European Transport \ Trasporti Europei (2023) Issue 91, Paper n° 6, ISSN 1825-3997

https://doi.org/10.48295/ET.2023.91.6

Multivariate Regression of Road Segments' Accident Data in Italian Rural Networks

Nicola Baldo[1]*, Valentina Indri[1], Fabio Rondinella[1], Fabiola Daneluz [1]

[1]Polytechnic Department of Engineering and Architecture (DPIA), University of Udine, Via del Cotonificio 114, 33100 Udine, Italy

Abstract

Increasing traffic flows on road infrastructures and the associated comfort and safety problems have led to an increased risk of accidents for road users. To take the proper corrective actions, it is fundamental to analyze the accident phenomenon in all its aspects. The purpose of the current paper was the development of an accident prediction model for rural road segments of Friuli-Venezia Giulia (FVG) Region. The model predicts the accident frequency as a function of Annual Average Daily Traffic (AADT), segment length, and both geometrical and environmental features related to the targeted road segment. The procedure is based on the Empirical Bayes (EB) method. The statistical model used to express the road segments' safety was the multivariate regression structure of the Safety Performance Functions. Results of a CURE plots analysis verified that the model is highly reliable in predicting the accident dataset for AADT up to 12500 vehicles per day.

Keywords: Road networks; Accident analysis; Prediction model; Road safety.

1. Introduction

The health of a community is based both on its environment safety and the quality of its members relationships. In this context, mobility plays a crucial role in terms of personal safety, public health, and environmental consequences. The continuous growth in travel needs over time has led to an exponential increase in the demand for both individual and collective transport. Road infrastructure has been part of this evolution as road transport is the most widely used for moving people and goods. In particular, the increase in traffic leads to an increased risk of accidents on road infrastructures, that is more pronounced if these are unable to handle the volume of traffic and/or are poorly designed or maintained. Road system interacts with the environment by means of the "Road-Vehicle-Driver" trinomial (AASHTO, 2014). Various aspects of this road system must be considered, ranging from problems related to environmental impact to purely operational aspects.

In 2001, the European Union (EU) published an action program intended to reduce the number of accidents with the goal of a 50% reduction by 2010 and this goal has been

* Corresponding author: Baldo Nicola (nicola.baldo@uniud.it)

fully achieved. In 2010, the EU renewed its effort in improving road safety by setting a new goal: a 50% reduction in road deaths by 2020. Comparing 2018 with 2010 (the baseline year for road safety), fatalities were reduced by 21% in Europe and 19% in Italy. From 2001 to 2010, in Friuli Venezia Giulia (FVG) (CMRSS, 2016), there was a reduction in road fatalities (-50.2%) higher than the national average (-42.0%), whereas from 2010 to 2018 there were a reduction of -25.2% and -19%, respectively. Moreover, from 2010 to 2018, the fatality rate per 100 accidents decreased from 2.6 to 2.3 deaths in the region, whereas the national average stayed basically the same (1.9).

The most serious accidents occurred on highways and rural roads. In order to achieve the objectives set, along with identifying strategies to remove or mitigate risk factors, in FVG efforts are focused on improving the road safety management system.

2. Methodology

To evaluate the effectiveness of a mitigation intervention, it is necessary to organize the analysis into two steps: estimating the expected accident frequency and predicting the accidents number. These problems can be solved by means of the Empirical Bayes (EB) approach (Hauer et al., 2002). It allows the regression-to-the-mean phenomenon to be solved and the number of accidents to be predicted for each reference period. To implement this methodology, it is essential to define a reference population, i.e., a group of sites that share the same safety characteristics with the analyzed road.

For each location j, assigning k_j and K_j as the predicted and recorded accidents in the reference period respectively, and denoting K as the accidents number occurred at the case-study site, a subset of the population consisting of the roads where exactly K collisions occurred and whose mean and variance are $E = \{k|K\}$ and $VAR = \{k|K\}$ can be built. Since the case-study site has the same characteristics and the same accidents number as those belonging to the subgroup, the number of predicted accidents k is also expected to be coincident.

