A comparative analysis of identification of hazardous locations in regional rural road network

V. Valentová J. Ambros Z. Janoška

Transport Research Centre, Líšeňská 33a, Brno, Czech Republic
email: veronica.valentova@cdv.cz, jiri.ambros@cdv.cz, zbynek.janoska@cdv.cz

subm. 14th January 2013 approv. after rev. 24th April 2014

Abstract

The first step of the road safety management cycle is the identification of hazardous road locations. Traditionally, the identification criterion in the Czech Republic has been recorded crash frequency; however, it has tended to omit the important influence of natural variations known as regression to the mean. Using the expected number of crashes and empirical Bayes adjustment is the recommended solution. This number is calculated with the use of a crash prediction model, taking into account available explanatory factors and controlling for potentially confounding variables at the same time.

The aim of this study was to compare both traditional and empirical Bayes approaches to identification of hazardous road locations. A crash prediction model for the regional rural road network was developed for the purpose of the study. The whole 2nd class road network in Czech region of South Moravia was used. The resulting expected injury crash frequency was further adjusted by the empirical Bayes estimate.

The empirical Bayes estimate was used as a criterion for the ranking of hazardous road locations list. At the same time, another list was produced using the Czech traditional criterion of recorded crash frequency. Three pairs of lists were developed for three time periods (2007 – 2009, 2008 – 2010 and 2009 – 2011) and compared.

The results showed that the predictions models with EB adjustment offered the results which were more stable in time compared to the traditional approach. This proves the potential suitability of empirical Bayes approach identification in the road safety management practice.

Keywords – hazardous road location, crash prediction model, empirical Bayes, regional road, rural road

1. Introduction

Traffic is an inevitable part of everyday life of humans all over the world; however, its negative outcomes include traffic crashes. Road safety is thus a major social issue, calling for specific interventions, framed within a road safety management system [9].

The first step of the road safety management cycle is the identification of hazardous road locations. An identification criterion is needed to detect a hazardous road location. According to a survey of identifying criteria of hazardous road locations in a number of European countries [6], most of them rely on recorded number of crashes.

However, this criterion has tended to omit the important influence of natural variations known as regression to the mean [10]. Therefore, it is, compared to state-of-the-art techniques for identifying hazardous road locations, likely to involve substantial inaccuracies [6].

To this end, state-of-the-art techniques based on the empirical Bayes (EB) adjustment have been promoted by a number of authors [1, 8, 10, 12, 14]. Their calculation uses expected number of crashes derived from a crash prediction model with EB adjustment taking into account several explanatory factors and controlling for potential confounding variables at the same time.

For that reason, crash prediction models were developed. Three time periods (2007 – 2009, 2008 – 2010, 2009 – 2011) were investigated. For each of these time periods, a separate prediction model was developed. Temporal continuity in the identification of black spots was compared with the prediction model and the currently used approach, based on traditional definition. The study was conducted on a part of Czech regional rural road network. The whole 2nd class road network in one of the Czech regions (South Moravia) was used. This network provides a traffic connection between towns and larger territorial units in the region.

The study was undertaken in the following steps:

1. Data collection and development of crash prediction models, calculation of expected crash frequency and EB estimate.
2. Identification of hazardous road locations based on the EB estimates.
3. Identification of hazardous road locations based on traditional criterion.
4. Comparison of the hazardous road location lists.

These steps are described in the following text. The aim of this study was to compare these two methods based on the percentage of the agreement in the identified locations in different time periods. Theoretically, the ideal method of identifying hazardous road locations should identify true hazardous locations which should remain the same in time (provided there are no changes in the infrastructure). The prediction models and the EB approach control for the influence of exposure and regression to the mean, therefore, the results should be more consistent.

