

# Accounting for Dispersion and Correlation in Estimating Safety Performance Functions. An Overview Starting from a Case Study

Orazio Giuffrè<sup>1</sup>, Anna Granà<sup>1</sup>, Tullio Giuffrè<sup>2</sup> & Roberta Marino<sup>1</sup>

<sup>1</sup> Department of Civil, Environmental, Aerospace and Materials Engineering, Università degli Studi di Palermo, Palermo, Italy

<sup>2</sup> Faculty of Engineering and Architecture, Kore University, Enna, Italy

Correspondence: Anna Granà, Department of Civil, Environmental, Aerospace and Materials Engineering, Università degli Studi di Palermo, Viale delle Scienze al Parco d'Orleans, Palermo 90128, Italy. Tel: 39-91-2389-9718. E-mail: anna.grana@unipa.it

Received: November 20, 2012 Accepted: December 18, 2012 Online Published: January 15, 2013

doi:10.5539/mas.v7n2p11

URL: <http://dx.doi.org/10.5539/mas.v7n2p11>

## Abstract

In statistical analysis of crash count data, as well as in estimating Safety Performance Functions (SPFs), the failure of Poisson equidispersion hypothesis and the temporal correlation in annual crash counts must be considered to improve the reliability of estimation of the parameters. After a short discussion on the statistical tools accounting for dispersion and correlation, the paper presents the methodological path followed in estimating a SPF for urban four-leg, signalized intersections. Since the case study exhibited signs of underdispersion, a Conway-Maxwell-Poisson Generalized Linear Model (GLM) was fitted to the data; then a quasi-Poisson model in the framework of Generalized Estimating Equations (GEEs) was performed in order to account for correlation.

Results confirm that dispersion and correlation are phenomena that cannot be eluded in the estimation of SPFs under penalty of loss of efficiency in estimating model parameters. Generalized Estimating Equations overcome this problem allowing to incorporate together dispersion and temporal correlation when a quasi-Poisson distribution is used for modeling crash data. Moreover, whereas GEE regression is handy (many statistical software packages have already implemented GEE functions), the interest of COM-Poisson regression, because of difficulties in interpreting the model parameters and in arranging COM-Poisson codes, is still limited to the research field.

**Keywords:** safety performance function, signalized intersections, COM-Poisson model, road safety

## 1. Introduction

It is well-known that Safety Performance Functions (SPFs) allow to estimate the expected number of crashes on entities (road sections or intersections). Differently from models to evaluate the potential accident rate (Mauro & Cattani, 2004), through a mathematical equation SPFs express the average crash frequency of a site as a function of traffic flow and other site geometric and/or functional characteristics. In statistical analysis of crash count data some problems must be addressed to improve the reliability of the estimations. Data and methodological issues associated with crash-frequencies are widely discussed by Lord and Mannering (2010) and by Turner et al. (2011); two of these issues will be focused in the following. First, the data structure can invalidate equidispersion hypothesis on which the Poisson model is founded; moreover, the same crash-data generating process makes inefficient estimations based on the traditional Poisson model. In order to relax the Poisson assumption of equidispersion, linear exponential family models incorporating a dispersion parameter, as well as quasi-likelihood methods, represent a potential solution to this question. Second, with many years of data, it is necessary to account for the year-to-year variations in crash counts because of the influence of factors that can change every year. This creates a temporal correlation that affects the reliability of the SPF estimates obtained through traditional estimation procedures of the model.

Starting from these considerations, the main purpose of the paper is to show how estimates efficiency can be improved taking into account either dispersion and temporal correlation in annual crash counts. In case of overdispersion, the Generalized Estimating Equations (GEEs) method with Negative Binomial distribution (NB1

or NB2) is an effective tool to address the mentioned problems and to increase efficiency of estimates (Lord, 2006); these models take into account for overdispersion by means of a parameter called overdispersion parameter  $\alpha$  (with  $\alpha > 0$ ); moreover, it is possible to account for correlation, too. Underdispersion, instead, is a phenomenon which has been less convenient to model directly than overdispersion, mainly because it is less commonly observed. Oh et al. (2006) report that crash data show characteristics of underdispersion especially in cases where the sample is small and the sample mean is very low; underdispersion can be also caused by the data generating process that is independent from the size of the sample or its mean. To further improve estimation of the parameters in case of underdispersion, the Conway-Maxwell-Poisson (COM-Poisson) distribution can be introduced (Lord et al., 2010; Giuffrè et al., 2011). Unfortunately, currently COM-Poisson distribution cannot be used in GEE context; so, the analyst that decides to estimate a SPF with GLM method and COM-Poisson distribution is likely to overlook correlation phenomenon running the risk of impairing estimates efficiency.

Focusing on the latter question, the paper presents a case study that exhibited signs of underdispersion. The convenience of handling underdispersion through a GLM COM-Poisson model rather than a GEE quasi-Poisson model is evaluated comparing the efficiency of the obtained estimates. It has to be noted that the small sample size used in the study could have affected the estimation of model parameters (coefficients and dispersion parameter). Therefore, though results can help to highlight the potential of COM-Poisson model and of GEE quasi-Poisson model in handling under-dispersed data, further researches should be carried out using different dataset (namely larger sample size) to confirm them.

## 2. Method

The methodological approach applied for estimating a Safety Performance Functions at 4-leg signalized intersections is described in the following sections. The intersections here analyzed are characterized by being inserted in urban area, factor that may directly affect the expected number of crashes. Model performance measures used to verify the suitability of the predictive model are also introduced. Before introducing the case study, the main issues related to the dispersion and the temporal correlation of the data to be considered in the treatment of crash data are extensively accounted for and discussed.

