Modeling traffic safety at urban four leg-signalized intersections

O. Giuffrè[1], A. Granà[1], T. Giuffrè[2] & R. Marino[1]
[1]*Department of Civil, Environmental, Aerospace and Materials Engineering, Palermo University, Italy*
[2]*Faculty of Engineering and Architecture, Kore University, Enna, Italy*

Abstract

According to the state-of-the-art of the methodologies, the development of safety performance functions (SPFs) for road sections and intersections requires the employment of statistical models to predict expected crash frequencies on the basis of traffic volumes and site characteristics to be surveyed and used as input to models. Nevertheless, literature reports several studies on issues deriving from data features or methodological approaches that may invalidate the efficiency of the models and the accuracy of the estimates. Drawing inspiration from the above mentioned considerations, the objective of this study is to develop safety performance functions for a sample of urban four leg-signalized intersections on the basis of 8 years of crash data in Palermo, Italy. Applications of the Conway-Maxwell model are presented for analyzing traffic crash data exhibiting underdispersion. Results comforted authors on the potential of the Conway-Maxwell model to account for dispersion phenomenon and to provide a good goodness-of-fit, as long as the temporal correlation in the data is not considered. In this regard, the GEE model, incorporating the time trend, allowed to gain further methodological insights compared to models that do not accommodate the temporal correlation in crash data.

Keywords: transportation safety, crash analysis, urban signalized intersections, safety performance function.

1 Introduction

Safety performance functions (SPFs) are essentially mathematical equations explaining interactions between road elements and crash frequencies. The

WIT Transactions on The Built Environment, Vol 130, © 2013 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/UT130591

development of SPFs for road sections and intersections requires the employ of statistical models to predict expected crash frequencies on the basis of geometric and traffic-related explanatory variables to be surveyed and used as input to models; (see e.g. [1–4]). Literature reports several studies on key issues deriving from crash data features or associated with the modeling of traffic crashes that may invalidate the efficiency of the models and the accuracy of the estimates; (see e.g. [5]). Some studies have focused particularly on the issue of the appropriate model form, i.e. the functional form linking the dependent variable to the explanatory variables; this is because the result of a regression model can be dependent on the choice of model function [6]. Poisson distribution models are usually the first choice in the modeling of traffic crashes because of the non-negative, discrete and random features [7]. Poisson regression model, however, has only one distribution parameter, requiring that the mean and the variance of the crash frequency are identical. The applicability of the Poisson models is therefore limited: in most of cases, the variance of the crash frequency exceeds the mean and crash data are overdispersed; in a few cases, the variance can result less than the mean and data exhibit underdispersion. In order to relax the Poisson assumption of equidispersion, quasi-likelihood methods represent a potential solution. Thus a quasi-Poisson distribution can be used to model crash data: the mean is the same of the Poisson mean; the variance is now a function of the mean: $v_{tj} = \phi\mu_{tj} = (1 + \alpha)\mu_{tj}$ where α is named the dispersion parameter. In the case of under-dispersion α <0 (and 0< ϕ <1); in the opposite case ($\alpha > 0$ and $\phi >$ 1), data are overdispersed. Several authors have addressed the overdispersion issue by using the Negative Binominal regression model; (see e.g. [3, 8]). Properties of the traditional NB models have been illustrated by Cameron and Trivedi [9]. Lord and Mannering [5] by reviewing and assessing some alternative methods for the statistical analysis of crash data have explained that it cannot be used in the case of underdispersion; moreover, the dispersion parameter of the negative binomial model is incorrectly estimated when data are characterized by small sample size and low sample mean [10]. It should be also said that, in modeling underdispersion, traditional count-data models can produce incorrect parameter estimates; moreover, options in selecting the appropriate distribution are more limited [5]. Among the latter, the Conway-Maxwell-Poisson (COM-Poisson) distribution, introduced in 1962 by Conway and Maxwell, has been recently re-introduced by statisticians to model count data characterized by either over- or under-dispersion and evaluated in the context of a Generalized Linear Model (GLM); (see e.g. [5, 11–13]). Since it was revalued, it has been further developed in several directions and applied in multiple fields [14]. Concerning other issues associated with crash-frequency data, it should be noted that count data often consist of observations over several time periods; thus, with many years of data, it is necessary to account for the year-to-year variations in crash counts because of the influence of factors that can change every year. This can create a temporal correlation that affects the reliability of the SPF estimate obtained through traditional model calibration procedures [15]. Generalized Estimating Equations (GEEs) overcome this problem, incorporating together dispersion and temporal correlation; a GEE model, indeed, can estimate the parameters of a generalized linear model with a possible unknown correlation between outcomes [16, 17]. Literature refers on several applications by GEE models; see e.g. [18–20]. Recently there is much debate on another matter arising under modeling of traffic crashes and concerning the crash model transferability; the reader is invited to consult (e.g. [21]).

