Crash modeling for urban roundabouts: a case study

O. Giuffrè & A. Granà
Department of Civil, Environmental, Aerospace, Materials Engineering, University of Palermo, Italy

Abstract

In many cities and towns, a large number of intersections are considered sites with promise for safety and operational improvements. Several studies have been carried out in many countries to establish relationships between crashes and flow and non-flow explanatory variables, using statistical tools to investigate factors critical to road safety. Starting from a brief review of existing information and analysis on the issue, this article summarizes the findings of an exploratory analysis aimed at modeling injury crashes for a sample of urban roundabouts. The methodological path followed in this research allowed to handle issues associated with the estimation of a safety performance function, and also introduce concerns related to the crash model transferability. At last, results can supply methodological insights that may be useful in the subsequent quantifying of benefits obtainable by engineering measures aimed at enhancing traffic safety in built up areas.

Keywords: road safety, crash, roundabout, flow-only models.

1 The background

The relationship between crash frequency and traffic/geometric variables for roadway segments and intersections has been the subject of study for many years. A wide number of research efforts have examined this relationship with the purpose of determining the effect of road and intersection design on the frequency of crashes. Technique of generalized linear models (GLMs) has been recognized able to offer a soundly-based approach for analyzing this kind of data and fitting predictive crash models. Due to the nature of the occurrence of traffic crashes, the assumption of a Poisson distribution for the crash frequency in a

WIT Transactions on Ecology and The Environment, Vol 179, © 2013 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/SC130852

given time period at any one site has proven to be the best choice to model the process [1]. Crash-frequency data are non-negative integers and the assumption of a continuous dependent variable, on which the application of standard ordinary least squares regression is based, is not appropriate [2]. Assuming the Poisson model, the functional forms of relationships can be estimated using the technique of generalized linear models (GLMs) [3]. However, crash data characteristics and methodological-technical issues may impair the efficient use of the Poisson model, which thus could produce considerable bias in parameters estimates and possible erroneous inferences [2, 4]. The use of Poisson assumes that the mean and the variance of the distribution are equal; since the mean of the crash counts on the road entities and the variance are not approximately equal, the equidispersion assumption for crash-frequency data is not appropriate. Evidence suggests that crash data counts may be overdispersed (the variance exceeds the mean of the crash counts on road entities), otherwise the data may be underdispersed (the mean is greater than the variance). The Poisson model, indeed, cannot take account of overdispersion (and underdispersion); it also can be affected by the low sample-mean and small sample size bias [2]. For a comprehensive review of data and methodological issues in the statistical analysis of crash-frequency data again [2] can be seen.

The estimation of the regression coefficients of the model can be performed using standard maximum-likelihood procedures within the framework of GLMs [5, 6]. This method selects the set of values of the model parameters that maximizes the likelihood function; for discrete random variables it maximizes the probability of the observed data under the resulting distribution. Because crash counts are often considered over many years of data and thus the explanatory variables are time-varying over the same period due to the influence of factors that can change every year, the potential within-period variation in explanatory variables has to be considered. In order not to lose important explanatory information, the number of crashes of each year is considered as a single observation. However, the temporal correlation can affect the reliability of the SPF estimate obtained through traditional model calibration procedures; the likelihood function, indeed, becomes very complicated to solve [6, 7]. Generalized Estimating Equations (GEEs) overcome this problem [8]; parameter estimates from the GEEs are consistent even when the covariance structure is misspecified [9].

Several empirical relationships between crashes and flow and non-flow independent variables have been studied using statistical models to investigate factors critical to road safety; see e.g. [10, 11] for roundabouts. However, there is still some uncertainty around the influence of non-flow variables because some potentially important predictor variables can be sometime overlooked [12]. Moreover, consideration of human factors should be enhanced; differently from the traditional human-factor approach, the identification of road geometric conditions contributing to produce crashes is now useful to correct and/or remove them [4].

