Studying the effect of weather conditions on daily crash counts using a discrete time-series model

BRIJS, Tom; KARLIS, Dimitris & WETS, Geert (2008) Studying the effect of weather conditions on daily crash counts using a discrete time-series model. In: ACCIDENT ANALYSIS AND PREVENTION, 40(3). p. 1180-1190.

DOI: 10.1016/j.aap.2008.01.001
Handle: http://hdl.handle.net/1942/8234

Studying the Effect of Weather Conditions on Daily Crash Counts Using a Discrete Time Series Model

Tom Brijs[†1], Dimitris Karlis[‡] and Geert Wets[†]

[†]Transportation Research Institute

Hasselt University

Wetenschapspark 5 — gebouw 6, B-3590 Diepenbeek, BELGIUM

Fax: +32 11 269199

{tom.brijs@uhasselt.be, geert.wets@uhasselt.be}

[‡] Department of Statistics

Athens University of Economics and Business

76, Patission Str., 10434, Athens, GREECE

Fax: +30 210 8203681

{karlis@aueb.gr}

Abstract

In previous research, significant effects of weather conditions on car crashes have been found. However, most studies use monthly or yearly data and only few studies are available analyzing the impact of weather conditions on daily car crash counts. Furthermore, the studies that are available on a daily level do not explicitly model the data in a time-series context, hereby ignoring the temporal serial correlation that may be present in the data. In this paper, we introduce an Integer Autoregressive model for modelling count data with time interdependencies. The model is applied to daily car crash data, metereological data and traffic exposure data from the Netherlands aiming at examining the risk impact of weather conditions on the observed counts. The results show that several assumptions related to the effect of weather conditions on crash counts are found to be significant in the data and that if serial temporal correlation is not accounted for in the model, this may produce biased results.

Keywords: Crashes; Accident analysis; Weather conditions; INAR

[1]corresponding author

1 Introduction

The last few years, road accidents statistics are the subject of increased interest both on the part of policy makers and academia. The objective is to better understand the complexity of factors that are related to road accidents in order to take corrective actions to remedy this situation. In this context, the modelling of crashes over time has obtained considerable attention by researchers in the past. For instance, several researchers have analyzed the effect of policies, economic climate and social conditions on the year-to-year changes in crash risk (Chang and Graham, 1993; Oppe, 1991). Other researchers have looked at month-to-month changes in accident levels (Van den Bossche *et al.*, 2005, 2004; Keeler, 1994; Fridstrøm and Ingebrigtsen, 1991). However, there are only few studies that have looked at changes in crash counts at a more disaggregate level. For instance, Levine *et al.* (1995a, 1995b) and Jones *et al.* (1991) studied daily changes, whilst Ceder and Livneh (1982) examined hourly fluctuations in crashes. Both approaches, high-level or low-level data aggregation, have advantages and disadvantages. While changes in crash counts on a highly aggregated level can be explained by structural changes, they cannot easily pick-up patterns of seasonality or weather effects. In contrast, the lower the level of aggregation, the more it is possible to study the effects of weather conditions, traffic volume, holidays etc. on changes in crash counts. Several authors have therefore warned for biases being introduced by modelling crash counts at high levels of aggregation (Golob *et al.*, 1990; Jovanis and Chang, 1989). In this paper, we study the effects of weather conditions on daily crashes for 3 large cities in the Netherlands (Dordrecht, Haarlemmermeer and Utrecht) in the year 2001. The use of weather conditions is motivated by earlier research where significant influences of weather conditions on road crashes were found (e.g. Andrey and Knapper, 2003).

From a methodological perspective, a number of approaches have been suggested by researchers to model time-series crash count data. More specifically, serial correlation between successive daily crash counts, i.e. autocorrelation, is reported as an important challenge for all accident models (Levine *et al.*, 1995; Fridstrøm *et al.*, 1995, 1991). For instance, Miaou and Lord (2003), Shankar *et al.* (1998) and Fridstrøm *et al.* (1995) use a Negative Binomial (NB) model to account implicitly for temporal serial correlation. Ulfarsson and Shankar (2003) use the Negative Multinomial (NM) model to predict the number of median crossover crashes using a multi-year panel of cross-sectional roadway data with roadway section-specific serial correlation across time.

However, the above models do not explicitly take into account the large and significant auto-correlation that is present in the data. Although, according to Fridstrøm *et al.* (1995), this has probably little effect on the statistical consistency of the coefficient estimates, they mention that it produces standard estimates that are too optimistic and thus not taking account of autocorrelation presents a potentially serious source of inefficiency in the modelling of cross-section/time-series

data. In response to these problems, we therefore present in this paper, a first-order autoregressive (AR1) time-series model for Poisson distributed data (see section 2) and compare it to some of the classical models found in the literature. The Poisson AR(1) model was first developed by Al-Osh and Alzaid (1987) and McKenzie (1985). Joe (1996) later generalized the approach. Weather effects in our model are easily incorporated as covariates via a link function as in standard GLM models.

The remaining of the paper proceeds as follows: in section 2, a detailed description of the INAR model is given. The EM algorithm to estimate the parameters of the model is described in section 3. Section 4 provides a description of the data. Section 5 contains information on the model formulation and estimation on our data. In section 6, detailed results are given. Finally, concluding remarks and some limitations of the research can be found in section 7 and 8.

2 Integer Autoregressive Models

Starting from the well-known simple AR(1) model for continuous data, we assume that $X_t = \phi X_{t-1} + \epsilon_t$, where $|\phi| < 1$ and $\epsilon_t \sim N(0, \sigma^2)$ independently. In other words, the current observation at time t depends for some part on the previous observation at time $t-1$. This model, while suitable for continuous random variables, cannot be used directly for discrete data. However, models that capture the same idea, but suitable for count data, can be also constructed. McKenzie (1985) and Al-Osh and Alzaid (1987) defined an analogous process for discrete data, called the Integer-valued autoregressive (INAR) process as follows:

Definition: A sequence of random variables $\{X_t\}$ is an $INAR(1)$ process if it satisfies a differential equation of the form

$$X_t = \alpha \circ X_{t-1} + R_t, \quad t = 1, 2, \ldots \tag{1}$$

where R_t is a sequence of uncorrelated non-negative integer-valued random variables having mean μ and finite variance σ^2 and X_0 represents an initial value of the process while the operator " $\circ$ " denotes the binomial thinning operator defined by

$$\alpha \circ X = \sum_{t=1}^{X} Y_t,$$

where Y_t are Bernoulli random variables with $P(Y_t = 1) = \alpha = 1 - P(Y_t = 0)$, $\alpha \in [0, 1]$. One can easily see that the binomial operator replaces the multiplication used for the normal time series autoregressive model so as to ensure that only integer values will occur. This implies that the Poisson AR model can be interpreted as a birth and death process, see Ross (1983, Section 5.3). Each individual at time $t-1$, has probability α of continuing to be alive at time t, and at each time t, the number of births R_t follows a Poisson distribution with mean μ.

Thus, conditional on X, $\alpha \circ X$ is a binomial random variable, where X denotes the number of trials and α denotes the probability of success in every trial. The term R_t is referred to as the *innovation term* and must be independent of $\alpha \circ X_{t-1}$ and follows any discrete distribution (in order for X_t to be counts).