Drawing inspiration from the above considerations, this paper summarizes the results of a research aimed at developing SPFs for a sample of urban 4 leg-signalized intersections on the basis of 8-years of crash data in Palermo, Italy. Applications of the COM-Poisson model are presented for handling underdispersion as exhibited by crash data. This dataset, indeed, showed to exhibit underdispersion when models linking crash data to different explanatory variables were estimated. Results confirmed the potential of the COM-Poisson model to account for dispersion phenomenon and provided a good goodness-of-fit. However, COM-Poisson regression models, despite their benefits, have disadvantages in terms of model estimation, also as a result of difficulties in accounting for temporal correlation in the data; (see [22]). In order to incorporate the time trend in crash count data, a different approach based on GEEs was also applied in the developing of SPFs. In this regard, the GEE model was able to give interesting methodological insights compared to models that do not accommodate the temporal correlation in crash data. Explanations on the modeling approach for the SPFs development will be reported in the following sections; but first, a brief introduction on the characteristics of crash data at the intersections that were examined for developing SPFs will be presented.

2 Crash data analysis

In Palermo City, Italy, similarly to other cities, a large number of signalized intersections are considered sites with promise for safety and operational improvements. To address the screening of these intersections, SPFs, predicting expected crash frequencies on the basis of traffic volumes and site characteristics, need to be developed and used. Despite the great deal of efforts associated with the data collection process due limitations on data availability in computerized records, crash data from a sample of urban 4-leg signalized intersections were directly collected from reports available at the Municipal Police Force in Palermo, Italy. Crash data were obtained for the same time period of eight years (years 2000–2007) for which data were available for all study intersections. Crashes occurred along major and minor roads on the intersection approaches were recorded as being at the intersection if within 20 meters of the intersection center. So 558 crashes were considered at 19 urban 4-leg signalized intersections (91 percent of multiple-vehicle crashes and 9 percent of single-vehicle crashes). Figure 1 shows fatal and injury crashes at each intersection examined in years 2000–2007. Some of the intersection features that were on-field surveyed and considered directly related to the safety and operational effectiveness are shown in table 1; it shows, indeed, the percentage distribution of intersections having the road characteristics specified in table, with reference to major- and minor-roads.

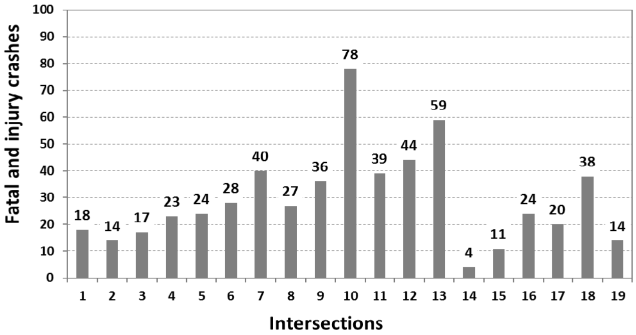

Figure 1: Fatal and injury crashes at each intersection in years 2000–2007.

Extensive traffic surveys were also carried out in 2007 and gave some information about the type of vehicles and maneuvers (through vehicles, left- and right-turning vehicles). Major- and minor-road annual average daily traffic (AADT) was estimated; AADT values of each year from 2000 through 2006 were computed using Italian vehicle registrations, as widely reported in [23].

Table 1: Percentage distribution of road elements for all intersections.

Road element		major street	minor street
Number of lanes	1 lane	21 %	21 %
	2 lanes	47 %	63 %
	3 lanes	32 %	16 %
Roadway width [m]	*w≤10 m*	16 %	26 %
	10 m <w≤15 m	58 %	58 %
	3*w> 15 m*	26 %	16 %
Permitted way system	one-way only	47 %	63 %
	two-way	53 %	37 %

Figure 2 shows mean values of crashes and the annual average daily traffic on major roads ($AADT_{major}$) and annual average daily traffic on minor roads ($AADT_{minor}$) at the surveyed intersections; values were averaged over the 8-years of observation. A study was also made to test whether data distribution tended or not to follow the shape of a Poisson distribution (or the Poisson distribution was an appropriate model for the data set). The findings were presented in an extensive way in [23].

