Periodica Polytechnica Transportation Engineering, 51(2), pp. 140–146, 2023

Generalized Linear Modeling of Crashes on Urban Road Links

Anteneh Afework Mekonnen[1*], Tibor Sipos[1], Zsombor Szabó[1]

[1] Department of Transport Technology and Economics, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, H-1111 Budapest, Hungary
[*] Corresponding author, e-mail: anteneh.mekonnen@kjk.bme.hu

Received: 22 August 2022, Accepted: 08 January 2023, Published online: 01 February 2023

Abstract

As crash data have distinctive behavior like over-dispersion, researchers have used statistical methods to deal with this unique behavior of crash data specifically. This study employed generalized linear modeling techniques to develop the model. It was assumed that the accident counts followed negative-binomial distribution, and the link function was chosen to be the log link function. Negative-binomial modeling technique was chosen over Poisson distribution because it is the most used technique by many researchers as crash data may encounter over-dispersion. The accident data set showed greater variability between its variance and mean. The accident frequency distribution is shown in this study that it is highly skewed, with a very high number of road segments registering zero accidents. Negative binomial distribution was chosen over Poisson distribution after comparing Akaike's Information Criterion (AIC) and Bayesian Information Criteria (BIC). The method is widely applied to count data. Twenty-two parameters were estimated in the model. Since $p < 0.05$ in the omnibus test, the null hypothesis is rejected, which indicates that the model is reasonably fit. The strongest variables in the model were witnessed to be the length of the links, number of lanes, average daily traffic, bus lane, number of buses and trolleys, and HGVs.

Keywords

generalized linear modeling, road traffic crashes, Poisson distribution, negative binomial distribution, road safety

1 Introduction

A road traffic accident has become the 8th leading cause of death for all age groups, from the 9th leading cause of death. Now it is also the leading cause of death for children and young adults aged 5–29 years. Its impact on peoples' mobility choices is overlooked when we think of the problems due to a traffic accident. In addition to the physical and property damage, people are less likely to bike or walk around unsafe conditions (World Health Organization, 2018). The threat of road crashes and significant public health and economic problem to communities also influence people's travel choices (Kiss et al., 2013).

In road safety research, analysts are interested in predicting road crashes in terms of location, frequency, pattern, and severity in order to ensure better traffic operations and save lives (Mussone et al., 2017).

Analyzing the association between relevant contributing factors and traffic accidents is crucial in traffic safety management (Ye et al., 2018). Regression models play a significant role in road safety. These models can be used for various purposes, such as establishing relationships between motor vehicle crashes and different covariates (i.e., understanding the system), predicting values or screening variables (Geedipally et al., 2012).

Accident rates can be related to selected factors with models (Greibe, 2003). Accident data is known to be count data. It is known that the traditional regression model cannot account for the differences among observations (i.e. heterogeneity) (Elvik, 2011). The most common approach used in modeling this type of data is the Poisson regression. However, because of overdispersion associated with the use of the Poisson model, researchers usually consider the negative binomial model as a potential alternative (Agresti, 2007).

Though the selection of a specific model hinges on the research's objective and the nature of the response compared to statistical modeling techniques (Mekonnen and Sipos, 2022) many studies have been carried out over the course of years using statistical modeling techniques to solve problems related to road traffic accidents. Lord et al. (2008) applied Conway-Maxwell-Poisson generalized linear model to analyze motor vehicle crashes; Prieto et al. (2014) analyzed accident blackspots with discrete generalized pareto

Cite this article as: Mekonnen, A. A., Sipos, T., Szabó, Z. (2023) "Generalized Linear Modeling of Crashes on Urban Road Links", Periodica Polytechnica Transportation Engineering, 51(2), pp. 140–146. https://doi.org/10.3311/PPtr.19119

distribution; Geedipally et al. (2012) employed negative binomial-Lindley generalized linear model to analyze crash data. As crash data have distinctive behavior like over-dispersion, researchers have used statistical methods to specifically deal with this unique behavior of crash data. This unique behavior also inspired researchers to come up with different new modeling techniques (Shirazi et al., 2016). The negative binomial distribution is quite the same as the Poisson distribution, however, unlike the Poisson distribution, the variance of the negative binomial distribution exceeds its mean. As such, if the variance of the observed outcome is suspected to be larger than its mean, the negative binomial distribution is more suitable compared to the Poisson distribution.

