Road casualties and enforcement:

Distributional assumptions of serially correlated count data

George Yannis*, Constantinos Antoniou and Eleonora Papadimitriou

National Technical University of Athens,

Department of Transportation Planning and Engineering,

5 Iroon Polytechniou Street, 157 73 Zografou, Athens, Greece

Tel: +30.210.7721326, Fax +30.210.7721454

geyannis@central.ntua.gr, antoniou@central.ntua.gr, nopapadi@central.ntua.gr

* Corresponding author

ABSTRACT

Objective: Road safety data are often in the form of counts and usually temporally correlated. The objective of this research is to investigate the distributional assumptions of road safety data in the presence of temporal correlation.

Methods: Using the generalized linear model framework, four distributional assumptions are considered: normal, Poisson, quasi-Poisson and negative binomial, and appropriate models are estimated. Monthly casualty and police enforcement data from Greece for a period of six years (January 1998 – December 2003) have been used. The developed models include sinusoidal latent terms to capture the temporal serial correlation of observations. Several statistical goodness-of-fit diagnostic tests have been performed for the results of the estimated models, and the predictive capabilities of the models are investigated.

Results: The residuals of the quasi-Poisson and negative binomial models do not show any serial correlation. The signs of the estimated coefficients for all models are consistent and intuitive. In particular, a negative coefficient value for the number of breath alcohol controls indicates that the number of persons killed and seriously injured decreases as the intensity of breath alcohol controls increases. The Poisson model fails to capture the overdispersion in the data, thus underestimating the standard errors of the estimated coefficients.

Conclusions: The results suggest that the quasi-Poisson and negative binomial outperform the normal and Poisson models in this application. The findings of this research demonstrate a clear link between the intensification of police enforcement and the reduction of traffic accident casualties. In particular, an increase in the number of breath alcohol controls in Greece after 1998 contributed to a reduction in the number of persons killed and seriously injured from traffic accidents.

Keywords: road safety, enforcement, serial correlation, generalized linear models, Poisson, negative binomial

INTRODUCTION

Many statistical techniques assume independence of observations. Road safety data, however, are often in the form of time-series of counts observed during successive time periods, e.g. days, months or years. In practice, such observations often tend to be correlated with the respective observations from previous years, months or days, i.e. are usually temporally correlated. The linear regression model -an attractive and simple method- has stringent assumptions that are therefore usually violated when applied to road safety data. In this research, alternative modeling assumptions are evaluated within the more flexible generalized linear modeling framework. The presented approach is demonstrated through models capturing the impact of the intensification of police enforcement on the reduction of traffic accident casualties.

While the linear regression model is simple (to run and interpret), elegant and efficient, it is subject to the fairly stringent Gauss-Markov assumptions (Washington et al., 2003). If these assumptions hold, it can be shown that the solution obtained by minimizing the sum of squared residuals ('least squares') is BLUE, i.e. best linear unbiased estimator. In other words, it is unbiased and has the lowest total variance among all unbiased linear estimators.

The basic Gauss-Markov assumptions require:

- Linearity (in the parameters; nonlinearity in the variables is acceptable);
- Homoscedasticity;
- Exogenous independent variables;
- Uncorrelated disturbances; and
- Normally distributed disturbances

These assumptions, however, are often violated in practice. In this research, two of these violations are explicitly considered, in particular correlated disturbances; and non-normal error structures. The choice of these two violations is not arbitrary; instead it is motivated by the fact that these two violations are more relevant to the nature (time-series count data) of the road

safety data. Generalized linear models (GLM), a generalization of the linear regression, can be used to overcome the restriction on the normality of the error structure (McCullagh and Nelder, 1989, Dobson, 1990, Gill, 2000). Specific treatment of the application of GLM in the presence of serially correlated count data is also presented.

The objective of GLM is to allow for more flexible error structures, besides the Gaussian which is assumed by –linear and nonlinear– regression. A further discussion of the distributional assumptions allowed by the GLM, as well as an overview of the approach are deferred until the next section.

The Poisson distribution has been considered suitable to counts of car crashes for a long time (Nicholson and Wong, 1993). However, the Poisson model -while arguably more appropriate than the Gaussian- is not without weaknesses and technical difficulties. For example, the assumption of a pure Poisson error structure may prove inadequate in the presence of "overdispersed" data (Maycock and Hall, 1984). Overdispersion reflects more variation in the response than what is expected by the Poisson assumption, which assumes that the variance equals the mean. An implication of overdispersion is that the estimates of the standard errors of the parameters will not be correct, and in fact the standard errors will be underestimated.

A straightforward approach to overcome this issue is to use a quasi-Poisson model; i.e. estimate a dispersion parameter for the Poisson model, thus allowing it to take values other than one. Maycock and Hall (1984) showed that the negative binomial model could also be used as an extension to the Poisson. Miaou (1994) and Wood (2002) have also used the negative binomial model for road safety applications. Maher and Summersgill (1996) mention that, quite often, the two approaches (quasi-Poisson and negative binomial) may provide very similar estimation results. One may then be tempted to think that the two models are equivalent and that it does not really matter which model is selected. Maher and Summersgill further warn that this may not be the case, as the two models may have different prediction properties, as measured, e.g. by the prediction error variance. Lord et al. (2005) examine the applicability of

different models, including Poisson, negative binomial (or Poisson-gamma) and zero-inflated Poisson and negative binomial models, to the modeling of accident data.

Furthermore, few processes are adequately modeled by linear models in practice. For example, several researchers have shown that conventional linear regression models lack the distributional property to adequately describe collisions. This inadequacy is due to the random, discrete, non-negative, and typically sporadic nature that characterize the occurrence of vehicle collisions. Several researchers (including Hauer et al.1988, Hakim et al., 1991; Cameron et al., 1993; Newstead et al., 1995), using road accident statistics, have presumed that the explanatory variables have a multiplicative effect on accidents, i.e. $y = ax_1^b x_2^c$ (as opposed to e.g. additive, i.e. $y = a + bx_1 + cx_2$).

