Transportation Research Board
94th Annual Meeting
January 11–15, 2015
Paper No. 15-1242

Investigation of Difference Between Network Screening Results Based on Multivariate and Simple Crash Prediction Models

Jiří Ambros*, Veronika Valentová, Zbyněk Janoška

Centrum dopravního výzkumu, v.v.i. (Transport Research Centre), Líšeňská 33a, 636 00 Brno, Czech Republic

* jiri.ambros@cdv.cz

INTRODUCTION

Identification of hazardous road locations (network screening) for safety investigation is the first step of road network safety management process and one of the most frequent tasks (*1*).

Screening has to be efficient and should enable ranking the locations according to their potential for safety improvement (*2*).

Screening is in the responsibility of road agency – should not be demanding in terms of data requirements and staff qualification.

Empirical Bayes (EB) approach is recommended (*3*).
It uses a crash prediction model (safety performance function) – simple or multivariate? (*4*)

	SIMPLE MODELS		MULTIVARIATE MODELS
data	only traffic volume and segment length	or	also other variables, usually geometric characteristics
demands	these variables are more easily area-wide available and able to be updated		require area-wide databases of number of variables
use	several road agencies in the world use simple models		several researchers warn against using simple models (omitted-variable bias)
preferred by...	***PRACTITIONERS***		***RESEARCHERS***

Research question:

What is the difference between network screening results based on multivariate and simple crash prediction models?

– regional road network of South Moravia (Czech Republic)
– in total approx. 1,000 km of road sections (excluding intersections) of single carriageway two-lane paved rural roads ("secondary roads")'
– simple and multivariate crash prediction models developed
– models applied according to EB approach in network screening
– identified segments compared and differences discussed

ACKNOWLEDGMENTS
– Consultations on crash prediction modeling:
Salvatore Cafiso (University of Catania, Italy)
Rune Elvik (Institute of Transport Economics, Norway)
Bhagwant Persaud (Ryerson University, Canada)
– Funding: Ministry of Interior 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*. TRB, 2001.
2. *Highway Safety Manual. First Edition*. AASHTO, 2010.
3. Hauer, E. *et al.* Estimating Safety by the Empirical Bayes Method: A Tutorial. *TRR*, 1784, 126-131.
4. Hauer, E. *Observational Before-After Studies in Road Safety: Estimating Effect of Highway and Traffic Engineering Measures on Road Safety*. Pergamon Press, 1997.
5. Cafiso, S. *et al.* Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. *AAP*, 42, 1072-1079.
6. Hauer, E., and J. Bamfo. Two tools for finding what function links the dependent variable to the explanatory variables. In *ICTCT*, 1997.
7. Persaud, B. *et al.* Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. *TRR*, 1665, 7-12.
8. Hauer, E. *The art of regression modeling in road safety*. 2014.

DATA AND MODELING

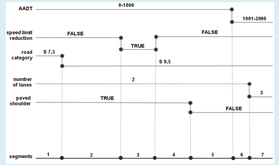
Segmentation

– road sections divided into homogeneous segments with respect to 5 variables
– a change of any of these variables = the end of a segment and beginning of another (*5*)

In order to obtain segment lengths which will be practical for follow-up safety inspections, segments > 500 m were divided into 250 m parts.

Results of segmentation:
5% < 250 m
78% = 250 m
17% > 250 m

Total 3,764 segments



Variables

Variables of safety-related features were assigned to segments:

Response variable: reported frequency of injury crashes (Traffic Police, 2009-2012)

Explanatory variables (exposure data, road/traffic/context/environment variables)
– traffic volumes from national traffic census 2010
– other characteristics from Road and Motorway Directorate database (2010)
– quality of road pavement (2011)
– additional variables computed or collected from on-line maps

	Description	Data type and unit
AADT	Annual average daily traffic	continuous [vehicle per day]
HGV	HGV percentage	continuous
L	Segment length	continuous [metres]
CAT	Road category	binary (0 = 7.5 m; 1 = 9.5 or 11.5 m wide)
SH	Paved shoulder	binary (0 = FALSE; 1 = TRUE)
PAV	Pavement quality	continuous
CCR	Average curvature change rate	continuous [gon per km]
INT	Density of intersections with minor rural roads	continuous [number per km]
FAC	Density of roadside facilities	continuous [number per km]
FOR	Forest environment	binary (0 = FALSE; 1 = TRUE)

Modeling

Explanatory variables

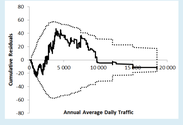
– variables with low variability removed (number of lanes, speed limit reduction)
– also inter-correlated variables removed (road category)
– negative binomial regression with log link function (generalized linear modeling in SPSS) ... estimation of regression parameters b_0, b_1, b_2, b_3...