### 2.1 Accounting for Dispersion

Poisson and Negative Binomial (NB) distributions in the context of Generalized Linear Models (GLM) have been widely used for some time to analyze crash count data in the estimation of Safety Performance Functions (SPFs). The basic regression model for count data is in fact the Poisson model:  $y_{ij} | x_{ij} \sim \text{Poisson}$  with  $E(y_{ij}|x_{ij}) = \mu_{ij} = \exp(\mathbf{x}_{ij}'\boldsymbol{\beta})$ , where  $y_{ij}$  denotes the crash count data at year  $t$  and at site  $j$ ;  $\mathbf{x}_{ij} = [x_{1ij}, \dots, x_{kij}]$  is a  $k$  dimensional vector of covariates at year  $t$  and at site  $j$ ;  $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_k]$  is the vector of parameters to be estimated.

In statistical analysis of crash count data, the Poisson model is often inadequate because of its implicit restriction on the distribution of observed crash counts; specifically in the Poisson model the variance of the random variable is constrained to equal the mean; this property is usually called equidispersion. Different factors can invalidate equidispersion hypothesis in the data: consequently, data often exhibit overdispersion (i.e. the variance is larger than the mean) and, occasionally they exhibit underdispersion with the mean exceeding the variance. The overdispersion can be caused by various factors, such as data clustering, unaccounted temporal correlation, model misspecification, but it has been shown to be mainly attributed to unobserved heterogeneity: counts are viewed as being generated by a Poisson process, but the analyst ignores that the rate parameter ( $\mu$ ) is itself a random variable and so he (or she) does not specify any distribution for it. In order to relax the Poisson assumption of equidispersion, quasi-likelihood methods represent a potential solution. By this way few assumptions about the distribution for the dependent variable are required; only the relationship between the outcome mean and the covariates, and between the mean and variance must be specified (McCullagh & Nelder, 1989): the variance  $v_{ij}$  of the observation  $y_{ij}$  is expressed as a known function,  $g$ , of the expectation  $\mu_{ij}$ , i.e.  $v_{ij} = \phi g(\mu_{ij})$  where  $\phi$  is the scale parameter. Then a quasi-Poisson distribution can be used to model crash data: the mean is the same of the Poisson mean and  $g(\mu_{ij}) = \mu_{ij}$ ; the variance is now a function of the mean:  $v_{ij} = \phi \mu_{ij} = (1 + \alpha)\mu_{ij}$ , where  $\alpha$  is the dispersion parameter. In the case of under-dispersion  $\alpha < 0$  (and thus  $0 < \phi < 1$ ); in the opposite case ( $\alpha > 0$  and  $\phi > 1$ ) the data are overdispersed. Although different Poisson-based distributions have been developed to accommodate over-dispersion, the most common distribution for crash count data remains the Poisson-gamma or Negative Binomial (NB) distribution. Properties of the traditional NB model were illustrated by Cameron and Trivedi (1998). The NB model arises mathematically by assuming that unobserved heterogeneity across sites is gamma distributed, while crashes within sites are Poisson distributed. According to Hauer (1997), the NB distribution offers a simple way to handle overdispersion, especially when the final equation has a closed form; in addition, the mathematics to manage the relationship between the mean and the

variance structures is relatively simple. There are two well-known nonnested forms of the negative binomial model, denoted as NB1 and NB2 (Greene, 2007). In overdispersed mixture models the Poisson mean is a random variable; when it is gamma distributed the model form will be  $y_{ij}|x_{ij} \sim \text{NB}$  with  $E(y_{ij}|x_{ij}) = \mu_{ij}$ ,  $\text{var}(y_{ij}|x_{ij}) = \mu_{ij} + \alpha\mu_{ij}^p$ ; for  $p$  equals 1 we obtain the quasi-Poisson distribution above discussed. The choice of the specific form of NB depends on data set; NB2 can only accommodate overdispersion and it cannot be used in the case of underdispersion. Application of mixture models for crash data are referred by Park and Lord (2009). In NB regression the overdispersion parameter  $\alpha$  is commonly assumed the same for all entities (intersection and/or road segment). Modeling crashes on road sections, Hauer (2001) showed that this assumption may have undesirable consequences when road sections in the data base differ in length; if model parameters are estimated by maximum likelihood, the relative influence of long road sections is much diminished whereas very short road sections exert an unduly large influence. The assumption of an overdispersion parameter constant within the sites can also lead to an inconsistency when the safety of a road section by the Empirical Bayes method is estimated (Hauer, 2001). In modeling crash-flow relationships for urban intersections, Miaou and Lord (2003) also challenged the assumption of fixed dispersion parameter. Because of the complexity and interaction of traffic flow in and around intersections, they supposed that the unmodeled heterogeneity of the mean of crash counts would be spatially structured. This means that the variance of NB models is not a simple function of mean, but contains a dispersion function that depends on site-specific characteristics such as major and minor road traffic flows. As well as overdispersion, underdispersion can violate some basic count data modeling assumptions; Winkelmann et al. (1995) proposed a correction for an underdispersed event count probability distribution. Several researchers recently have proposed new and innovative methods for analyzing under-dispersed crash data. Moreover, the Conway-Maxwell-Poisson (COM-Poisson) distribution has been re-introduced by statisticians to model count data characterized by either over- or under-dispersion (Shmueli et al., 2005; Guikema & Coffelt, 2008; Lord et al., 2010; Zou et al., 2011). The COM-Poisson distribution was first introduced in 1962 by Conway and Maxwell; only in 2008 it was evaluated in the context of a GLM by Guikema and Coffelt (2008), Lord et al. (2008) and Sellers and Shmueli (2010). The COM-Poisson distribution is a two parameter generalization of the Poisson distribution that is flexible enough to describe a wide range of count data distributions (Sellers & Shmueli, 2010); since its revival, it has been further developed in several directions and applied in multiple fields (Sellers et al., 2011).

