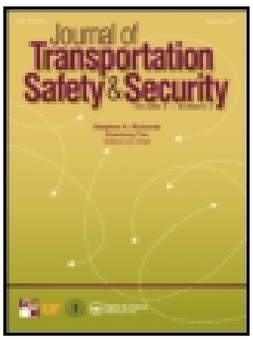

Journal of Transportation Safety & Security

ISSN: 1943-9962 (Print) 1943-9970 (Online) Journal homepage: http://www.tandfonline.com/loi/utss20

An accident prediction model for urban road networks

Fancello Gianfranco, Stefano Soddu & Paolo Fadda

To cite this article: Fancello Gianfranco, Stefano Soddu & Paolo Fadda (2017): An accident prediction model for urban road networks, Journal of Transportation Safety & Security, DOI: 10.1080/19439962.2016.1268659

To link to this article: https://doi.org/10.1080/19439962.2016.1268659

Accepted author version posted online: 22 Dec 2016.
Published online: 22 Mar 2017.

JOURNAL OF TRANSPORTATION SAFETY & SECURITY
http://dx.doi.org/10.1080/19439962.2016.1268659

Taylor & Francis
Taylor & Francis Group

An accident prediction model for urban road networks

Fancello Gianfranco, Stefano Soddu, and Paolo Fadda

Department of Civil Engineering, Environment and Architecture, University of Cagliari, Cagliari, Italy

ABSTRACT

For several years now increasingly analytical prediction models have been developed that are able to correlate accident frequency with infrastructure characteristics to support the planning and design of measures for enhancing road safety. The models developed so far, though useful in the context for which they have been calibrated, are limited by the fact that they are not transferable to other contexts because of different traffic regulations. The present work aims to develop a predictive model for urban roads that is able to estimate the number of accidents for three situations in an urban road network, a roundabout, a three- or four-way junction, and a straight stretch of road. The models constructed are based on Poisson's and negative binomial algorithms and can be readily applied for accident prediction or identification of black spots.

KEYWORDS

Poisson distribution; field research; accident prediction; urban accidents

1. Introduction

According to the latest European Union (EU) data, 1,054,754 road accidents were recorded in 2013 resulting in 25,938 fatalities and 1,387,957 injuries. In Italy road accidents resulting in injuries numbered 181,227 with 3,385 deaths and 257,421 injured. Seventy-five percent of accidents recorded occurred on urban roads, 20% on nonurban roads, and 5% on motorways.

These data demonstrate the need to identify strategies for reducing the number of accidents. One of the main research lines concerns the development of accident prediction models that are able to correlate accident frequency with infrastructure characteristics. This approach is also upheld by the *Highway Safety Manual* (*HSM*) (2010) that proposes different approaches for road safety management, including the use of predictive models.

Several types of prediction models have been developed over the years, but they all have one limitation, namely, they are not transferable to contexts other than the one for which they have been implemented and calibrated. The reason for this lies in the differences in urban areas such as for instance urban structure, traffic

CONTACT Fancello Gianfranco fancello@unica.it Department of Civil Engineering, Environment and Architecture, University of Cagliari, Via Marengo 2, Cagliari 09121, Italy.

regulations and management, number and type of vehicles circulating in the area, and so on. Thus a new model needs to be constructed ad hoc for each specific situation.

The aim of this work is to develop a model for each major portion of road network, namely, roundabouts, junctions (three- or four-way), and straight stretches of road, that, on the basis of specific variables, is able to predict the number of accidents likely to occur.

The article is organized into four sections: Section 2 examines the state of the art of similar models reported in the literature, Section 3 briefly describes the method used to calibrate the models, Section 4 presents the results of the calibration, Section 5 concludes the article.

2. State of the art

Several methods have been developed in recent years for analyzing accident frequency (Lord & Mannering, 2010; Mannering & Bhat, 2014; Savolainen, Mannering, Lord, & Quddus, 2011). Crash-frequency data are non-negative integers, so the application of standard ordinary least-squares regression (which assumes a continuous dependent variable) is not appropriate, though Rakha, Arafeh, Abdel-Salam, Guo, and Flintsch (2010) have actually demonstrated that it is possible to obtain valid results also with a linear regression model using manipulated data.

One classical and common approach to modelling accident frequency is to use a generalized linear model (GLM) considering a Poisson or negative binomial (NB) error distribution. Greibe (2003) used the Poisson model that is simpler to apply than the NB model, whereas Bauer and Harwood (2000) and Canale, Leonardi, and Pappalardo (2005) examined Poisson or NB log-normal and log-linear regressions to determine the most suitable, depending on accident data distribution. Shankar, Mannering, and Barfield (1995), Cafiso, Di Graziano, Di Silvestro, and La Cava (2010), Ackaah and Salifu (2011), Šenk, Ambros, Pokorný, and Striegler (2013), Van Petegem and Wegman (2014) and Eustace, Aylo, and Mergia (2015) on the other hand used the NB model.