2. Data collection and crash prediction modelling

A crash prediction model for the regional rural road network was developed in previous studies [16, 17]. The original model was prepared for time period 2009 - 2011 for all crashes in Czech Traffic Police records regardless of their severity. For the purpose of this study it had to be modified by using data from time period 2007 - 2011. In addition, there was a crash reporting threshold change in 2009: the value of property damage only (PDO) crashes increased from the value of CZK 50 000 (2000 €) to CZK 100 000 (4000 €). In order to avoid the resulting incompatibility in time series of numbers of property damage only crashes, only injury crashes were used. New forms of prediction models were developed as well.

2.1. Segmentation

The segmentation was done in a way similar to Cafiso et al. [3]. The road sections, excluding intersections, were divided into segments which were homogenous with the respect to the several following variables:

- annual average daily traffic (AADT)
- presence of speed limit reduction
- road category
- number of lanes
- presence of paved shoulder

A change of any of these variables marked the end of the segment and the beginning of another one. There were 839 segments with the length ranging from 51 to 6456 m.

Therefore, segments longer than 500 m were divided into 250 m parts. Finally, the length of the segments was between 50 and 500 m. Most of the segments (2925) are 250 m long, 196 segments are shorter and 643 segments are longer. These segments were assigned with specific values of a response variable (injury crash frequency) and various explanatory variables. The length of 250 m was chosen due to the similarity to the traditional criterion, where the sliding window of the length of 250 m has been used.

2.2. Variables

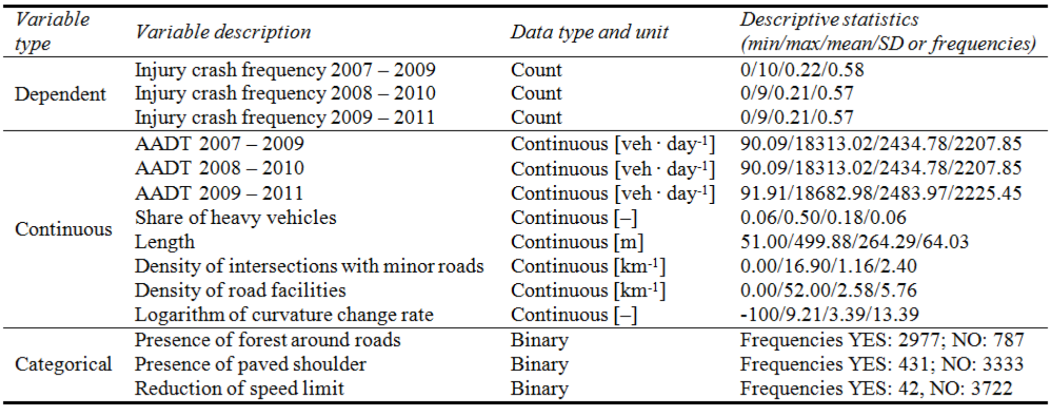
In this part, variables used for statistical modelling are briefly described. Table 1 summarizes variables used in analysis. Crash data were provided by the Police of the Czech Republic. This data contain the information on the crash localization recorded by GPS since 2007. At the time of writing 2011 was the last available year, thus total period 2007 – 2011 was available. Since common practice of crash analyses is using 3-year periods, 3 such periods were chosen: 2007 – 2009, 2008 – 2010, 2009 – 2011.

Exposure variables were represented by annual average daily traffic and the share of heavy vehicles. These data were provided by the Czech Road and Motorway Directorate, based on the results of national road traffic census in 2010. Data was adjusted by a growth factor for different time periods. The road characteristics, context and environment variables were obtained from the database of the Czech Road and Motorway Directorate. The used variables included presence of a paved shoulder, curvature change rate, reduced speed limit, density of intersections with minor rural roads (per 1 km), density of road facilities (per 1 km), and presence of forest around road.

All the three national databases are periodically updated (crashes annually, exposure each five years, road data twice a year) and are generally consistent and reliable. The only inconsistency related to mentioned PDO crashes reporting threshold change in 2009 – it was circumvented by considering injury crashes only which are not influenced by this change. The stability was also proven with comparison of trend of injury crashes in the whole Czech Republic.