This model belongs to a more general family of autoregressive models discussed in Grunwald *et al.* (2000). The basic ingredient of the INAR model is that it assumes that the realization of the process at time t is composed by two parts, the first one clearly relates to the previous observation, while the second one is independent and depends only on the current time point. Although it is possible to incorporate higher-order lags into the model, we do not pursue them here since it has been shown that their interpretation is not straightforward (see Jin-Guan and Yuan, 1991). Therefore, in this paper we will confine ourselves to the first-order case.

The mean and variance of a stationary INAR(1) process are constants given by the formulae

$$\mu_X = E(X_t) = \frac{\mu_R}{1-\alpha} \quad \text{and} \quad \sigma_X^2 = Var(X_t) = \frac{\alpha\mu_R + \sigma_R^2}{1-\alpha^2}, \tag{2}$$

where μ_R and σ_R^2 are respectively the (assumed finite) mean and variance of the i.i.d. innovations. The auto-covariance function of a stationary INAR(1) process $\{X_t\}_{t\in Z}$ is given by the formula

$$\gamma_X(k) = Cov(X_t, X_{t-k}) = \alpha^{|k|}\sigma_X^2 \quad , \quad k \in Z. \tag{3}$$

From the covariance function, it is easy to obtain the autocorrelation function $\rho(k)$ as follows:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \alpha^{|k|}.$$

Thus, the autocorrelation function $\rho(k)$ decays exponentially with lag k and for $k = 1$ we obtain that the parameter α represents the correlation between successive time points. By specifying the distributional form of the innovation term (R_t), a large number of different models can arise. The simplest and most common choice is to assume a Poisson distribution for the innovation term R_t. However, generalizations of the basic INAR model can be based on either other distributional forms for R_t, e.g. McKenzie (1986) or by replacing the binomial thinning operator with other kind of operators based on similar arguments (e.g. Al-Zaid and Al-Osh, 1993).

The simple Poisson INAR model can be extended to a INAR Poisson regression model by adding covariates to both the innovation term R_t and/or the autocorrelation parameter α. The model then takes the form

$$
\begin{aligned}
X_t &= \alpha_t \circ X_{t-1} + R_t \\
R_t &\sim Poisson(\lambda_t) \\
\log \lambda_t &= \mathbf{z}_t' \beta \\
\log \left(\frac{\alpha_t}{1-\alpha_t} \right) &= \mathbf{w}_t' \gamma,
\end{aligned} \tag{4}
$$

for $t = 1, \ldots, T$ where $\mathbf{z}_t$ and $\mathbf{w}_t$ are vectors of covariates at time t while β and γ are the associated regression coefficients. Note that the covariates for the two parts of the model must not necessarily be the same.

The well-known Poisson regression model corresponds to the case when $\alpha_t = 0$ for all t and thus the INAR(1) model is a natural extension of the standard Poisson regression model when autocorrelation in time series counts is present. The model also assumes that the correlation between successive points (α_t) may depend on some variables, i.e. it is not constant across time.

Clearly, the above model offers great flexibility for modelling count data with serial correlation. Other models to cover the time series nature of discrete valued data can be found in MacDonald and Zucchini (1997). The models discussed include Markov chains, higher-order Markov chains, models based on mixtures and models based on the idea of thinning. Also the model of Zeger (1988) offers a framework for discrete time series models.

However, those models usually model the time dependency through the relationships of their parameters (so-called parameter-driven models). As a result, the interpretation of the autocorrelation structure is not so easy since they induce it in the mean process and not directly in the observations, like in the INAR model. On the other hand, the advantage of the parameter-driven approach is that it accounts for overdispersion in the model, which is not straightforward in the INAR model. In fact, the INAR model assumes Poisson marginal distributions and hence it does not allow for overdispersion. Some extensions to allow for overdispersion can be found in Franke and Sellingman (1993), Karlis and Xekalaki (2001) and Gourieroux and Jasiak (2004). However, when the overdispersion in the data is small, like in our case, the INAR(1) model suffices to describe the data and offers easy interpretation of the results. Furthermore, according to Böckenholt (1999), the INAR Poisson regression model uses a dependence structure that is more parsimonious requiring only a small number of parameters.

Finally, the interpretation of the model is also suitable for accident data. The current count is split in two parts, the one part ($\alpha_t \circ X_{t-1}$) reflecting common elements with previous counts, like infrastructure, and the second part (R_t) reflecting a random process that generates accidents. Indeed, for our accident data, where we deal with the daily number of crashes for 3 city regions, it is reasonable to assume correlation between successive crash counts as a result of a structural

underlying level of risk that is region-specific and which depends, for instance, also on the characteristics of the road infrastructure. Indeed, given all other influential factors (like differences in weather or exposure) unchanged, we may expect the number of crashes of the current day to be similar to the number of crashes of yesterday due to a certain underlying level of unsafety that is determined by the intrinsic safety level of the infrastructure (type of roads, length of the road network, existence of black spots, etc). However, additionally, the current observation also depends on day-to-day differences in e.g. weather, exposure, etc. that may influence the unsafety level on the current day (R_t).

3 Estimation

From the preceding formulation of the INAR model one can easily see that the conditional distribution of $X_t \mid X_{t-1} = x_{t-1}$ takes the form

$$P(x_t \mid x_{t-1}, \alpha_t, \lambda_t) = \sum_{k=0}^{s} \frac{\exp(-\lambda_t){\lambda_t}^{x_t-k}}{(x_t-k)!} \left(\begin{array}{c} x_{t-1} \\ k \end{array} \right) {\alpha_t}^k (1-\alpha_t)^{x_{t-1}-k},$$

where α_t and λ_t are defined previously and $s = \min(x_t, x_{t-1})$. This probability function is the convolution of a Poisson with a binomial random variable (see, Shumway and Gurland, 1960).

The likelihood for model defined in (4), conditional on some initial value X_0, takes the form

$$L(\theta) \quad = \quad \prod_{t=1}^{T} P(x_t \mid x_{t-1}, \alpha_t, \lambda_t)$$

where $\theta = (\beta, \gamma)$ denotes the vector of parameters. The likelihood is rather complicated as it involves multiple summations and, hence, direct maximization of the likelihood is not easy. ML estimation for the model including covariates has been discussed in Böckenholt (1999). He proposed a Newton-Raphson approach for maximizing the likelihood. For the model without covariates, see the contributions by Al-Osh and Al-Zaid (1987), Ronning and Jung (1992) and Freeland and McCabe (2002). Karlis and Xekalaki (1999) provided an EM algorithm for the simple Poisson INAR model. We now extend this algorithm to the INAR Poisson regression model by including covariates.

The EM algorithm has had a tremendous influence on recent statistical practice as it can provide ML estimates to a broad range of problems that either containing missing values or that can be considered as containing missing values. For our formulation of the model, we can rewrite the observation at point t as $X_t = Y_t + R_t$ where $Y_t = \alpha_t \circ X_{t-1}$. In fact, we have observed data X_t while we cannot observe the latent variables Y_t and R_t. Note that if we could observe those values, then the estimation of the complete data (Y_t, R_t) would be straightforward as it comprises simple MLE in GLM models. Recall that $Y_t \sim Binomial(X_{t-1}, a_t)$ while $R_t \sim Poisson(\lambda_t)$.