The Poisson model lends itself well to analyzing accident frequency inasmuch as the accident data consist of non-negative integers. However this model does not perform well when overdispersion occurs (in the Poisson model variance and mean must be equal). In the latter case the NB model, also called Poisson gamma, is used to handle overdispersion (Lord & Mannering, 2010).

Over the years these models have been further refined to solve a number of problems that had emerged with the original versions. One of these problems concerned excess zeros (Lord & Mannering, 2010), associated with the lack of accident data for certain stretches of road (either because no accidents had occurred or no observations existed). To overcome this problem specific models were developed (Kumara & Chin, 2003; Miaou, 1994; Shankar, Milton, & Mannering, 1997) such as the zero-inflated Poisson (ZIP) or the zero-inflated negative binomial (ZINB),

that made it possible to "distinguish sections of roadway that are truly safe (near zero-accident likelihood) from those that are unsafe but happen to have zero accidents observed during the period of observation (e.g. one year)" (Shankar, Milton, & Mannering, 1997, p. 829). The traditional model (Poisson and NB) does not account for this distinction and thus can produce biased coefficient estimates because of the preponderance of zero-accident observations. Miaou (1994) performed a comparative analysis of the Poisson, NB, and ZIP models and found that the ZIP regression model appeared to be a serious candidate model when data exhibit excess zeros, for example, due to underreporting, in so far as the interpretation of the ZIP model can be difficult.

Shankar, Milton, and Mannering (1997) examined under what conditions the ZIP and ZIBN methods proved to be more appropriate with respect to the Poisson and NB models. Again, Kumara and Chin (2003) used the ZIBN structure to identify those factors affecting road accidents at a number of signalized junctions in Singapore: these too had some shortcomings (Lord, Washington, & Ivan, 2005) associated with the "zero" state whose distribution has a long-term average of zero (theoretically impossible). This led to the development of new models, such as the one proposed by Malyshkina and Mannering (2010) based on "a two state Markov switching count data model" that compared to the "zero" inflated models has the advantage of being able to directly provide a statistical estimation of the state of the road stretch concerned. Geedipally, Lord, and Dhavala (2012) on the other hand used a generalized linear model adopting the Lindley NB distribution to describe error distribution. This approach has the advantage of being able to handle databases containing overdispersed or zero inflated count data while maintaining characteristics similar to the NB distribution.

Another issue concerns the spatial and temporal correlation of the data. Two approaches have been adopted to solve this: one consists in using random effect models, the other in replacing the classical GLM procedure with the new generalized equation estimating (GEE) procedure.

The first to introduce random effect as opposed to fixed effect models were Husman, Hall, and Griliches (1984). Random effects have been studied by several researchers including (Johansson, 1996) who explored the effects of reducing the speed limit on the number of accidents in Sweden and Shankar, Albin, Milton, and Mannering (1998) who performed a comparative analysis of the NB and the random effect NB model. They found that the latter model yielded major benefits in the case where the spatio-temporal correlation is unobserved. Hosseinpour, Yahaya, and Sadullah (2014) on the other hand compared seven types of models including the Poisson, NB, ZIP, ZINB, and random effect NB models. They found that the most suitable for studying accident frequency was actually the random effect NB model.

Regarding the GEE, Lord and Persaud (2000) demonstrated that the GEE model that incorporates the temporal trend performs better than models that do not, or that do not account for the temporal correlation of accident data. Wang and Abdel-Aty (2006) studied the temporal, spatial, and local correlation for accidents

as a whole, for rear-end, head-on or side collisions, and accidents caused by left turns and constructed a series of GEE models. Ma, Yan, Abdel-Aty, Huang, and Wang (2010) used the GEE model to identify the risk factors associated with the serious accidents recorded on the streets in Beijing whereas Mohammadi, Samaranayake, and Bham (2014) compared the GEE model with a NB model that uses the maximum likelihood method that does not account for temporal correlation and Barbosa, Cunto, Bezerra, & Nodari (2014) used the GEE procedure for constructing a prediction model valid for 32 intersections in Brasilia.

Lastly, many authors have proposed using either bivariate (Caliendo, De Guglielmo, & Guida, 2013; Maher, 1990; Ng, Ong, & Srivastava, 2010; Wang, Lee, Yau, & Carrivick, 2003;), or multivariate models (Aguero-Valverde, 2013; Dong, Clarke, Yan, Khattak, & Huang, 2014; El-Basyouny & Sayed, 2009) for modelling different types of accidents. For example Caliendo et al. (2013) and Dong, Richards, Clarke, Zhou, and Ma (2014) used the bivariate and multivariate model respectively to estimate accident frequency, discriminating accidents according to the severity of the consequences (only material damage, bodily injury, death).

As mentioned in the introduction, one problem shared by all these models is the need to be able to use them in contexts other than the one for which they have been calibrated. This difficulty stems from the fact that urban structures, traffic regulations and management, number and type of vehicles circulating in the area, and so on differ substantially.