Both trends are relatively similar (see Figure 1). The trends consistency was furthermore tested using time series sample odds ratios and confidence interval according to Hauer [10]: calculated 95% confidence interval was (0.84, 1.12), i.e. close to 1.0 and including the value of 1.0, as requested.

Tab. 1 - Overview of variables and their descriptive statistics



Variable type	*Variable description*	*Data type and unit*	*Descriptive statistics (min/max/mean/SD or frequencies)*
Dependent	Injury crash frequency 2007 – 2009	Count	0/10/0.22/0.58
	Injury crash frequency 2008 – 2010	Count	0/9/0.21/0.57
	Injury crash frequency 2009 – 2011	Count	0/9/0.21/0.57
Continuous	AADT 2007 – 2009	Continuous [veh · day^{-1}]	90.09/18313.02/2434.78/2207.85
	AADT 2008 – 2010	Continuous [veh · day^{-1}]	90.09/18313.02/2434.78/2207.85
	AADT 2009 – 2011	Continuous [veh · day^{-1}]	91.91/18682.98/2483.97/2225.45
	Share of heavy vehicles	Continuous [–]	0.06/0.50/0.18/0.06
	Length	Continuous [m]	51.00/499.88/264.29/64.03
	Density of intersections with minor roads	Continuous [km^{-1}]	0.00/16.90/1.16/2.40
	Density of road facilities	Continuous [km^{-1}]	0.00/52.00/2.58/5.76
	Logarithm of curvature change rate	Continuous [–]	-100/9.21/3.39/13.39
Categorical	Presence of forest around roads	Binary	Frequencies YES: 2977; NO: 787
	Presence of paved shoulder	Binary	Frequencies YES: 431; NO: 3333
	Reduction of speed limit	Binary	Frequencies YES: 42, NO: 3722

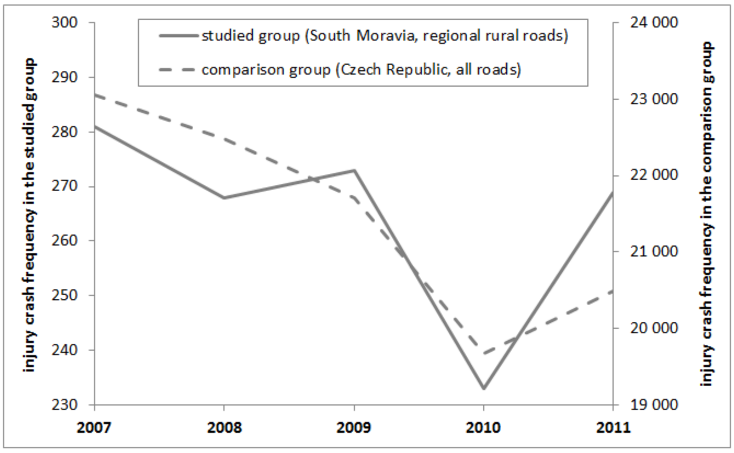


Fig. 1 - Comparison of annual injury crash frequency trends in the studied group (South Moravian regional rural roads) and the comparison group (all Czech roads)

2.3. Modelling

The statistical models were developed using generalized linear modelling according to recommendations provided e.g. by Hauer [13] or Reurings et al. [15]. The injury crash frequency in each of time periods was a dependent variable. The model form was negative binomial with a log-linear link function.

The general form of the models is as follows:

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

where

$\widehat{N_\iota}$ = expected number of crashes;

$AADT_i$ = annual average daily traffic;

β_0 = intercept;

β_i = model coefficients;