1.1 Purposes and objectives of the study

With no claim of being exhaustive, this paper deals with the issues associated with the estimation of a safety performance function (SPF) for a sample of urban roundabouts. Data description for a representative sample of urban roundabouts (10 of 35 total over a 5-year period) is presented in [13]; a criterion for the preliminary risk analysis by means of an infrastructural scenarios method was proposed. The focus is now on flow-only models developed for a set of empirical injury crash data derived from a survey on these urban intersections, selected in the road network of Palermo City, Italy. According to Turner *et al.* [12] the decision to explore flow-only models was made in the belief that traffic volume is found to be the most important predictor variable; in so doing, the comparison with other models, selected basing on the knowledge of the authors, was aided. The comparison was also made to highlight the presence or absence of similarities in safety experience at roundabouts regardless of the non-flow variables, which inevitably are affected by differences in design standards and practices in force in different countries of the world. Next section will summarize the methodological path followed in this research for estimating the flow-only model; then a comparison with other flow-only models for roundabouts will be presented.

2 SPF estimation

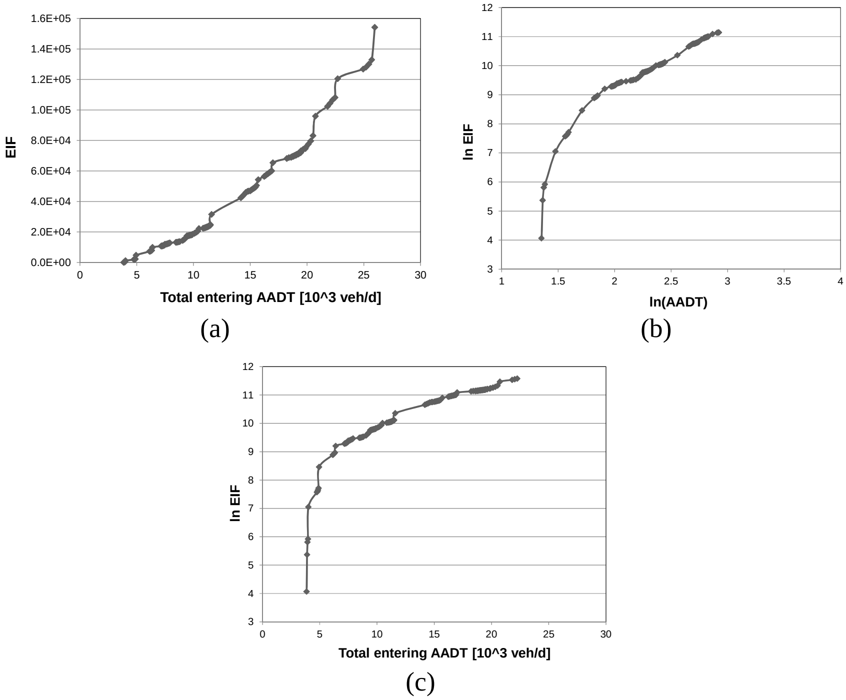
This section describes the characterization of the functional form linking the dependent variable to the explanatory variable, the methods of parameters estimation and the goodness-of-fit statistics, and the comparison with other flow-only models for roundabouts as a whole.

2.1 The selection of the functional form for the model

The Integrate-Differentiate method was proposed to recognize the suitable functional model form behind the empirical integral function (EIF) even when no pattern is discernible from the data scatterplot [14]. A link between the expected crash frequency and the total entering annual average daily traffic as explanatory variable can be derived from the exploratory data analysis; that is one can recognize the integral function by examining the empirical integral function and thus the functional form of the explored relationship follows as the derivative of the integral function [14].

Figure 1 shows the functional forms among these investigated: a) the EIF as a function of the total entering AADT; b) the ln(EIF) as a function of ln(AADT); c) the ln(EIF) as a function of the total entering AADT.

The results revealed that for the roundabouts under examination the explored relationship could be described both by the power function and the exponential function; however, the former was assumed as the functional model form in analogy to international crash prediction models for roundabouts afterwards considered for comparison purposes (see section 2.3).



(a)

(b)

(c)

Figure 1: Functional model form: (a) EIF for data sample; (b) power form; (c) exponential form.