Accidents are rare and unpredictable events so, for a given site, the accident frequency naturally fluctuates over time. For this reason, the average "short-term" accident frequency can significantly change from the "long-term" one. This trend is known as "Regression-to-Mean" (RTM) and it is also associated with the high probability that a low-frequency period will be followed by a high-frequency one. Neglecting the RTM phenomenon can lead to errors known as "RTM bias" or "selection bias". For this reason, it is advisable to conduct analyses based on accident frequency data observed over a long period of time (AASHTO, 2014).

To perform a statistical evaluation of safety performance, it is necessary to acquire data on accidents that occurred on the road network under investigation, such as accident type, date and time of the event, weather and pavement surface conditions, number of vehicles involved and number of injuries and/or fatalities. For this purpose, FVG Region has established the Regional Road Safety Monitoring Centre (CRMSS, 2016) aimed at collecting and processing accidents data recorded by the Police and linking them to the relevant health databases. This has resulted in an Integrated Monitoring System called MITRIS.

Once the site to be investigated has been located, it was necessary to select the accidents occurred in the surrounding area; to do this, the partitioning into arcs and nodes is used. It was also necessary to identify the road segments proper length in order to model the case-study accident phenomenon. Excessive length could result in a very low concentration of accidents; conversely, using a too short length, could result in most

segments showing zero accidents (not very significant for the analysis purposes). It was decided to use a constant length of the road segments equal to 1000 m.

As shown in figure 1, the focus was on rural road segments covered by the TrIM station network, under the responsibility of the company FVG Strade S.p.A, and characterized according to Annual Average Daily Traffic (AADT). FVG model proposed here is presented as a case study, but the methodology used can be easily extended to other investigation sites.

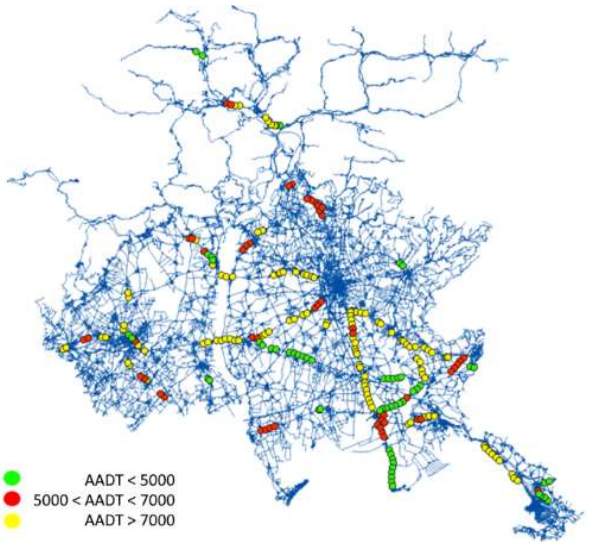

Figure 1: Road segments investigated.

3. Modeling

3.1 Background

Accident Prediction Models (APMs) are reliable tools for performing quantitative safety assessment; they are supported by mathematical equations that allow road engineers and national road authorities to correlate the number of accidents expected at a specific site with its geometric and environmental characteristics (Yannis et al., 2016). They also allow preventive interventions evaluation to be performed by estimating their safety benefits and their execution priorities. The APM proposed by the Highway Safety Manual (HSM) is a widely used approach to estimate the average accident frequency of a specific network or site. According to this procedure, the baseline accident frequency of a segment can be estimated by applying a regression model to accident data observed on a large number of sites.

Regression equations (also called Safety Performance Functions – SPFs (Baldo and Miani, 2020)) are presented for several baseline conditions that account for the specific geometric design and traffic control characteristics of the baseline site. Under these conditions, the accident frequency depends only on the AADT and segment length (AASHTO, 2014).