Examples of road safety applications involving the use of GLM in temporally correlated data include before/after analysis on the impact of red-light camera presence in crashes (Retting and Kyrychenko, 2001), investigation of relationships between accidents, flows and road or junction geometry, allowing for the presence of a trend over time in accident risk (Maher and Summersgill, 1996), traffic safety comparisons among several counties in France, where the time trend of each index (incidence and severity) is the same across counties and across road types (Amoros et al., 2003), and estimation of expected junction accidents (both in total and disaggregated by severity, road surface condition and lighting condition), which allow for the possibility of accident risk varying over time (Mountain et al., 1998). White and Washington (2001) developed a logistic regression model to gain insight into the relationship between enforcement and the use of safety restraint.

In this research, the suitability of several distributions for modeling road safety data that are temporally correlated is investigated using casualty data from Greece. More precisely, the correlation between accident casualties and police enforcement data is examined. An overview of generalized linear models is presented first, while specific issues that relate to the application of GLMs in the presence of serially correlated data are discussed next. The model data and

specification are described next, followed by model fit and diagnostics. In the last section the obtained results are discussed.

GENERALIZED LINEAR MODELS

Generalized linear models facilitate the analysis of the effects of explanatory variables in a way that closely resembles the analysis of covariates in a standard linear model, but with less confining assumptions. This is achieved by specifying a *link function*, which links the systematic component of the linear model with a wider class of outcome variables and residual forms.

A key point in the development of GLM was the generalization of the normal distribution (on which the linear regression model relies) to the exponential family of distributions. This idea was developed by Fisher (1934). Consider a single random variable y whose probability (mass) function (if it is discrete) or probability density function (if it is continuous) depends on a single parameter θ. The distribution belongs to the exponential family if it can be written in the form (Eq. (1)):

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} \quad (1)$$

where a, b, s, and t are known functions. The symmetry between y and θ becomes more evident if Eq. (1) is rewritten as Eq. (2):

$$f(y;\theta) = \exp\left[a(y)b(\theta) + c(\theta) + d(y)\right] \quad (2)$$

where s(y)=exp[d(y)] and t(θ)=exp[c(θ)]. If a(y)=y then the distribution is said to be in the canonical form. Furthermore, any additional parameters (besides the parameter of interest θ) are regarded as nuisance parameters forming parts of the functions a, b, c, and d, and they are treated as though they were known. Many well-known distributions belong to the exponential family, including –for example– the Poisson, normal, and binomial distributions. On the other hand, examples of well-known and widely used distributions that cannot be expressed in this form are the student's t-distribution and the uniform distribution.

The generalized linear model can be defined in terms of a set of independent random variables $y_1, \ldots, y_n$, each with a distribution from the exponential family with the following properties:

1. The distribution of each y_i is of the canonical form and depends on a single parameter θ_i (not necessarily the same parameter for all variables) (Eq. (3)):

$$f(y_i;\theta_i) = \exp\left[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)\right] \tag{3}$$

2. The distributions of all the y_i s are of the same form (e.g. all normal or all binomial) so that the subscripts on b, c, and d are not needed.

The joint probability density function of $y_1, \ldots, y_n$ is then (Eq. (4)):

$$f(y_1,\ldots,y_N;\theta_1,\ldots,\theta_N) = \exp\left[\sum_{i=1}^{N}\left(y_i b(\theta_i) + c(\theta_i) + d(y_i)\right)\right] \tag{4}$$

When specifying a model, the N parameters θ_i are usually not of direct interest. Instead, for a GLM, a smaller set of p parameters $\beta_1, \ldots, \beta_p$ is considered (where $p < N$), such that a linear combination of the βs is equal to some function of the expected value μ_i of y_i, i.e. (Eq. (5)):

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \tag{5}$$

where

g is a monotonic, differentiable function called the link function;

$\boldsymbol{x}_i$ is a *(p x 1)* vector of explanatory variables (covariates and dummy variables for levels of factors); and

$\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$ is the *(p x 1)* vector of parameters.

To recapitulate, in the univariate case, a generalized linear model has three components:

1. A response variable y assumed to follow a distribution from the exponential family (which is a generalization of and includes the normal distribution);
2. A set of parameters $\boldsymbol{\beta}$ and explanatory variables $\mathbf{X} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T]^T$
3. A monotonic link function g such that $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$, where $\mu_i = E(y_i)$

GENERALIZED LINEAR MODELS IN THE PRESENCE OF SERIALLY CORRELATED DATA

Generalized linear models require uncorrelated observations. Time-series data require special consideration, since the observations typically fail to meet this assumption, as neighboring observations are likely to be correlated. It is often possible to include a large number of explanatory variables in a linear regression model, resulting in seemingly serially uncorrelated residuals (and, therefore, the linear model theory would apply). This strategy, however, is problematic, as it may not be easy to identify the appropriate explanatory variables that would reflect the serial correlation.

In a very different (with respect to road safety) context, Zeger (1988) introduced a method for regression when the outcomes are a time series of counts (as is often the case in road safety applications). Zeger concludes that generalized linear models with linear and log links can be extended to parameter-driven models that capture serial correlation. The serial correlation in the observed data is captured in this model through some unobserved (or latent) process. Conditional on this unobserved process, the counts are assumed to be independent. This is a reasonable assumption for road safety data, since the occurrence of an accident (or a fatality or injury) is *usually* not directly caused by another.

The data, however, are serially correlated because they are ordered in time, and other factors (also ordered in time) are affecting the underlying risk. A discussion on these properties, albeit in a totally different context, can be found in Campbell (1994), who also presents a practical application of the approach, where the only assumption that is made on the distribution of the error structure is that it is mean stationary. A process is called mean stationary if the mean of the process is stationary, i.e. $E(y_t) = E(y_{t+m})$ for any *t* and *m* (Pindyck and Rubinfeld, 1997). Davis et al. (2000) developed a practical approach to diagnose the existence of a latent stochastic process in the mean of a Poisson regression model.

For the Poisson model, the covariance matrix, and hence the standard errors of the parameter estimates, are estimated under the assumption that the Poisson model is appropriate. Occasionally one may observe overdispersion, i.e. more variation in the response than what is expected by the Poisson assumption, which assumes that the variance equals the mean. An implication of overdispersion is that the estimates of the standard errors of the parameters will not be correct, and in fact the standard errors will be underestimated. Underdispersion (less variation than expected) is also possible, although not as common.