The EM algorithm proceeds by estimating the unobserved data by their conditional expectations given the data and the current values of the parameters and then it maximizes the complete data likelihood using the expectations of the unobserved data taken at the previous step. The algorithm has some interesting properties like monotonic, but slow, convergence, parameters always in the admissible range etc. Multiple runs are suggested in order to ensure that the global maximum has been located. More details on the EM algorithm can be found in McLachlan and Krishnan (1997).

In our case, the algorithm has to be constructed so as to estimate, at the E-step, the conditional expectations of Y_t and R_t given the data and the current values of the estimates and to maximize, at the M-step, the complete likelihood. The latter is equivalent to maximizing the likelihood of a standard GLM model for the binomial distribution and the likelihood of a GLM model for the Poisson distribution. Statistical packages now offer procedures to fit those models. Hence the algorithm can be described as follows.

- *E-step*: Using the current values of the estimates, say $\theta^{old} = (\beta^{old}, \gamma^{old})$, calculate

$$\begin{aligned} s_t &= E(R_t \mid x_t, x_{t-1}, \theta^{old}) \\ &= \sum_{z=0}^{\infty} z P(R_t = z \mid x_t, x_{t-1}, \theta^{old}) \\ &= \sum_{z=0}^{\infty} z \frac{P(R_t=z)P(Y_t=x_t-z)}{P(x_t|x_{t-1},\alpha_t^{old},\lambda_t^{old})} \end{aligned}$$

  where $P(R_t = z)$ and $P(Y_t = x_t - z)$ are the probability functions of a Poisson and a binomial distribution respectively using the convolution representation discussed above, defined in the appropriate range. After some algebraic manipulation we can derive the relatively simpler formula

$$s_t = \lambda_t^{old} \frac{P(x_t-1|x_{t-1},a_t^{old},\lambda_t^{old})}{P(x_t|x_{t-1},\alpha_t^{old},\lambda_t^{old})},$$

  for $t = 1, \ldots, T$, where according to the model $\lambda_t^{old} = \exp\left(\mathbf{z}_t \beta^{old}\right)$ and $\alpha_t^{old} = \exp(\mathbf{w}_t \gamma^{old})/(1+\exp(\mathbf{w}_t \gamma^{old}))$.

  The conditional expectation of Y_t given the data and the current values of the estimates can be determined by simple subtraction, as

$$c_t = E(Y_t|x_t, x_{t-1}, \theta^{old}) = x_t - s_t.$$

- *M-Step*: Update the parameters in θ by fitting two GLM models. Namely, update β by fitting a Poisson regression model with response variables c_t and design matrix $\mathbf{z}$, while γ can be updated by fitting a binomial logit model with response s_t and design matrix $\mathbf{w}$.

- Stop iterating when some convergence criterion is satisfied, otherwise, go back to the E-step.

The above algorithm has all the pros and cons of the standard EM. Initial values for β can be retrieved by fitting a simple Poisson GLM model to the data. This algorithm was extensively used in our data analysis and we did not face any problems.

4 Data Description

This study is based on the daily crash counts that were obtained from the major roads covered by the surface of 3 big cities (Utrecht, Dordrecht and Haarlemmermeer) in the Netherlands in the year 2001. The cities were selected based on two criteria. Firstly, their proximity to some national weather station such that accurate daily weather conditions for each city could be obtained. Secondly, the cities were selected so that they are far enough apart in order to prevent that weather conditions would be identical for the different sites for too many of the observations.

Information on daily traffic exposure in 2001 was obtained from the Dutch Ministry of Transport. More specifically, for each city region, daily vehicle counts were obtained for each road segment of the major road network based on loop detector data. Taking into account the length of the road segments, this enabled us to calculate the day-to-day total amount of vehicle kilometers driven on the major road network of each city region. Later on in this paper we will show that if information on daily traffic exposure is not be available for some reason, day-of-the-week dummies may also account quite well for the day-of-the-week variability in exposure and still produce consistent results for the weather effects. In fact, the use of dummies has also been proposed in other studies (see e.g. Martin, 2002; Levine *et al.*, 1995; Jones *et al.*, 1991; Tanner, 1967).

In any case, it is necessary to account for differences in exposure in the model in order to separate the direct effect from the indirect effect (through exposure) that weather may have on crashes. Since a measure of exposure is included in our model, the results in this paper will thus show the direct effect of weather on crashes, conditional on a certain level of exposure.

With respect to weather conditions, the daily weather observations were obtained from the Dutch National Metereological Institute. More specifically, the following variables were created from the data and considered for inclusion in the model. The choice of variables was based on previous research where they have demonstrated to be important/significant or at least hypothesized as being influential towards predicting the number of crashes. Note that the data are daily averages and thus they do not reflect instant weather conditions.

- *wind.* Variables related to wind velocity have been used by Lian *et al.* (1998), Levine *et al.* (1995) and Baker and Reynolds (1992). The literature shows that wind is usually not found to be significant, except for heavy storms and for large vehicles. Nevertheless, we use the prevailing wind direction in degrees 360=North, 180=South, 270=West, 0=calm/variable),

the daily mean wind speed in 0.1 m/s, the maximum hourly mean wind speed in 0.1 ms/s and the maximum wind gust in 0.1 m/s.

- *temperature.* Temperature has found to be important, especially in combination with snowfall or rain (e.g., Branas and Knudson, 2001; Brown and Baass, 1997; Fridstrøm *et al.*, 1995; Fridstrøm and Ingebrigtsen, 1991). We use the daily mean temperature in 0.1 degrees Celsius, the minimum temperature in 0.1 degrees Celsius and the maximum temperature in 0.1 degrees Celsius. However, since the same absolute temperature during summer and winter may have a different effect on crashes, we also created a relative temperature variable, being the deviation of the mean daily temperature from the monthly temperature, to cancel out potential seasonal effects. Finally, since the effect of temperatures may be nonlinear, the daily mean temperature was discretized into four non-overlapping intervals, i.e. $T < 0, \quad 0 \leq T < 10, \quad 10 \leq T < 20$ and $T \geq 20$. The latter also enables to treat temperatures below zero as a separate category.

- *sunshine.* The amount of sunshine was found to be an important variable in the prediction of crashes. For instance, in Fridstrøm *et al.* (1995), it was found that an extra hour of daylight between 7 A.M and 11 P.M decreased the the number of crashes in Norway by 4%. We use sunshine duration in 0.1 hour and percentage of maximum possible sunshine duration. The latter variable accounts for seasonal differences in the amount of daylight due to different sunrise and sunset hours. Furthermore, an additional dummy variable was created to account for sun dazzle effects. Typically, in Northern countries and during fall and the winter period, the sun is very low above the horizon during certain periods of the day causing crashes by drivers who get dazzled by the sun. This dummy variable takes the value of 1 if the month is between September and February, maximum possible sunshine duration is above 70% and cloud cover is less than 4, approximating in this way a bright day with a lot of sunshine during fall or the winter period and 0 otherwise.

- *precipitation.* Rainfall has found to be a significant predictor for road crashes in many studies (see e.g. Fridstrøm *et al.*, 1995; Levine *et al.*, 1995; Satterthwaite, 1976). We use precipitation duration in 0.1 hour and daily precipitation amount in 0.1 mm. Moreover, an additional variable was created that expresses the intensity of rain, calculated as the ratio of the precipitation amount divided by the precipitation duration. High values for this variable indicate heavy rains during small time periods. Also, a lagged variable was created indicating the number of days since it has last rained. In fact, it was hypothesized recently (Eisenberg, 2004) that the risk imposed by precipitation increases dramatically as the time since last precipitation increases. Finally, the amount of rainfall was also discretized into several non-overlapping intervals in addition to a binary variable indicating whether it has rained or not

during that day.