Generalized Linear Models (GLM) with negative binomial distribution for errors have been commonly used to estimate safety at the level of transportation planning (Gardner et al., 1995; Geedipally et al., 2012; Hilbe, 2011; Shirazi et al., 2016; Wood, 2002). GLMs with log link function and negative binomial distributions are widely used to relate accidents with explanatory variables (Maher and Summersgill, 1996; Wood, 2005).

Few or no such studies have been conducted in Budapest, except for Harmati et al. (2008) developed energy distribution model of car body deformation using linear parameter varying (LPV) representations and fuzzy reasoning in 2008.

2 Description and preparation of data

2.1 Description of data

The total number of road links in Budapest that are considered in this model is 4,701. The total length of road networks in the dataset is 1,310 km. The average link length is 0.28 km. 1,586 accidents that occurred on the road links in the city over three years from 2016 to 2018 were considered in the model. That is a huge dataset despite the fact that it is usually witnessed in the field of road safety research, that one of the challenges is limited data.

Many studies used hundreds or even dozens of observations due to the limited availability of crash and highway data (Wu and Lord, 2017). Fig. 1 is a map from the ArcMap 10.4.1 showing the distribution of crashes on the road links for the city of Budapest for the three study years.

2.2 Data preparation

To conduct sound traffic safety study and practice, researchers should have a better understanding of their data (Zhao et al., 2019). Information about road elements is normally collected in the form of nodes and links, with nodes as intersections and links as segments (Qin and Wellner, 2012). The Budapest road network shape files and accident datasets were imported to ArcMap 10.4.1. Crashes are not completely randomly distributed but highly associated to the traffic condition and geometric features of the road. This postulate leads researchers to conduct segment based traffic safety analysis (Zhao et al., 2019). A buffer radius of 50 m was taken at every node or junction as the sample is shown in Fig. 2. The segment which is out of the buffer radius, or the midblock section was considered as road link. After buffering, the intersection point between the buffer and the links were found and segmentation was completed.

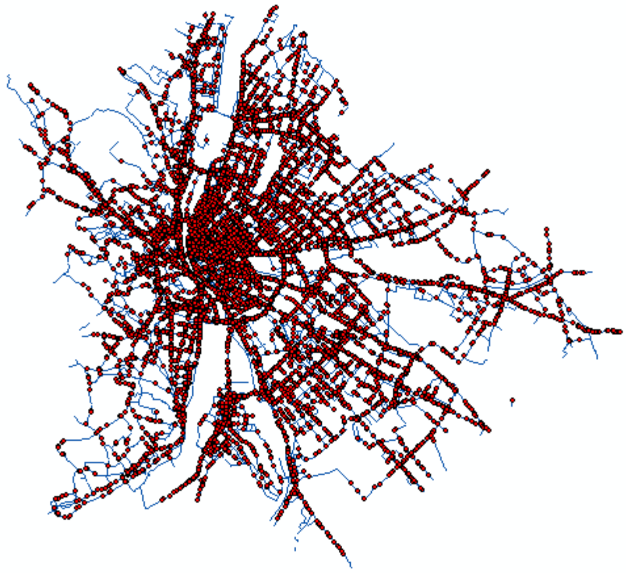

Fig. 1 Crash distribution on Budapest road links from 2016 to 2018 (Source: Authors)