3 Statistical models

This section focuses on the methodological approach followed to estimate SPFs for the urban 4-leg signalized intersections under examination.

The selection of explanatory variables to be used, the search for the best model form, the estimate of the regression parameters and the model validation were the main working steps.

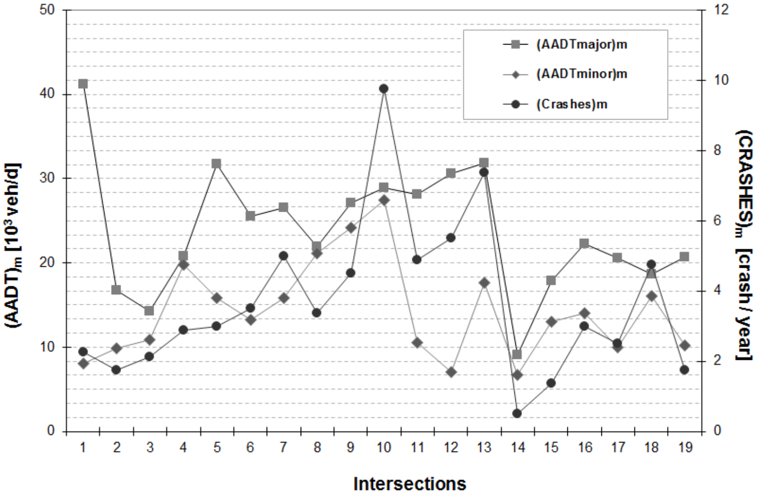

Figure 2: Mean values of $AADT_{major}$, $AADT_{minor}$ and crashes at intersections.

It should be noted here that this dataset exhibited signs of underdispersion when models linking crash data to different covariates were estimated; so, underdispersion in crash data was explored and Conway-Maxwell-Poisson model in GLM context was fitted to the data. After testing the potential of the aforementioned distribution, time trend was also examined. In order to account for correlation, quasi-Poisson models in the framework of Generalized Estimating Equations were then performed. GEEs, indeed, allow us to incorporate together dispersion and temporal correlation when a quasi-Poisson distribution is used for modeling crash data. At last, different goodness-of-fit criteria were applied to evaluate predictive performance of models and to find the model that best explains the data among all estimated models. The methods used will be introduced in the following.

3.1 Crash modeling and model validation

Different model forms were investigated considering the combinations of the variables considered to be significant. For the case study, the chief road variables are the traffic volumes at intersection approaches; besides traffic volumes, geometric variables affecting safety are related to major- and minor-road characteristics. Only variables found to be statistically significant at the 15% significance level were included in the model specification. The exploratory analysis also revealed that two functional model forms could be used to explain the relationship between crashes and the significant covariates, as follows:

$$\text{Model 1: } y_{tj} = \beta_0 \times F_{1tj}^{\beta_1} \times RW_{2j}^{\beta_2} \times e^{\beta_3 PW_{1j}}$$

$$\text{Model 2: } y_{tj} = \beta_0 \times \left(F_1 + F_2\right)_{tj}^{\beta_1} \times RW_{2j}^{\beta_2} \times e^{\beta_3 PW_{1j}}$$

where:

y_{tj} =expected number of crashes for the year *t* and the intersection *j*;
F_{1tj} =annual average daily traffic on major road for the year *t* and the intersection *j*;
$(F_1 + F_2)_{tj}$ =sum of annual average daily traffic on major- and minor-roads for the year *t* and the intersection *j*;
RW_{2j} =minor-road roadway width at the intersection *j*;
PW_{1j} =major-road permitted ways at the intersection j (PW_1=0 for one way only, PW_1=1 for two ways or more);
$\beta_0, \beta_1, \beta_2, \beta_3$ =parameters to be estimated.

First, in order to address the issue of underdispersion in the data, model 1 and model 2 were estimated considering Poisson, quasi-Poisson and COM-Poisson distributions. Model regression coefficients and the associate standard errors were estimated in GLM context assuming Poisson and quasi-Poisson distributions; in both cases, GenStat software was used. COM-Poisson model was estimated using *R* software, by means of codes arranged by Sellers and Shmueli [11]. Table 2 provides a very brief overview on COM-Poisson distribution properties as discussed in [11].