In order to account for the different geometric design and traffic control characteristics of the case-study site, Crash Modification Factors (CMFs) are introduced. Such multiplying factors are used to adjust the accident frequency with regard to the differences between the baseline situation and the case-study. They are usually identified by estimating the impact that an intervention is expected to have on safety by means of a

Before-After approach (La Torre et al., 2019). Many factors affect the accident frequency: climate, driver behavior, accident reporting method, etc. Therefore, a calibration factor (C) is usually implemented to improve HSM model prediction accuracy when applied to different road networks. The formulation of an APM by means of a SPF consists in the development of a statistical model that provides estimates of the average accident frequency of a unit (road segment, crossing, etc.) as a function of its features (traffic, geometry). Among the different statistical techniques, multivariate analysis has assumed a relevant importance in the accidents analysis.

The first models developed with this technique were based on multiple linear regression while the approach currently in use is based on the generalized linear model (GLM) technique, which allows linear modeling to be extended to stochastic variables that are not normally distributed with a constant variance. Within stochastic processes, the accidents frequency at a specific location is generally treated as a random variable distributed according to the Poisson distribution and with a negative binomial error structure (Roque and Cardoso, 2014).

The proposed accident prediction model consists of a basic SPF (Donnell et al., 2014) developed as a function of AADT combined with CMFs. It is developed using a generalized linear modeling approach, assuming an error structure described by a negative binomial distribution. The full model was then calibrated based on the total number of accidents observed in the full dataset.

3.2 APM modeling

APMs were developed according to the following scenario:

Road type	Single carriageway
Site type	Rural section (excluding crossings, junctions, etc.)
Section	One lane for each direction
Accident type	All accidents
Accident severity	Accidents with fatalities, with injuries, and with only property damage

The model to estimate the predicted average accident frequency ($N_{predicted}$) was developed following the HSM approach, combining both good flexibility and adaptability to local conditions with reliable accident prediction capabilities. It consists of a base SPF multiplied by CMFs and a calibration factor C, according to the equation:

$$N_{predicted,x} = N_{spf,x} \cdot (CMF_{1x} \cdot CMF_{2x} \dots \cdot CMF_{ix}) \cdot C_x \quad (1)$$

where:

x	Specific road segment
$N_{predicted,x}$	Expected average accident frequency for the case-study road section (referred to a specific year)
$N_{spf,x}$	Predicted average accident frequency for the case-study road section determined for base conditions
CMF_{ix}	Crash modification factors specific to SPF for road section
C_x	Calibration factor to adjust the SPF to local conditions for road section

Since the dataset size was relatively limited, focus was on three main parameters: road width, road type, and road hazard level. Sites analyzed are straight sections of main rural roads belonging to the FVG road network: they all have a homogeneous length of 1 km. The key factors behind the present study are:

Accident Frequency, defined as the average number of accidents that occurred during the 10-year time period. It is calculated as:

$$f_j = \frac{N_j}{n} \tag{2}$$

where:

f_j	Accident frequency of site j (accidents/year)
N_j	Number of accidents occurred at the j-th site during the analysis period
n	Number of years composing the analysis period

Accident Rate, defined as the ratio between the number of accidents recorded on a given section in a specific time period and any exposure measure. In road safety studies, the most commonly used exposure measure is traffic volume. Therefore, the accident rate is calculated as:

$$R_j = \frac{f_j \, 10^6}{365{,}25 \, n \, L_j \, Q_j} \tag{3}$$

where:

R_j	Accident rate of j-th site (accidents/million vehicles per km)
L_j	Length (km) of j-th road section belonging to the reference population
Q_j	AADT value of j-th site (vehicles/day)

Annual Average Daily Traffic (AADT), obtained from the combination of heavy and light vehicles number passing through a given road section for a year divided by 365 days.

Shoulder size, defined as the part of the road free from any obstacle (vertical signs, reflectors, restraint devices), between the roadway edge and the nearest of the following longitudinal elements: sidewalk, central reservation, embankment, inner edge of the bump, upper edge of the embankment.

Road Hazard Rating (RHR), evaluated on a scale of seven possible levels, represent the different scenarios that can occur beyond the end of the roadway. They must be considered because they also affect accidents from the driver's psychological viewpoint.