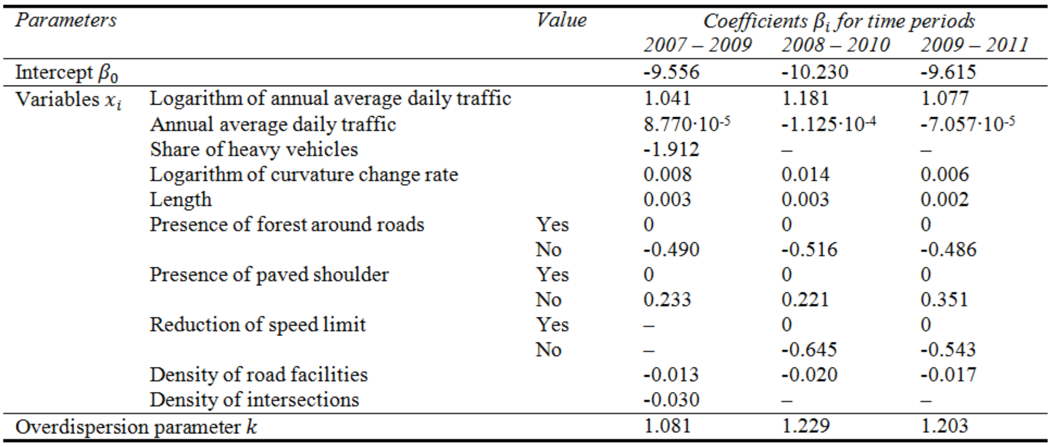
x_i = explanatory variable.

Based on literature survey, a general model form was modified to $\beta_0 \cdot AADT_i^{\beta i} \cdot \exp(\beta_2 \cdot AADT_i)$. This form is better suitable for functional representation of larger *AADT* values [13]. Since generalized linear modelling with log-linear link function was used, coefficients β_i for equation $\ln\beta_0 + \beta_1 \cdot \ln AADT_i + \beta_2 \cdot AADT_i$ were estimated in the first step. Logarithm function was used for *AADT* and curvature change rate, consistently with other studies [3, 13, 15]. Since logarithm for zero is not defined (it approaches $-\infty$), straight sections with zero curvature change rate were assigned value $\ln\mathrm{CCR} = -100$. This value corresponds to $\mathrm{CCR} = 3.72 \cdot 10^{-44}$ which is sufficiently close to zero.

In further steps remaining explanatory variables were added in turn and their coefficients were estimated using datasets of three time periods 2007 – 2009, 2008 – 2010, 2009 – 2011. Only variables with 95% statistically significant influence ($p < 0.05$) were kept in the model. Resulting coefficients are listed in Table 2.

The signs of their values (explanatory variable exponents) provide interpretation of increase or decrease of predicted variable (injury crash frequency) [18].

Tab. 2 - Models for injury crashes on rural 2nd class roads in South Moravia in different time periods



Parameters		*Value*	*Coefficients β_i for time periods*		
			2007 – 2009	*2008 – 2010*	*2009 – 2011*
Intercept β_0			-9.556	-10.230	-9.615
Variables x_i	Logarithm of annual average daily traffic		1.041	1.181	1.077
	Annual average daily traffic		$8.770 \cdot 10^{-5}$	$-1.125 \cdot 10^{-4}$	$-7.057 \cdot 10^{-5}$
	Share of heavy vehicles		-1.912	–	–
	Logarithm of curvature change rate		0.008	0.014	0.006
	Length		0.003	0.003	0.002
	Presence of forest around roads	Yes	0	0	0
		No	-0.490	-0.516	-0.486
	Presence of paved shoulder	Yes	0	0	0
		No	0.233	0.221	0.351
	Reduction of speed limit	Yes	–	0	0
		No	–	-0.645	-0.543
	Density of road facilities		-0.013	-0.020	-0.017
	Density of intersections		-0.030	–	–
Overdispersion parameter k			1.081	1.229	1.203

Based on this it is obvious that:

- Crash frequency increases with traffic volume, segment length, curvature change rate, presence of forest and paved shoulder. All these influences are logical and consistent with general literature [7, 13, 15].
- Crash frequency decreases with share of heavy vehicles, speed limit reduction, density of road facilities and density of intersections. Apart from speed limit reduction these findings are not consistent with general knowledge; they may be therefore influenced by other variables not controlled for.

3. Identification of hazardous road locations based on the empirical Bayes estimate

The EB approach was used to adjust estimates of the multivariate crash prediction models. This approach is recommended as a state-of-the-art approach to estimate expected number of crashes [1, 8, 10, 12, 14]. It increases the precision of estimation and corrects the regression to the mean bias.