2.2 Baseline regression flow-only models for roundabouts as a whole

Since crash data are non-negative integers, Poisson regression model was considered as the starting point. In general, the observed number of crashes at a site *i* is y_i, where y_i is assumed to be Poisson distributed about a mean of μ_i; the latter in turn is assumed to be proportional to the length of the time period T_i. The Poisson regression model specifies that the probability of a site *i* having y_i crashes (y_i = 0, 1, 2,…) per some time period is expressed by:

$$P(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \qquad (1)$$

with mean $\mu_\mathrm{i} = \lambda_\mathrm{i} T_\mathrm{i}$, which is the Poisson parameter for site *i*, where λ_i is equal to site *i*'s expected number of crashes per year, $E(y_i)$ [2]. Furthermore, the expected number of crashes per period, μ_i is linked to the explanatory variables $\boldsymbol{X}_i$ through a log link function: $\mu_i = \lambda_i T_i = \exp(\eta) = \exp(\boldsymbol{\beta}^\mathrm{T} \boldsymbol{X}_i)$, where η is linear predictor, and $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$ are vectors containing the values of the explanatory variables (traffic flows, geometric characteristics of intersections) and the estimable parameters, respectively [1]. The first term in the vector $\boldsymbol{\beta}$ is the constant term or the intercept, since the first component of $\boldsymbol{X}$ is 1. However, for a Poisson distribution the variance is equal to the mean; this may be quite restrictive for crash data which exhibit more variation than given by the mean and are often overdispersed. A way to relax the equidispersion assumption and take into account the resulting variability, as well as avoid model specification errors, is to use a generalized linear model framework [3] where the most common approach is a quasi-Poisson model (a quasi-likelihood with Poisson-like assumptions [15]) or a Negative Binomial model; they derive from the Poisson model and deal with overdispersion for counts since allow the mean to differ from the variance [2, 4].

They often give similar results, but there can be significant differences in the estimation of the effects of the covariates [16]. In a generalized linear model, the variance, var(y_i), of an observation y_i with mean μ_i is assumed to be a linear function of the mean: $E(y_i)=\mu_i$ and $var(y_i)=\phi\mu_i$, where $E(y_i)$ is the expectation of y_i, with $\mu_i > 0$; ϕ is equal to $(1+\alpha)$, where $\alpha < 0$ ($0 < \phi < 1$) denotes underdispersion and $\alpha > 0$ ($\phi > 1$) denotes overdispersion in turn. The relationships above introduced, along with the use of a log link function, allow us to name this a quasi-Poisson model: $y \sim Poi(\mu, \phi)$. In contrast to the latter, for a negative binomial regression the variance of y_i is assumed to be a quadratic function of the mean: $var(y_i) = \mu_i + \mu_i^2/k = \mu_i + \alpha\,\mu_i^2$ where $\mu_i > 0$, and α is the dispersion parameter; the parameter k must be also positive, and as it increases to infinity the distribution approaches the Poisson distribution. Here, the amount in excess of μ (i.e. the overdispersion) is the factor $(1+\mu/k)$ depending on μ (the observation subscript is omitted where no ambiguity should be caused). Thus one could use the quasi-Poisson or the Negative Binomial distribution to represent the distribution of overdispersed crash counts; again one can adopt predictive or goodness-of-fit criteria to choose a quasi-Poisson or a NB model for analyzing overdispersed data; for further details on this topic see e.g. [16].

Despite many model forms can be introduced for SPFs, the power function was here assumed $E[y] = \beta_0 X^{\beta 1}$ or the linear version $ln\ E[y] = ln\ \beta_0 + \beta_1\ lnX$, where X is the AADT (exposure variable) and β_0, β_1 are coefficients to be estimate; these coefficients were estimated by the maximum-likelihood procedure. First, trend effects (i.e. the phenomenon is stationary) were excluded and GLMs were applied. Then, the correlation within responses was accounted for and the flow-only model was also estimated through GEEs, assuming that data consisted of repeated measures over time, possibly correlated within an entity [17]. In GEE, the correction for within-subject correlations is made by assuming *a priori* a correlation structure for the repeated measurements; for this purpose we looked for the simplest structure that fitted data well. In both cases GenStat software was used. The statistical performance of the model was assessed by the methods briefly explained in table 1.

This application led to the results showed in table 2, which shows the parameter estimates with three different distributions in GLM framework. In the same table the corresponding measures of goodness-of-fit are also reported. MPB, MAD and MSPE values in table 2 have slight differences: the MPB values of the NB model highlight that the model slightly underestimates crashes; the

MSPE values of the quasi-Poisson model, being closer to 0 than the NB model, highlight that the model have good prediction accuracy.

Table 1: Goodness-of-fit methods.