3.3 Classes identification

By statistically analyzing the available dataset, standard conditions were identified using frequency distribution plots for each analysis variable. Road sections where the

CMF is within ± 10% of the calculated maximum value are considered standard and are used for baseline model development. After defining the standard conditions for each analyzed variable and evaluating the corresponding CMF values, a correction coefficient was calculated to adjust the SPFs for conditions different from the HSM baseline (AASHTO, 2014) according to:

$$\prod_{i=1}^{n} CMF_i \tag{4}$$

Parameters described in the previous paragraph were then calculated, defining homogeneous classes of data so that road segments belonging to the same class would have the same characteristics.

To define the classes related to shoulder size, it was determined how often a certain value of shoulder occurs in the available data; it was observed that the frequencies are nearly in accordance with a Gaussian distribution. Subsequently, road segments evaluated as critical were identified for each class by means of the following procedure. Road segments were sorted in descending order with respect to accident frequency value within each reference population. A maximum accidental frequency was established for each reference population. In this way, by comparing it with the accidental frequency of the investigated sites, it became possible to identify those that showed too high and therefore not acceptable risk levels.

For each reference population, since no nationally collected frequency values are available, critical accidental frequency can be calculated according to:

$$f_{crp} = f_{mp} + K\sqrt{\frac{n_e\, f_{mp}\, 10^6}{365{,}25\, n\, \sum_{j=1}^{n_e}(L_j\, Q_j)}} + \frac{n_e\, 10^6}{730{,}5\, n\, \sum_{j=1}^{n_e}(L_j\, Q_j)} \tag{5}$$

where:

f_{crp}	Critical accident frequency of the reference population the site belongs to
f_{mp}	Average accident frequency of the reference population (accidents/year)
K	Probability constant equal to: 1,036 for an 85% confidence level; 1,282 for a 90%; 1,645 for a 95%; 2,326 for a 99%;
n_e	Number of elements belonging to the reference population

For each investigated site, the above-mentioned factors were determined by calculating the overall mean and then the critical indicator assuming a Poisson distribution and a confidence level of 85%.

3.4 Model fitting

The baseline SPF was identified by means of negative binomial regression models estimated using a generalized linear model technique (Dong et al., 2014). In fact, according to the HSM, the response variable is assumed to follow a negative binomial

distribution, allowing for data over-dispersion, whereas link function is assumed to be logarithmic.

To correlate the accident frequency with the characteristics outlined in the above paragraphs, the accident frequency for sites with baseline conditions ($N_{spf,x}$) has been identified. To do this, a statistical data analysis software, called R, was employed. It returned an intercept value and a coefficient to be applied to the AADT ($b_{AADTsegments}$) equal to $1{,}430$ and $8{,}557 \cdot 10^{-5}$, respectively.

Using Equation 1 and the results obtained from the software, the accident frequency related to average daily traffic for sites with baseline conditions (grassed shoulder with a hazard level of 3 and less than 1 meter wide) is calculated as:

$$N_{spf,x} = e^{1{,}430+8{,}557 \cdot 10^{-5} \cdot AADT} \tag{6}$$

The coefficient of determination R^2 was determined to evaluate the statistical accuracy of the results. In addition, Cumulative Residuals (CURE) plots were generated to visually figure out if and where the model overestimates or underestimates the actual data. By sorting data according to traffic volumes and calculating for each site the difference between accidents recorded and those predicted, cumulative residuals trend can be plotted. An increasing graph refers to areas where the observed accidents are greater than those predicted, therefore where the model equation underestimates the results. Conversely, a decreasing graph refers to areas where there is an overestimation of the results. A fluctuation of the CURE plot is acceptable as long as it is limited within the equation curves:

$$\pm 2\sigma_s'(i) = 2 \cdot \left(\pm \sigma_s(i) \sqrt{1 - \frac{\sigma_s^2(i)}{\sigma_s^2(n)}} \right) \tag{7}$$

where $\sigma_s(i)$ and $\sigma_s(n)$ are the standard deviations of i-th and n-th element (the last in terms of AADT in ascending order), respectively.