MODEL DATA AND SPECIFICATION

The use of generalized linear models for road safety research is demonstrated using accident casualties (persons killed or seriously injured) and police enforcement data from Greece. Data from Athens and Thessaloniki were excluded because traffic conditions in these large agglomerations are much more complex as are the parameters potentially describing the road safety phenomenon. The number of vehicles in circulation in the studied areas is also added into the model specification as an offset (i.e. its coefficient is not estimated but is constrained to one). Ideally, one would consider vehicle-kilometers instead of vehicles. However, this data is not available. Average vehicle-kilometer data collected from the SARTRE 2 and SARTRE 3 projects in 1996 and 2002 respectively (SARTRE 3, 2004), however, suggest that the number of kilometres travelled does not present a statistically significant change during this period (average kilometres travelled was equal to 15 231km in 1996 and 15 070 in 2002) and therefore, their use is acceptable for the purposes of the specific research.

Monthly data from January 1998 to December 2003 have been used for this research (Figure 1). The data of the first five years (60 observations) are used for the model estimation, while the data for the last year (12 observations) are used for the validation of the estimated model.

The model specification comprises three main effects: trend, seasonal effects, and explanatory variables. The trend captures the evolution of the dependent variable over time. This is captured in the specification by the addition of the "Month" variable, which ranges from 1 (for the first month, i.e. January 1998) to 72 (for December 2003). This variable was shown to be statistically

insignificant and has not been retained in the final model, detailed in the next section. Seasonal effects are captured by the incorporation of sinusoid components (similar to those used e.g. by Zeger, 1988, and Campbell, 1994). Several frequencies have been investigated (from 1 to 15 months), but the most useful proved to be the annual and its first (six month) harmonic.

Furthermore, besides specifying trend and seasonal components, the impact of explanatory variables is also tested, with an emphasis on enforcement data (number of breath alcohol controls per month) and (the log of) vehicles in circulation. Acknowledging that there may be a lot of other intervening parameters, it could be argued that police enforcement could in some cases be influenced by total traffic flow, which might be indirectly affected by vehicle fleet. To account for the delayed impact of enforcement in road safety (as the word-of-mouth spreads) the number of breath alcohol controls has been lagged by two intervals, capturing the impact of enforcement intensification two months after it occurs. The log of vehicles in circulation has been entered as an offset. Naturally, the two major Greek urban areas excluded from the casualty data have also been excluded from the data of breath alcohol controls and registered vehicles. The number of registered vehicles has been interpolated from annual figures. Finally, a high number of casualties was observed during the month of August. Therefore, a dummy variable has been introduced, that takes the value of one for August and zero otherwise. Further exploration of the available monthly data did not reveal any new insight in the seasonality of the road safety phenomenon. The "August phenomenon" remained predominant.

[INSERT FIGURE 1 ABOUT HERE]

Seasonality (August peak) observed mainly in the persons killed and seriously injured but also on the enforcement can be attributed to increased summer traffic in Greece as a holiday destination. The exceptional enforcement low value on December 2001 cannot be explained by any other reason than the internal enforcement programming of the Police.

MODEL FIT AND DIAGNOSTICS

In this section, different error structures -that are allowable within the GLM framework and are also theoretically supported- are applied. Model estimation and analysis has been performed using the R Software for Statistical Computing (RDCT, 2006). First, the Gaussian (Normal) distribution is used. If the identity link function was used, then as the model specification is linear additive, this would be equivalent to the linear regression model. A Poisson model is also fitted, along with a quasi-Poisson that relaxes the assumption that the dispersion parameter is equal to one. Finally, a negative binomial model is fitted. A log link function is used for all models.

Estimation results and model fit for the four model families are shown in Table 1. A sinusoid term with an annual frequency and its (6 month) harmonic capture periodicity. A negative coefficient value for the number of breath alcohol controls indicates that the number of persons killed and seriously injured decreases as the intensity of breath alcohol controls increases, which is an intuitive result. More visible police presence on the roads results in safer driver behavior (less speeding, less aggressive driving, etc.). However, police enforcement results require some time before becoming significant, as driver perception of enforcement grows by the continuous visibility of policemen; e.g. a lagged effect of two months was revealed in this paper.

[INSERT TABLE 1 ABOUT HERE]

A dummy variable, taking the value of one for August and zero otherwise, was also found to be significant. General traffic increase over years refers to the overall network with emphasis on urban and suburban areas (often congested), resulting in lower speeds and less accidents and related casualties. However, summer traffic increase refers to rural - often not congested - areas, where speed is not necessarily reduced. Furthermore, a large proportion of this traffic refers to Greek and foreigner tourists not acquainted with the local network and traffic patterns and with more accident prone trip characteristics (late night entertainment, alcohol consumption, etc.). The macroscopic relation between accident casualties and traffic volumes is a very complex phenomenon, which certainly requires further investigation.

Other explanatory variables (such as the number of speeding violations) were also originally entered into the model. However, explanatory variables relating to enforcement were highly correlated (in particular the number of breath alcohol controls and speeding violations had a correlation of 0.97). Therefore, while using either variable resulted in intuitive results, their combination resulted in multicollinearity problems.

The coefficient signs, however, are consistent for all models and all retained parameters are significant at the 1% level (with the exception of the enforcement data in the quasi-Poisson and negative binomial models, which are still significant at the 10% level). Due to the use of the log link function in all models, the magnitude of the estimated coefficients is rather close for all four models. A comparison of the computed standard errors shows that the values obtained for the Poisson model are significantly lower than those obtained from the normal, the quasi-Poisson and the negative binomial. Therefore, the z-values obtained for the Poisson model seem unusually high. A closer look at the model statistics suggests that the data may be overdispersed.