Methods	Description
$MPB = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$	The mean prediction bias considers the differences between predicted and actual values; a positive (or negative) value indicates that the model overpredicts (or undepredicts) crashes. Smaller absolute values of MPB indicate a better predictive model [18].
$MAD = \frac{1}{N}\sum_{i=1}^{N}\lvert\hat{y}_i - y_i\rvert$	The mean absolute deviation is the average dispersion of the model; an estimate close to 0 suggests that the model predicts the actual values well [18].
$MSPE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$	The mean square prediction error is an assessment of the error associated with the validation dataset and is the sum of the squared differences between predicted and actual values. A model that provides MSPE closer to zero is considered to be the best model among all the available models [18].
$R_m^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$	R^2-like measure of fit provides an estimate similar to R^2 used in linear regression but is not appropriate for GLMs and is calculated from the residual sum of squares and total sum of squares after the model is applied to the data. The statistics represents the amount of variance in the response variable which is explained by the fitted model [9].

Table 2: Parameter estimates with three different distributions in GLM framework and measures of goodness-of-fit.

Parameter	Poisson			Quasi-Poisson			Negative Binomial		
	est (s.e.)	t	t pr.	est (s.e.)	t	t pr.	est (s.e.)	t	t pr.
Constant (β_0)	-7.10* (0.64)	-11.17	<.001	-7.10* (0.86)	-8.22	<.001	-6.84** (0.90)	-7.60	<.001
ln(AADT) (β_1)	0.91 (0.06)	14. 46	<.001	0.91 (0.08)	10.64	<.001	0.88 (0.09)	9.77	<.001
α	-			0.85			0.12		
MPB	0.00			0.00			-0.03		
MAD	2.84			2.84			2.82		
MSPE	14.93			14.93			15.01		

(*) antilog of constant estimate: 0.0008242; (**) antilog of estimate: 0.001072.

According to [16], $(y_i - \mu_i)^2$ were plotted against μ_i, binning μ_i into 11 mean categories and averaging $(y_i - \mu_i)^2$ within categories; this plot should help to diagnose a linear or quadratic relationship between the mean and variance. The diagnostic plot of the empirical fit of the variance (using average squared residuals) to mean relationship in fig. 2 suggests that for small values of $\hat{\mu}_i$ the negative binomial error structure fits better and for larger values of $\hat{\mu}_i$ the quasi-Poisson fits slightly better; figure 2 also suggests that for means of less than 15 crashes, the quasi-Poisson will have a higher variance, and for means above 15, the negative binomial will have a higher variance. Thus the comparison analysis for the variance estimated by the quasi-Poisson and NB models, linking the number of crashes to the entering traffic flows at the intersections, shows that the quasi-Poisson model was able to capture the variance as well as the NB model.

Table 3 shows in turn GEE regression results only with the NB model; GEE regressions were fitted assuming that repeated observations were correlated under four different working correlation matrices; table 3 also reports the measures of goodness-of-fit. Differences in measures of goodness-of-fit can be observed for the different working correlation matrixes; the marginal R^2_m test is now introduced and used [9]. MPB, MAD and MSPE values in table 3 provide insights on the best correlation structure. GEE regression, indeed, assuming that repeated observations are correlated in different ways, allows us to gain a better understanding of the proper correlation structure in crash counts.

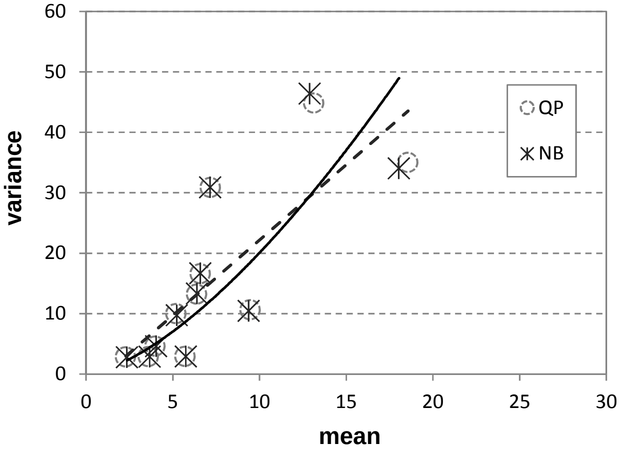


Figure 2: Estimated variance-to-mean relationship for the data set. (*Note: the dashed line is the quasi-Poisson regression model; the solid line is the NB regression model.*)

Table 3: GEE regression results under different working correlation matrices and measures of goodness-of-fit.