- *air pressure.* In Roer (1974) and Orne & Yang (1972), falling barometric pressure was found to produce a significant increase in crash rate. We use the daily mean surface air pressure in 0.1 hPa.

- *visibility.* We use minimum visibility (0=less than 100m, 1=100-200m, 2=200-300m,...,) and cloud cover in octants (9=sky invisible).

In addition, similar to Fridstrøm *et al.* (1995), we also introduced city-specific dummy variables to account for city-specific differences in the number of crashes not accounted for by weather or traffic exposure (e.g. different physical conditions of the road network).

5 Model Formulation and Estimation

Let us now formulate the model in a mathematical notation. The variable X_{it} denotes the number of crashes for site i at time t, with $i = 1, 2, 3$ and $t = 2, \ldots, 365$. The first observation X_{i1} for each site is considered as the initial value. We use a model of the form

$$
\begin{aligned}
X_{it} &= \alpha_{it} \circ X_{i,t-1} + R_{it} \\
R_{it} &\sim Poisson(\lambda_{it}) \\
\lambda_{it} &= exp\left(\mathbf{z}'_{it}\beta\right) \\
\log\left(\frac{\alpha_{it}}{1-\alpha_{it}}\right) &= \mathbf{w}'_{it}\gamma
\end{aligned}
$$

where $\mathbf{z}_{it}$ and $\mathbf{w}_{it}$ are vectors of parameters at time t for site i while β and γ are the associated regression coefficients. Our model assumes different α's for each site, because preliminary analysis showed that the sites have different autocorrelations. Thus vector $\mathbf{w}$ consists of dummy variables for the 3 different sites. No weather or exposure covariates were used for the α's in order to avoid confusion about the effect of the weather conditions and exposure on the mean crash count. Recall that for the INAR model the marginal mean equals $\lambda/(1-\alpha)$ and hence using the same covariates for both the nominator and the denominator can lead to results without simple and useful interpretation. In the next section, we do not report the estimates for the vector γ but the parameters α_j, $j = 1, 2, 3$ corresponding to the three different sites.

Note that in our model we estimate only one parameter for each weather effect and exposure for all cities. Clearly, one can assume different regression parameters for each city, i.e. an interaction of weather conditions and city. For example, we may assume that $\lambda_{it} = exp\left(\mathbf{z}'_{it}\beta_i\right)$, so changes in β_i show the interaction of the particular city to the weather parameters and exposure. Alternatively,

such effect can be incorporated into the model by using dummy variables to reflect different cities. Preliminary examination of the data did, however, not show such effect. Generalization to this case is straightforward but it will not be treated in this paper. We use dummy variables for the 3 sites in order to account for the different mean crash counts observed in the data, but the regression parameters are assumed constant across sites.

Finally, since there is significant multi-collinearity in the data, especially for those variables referring to the same weather characteristic (e.g. minimum and maximum temperature during the day) we adopted a stepwise selection procedure during model estimation. Details about this stepwise selection can be found in section 6.2.

6 Results

6.1 Preliminary Analysis

Figure 1 shows the crash number series and daily traffic exposure for the three sites. It is apparent that the mean daily crash count and exposure are different between the three sites (see also table 1), which clearly motivates the use of site-specific dummy variables in the model.

Table 1 shows some of the crash data characteristics for the 3 sites. Clearly, there are differences between the sites. Firstly, for all three sites the ratio of the variance to the mean is larger than 1 implying overdispersion relative to the simple Poisson distribution. An overdispersed INAR model, like the negative binomial regression INAR model could be used to account for the overdispersion. However, after fitting the INAR Poisson regression model, it turned out that the remaining overdispersion is no longer significant and for this reason there is no need to use the more complicated negative binomial model. The reason is that the covariates used for modelling the data explain the overdispersion to a large extent. Secondly, there is a large difference between the autocorrelation of the three sites. The autocorrelations reported are of the first order. Higher-order autocorrelations were not large, apart from the autocorrelation for lag=7, which is however not statistically significant, and which in some sense indicates the effect of the day. For this reason, we fitted different autocorrelation parameters for the three sites.

Figure 2 shows the time series for some of the weather variables. The columns correspond to a site and the rows to a weather variable. The four variables presented in the plot are the mean temperature, the precipitation duration, the daily precipitation amount and the mean wind speed. We used the same scale to enable a fair comparison between the sites. Although the general pattern for each variable is similar across the different sites, several differences between sites for the same day are observable. For this reason, we expect that the weather effect obtained through the

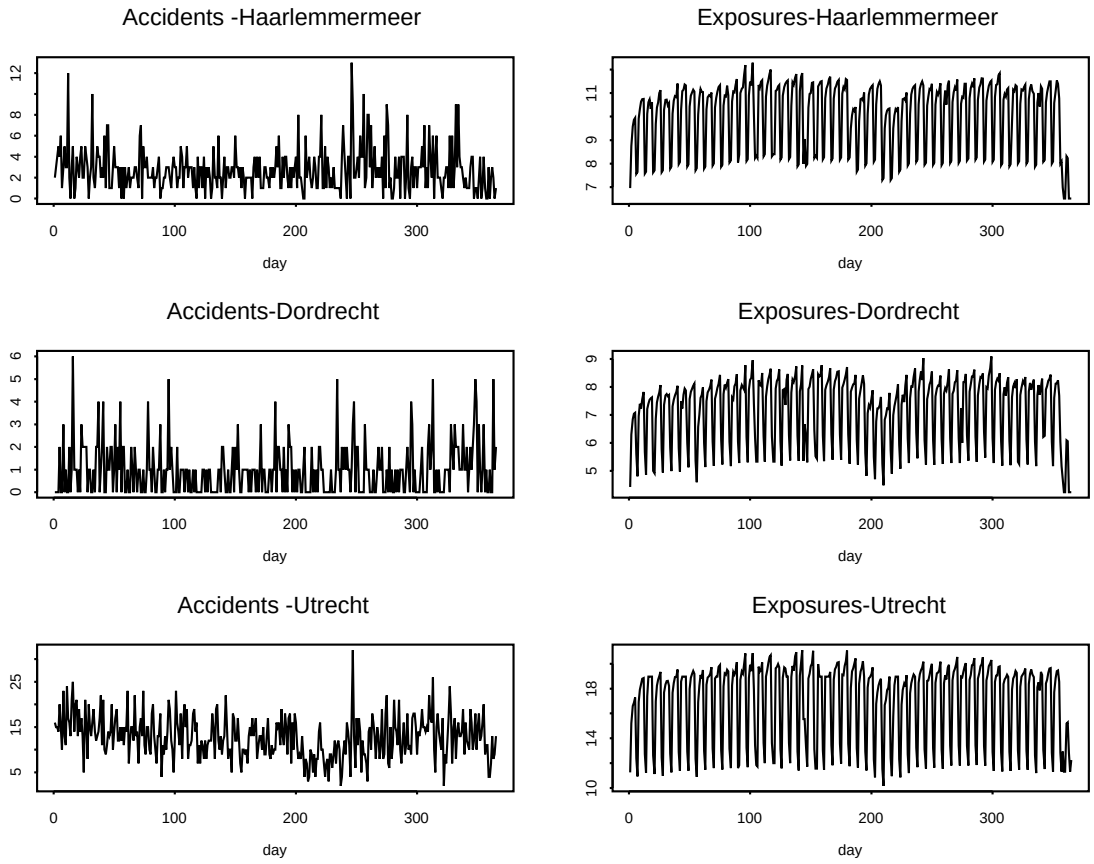

Figure 1: Crash counts and daily traffic exposure for the three sites (daily data for 2001)

analysis may have a more general interpretation since we have selected sites with varying weather conditions.