Table 2: Conway–Maxwell–Poisson distribution [11].

parameters	$\lambda > 0$, and $\nu \geq 0$	$\lambda_{tj} =$ a centering parameter [11]; $\nu =$ the dispersion parameter (ν<1 for over-dispersion; ν>1 for under-dispersion).
support	$y_{tj} \in \{0,1,2,\ldots\}$	y_{tj} = a discrete count at year t and at site j;
probability mass function	$P\left(Y_{tj} = y_{tj}\right) = \dfrac{\lambda_{tj}^{y_{tj}}}{(y_{tj}!)^{\nu} Z(\lambda_{tj},\nu}$	$Z(\lambda_{tj},\nu) = \sum_{s=0}^{\infty} \dfrac{\lambda_{tj}^{s}}{(s!)^{\nu}}$
mean	$E\left(Y_{tj}\right) = \lambda_{tj} \dfrac{\partial \log Z\left(\lambda_{tj},\nu\right)}{\partial \lambda_{tj}} \approx \lambda_{tj}^{1/\upsilon} - \dfrac{(\nu - 1)}{2\nu}$	good approximation for ν ≤ 1 or $\lambda_{tj} > 10^{\nu}$; $E(Y^{\nu})=\lambda$
variance	$var\left(Y_{tj}\right) = \dfrac{\partial E\left(Y_{tj}\right)}{\partial \log \lambda_{tj}}$	

Table 3 shows coefficient estimates and goodness-of-fit for model 1 and model 2. For an explanation about goodness-of-fit criteria (i.e. MPB, MAD and MSPE) see [24]. It should be noted that COM-Poisson coefficients are for the centering parameters $\hat{\lambda}_{tj} = exp\left(x_{tj}^{'}, \hat{\beta}\right)$, as computed by E($Y_{tj}$) in table 2; so they cannot be directly compared with Poisson/quasi-Poisson coefficients. MPB, MAD and MSPE values in table 3 show that the COM-Poisson regression fits the data better the Poisson/quasi-Poisson models for model 1: the MPB values of the COM-Poisson highlight that the model fairly estimates crashes; the MAD and the MSPE values of the COM-Poisson model, being closer to 0 than the quasi-Poisson model, highlight that the model have good prediction accuracy.

Table 3: Parameter estimates and goodness-of-fit for model 1 and model 2.

variables	Poisson		quasi-Poisson		COM-Poisson	
	model 1	**model 2**	**model 1**	**model 2**	**model 1**	**model 2**
constant	-6.56 (0.84)[a]	-6.94 (0.77)[a]	-6.56 (0.58)[a]	-6.94 (0.60)[a]	-14.28[b] (2.09)[a]	-12.41 (1.72)[a]
F_1	1.74 (0.15)	-	1.74 (0.11)	-	3.91 (0.52)	-
$F_1 + F_2$	-	1.69 (0.14)	-	1.69 (0.11)	-	3.18 (0.40)
RW_2	0.86 (0.20)	0.73 (0.17)	0.86 (0.14)	0.73 (0.14)	1.94 (0.38)	1.38 (0.28)
PW_1	0.25 (0.09)	0.22 (0.09)	0.25 (0.06)	0.22 (0.07)	0.54 (0.15)	0.39 (0.13)
ν	-	-	-	-	2.41 (0.30)	1.99 (0.25)
α	-	-	-0.53 (0.06)	-0.46 (0.06)	-	-
$MPB = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$	-0.42	0.00	-0.42	0.00	0.01	0.01
$MAD = \frac{1}{N}\sum_{i=1}^{N}\lvert\hat{y}_i - y_i\rvert$	1.35	1.09	1.35	1.09	1.03	1.09
$MSPE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$	4.47	1.98	4.47	1.98	1.80	1.97
$AIC^c = -2 \log L + 2$	490	543	490	543	455	519

[a] *standard error;* [b] *model parameters to be used for determining* $\hat{\lambda}_{t;}$ *;*
[c] *L is the maximized value of the likelihood function for the estimated model; p is the number of parameters in the statistical model; see* [25].