For a random variable  $Y_{ij}$  (e.g., a discrete count at year  $t$  and at site  $j$ ) COM-Poisson probability distribution function is given by the equation:

$$P(Y_{ij} = y_{ij}) = \frac{\lambda_{ij}^{y_{ij}}}{(y_{ij}!)^\nu Z(\lambda_{ij}, \nu)} \quad (1)$$

where:

$$Z(\lambda_{ij}, \nu) = \sum_{s=0}^{\infty} \frac{\lambda_{ij}^s}{(s!)^\nu} \quad (2)$$

$\lambda_{ij}$  = a centering parameter, denoting the expected value under a Poisson distribution associated with the generic observation at year  $t$  and at site  $j$  (Sellers and Shmueli, 2010);

$\nu (\geq 0)$  = the dispersion parameter (where  $\nu < 1$  for over-dispersion and  $\nu > 1$  for under-dispersion).

This formulation allows for a non-linear decrease in ratios of successive probabilities in the form:

$$\frac{P(Y_{ij} = y_{ij} - 1)}{P(Y_{ij} = y_{ij})} = \frac{y_{ij}^\nu}{\lambda_{ij}} \quad (3)$$

Shmueli *et al.* (2005) refer that the serie  $\lambda^s/(s!)^\nu$  converges for any  $\lambda > 0$  and  $\nu > 0$ , since the ratio of two subsequent terms of the serie  $\lambda/s^\nu$  tends to 0 as  $s \rightarrow \infty$ . Moments of COM-Poisson distribution can be expressed using the recursive formula (Shmueli et al., 2005):

$$E\left(Y_{ij}^{r+1}\right)=\begin{cases} \lambda_{ij} E^{1-\nu}\left(Y_{ij}+1\right) & r=0 \\ \lambda_{ij} \frac{\partial}{\partial \lambda_{ij}} E\left(Y_{ij}^r\right)+E\left(Y_{ij}\right) \cdot E\left(Y_{ij}^r\right) & r>0 \end{cases} \quad (4)$$

Using an asymptotic approximation for  $Z(\lambda_{ij}, \nu)$ ,  $E(Y_{ij})$  can be closely approximated by:

$$E\left(Y_{ij}\right)=\lambda_{ij} \frac{\partial \log Z\left(\lambda_{ij}, \nu\right)}{\partial \lambda_{ij}} \approx \lambda_{ij}^{1/\nu}-\frac{(\nu-1)}{2\nu} \quad (5)$$

This approximation is especially good for  $\nu \leq 1$  or  $\lambda_{ij} > 10^\nu$ . Once the COM-Poisson regression model has been estimated, fitted values can be computed by equation 5, setting:

$$\hat{\lambda}_{ij}=\exp\left(x_{ij}'\hat{\beta}\right) \quad (6)$$

## 2.2 Accounting for Correlation

Count data often consist of observations over several time periods, these are usually referred to as longitudinal data or panel data. A data structure of this kind creates a specific problem in safety modeling because of the failure of the independence hypothesis for the variate response. With respect to the precision of the parameters estimation, this is a serious issue in safety modeling; that is why elusion of the correlation within responses can lead to misleading conclusions in model interpretation on the basis of incorrect estimates of the variances and of an inefficient or biased estimate of the regression coefficients (Diggle et al., 2002; Giuffrè et al., 2007). Literature refers on several applications by GEE models (see by way of example, Lord & Persaud, 2000; Cafiso & D'Agostino, 2012).

It is well known that a robust model estimation based on Generalized Estimating Equations (GEEs) can still supply consistent estimates of the regression parameters even if the correlation matrix is incorrectly specified (Fitzmaurice, 1995). Standard application of GEEs to safety analysis uses robust (or sandwich) estimates of regression coefficients under an independence hypothesis for the working correlation matrix. Nevertheless it has been demonstrated that the efficiency of estimators declines as the correlation increases, and the decline becomes appreciable when the correlation is greater than 0.4 (Fitzmaurice, 1995). Furthermore, efficiency losses - when independence is a false assumption - will seriously compromise the significance of estimates for within-subject correlations greater than 0.5. Errors are particularly large when the correlation is highly positive or highly negative. Other researchers think that the search for the right correlation matrix becomes important only when marginal models are estimated by using data with missing values (Lord & Persaud, 2000); nevertheless, they agree that standard errors of the coefficients usually are underestimated when temporal effects are not included in the modeling framework (Hardin & Hilbe, 2003). As above referred, quasi-likelihood methods allow to relax Poisson assumption of equidispersion. The GEEs method is an extension of the quasi-likelihood approach (Liang & Zeger, 1986); by this method, once a proper distribution has been selected for the data set (e.g., Poisson, Quasi-Poisson, Negative Binomial), it is possible to improve the efficiency of parameters estimation specifying a "working" correlation matrix, in order to explicitly take into account for the correlation within observations. Parameters estimates can be found by solving the following estimating equation:

$$\sum_{t=1}^m \sum_{j=1}^n D_{ij}' V_{ij}^{-1} \left[ Y_{ij} - \mu_{ij} \right] = 0 \quad (7)$$

for the year  $t$  ( $t=1, 2, \dots, m$ ) and for the entity  $j$  ( $j=1, 2, \dots, n$ )

$$D_{ij}=\frac{\partial \mu_{ij}}{\partial \beta}=\begin{bmatrix} \frac{\partial \mu_{ij}}{\partial \beta_1} & \frac{\partial \mu_{ij}}{\partial \beta_2} & \dots & \frac{\partial \mu_{ij}}{\partial \beta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_{mj}}{\partial \beta_1} & \frac{\partial \mu_{mj}}{\partial \beta_2} & \dots & \frac{\partial \mu_{mj}}{\partial \beta_k} \end{bmatrix} \quad (8)$$

where:

$V_{ij}$  = covariance matrix at year  $t$  and site  $j$

$\beta$  = vector of regression parameters  $(\beta_1, \dots, \beta_k)'$

$\mu_{ij} = n \times I$  vector  $(\mu_{1j}, \mu_{2j}, \dots, \mu_{mj})'$  of expected values for the  $j^{\text{th}}$  site.