Marchionna, Perco, and Taverna (2008) tested four models in the city of Trieste (Italy): Canale et al. (2005) developed in Catania (Italy), Bauer and Harwood (2000) in the United States, Summersgill, Kennedy, Hall, Hickford, and Barnad (2001) in Great Britain, and Greibe (2003) developed in Denmark, coming to the conclusion that none of them was able to provide a sufficiently reliable estimate of accident frequency. Sacchi, Persaud, and Bassani (2012), Martinelli, La Torre, and Vada (2009), Marchionna, Perco, and Falconetti (2012) applied the calibration procedure suggested by the *HSM* but again obtaining unsatisfactory results. Thus there is clearly a need to develop new models that can be applied to different urban contexts.

3. Method

Based on the relevant literature and on the heterogeneous nature of the accident data analyzed, relative to the types of accident, road and vehicles involved, and so on, we opted for a methodology that first divided the accident data into homogeneous clusters, by means of cluster analysis, so as to make it possible to develop a Poisson or NB model for each cluster.

3.1. Cluster analysis

Cluster analysis classifies the cases into previously unknown groups. However, to obtain good results, the variables need to be carefully chosen, as the omission of

important variables may lead to totally incorrect analytical results. The analysis is based on analogous, be they contrasting, concepts of similarity and distance: the shorter the distance, the greater the similarity.

The most common method for measuring the distance between cases is the squared Euclidean distance, defined as the sum of squared distances between all the variables in two different groups:

$$d_{hk} = \sum_{v=1}^{p} w_v (x_{hv} - x_{kv})^2$$

Where x_{hv} and x_{kv} are the coordinates of the two points, P_h and P_k in the Cartesian plan on the variable x_v, And w_v is the weight assigned to the variable x_v.

Prior to the analysis, the variables are standardized (divided by the standard deviation) such that the unit of measure does not affect their distance. Essentially, this involves working with standard deviations (z). Once the cases have been classified into groups, the absolute value of the correlation coefficient is widely used to measure the degree of similarity. Hierarchical clustering is the most common method for producing clusters whereby, once formed, the cluster does not undergo further partitioning. There are two types of hierarchical clustering

- agglomerative: that groups together the closest elements to form a single large cluster
- divisive: that starts with a single large cluster splitting it into smaller ones for each case.

The cases and the clusters are grouped together using criteria adopted at each step of the agglomeration, based on the distance and similarity matrix between all cases. Here we adopt agglomerative hierarchical clustering based on the inertia criterion and on Ward's method (Ward, 1963) that combines clusters in such a way that at each agglomeration the two clusters merged are those with the smallest increment in the sum of squared distances (within-cluster standard deviation).

$$DEV_T = \sum_{s=1}^{p} \sum_{i=1}^{n} (x_{is} - \overline{x}_s)^2 = \sum_{i=1}^{n} \sum_{s=1}^{p} (x_{is} - \overline{x}_s)^2$$

Where $\overline{x}_s$ is the mean of the variable s referred to the data set as a whole.

Given a partitioning into "g" groups, this deviation can be decomposed into:

$$DEV_{IN} = \sum_{k=1}^{g} \sum_{s=1}^{p} \sum_{i=1}^{n_k} \left(x_{is} - \overline{x}_{s,k}\right)^2$$

which is the deviation between groups referred to the p variables for the group "k,"

and where, $\overline{x}_{s,k}$ is the mean of the variable for the group "k";

$$DEV_{OUT} = \sum_{s=1}^{p}\sum_{k=1}^{g}\left(\overline{x}_{s,k} - \overline{x}_s\right)^2 n_k$$

which is the deviation between groups. As is known, $DEV_T = DEV_{IN} + DEV_{OUT}$.

Passing from k+1 to k groups (agglomeration) DEV_{IN} increases, while clearly DEV_{OUT} decreases.

The classes of homogeneous cases are thus identified by determining the distance between them. This method, starting from a single class for each case proceeds by successive aggregation until such time as a single "representative" group is created. The clustering output is visualized in a dendrogram that shows the newly formed clusters and the distance between them. The number of clusters is chosen using the distance between the two clusters being merged, easily drawn from the dendrogram. If when passing from k to k+1 groups the merging distance increases substantially, then the dendrogram needs to be pruned to k groups. To determine the entity of this increase, we can calculate the relative increment of the merging distance:

$$\delta_k = \frac{(d_k - d_{k+1})}{d_{k+1}}$$

setting K for maximum δ_{k} .

3.2. The regression model

The model estimated here is a generalized linear model (GLM), consisting of three components:

Error distribution: it is generally not normal distributed (as in classical linear models) but Poisson, NB, gamma, exponential, and so on. In this case the Poisson distribution is considered, while the NB distribution is used when overdispersion occurs.

Linear predictor: the model is structured such that each observed value of y is given as a predicted value, obtained by transforming the value yielded by the linear part of the model. This is denoted with η and is given by:

$$\eta = X\beta^T$$

that can be written component wise as follows:

$$\eta_i = \sum_{i=0}^{p} X_{ij}\beta_i$$

where, X denotes the values of the explanatory variables p and β are the unknown

parameters to be estimated. The right hand side of the above equation is known as linear structure.

