STUDYING THE EFFECT OF WEATHER CONDITIONS ON DAILY CRASH COUNTS

Tom Brijs[1], Dimitris Karlis[2] and Geert Wets[1]

[1]Hasselt University - Transportation Research Institute
Wetenschapspark 5, 3590 DIEPENBEEK, Belgium
Phone: +32 11269155 Fax: +32 11269199
Email: tom.brijs@uhasselt.be, geert.wets@uhasselt.be

[2] Athens University of Economics and Business - Department of Statistics
Patission Str. 76, 10434, ATHENS, Greece
Phone: +30 2108203920 Fax: +30 210 8203681
Email: 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 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 and metereological 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 an appropriate statistical model should be used to account for the existing autocorrelation in the data.

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; 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). Therefore, 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 (see section 3).

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 autocorrelation 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. Section 3 provides a description of the data. Section 4 contains information on the model formulation. In section 5, detailed results are given. Finally, concluding remarks and some limitations of the research can be found in section 6 and 7.

2 INTEGER AUTOREGRESSIVE MODELS

Starting from the well-known simple AR(1) model for continuous data, we assume that $X_t = \phi X_{t-1} + \varepsilon_t$, where $|\phi| < 1$ and $\varepsilon_t \sim N(0, \sigma^2)$ 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

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

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 mimics 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).

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

$$\begin{aligned} X_t &= \alpha_t \circ X_{t-1} + R_t \\ R_t &\sim Poisson(\lambda_t) \\ \log \lambda_t &= \mathbf{z}_t' \beta \\ \log\left(\frac{\alpha_t}{1-\alpha_t}\right) &= \mathbf{w}_t' \gamma, \end{aligned} \tag{2}$$

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.

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, 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 depend on the

number of crashes of yesterday due to a certain 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 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 in order to obtain accurate daily weather conditions for each city. 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.

With respect to weather conditions, daily weather observations were obtained from the Dutch National Metereological Institute. More specifically, the following list variables were created from the data and considered for inclusion in the model. This was based on previous research where they have shown 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 windspeed in 0.1 m/s, the maximum hourly mean windspeed 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., 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. $T < 0$, $0 \leq T < 10$, $10 \leq T < 20$ 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).

Information on daily traffic exposure in 2001 was obtained from the Dutch Ministry of Transport. More specifically, for each city region, daily vehicle kilometers driven were calculated for each road segment of the major road network based on loop detector data. This enabled us to calculate the day-to-day total amount of vehicle kilometers driven for each city region. However, if information on daily traffic exposure is not be available, we show later in this paper that one can also include dummy variables in the model for the different days of the week in order to account for day-of-the-week variability in exposure (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 the model, the results in this paper will thus show the direct effect of weather on crashes. 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).

4 MODEL FORMULATION

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 had different autocorrelations. Thus vector $\mathbf{w}$ consists of dummy variables for the 3 different sites. No weather variables were used for the α's in order to avoid confusion about the effect of the weather conditions on the mean crash count. Indeed, one can show 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 our model assumes the same parameters related to weather conditions for all sites. Clearly, one can assume different regression parameters for each site, i.e. an interaction of weather conditions and site. For example, we may assume that $\lambda_{it} = exp(\mathbf{z}'_{it}\beta_i)$, so changes in β_i show the interaction of the particular site to those weather parameters. Alternatively, such effect can be incorporated into the model by using dummy variables to reflect different sites. Preliminary examination of the data did, however, not show any such effect. Generalization to this case is straightforward and it will not 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 (see section 5.2).

5 RESULTS

5.1 Preliminary analysis

Table 1 shows some of the 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 constructed 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 1 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 windspeed. 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 analysis may have a more general interpretation since we have selected sites with varying weather conditions.