3.5 Crash Modification and Calibration Factors definition

To account for site-specific conditions that differ from the base conditions, CMFs have been defined. CMF estimation involves calculating the ratio between averaged accident frequency of the segments belonging to the base class and that of the segments belonging to the i-th class analyzed.

$$CMF = \frac{N_{with}}{N_{without}} \tag{8}$$

where:

N_{with}	averaged accident frequency of the segments belonging to the base class
$N_{without}$	averaged accident frequency of the segments belonging to the i-th class analyzed

When HSM model is transferred to road networks different from the one which the model was developed for, a calibration factor C is usually considered. It is defined as the ratio between the observed accidents (N_{obs}) and the accidents predicted by the uncalibrated model (N_{pred}).

$$C = \frac{\sum_{i=1}^{allsites} \sum_{j=1}^{allyears} N_{obs,ij}}{\sum_{i=1}^{allsites} \sum_{j=1}^{allyears} N_{pred,ij}} \quad (9)$$

The calculated calibration factor is equal to 0.998 (this value may change for a different network). Since it is slightly less than 1, the model tends to slightly overestimate the accident frequency. Then, the model was validated by applying calibration and crash modification factors to all road sections.

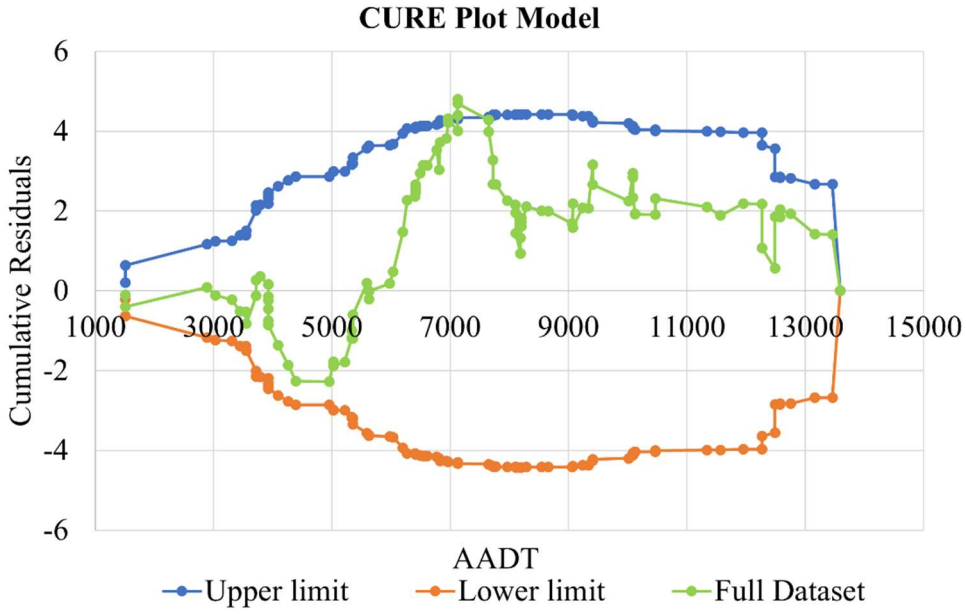

Figure 2: Cumulative residuals plot model.

For an AADT between 1000 and 5000 vehicles per day, the model overestimates accident frequency; between 5000 and 7000 the trend is reversed; above this threshold, predicted frequencies are again higher than the observed (Figure 2).

Figure 3 shows a correlation value between the predicted data of 0.309, consistent with the low correlation shown between the observed data and the relatively small sample size.