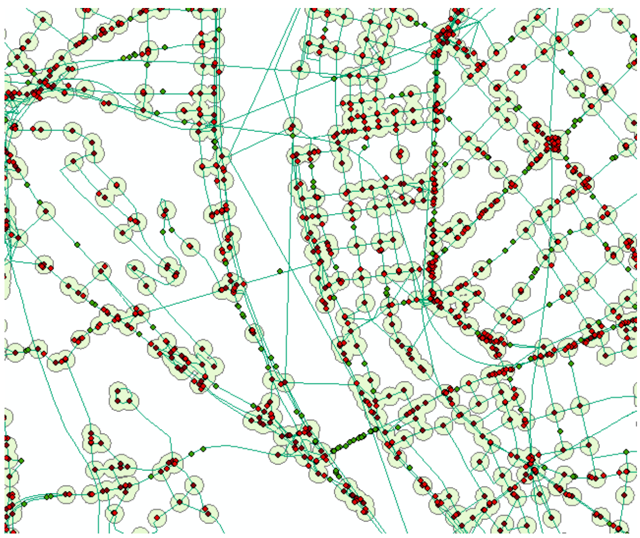

Fig. 2 Buffering of crashes with 50 m radius from nodes (Source: Authors)

Roadway segmentation is partitioning roadways into segments to create the observations for modelling. Segmentation heavily affects the modelling of roadway crash frequencies because crashes are rare events. Very short segments result in many segments with zero crashes, which leads to over-dispersion. Over-dispersion means that the variance of crash data exceeds its mean. For the accident data in this study, the variance was found to be 1.992, and the mean is 0.66, which shows a greater variability. It creates challenges for statistical inference because it is difficult to accurately assign a crash to a segment if segment lengths are very small (Lord and Mannering, 2010).

The next step was to assign crashes, as the sample is shown in Fig. 3 and Fig. 4, to individual road elements. After assigning crashes to segments, an attribute table with crash information was exported to a separate Excel file to prepare it for modelling.

Accidents and frequency of road links are summarized in Table 1.

The event dealt with in this study is rare and the distribution of number of accidents is highly skewed, with frequencies showing peak for lower number of accidents and sharply declines towards the right end for higher number of accidents.

Fig. 5 shows the frequency distribution of accidents occurred in Budapest City from 2016 to 2018 on road segments.

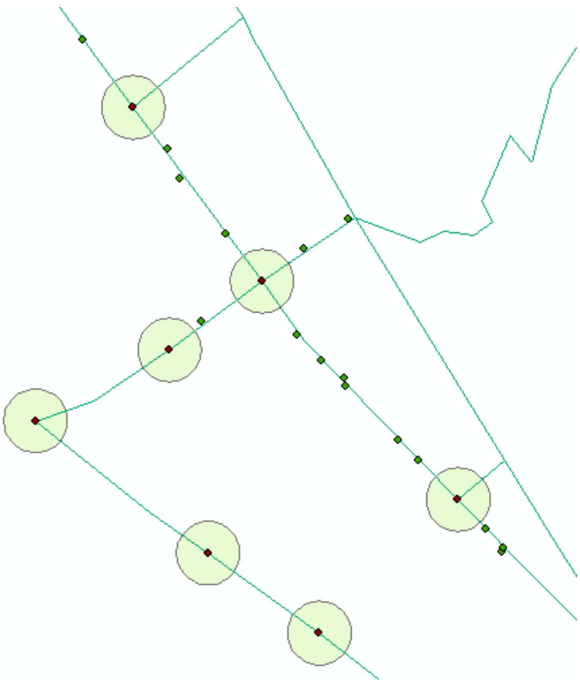

Fig. 3 Crashes outside the buffer zone (Source: Authors)