– *simple* model: AADT and segment length

$$P = \beta_0 \cdot AADT^{\beta_1} \cdot \exp(\beta_2 x_2 + \beta_3 x_3 + ...)$$

– *multivariate* model: AADT and segment length + other statistically significant variables
(insignificant variables removed: HGV percentage, roadside facility density, intersection density)

– check of *model quality*: Akaike information criterion (*AIC*) and over-dispersion parameter (*k*) ... the smaller the better
and plot of Cumulative Residuals (CURE) (*6*)

Model equation	*AIC*	*k*	CURE plot
Simple model $P_i = \exp(-7.978) \cdot AADT_i^{0.776} \cdot \exp(0.003 \cdot L_i)$	4671	1.236	
Multivariate model $P_i = \exp(-9.032) \cdot AADT_i^{0.895} \cdot \exp(0.002 \cdot L_i) \cdot \exp(0.002 \cdot CCR_i) \cdot \begin{bmatrix} \exp(-0.309) & if\ FOR_i = FALSE \\ 1 & if\ FOR_i = TRUE \end{bmatrix} \cdot \exp(-0.173 \cdot PAV_i) \cdot \begin{bmatrix} \exp(0.370) & if\ SH_i = FALSE \\ 1 & if\ SH_i = TRUE \end{bmatrix}$	4599	1.062	

Network screening

Models were used to obtain predicted crash frequency (*P*) for each segment (*i*).

Further calculation steps: (*2, 3, 7*)

– empirical Bayes estimate (*EBi*) $EB_i = w_i \cdot P_i + (1 - w_i) \cdot R_i$

– weight (*wi*) $w_i = \frac{k_i}{k_i + P_i}$

– length-dependent over-dispersion parameter (*ki*) $k_i = k \cdot L_i$

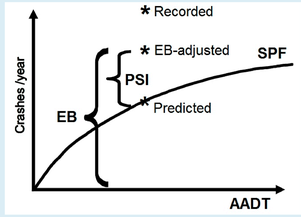
– potential for safety improvement (*PSIi*) $PSI_i = EB_i - P_i$

PSI values for both simple and multivariate models were used for network screening.

After their descending sorting, two sets of segment numbers were developed.

1%, 2.5% and 5% upper tails were further used in order to investigate the differences.

Recorded → Predicted → EB-adjusted → EB and PSI



COMPARISON AND RESULTS

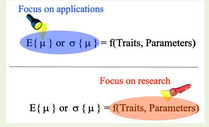
Six ranked lists of segments: from *simple* model and from *multivariate* model (for both 1%, 2.5% and 5% upper tails)

Several comparisons were made in order to answer following questions:

Question	Method	Results
Are identified segments the same?	Comparison of the lists of segment numbers with Spearman's rank correlation coefficient	Coefficients for all the upper tails > 0.9 (statistically significant at 99% confidence level, two-tailed)
Are identified segments equally safe?	Mann-Whitney U test of equality of PSI distributions	Equal distributions across both simple and multivariate results, with exception of the list based on 5% upper tail (statistically significant at 95% confidence level)
How many segments were 'lost' due to using simple model only?	Comparison in % of number of segments	Between 2% and 17% of segments (depending on size of upper tail) unidentified due to using simple model instead of multivariate
How unsafe were the 'lost' segments?	Comparison in % of total PSI	The unidentified segments contained between 1% and 10% of total PSI (depending on size of upper tail)

– There are small 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) **were identified by both models.**
– **The segments 'lost'** due to using simple model instead of multivariate **did not contain more than 10% of total PSI.**

Findings are valid only for *prediction* (not *causation*) ! (*1, 8*)



prediction models
(e.g. for network screening)

causation models
(e.g. for developing CMFs)

CONCLUSIONS

In the presented case simple prediction model may be used in network screening without causing a significant bias.

For the authority of regional road network of South Moravia it may be recommended to rely on network screening with simple crash prediction model.

Compared to multivariate modeling it will reduce their time, staff and data demands.