With the usual notation we can put  $\mu_{ij} = g^{-1}(X_{ij} \beta)$ , where  $g$  is referred to as the “link” function. The temporal correlation in responses is described by means of a  $m \times m$  matrix  $R(\gamma)$ , where  $\gamma$  represents the type of correlation with  $\gamma = (\gamma_1, \dots, \gamma_{m-1})'$  and  $\gamma_j = \text{corr}(Y_{t_r, j}, Y_{t_s, j})$  for  $r, s = 1, \dots, m-1$  and  $r \neq s$ . The covariance matrices can be expressed as follows:

$$V_{ij} = A_{ij}^{1/2} R_{ij}(\gamma) A_{ij}^{1/2} \quad (9)$$

$$\text{cov}(\hat{\beta}) = \sigma^2 \left[ \sum_{t=1}^m \sum_{j=1}^n D_{ij}' V_{ij}^{-1} D_{ij} \right]^{-1} \quad (10)$$

where  $A_{ij}$  is a diagonal matrix containing the variances of the elements of  $Y_{ij}$ , expressed in terms of  $\beta$ . The simultaneous solution of above equations with the iterative weighted least squares method gives the GEEs estimates for  $\beta$  and for the correlation type ( $\gamma$ ). Since it is not possible to know the proper correlation type for repeated observations, Liang and Zeger (1986) proposed the use of a “working” correlation matrix by replacing in the upper basic equation  $V_{ij}$  with  $\hat{V}_{ij}$  based on the correlation matrix  $\hat{R}_{ij}$ .

Then, from the above-mentioned structure of the GEE procedure it has to be highlighted that the specification of the form of response correlation represents the central issue in obtaining more efficient estimates. In fact, although GEE models are generally robust to misspecification of the correlation structure (Liang & Zeger, 1986), when the specified structure does not incorporate all the information on the correlation of measurements within the subjects, loss of efficiency in estimates can be expected (Ballinger, 2004). In order to get the true correlation structure, it is necessary to test on different hypotheses of within-subject correlation. It is usually possible to choose from different types of correlation structures (e.g., *independence*, *exchangeable*, *unstructured*, *autoregressive*, *m-order dependence*). Assuming the independence structure (i.e. assuming subjects are independent of each other), one has to sacrifice the advantage of using GEE (because it does not consider the within-subject correlation); nevertheless, the independence structure can still be useful in fitting a base model. The exchangeable structure supposes no logical ordering for within-entity observations; when an unstructured working correlation matrix is chosen, estimates of all possible correlations of within-entity responses are made and they are included in the estimates of the variance. The m-order dependence structure implies that the  $\gamma_s$  take different values at different time points. Finally, for data that are correlated within cluster over time, an autoregressive correlation structure can be appropriated; in this case correlation within subject is specified as an exponential function of the lag time period (Ballinger, 2004).

In general, decisions about correlation structure should be guided first by theory; there are specific correlation structures that are appropriate for time-dependent correlation structures (e.g. autoregressive) and some that are not (e.g. exchangeable). For cases in which analyst may be undecided between few structures, Pan (2001) proposed a test that extends the Akaike’s Information Criterion to allow comparison of covariance matrices under GEEs models to the covariance matrix generated under the independence hypothesis (Quasi-likelihood under the Independence model Criterion, QIC). The correlation structure with the QIC score closest to zero is judged to be the most appropriate. Applications of QIC in choosing the best correlation structure for a marginal GEEs model, as well as some useful general guidelines, are given by Hardin & Hilbe (2003).

### 2.3 Model Performance Measures

Technical literature suggests different goodness-of-fit methods to evaluate predictive performance of models and to find the model that best explains the data among all estimated models. The methods used in this paper include the following (where the subscript “ $i$ ” denotes the generic observation at year  $t$  and at site  $j$ ):

#### 2.3.1 Mean Prediction Bias (MPB)

MPB gives a measure of the magnitude and direction of the average model bias (Oh et al., 2003). If the MPB is positive then the model over-predicts crashes and if the MPB is negative then the model under-predicts crashes. It is computed using the following equation:

$$MPB = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i) \quad (11)$$

where  $N$  is the sample size,  $\hat{y}_i$  and  $y_i$  are the predicted and observed crashes at site  $i$  respectively.

### 2.3.2 Mean Absolute Deviance (MAD)

MAD gives a measure of the average mis-prediction of the model (Oh et al., 2003). The model that provides MAD closer to zero is considered to be the best among all the available models. It is computed using the following equation:

$$MAD = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (12)$$

### 2.3.3 Mean Squared Predictive Error (MSPE)

MSPE is typically used to assess the error associated with a validation or external data set (Oh et al., 2003). The model that provides MSPE closer to zero is considered to be the best among all the available models. It can be computed using the following equation:

$$MSPE = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (13)$$

### 2.3.4 Akaike Information Criterion (AIC)

The AIC (Akaike, 1974) is a measure of the goodness-of-fit of an estimated statistical model and is defined as:

$$AIC = -2 \log L + 2p \quad (14)$$

where:

$L$  = the maximized value of the likelihood function for the estimated model;

$p$  = the number of parameters in the statistical model.