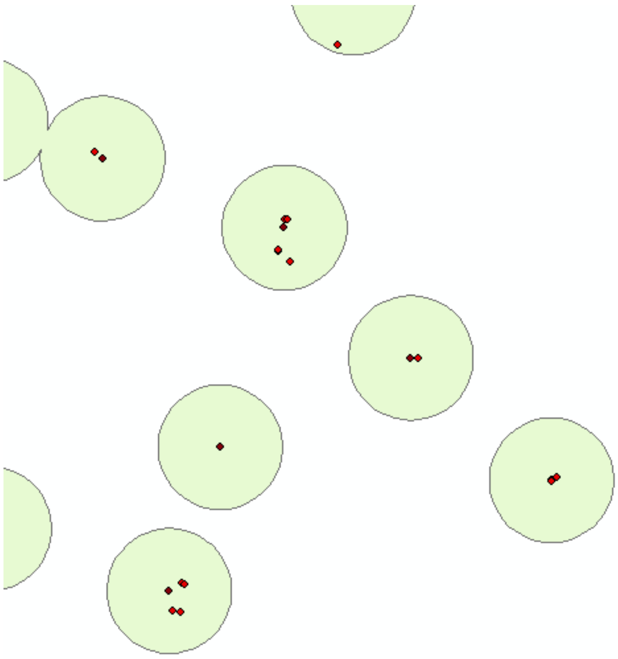

Fig. 4 Crashes within the buffer zone (Source: Authors)

Table 1 Crashes and frequency of road links (Source: Authors)

No. of accidents	No. of road links	%	No. of accidents	No. of road links	%
0	3115	66.3	9	3	0.1
1	909	19.3	10	3	0.1
2	350	7.4	11	2	0.0
3	141	3.0	12	1	0.0
4	75	1.6	15	1	0.0
5	49	1.0	16	1	0.0
6	19	0.4	17	1	0.0
7	19	0.4	18	1	0.0
8	10	0.2	28	1	0.0
Total				4701	100.0

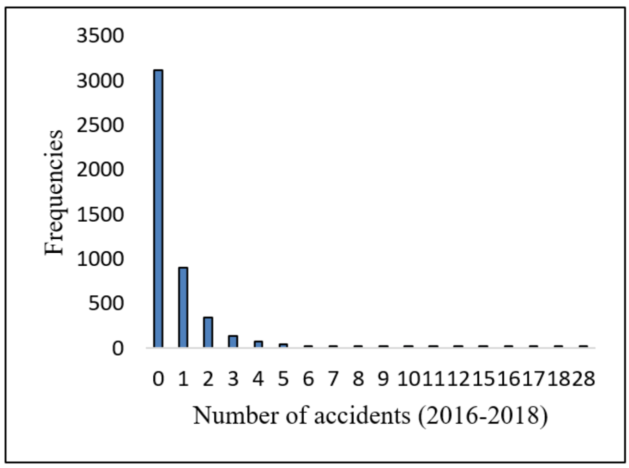

Fig. 5 Number of crashes on Budapest road links from 2016 to 2018 (Source: Authors)

3 Modeling

3.1 Model structure

Generalized linear modelling techniques were employed to develop the model. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. The GLM consists of three elements. An exponential family of probability distributions, a linear predictor, and a link function (Aliakbar Golkar and Valizadeh Haghi, 2011).

The random component of a GLM consists of a response variable Y with independent observations (y_1, …, y_N) from a distribution in the natural exponential family. This family has probability density function or mass function of the form:

$$f\left(y_i;\theta_i\right)=a\left(\theta_i\right)b\left(y_i\right)\exp\left[y_i Q\left(\theta_i\right)\right]. \quad (1)$$

Several important distributions are special cases, including the Poisson and binomial. The value of the parameter θ_i may vary for i = 1, …, N, depending on the values of explanatory variables. The term $Q(\theta)$ is called the natural parameter.

Poisson regression is utilized when a researcher has a count data on some dependent measure representing the incidence rate of some event (Moksony and Hegedűs, 2014). However, when working with this this distribution one can assume that the mean of the distribution is equal to its variance. When the variance of a distribution is greater than its mean, the distribution is said to be over-dispersed.