The procedure combines the crash record of every segment and the injury crash frequency expected by the prediction model for similar segments. The basic functional form is as follows:

$$EB_i = w_i \cdot \hat{N}_i + (1 - w_i) \cdot N_i \tag{2}$$

where

EB_i = EB estimate;

w_i = weight;

$\hat{N}_i$ = expected number of crashes estimated by prediction model;

N_i = recorded number of crashes.

The weight indicates the significance of the crash data and the model.

It shows the difference between the safety of a specific segment and the average specified by the prediction model.

The weight was calculated as follows:

$$w_i = \frac{1}{1+\hat{N}_i/k_i} = \frac{k_i}{k_i+\hat{N}_i} \tag{3}$$

where

k_i = overdispersion parameter;

$\widehat{N}_i$ = expected number of crashes estimated by a model.

According to recommendations of Hauer [11] overdispersion parameter was calculated for each segment as follows:

$$k_i = k \cdot L \tag{4}$$

where k is an overdispersion parameter obtained during modelling and

L_i is the segment length.

Values of k for prediction models were 1.081 for time period 2007 – 2009, 1.229 for time period 2008 – 2010, and 1.203 for time period 2009 – 2011.

4. Identification of hazardous road locations based on traditional criterion

The methodology, which is currently employed for the identification of hazardous road locations in the Czech Republic, is based on a calculation of recorded crash frequency at a given section.

According to this methodology of Andres et al. [2], hazardous locations are defined by sliding window with the length of 250 m, which meets at least one of the following conditions:

- at least three injury crashes in one year,
- at least three injury crashes of the same type in three years,
- at least five crashes of the same type within one year,
- while used limits (three and five) have been determined empirically.

Injury crashes are defined as crashes involving an injury or death of one or more people involved in a crash. There are three severity levels of injury crashes: slight injury, severe injury or fatal injury.

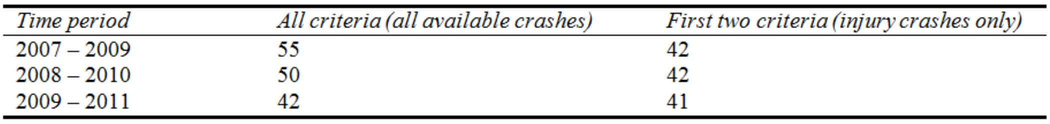
Crash type is defined by the main cause of a crash and can be one of the following types: not caused by driver; speeding; faulty overtaking; denying the right of way; faulty driving style; technical fault of vehicle.

The definition above shows that there are serious shortcomings when it comes to the identification of hazardous road locations. One can imagine the following examples:

- a road segment, where one crash with slight injury happened in each of three consequent years, will be identified as hazardous.
- a road segment, where two slight injuries, two serious injuries and two deaths occurred within three consequent years, will not be identified as hazardous.

Furthermore due to the methodological changes in 2009, which were described above, it made sense to use only the injury crashes for the comparison with the prediction models. Therefore, only first two criteria of the definition were used for the identification of hazardous road locations.

Tab. 3 - Number of identified hazardous road locations using traditional approach



Time period	*All criteria (all available crashes)*	*First two criteria (injury crashes only)*
2007 – 2009	55	42
2008 – 2010	50	42
2009 – 2011	42	41

Table 3 shows that these results were compared to results in a situation when all crashes would have been used. It is obvious that after the methodological change in Police records the omission of the third criterion does not cause any significant change. Thus resulting 42 identified segments will be used in further comparisons and referred to as the “worst” segments. They will be compared with 42 highest values according to descending values of EB estimates.