The AIC methodology is used to find the model that best explains the data with a minimum of free parameters, penalizing models with a large number of parameters. The model with the lowest AIC is considered to be the best model among all available models. The following expressions of log-likelihood were used in order to compute AIC:

Poisson and quasi-Poisson distributions:  $\log L = y_i \log \mu_i - \log (y_i!)$

COM-Poisson distribution:  $\log L = \sum_{i=1}^n y_i \log \lambda_i - \nu \sum_{i=1}^n \log (y_i!) - \sum_{i=1}^n \log Z(\lambda_i, \nu)$

*Quasilikelihood under the Independence model Criterion (QIC)*

As above referred, since the GEE method is a quasi-likelihood based method, an extension of the Akaike's information criterion is needed to compare covariance matrices under GEE models to the covariance matrix generated under the independence hypothesis. So AIC statistic is replaced by the QIC statistic, defined as (Pan, 2001):

$$QIC = -2Q + 2p \quad (15)$$

where  $Q$  is the quasi-likelihood function ( $Q = L/\phi$ ) and  $p$  is the number of parameters in the statistical model. The model with the lowest QIC is considered to be the best model among all available models.

*The marginal  $R^2$ -test*

The marginal  $R^2$ -test supplies a measure of improvement in fit between the estimated model and the intercepted-only one; it compares predicted values from the model (after it is estimated) against the actual values (observations) and against the squared deviations of the observations from mean values for the response variable:

$$R_m^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y}_i)^2} \quad (16)$$

$R^2$  statistics can be interpreted as the amount of variance in the response variable that is explained by the fitted model (Hardin & Hilbe, 2003).

## 3. Estimating SPF for Urban Four-Leg Signalized Intersections - A Case Study

### 3.1 Data Description

This section provides an overview of the characteristics of the data set used in this study.

The data collected at nineteen urban, four-leg, signalized intersections in the Municipality of Palermo City road network included: crash occurred from 2000 to 2007, crash-related maneuvers, infrastructure characteristics and traffic volumes; crash data were directly acquired from reports available at the Municipal Police Force; only fatal and injury crashes were considered. Note that, since eight years of crash data were considered as distinct observations, there were  $8 \times 19 = 152$  observations. Table 1 shows yearly and 8-year crash statistics (minimum, median, mean, maximum), as well as total crashes for the entire dataset. Table 2 reports mean values over 8-years (2000-2007) of crashes and Annual Average Daily Traffic (AADT) for each site.

Table 1. Annual Crash Statistics, All Collision Types, 2000 - 2007

Year	Minimum	Median	Mean	Maximum	Total
2000	0	3	3.74	13	71
2001	0	3	3.63	7	69
2002	0	4	4.21	10	80
2003	0	3	3.58	8	80
2004	0	3	3.47	9	66
2005	1	3	3.42	10	65
2006	0	3	3.63	9	69
2007	0	3	3.68	13	70
2000-07	0	3	3.67	13	558

Description: Total crashes, yearly and 8-year crash statistics for the entire dataset.

Table 2. Mean values of AADT<sub>major</sub>, AADT<sub>minor</sub> at intersections and crashes

Intersection	(AADT <sub>major</sub> ) <sub>m</sub> [ $10^3$ veh/d]	AADT <sub>minor</sub> [ $10^3$ veh/d]	(Crashes) <sub>m</sub> [crash/year]
1	41.257	8.010	2.25
2	16.772	9.852	1.75
3	14.237	10.864	2.13
4	20.777	19.748	2.88
5	31.688	15.798	3.00
6	25.499	13.194	3.50
7	26.499	15.866	5.00
8	21.910	21.164	3.38
9	27.085	24.153	4.50
10	28.945	27.403	9.75
11	28.092	10.588	4.88
12	30.653	7.000	5.50
13	31.825	17.615	7.38
14	9.056	6.713	0.50
15	17.876	12.975	1.38
16	22.218	13.991	3.00
17	20.592	9.986	2.50
18	18.629	16.101	4.75
19	20.677	10.187	1.75

Description: 2000-2007 crashes and Annual Average Daily Traffic (AADT) by site.

### 3.2 Model Selection and Covariates

It is well known that the development of a Safety Performance Function (SPF) involves: i) which explanatory variables should be used; ii) how variables should enter into the model, i.e. the best model form. The results of

these tasks can be summarized as follows:

- all the covariates explored and the significant ones (at the 15% confidence level) are listed in Table 3;
- different model forms were investigated considering the combinations of all the significant variables listed in Table 3. The exploratory analysis suggested to insert into the model only the covariates listed on the right column in Table 3; it also revealed that the functional model form could be described using the power function for the variables  $(F_1 + F_2)$  and  $RW_2$ , and the exponential function for the variable  $PW_1$ . Note that the variable  $F_1$  was not introduced in the model because it was accounted in variable  $(F_1 + F_2)$ ;  $R_1$  was not selected because its introduction in the model together with the other significant variables, either in the power or in the exponential form, produced no appreciable benefits on the model performance. Then the final selected model had the form:

$$y_{ij} = \beta_0 (F_1 + F_2)_{ij}^{\beta_1} RW_{2j}^{\beta_2} e^{\beta_3 PW_{1j}} \quad (17)$$

where:

$y_{ij}$  = expected number of crashes for the year  $t$  and the intersection  $j$ ;

$(F_1 + F_2)_{ij}$  = sum of Annual Average Daily Traffic on major- and minor-road for the year  $t$  and the intersection  $j$ ;

$RW_{2j}$  = minor-road roadway width at the intersection  $j$ ;

$PW_{1j}$  = major-road permitted ways at the intersection  $j$  ( $PW_1 = 0$  for one way only,  $PW_1 = 1$  for two ways or more);

$\beta_0, \beta_1, \beta_2, \beta_3$  = parameters to be estimated.