An alternative strategy for modeling involves using the negative binomial distribution. It has the same "sample space" as the Poisson distribution but also has an additional parameter used to model the variance. This parameter is referred to, unsurprisingly, as the dispersion parameter (Gardner et al., 1995).

The systematic component of a GLM relates a vector (η_1, …, η_N) to the explanatory variables through a linear model. Let x_{ij} denote the value of predictor j(j = 1, 2, …, p) for subject i. Then

$$\eta_i=\sum_j \beta_j x_{ij},\ i=1,\ldots,N \quad (2)$$

This linear combination of explanatory variables is called the linear predictor. Usually, one x_{ij} = 1 for all i, for the coefficient of an intercept (often denoted by a) in the model.

The third component of a GLM is a link function that connects the random and systematic components. Let μ_i = $E(Y_i)$, i = 1, …, N. The model links μ_i to η_i by $\eta_i = g(\mu_i)$, where the link function g is a monotonic, differentiable function. Thus, g links $E(Y_i)$ to explanatory variables through the formula" (Agresti, 2002):

$$g\left(\mu_i\right)=\sum_j \beta_j x_{ij},\ i=1,\ldots,N. \quad (3)$$

In this study, it was assumed that the accident counts follow negative-binomial distribution, and the link function was chosen to be log link function. Negative-binomial modeling technique was selected because it is the most used technique by many researchers as crash data may encounter over-dispersion.

Overdispersion may cause standard errors of the estimates to be deflated or underestimated, i.e. a variable may appear to be a significant predictor when it is not actually significant (Hilbe, 2011).

Assumption:

$$\left(Y/X\right)\sim \mathrm{NB}\left(m,a\right),\ Var\left(y_i/x\right)=\mu_i+\alpha\mu_i^2, \quad (4)$$

where α is the dispersion parameter that indicates the degree of over-dispersion.

In negative binomial regression, the mean of y is determined by the exposure and a set of k regressor variables (the x's). The expression that creates a relationship among these quantities is

$$\mu_i=\left(E\left(Y_i/x_i\right)\right)=\exp\left[\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\ldots+\beta_kX_{ki}\right], \quad (5)$$

where β_0 is called the intercept. The regression coefficients β_1, β_2, …, β_k are unknown parameters that are estimated from a set of data. Their estimates are symbolized as b_1, b_2, …, b_k.

3.2 Modeling procedure

Initially, all the independent variables were considered in the model. Then all insignificant variables were eliminated based on likelihood ratio analysis and standard errors of the parameter estimate values. The correlation matrix was analyzed, and if two variables were correlated to each other strongly, one of them was eliminated from the model. This helps avoid the confounding effect in our model. Even if a variable was found to be significant but had less contribution to the decrease in deviance, it was eliminated from the model.

Categorical variables were interacted with and coupled together in the model analysis. In the generalized linear modeling, Negative binomial distribution and log link function were considered.

Simplification of some explanatory variables was carried out. Free flow speed values were categorized in four groups: the first category is less than 20 km/h, the second between 20 and 50 km/h, the third between 50 and 90 km/h and the last greater than 90 km/h. The original dataset contains four categories of HGVs, and they are HGVs < 3.5 ton, 3.5–7.5 ton, 7.5–12 ton, and > 12 ton, respectively. Then the first category, which is < 3.5 ton is grouped with taxis. The following two categories of HGVs are grouped, and the last HGVs category, which is > 12 ton is considered as it is. Each parameter was also identified as a covariate and factor in the model development.

Overall, the modelling process involves the following four steps:

1. Specifying models in two parts: equations linking the response and explanatory variables and the probability distribution of the response variable.
2. Estimating parameters used in the models.
3. Checking how well the models fit the actual data.
4. Making inferences (Dobson, 2002).

3.3 Model test

Goodness of fit for a generalized linear model is traditionally assessed using either the scaled deviance G2 (twice the difference between log likelihood of saturated model and loglikelihood of current model) or Pearson's χ^2 statistic (sum of squares of standardized observations). Here the models are assumed to be nested, and the larger model is that with, the greater number of parameters. If the data is normally distributed, these statistics follow chi-square distributions with degrees of freedom equaling the difference between the number of parameters in the larger model and the number in the smaller model (Wood, 2002).

