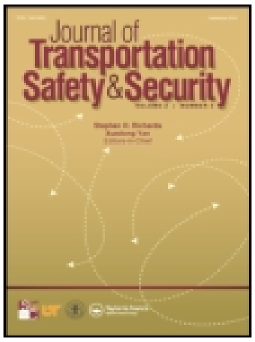

Taylor & Francis
Taylor & Francis Group

Journal of Transportation Safety & Security

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

Handling Underdispersion in Calibrating Safety Performance Function at Urban, Four-Leg, Signalized Intersections

Orazio Giuffrè , Anna Granà , Marino Roberta & Ferdinando Corriere

To cite this article: Orazio Giuffrè , Anna Granà , Marino Roberta & Ferdinando Corriere (2011) Handling Underdispersion in Calibrating Safety Performance Function at Urban, Four-Leg, Signalized Intersections, Journal of Transportation Safety & Security, 3:3, 174-188, DOI: 10.1080/19439962.2011.599014

To link to this article: https://doi.org/10.1080/19439962.2011.599014

Published online: 02 Sep 2011.

Journal of Transportation Safety & Security, 3:174–188, 2011
Copyright © Taylor & Francis Group, LLC and The University of Tennessee
ISSN: 1943-9962 print / 1943-9970 online
DOI: 10.1080/19439962.2011.599014

Handling Underdispersion in Calibrating Safety Performance Function at Urban, Four-Leg, Signalized Intersections

ORAZIO GIUFFRÈ,[1] ANNA GRANÀ,[1] MARINO ROBERTA,[2] AND FERDINANDO CORRIERE[2]

[1]Department of Civil, Environmental and Aerospace Engineering – Università degli Studi di Palermo, Palermo, Italy
[2]Department of Land and City – Università degli Studi di Palermo, Palermo, Italy

Poisson basic assumption of equidispersion is often too much restrictive for crash count data; in fact this type of data has been found to often exhibit overdispersion. Underdispersion has been less commonly observed, and this is the reason why it has been less convenient to model directly than overdispersion. Overdispersion and underdispersion are not the only issues that can be a potential source of error in specifying statistical models and that can lead to biased crash-frequency predictions; these issues can derive from data properties (temporal and spatial correlation, time-varying explanatory variables, etc.) or from methodological approach (omitted variables, functional form selection, etc.). This article focuses on the potential of the Conway-Maxwell (COM-Poisson) model in handling underdispersion that arose in the development of a Safety Performance Function for urban four-leg signalized intersections; other issues, as temporal data correlation, have been intentionally eluded to test the best way of handling underdispersion. Results confirmed that the COM-Poisson model properly handled crash data set for which neither Poisson nor negative binomial model were able to account for dispersion phenomenon; they also showed that the COM-Poisson model provided a good statistical performance and a better goodness-of-fit than the quasi-Poisson and the traditional Poisson model.

Keywords safety performance function, signalized intersections, COM-Poisson model, under-dispersion

1. Introduction

Safety Performance Functions (SPFs) have been used since long time in safety assessments such as establishing relationships between safety and roadway infrastructures, predicting crashes, and so on.

Technical literature shows several studies on the development of SPFs in relation to different types of road infrastructures. It also suggests a guidance for the calibration procedure (Harwood, Council, Hauer, Hughes, & Vogt, 2000; Persaud & Dzbik, 1993;

Address correspondence to Anna Granà, Department of Civil, Environmental and Aerospace Engineering – Università degli Studi di Palermo, Viale delle Scienze al Parco d'Orleans, Palermo, Italy 90128. E-mail: anna.grana@unipa.it

Vogt, 1999; Vogt & Bared, 1998). Bonneson, Zimmerman, and Fitzpatrick (2005) reported the comparison among SPFs properly calibrated considering different types of road traits. Specific SPFs were calibrated for road intersections in urban and rural area: for signalized intersections (Bauer & Harwood, 2000; Canale, Leonardi, & Pappalardo, 2005; Losa & Terrosi Axerio, 2005; Lyon, Haq, Persaud, & Kodama, 2005) and for stop-controlled intersections (Bauer & Harwood, 2000; Canale et al., 2005), as well as for roundabouts (Maycock & Hall, 1984; Giuffrè, Graná, Giuffrè, & Marino, 2007).

In highway safety studies, the Poisson distribution remains the most common probabilistic model used for analyzing crash data (Hauer, 2004; Lord, 2006). Nevertheless, Poisson basic assumption of equidispersion (i.e., equality between mean and variance values) is often too much restrictive for crash count data. In fact this type of data has been found to often exhibit overdispersion (i.e., the variance is greater than the mean); in few cases underdispersion (i.e., the variance is less than the mean) has been found. If the data exhibit overdispersion, the negative binomial model (Hauer, 2004; Lord, 2006) is commonly used to fit the data; it takes into account for overdispersion by a parameter α called overdispersion parameter (with $\alpha > 0$). However, the negative binomial (or Poisson-Gamma) model has limitations for its inability to handle underdispersed data; furthermore problems in estimation of dispersion parameter can arise when the sample mean value is low and the sample size is small (Clark & Perry, 1989; Dean, 1994; Lord, 2006; Lord & Mahlawat, 2009; Piegorsch, 1990).