Conversely, MPB, MAD and MSPE values for model 2 have slight differences; so significant information about the prediction accuracy of the model is not added. In all cases, AIC values for both models show that the COM-Poisson model can be considered the best among all estimated models. Considering that data consisted of repeated measures over time, possibly correlated within an entity, the correlation within responses was also accounted for. In order to consider simultaneously both the correlation and the underdispersion in the data, only the quasi-Poisson distribution was used due to difficulties in accounting for correlation in the data through the COM-Poisson model. GEE regressions were fitted assuming that repeated observations were correlated in different ways, i.e. under different working correlation matrices.

GenStat software was used again. The GEE regression results (i.e. parameter estimations and goodness-of-fit for model 1 and model 2) for three different correlation matrices are shown in table 4; the marginal R^2_m test is now introduced and used [17].

Slight differences in MPB, MAD and MSPE can be observed for the different working correlation matrixes. However, R^2_m values provide insights on the best correlation structure. GEE regression, indeed, using different working correlation matrices (i.e. assuming that repeated observations are correlated in different ways) allows to gain a better understanding of the proper correlation structure in crash counts.

Table 4: Parameter estimates and goodness-of-fit in GEEs.

variables	independence		unstructured		7-dependence	
	model 1	**model 2**	**model 1**	**model 2**	**model 1**	**model 2**
constant	-6.59 (0.67)	-6.94 (0.82)	-6.78 (0.55)	-7.63 (0.66)	-6.97 (0.51)	-7.35 (0.63)
F_1	1.74 (0.07)	-	1.83 (0.07)		1.84 (0.06)	
$F_1 + F_2$	-	1.69 (0.09)	-	1.78 (0.08)		1.77 (0.08)
RW_2	0.87 (0.23)	0.73 (0.24)	0.82 (0.17)	0.87 (0.18)	0.89 (0.20)	0.79 (0.19)
PW_1	0.25 (0.08)	0.22 (0.10)	0.36 (0.06)	0.31 (0.08)	0.33 (0.07)	0.29 (0.08)
α	-0.53	-0.47	$-0.52^{(*)}$	$-0,45^{(*)}$	-0.52	-0.48
MPB	0.00	0.00	0.13	1.19	0.22	0.27
MAD	1.04	1.09	1.08	1.14	1.09	1.15
MSPE	1.80	1.98	1.89	2.09	1.93	2.12
$R^2_m = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$	0.70	0.68	0.68	0.66	0.67	0.65

$^{(*)}$ *mean value of the dispersion parameter varying over time in the observation period.*

4 Conclusions

The paper describes methods applied to develop SPFs for urban 4-leg, signalized intersections. Data of the case study pertain to a sample of 19 4-leg signalized intersections in Palermo, Italy, in the years 2000–2007, directly processed from Municipal Police Force reports. The development of SPFs involved the selection of explanatory variables to be used, whether and how variables could be grouped, and how variables could enter into the model to choice the best model form. With regards to the functional model form, the power function seemed appropriate for the covariate F_1 for model 1 (and F_1+F_2 for model 2) and RW_2, while the exponential function was the best form for the variable PW_1. In order to test the Poisson assumption of equi-dispersion, a quasi-Poisson regression was implemented, assuming a linear relationship between the variance and the mean. Model output exhibited underdispersion; the difficulty to handle the dispersion phenomenon by the Negative-Binomial model led us to consider the COM-Poisson model due to its flexibility in holding both over- and under-dispersed data. Results confirmed that the COM-Poisson regression model provided a good statistical performance and a better goodness-of-fit than the Poisson/quasi-Poisson models, as long as the temporal correlation in the data is not considered. In this regard, GEE quasi-Poisson model, incorporating the time trend, was then performed under different working correlation matrices; it provided interesting methodological insights through estimates of model parameters compared to the correspondent GLM models (that do not account for the temporal correlation in the data). It has to be noted that the large costs associated with the data collection process likely conditioned the estimation of model parameters due to the failure in the large-sample properties of some parameter-estimation techniques [5]. Therefore, though results can help to test the potential of COM-Poisson model in handling underdispersed data, researches should be carried out using larger sample size to confirm them. At last, applications of the COM-Poisson model in GEE context are also desirable to obtain correct estimates for model parameters accounting simultaneously both for correlation and for dispersion in the data; in this way the interest of findings may not be limited to the research field only.

References

[1] Persaud, B. & Dzbik, L., Accident prediction models for freeways. *Transportation Research Board*, **1401**, pp.55–60, 1993.

[2] Cafiso, S., Di Graziano, A., Di Silvestro, G., La Cava, G. & Persaud, B., 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, 2010.