Parameter	independent		unstructured		autoregressive		exchangeable	
	est (s.e.)	t	est (s.e.)	t	est (s.e.)	t	est (s.e.)	t
Constant (β_0)	-6.84 (0.96)	-7.12	-7.22 (0.92)	-7.84	-6.87 (1.08)	-6.36	-6.33 (1.38)	4.58
ln(AADT) (β_1)	0.88 (0.09)	9.78	0.91 (0.09)	10.12	0.89 (0.11)	8.09	0.79 (0.13)	6.07
MPB	-0.03		-1.12		0.19		-2.19	
MAD	2.82		2.81		2.85		3.17	
MSPE	15.00		16.76		14.98		22.86	
R^2_m	0.53		0.47		0.53		0.28	

The goodness-of-fit of the model was also explored using the method of cumulate residuals (CURE), in which the cumulative residuals (i.e. the difference between the actual and the fitted values for each roundabout) were plotted in increasing order for each covariate separately [12, 14]. The graph in figure 3 shows how the model under the unstructured correlation matrix fits data with respect to the selected covariate. The indication is that the fit for the covariate is fairly good in that the cumulative residuals, though oscillating slightly around the value of 0, lie between the limit values of the standard deviation (± 2 σ*).

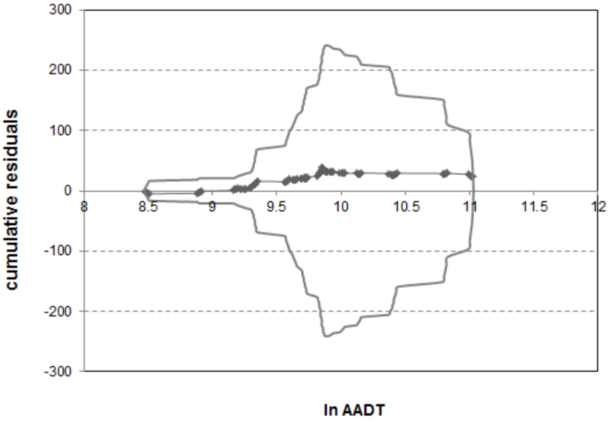


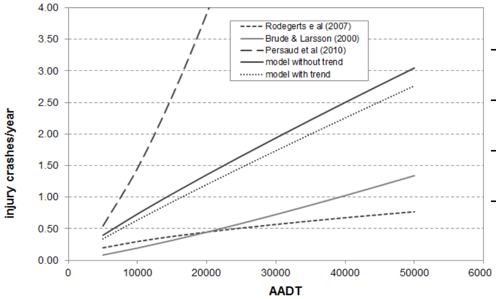
Figure 3: Cumulative residuals and the ±2 σ* band.

2.3 Comparison among flow-only models for roundabouts as a whole

Despite the data set is limited to a sample of 35 four-leg roundabouts operating in the road network of Palermo City, Italy, and it cannot be considered representative of the national situation (due to atypical features detected in the geometric layouts and driver behavior [13]), a comparison of the above models with other flow-only models for roundabouts known to the authors has been attempted. Figure 4 shows details of these models from the US [19], Sweden [20] and Canada [21]. This figure shows that the model developed in this case study predicts more crashes than the Swedish and US models, especially when the trend is neglected; it predicts less crashes than the Canadian model which is, however, based on more recent data.

Transferability can be of interest given that some countries may not have sufficient data to calibrate crash predictive models; however, crash models reflecting safety experience in different countries are not similar enough that they can be transferred straight from one country to another. According to [12] possible differences among countries in road safety are due to crash reporting rates, definition of intersection crashes, climatic conditions, speed limits. Thus an important issue for future research is how to transfer a model to other countries where design standards and driver characteristics are different.

For making a transferability assessment of crash predictive models the reader is advised to consult the current literature on the subject; see e.g. [22]. For a discussion on design issues for converting circular intersections into modern roundabouts, often simultaneously present in the urban road network of Italian cities and towns e.g. [23] is cited.



models	Data period
1.30E-03 (AADT)$^{0.59}$ [19]	1996–2001
3.08E-06 (AADT)$^{1.2}$ [20]	1994–1997
3.05E-06(AADT)$^{1.42}$ [21]	1998–2004
1.07E-03 (AADT)$^{0.88*}$ 7.32E-04(AADT)$^{0.91+}$	1995–2000

**without trend; + with trend*

Figure 4: Comparison for only-flow models for injury crashes for roundabouts.