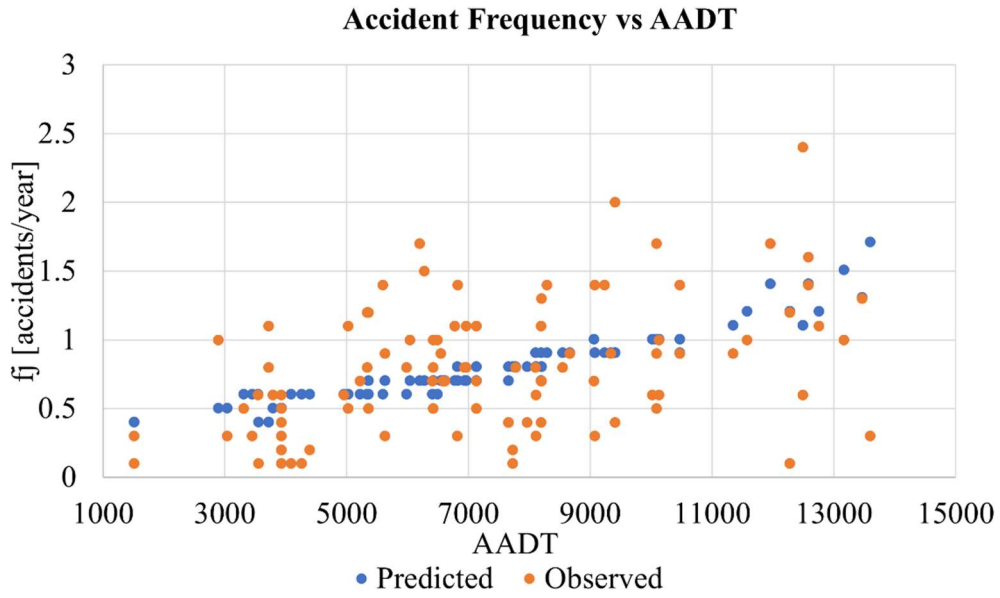

Figure 3: Accident frequency vs AADT.

Figure 4 shows the observed and predicted accident frequencies for the defined model; the model underestimates the accidents number in sections where the observed arrests are higher than the average values.

Finally, Figure 5 shows a non-random residuals distribution, typical of negative binomial models in which residuals increase as predicted accidents increase.

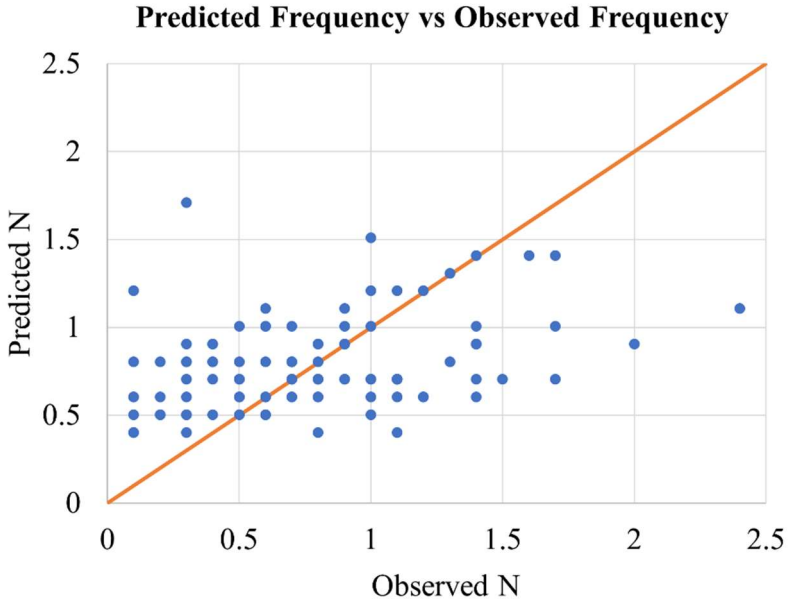

Figure 4: Predicted frequency vs observed frequency.