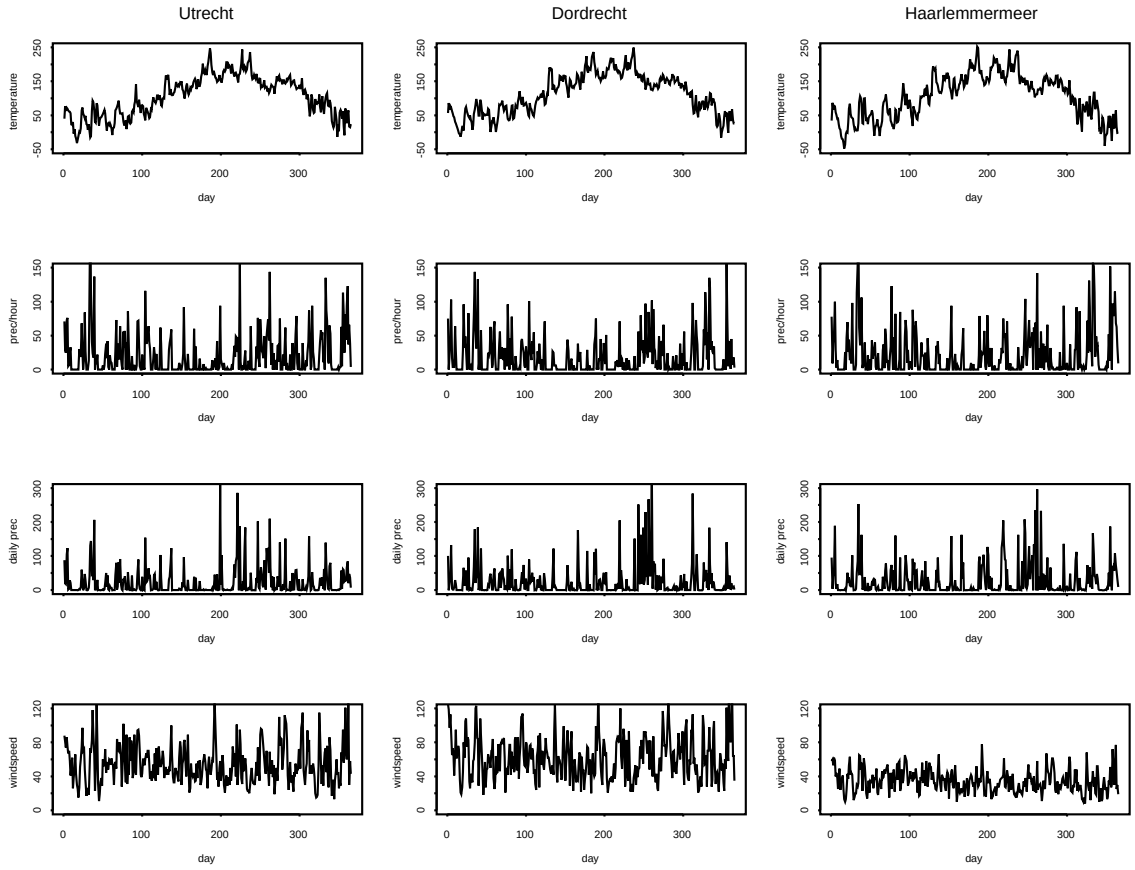

Figure 1: Plot of some of the weather variables for the three sites. There are notably different weather conditions

5.2 Estimation and results

Table 2 shows the results after estimation using an EM-type algorithm (Karlis and Xekalaki, 2001). More precisely, two models were fitted (see table 2): the first model using dummies to reflect day-of-the-week differences when exposure is not known, and the second model including exposure as an additional variable in the model (without the day-of-week dummies). 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 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