3 Conclusions

This paper presents an exploratory analysis aimed at modeling the crash phenomenon for a set of injury crash data of roundabouts operating in the road network of Palermo City, Italy. With no claim of being exhaustive, issues associated with the estimation of a crash predictive model were addressed and discussed. Focusing on flow-only models in the belief that traffic volume is the most important predictor variable, a comparison with other models, specifically selected basing on the knowledge of the authors, was performed; the comparison was also made to highlight possible similarities or not in safety experience at roundabouts regardless of the non-flow variables, which inevitably are affected by differences in design standards and practices in force in different countries around the world. Even though the sample is limited to derive outcomes of general validity, the practical usefulness of the results is due to the methodological path followed in this research; moreover, results have allowed to highlight concerns and difficulties in transferring crash models from one country to another, and the usefulness of methods and predictive tools to be developed for specific territorial situations (or preferably at a national level) for quantifying both benefits of engineering measures aimed at enhance traffic safety in built up areas and then the social costs associated with the crash phenomenon.

References

[1] Maher, M. J. & Summersgill, I., A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis & Prevention*, **28(3)**, pp. 281–296, 1996.

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

[3] McCullagh, P. & Nelder, J.A., *Generalized linear models*. Chapman and Hall: London and New York, 1989.

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

[5] Greene, W.H., *Economic Analysis*, Macmillan Publishing Co., New York, 1993.

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

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

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

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

[10] Maycock, G. & Hall, R.D. *Accidents at 4-Arm Roundabouts*, Laboratory Report LR1120, TRL, Crawthorne, Berkshire, U.K. 1984.

[11] Rodegerdts, L., Blogg, M., Wemple, E., Myers, E., Kyte, M., Dixon, M., List, G., Flannery, A., Troutbeck, R., Brilon, W., Wu, N., Persaud, B., Lyon, C., Harkey, D. & Carter., D., *NCHRP Report 572: Roundabouts in the United States*. TRB, Washington, D.C., USA, 2007.

[12] Turner, S., Persaud, B., Chou, M., Lyon, C. & Roozenburg, A., International crash experience comparisons using prediction models. *Proc. of the TRB 2007 Annual Meeting*, Washington, D.C., USA, 2007.

[13] Granà, A., Safety valuation methods at urban atypical intersections. Analysis of infrastructural scenarios. *International Journal of Sustainable Development and Planning*, ***2(3)***, pp. 271–286, 2007.

[14] Hauer, E. & Bamfo, J., Two tools for finding what function links the dependent variable to the explanatory variables. *Proc. of ICTCT 97 Conference*, November 5-7, Lund, Sweden, 1997.

[15] Wedderburn, R.W.M., Quasi-likelihood functions, generalized linear models and the gauss-Newton method. *Biometrika*, **61**, pp. 439–447, 1974.

[16] Ver Hoef, J.M. & Boveng, P.L., Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, ***88(11)***, pp. 2766–72, 2007.

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

[18] 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.

[19] Rodegerdts, L., Blogg, M., Wemple, E., Myers, E., Kyte, M., M. Dixon, List, G., Flannery, A., Troutbeck, R., Brilon, W., Wu, N., Persaud, B., Lyon, C. & Harkey, D., *Application of roundabouts in the US. NCHRP Report 572: Roundabouts in the US*, TRB, Washington, DC, 2007.

[20] Brude, U. & Larsson, J., What roundabout design provides the highest possible safety?, *Nordic Road & Transport Research*, ***2***, pp. 17–21, 2000.

[21] Persaud, B., Lyon, C. & Chen, Y., Tools for estimating the safety and operational impacts of roundabouts, Report for Transport Canada, Ryerson University, Toronto, Canada, 2010.

[22] Highway Safety Manual, First Edition, AASHTO, 2010.

[23] Pratelli, A. & Souleyrette, R.B. Visibility, perception and roundabout safety. *WIT Transactions on the Built Environment*, **107**. In: Brebbia, C.A. (ed.), UK. Proceedings Urban Transport XV, pp. 577–588, 2009.