6.2 Model Estimation and Results

Table 2 shows the results after estimation. More precisely, three models were fitted. Weather variables are always included, but the way how traffic exposure is included in the model is different for each model. In the first model, day-of-the-week dummies are used to reflect differences in exposure in case actual traffic exposure data for each day of the year are not available. In the second model, actual traffic exposure is included as a covariate in the model, but no day-of-the-week dummies. Finally, in the third model, both the actual traffic exposures and the day-of-the-week dummies are included. Comparison of the results between the three models will indeed enable us to evaluate if day-of-the-week dummies are able to act as some kind of proxy variable for real traffic exposure, or not.

site	mean	variance	autocorrelation	variance/mean
Utrecht	2.747	4.227	0.0276	1.54
Dordrecht	0.950	1.239	0.0956	1.30
Haarlemmermeer	12.819	21.950	0.222	1.71

Table 1: Descriptive measures for the 3 series

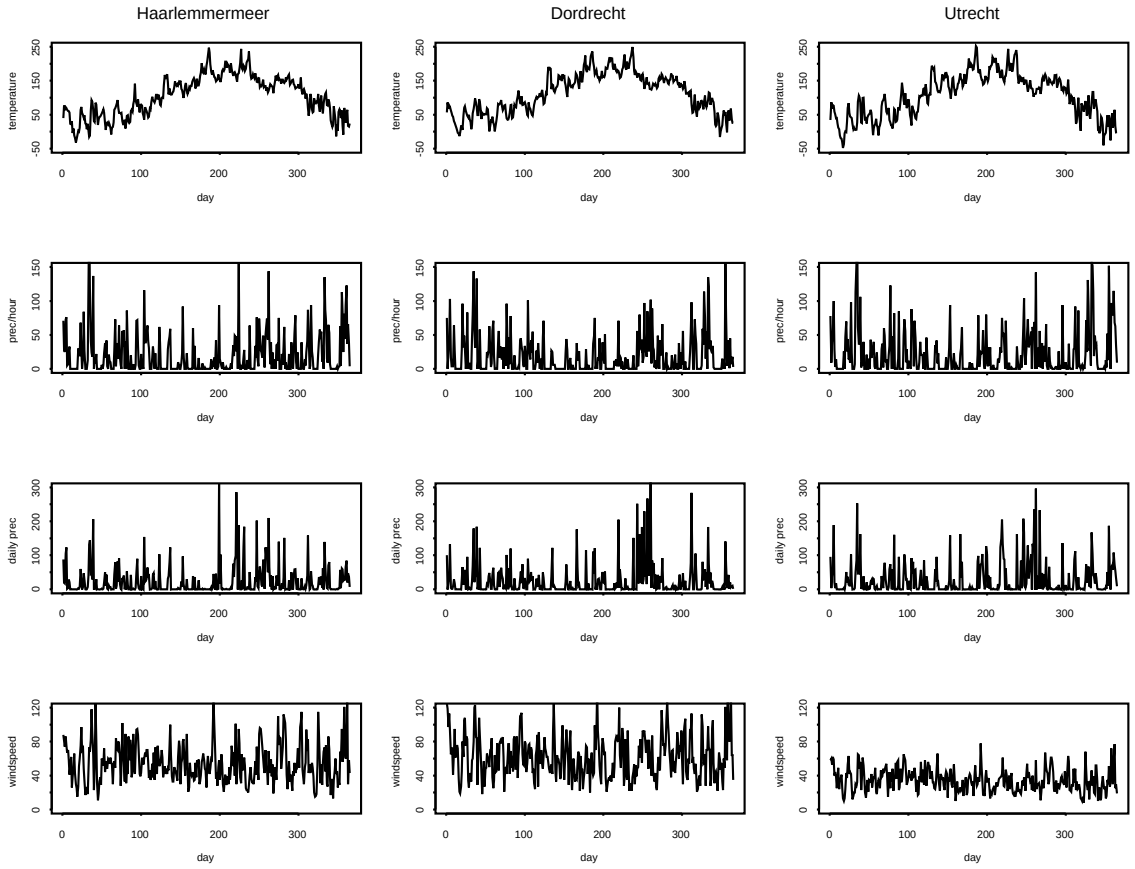

Figure 2: Plot of some of the weather variables for the three sites. There are notably different weather conditions

Given the large amount of explanatory variables available in this study and the problem of multi-collinearity associated with it, a stepwise model selection procedure was carried out. More specifically, the selection of the variables for the model was based on a forward search technique. In the first step, for each set of related variables (e.g. those related to wind, precipitation, temperature, etc.) one variable was selected from each set; the one with higher correlation with the response variable is selected in order to avoid multi-collinearity effects. After estimating this model and

	Model 1		**Model 2**		**Model 3**	
	coefficient (s.e.)	p-value	coefficient (s.e.)	p-value	coefficient (s.e.)	p-value
Constant	1.7048 (0.1072)	0.000	1.0084 (0.1520)	0.000	0.8375 (0.1942)	0.000
Utrecht	-1.3897 (0.0684)	0.000	-0.9964 (0.0900)	0.000	-0.9021 (0.1142)	0.000
Dordrecht	-2.5186 (0.0852)	0.000	-1.9518 (0.1140)	0.000	-1.8300 (0.1531)	0.000
Mean temperature		< 0.001		< 0.001		< 0.001
< 0	0.5238 (0.1135)	< 0.001	0.5742 (0.1120)	0.000	0.5666 (0.1141)	0.000
$[0, 10]$	0.3214 (0.0826)	< 0.001	0.3253 (0.0810)	0.000	0.3191 (0.0827)	0.000
$[10, 20]$	0.2516 (0.0798)	0.000	0.2353 (0.0780)	0.003	0.2160 (0.0801)	0.007
> 20	0	-	0	-	0	-
Dev of mean temp	0.0010 (0.0006)	0.081	0.0012 (0.0010)	0.042	0.0012 (0.0006)	0.0366
Precipitation duration	0.0027 (0.0005)	0.000	0.0029 (0.0000)	0.000	0.0033 (0.0005)	0.000
Intensity of rain	0.0321 (0.0143)	0.025	0.0304 (0.0140)	0.033	0.0344 (0.0144)	0.0168
Sun dazzle	0.1948 (0.0790)	0.013	0.1921 (0.0780)	0.014	0.1815 (0.0791)	0.0217
% max. possible sunsh. dur.	0.0012 (0.0006)	0.068	0.0011 (0.0010)	0.084	0.0013 (0.0006)	0.0403
Day of the week		0.000	-	-	-	0.001
Monday	0.3448 (0.0586)	0.000	-	-	0.0144 (0.0801)	0.8569
Tuesday	0.4467 (0.0574)	0.000	-	-	0.0452 (0.0880)	0.6077
Wednesday	0.2659 (0.0590)	0.000	-	-	-0.1559 (0.0926)	0.0922
Thursday	0.2886 (0.0587)	0.000	-	-	-0.1527 (0.0942)	0.1050
Friday	0.4364 (0.0570)	0.000	-	-	-0.0107 (0.0933)	0.9091
Saturday	0.0812 (0.0623)	0.192	-	-	0.0030 (0.0640)	0.9625
Sunday	0	-	-	-	-	-
Exposure	-	-	0.0581 (0.0060)	0.000	0.0687 (0.0117)	0.000
Autocorrelation parameters						
Utrecht	0 (0.0375)	1	0.0 (0.037)	1	0 (0.0371)	1
Dordrecht	0.0759 (0.0432)	0.078	0.0689 (0.044)	0.116	0.0720 (0.0438)	0.1001
Haarlemmermeer	0.1403 (0.0391)	0.003	0.1223 (0.039)	0.002	0.1402 (0.0410)	0.001