Potential overdispersion can be identified by dividing the residual deviance (defined -up to a constant- as twice the log-likelihood ratio statistic) by the residual degrees of freedom (i.e. the number of observations minus the number of parameters in the model). The resulting measure is an approximately unbiased estimator of the dispersion parameter (Venables and Ripley,

2002). If the deviance is equal to the degrees of freedom then there is no evidence of overdispersion. Note that a dispersion parameter not equal to one does not necessarily imply overdispersion, but could also indicate other problems, such as an incorrectly specified model or outliers in the data. An incorrectly specified model can be due to an incorrectly specified functional form, e.g. an additive error term ($y = f(x) + \varepsilon$) rather than a multiplicative error term ($y = f(x) \cdot \varepsilon$) may be appropriate, or, more likely, that important explanatory variables (or interactions) are missing from the model. However, overdispersion can also be a property of the data, typically indicating a lack of independence or heterogeneity among observations, sampling issues, etc.

The dispersion factor for the data at hand is equal to 151.11/51=2.96, which is significantly different from one. The assumption of a Poisson model (with a dispersion parameter equal to one) is therefore unlikely to be realistic. A quasi-Poisson model (an extension of the Poisson model, in which the dispersion parameter is allowed to vary from one) has also been estimated. The estimation is based on the iterative algorithm proposed by Breslow (1984) for fitting overdispersed log-linear Poisson models. The magnitude of the estimated coefficient values is similar to that obtained by the Poisson model, and the signs are the same. The significance of the coefficients, however, has significantly decreased, indicating that in the Poisson model the standard errors were underestimated due to the overdispersion. As expected, the dispersion parameter for the quasi-Poisson model is 51.38/51=1.01, i.e. very close to one.

Finally, a negative binomial model has been fitted. The estimated coefficients were similar to those obtained from the quasi-Poisson. This confirms the findings of Maher and Summersgill (1996) who state that the two approaches may provide similar estimation results. Slightly lower standard errors for the binomial, however, lead to more significant statistics.

Further model diagnostics are presented in Figures 2 through 5. Normal scores plot (QQ plot) of standardized deviance residuals is presented in the top subfigure of each figure. The x-axis represents the standardized deviance residuals, while the y-axis represents the quantiles of the standard normal. The dotted line in the QQ plot (top) is the expected line if the standardized

residuals are normally distributed, i.e. it is the line with intercept 0 and slope 1. If the deviance residuals were normally distributed, all points on the plot would fall on this dotted line. The deviance residuals of the normal model are far from normally distributed. The Poisson model is a slight improvement, but still far off. The quasi-Poisson and the negative binomial model deviance residuals, on the other hand, are practically normally distributed.

On the bottom subfigure is a plot of the Cook statistics against the standardized leverages. The standardized leverage of the i-th observation x_i can be computed as (Belsley et al., 1980, Eq. (6)):

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x}_i)}{(n-1)s_x^2} \qquad (6)$$

where *n* is the number of observations, the overbar indicates the predicted value, and s_x is the standard error. There are two dotted lines on each plot. The horizontal line is at 8/(n-2p) where *n* is the number of observations and *p* is the number of parameters estimated. Points above this line may be points with high influence on the model. The vertical line is at 2p/(n-2p) and points to the right of this line have high leverage compared to the variance of the raw residual at that point. If all points are below the horizontal line or to the left of the vertical line then the line is not shown. For example, in the quasi-Poisson (Figure 4) and negative binomial (Figure 5) plots, the horizontal line is not present, since no point lies above it.

The number of (high) leverage points is the union of points to the right and top of the two dashed lines. Therefore, most observations in the Gaussian model (Figure 2) appear to be leverage points (as they are above the horizontal line). For the Poisson case (Figure 3), the points that are either above the horizontal dashed line or to the right of the vertical dashed line are seven. Only three points are to the right of the vertical dashed line for the quasi-Poisson (Figure 4) and negative binomial (Figure 5) models (the horizontal dashed line is not drawn in these two figures as no point lies above it), which provides additional evidence that (a) these

models are more appropriate for this application and (b) that they provide comparable fit in this application.

The estimation results and the model diagnostics suggest that the quasi-Poisson and the negative binomial assumptions are more valid for the considered problem (while this may not be always the case). The output of the resulting models is very similar and therefore a clear decision regarding the most appropriate model cannot be made. One observation relates to the estimated standard errors, which are higher for the quasi-Poisson. Choosing to err in the side of caution, one could retain this model.

It should be noted that the usual tests for comparing nested models estimated using maximum likelihood estimation, such as the Akaike Information Criterion, AIC, (Akaike, 1973) or the Schwarz/Bayesian Information Criterion, BIC, (Schwarz, 1978), are not suitable for comparison across these (non-nested) models. For example, AIC or BIC could be used to compare models with different numbers of parameters and the same likelihood function (except for the number of parameters), e.g. two normal or two Poisson models, but not one normal and one Poisson.

[INSERT FIGURES 2 THROUGH 5 ABOUT HERE]

An important consideration when dealing with serially correlated data is the autocorrelation of the residuals. Residual plots for the four estimated models reveal that while there is still some autocorrelation present in the Gaussian model, the residuals of the quasi-poisson and negative binomial models do not show any serial correlation. The lack of serial correlation for the quasi-Poisson and the negative binomial models has also been confirmed from plots of residuals vs. time, as well as appropriate autocorrelation (ACF) and partial autocorrelation function (PACF) plots.

Figure 6 shows the values predicted by the quasi-Poisson model. The dashed line shows the actual observed number of persons killed and seriously injured in Greece (excluding the two major metropolitan areas of Athens and Thessaloniki). The thick solid line represents the model predictions and 95% confidence intervals are also shown with thinner solid lines. The fitted and

predicted values confirm the hypothesis that the summer peak is captured through the "August" variable.