Underdispersion is a phenomenon that has been less convenient to model directly than overdispersion, mainly because it is less commonly observed. According to Oh, Washington, and Nam (2006), the underdispersion can be data related (i.e., small sample and low sample mean); it can be also caused by the data-generating process that is independent from the size of the sample or its mean.

Differently from the case of overdispersed data, options in selecting a proper distribution are limited when modeling underdispersed count data, e.g., the gamma model proposed by Winkelman (1995) and the negative binomial 1 (NB1) model (Cameron & Trivedi, 1998). Several researchers recently have proposed new and innovative methods for analyzing underdispersed crash data. As part of these new methods the Conway-Maxwell-Poisson (COM-Poisson) distribution has been reintroduced by statisticians to model count data that are characterized by either over- or underdispersion (Geedipally, Guikema, Dhavala, & Lord, 2008; Guikema & Coffelt, 2008; Kadane, Shmueli, Minka, Borle, & Boatwright, 2006; Lord, Geedipally, & Guikema, 2010; Shmueli, Minka, Kadane, Borle, & Boatwright, 2005). The COM-Poisson distribution was first introduced in 1962 by Conway and Maxwell, but only in 2008 it was evaluated in the context of a generalized linear model (GLM) by Guikema and Coffelt (2008), Lord, Guikema, and Geedipally (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); since its revival, it has been further developed in several directions and applied in multiple fields (Sellers, Borle, & Shmueli, in press).

Overdispersion and underdispersion are not the only issues that can be a potential source of error in specifying statistical models and that can lead to biased crash-frequency predictions; these issues can derive from data properties (temporal and spatial correlation, time-varying explanatory variables, etc.) or from methodological approach (omitted variables, functional form selection, etc.). A comprehensive discussion on data and methodological issues in statistical analysis of crash data can be found in Lord and Mannering (2010).

This article focuses on the issue of underdispersion that arose in the development of a SPF at urban four-leg signalized intersections; at the moment other issues have been intentionally eluded to test the best way of handling underdispersion. With regard to the temporal data correlation, due to the use of each year as distinct observation, an in-depth discussion can be found in Giuffrè et al. (2007).

The SPF was calibrated considering Poisson, quasi-Poisson, and COM-Poisson models; comparisons among the statistics of the parameters estimates and the model goodness-of-fit were made. Results indicated that the COM-Poisson model can properly handle crash data when the model output shows signs of underdispersion; they also showed that the COM-Poisson model provides a good statistical performance and a better goodness-of-fit than the quasi-Poisson and the traditional Poisson model.

2. Data Description

To accomplish the study, 19 urban four-leg signalized intersections have been selected in the Palermo City, Italy, road network. For each intersection crashes occurred from 2000 to 2007 were directly collected from reports available at the Municipal Police Force.

Only fatal and injury crashes were considered; the data set included the date and the time of day when crashes occurred, the condition of traffic lights (working or not working), the environmental conditions regarding the pavement and the presence of work zones, the type and the number of involved users, the maneuvers and the road (major or minor) the users came from. In the observation period 558 crashes in total were collected.

Table 1 shows yearly and 8-year crash statistics (minimum, median, mean, maximum) for the entire data set; Figure 1 shows the 8-year fatal and injury crashes occurred at the selected urban, four-leg, signalized intersections.

The analysis by crash type was carried out only on a part of the sample, that is, for crashes occurred during the 4-year period from 2003 to 2006. Results of this analysis are shown in Table 2. Multiple-vehicle crashes occurred much more frequently than single-vehicle crashes; within multiple-vehicle crashes, right-angle collisions and left-turn

Table 1
Annual crash statistics, all collision types, 2000–2007

	Total Crashes				
Year	Minimum	Median	Mean	Maximum	Total
	19 urban, four-leg, signalized intersections				
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	68
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
2000–2007	0	3	3.67	13	558