Table 2: Results based on the fitted model INAR regression model

evaluating the significance of each of the included variables, we added covariates to the model in additional steps by finding in each step the one that improves most the likelihood when added. Note that our EM algorithm provides an efficient tool for fitting models with similar structure since if good initial values are available, the algorithm converges quite fast. This implies that a large number of models were fitted. In several cases, in order to add flexibility to the model, we moved from simple linear relationships of the variable to the logarithm of the response variable by fitting non-linear relationships and/or discretized versions (see also the discussion in section 4).

Table 2 shows the results of the final models. The reported standard errors are asymptotic standard errors based on differentiating twice the loglikelihood. For the categorical variables, the reported statistic is the likelihood ratio test statistic measuring the improvement on the loglikelihood while adding this categorical variables. The reported p-values are based on the well known χ^2

distribution of the test statistic. For the other parameters, the reported p-values are based on the standard normal approximation of the t-values reported. Comments for all the variables included in the model follow:

- *exposure.* The first model shows that the proxies for exposure (day-of-the-week effect) are highly significant. It is apparent from the table that weekdays are more dangerous than weekend days (Sunday being the reference day). The difference between Saturday and Sunday is, however, not significant. Tuesday (0.45) and Friday (0.44) are the most dangerous days of the week in terms of the number of crashes. In fact, from an overall perspective, table 2 shows that the variable 'day of the week' is a highly significant variable and thus it should not be removed from the model. These results are consistent with findings in earlier research (e.g. Levine *et al.*, 1995a, 1995b) where differences between weekdays and weekend days were also found. However, the second model shows that when daily exposure data are included in the model as covariate, this variable is highly significant and when exposure increases one can expect a higher number of crashes. More interestingly, the results of model 3 show that once real traffic exposure information is included in the model as a covariate, the individual day-of-the-week dummies are no longer significant, although overall they are still significant in the model. Moreover, it appears that the results for the weather variables are only very slightly influenced by the way how exposure is being included in the models. This shows that day-of-the-week dummies can be seen as some kind of proxy variable for exposure when real traffic exposure information is missing.

- *precipitation.* Rainfall is also highly significant with respect to the number of crashes in all of the three models. The variable 'intensity of rain', being the ratio between daily precipitation amount and daily precipitation duration, is highly significant and shows that if the intensity of the rain increases, then this leads to a higher number of crashes. The same is true for the variable 'precipitation duration'. In fact, one can see a positive relationship between the number of hours of rainfall per day and the number of crashes. The interpretation for the coefficient of precipitation duration is that if the duration increases by one unit (0.1 hour per day) we expect an increase in the mean number of crashes by between 0.27% and 0.33%. However, we did not find any support for a lag-effect (Eisenberg, 2004), indicating that the risk imposed by precipitation increases as the time since last precipitation increases.

- *temperature.* The relationship between temperature and crashes is not straightforward. In fact, the relationship between the absolute temperature and the number of crashes is negative, highly significant and nonlinear. Indeed, relative to the base category (temperatures above 20), lower temperatures result in more crashes, with temperatures below zero being the most

significant. However, when looking at the deviance from the monthly mean temperature (significant at the 10% level), a different effect is observed. Indeed, when the daily mean temperature exceeds the monthly mean temperature, we expect more crashes. In other words, although on average a daily temperature of say 10 degrees leads to a higher number of crashes compared to temperatures above 20 degrees, the deviation from the monthly mean temperature may indicate the reverse. For instance, during a winter month when the monthly mean temperature is below 10 degrees, a temperature of 10 degrees produces more crashes, whereas during a hot summer month with monthly mean temperatures above 20 degrees, we expect less crashes.

- *sunshine.* In absolute terms, the amount (in hours) of sunshine was not found to be significant towards predicting the number of crashes. However, the relative amount of sunshine, as measured by the percentage of maximum possible sunshine duration, was found significant (at the 10% level) and positive. This means that, after correcting for seasonal differences in maximum possible sunshine duration, we can say that there is a positive effect between the number of hours of sunshine and the number of crashes. Finally, also sun dazzle during winter months was found to be highly significant and positive towards the number of crashes.

- *city-specific dummies.* The city-specific estimates are highly significant and indicate that relative to Haarlemmermeer, both Utrecht and Dordrecht show a lower number of crashes overall.

All other weather variables, discussed in section 4, such as air pressure, wind, sky visibility and lagged precipitation effects were not found to be significant in the model.

The last part of the Table 2 contains the autocorrelation parameters directly and not on a regression form. The table shows that the autocorrelation for Haarlemmermeer is the highest (around 0.14), whereas the correlation for Utrecht is not significant and equals 0. This indicates that autocorrelation is present in the data and should be taken into account for correct assessment of the effect of the variables. This can be shown when different competing models are compared against our fitted models in the next section.

6.3 Comparison with competing models

Table 3 presents a series of competing models. It can be seen that the simple Poisson regression model (ignoring autocorrelation) is the worst. Next, the negative binomial regression improves with respect to the Poisson regression model due to the small amount of overdispersion. For the three cases presented with different variables included in the model the INAR model gives

Model	Statistic	Poisson Regression	Negative Binomial Regression	INAR
Model 1	LL	-2246.496	-2243.034	-2238.961
	AIC	4526.992	4522.068	4515.922
Model 2	LL	-2239.109	-2234.123	-2232.950
	AIC	4502.218	4494.246	4493.900
Model 3	LL	-2227.122	-2221.087	-2220.190
	AIC	4490.244	4481.614	4480.380

Table 3: Comparison of different competing models

the best log-likelihood. Using likelihood ratio test statistics the INAR model is preferable to the simple Poisson model for each of the 3 models. This implies that the autocorrelation present in the data is significant and improves the fit of the models. We have also tested whether a common autocorrelation parameter could be used but the results support the use of different autocorrelations for each city.

A formal comparison with the negative binomial is not applicable as the models are not nested but one can see that the INAR models provide better log-likelihood while they have one more parameter. Using Akaike information criterion (AIC) to compare the models therefore shows that the INAR model is preferred over the negative binomial for all the three estimated models.

Table 4 shows the coefficients estimates and the associated standard errors of the regression parameters for the simple Poisson regression model and the INAR Poisson regression model considered for the model where both the exposure and the weekday dummies are included in the model.