[INSERT FIGURE 6 ABOUT HERE]

Elasticities are another useful tool in interpreting the impact of the model parameters on the response variable. Unlike estimated coefficients, elasticities are dimensionless. Among the estimated model parameters, meaningful and interpretable elasticities can be computed for the alcohol controls, and not for the intercept, the vehicles in circulation (which have been entered as an offset), the dummy variable for August, and the sinusoidal/cosinusoidal curves (a change in which would different to conceptualize). For the quasi-Poisson model (with a log link) the elasticity for parameters that enter the formulation linearly (such as the alcohol controls) is obtained as Eq. (7):

$$E_{x_{ik}}^{\lambda_i} = \frac{\partial \lambda_i}{\lambda_i} \cdot \frac{x_{ik}}{\partial x_{ik}} = \beta_k \cdot x_{ik} \qquad (7)$$

where E represents the elasticity, x_{ik} is the value of the *k*th independent variable for observation *i*, β_k is the estimated parameter for the *k*th independent variable and λ_i is the expected frequency for observation *i* (Washington et al., 2003). A common way, however, to report a single elasticity figure is to average these elasticities. Such an approach yields an elasticity for the (lagged) alcohol controls equal to -0.15, i.e. a 1% increase in the number of alcohol controls would result in a 0.15% decrease in the number of killed and seriously injured two months into the future.

CONCLUSION

The impact of different distributional assumptions for the dependent variables on the model estimation results is demonstrated in this research within the unified framework of generalized linear models. Due to the time-series nature of the data, a modeling approach to capture serial correlation through the introduction of sinusoid latent processes has also been demonstrated.

The estimated coefficients for the Poisson model are close to those estimated by the other three models, but the standard errors are severely underestimated (due to overdispersion), leading to artificially high z-statistic values. Even though these values were indeed significant in this application, this issue could have led to incorrect retention of insignificant variables in the Poisson model. Furthermore, even though the magnitude of the estimated coefficients for the quasi-Poisson and negative binomial is very similar, the different models may have different predictive properties and therefore may not –in general- be used interchangeably.

The estimated model includes an intercept, a zero-one dummy for the month of August, two sinusoid terms, an enforcement-related explanatory variable about the number of breath alcohol controls and the log of the vehicles in circulation (entered as an offset). A positive intercept captures the baseline number of casualties (persons killed and seriously injured). A positive coefficient for the dummy variable for August confirms that there is a higher number of fatalities and serious injuries in the month of August. While it is not easy to directly interpret the coefficients for the sinusoid terms, it becomes evident that the two corresponding latent processes are appropriate for this model.

The significant negative coefficient associated with the number of breath alcohol controls suggests that there is a negative correlation between the number of alcohol controls and the number of persons killed and seriously injured in road accidents. An increase in the number of breath alcohol controls can therefore lead to a reduction of the persons killed and seriously injured due to traffic accidents. This is a useful empirical finding that supports the argument that intensification of police enforcement can lead to an improvement in road safety.

The calculation of elasticities allowed for the quantification of the comparative impact of the parameters examined to the improvement of road safety at the national level. For example, policy makers can consider in their decisions the fact that in order to obtain a 10% decrease in the number of killed and seriously injured an enforcement increase of 66% is necessary. In this way, decision makers can better judge the importance of their actions and better adjust them into the overall transport environment.

Any type of count data where change over time might be observed can be modeled using these distributions. For example, traffic volume and headways can be modeled by the use of generalized linear models (especially negative binomial and quasi-Poisson distributions). Tolle (1976), Cowan (1975) and Leuzbach (1988) are some of the researchers that used the negative exponential distribution for modeling headway distribution in the uncongested regime, whereas variations of the negative binomial distribution (Akcelik and Chung, 1994, Griffiths et al., 1991) can be used under congested conditions. In conclusion, the researcher could be assisted by the methodological findings of this research, in order to select the model that better suits the particularities of each case.

REFERENCES

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory (B. Petrox and F. Caski, eds.), 267–281. Akademia Kiado, Budapest. (Reprinted in Breakthroughs in Statistics, eds Kotz, S. & Johnson, N. L. (1992), volume I, pp. 599–624. Springer, New York.
2. Akcelik, R. and E. Chung. Calibration of the Bunched Exponential Distribution of Arrival Headways. Road & Transport Research, Vol.3 No.1 pp. 42-59, March 1994.
3. Amoros, E, Martin, J., and B. Laumon (2003). Comparison of road crashes incidence and severity between some French counties. Accident Analysis and Prevention, Vol 35, pp. 537-547.
4. Belsley, D., Kuh, E., and Welsch, R. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley and Sons, New York.
5. Breslow, N.E. (1984), Extra-Poisson variation in log-linear models, Applied Statistics, 33, 38–44.
6. Cameron, MH, Haworth, N, Oxley, J, Newstead, S, & Le, T (1993). Evaluation of Transport Accident Commission road safety television advertising. Report No.52, Monash University Accident Research Centre.

7. Campbell, M. J. (1994). Time Series Regression for Counts: An Investigation into the Relationship between Sudden Infant Death Syndrome and Environmental Temperature. Journal of the Royal Statistical Society, Series A (Statistics in Society), Vol. 157, No. 2, pp. 191-208.

8. Cowan, R. J. Useful Headway Models. Transportation Research, Vol. 9, pp. 371-375, Pergamon Press 1975.

9. Davis, R., Dunsmuir, W. and Wang, Y. (2000). On autocorrelation in a Poisson regression model. Biometrika, Vol. 87, No. 3, pp. 491-505.

10. Dobson, A.J. (1990), *An Introduction to Generalized Linear Models*. Second edition, Chapman and Hall, London.

11. Fisher, R. A. (1934), *Two new properties of mathematical likelihood.* Proceedings of the Royal Society A, 144, pp 285-307.

12. Gill, J. (2000), *Generalized Linear Models: A Unified Approach*, Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-134, Thousand Oaks, CA: Sage.

13. Griffiths, J. D. and J. G. Hunt. Vehicle Headway in Urban Areas. Traffic Engineering and Control, pp. 458-462, Oct. 1991.

14. Hakim, S., Shefer, D., Hakkert, A. S. and Hocherman, I. (1991). A critical review of macro models for road accidents. Accident Analysis and Prevention, 23 (5), 379-400.

15. Hauer, E., Ng, J. C. N., and Lovell, J. (1988). "Estimation of Safety at Signalized Intersections", Transportation Research Record, 1185, Transportation Research Board, National Research Council, Washington, D.C., pp. 48-61.