Table 3. Variables explored and selected

Variables	Abbreviation	Significant variables	Selected
Annual Average Daily Traffic on major-road	$F_1$	✓	
Annual Average Daily Traffic on minor-road	$F_1 + F_2$	✓	✓
Major-road roadway width	$RW_1$		
Minor-road roadway width	$RW_2$	✓	✓
Major-road number of lanes	$NL_1$		
Minor-road number of lanes	$NL_2$		
Major-road red light time	$R_1$	✓	
Minor-road red light time	$R_2$		
Major-road permitted ways	$PW_1$	✓	✓
Minor-road permitted ways	$PW_2$		

Description: Variables represent all the possible covariates; in the third column only variables explored as significant are reported. The fourth column includes only the variables selected as covariates of the model.

### 3.3 Accounting for Dispersion in GLM Regression

Generalized Linear Model was performed to estimate model coefficients assuming a Poisson error distribution and a quasi-Poisson distribution.

After model estimation with quasi-Poisson distribution it was observed that data clearly exhibited underdispersion ( $\alpha < 0$ ). Then, to further improve parameters estimates, the COM-Poisson distribution was used.

For estimating Poisson and quasi-Poisson regression coefficients and standard errors GenStat software was used; since GenStat does not provide standard error for dispersion parameter, it was estimated iterating the following auxiliary regression (Cameron & Trivedi, 1998):

$$\frac{(y_{ij} - \hat{y}_{ij})^2 - y_{ij}}{\hat{y}_{ij}} = \alpha + \varepsilon_{ij} \quad (18)$$

where  $\hat{y}_{ij}$  is the fitted value estimated by the Poisson model. Once the dispersion parameter was estimated its value was used to obtain new estimates of model parameters; this iteration was performed until all the values (i.e., dispersion parameter, coefficients) converged.

COM-Poisson model was estimated using *R* software, after implementation of codes arranged by Sellers and Shmueli (2010), available at [www9.georgetown.edu/faculty/kfs7/research](http://www9.georgetown.edu/faculty/kfs7/research).

Table 4 shows coefficients estimates, standard errors and goodness-of-fit statistics for the Poisson model (model 1), the quasi-Poisson model (model 2), the COM-Poisson model (model 3). COM-Poisson coefficients are for centering parameter  $\lambda$  (see equation 6) and not for mean as in the case of Poisson/quasi-Poisson model (Lord *et al.* 2010; Sellers & Shmueli 2010); this is why COM-Poisson coefficients cannot be directly compared with the Poisson/quasi-Poisson ones.

Table 4. Coefficients estimates and goodness-of-fit for the three models

variables	Model 1			Model 2			Model 3		
	est	s.e.	t	est	s.e.	t	est <sup>(*)</sup>	s.e.	t
Constant ( $\beta_0$ )	-6.94	0.77	-9.01	-6.94	0.60	-11.49	-12.41	1.72	-7.22
$F_1 + F_2$ ( $\beta_1$ )	1.69	0.14	12.15	1.69	0.11	15.49	3.18	0.40	7.75
$RW_2$ ( $\beta_2$ )	0.73	0.17	4.23	0.73	0.14	5.40	1.38	0.28	4.79
$PW_1$ ( $\beta_3$ )	0.22	0.09	2.38	0.22	0.07	3.04	0.39	0.13	2.96
$\nu$	-	-	-	-	-	-	1.99	0.25	-
$\alpha$	-	-	-	-0.46	0.06	-	-	-	-
<i>MPB</i>	0.00	-	-	0.00	-	-	0.01	-	-
<i>MAD</i>	1.09	-	-	1.09	-	-	1.09	-	-
<i>MSPE</i>	1.98	-	-	1.98	-	-	1.97	-	-
<i>AIC</i>	543	-	-	543	-	-	519	-	-

<sup>(\*)</sup> model parameters to be used for determining  $\hat{\lambda}_i$  according to equation 6.

Description: Coefficients estimates, standard errors and goodness-of-fit statistics for the three estimated models: the Poisson model (model 1), the quasi-Poisson model (model 2), the COM-Poisson model (model 3).

From results showed in Table 4, although there is no difference between parameters estimates (and GOF) for Poisson and quasi-Poisson models, it can be seen that the consideration of underdispersion in the data improves the estimates accuracy, as it is shown by the reductions in the standard errors values.

The shape parameter of the COM-Poisson distribution again shows under-dispersion ( $\nu > 1$ ). This confirms that the Poisson distribution is not appropriate to interpret the data-set. *MPB*, *MAD* and *MSPE* values of the models have slight differences, so they do not add any significant information about the models prediction capacity; on the contrary, *AIC* values indicate that the COM-Poisson model has to be considered the best among all estimated models.

### 3.4 Accounting for Correlation through GEE Regression

Considering that data consisted of repeated measures over the years that could be correlated within an entity, it seemed appropriate to account for the correlation within responses. For this reason GEE regressions were fitted under different working correlation matrices, these are assuming that repeated observations were correlated in different ways.

Again GenStat software was used for this purpose. As unfortunately up to now software packages do not allow to

perform a GEE model using a COM-Poisson distribution, we used the quasi-Poisson distribution to consider simultaneously both the correlation and the under-dispersion in the data.

As mentioned earlier, three forms of correlation were explored starting from the simplest one (*independence structure*) for which observations are thought (unrealistically) to be uncorrelated. In contrast to the hypothesis of independence, an *unstructured* working correlation matrix was stated to allow the free estimates on the within-site correlation from the data. A correlation structure of a stationary *7-dependence process* was assumed, too. The GEE regression results under the three named working correlation matrices are summarized in Table 5, in which both  $R_m^2$  and the Pan statistic (QIC) are shown.