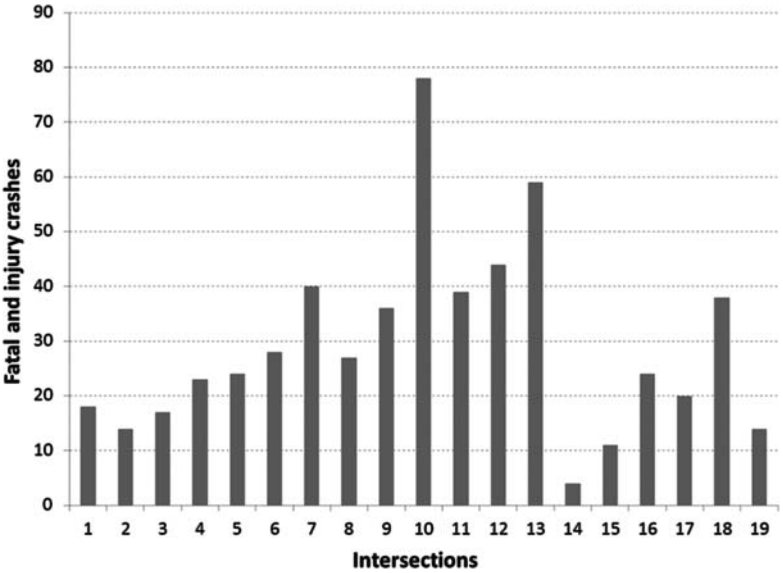

Figure 1. Crashes at selected intersections, 2000–2007.

collisions were characterized by the highest percentage values. Collisions with pedestrians were much more frequent within single-vehicle crashes. Considering the time band, 45% of crashes occurred in the afternoon and in the evening (3.00 p.m.–11.00 p.m.), hours generally characterized by high traffic volume.

Table 2
Distributions of fatal and injury crashes by type, 2003–2006

Crash Type	Number (percentage) of fatal and injury crashes	Number (percentage) of injury crashes
	Single-vehicle crashes	
Collision with pedestrian	9 (5)	9 (7)
Collision with parked car	4 (2)	2 (2)
Collision with fixed object	2 (1)	0 (0)
Ran off road	1 (1)	0 (0)
Total single-vehicle crashes	16 (9)	11 (8)
	Multiple-vehicle crashes	
Right-angle collision	74 (43)	58 (44)
Left-turn collision	46 (26)	38 (29)
Rear-end collision	15 (9)	10 (8)
Right-turn collision	11 (6)	6 (5)
Head-on collision	0 (0)	0 (0)
Sideswipe collision	3 (2)	3 (2)
Other multiple-vehicle collision	9 (5)	9 (4)
Total multiple-vehicle crashes	158 (91)	120 (92)

Table 3
Geometric and operational characteristics at the selected intersections

	Number of intersections			Number of intersections			Number of Intersections	
Feature	Major street	Minor street	Feature	Major street	Minor street	Feature	Major street	Minor street
Number of lanes			Roadway width (m)			Permitted way system		
1	4	4	a[a]	3	5	One-way only	9	11
2	9	12	b[a]	11	11	Two-way	1	8
3	6	3	c[a]	5	3			

[a] a: $w \leq 10$ m; b: 10 m $< w \leq 15$ m; c: $w > 15$ m.

Table 3 shows geometric and operational characteristics of major and minor roads at the examined urban, four-leg, signalized intersections. According to operational conditions, intersections can also be classified in four types as follows:

- Type 1: intersections with two-way at major road and minor road (15%)
- Type 2: intersections with two-way at major road and one-way at minor road (32%)
- Type 3: intersections with one-way at major road and two-way at minor road (32%)
- Type 4: intersections one-way at major road and minor road (21%).

To complete the sample characterization traffic data surveys were carried out during 2007. The type of vehicle and its maneuvers (through vehicle, right- and left-turning vehicle) were also distinguished. Major-road Annual Average Daily Traffic (AADT) and minor-road AADT were estimated using method proposed by Association Suisse des Professionnels de la route et du traffic (1995).

Starting from traffic count data, AADT values in the previous years were estimated for each intersection and for each year according to vehicle registrations from 2000 to 2006 (Automobile Club of Italy [ACI], 2005, 2008). Figure 2 shows crash mean values per year and AADT for each intersection for major road and for minor road. Crashes and AADT values in Figure 2 are mean values over 8 years (2000–2007), including traffic counts performed in 2007 and the estimates in the previous years.

3. Modeling Approach, Issues, and Results

This section shows the methodological path followed to develop a SPF for the examined case study. Yearly fatal and injury crashes occurred at the selected intersections in the observation period (19 × 8 observations in total) were used.