Hilbe (2011) mentions Akaike's Information Criterion (AIC) as one of the most used fit statistics. It has the following formulation:

$$\mathrm{AIC} = 2k - 2\ln(\mathcal{L}), \tag{6}$$

where k is the number of estimated parameters in the model, including the intercept, and $\mathcal{L}$ is the maximum value of the likelihood function for the model.

The one with the minimum AIC value is preferred for the given set of models.

Hilbe (2011) mentions the Bayesian Information Criterion (BIC) as another common fit statistic. It has the following formulation:

$$\mathrm{BIC} = \ln(n)k - 2\ln(\mathcal{L}), \tag{7}$$

where n is the number of data sets or the number of observations, k is the number of estimated parameters in the model, including the intercept, and $\mathcal{L}$ is the maximum value of the likelihood function for the model.

An omnibus test was made to test whether the explained variance in a dataset was significantly greater than the unexplained variance overall. Likelihood ratio chi-square test was made to the model.

The following two hypotheses were tested to imply whether the model is fit or not:

$$H_0: \ \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_1: \ \text{at least one } \beta_j \neq 0 .$$

4 Results and discussion

IBM SPSS Statistics 25 (Statistical Package for Social Sciences) software was employed in developing a generalized linear model based on the negative binomial distribution in this study. IBM SPSS Statistics is the most widely used software to analyze quantitative data. IBM SPSS is a comprehensive system for generating descriptive statistics and complex statistical analysis (Vilaça et al., 2017).

Road accident datasets were prepared in a single database, and the following parameters were analyzed with IBM SPSS.

In the case of the Poisson distribution, the Pearson Chi-Square value divided by the degree of freedom gives 2.074, which is greater than one. We understand that the data are over-dispersed, and negative binomial distribution is suitable. The negative binomial distribution model greatly improved the model with Pearson Chi-Square ratio to the degree of freedom value of 1.264. Table 2 shows the comparison of the two models from Poisson and Negative binomial distributions based on multiple criteria.

Based on all three criteria, the Negative binomial distribution improved the model significantly. It was found to be suitable to model the over-dispersed data.

The parameter estimates from the IBM SPSS are presented in Table 3.

Table 2 Comparing the Poisson and Negative binomial distributions (Source: Authors)

Criterion	Distribution	
	Poisson	Negative binomial
AIC	10,015	9,242
BIC	10,157	9,384
Pearson Chi-square/df	2.074	1.264

Table 3 Parameter estimates (Source: Authors)

Parameter	B	Sig.
(Intercept)	−3.821	**
[CAT_FF_SP = 2] * [NUMLANES = 1]	2.910	**
[CAT_FF_SP = 3] * [NUMLANES = 1]	−26,050a	
[CAT_FF_SP = 3] * [NUMLANES = 7]	0b	
BUS_LANE	0.417	***
BUS_TROL_4	0.000	***
HGV_12_D	0.074	.
BIKE_VOL	0.000	***
LN_ADT	0.282	***
Ratio_HGV	−4.142	***
LN_LENGTH	1.473	***
(Scale)	1c	
(Negative binomial)	1c	

Sig. (. if $p < 0.3$; * if $p < 0.05$; ** if $p < 0.01$; *** if $p < 0.001$)

The generalized linear model was developed to predict the probability of the number of accidents that will occur in the future.

Since $p < 0.05$ in the omnibus test, the null hypothesis is rejected. Therefore, the model is proven to be reasonably fit.

The model shows length of the links, average daily traffic, bus lane, number of busses and trolleys, and bike volumes with *** rating which has highest significance for the contribution of the dependent variable followed by interaction of free flow speed category 2 and number of lanes category 1 with ** significance rating. But all the parameters in the final selected model have relevance in predicting the dependent variable to an acceptable degree.