Table 5. Coefficients estimates and goodness-of-fit in GEEs

variables	independence			unstructured			7-dependence		
	est	s.e.	t	est	s.e.	t	est	s.e.	t
Constant ( $\beta_0$ )	-6.94	0.82	-8.46	-7.63	0.66	-11.56	-7.35	0.63	-11.67
$F_1 + F_2$ ( $\beta_1$ )	1.69	0.09	18.78	1.78	0.08	22.25	1.77	0.08	23.11
$RW_2$ ( $\beta_2$ )	0.73	0.24	3.04	0.87	0.18	4.83	0.79	0.19	4.21
$PW_1$ ( $\beta_3$ )	0.22	0.10	2.20	0.31	0.08	3.88	0.29	0.08	3.42
$\alpha^{(*)}$	-0.47			-0.45			-0.48		
MPB	0.00			1.19			0.27		
MAD	1.09			1.14			1.15		
MSPE	1.98			2.09			2.12		
QIC	1009			1008			1032		
$R_m^2$	0.68			0.66			0.65		

(\*) the unstructured working correlation matrix allows a dispersion parameter varying over time in the observation period; the mean value is reported in the table.

Description: GEE regression results under the *independence*, *unstructured* and *7-dependence* working correlation matrices and model performance measures.

The results in Table 5 show that unstructured and 7-dependence working correlation matrices give parameters estimates more accurate than under independence hypothesis; that is particularly evident with regard to the traffic variable ( $F_1 + F_2$ ), which most affects the model (i.e. the one with the highest parameter value). Nevertheless, the hypothesis of a correlation matrix different from the independent one does not provide any significant improvement in QIC and do not allow the best correlation structure to be determined clearly.

Thus it was decided to thoroughly analyze the model adequacy from another point of view.

According to the purpose of the current research it was thought that the width of the confidence interval for the mean could represent a criterion for model selection, as well as for deciding about the correlation structure of response.

As estimates of GEE parameters ( $\beta_{GEE}$ ) are asymptotically normal, it can be derived that an asymptotic  $(1-\alpha)100\%$  confidence interval for the mean ( $x'_{\beta_{GEE}}$ ) is given by:

$$x'_{\beta_{GEE}} \pm z_{1-\frac{\alpha}{2}} \sqrt{x'V_{GEE}x}$$

Table 6 shows  $\sqrt{x'V_{GEE}x}$  values for models estimated under unstructured, dependence and independence hypothesis for the working correlation matrix.

It can be easily seen that relaxing the independence hypothesis, the confidence interval is considerably reduced both in the case of unstructured and dependence correlation matrix.

That confirms benefits in explicitly considering correlation in the data as allowed by a GEE procedure.

Table 6.  $\sqrt{x'V_{GEE}x}$  sample values for different hypothesis of correlation matrix

correlation matrix	min	mean	max	median
unstructured	0.042	0.079	0.127	0.077
dependence	0.043	0.078	0.121	0.079
independence	0.054	0.093	0.156	0.091

Description: Above values are related to models estimated under unstructured, dependence and independence hypothesis for the three working correlation matrix.

#### 4. Discussion and Conclusion

Results reported in the previous section highlight that the dispersion and the correlation are phenomena that cannot be eluded in the estimation of SPFs under penalty of loss of efficiency in estimating model parameters. Generalized Estimating Equations (GEEs) procedure overcomes this problem because it allows to incorporate together the dispersion in the data and the temporal correlation. Moreover, GEE regression using different working correlation matrices (i.e. assuming that repeated observations are correlated in different ways) allows to gain a better understanding of the proper correlation structure in the crash count data.

The paper presents an application of the GLM and the GEE procedures in developing a SPF; data of the case study pertain to a sample of nineteen urban, four-leg, signalized intersections in Palermo, Italy, for the years 2000-2007. Since data were found to exhibit clear signs of underdispersion together with a quasi-Poisson distribution a COM-Poisson was considered in the GLM context; then a GEE quasi-Poisson model was performed under three different working correlation matrices.

Through the case study it was shown that:

- quasi-Poisson and COM-Poisson GLM regression model allow to handle underdispersion and to obtain more accurate estimates for the model parameters than the traditional Poisson model;
- according to the Akaike Information Criterion, COM-Poisson regression further improves the predictive performance of the proposed model and it provides a better goodness-of-fit than the quasi-Poisson model, at least for the case study dataset;
- GEE quasi-Poisson model, especially under working correlation matrices different from the independent one, allows more accurate estimates of model parameters than the correspondent GLM model (that do not account for the temporal correlation in crashes).

It has to be noted that regression models based on COM-Poisson distribution, despite their benefits, have disadvantages in terms of model estimation due to:

- difficulties in interpreting the model parameters ( $\lambda$ ,  $\nu$ ), in obtaining fitted values and in comparing coefficients from a COM-Poisson regression model to those from other models (in fact, the comparison is possible only in terms of fitted values);
- difficulties in arranging COM-Poisson codes, that is actually possible using few statistical softwares just in GLM context (for example *R* software or WinBUGS package); this makes the interest in using COM-Poisson regression limited to the research field;
- difficulties in accounting for temporal correlation in the data since up to now it cannot be used in GEE context.

It follows that the practical utility of COM-Poisson regression may be limited only to the cases of underdispersed data. On the contrary, in more frequent cases of overdispersion, Negative-Binomial GEE models allow to obtain correct estimates for model parameters accounting simultaneously both for correlation and for dispersion in the data. This is handy since many statistical software packages have already implemented GEE functions.