Consider INAR1 model. The marginal mean is given as $\lambda/(1-\alpha)$. The simple Poisson model assumes $\alpha = 0$ and thus the coefficients for each variable are not the true ones but they overestimate or underestimate the true relationship of each variable. The difference between the estimated parameters under the two models depends on the amount of autocorrelation at each of the variables. For example, variables that are highly autocorrelated (as it is expected for weather data) in some sense contribute to the autocorrelation of the observed counts. So, including a covariance parameter changes more drastically their estimates.

For some variables, like the variable "sun dazzle" the relative difference is larger than 10% between the estimated coefficient of the simple Poisson model and an INAR Poisson model. This clearly demonstrates the importance for taking into account the autocorrelation in order to find the correct effect for each of the variables as otherwise the presence of autocorrelation can mask

such effects.

On the other hand, looking at the standard errors, one can see that the standard errors provided by the INAR model are larger. This is due to the fact that the INAR model takes into account both the serial correlation and the overdispersion and hence provides standard errors of more reasonable scale, while the simple Poisson model ignoring those aspects results to smaller but incorrect standard errors. The latter can be misleading as it leads to more significant results, thus, finding significant effects when they do not exist.

7 Conclusions

The effect of weather conditions on crashes has been a topic of debate for some years already and different studies tend to find conflicting results, depending on the granularity of the data, both in time and space, depending on the operationalization of the variables and finally depending on the kind of models being used. In this paper, we have argued that when autocorrelation is present in the data, a suitable statistical method should be used to model the time-dependencies in the data. To this end, we presented the Poisson Integer Autoregressive Model (INAR) for count data and used a number of covariates related to exposure and weather aspects (e.g. wind, temperature, sunshine, precipitation, air pressure, etc.) to estimate their impact on the number of crashes.

From the practical point of view, we showed that apart from exposure, weather effects do have an influence on the number of crashes but that depending on the operationalization of the variables, different effects can be found. More research is therefore needed, i.e. on more data sets, different variable operationalizations, different levels of granularity in time and space, to distinguish between the different impacts that weather may have on crashes. Furthermore, these results can be used in dynamic traffic management, information campaigns, In case of inclement weather, measures can be taken temporarily and locally by means of roadside variable message signs or via onboard navigation systems, for example indicating a lower maximum speed or by warning for reduced visibility. It must be said, however, that the effectiveness of such warning systems is, to some extent, still an open debate (Al-Ghamdi, 2007; Andrey and Knapper, 2003). For certain regions with a significantly higher impact of a weather element, structural measures can be taken.

From the statistical modelling point of view, we presented the Poisson INAR regression model and an efficient EM algorithm to estimate the model. Moreover, we showed that significance tests based on the likelihood ratio test indicate that, for our data set, the INAR Poisson regression model outperforms the simple Poisson regression model, which shows that autocorrelation indeed matters. Based on the AIC information criterion, the Poisson INAR model also slightly outperforms the NB regression model. Moreover, we demonstrated two interesting features of the proposed INAR model.

	INAR		Poisson	
	coeff	st. err.	coeff	st.err
Constant	0.8375	0.1942	1.0834	0.1554
Utrecht	-0.9021	0.1142	-1.0921	0.0813
Dordrecht	-1.8300	0.1531	-1.9608	0.1171
Mean temperature				
< 0	0.5666	0.1141	0.5686	0.1016
$[0, 10]$	0.3190	0.0827	0.3334	0.0737
$[10, 20]$	0.2160	0.0801	0.2224	0.0715
> 20	-	-	-	-
Dev of mean temp	0.0012	0.0006	0.0011	0.0005
Precipitation duration	0.0033	0.0005	0.0030	0.0005
Intensity of rain	0.0344	0.0144	0.0306	0.0132
Sun dazzle	0.1815	0.0791	0.1610	0.0716
% max. possible sunsh. dur.	0.0013	0.0006	0.0012	0.0006
Day of the week	-	-	-	-
Monday	0.0144	0.0801	-0.0205	0.0734
Tuesday	0.0452	0.0880	0.0349	0.0810
Wednesday	-0.1559	0.0926	-0.1376	0.0843
Thursday	-0.1527	0.0942	-0.1445	0.0866
Friday	-0.0107	0.0933	-0.0176	0.0863
Saturday	0.0030	0.0640	0.0394	0.0550
Sunday	-	-	-	-
Exposure	0.0686	0.0117	0.0635	0.0104

Table 4: Comparison of the estimated covariate coefficients for the simple Poisson regression and the Poisson INAR model

Firstly, the estimated standard errors of the regression coefficients, when the serial correlation is ignored, underestimate the true ones, leading more often to incorrect significant results. Using the INAR models this serial correlation is taken into account and hence the standard errors are larger reflecting more correctly the uncertainty on the estimated parameters. Secondly, the model with covariates in the autocorrelation part, as the one fitted in the paper, leads to non-Poisson marginal distributions and hence it can describe the overdispersion present to the data. Thus, the proposed INAR model seems to correct together for overdispersion and autocorrelation issues producing more reliable standard errors for the parameters.

8 Limitations

Firstly, the use of climatological weather data instead of using crash records to describe weather conditions may introduce a measurement problem since weather conditions (like rainfall) may be very local. However, since we model the number of crashes on the level of a larger geographical area (i.e. a major city), we think that it is more efficient to use data from a nearby weather station. Furthermore, at the time these data were collected, information from RWIS sites were not available unfortunately. Probably, this would have added even more detail to this study.

Secondly, our model does not distinguish between different types of crashes (fatal, severe, slight) and precipitation (snow, rain, hail). In fact, earlier research showed that some of the weather effects may have a different impact with respect to the type of injury. However, since the number of injuries of different types are not independent from each other, they should be studied preferably within a multivariate model, which is not straightforward. Furthermore, it introduces much smaller count numbers and thus may introduce additional complexity during model estimation. This will be the subject for future research.

Thirdly, in the present paper we used discretized versions of the covariates (e.g. for temperature) while non-parametric function could have been used, like splines. However, we think that such functions could create problems in the interpretability of the model and its computational stability.

Finally, our model directly fits serial autocorrelation by considering a time series model. In the literature there are several other models that could be used to fit correlated data, in an implicit way like random effects models. We do not use such model and consider only a simple time series model to explicitly fit the serial correlation.

References

Al-Ghamdi, A.S. (2007). Experimental evaluation of fog warning system. *Accident Analysis and Prevention*, **39(6)**, 1065-1072.

Al-Osh, M.A. and Al-Zaid, A.A. (1987). First Order Integer Valued Autoregressive Process. *Journal of Time Series Analysis*, **8**, 261-275.

Al-Zaid, A.A. and Al-Osh, M.A (1993). Some Autoregressive Moving Average Processes with Generalized Poisson Marginal Distributions. *Annals of the Institute Statistics Mathematics*, **45**, 223-232.

Andrey, J., and Knapper, C.K. (2003). *Weather and Transportation in Canada*. Department of Geography publication series, no. 55. ISBN: 0-921083-65-3

Baker, C.J. and Reynolds, S. (1992). Wind-induced Accidents of Road Vehicles. *Accident Analysis and Prevention*, **24(6)**, 559-575.

Branas, C. and Knudson, M. (2001). Helmet Laws and Motorcycle Rider Death Rates. *Accident Analysis and Prevention*, **33(5)**, 641-648.

Brown, B. and Baass, K. (1997). Seasonal Variation in Frequencies and Rates of Highway Accidents as Function of Severity. *Transportation Research Record* **1581**, 59-65.