It is well known that the development of the SPF involves (1) which explanatory variables should be used, (2) whether and how variables should be grouped, and (3) how variables should enter into the model, that is, the best model form. In general, a safety performance function has the following form:

$$\mathrm{E}\{\kappa_{\mathrm{tj}}\} = \mathrm{f}(\mathbf{X},\ \beta) \tag{1}$$

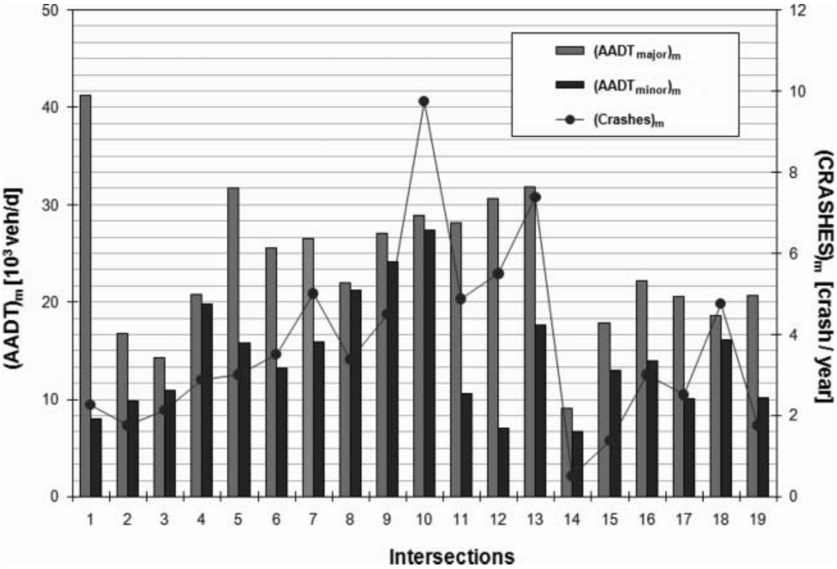

Figure 2. Mean values of crashes and AADT$_{major}$ and AADT$_{minor}$ at the intersections. AADT$_{major}$ = annual average daily traffic on major-road; AADT$_{minor}$ = annual average daily traffic on minor-road.

where,

$E\{\kappa_{tj}\}$ = expected number of crashes per unit of time t at the site j
$\mathbf{X}$ = vector of covariates, $x_{1tj}, x_{2tj}, \ldots, x_{ptj}$
β = coefficients to be estimated, $\beta_0, \beta_1, \ldots, \beta_p$.

Equation 1 is used to predict the number of crashes per unit of time on a given transportation facility. The main goal of Equation 1 is to estimate the coefficients β associated with the covariates (or explanatory variables). The techniques for finding these coefficients are very well developed and they typically use GLM through maximum likelihood methods.

Coefficients estimates are primarily affected by the included covariates and the functional model form selected to explain the relationship between $E\{\kappa_{tj}\}$ and each covariate.

The covariates explored in the current study are listed in Table 4; they were selected having in mind statistical models of intersection crashes referred in literature over last 10 years (Bauer & Harwood, 2000; McGee, Taori, & Persaud, 2003; Lyon et al., 2005). The variables marked on the right column in Table 4 were the only found to be significant at the 15% confidence level and then were included in the model specification.

Investigated model forms are reported in Table 5; they were calibrated considering the combinations of all the variables listed in Table 4. The results of this exploratory analysis revealed that the functional relationship between crashes and the significant covariates could be described using the power function for the variables AADT$_1$ and RW$_2$, and the exponential function for the variable PW$_1$. Then the final selected model had the form:

$$y_{tj} = \beta_0 \mathrm{AADT}_{1tj}^{\beta_1} \mathrm{RW}_{2j}^{\beta_2} e^{\beta_3 \mathrm{PW}_{1j}} \tag{2}$$

where,

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

Table 4
Variables explored and selected

Variables	Abbreviation	Significant Variables
Annual Average Daily Traffic on major-road	$AADT_1$	√
Annual Average Daily Traffic on minor-road	$AADT_2$	
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	

√ significant at the 15% confidence level.

$AADT_{1tj}$ = Annual Average Daily Traffic on major 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.

Generalized linear model (Cameron & Trivedi, 1988; Nelder & Wedderburn, 1972; McCullagh & Nelder; 1989) was performed to estimate model coefficients using the software package Genstat (Payne, Lane, & Digby, 2008) and assuming as base model the Poisson error distribution, consistent with the state of research in developing these models (Hauer, 2004; Lord, 2006).

On the other hand, to test, a priori, whether the Poisson distribution is an appropriate model for the data set, the Hoaglin's Poissonness plot (Hoaglin, 1980) was performed. This plot is used to determine if the specified distribution (Poisson) is appropriate for the data set. When data follow a Poisson distribution the Hoaglin's plot will give points along a straight line; on the contrary, when the data deviate from a Poisson, the points will be curved. Figure 3 shows the plot for the data set of the current study and the associated 95% confidence interval. According to Hoaglin, crash data were split into classes of assigned yearly size; in this way, the x-coordinate of the plot was the "class" of occurred crashes, the y-coordinate was

Table 5
Model forms investigated

Name	Model Form
Power function	$y = \beta_0 X^{\beta}$
Exponential function	$y = \beta_0 e^{\beta X}$
Gamma function	$y = \beta_0 X^{\beta 1} e^{\beta 2X}$