16. Koornstra, M. J. (1992) The evolution of road safety and mobility, IATSS Research, 16: 129-148.

17. Koornstra, M. J. (1997) Trends and forecasts in motor vehicle Kilometrage, road safety, and environmental quality, pp: 21-32 in Roller, D., (ed.) The motor vehicle and the environment – Entering a new century. Proceedings of the 30th International Symposium on Automotive Technology & Automation, Automotive Automation Limited, Croydon.

18. Leutzbach, W. Introduction to the Theory of Traffic Flow. Springer-Verlag, Berlin Heidelberg, pp.68-72, 1988.

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

20. Maher M.J. and I. Summersgill (1996). A comprehensive methodology for the fitting of predictive accident models. Accident Analysis and Prevention 28(3), pp. 281-296.

21. Maycock, G., and Hall, R. D. (1984). "Accidents at 4-Arm Roundabouts." TRRL Laboratory Report 1120, Transport and Road Research Laboratory, Crowthorne, Berkshire, UK.

22. McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*. Second edition. Chapman Hall, New York.

23. Miaou, S. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Proceedings of the 73rd Annual Meeting of the Transportation Research Board, Washington, D.C.

24. Mountain, L., Maher, M., and B. Fawaz (1998). The influence of trend on estimates of accidents at junctions. Accident Analysis and Prevention, Vol. 30, No. 5, pp. 641-649.

25. Newstead, S., Cameron, M. H., Gantzer, S. and Vulcan, P. (1995). Modeling of some major factors influencing road trauma trends in Victoria 1989 - 93. Report No. 74, Monash University Accident Research Centre.

26. Nicholson, A., and Y-D. Wong (1993). Are accidents Poisson distributed? A statistical test. Accident Analysis & Prevention, Volume 25, Issue 1, February 1993, Pages 91-97

27. Pindyck, R. S. and D. L. Rubinfeld (1997). Econometric models and economic forecasts. Fourth edition. Irwin/McGraw-Hill.

28. R Development Core Team (RDCT, 2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org (accessed Jan 26, 2006).

29. Retting, R.A. and Kyrychenko, S.Y. (2001). Crash Reductions Associated with Red Light Camera Enforcement in Oxnard, California. Insurance Institute for Highway Safety, Arlington, VA.

30. SARTRE 3 (2004). European drivers and road risk. Part 1: Report on principal results. Part 2: Report on in-depth analyses. INRETS, Arcueil.

31. Schwarz, G. (1978). Estimating the Dimension of a Model. Annals of Statistics. 6 461–464.

32. Tolle, J. E. Vehicular Headway Distributions: Testing and Results. Transportation Research Record No. 567, pp. 56-64, Washington, D.C., 1976.

33. Venables, W. N., and Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth edition, Springer-Verlag, New York.

34. Washington, S. P., Karlaftis, M. G., and F. L. Mannering (2003). Statistical and Econometric Models for Transportation Data Analysis. Chapman & Hall/CRC.

35. White, D. J. and S. P. Washington (2001). Safety restraint use as a function of law enforcement and other factors, Transportation Research Record, Journal of the Transportation research Board, 1779: pp. 109-115.

36. Wood, G.R. (2002). Generalized Linear Accident Models and Goodness of Fit Testing. Accident Analysis & Prevention, Vol. 34, pp. 417-427.

37. Zeger, S. (1988). A Regression Model for Time Series of Counts. Biometrika, Vol. 75, No. 4, pp. 621-629.

List of Figures

Figure 1. Dataset overview

Figure 2. Model fit diagnostic plots (Gaussian)

Figure 3. Model fit diagnostic plots (Poisson)

Figure 4. Model fit diagnostic plots (Quasi-Poisson)

Figure 5. Model fit diagnostic plots (Negative binomial)