The link function: links the mean value of the response variable y to its linear predictor:

$$\eta = g(\mu)$$

In other words, the linear predictor is given by the sum of the terms for each of the p parameters. Clearly this does not yield the value of y, except in the particular case where the link function is the identity function. The value of η is obtained by transforming the value of y by means of the link function, while the predicted value of y is obtained by applying the inverse of the link function used to η.

Under the assumption that the dependent variable follows a non-Gaussian distribution, the parameters $\beta_0, \ldots, \beta_p$ have been estimated with the maximum likelihood method, rather than with the least squares method, whereby the β_i are estimated by maximising the log-likelihood function, given by:

$$l(\mu, Y) = \sum_{j=1}^{n} l_j\left(Y_J; \mu_j\right) = \sum_{j=1}^{n} logf_j\left(Y_J; \mu_j\right)$$

where, f_j is the distribution of Y_j.: to each observation j is associated a log-likelihood value equal to the logarithm of the value that the distribution $f_j(Y)$ takes for μ_j.

The regression analysis consists basically of three stages:

1. Identify the explanatory variables associated with accident occurrence
2. Evaluate the significance of the variables
3. Evaluate the models obtained using goodness-of-fit indicators.

To identify the explanatory variables we chose arbitrarily one of the following two procedures depending on the circumstances:

the "step-out" procedure that, starting from the initial model implementation that includes all possible explanatory variables, successively eliminates those that are not significant, the elimination process starting from the variables with the highest p

the "step-in, procedure that, starting from a univariate model, tests all the possible explanatory variables one at a time, considering valid only those models where the variable introduced is significant.

We used the Wald (or z-test) and the p value tests to determine whether the independent variable was related in a statistically significant way to the dependent variable. We set a threshold of 0.05 as the significance level α above which the variable was considered not to be significant, corresponding to a 95% confidence interval.

Lastly, to measure the model's goodness-of-fit we used the following indicators:

R^2: generally measures the proportion of variance explained by the model. It varies between 0 and 1 and is calculated as follows:

$$R^2 = \frac{\sum_{i=1}^{n}(\mu_i{}^2) - \sum_{i=1}^{n}(\mu_i - y_i)^2}{\sum_{i=1}^{n}(\mu_i^2)}$$

where, μ_i = is the number of expected accidents, and y_i = is the number of observed accidents.

In Poisson regression R^2 is calculated as above but refers to the proportion of variance explained by the model:

$$\text{pseudo } R^2 = 1 - \frac{\lambda \text{ (complete model)}}{\lambda(\text{model with one parameter})}$$

where, λ represents the maximum likelihood estimate of variance.

AIC (Akaike's Information Criterion): provides a measure of the estimate of the statistical model taking into account both the goodness of fit and model complexity. It is defined as AIC = 2k-2ln(L), where k is the number of parameters in the statistical model and L the maximized value of the likelihood function. In choosing between the two models we opted for the model with the lower AIC.

4. Results

4.1. Data sources

Since the 1970s road traffic accident data in Italy have been collected using the report forms created by Italian Institute of Statistics (ISTAT). The database is compiled from various sources—ISTAT, Italian Automobile Club d'Italia (ACI) Ministry for Home Affairs, traffic police, Carabinieri, municipal police, statistics departments of provincial capitals and of certain provinces.

4.1.1. ISTAT CTT/INC report form ISTAT CTT/INC

Accident data are recorded by the authority present at the scene of the accident, by completing the CTT/INC report form Report of Personal Injury Road Accidents and subsequently notifying ISTAT. The report form contains all the details necessary for describing the accident. It consists of a general information section concerning location, day and time of the accident, as well as the reporting body and coordinator, plus nine sections containing the following information:

1. accident location: road class (urban, highway, motorway, etc.), number or name of road, distance mark or, if the accident has occurred in a built up area, outside house number

2. accident site: cross-section characteristics, condition of carriageway, type of road (junction, roundabout, straight stretch) conditions of road surface (dry, slippery, etc.), road signs/markings (none, vertical, horizontal or vertical and horizontal) weather conditions (fine, raining, snowing, high winds, etc.)
3. nature of the accident: collision type (head-on collision, rear-end collision)
4. type of vehicles involved: motorcycle, car, heavy duty vehicle (HDV), tram, etc.
5. presumed circumstances of the accident: this can help to reconstruct the dynamics of the accident selecting from a predefined list of possibilities (e.g., speeding, dangerous overtake)
6. vehicles involved: number plate, year of registration
7. consequences of accident on persons involved: information about drivers of vehicles involved (age, gender, type of driving license, year of issue, any injured or fatalities) or any passengers or pedestrians involved
8. name of person(s) killed
9. name(s) of injured and where hospitalized.

The most interesting information for statistical analyses is contained in sections 1 and 2, the accident location and site.

4.1.2. Traffic data collection

The most significant variable for a Safety Performance Function is the "average daily traffic flow," which is calculated as the total number of vehicles passing in both directions through a road segment per year divided by 365 (days in a year). Data are collected either by direct measurements with the aid of video cameras and traffic counters or by means of appropriate traffic models that, however, require long calibration procedures to be adapted to specific contexts.