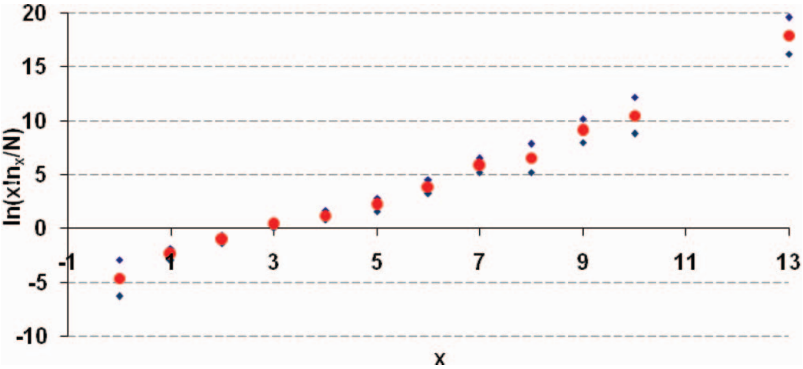

Figure 3. Hoaglin's Poissonnes plot for the data (color figure available online).

computed by:

$$\log\left(\frac{x!n_x}{N}\right) \tag{3}$$

where,

x = the "class" of the occurred crashes (i.e., the yearly crashes size)
n_x = x frequency
N = total sample size.

It can be seen in Figure 3 that the points are arranged along a straight line and so a Poisson distribution is reasonable for the data set. Table 6 reports the results of the coefficients estimates using the GLM and assuming a Poisson error distribution.

3.1. Observation of Dispersion in the Data

As referred in the introduction and according to the state of research, Poisson basic assumption of equidispersion is often too much restrictive for crash count data. To relax the Poisson assumption of equidispersion, the final model form (Equation 2) was recalibrated using a quasi-Poisson distribution, where the mean is the same of the Poisson mean, whereas the

Table 6
Coefficients estimates assuming a Poisson error distribution

Variables	Estimate	*SE*	*t*	*t* Probability
Constant (β_0)	−6.56	0.84	−7.81	<0.001
AADT$_1$ (β_1)	1.74	0.15	11.37	<0.001
RW$_2$ (β_2)	0.86	0.20	4.24	<0.001
PW$_1$ (β_3)	0.25	0.09	2.79	0.005

AADT$_1$ = Annual Average Daily Traffic on major-road; RW$_2$ = Minor-road roadway width; PW$_1$ = Major-road permitted ways.

Table 7
Model parameters assuming quasi-Poisson error distribution

Variables	Estimate	*SE*	*t*	*t* Probability
Constant (β_0)	−6.56	0.58	−11.40	< 0.001
AADT$_1$ (β_1)	1.74	0.11	16.58	<0.001
RW$_2$ (β_2)	0.86	0.14	6.19	< 0.001
PW$_1$ (β_3)	0.25	0.06	4.07	<0.001
α	−0.53	0.06	−8.68	<0.001

AADT$_1$ = Annual Average Daily Traffic on major-road; RW$_2$ = Minor-road roadway width; PW$_1$ = Major-road permitted ways.

variance is now a function of the mean, with:

$$\omega_{tj} = (1 + \alpha) \cdot \mu_{tj} \tag{4}$$

where,

ω_{tj} = variance function;
μ_{tj} = the mean value for the year *t* and the intersection *j*;
α = the *dispersion parameter*.

Estimates of α for the quasi-Poisson variance function are interpreted as follows (Cameron & Trivedi, 1998):

- if α is less than 0, there is evidence of underdispersion
- if α is between 0 and 1, there is evidence of modest overdispersion
- if α is greater than 1, there is evidence of considerable overdispersion.

Table 7 reports the values of model parameters estimated by GLM, where a quasi-Poisson error distribution was assumed. For this purpose Genstat software was used again (Payne et al., 2008). It should be noted that the dispersion parameter α in Table 7 and its *SE* were estimated iterating the following auxiliary regression (Cameron & Trivedi, 1998):

$$\frac{\left(y_{tj} - \hat{\mu}_{tj}\right)^2 - y_{tj}}{\hat{\mu}_{tj}} = \alpha + \varepsilon_{tj} \tag{5}$$

where, $\hat{\mu}_{tj}$ 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 values (i.e., dispersion parameter, coefficients) converged.

Results showed in Table 7 highlight that the data are underdispersed ($\alpha < 0$). Moreover, the quasi-Poisson model fits the data better than the Poisson one; in fact, the consideration of underdispersion in the data improves the estimates accuracy, as it is shown by the reductions in the standard error values.