Figure 6. Quasi-Poisson model predictions

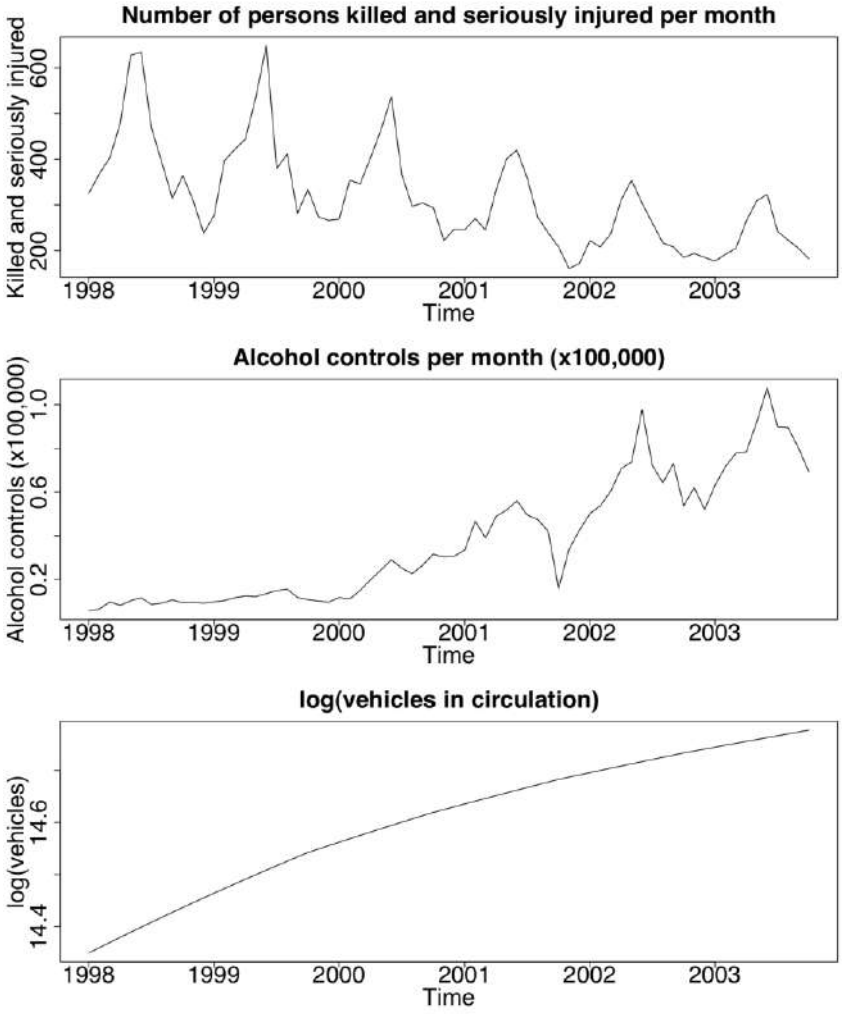

Figure 1. Dataset overview

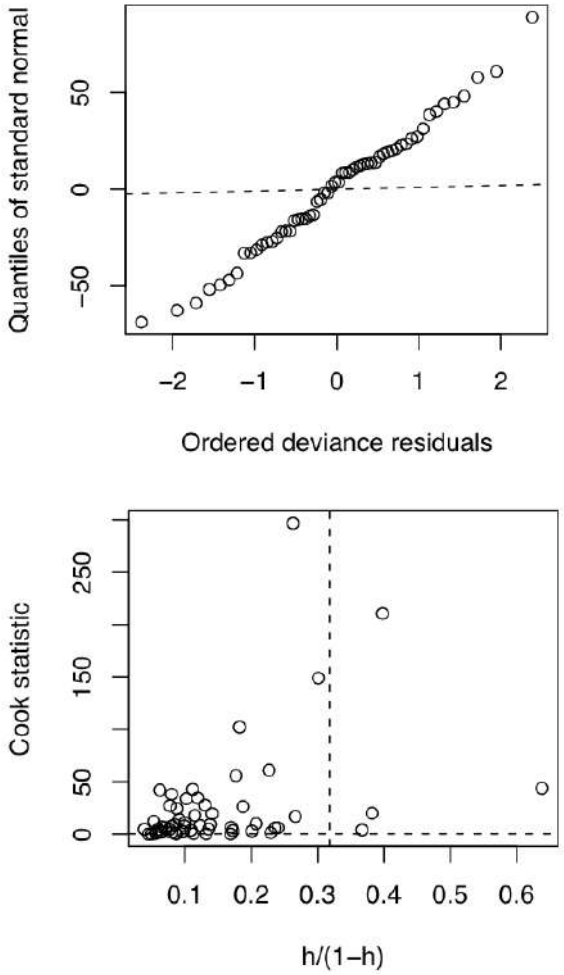

Figure 2. Model fit diagnostic plots (Gaussian)

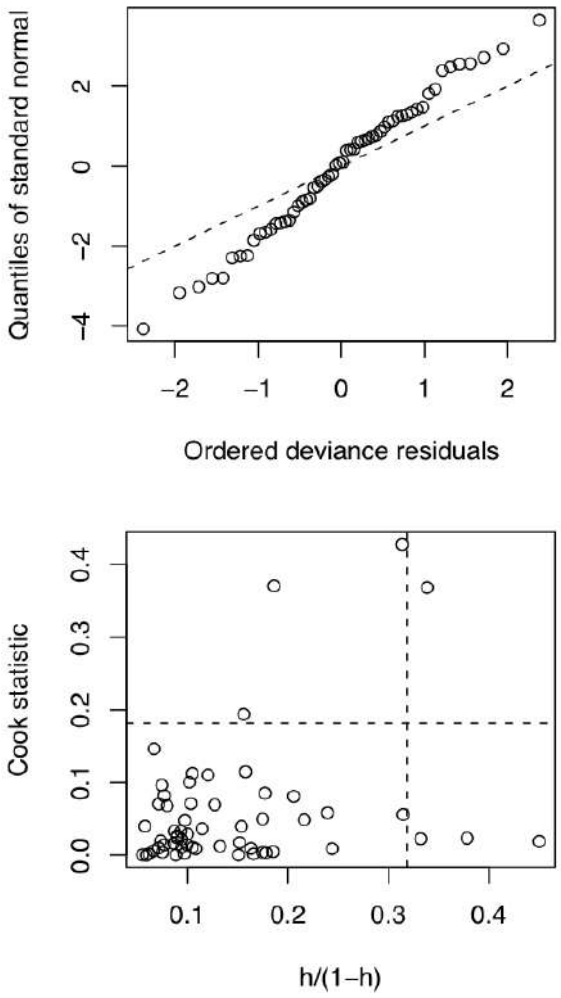

Figure 3. Model fit diagnostic plots (Poisson)

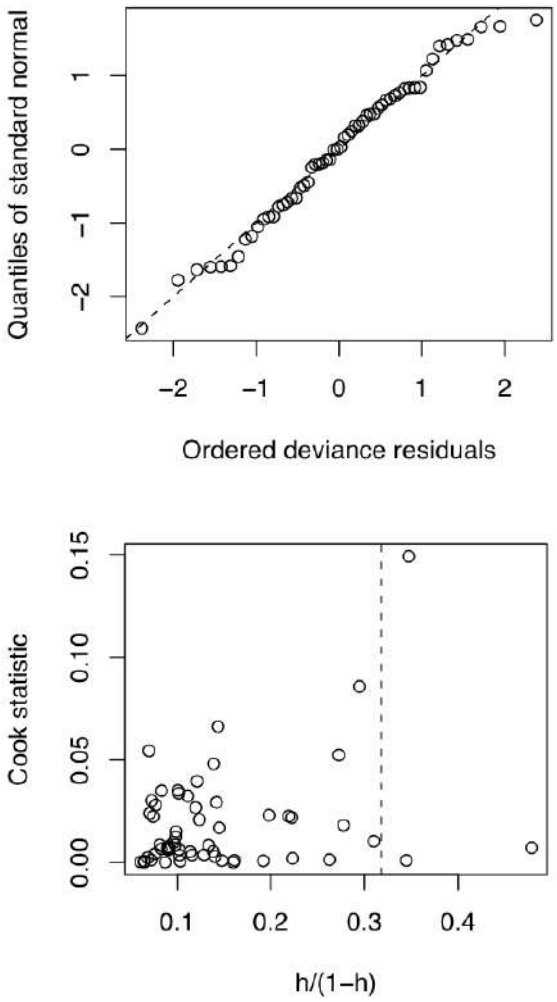

Figure 4. Model fit diagnostic plots (Quasi-Poisson)