4.1.3. Geometric data collection

Geometric data can be gleaned from a variety of sources:

Maps provided by local authorities
Topographical surveys
Road plans

or resorting to more approximate methods such as on site surveys or using information tools, though these are clearly less accurate.

The data collected are then processed to identify those factors most affecting accident frequency.

4.2. Data description

We analysed accident data for the period 2005 to 2010 in a small town in Italy with a population of around 20,000. A total of 414 accidents were reported during this time, of which four were fatal and 264 resulted in personal injuries. Each accident was described using 28 explanatory variables.

Besides accident data, average daily traffic (ADT) counts were also collected. For intersections, rather than multiply the ADT for the roads converging there (Greibe, 2003), we considered the total ADT, in other words the total volume of traffic entering the junction or roundabout.

The following considerations may be drawn from the preliminary analysis:

- The most common types of road accidents were broadside collisions (38.65%), followed by side collisions (19.81%) and rear-end collisions (16.43%)
- 38.16% of accidents occurred on two-way single carriageways, 31.16% on one-way single carriageways, and 29.23% on dual carriageways
- 48.07% of accidents occurred at junctions, 36.71% on straight stretches, and only 11.84% at roundabouts.

On the basis of the data gathered, three different prediction models were constructed for accidents at roundabouts, junctions and straight road segments.

Our starting point was the basic generalized linear model:

$$y = a\ ADT^{p_1} e^{\sum_{i=1}^{n} X_i \beta_i}$$

where ADT = average daily traffic (vehicles per day), X_i = are the explanatory variables, and p_1, β_i = the coefficients to be estimated.

The basic model was then adapted for the three situations examined (roundabouts, junctions, and straight sections).

4.3. Roundabout model

For implementing the model we followed the "step out" approach, starting from the basic model with three explanatory variables, namely, ADT, internal diameter, and number of roundabout arms.

The above model was rewritten in a different form, raising the ADT to a power by calculating its natural logarithm:

$$y = e^{\beta_0 + X_1\beta_{1+} X_2\beta_{2+} X_3\beta_3}$$

y = number of accidents occurring over the observed period;
X_1 = natural logarithm of average daily traffic (ADT);
X_2 = number of roundabout arms;
X_3 = internal diameter;
$\beta_0 \ldots\ldots \beta_3$ = coefficients to be determined.

We ran the model using the Poisson error distribution and taking "log" as the link function, then checked for any overdispersion. The following outputs are indicated:

- The position index of the estimated residuals

Table 1. iterations for roundabout model.

	1st Iteration				2nd Iteration				3rd Iteration			
Coefficients	Estimate	*SE*	z value	Pr(>|z|)	Estimate	*SE*	z value	Pr(>|z|)	Estimate	*SE*	z value	Pr(>|z|)
Intercept	−20.8377	15.43294	−1.35	0,177	−26.9065	10.6521	−2.526	0.0115	−19.589	5.5478	−3.531	0.00041
ADT	2.50991	1.8191	1.38	0,168	3.2229	1.2604	2.557	0.0106	2.314	0.5766	4.013	5.98E−05
Arms	−0.19922	0.35183	−0.566	0,571	−0.2769	0.3315	−0.835	0.4035	—	—	—	—
Internal diameter	0.01262	0.02464	0.512	0,609	—	—	—	—	—	—	—	—
Null deviance	1.7031 e+01 on 3 degrees of freedom				17.03085 on 3 degrees of freedom				17.03085 on 3 degrees of freedom			
Residual deviance	−1.9984 e-15 on 0 degrees of freedom				0.26973 on 1 degrees of freedom				0.96536 on 2 degrees of freedom			
Akaike Information Criterion	24.727				22.996				21.692			

ADT = average daily traffic.

Table 2. First four iterations junction model.

	Junction Model							
	1st Iteration				2nd Iteration			
Coefficients	Estimate	—	z value	Pr(>|z|)	Estimate	—	z value	Pr(>|z|)
Intercept	0.612611	1.648258	0.372	0.710	0.57949	1.63601	0.354	0.723
ADT	0.044337	0.15803	0.281	0.779	0.04894	0.15492	0.316	0.752
Width road A	0.032835	0.032997	0.995	0.32	0.03361	0.03244	1.036	0.300
Width road B	0.004782	0.034355	0.139	0.889	—	—	—	—
Arms	−0.077420	0.26079	−0.297	0.767	−0.07068	0.25592	−0.276	0.782
Null deviance		19.285 on 31 degrees of freedom				19.285 on 31 degrees of freedom		
Residual deviance		17.752 on 27 degrees of freedom				17.771 on 28 degrees of freedom		
Akaike Information Criterion		118.97				116.98		
	3rd Iteration				4th Iteration			
Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.26568	1.16776	0.228	0.820	—	—	—	—
ADT	0.06259	0.14572	0.430	0.668	0.0948	0.03397	2.791	0.0053
Width road A	0.02986	0.02956	1.010	0.313	0.02967	0.02950	1.006	0.3146
Null deviance		19.285 on 31 degrees of freedom				89.418 on 32 degrees of freedom		
Residual deviance		17.848 on 29 degrees of freedom				17.899 on 30 degrees of freedom		
Akaike Information Criterion		115.06				113.11		

ADT = average daily traffic.