According to Lord et al. (2010), to further improve the parameters estimates, the COM-Poisson distribution was used. As referred in Section 1, this distribution has recently been reintroduced by statisticians for modeling count data that are characterized by either over- or underdispersion. The COM-Poisson distribution belongs to the exponential family and 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). As referred by Shmueli et al. (2005) for a random variable Y_{tj} (e.g., a discrete count at year t and at intersection j), COM-Poisson probability distribution function is given by the equation:

$$P\left(Y_{tj} = y_{tj}\right) = \frac{\lambda_{tj}^{y_{tj}}}{(y_{tj}!)^{\nu} Z(\lambda_{tj}, \nu)} \tag{6}$$

where,

$$Z(\lambda_{tj}, \nu) = \sum_{s=0}^{\infty} \frac{\lambda_{tj}^{s}}{(s!)^{\nu}} \tag{7}$$

λ_{tj} = a centering parameter, denoting the expected value under a Poisson distribution associated with observation tj (Sellers & Shmueli, 2010);
ν ($\geq$0) = the dispersion parameter (where $\nu < 1$ for overdispersion and $\nu > 1$ for underdispersion).

Moments of COM-Poisson distribution can be expressed using the recursive formula (Shmueli et al., 2005):

$$E(Y_{tj}^{r+1}) = \begin{cases} \lambda_{tj}\ E^{1-V}\ (Y_{tj} + 1) & r = 0 \\ \lambda_{tj} \dfrac{\partial}{\partial \lambda_{tj}} E(Y_{tj}^{r}) + E(Y_{tj}) \cdot E(Y_{tj}^{r}) & r > 0 \end{cases} \tag{8}$$

Using an asymptotic approximation for $Z(\lambda_{tj},\nu)$, as referred by Shmueli et al. (2005), E(Y_{tj}) can be closely approximated by:

$$E(Y_{tj}) = \lambda_{tj} \quad \frac{\partial \log\ Z(\lambda_{tj}, \nu)}{\partial \lambda_{tj}} \approx 0\lambda_{tj}^{1/\upsilon} - \frac{(\upsilon - 1)}{2\upsilon} \tag{9}$$

It has been noted that the approximation is especially good for $\nu \leq 1$ or $\lambda_{tj} > 10^{\nu}$.

After estimating the COM-Poisson regression model, fitted values can be computed by Equation 9, setting:

$$\hat{\lambda}_{tj} = \exp\left(\mathrm{x}_{\mathrm{tj}}' \hat{\beta}\right) \tag{10}$$

3.2. Results and Goodness-of-Fit

Because the data, as well as model output, exhibited underdispersion, model calibration was done with particular attention on the dispersion phenomenon. For this purpose, the model in Equation 2 was recalibrated using the COM-Poisson distribution. To estimate the COM-Poisson regression coefficients and the related standard errors, R software (R Development Core Team, 2010) was used through the codes arranged by Sellers and Shmueli (2010), available at http://www9.georgetown.edu/faculty/kfs7/research.

The statistics of the parameters estimates and the standard errors under COM-Poisson hypothesis were compared with those obtained using the Poisson and quasi-Poisson distributions (Table 8). It must be noted that COM-Poisson coefficients are for centering parameter λ (Equation 10) and not for mean as in the case of Poisson/quasi-Poisson model (Lord et al., 2010; Sellers & Shmueli, 2010). The shape parameter of the COM-Poisson

Table 8
Coefficients estimates and goodness-of-fit for the three models

	Poisson			Quasi-Poisson			Com-Poisson		
Variables	Est	*SE*	*t*	Est	*SE*	*t*	Est[a]	*SE*	*t*
Constant (β_0)	−6.56	0.84	−7.81	−6.56	0.58	−11.40	−14.28	2.09	−6.83
AADT$_1$ (β_1)	1.74	0.15	11.37	1.74	0.11	16.58	3.91	0.52	7.51
RW$_2$ (β_2)	0.86	0.20	4.24	0.86	0.14	6.19	1.94	0.38	5.10
PW$_1$ (β_3)	0.25	0.09	2.79	0.25	0.06	4.07	0.54	0.15	3.60
ν	—			—			2.41	0.30	
α	—			−0.53	0.06	−8.68	—		
MPB	−0.42			−0.42			0.01		
MAD	1.35			1.35			1.03		
MSPE	4.47			4.47			1.80		
AIC	490			490			455		

Est = Estimate; AADT$_1$ = Annual Average Daily Traffic on major-road; RW$_2$ = Minor-road roadway width; PW$_1$ = Major-road permitted ways; MPB = mean prediction bias; MAD = mean absolute deviance; MSPE = mean squared predictive error; AIC = Akaike Information Criterion.
[a]Model parameters to be used for determining $\hat{\lambda}_i$ (Equation 10).

distribution clearly indicates underdispersion ($\nu > 1$). This further confirms that the Poisson distribution is not appropriate to interpret the data set. To value the quasi-Poisson and COM-Poisson models by the point of view of the goodness-of-fit, the indicators listed below were calculated and also reported in Table 8:

- Mean prediction bias (MPB): it provides a measure of the magnitude and direction of the average model bias. If the MPB is positive then the model overpredicts crashes and if the MPB is negative then the model underpredicts accidents. It is computed using the following equation:

$$\text{MPB} = \frac{1}{n}\sum_{t=1}^{k}\sum_{j=1}^{m}(\hat{y}_{tj} - y_{tj}) \tag{11}$$

where, n is the sample size, $\hat{y}_{tj}$ and y_{tj} are the predicted and observed crashes at year t and at intersection j respectively (Oh, Lyon, Washington, Persaud, & Bared, 2003);

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

$$\text{MAD} = \frac{1}{n}\sum_{t=1}^{k}\sum_{j=1}^{m}|\hat{y}_{tj} - y_{tj}| \tag{12}$$

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

$$\text{MSPE} = \frac{1}{n}\sum_{t=1}^{k}\sum_{j=1}^{m}\left(\hat{y}_{tj} - y_{tj}\right)^2 \tag{13}$$

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

$$\text{AIC} = -2\ \log\ L + 2p \tag{14}$$

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. In order to compute AIC the following expressions of log likelihood were used (in the following the subscript *i* denotes the generic observation at year *t* and at intersection *j*):

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

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

The goodness-of-fit indicators in Table 8 show that the COM-Poisson regression fits the data better than the quasi-Poisson model:

- the MPB value of the COM-Poisson indicates that the model fairly estimates crashes
- the MAD and the MSPE values of the COM-Poisson model, closer to 0 than the quasi-Poisson model, indicate that the model has good prediction capacity.

4. Conclusions

The article describes the methodological path followed to generate a SPF for urban, four-leg, signalized intersections. Data set included crashes occurred in the years 2000 to 2007 at 19 intersections in the Palermo City road network, directly collected from Municipal Police Force reports. A set of potential covariates was examined, but only three of them appeared to be significant at the 15% confidence level in the preliminary analysis:

- Annual Average Daily Traffic on major road (AADT_1)
- Minor road roadway width (RW_2)
- Major road number of permitted ways (PW_1).

With regards to the functional model form the power function seemed appropriate for the covariates AADT_1 and RW_2, whereas the exponential function was the best form for the variable PW_1. So the functional form used for the model was the following (see Equation 2):

$$\mathrm{y}_{\mathrm{tj}} = \beta_0 \mathrm{AADT}_{1\mathrm{tj}}^{\beta_1} \mathrm{RW}_{2\mathrm{j}}^{\beta_2} \mathrm{e}^{\beta_3 \mathrm{PW}_{1\mathrm{j}}}$$

To test the Poisson assumption of equidispersion, a quasi-Poisson regression was implemented, that is, assuming a linear relationship between the variance and the mean. As the model output showed clear signs of underdispersion, a COM-Poisson model was performed and evaluated in GLM framework thanks to its statistical properties and in particular its flexibility in handling over- and underdispersed data.

The results of the current study confirm that COM-Poisson GLM regression model can properly handle underdispersed crash data; they also show that quasi-Poisson and COM-Poisson GLM regression model allow to obtain more precise estimates for the model parameters than the traditional Poisson model. Moreover, COM-Poisson regression further improves the predictive performance of the proposed model and, at least for the case study data set, it provides a better goodness-of-fit than the quasi-Poisson model.

As said above, it has to be noted that model evaluation ignored temporal correlation within observations; that is why the modeling framework for COM-Poisson model is not yet available in GEE. Development of research in this direction by the statistical community are desirable.

At last, the reader must be advised that the small sample size used in the study could have affected the estimation of model parameters (coefficients and dispersion parameter): for example, the desirable large-sample properties of some parameter-estimation techniques (e.g., maximum likelihood estimation) are not realized (Lord & Mannering, 2010). Therefore, though results can help to highlight the potential of COM-Poisson model in handling underdispersed data, further researches should be carried out using different data set (namely larger sample size) to confirm them.

References

Automobile Club of Italy. (2005). *Car trend 2000-2005, Statistic area, studies & searches direction*. Management Control Office. Automobile Club of Italy & National Institute of Statistics, Rome.

Automobile Club of Italy. (2008). *Car trend 2006-2008, statistic analysis on car-market trends in Italy, Statistic area, studies & searches direction*. Management Control Office. Automobile Club of Italy & National Institute of Statistics, Rome.

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

Association Suisse des Professionnels de la Route et du Trafic. (1995). *Courbes de variation caractéristiques et traffic journalier moyen* (TJM, SN 640 005), Zurich, Switzerland.

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, Federal Highway Administration.

Bonneson, J., Zimmerman, K., & Fitzpatrick, K. (2005). *Roadway safety design synthesis (*FHWA*/TX-05/0-4703-P1)*. Washington, DC: U.S. Department of Transportation, Federal Highway Administration.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data* (Econometric Society Monographs No. 30). Cambridge, UK: Cambridge University Press.