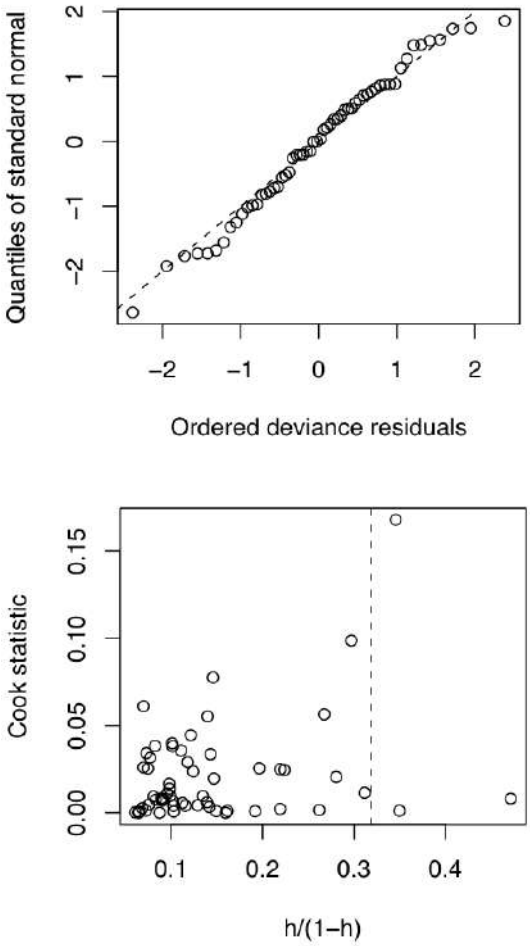

Figure 5. Model fit diagnostic plots (Negative binomial)

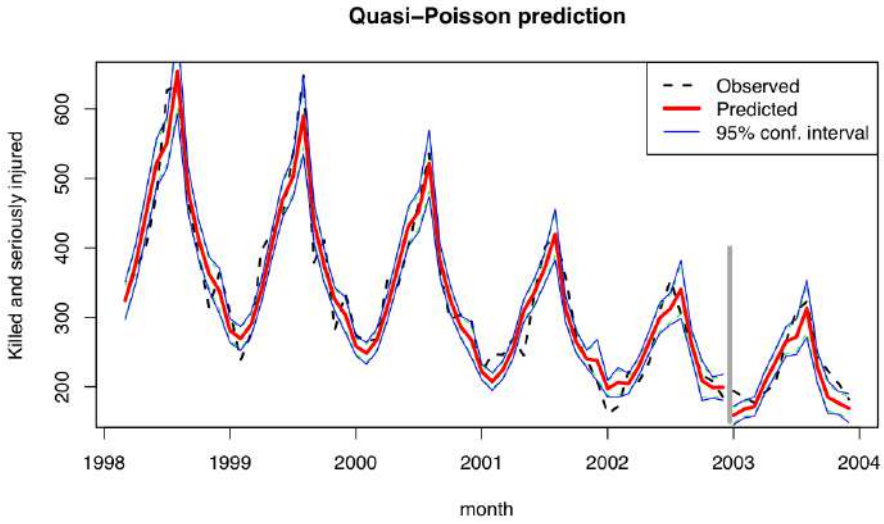

Figure 6. Quasi-Poisson model predictions

List of Tables

Table 1. Estimation results

Table 1. Estimation results

		Normal	
Coefficient	Estimate	Std. error	t-value
Intercept	-7.9608	0.1175	-67.763
Trend	-0.0154	0.0017	-9.054
August dummy	0.1995	0.0355	5.628
sin(pi*Month/6)	-0.2279	0.0215	-10.580
sin(pi*Month/12)	-0.5326	0.1826	-2.917
cos(pi*Month/6)	-0.4434	0.0781	-5.674
Laggedx2 alcohol controls (x100,000)	-0.2949	0.1481	-1.992
Null deviance:		*1 077 519*	*(57 d.o.f.)*
Residual deviance:		*52 892*	*(51 d.o.f.)*
		Poisson	
Coefficient	Estimate	Std. error	z-value
Intercept	-7.9881	0.0641	-124.548
Trend	-0.0157	0.0010	-15.921
August dummy	0.1919	0.0241	7.963
sin(pi*Month/6)	-0.2229	0.0123	-18.162
sin(pi*Month/12)	-0.4859	0.0985	-4.932
cos(pi*Month/6)	-0.4214	0.0430	-9.803
Laggedx2 alcohol controls (x100,000)	-0.2629	0.0821	-3.201
Null deviance:		*3 127.47*	*(57 d.o.f.)*
Residual deviance:		*151.11*	*(51 d.o.f.)*
		Quasi-Poisson	
Coefficient	Estimate	Std. error	z-value
Intercept	-8.0038	0.1066	-75.068
Trend	-0.0159	0.0017	-9.470
August dummy	0.1838	0.0466	3.949
sin(pi*Month/6)	-0.2206	0.0212	-10.427
sin(pi*Month/12)	-0.4582	0.1623	-2.824
cos(pi*Month/6)	-0.4087	0.0718	-5.692
Laggedx2 alcohol controls (x100,000)	-0.2410	0.1368	-1.761
Null deviance:		*1 004.67*	*(57 d.o.f.)*
Residual deviance:		*51.38*	*(51 d.o.f.)*
		Negative binomial	
Coefficient	Estimate	Std. error	z-value
Intercept	-8.0027	0.1007	-79.434
Trend	-0.0159	0.0016	-10.022
August dummy	0.1843	0.0436	4.229
sin(pi*Month/6)	-0.2208	0.0199	-11.071
sin(pi*Month/12)	-0.4602	0.1534	-2.999
cos(pi*Month/6)	-0.4096	0.0678	-6.038
Laggedx2 alcohol controls (x100,000)	-0.2425	0.1293	-1.875
Null deviance:		*1 183.74*	*(57 d.o.f.)*
Residual deviance:		*58.07*	*(51 d.o.f.)*