5. Comparison of hazardous road location lists

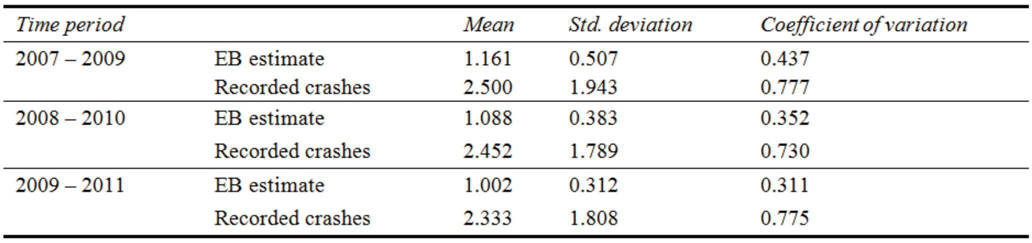
Regarding effective investments in road safety measures, it is necessary to determine which locations are truly hazardous. However using traditional hazardous locations criteria, one cannot distinguish whether hazardous road location is true or not. It is practically impossible to decide without a detailed on-site inspection, ideally also with comparison to similar location without crashes [4].

In a hypothetical case when only true hazardous road locations are identified, they do not vary in time. In order to minimize the variations, all potential risk factors should be ideally controlled for by including them in a prediction model. However, this goal is not practically feasible – the data covering all risk factors are unlikely to be available, therefore, the variation explained is never complete. The prediction models in this study included the influence of traffic and infrastructure. The expected crash frequency and EB approach were used to control for the regression to the mean. These results should contain less variability than the results obtained by the traditional approach, which is based on a frequency of recorded crashes. In order to compare two methods of calculation standard deviations and coefficients of variation were computed for the 42 worst segments as well as using recorded crash frequency. Results given in Table 4 show approximately half of EB estimate variability (coefficient of variation) in comparison to recorded crash frequency.

The comparison of the number of locations identified in all three time periods or just in some of them is visualized in Figure 2 and Figure 3. The figures show three circles which represent sets of locations identified in the three time periods. The circles are overlapping, while overlaps represent sets of locations identified in more than one period and size of overlap area indicates amount of locations indicated in the time periods. Numbers of identified locations are also reported in the original sets as well as in the overlaps. Based on these number conclusions are as follows:

- According to the EB approach in total 64 different hazardous locations were identified. 21 of them were identified in all three time periods, which makes up 50% of 42 worst segments.
- According to the traditional approach in total 72 different hazardous locations were identified. 36 of them were identified in all three time periods, which makes up 36% of 42 worst segments.

Tab. 4 - Average values of descriptive statistics for 42 worst segments



Time period		*Mean*	*Std. deviation*	*Coefficient of variation*
2007 – 2009	EB estimate	1.161	0.507	0.437
	Recorded crashes	2.500	1.943	0.777
2008 – 2010	EB estimate	1.088	0.383	0.352
	Recorded crashes	2.452	1.789	0.730
2009 – 2011	EB estimate	1.002	0.312	0.311
	Recorded crashes	2.333	1.808	0.775

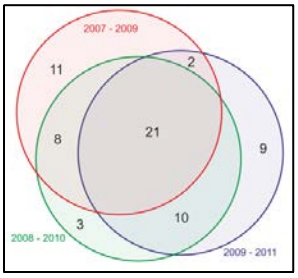


Fig. 2 - Schematic representation of sets of locations identified in the three time periods according to the EB approach

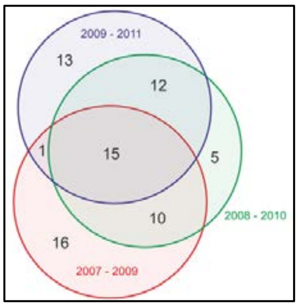


Fig. 3 - Schematic representation of sets of locations identified in the three time periods according to the traditional approach

The results confirm the original hypothesis stating that the EB approach identifies temporarily stable locations; while locations identified using the traditional approach vary significantly more in time. Identified segments are displayed on maps in Figure 4 and Figure 5.