[3] Poch, M. & Mannering, F., Negative binomial analysis of intersection-accident frequencies. *Journal of Transportation Engineering*. **122(2)**, pp. 105–113, 1996.

[4] Bauer, K. M. & Harwood, D. W., Statistical models of at-grade intersection accidents – Addendum; U.S. Department of Transportation, FHWA, Washington, DC: 2000.

[5] Lord, D., & Mannering F., The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A*. **44**, pp. 291–305, 2010.

[6] Hauer, E. & Bamfo, J., Two tools for finding what function links the dependent variable to the explanatory variables. *Proc. of the International*

Cooperation on Theories and Concepts in Traffic Safety, eds Lund University, 1997: Lund, Sweden, pp. 1–18, 1997.

[7] Hauer, E. Statistical road safety modelling. *Transportation Research Board*, 1897, pp.81–87, 2004.

[8] Miaou S. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative Binomial regressions, Accident *Analysis and Prevention*, **26(4)**, pp. 471–482, 1994.

[9] Cameron, A.C., & Trivedi, P.K., *Regression Analysis of Count Data*, Cambridge University Press; Cambridge, 1998.

[10] Lord, D., Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the Estimation of the fixed dispersion parameter. *Accident Analysis and Prevention*, **38(4)**, 751–766, 2006.

[11] Sellers, K.F. & Shmueli, G., A flexible regression model for count data. *The Annals of Applied Statistics*, **4(2)**, pp. 943–961, 2010.

[12] Lord, D., Geedipally, S. R. & Guikema, S. D., Extension of the Application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crash Data Exhibiting Underdispersion. *Risk Analysis*, **30(8)**, pp. 1268–1276, 2010.

[13] Lord, D. & Guikema, S.D., The Conway-Maxwell-Poisson model for analyzing crash data. *Applied Stochastic Models in Business and Industry*. **28(2)**, pp. 122–127, 2012.

[14] Sellers, K., Borle, S. & Shmueli, G. The COM-Poisson Model for Count Data: A Survey of Methods and Application. *Applied Stochastic Models in Business and Industry*, **28(2)**, pp. 104–116, 2012.

[15] Diggle, P. J., Heagerty, P., Liang, K.-Y. & Zeger, S. L. *Analysis of Longitudinal Data*, 2nd ed. Oxford University Press: Oxford, United Kingdom, 2002.

[16] Liang, K.Y. & Zeger, S. Longitudinal data analysis using generalized linear models. *Biometrika*, **73 (1)**, pp. 13–22, 1986.

[17] Hardin, J. & Hilbe, J., *Generalized Estimating Equations*. Chapman and Hall/CRC: London, 2003.

[18] Lord, D. & Persaud, B., Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations (GEE) Procedure. *Transportation Research Record*, **1717**, pp. 102–108, 2000.

[19] Giuffrè, O., Granà, A., Giuffrè, T. & Marino, R., Improving Reliability of Road Safety Estimates Based on High Correlated Accident Counts. *Transportation Research Record: Journal of the Transportation Research Board*, **2019**, pp. 197–204, 2007.

[20] Cafiso, S.D. & D'Agostino, C., Safety performance function for motorways using generalized estimation. *Procedia-Social and Behavioral Sciences*, **53**, pp. 901–910, 2012.

[21] Turner, S., Persaud, B., Lyon, C., Bassani, M. & Sacchi, E., International crash experience comparisons using prediction models. *Road and Transport Research*, **20(4)**, pp. 16–27, 2011.

[22] Giuffrè, O., Granà, A., Giuffrè, T. & Marino, R., Accounting for Dispersion and Correlation in Estimating Safety Performance Functions. An Overview

Starting from a Case Study. *Modern Applied Science*, **7(2)**, pp. 11–23, 2013.

[23] Giuffrè, O., Granà, A., Marino, R. & Corriere, F., Handling Underdispersion in Calibrating Safety Performance Function at Urban, Four-Leg, Signalized Intersections, *Journal of Transportation Safety & Security*, **3(3)**, pp. 174–188, 2011.

[24] Oh, J., Lyon, C., Washington, S.P., Persaud, B.N. & Bared, J., Validation of the FHWA Crash Models for Rural Intersections: Lessons Learned. *Transportation Research Record*, **1840(1)**, pp. 41–49, 2003.

[25] Akaike, H., A new look at the statistical model identification. IEEE Transactions on Automat. Control, **19(6)**, pp. 716–723, 1974.