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

	Without exposure		**With exposure**	
	coefficient (s.e.)	p-value	coefficient (s.e.)	p-value
Constant	1.7048 (0.1072)	0.000	1.0084 (0.1520)	0.000
Utrecht	-1.3897 (0.0684)	0.000	-0.9964 (0.0900)	0.000
Dordrecht	-2.5186 (0.0852)	0.000	-1.9518 (0.1140)	0.000
Mean temperature		< 0.001		< 0.001
< 0	0.5238 (0.1135)	< 0.001	0.5742 (0.1120)	0.000
$[0, 10]$	0.3214 (0.0826)	< 0.001	0.3253 (0.0810)	0.000
$[10, 20]$	0.2516 (0.0798)	0.000	0.2353 (0.0780)	0.003
> 20	0	-	0	-
Dev of mean temp	0.0010 (0.0006)	0.081	0.0012 (0.0010)	0.042
Precipitation duration	0.0027 (0.0005)	0.000	0.0029 (0.0000)	0.000
Intensity of rain	0.0032 (0.0143)	0.025	0.0304 (0.0140)	0.033
Sun dazzle	0.1948 (0.0790)	0.013	0.1921 (0.0780)	0.014
% max. possible sunsh. dur.	0.0012 (0.0006)	0.068	0.0011 (0.0010)	0.084
Day of the week		0.000	-	-
Monday	0.3448 (0.0586)	0.000	-	-
Tuesday	0.4467 (0.0574)	0.000	-	-
Wednesday	0.2659 (0.0590)	0.000	-	-
Thursday	0.2886 (0.0587)	0.000	-	-
Friday	0.4364 (0.0570)	0.000	-	-
Saturday	0.0812 (0.0623)	0.192	-	-
Sunday	0	-	-	-
Exposure	-	-	0.0581 (0.0060)	0.000
Autocorrelation parameters				
Utrecht	0 (0.0375)	1	0.0 (0.037)	1
Dordrecht	0.0759 (0.0432)	0.078	0.0689 (0.044)	0.116
Haarlemmermeer	0.1403 (0.0391)	0.003	0.1223 (0.039)	0.002

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

Model	Log-likelihood
Poisson regression	-2246.496
Negative Binomial Regression	-2243.034
Poisson INAR regression without exposure	-2238.961
Poisson INAR regression with exposure	-2232.950

Table 3: Comparison of different competing models

and/or discretized versions (see also the discussion in section 3). Adding more explanatory variables did not show any statistically significant improvement according to the likelihood ratio test. From table 2, it becomes clear that both models (with and without exposure) show very similar results for the different weather variables. This is an important observation because it shows that when traffic exposure information is not available, the use of day-of-the-week dummies as proxy variables for exposure still provides valid results for the effect of weather conditions on road safety. 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. However, when daily exposure information is available, the second model shows that indeed this variable is highly significant and when exposure increases one can expect a higher number of crashes. 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.
- *precipitation*. Rainfall is also highly significant with respect to the number of crashes. 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) according to the first model we expect an increase in the mean number of crashes by 0.27%. However, we did not find any support for a lag-effect (Eisenberg, 2004), implying that the risk imposed by precipitation does not increase 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, 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 less crashes, whereas during a hot summer month with monthly mean temperatures above 20 degrees, we expect more 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 (-1.39) and Dordrecht (-2.52) show a lower number of accidents overall.

All other weather variables, discussed in section 3, 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. The table shows that the autocorrelation for Haarlemmermeer is the highest (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.

5.3 Comparison with competing models

Table 3 presents a series of competing models fitted to the same data (i.e. using the same weather covariates for all models) to allow for comparisons and hypothesis testing. 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. Furthermore, the INAR Poisson regression model without exposure is preferable to the Poisson regression model (LRT statistic is 15.07 with 3 degrees of freedom, p-value$=0.0015$). It is also slightly better than the negative binomial model (LRT is 8.146 with 2 degrees of freedom, p-value$=0.017$). However, the best model is the INAR Poisson regression model with exposure included since compared with the INAR Poisson regression model without exposure it has a better loglikelihood even with 5 parameters less.

6 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 methods being used. In this paper, we have shown that when autocorrelation is present in the data, suitable statistical methods should be used to model the time-dependencies in the data. To this end, we presented the Integer Autoregressive (INAR) Poisson Model for count data and used a number of covariates related to traffic exposure and different weather aspects (e.g. wind, temperature, sunshine, precipitation, air pressure, etc.) to estimate their impact on the number of crashes.

From the technical point of view, 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 and thus that autocorrelation matters. Furthermore, we showed that the model including daily traffic exposures outperforms the INAR Poisson regression model with day-of-the-week dummies, although the effect of the weather variables on daily crashes remains essentially unchanged.

From the practical point of view, we showed that weather effects indeed 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 datasets, 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, etc. In case of 'dangerous' weather, measures can be taken temporarily - by means of dynamic overhead traffic signs for example indicating a lower maximum speed - or geographically - for certain regions with a significantly higher impact of a weather element - structural measures can be taken.

7 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.

Finally, 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.

REFERENCES

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

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

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.

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.

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

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.

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.

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.

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.

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.

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. (2005). Role of Exposure in Analysis of Road Accidents: A Belgian Case Study. *Transportation Research Record* **1908**, 96-103.