5 Conclusion

The study aimed at developing a traffic accident model which would predict the possible number of accidents in the future at a specific road link in the city of Budapest. In addition, the purpose was to identify factors that affect traffic safety and contribute to an accident on the road links. Multiple models were run based on 1,586 accidents from 1,310 km of road links.

The accident data was highly skewed and showed greater variability between its mean and variance. Because of that, Negative binomial distribution was chosen over Poisson as it is better to deal with over-dispersed data. The strongest variables in the model which contribute to traffic accidents were the length of the links, average daily traffic, bus lane, number of busses and trolleys, and bike volumes.

Since internal correlation within the data has been witnessed to be a major problem, the safety effects from a single explanatory variable were difficult to estimate as it may be affected by other variables in the model. One of the challenges during the statistical analysis of this study was association of quite many road links with zero accidents. Out of the total 4,701 road links, 3,115 of them showed zero number of accidents recorded. Modeling and analysis of accidents at junctions w not the scope of this study which is recommended to be compared with the safety issues at road links.

References

Agresti, A. (2002) "Categorical data analysis", 2nd ed., John Wiley & Sons, Inc. ISBN 9780471360933
https://doi.org/10.1002/0471249688

Agresti, A. (2007) "An Introduction to Categorical Data Analysis", 2nd ed., John Wiley & Sons, Inc. ISBN 9780471226185
https://doi.org/10.1002/0470114754

Aliakbar Golkar, M., Valizadeh Haghi, H. (2011) "Using a Multivariate DOE Method for Congestion Study in Distribution Ssystems under Impacts of Plug-in Electric Vehicles", [pdf] In: 21st International Conference on Electricity Distribution, Frankfurt, Germany, 0718. Available at: http://www.cired.net/publications/cired2011/part1/papers/CIRED2011_0718_final.pdf [Accessed: 13 October 2020]

Dobson, A. J. (2002) "An introduction to generalized linear models", [pdf] 2nd ed., Chapman & Hall/CRC. ISBN 1-58488-165-8 Available at: https://reneues.files.wordpress.com/2010/01/an-introduction-to-generalized-linear-models-second-edition-dobson.pdf [Accessed: 17 October 2020]

Elvik, R. (2011) "Assessing causality in multivariate accident models", Accident Analysis & Prevention, 43(1), pp. 253–264.
https://doi.org/10.1016/j.aap.2010.08.018

Gardner, W., Mulvey, E. P., Shaw, E. C. (1995) "Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models", Psychological Bulletin, 118(3), pp. 392–404.
https://doi.org/10.1037/0033-2909.118.3.392

Geedipally, S. R., Lord, D., Dhavala, S. S. (2012) "The negative binomial-Lindley generalized linear model: Characteristics and application using crash data", Accident Analysis & Prevention, 45, pp. 258–265.
https://doi.org/10.1016/j.aap.2011.07.012

Greibe, P. (2003) "Accident prediction models for urban roads", Accident Analysis & Prevention, 13(2), pp. 273–285.
https://doi.org/10.1016/S0001-4575(02)00005-2

Harmati, I. Á., Rövid, A., Szeidl, L., Várlaki, P. (2008) "Energy Distribution Modeling of Car Body Deformation Using LPV Representations and Fuzzy Reasoning", [pdf] WSEAS Transactions on Systems, 7(11), pp. 1228–1237. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=9a5fdf2df6365ab0657e18f7ba418d1ee1efb96d [Accessed: 18 October 2020]

Hilbe, J. M. (2011) "Negative Binomial Regression", 2nd ed., Cambridge University Press. ISBN 978-0-521-19815-8

Kiss, Á. O., Sesztakov, V., Török, Á. (2013) "Road Safety Analysis in Győr", Periodica Polytechnica Transportation Engineering, 41(1), pp. 51–56.
https://doi.org/10.3311/PPtr.7098