Table 3. 5th iteration junction model.

	5th Iteration			
Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
ADT	0.12581	0.01276	9.862	<2e-16
Null deviance	89.418 on 32 degrees of freedom			
Residual deviance	18.856 on 31 degrees of freedom			
AIC	112.07			

ADT = average daily traffic; AIC = Akaike.

- A block for the coefficients β that comprise: the estimated values, standard error, result of the Wald test, and the p value
- Deviance of the null model (containing just the intercept)
- Deviance of the adapted model (containing all the introduced variables)
- AIC (Akaike) index

The results are shown in Table 1 for each iteration.

4.4. Junction model

For reasons of statistical significance we only considered those junctions where more than one accident had occurred. We chose as explanatory variables ADT, width of road A, width of road B, and number of arms. Here too we adopted a step out approach, starting with the model having the same functional form as the roundabout model:

$$y = e^{\beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4}$$

where y = number of accidents occurring over the observed period, X_1 = natural logarithm of average daily traffic (ADT), X_2 = width road A, X_3 = width road B, X_4 = number of arms, and $\beta_0 \ldots\ldots \beta_3$ = coefficients to be determined.

Similarly to the roundabout model we ran the model using the Poisson error distribution and taking "log" as the link function, then checked for any overdispersion. The results obtained for the different iterations are shown in Tables 2 and 3.

4.5. Straight section model

Here again we disregarded those cases where only one accident had occurred. The explanatory variables chosen in this case were ADT, number of access points along the road, and number of pedestrian crossings. Width of road was not taken into account as the entire length of the straight stretches was considered.

We used a different approach to construct this model, opting for the step in procedure. Thus, initially we ran the model in univariate form, evaluating the need or not to add further explanatory variables. Here too the basic model was rewritten

Table 4. First 4 iterations for straight section model (Poisson model).

	Straight Section Model							
	1st Iteration				2nd Iteration			
Coefficients	Estimate	*SE*	z value	Pr(>|z|)	Estimate	*SE*	z value	Pr(>|z|)
Intercept	0.6891	1.1942	0.577	0.564	-	-	-	-
ADT	0.1547	0.1417	1.092	0.275	0.23607	0.0113	20.81	<2e-16
Null deviance	56.732 on 14 degrees of freedom				301.09 on 15 degrees of freedom			
Residual deviance	55.547 on 13 degrees of freedom				55.88 on 14 degrees of freedom			
Akaike Information Criterion	113.71				112.04			
	3rd Iteration				4th Iteration			
Coefficients	Estimate	*SE*	z value	Pr(>|z|)	Estimate	*SE*	z value	Pr(>|z|)
ADT	0.17967	0.02046	8.780	<2e-16	0.17902	0.0204	8.77	<2e-16
Access points	0.03838	0.01036	3.704	0.0002	0.03149	0.0167	1.884	0.0595
Pedestrian crossing	-	-	-	-	0.01058	0.0202	0.524	0.6004
Null deviance	301.091 on 15 degrees of freedom				301.091 on 15 degrees of freedom			
Residual deviance	43.283 on 13 degrees of freedom				43.007 on 12 degrees of freedom			
Akaike Information Criterion	101.45				103.17			

ADT = average daily traffic.

Table 5. 5th iteration for straight section model (negative binomial model).

	5th Iteration (using negative binomial)			
Coefficients	Estimate	*SE*	z value	Pr(>|z|)
ADT	0.17217	0.03307	5.206	1.93E-07
Access points	0.04380	0.01870	2.342	0.0192
Null deviance		179.057 on 15 degrees of freedom		
Residual deviance		13.642 on 13 degrees of freedom		
Akaike Information Criterion		87.822		
Theta		4.17		
SE		2.22		
2 × log likelihood		−81.822		

ADT = average daily traffic.

Table 6. Models obtained.

Road Type	Error Distribution	Functional Form	Pseudo R^2
Roundabout	Poisson	$y = e^{(-19.5890 + 2.3140\ X1)}$	0.94
Junction	Poisson	$y = e^{(0.12581\ X1)}$	0.78
Straight	Negative Binomial	$y = e^{(0.17217\ X1 + 0.04380\ X2)}$	0.92

raising the ADT to a power by calculating its natural logarithm:

$$y = e^{\beta_0 + X_1\beta_1}$$

where y = number of accidents occurring over the observed period and X_1 = is the natural logarithm of ADT.

The results of the iterations are shown in Table 4.

Thus we opted for the model that included the variables *ADT* and *access points*. Examining the residual deviance we observed however a strong overdispersion ("dispersion test" = 4.11, four times the average) and for this reason chose the NB distribution (Table 5).