#### References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723. <http://dx.doi.org/10.1109/TAC.1974.1100705>
- Ballinger, G. A. (2004). Using generalized Estimating equations for longitudinal data analysis. *Organizational Research Methods*, 7(2), 127-150. <http://dx.doi.org/10.1177/1094428104263672>

- Cafiso, S. D., & D'Agostino, C. (2012). Safety performance function for motorways using generalized estimation. *Procedia-Social and Behavioral Sciences*, 53(3 October, 2012), 901-910. <http://dx.doi.org/10.1016/j.sbspro.2012.09.939>
- Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*, *Econometric Society Monograph No. 30*. Cambridge, UK: Cambridge University Press.
- Conway, R. W., & Maxwell, W. L. A. (1962). Queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12, 132-136.
- Diggle, P. J., Heagerty, P., Liang, K. Y., & Zeger, S. L. (2002). *Analysis of Longitudinal Data* (2nd ed.). New York, NY: Oxford University Press Inc.
- Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometric*, 51(1), 309-317.
- Giuffrè, O., Granà, A., Giuffrè, T., & Marino, R. (2007). Improving reliability of road safety estimates based on high correlated accident counts. *Transportation Research Record*, 2019, 197-204. <http://dx.doi.org/10.3141/2019-23>
- Giuffrè, O., Granà, A., Marino, R., & Corriere, F. (2011). Handling Underdispersion in Calibrating Safety Performance Function at Urban, Four-Leg, Signalized Intersections. *Journal of Transportation Safety & Security*, 3(3), 174-188. <http://dx.doi.org/10.1080/19439962.2011.599014>
- Greene, W. (2007). Functional forms for the negative binomial model for count data. *Economics Letters*, 99(3), 585-590. <http://dx.doi.org/10.1016/j.econlet.2007.10.015>
- Guikema, S. D., & Coffelt, J. P. (2008). A Flexible Count Data Regression Model for Risk Analysis. *Risk Analysis*, 28(1), 213-223. <http://dx.doi.org/10.1111/j.1539-6924.2008.01014.x>
- Hardin, J. W., & Hilbe, J. M. (2003). *Generalized Estimating Equations*. London, UK: Chapman & Hall/CRC Press.
- Hauer, E. (1997). *Observational before-after studies in road safety. Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Oxford, UK: Pergamon Press.
- Hauer, E. (2001). Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation. *Accident Analysis & Prevention*, 33(6), 799,808. [http://dx.doi.org/10.1016/S0001-4575\(00\)00094-4](http://dx.doi.org/10.1016/S0001-4575(00)00094-4)
- Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13-22. <http://dx.doi.org/10.1093/biomet/73.1.13>
- Lord, D. (2006). Modelling motor vehicle crashes using Poisson-Gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis & Prevention*, 38(4), 751-766. <http://dx.doi.org/10.1016/j.aap.2006.02.001>
- Lord, D., Geedipally, S. R., & Guikema, S. D. (2010). Extension of the Application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crash Data Exhibiting Under-Dispersion. *Risk Analysis*, 30(8), 1268-1276. <http://dx.doi.org/10.1111/j.1539-6924.2010.01417.x>
- Lord, D., Guikema, S. D., & Geedipally, S. (2008). Application of the Conway-Maxwell- Poisson Generalized Linear Model for Analyzing Motor Vehicle Crashes. *Accident Analysis & Prevention*, 40(3), 1123-1134. <http://dx.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), 291-305. <http://dx.doi.org/10.1016/j.tra.2010.02.001>
- Lord, D., & Persaud, B. (2000). Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations (GEE) Procedure. *Transportation Research Record*, 1717, 102-108. <http://dx.doi.org/10.3141/1717-13>
- Mauro, R., & Cattani, M. (2004). Model to evaluate potential accident rate at roundabouts. *Journal of Transportation Engineering*, 130(5), 602-609. [http://dx.doi.org/10.1061/\(ASCE\)0733-947X\(2004\)130:5\(602\)](http://dx.doi.org/10.1061/(ASCE)0733-947X(2004)130:5(602))
- McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). London, UK: Chapman & Hall/CRC.

- Miaou, S. P., & Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. *Transportation Research Record, 1840*, 31-40. <http://dx.doi.org/10.3141/1840-04>
- Oh, J., Lyon, C., Washington, S. P., Persaud, B. N., & Bared, J. (2003). Validation of the FHWA Crash Models for Rural Intersections: Lessons Learned. *Transportation Research Record, 1840*(1), 41-49. <http://dx.doi.org/10.3141/1840-05>
- Oh, J., Washington, S. P., & Nam, D. (2006). Accident Prediction Model for Railway-Highway Interfaces, *Accident Analysis & Prevention, 38*(2), 346-56. <http://dx.doi.org/10.1016/j.aap.2005.10.004>
- Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics, 57*, 120-125.
- Park, B. J., & Lord, D. (2009). Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention, 41*(4), 683-691. <http://dx.doi.org/10.1016/j.aap.2009.03.007>
- Sellers, K. F., & Shmueli, G. (2010). A flexible regression model for count data, *The Annals of Applied Statistics, 4*(2), 943-961. <http://dx.doi.org/10.1214/09-AOAS306>
- Sellers, K. F., Borle, S., & Shmueli, G. (2011). The COM-Poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Business and Industry*. <http://dx.doi.org/10.1002/asmb.918>
- Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution, *Journal of the Royal Statistical Society: Part C, 54*(1), 127-142. <http://dx.doi.org/10.1111/j.1467-9876.2005.00474.x>
- Turner, S., Persaud, B., Lyon, C., Bassani, M., & Sacchi, E. (2011). International crash experience comparisons using prediction models. *Road and Transport Research, 20*(4), 16-27.
- Winkelmann, R., Signorino, C., & King, G. (1995). A Correction for an Underdispersed Event Count Probability Distribution, *Political Analysis, 5*, 215-228.
- Zou, Y., Lord, D., & Geedipally, S. R. (2011). Over- and Under-Dispersed Crash Data: Comparing the Conway-Maxwell-Poisson and Double-Poisson Distributions. *91st TRB Annual Meeting*, January 22-26, 2012. Retrieved from <http://ceprofs.civil.tamu.edu/dlord/>