietze, D. Ebersbach, C. Lippold, B. Schlag, G. Weller, A. Wyczynski, F. Iman, and C. Aydin. *Safety Handbook for Secondary Roads, RIPCORD-ISEREST project deliverable D13*. 2007.
23. Cafiso, S., A. Di Graziano, G. Di Silvestro, G. La Cava, and B. Persaud. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. *Accident Analysis and Prevention*, Vol. 42, 2010, pp. 1072-1079.
24. Gregor, J., and L. Mališ. *Stav povrchu vozovek silnic II. a III. třídy v Jihomoravském kraji k 30.4.2011*. PavEx Consulting, s.r.o., Brno, 2011. (in Czech)
25. Elvik, R., A. Høye, T. Vaa, and M. Sørensen. *The Handbook of Road Safety Measures, Second Edition*. Emerald, Bingley, 2009.
26. Dantas, A., M. Fowler, and G. Koorey. Effect of road network bendiness on traffic crash occurrence. In *IPENZ Transportation Group Technical Conference*, Tauranga, New Zealand, October 7-10, 2007.
27. Hauer, E. Statistical Road Safety Modeling. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 1897*, Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 81-87.

28. Hauer, E., and J. Bamfo. Two tools for finding what function links the dependent variable to the explanatory variables. In *Proceedings, ICTCT 97 Conference, International Cooperation on Theories and Concepts in Traffic Safety*, Lund, 5-7 November, 1997. ICTCT and the Department of Traffic Planning and Engineering, Lund University, Sweden, 1997.
29. Cenek, P. D., R. J. Henderson, M. Forbes, R. B. Davies, and A. Tait. *The relationship between crash rates and rutting*. NZ Transport Agency research report 545. NZTA, Wellington, 2014.
30. Hauer, E. Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation. *Accident Analysis and Prevention*, Vol. 33, 2001, pp. 799-808.
31. Persaud, B., C. Lyon, and T. Nguyen. Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 1665*, Transportation Research Board of the National Academies, Washington, D.C., 1999, pp. 7-12.
32. Field, A. *Discovering statistics using IBM SPSS Statistics, 4th Edition*. SAGE, Thousand Oaks, 2013.
33. Lu, J., K. Haleem, P. Alluri, and A. Gan. Full versus Simple Safety Performance Functions: A Comparison Based on Urban Four-Lane Freeway Interchange Influence Areas in Florida. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 2398*, Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 83-92.
34. Lord, D., P.-F. Kuo, and S. R. Geedipally. Comparison of Application of Product of Baseline Models and Accident-Modification Factors and Models with Covariates Predicted Mean Values and Variance. In *Transportation Research Record: Journal of the Transportation Research Board*, *No. 2147*, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 113-122.
35. Hauer, E. *The art of regression modeling in road safety*. 2014. (Unpublished manuscript)
36. Reurings, M., T. Janssen, R. Eenink, R. Elvik, J. Cardoso, and C. Stefan. *Accident prediction models and road safety impact assessment: a state-of-the-art, RIPCORD-ISEREST project deliverable D2.1*. 2005.