In order to assess its predictive performance model may be used to predict in the future [15]. In this regards model from 2007 – 2009 was used to identify hazardous locations in time period 2009 – 2011. As a result 20 hazardous locations were identified, as compared to 21 identified from original model for time period 2009 – 2011. Related to 42 worst segments, an "extended" model identified 48%, as compared to original 50%. Proximity of these results confirms acceptable predictive performance of the model.

6. Discussion and conclusions

The study compared two methods of the identification of hazardous road locations: the traditional approach based on traditional definition and the prediction models refined by the EB adjustment. The method used for the comparison was based on the percentage of agreement in identified locations. Several time periods could be used for this purpose. In the study three overlapping time periods were used; overlaps are not an issue, since the objective was relative comparison of both approaches using the same time periods. The results showed that the predictions models with EB adjustment ("EB approach") offer the results which are more stable in time compared to the traditional approach. From the perspective of the road administrator, the EB approach provides more reliable results – it offers a better chance of identifying hazardous locations which are stable in time, i.e. they are more likely to be true hazardous road locations, and not being identified due to random variations only.

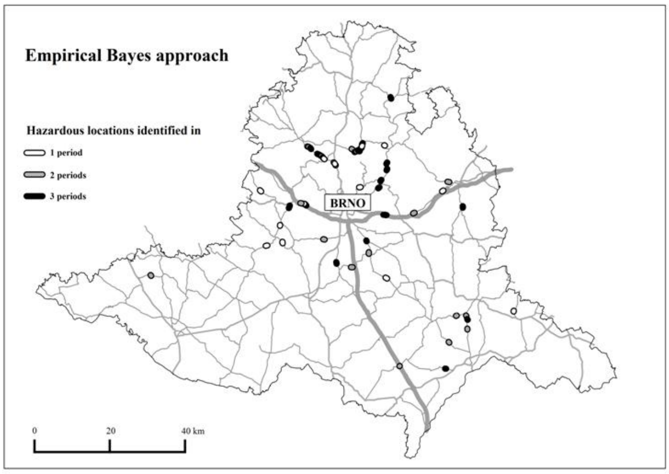


Fig. 4 - Hazardous locations identified by the EB approach in three time periods

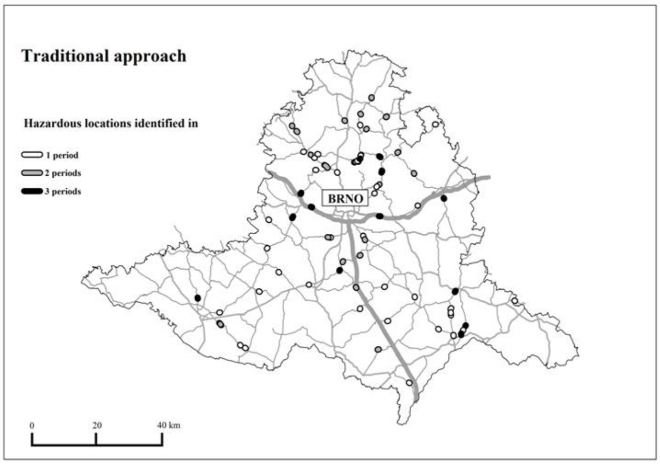


Fig. 5 - Hazardous road locations identified by the traditional approach in three time periods

This is apparent also from maps in Figure 4 and Figure 5: spatial dispersion of the hazardous road locations is much larger in the latter (obtained by the traditional approach) compared to the former (obtained by the EB approach).

The validity of the identification should be further checked by an on-site investigation, where the identified segments are compared to similar locations, which are not hazardous road locations. The percentage of true hazardous road locations between the locations, which were identified in all three periods, should turn out to be higher than in the case of locations identified only in one time period.

The study shows that the use of the crash prediction models is a step forward in road safety management. Identification of hazardous road locations (e.g. road segments as in the study) and their priority ranking should be the first step in management and maintenance processes.

Acknowledgements

Authors are grateful to professors Salvatore Cafiso, Rune Elvik and Bhagwant Persaud for valuable consultations. The study was conducted with the support of the Czech Ministry of Interior's research project No. VG20112015013 "Identification and treatment of high risk road spots and sections" (IDEKO) and also with support of the Czech Ministry of Educations, Youth and Sports' research project No. CZ.1.05/2.1.00/03.0064 "Transport R&D Centre".