Ceder, A. and Livneh, M. (1982). Relationships Between Road Accidents and Hourly Traffic Flow. *Accident Analysis and Prevention*, **14(1)**, 19-34.

Chang, B-H. and Graham, J.D. (1993). A New Method for Making Interstate Comparisons of Highway Fatality Rates. *Accident Analysis and Prevention*, **25(1)**, 85-90.

Eisenberg, D. (2004). The mixed effects of precipitation on traffic accidents. *Accident Analysis and Prevention*, **36(4)**, 637-647.

Freeland, K. and McCabe, B. (2002). Estimation and Testing of the Poisson Autoregression Model of Order 1 R. *Actuarial Research Clearing House*.

Franke, J. and Seligmann, T. (1993). Conditional maximum likelihood estimates for INAR(1) processes and their application to modeling epileptic seizure counts. *In "Developments in Time Series Analysis",ed. T. Subba Rao, Chapman and Hall*, 310-330.

Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R. and Krogsgard Thomsen, L. (1995). Measuring the Contribution of Randomness, Exposure, Weather, and Daylight to the Variation in Road Accident Counts. *Accident Analysis and Prevention*, **27(1)**, 1-20.

Fridstrøm, L. and Ingebrigtsen, S. (1991). An Aggregate Accident Model Based on Pooled, Regional Time-Series Data. *Accident Analysis and Prevention*, **23(5)**, 363-378.

Golob, T.F., Recker, W.W. and Levine, D.W. (1990). Safety of Freeway Median High Occupancy Vehicle Lanes: A Comparison of Aggregate and Dissagregate Analyses. *Accident Analysis and Prevention*, **22(1)**, 19-34.

Gourieroux, C. and Jasiak, J. (2004). Heterogeneous INAR(1) model with application to car insurance. *Insurance: Mathematics and Economics*, 34, 177–192

Grunwald, G.K., Hyndman, R.J., Tedesco, L. and Tweedie, R.L. (2000). Non-Gaussian Conditional Linear AR(1) Models. *Australian and New Zealand Journal of Statistics*, **42**, 479-495.

Jin-Guan, D. and Yuan, L. (1991). The Integer-Valued Autoregressive (INAR(p)) Model. *Journal of Time Series Analysis*, **12**, 129-142.

Joe, H. (1996). Time Series Models with Univariate Margins in the Convolution-Closed Infinitely Divisible Class. *Journal of Applied Probability*, **33**, 664-677.

Jones, B., Janssen, L. and Mannering, F. (1991). Analysis of the Frequency and Duration of Freeway Accidents in Seattle. *Accident Analysis and Prevention*, **23(4)**, 239-255.

Jovanis, P.P. and Chang, H-L. (1989). Disaggregate Model of Highway Accident Occurrence Using Survival Theory. *Accident Analysis and Prevention*, **21(5)**, 445-458.

Karlis, D. and Xekalaki, E. (1999). Maximum Likelihood Estimation for the Poisson-Binomial Distribution via the EM Algorithm with Applications to Discrete Time Series Models. *Technical Report No 62*, Department of Statistics, Athens University of Economics.

Karlis, D. and Xekalaki, E. (2001). ML Estimation For Integer Valued Time Series Models. In *HERCMA 2001 Proceeding, ed. E.A. Lipitakis*, pp 778-783.

Keeler, T.E. (1994). Highway Safety, Economic Behavior, and Driving Enforcement. *The American Economic Review*, **84(3)**, 684-693.

Levine, N., Kim, K.E. and Nitz, L.H. (1995a). Spatial Analysis of Honolulu Motor Vehicle Crashes: I. Spatial Patterns. *Accident Analysis and Prevention*, **27(5)**, 663-674.

Levine, N., Kim, K.E. and Nitz, L.H. (1995b). Daily Fluctuations in Honolulu Motor Vehicle Accidents. *Accident Analysis and Prevention*, **27(6)**, 785-796.

Lian, WL., Kyte, M., Kitchener, F. and Shannon, P. (1998). Effect of Environmental Factors on Driver Speed: A Case Study. *Transportation Research Record* **1635**, 155-161.

MacDonald, I.L. and Zucchini, Z. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman & Hall: New York.

Martin, J-L. (2002). Relationship Between Crash Rate and Hourly Traffic Flow on Interurban Motorways. *Accident Analysis and Prevention*, **34(5)**, 619-629.

McKenzie, E. (1985). Some Simple Models for Discrete Variable Time Series. *Water Resources Bulletin,* **21**, 645-650.

McKenzie, E. (1986). Autoregressive Moving-Average Processes with Negative Binomial and Geometric Marginal Distributions. *Advances in Applied Probability*, **18**, 679-695.

McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley: New York.

Miaou, S-P. and Lord, D. (2003). Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes. *Electronic Proceedings of the 82nd Transportation Research Board Annual Meeting*, Washington D.C., USA.

Oppe, S. (1991). Development of Traffic and Traffic Safety: Global Trends and Incidental Fluctuations. *Accident Analysis and Prevention*, **23(5)**, 413-422.

Orne, D.E. and Yang, A.H. (1972). An Investigation of Weather Factor Effects on Traffic Accidents. *Traffic Engineering*, **43**, 14-20.

Roer, P.O. (1974). Effects of Some Non-Transportation Factors on the Incidence and Severity of Traffic Accidents. *Log. Transportation Review*, **10**, 165-179.

Ronning, G. and Jung R. (1992). Estimation of a First-Order Autoregressive Process With Poisson Marginals For Count Data. *Advances in GLIM and Statistical Modelling*, Springer-Verlag: New York, 188-194.

Ross, S.M. (1983). *Stochastic Processes*. Wiley: New York.

Satterthwaite, S.P. (1976). An Assessment of Seasonal and Weather Effects on the Frequency of Road Accidents in California. *Accident Analysis and Prevention*, **8(3)**, 87-96.

Shankar, V.N., Albin, R.B., Milton, J.C. and Mannering, F.L. (1998). Evaluating Median Cross-Over Likelihoods with Clustered Accident Counts: An Empirical Inquiry Using the Random Effects Negative Binomial Model. *Transportation Research Record* **1635**, 44-48.

Shumway, R. and Gurland, J. (1960). Fitting the Poisson Binomial Distribution. *Biometrics*, **16**, 522-533.

Tanner, J.C. (1967). Casualty Rates at the Easter, Whitsun and August public holidays. *Ministry of Transport, Road Research Laboratory Report*, **LR74**.

Ulfarsson, G.F. and Shankar, V.N. (2003). An Accident Count Model Based on Multi-Year Cross-Sectional Roadway Data with Serial Correlation. *Electronic Proceedings of the 82nd Transportation Research Board Annual Meeting*, Washington DC, USA.

Van den Bossche, F., Wets, G. and Brijs, T. (2004). A Regression Model with ARMA Errors to Investigate the Frequency and Severity of Road Traffic Accidents. *Electronic Proceedings of the 83th Annual Meeting of the Transportation Research Board*, Washington D.C., USA.

Van den Bossche, F., Wets, G., and Brijs, T. (2005). Role of Exposure in Analysis of Road Accidents: A Belgian Case Study. *Transportation Research Record* **1908**, 96-103.

Zeger, S.L. (1988) A Regression Model for Time Series of Counts. *Biometrika*, **75**, 621-629.