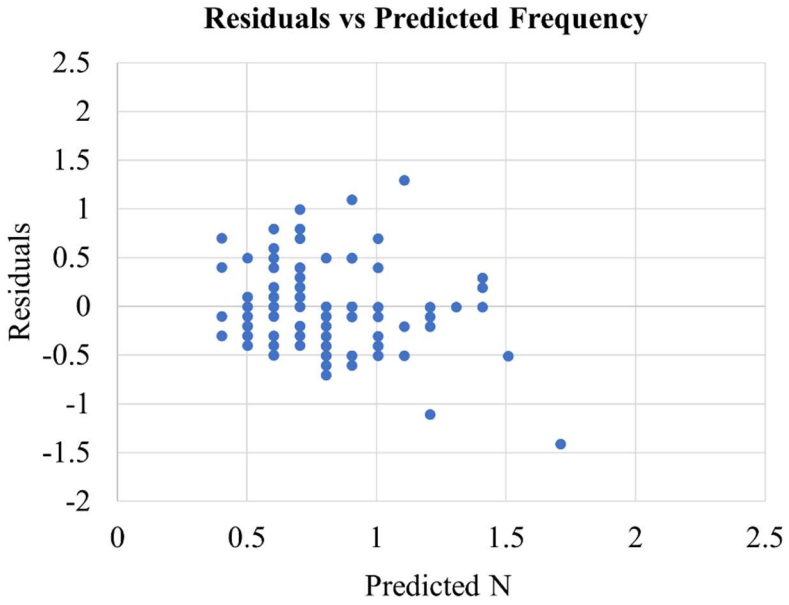

Figure 5: Residuals distribution.

4. Conclusions

The methodology developed in the current study allows the accident rate of a main rural road with specific geometric features to be estimated. The analysis has been characterized by the identification of each road segment main properties, focusing on accident frequency and accident rate. This allowed different sites to be classified according to their hazard level. CURE plots and correlation coefficients obtained confirm that the model correctly describes the available dataset up to AADT values of 12500 vehicles/day. This model has been applied to real values coming from FVG road network and provides a useful tool to perform safety assessments. It allows critical sites, i.e., those with a high accident frequency with respect to traffic volumes, to be identified. The model can also be used for new sites preliminary assessment, as well as a decision-making tool among multiple realization possibilities. A limitation of this approach is represented by the randomness of the accident data. Considering the same factors such as traffic flow, geometric characteristics and type of users, different accidents values can potentially be found. Finally, the model does not account for the human factor represented by driver behavior; however, this feature can be analyzed for further investigations.

References

AASHTO (2014) *Highway Safety Manual (HSM)*, American Association of State Highway and Transportation Officials, Washington DC.

CRMSS (2016) *Relazione sullo stato dell'incidentalità in Friuli-Venezia Giulia,* Centro Regionale di Monitoraggio della Sicurezza Stradale, Palmanova.

Hauer, E., Harwood, D. W., Council, F. M., Griffith M. S. (2002) "Estimating Safety by the Empirical Bayes Method: A Tutorial", *Transportation Research Record: Journal of the Research Board* 1784, pp. 126-131.

Yannis, G., Dragomanovits, A., Laiou, A., Richter, T., Ruhl, S., La Torre, F., Domenichini, L. Graham, D., Karathodorou, N., Li, H. (2016) "Use of Prediction

Model in Road Safety Management – An International Inquiry." *Transportation Research Procedia* 14, pp. 4257-4266.

Baldo, N., Miani, M. (2020) "Safety Performance Functions for Road Intersections in the Friuli Venezia Giulia Region", *European Transport – Trasporti Europei* 77, pp. 1-12.

La Torre, F., Meocci, M., Domenichini, L., Branzi, V., Tanzi, N. (2019) "Development of an Accident Prediction Model for Italian Freeways", *Accident Analisys and Prevention* 124, pp. 1-11.

Roque, C., Cardoso, J.L. (2014) "Investigating the Relationship Between Run-Off-The-Road Crash Frequency and Traffic Flow Through Different Functional Form", *Accident Analisys and Prevention* 63, pp. 121-132.

Donnell, E., Gayah, V., Jovanis, P. (2014) "Safety performance functions", *Pennsylvania Department of Transportation*, Bureau of Planning and Research, pp. 1-42.

Dong, C., Clarke, D. B., Jan, X., Khatak, A., Huang, B. (2014) "Multivariate Random-Parameters Zero-Inflated Negative Binomial Regression Model: An Application to Estimate Crash Frequencies at Intersections", *Accident Analisys and Prevention* 70, pp. 320-329.