References

1. American Association of State Highway and Transportation Officials (AASHTO). 2010. *Highway Safety Manual. First Edition*. Washington: AASHTO.
2. Andres, J., Mikulík, J., Rokytová, J., Hrubý, Z., Skládaný, P. 2001. *Metodika identifikace a řešení míst častých dopravních nehod*. Brno: Centrum dopravního výzkumu, v.v.i.
3. Cafiso, S., Di Graziano, A., Di Silvestro, G., La Cava, G., Persaud, B. 2010. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. *Accident Analysis and Prevention*, 42 (4), pp. 1072-1079.
4. Elvik, R. 2006. New Approach to Accident Analysis for Hazardous Road Locations. *Transportation Research Record*, 1953, pp. 50-55.
5. Elvik, R. 2007. *State-of-the-Art Approaches to Road Accident Black Spot Management and Safety Analysis of Road Networks*, report 883 [online]. Available from: http://www.toi.no/getfile.php/Publikasjoner/T%D8I%20rapporter/2007/883-2007/883-2007-nett.pdf
6. Elvik, R. 2008. Comparative Analysis of Techniques for Identifying Locations of Hazardous Roads. *Transportation Research Record*, 2083, pp. 72-75.
7. Elvik, R., Høye, A., Vaa, T., Sørensen, M. 2009. *The Handbook of Road Safety Measures. Second Edition*. Emerald.
8. Elvik, R. 2010. *Assessment and applicability of road safety management evaluation tools: Current practice and state-of-the-art in Europe*, report 1113 [online]. Available from: http://www.toi.no/getfile.php/Publikasjoner/T%D8I%20rapporter/2010/1113-2010/1113-2010-elektronisk.pdf
9. European Road Safety Observatory 2009. *Road Safety Management* [online]. Available from: http://ec.europa.eu/transport/road_safety/specialist/knowledge/pdf/road_safety_management.pdf
10. Hauer, E. 1997. *Observational Before-After Studies in Road Safety*. Bingley: Emerald.
11. Hauer, E. 2001. Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation. *Accident Analysis and Prevention*, 33 (6), pp. 799-808.
12. Hauer, E., Harwood, D. W., Council, F. M., Griffith, M. S. 2002. Estimating Safety by the Empirical Bayes Method - A Tutorial. *Transportation Research Record*, 1784, pp. 126-131.
13. Hauer, E. 2004. Statistical Road Safety Modeling. *Transportation Research Record*, 1897, pp. 81-87.
14. Persaud, B. N., Lyon, C., Nguyen, T. 1999. Empirical Bayes procedure for ranking sites for safety investigation by potential for safety improvement. *Transportation Research Record*, 1665, pp. 7-12.
15. Reurings, M., Janssen, T., Eenink, R., Elvik, R., Cardoso, J., Stefan, C. 2005. *Accident prediction models and road safety impact assessment: a state-of-the-art. RIPCORD-ISEREST project deliverable D2.1* [online]. Available from: http://ripcord.bast.de/pdf/RI-SWOV-WP2-R1-State_of_the_Art.pdf
16. Šenk, P., Ambros, J., Pokorný, P., Striegler, R. 2012. Use of Accident Prediction Models in Identifying Hazardous Road Locations. *Transactions on Transport Sciences*, 5 (4), pp. 223-232.
17. Striegler, R., Valentová, V., Pokorný, P., Ambros J., Šenk, P., Janoška, Z. 2012. *Identifikace kritických míst na pozemních komunikacích v extravilánu: metodika provádění*. Brno: Centrum dopravního výzkumu, v.v.i.
18. Turner, S., Wood, G. (2010). *Accident Prediction Modelling Down-under: A Literature Review* [online]. Available from: http://www.beca.com/services/~/media/publications/technical_papers/accident_prediction_modelling_downunder.ashx