5. Conclusions

The aim of the present study was to develop an accident prediction model for urban roads. The first step consisted in a review of the models reported in the literature from which it emerged that none was transferable to contexts different from the one for which they were calibrated. Having analyzed and simplified the available data, we then proceeded to develop three different models for predicting accident frequency at roundabouts, junctions, and straight stretches of road. Each model shows that accident rates vary with risk factors. Thus it is possible to identify appropriate countermeasures to be implemented for reducing the risk of road accidents.

Generally speaking, the most representative explanatory variable was ADT, especially for the two types of intersections, junctions and roundabouts. No other explanatory variable was found to be particularly important for these two cases. On the other hand, for straight stretches the number of access points along the road also proved to be a representative variable.

The calibrated models are shown in Table 6, where X_1 = is the natural logarithm of ADT and X_2 = is the number of access points.

The models obtained provide a fairly satisfactorily estimate of accident frequency, as the results of the pseudo R^2 test indicate.

To assess their adaptability for accident prediction, the models were tested on portions of the road network different from those for which they had been calibrated (outside the study area), obtaining the following results.

5.1. Roundabout model

- In 75% of cases the estimated number of accidents deviates from the observed value by just one unit
- In the remaining 25%, by only 3 units.

5.2. Junction model

- In 25% of cases the model estimate matches the observed data
- In 53% of cases the model estimate exceeds the observed value by one unit
- In 12.5% of cases model estimate is one unit lower than the observed value
- In 9.5% of cases the model overestimates the observed value by between 2 and 6 units.

5.3. Straight section model

- In 13% of cases the predicted value matches the observed data
- In 67% of cases the predicted value is overestimated by 1 to 7 units with respect to the observed data
- In 13% of cases the predicted value is underestimated by 1 to 4 units with respect to the observed data
- In 7% of cases the value is substantially underestimated.

Notwithstanding the fact that the prediction models obtained provided, after verification, fairly reliable estimates of accident frequency, we should stress that they are intended only as basic models to be used as a starting point for further development that envisages:

- the inclusion of new explanatory variables
- analysis of the relationship between accident frequency and human characteristics and behavior that are the primary cause of road accidents
- comparison of the basic models used with more sophisticated ones to assess the real need to use the latter for obtaining improved estimates.

References

AASHTO. (2010). *The highway safety manual*. Washington, DC: American Association of State Highway Transportation Professionals.

Ackaah, W., & Salifu, M. (2011). Crash prediction model for two-lane rural highways in the Ashanti region of Ghana. *IATTS Research*, *35*, 34–40.

Aguero-Valverde, J. (2013). Multivariate spatial models of excess crash frequency at area level: Case of Costa Rica. *Accident Analysis & Prevention*, *59*, 365–373.

Barbosa, H., Cunto, F., Bezerra, B., & Nodari, C. (2014). Safety performance model for urban intersection in Brasil. *Accident Analysis and Prevention*, *70*, 258–266.

Bauer, K. M., & Harwood, D. W. (2000). *Statistical models of at-grade intersection accidents—addendum* (FHWA-RD-99-094). Washington, DC: U.S. Department of Transportation.

Cafiso, S., Di Graziano, A., Di Silvestro, G., La Cava, G., & Persaud, B. (2010). Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. *Accident Analysis and Prevention*, *42*, 1072–1079.

Caliendo, C., De Guglielmo, M. L., & Guida, M. (2013). A crash-prediction model for road tunnels. *Accident Analysis & Prevention*, *55*, 107–115.

Canale, S., Leonardi, S., & Pappalardo, G. (2005). The reliability of the urban road network: Accident forecast models. In *Proceedings of 3rd International SIIV* (pp. 1–22). Bari, Italy, September 22–24, 2005.

Dong, C., Clarke, D. B., Yan, X., Khattak, A., & Huang, B. (2014). Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. *Accident Analysis & Prevention*, *70*, 320–329.

Dong, C., Richards, S. H., Clarke, D. B., Zhou, X., & Ma, Z. (2014). Examining signalized intersection crash frequency using multivariate zero-inflated Poisson regression. *Safety Science*, *70*, 63–69.

El-Basyouny, K., & Sayed, T. (2009). Collision prediction models using multivariate Poisson-lognormal regression. *Accident Analysis & Prevention*, *41*(4), 820–828.

Eustace, D., Aylo, A., & Mergia, W. Y. (2015). Crash frequency analysis of left-side merging and diverging areas on urban freeway segments—A case study of I-75 through downtown Dayton, Ohio. *Transportation Research Part C: Emerging Technologies*, *50*, 78–85.

Geedipally, S. R., Lord, D., & Dhavala, S. S. (2012). The negative binomial-Lindley generalized linear model: Characteristics application using crash data. *Accident Analysis and Prevention*, *45*, 258–265.

Greibe, P. (2003). Accident prediction models for urban roads. *Accident Analysis & Prevention*, *35*(2), 273–285.

Hosseinpour, M., Yahaya, A. S., & Sadullah, A. F. (2014). Exploring the effects of roadway characteristics on the frequency severity of head-on crashes: Case studies from Malaysian federal roads. *Accident Analysis and Prevention*, *62*, 209–222.