Lord, D., Guikema, S. D., Geedipally, S. R. (2008) "Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes", Accident Analysis & Prevention, 40(3), pp. 1123–1134.
https://doi.org/10.1016/j.aap.2007.12.003

Lord, D., Mannering, F. (2010) "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.
https://doi.org/10.1016/j.tra.2010.02.001

Maher, M. J., Summersgill, I. (1996) "A comprehensive methodology for the fitting of predictive accident models", Accident Analysis & Prevention, 28(3), pp. 281–296.
https://doi.org/10.1016/0001-4575(95)00059-3

Mekonnen, A. A., Sipos, T. (2022) "Crash Prediction Models and Methodological Issues", Periodica Polytechnica Transportation Engineering, 50(3), pp. 267–272.
https://doi.org/10.3311/PPtr.16295

Moksony, F., Hegedűs, R. (2014) "The use of Poisson regression in the sociological study of suicide", Corvinus Journal of Sociology and Social Policy, 5(2), pp. 97–114.
https://doi.org/10.14267/cjssp.2014.02.04

Mussone, L., Bassani, M., Masci, P. (2017) "Analysis of factors affecting the severity of crashes in urban road intersections", Accident Analysis & Prevention, 103, pp. 112–122.
https://doi.org/10.1016/j.aap.2017.04.007

Prieto, F., Gómez-Déniz, E., Sarabia, J. M. (2014) "Modelling road accident blackspots data with the discrete generalized Pareto distribution", Accident Analysis & Prevention, 71, pp. 38–49.
https://doi.org/10.1016/j.aap.2014.05.005

Qin, X., Wellner, A. (2012) "Segment length impact on highway safety screening analysis", In: Transportation Research Board 91st Annual Meeting, Washington, DC, USA, 12-0644. [online] Available at: https://www.semanticscholar.org/paper/Segment-Length-Impact-on-Highway-Safety-Screening-Qin-Wellner/0595e2188788508e4ccf89b51d4916945ff1b1af [Accessed: 10 November 2020]

Shirazi, M., Lord, D., Dhavala, S. S., Geedipally, S. R. (2016) "A semi-parametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data", Accident Analysis & Prevention, 91, pp. 10–18.
https://doi.org/10.1016/j.aap.2016.02.020

Vilaça, M., Silva, N., Coelho, M. C. (2017) "Statistical Analysis of the Occurrence and Severity of Crashes Involving Vulnerable Road Users", Transportation Research Procedia, 27, pp. 1113–1120.
https://doi.org/10.1016/j.trpro.2017.12.113

Wood, G. R. (2002) "Generalised linear accident models and goodness of fit testing", Accident Analysis & Prevention, 34(4), pp. 417–427.
https://doi.org/10.1016/S0001-4575(01)00037-9

Wood, G. R. (2005) "Confidence and prediction intervals for generalised linear accident models", Accident Analysis & Prevention, 37(2), pp. 267–273.
https://doi.org/10.1016/j.aap.2004.10.005

World Health Organization (2018) "Global status report on road safety 2018", World Health Organization. ISBN 9789241565684 [online] Available at: https://www.who.int/publications/i/item/9789241565684 [Accessed: 12 January 2019]

Wu, L., Lord, D. (2017) "Examining the influence of link function misspecification in conventional regression models for developing crash modification factors", Accident Analysis & Prevention, 102, pp. 123–135.
https://doi.org/10.1016/j.aap.2017.02.012

Ye, Z., Xu, Y., Lord, D. (2018) "Crash data modeling with a generalized estimator", Accident Analysis & Prevention, 117, pp. 340–345.
https://doi.org/10.1016/j.aap.2018.04.026

Zhao, X., Lord, D., Peng, Y. (2019) "Examining Network Segmentation for Traffic Safety Analysis With Data-Driven Spectral Analysis", IEEE Access, 7, pp. 120744–120757.
https://doi.org/10.1109/ACCESS.2019.2937001