Canale S., Leonardi S., & Pappalardo G. (2005, September 21-285). *The reliability of the urban road network: Accident forecast models*. Proceedings of the 3rd International SIIV Congress, Bari, Italy.

Clark, S. J., & Perry, J. N. (1989). Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics*, *45*, 309–316.

Conway, R. W, & Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, *12*, 132–136.

Dean, C. B. (1994). Modified pseudo-likelihood estimator of the overdispersion parameter in Poisson mixture models. *Journal of Applied Statistics*, *21*(6), 523–532.

Geedipally, S., Guikema, S. D., Dhavala, S., & Lord, D. (2008). *Characterizing the performance of a Bayesian Conway-Maxwell Poisson GLM*. Paper presented at Joint Statistical Meetings, 3–7 August, 2008. Denver, CO.

Giuffrè, O., Granà, A., Giuffrè, T., & Marino, R. (2007). Improving reliability of road safety estimates based on high correlated accident counts. *Journal of the Transportation Research Board*, *2019/2007*, 197–204.

Guikema, S. D., & Coffelt, J. P. (2008). A flexible count data regression model for risk analysis. *Risk Analysis*, *28*(1), 213–223.

Harwood, D. W., Council, F. M., Hauer, E., Hughes, W. E., & Vogt, A. (2000). *Prediction of the expected safety performance of rural two-lane highways (*FHWA-*RD-99-207)*. Washington, DC: U.S. Department of Transportation, Federal Highway Administration.

Hauer, E. (2004). Statistical road safety modelling. *Journal of the Transportation Research Board*, *1897*, 81–87.

Hoaglin, D. C. (1980). A Poissonness plot. *The American Statistician*, *34*(3), 146–149.

Kadane, J. B., Shmueli, G., Minka, T. P., Borle, S., & Boatwright, P. (2006). Conjugate analysis of the Conway-Maxwell-Poisson distribution. *Bayesian Analysis*, *1*, 363–374.

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 and Prevention*, *38*(4), 751–766.

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.

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.

Lord, D., & Mahlawat, M. (2009). Examining the application of aggregated and disaggregated Poisson-gamma models subjected to low sample mean bias. *Transportation Research Record*, *2136*, 1–10.

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.

Losa, M., & Terrosi Axerio, A. (2005). *Adapting safety performance functions for signalized four-legged intersections*. Proceedings of the 3rd International SIIV Congress, Bari, Italy. September 22–24.

Lyon, C., Haq, A., Persaud, B., & Kodama, S. (2005). *Development of safety performance functions for signalized intersections in a large urban area and application to evaluation of left-turn priority treatment*. Proceedings of the 84th TRB Annual Meeting, Washington, D.C. January 9–13.

Maycock, G., & Hall, R. D. (1984). *Accidents at 4-arm roundabouts* (TRRL Laboratory Report 1120). Berkshire, UK: Crowthorne.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models - second edition*. London, UK: Chapman and Hall, Ltd.

McGee, H., Taori, S., & Persaud, B. (2003). *Crash experience warrant for traffic signals* (NCHRP Report 491). Washington, DC: Transportation Research Board, National Cooperative Highway Research Program.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, *135*, 370–384.

Oh, J., Lyon, C., Washington, S. P., Persaud, B. N., and Bared, J. (2003). Validation of the FHWA crash models for rural intersections: Lessons learned. *Transportation Research Record*, *1840*, 41–49.

Oh, J., Washington, S. P., and Nam, D. (2006). Accident prediction model for railway-highway interfaces. *Accident Analysis and Prevention*, *38*(2), 346–356.

Payne, R. W, Lane, P. W., & Digby, P. G. N. (2008). *GenStat* (11th ed.). Reference Manual. Oxford, UK: Clarendon Press.

Persaud, B., & Dzbik, L. (1993). Accident prediction models for freeways. *Journal of the Transportation Research Board*, *1401*, 55–60.

Piegorsch, W.W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, *46*, 863–867.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Retrieved from http://www.Rproject.org.

Sellers, K. F., Borle, S. & Shmueli, G. (in press). The COM-Poisson model for count data: A survey of methods and applications. *Applied Stochastic Models in Business and Industry*. (http://galitshmueli.com/content/com-poisson-model-count-data-survey-methods-and-applications)

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

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*, 127–142.

Vogt, A. (1999). *Crash models for rural intersections: Four-lane by two-lane stop-controlled and two-lane by two-lane signalized* (FHWA-RD-99-128). Washington, DC: U.S. Department of Transportation, Federal Highway Administration.

Vogt, A., & Bared, J. (1998). *Accident prediction models for two-lane rural roads: segments and intersections* (FHWA-RD-98-133). Washington, DC: U.S. Department of Transportation, Federal Highway Administration.

Winkelman, R. (1995). Duration dependence and dispersion in count data model. *Journal of Business & Economical Statistics*, *13*(4), 467–474.