Husman, J. A., Hall, B., & Griliches, Z. (1984). Econometric models for count data with an application to the patents-R&D relationship. *Econometrica*, *52*(4), 909–938.

Johansson, P. (1996). Speed limitation and motorway casualties: A time series count data regression approach. *Accident Analysis & Prevention*, *28*(1), 73–87.

Kumara, S., & Chin, H. (2003). Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. *Traffic Injury Prevention*, *4*, 53–57.

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

Lord, D., & Persaud, B. N. (2000). Accident prediction models with and without trend: application of the generalized estimating equation (GEE) procedure. *Transportation Research Board*, *00–0496*, 14.

Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis and Prevention*, *37*(1), 35–46.

Ma, M., Yan, X., Abdel-Aty, M., Huang, H., & Wang, X. (2010). Safety analysis of urban arterials under mixed-traffic patterns in Beijing. *Transportation Research Record*, *2193*, 105–115.

Maher, M. J. (1990). A bivariate negative binomial model to explain traffic accident migration. *Accident Analysis & Prevention*, *22*(5), 487–498.

Malyshkina, N. V., & Mannering, F. L. (2010). Zero state Markov switching count-data models: An empirical assessment. *Accident Analysis and Prevention*, *42*(1), 131–139.

Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future direction. *Analytic Methods in Accident Research*, *1*, 1–22.

Marchionna, A., Perco, P., & Falconetti, N. (2012). Evaluation of the applicability of IHSDM crash prediction module on Italian two-lane rural roads. *SIIV—5th International Congress—Sustainability of Road Infrastructures*, *53*, 933–942. Roma: Procedia—Social and Behavioral Sciences.

Marchionna, A., Perco, P., & Tavernar, F. C. (2008). Transferability of accident prediction models for urban intersections. *Proceedings of the 17th National SIIV Congress*, Enna, Italy, 10–12 September 10–12, 2008.

Martinelli, F., La Torre, F., & Vada, P. (2009). Calibration of the highway safety manual's accident prediction model for Italian secondary road network. *Transportation Research Record: Journal of the Transportation Research Board, 2103*, 1–9.

Miaou, S. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regression. *Accident Analysis and Prevention*, *26(*4), 471–482.

Mohammadi, M. A., Samaranayake, V. A., & Bham, G. H. (2014). Crash frequency modeling using negative binomial models: An application of generalized estimating equation to longitudinal data. *Analytic Methods in Accident Research*, *2*, 52–69.

Ng, C. M., Ong, S.-H., & Srivastava, H. M. (2010). A class of bivariate negative binomial distributions with different index parameters in the marginals. *Applied Mathematics and Computation*, *217*(7), 3069–3087.

Rakha, H., Arafeh, M., Abdel-Salam, A. G., Guo, F., & Flintsch, A. M. (2010). *Linear regression crash prediction models: Issue and proposed solutions*. Washington, DC: Virginia Tech Transportation Institute—U.S. Department of Transportation.

Sacchi, E., Persaud, B., & Bassani, M. (2012). Assessing international transferability of highway safety manual crash prediction algorithm and its components. *Transportation Research Record*, *2279*, 90–98.

Savolainen, P. T., Mannering, F., Lord, D., & Quddus, M. A. (2011). The statistical analysis of crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis and Prevention*, *43*(5), 1666–1676.

Šenk, P., Ambros, J., Pokorný, P., & Striegler, R. (2013). Use of accident prediction models in identifying hazardous road locations. *Transactions on Transport Sciences*, *5*(4), 223–232.

Shankar, V. N., Albin, R. B., Milton, J. C., & Mannering, F. L. (1998). Evaluating median crossover likelihoods with clustered accident counts: An empirical inquiry using the random effects negative binomial model. *Transportation Research Record*, *1635*, 44–48.

Shankar, V., Mannering, F., & Barfield, W. (1995). Effects of roadway geometrics and environmental factors of rural freeway accident frequencies. *Accident Analysis and Prevention*, *27*(3), 371–389.

Shankar, V., Milton, J., & Mannering, F. (1997). Modeling accident frequencies as zero-altered probability process an empirical inquiry. *Accident Analysis and Prevention*, *29*(6), 829–837.

Summersgill, I., Kennedy, J., Hall, R. D., Hickford, A. J., & Barnad, S. R. (2001). *Accidents at junctions on one-way urban roads*. Wokingham, England: Transport Research Laboratory.

Van Petegem, J. H., & Wegman, F. (2014). Analyzing road design risk factors for run-off-road crashes in the Netherlands with crash prediction models. *Journal of Safety Research*, *49*, 121–127.

Wang, K., Lee, A. H., Yau, K. K., & Carrivick, P. J. (2003). A bivariate zero-inflated Poisson regression model to analyze occupational injuries. *Accident Analysis & Prevention*, *35*(4), 625–629.

Wang, X., & Abdel-Aty, M. (2006). Temporal and Spatial analyses of rear-end crashes at signalized intersection. *Accident Analysis and Prevention*, *38*(6), 1137–1150.

Ward, J. H. Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.